MATH1318 - Time Series Analysis

Assignment 1 Project Report

Time Series Analysis on Thickness of Ozone layer data

By
Ayaz Aziz Mujawar (s3751555)

Abstract

The report aims to find the best fitting model to analyze and predict the thickness of the ozone layer and to propose the set of possible ARIMA(p,d,q) models. Two regression models namely Linear and Quadratic trend models were used to analyze the series of data. Overall the result indicates that the Quadratic model outperforms the Linear model but both models were not the best fit on the given dataset. It is recommended to use the ARIMA model on the datasets to get better results. Also for the second task, we proposed 5 sets of ARIMA(p,d,q) models using EACF and BIC tools.

Introduction

The ozone layer is the shield over the earth's stratosphere that absorbs most of the sun's radiations. The thickness of the ozone layer differs according to the region. Thickness refers to how much ozone is present in that section and reasons for the thickness variations might be due to atmospheric circulation patterns and solar intensity. Time Series Analysis techniques are the best way to measure and predict the changes in the thickness of the ozone layer over time.

Methodology

The dataset provided represents the changes in the thickness level of the ozone layer from 1927 - 2016 in doubson unit. The positive and negative values in the series represent an increase and decrease in the thickness level respectively. After reading the data into the data frame it is then converted into time series objects and then applied the time series plot on it. Observations of the plot are done based on trends, variance, seasonality, autocorrelation, and intervention. Based on that we had done our further analysis.

General Analysis

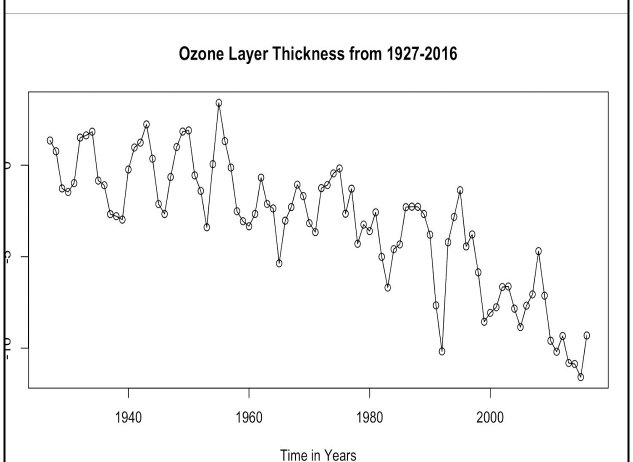

Fig: 1.1

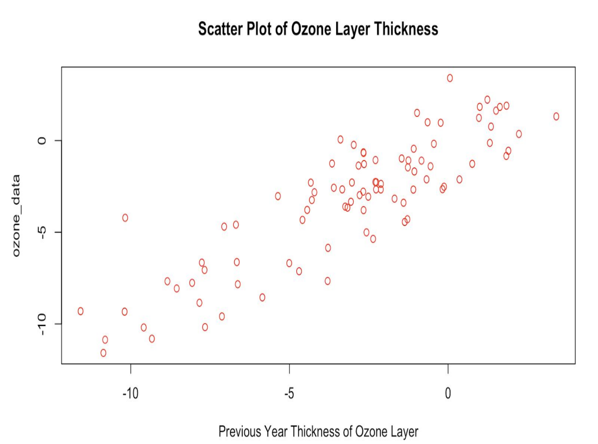

Fig: 1.2

From the figure(fig 1.1), We can observe that there seems to be a downward trend as well as a slight change in the value of mean over time. There is no obvious seasonality present in the graph. Also, a change in variance is not observed in the series. There seem to be fluctuations around the mean value and also successive points are observed through the time in the series which leads to moving average and Auto-Regressive behavior. There seems to be a minor intervention in the series during the year 1988 - 1992 because the emission of ozone-depleting substances from manmade and natural resources is weighted highest during that period up to 1.4 million tonnes. By applying the correlation analysis I found that there is a strong correlation between the yearly thickness of the ozone layer.

Modelling

Linear Model

A deterministic linear model is defined by $\mu t = \beta 0 + \beta 1t$ where $\beta 0$ = intercept and $\beta 1$ = slope of linear trend.

```
Call:
lm(formula = ozone_data ~ time(ozone_data))

Residuals:
    Min      1Q  Median      3Q     Max
-4.7165 -1.6687  0.0275  1.4726  4.7940

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      213.720155  16.257158   13.15   <2e-16 ***
time(ozone_data)  -0.110029   0.008245  -13.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.032 on 88 degrees of freedom
Multiple R-squared:  0.6693,    Adjusted R-squared:  0.6655
F-statistic: 178.1 on 1 and 88 DF,  p-value: < 2.2e-16
```

Fig: 1.3

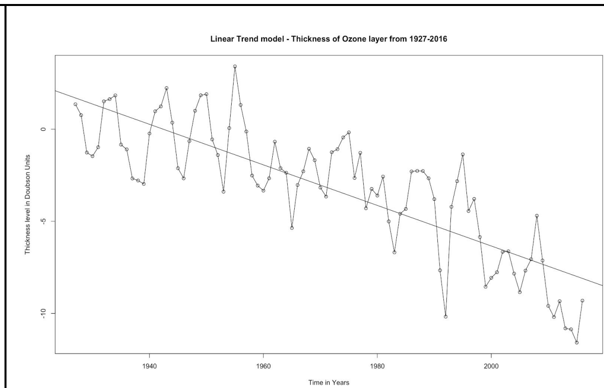

Fig: 1.4

From the statistics summary (Fig 1.3) of the linear model, we conclude that the estimation of slopes is -0.110 and the intercept is 213.720. Also, we found that the p-value is statistically significant as it is less than 0.05. I also look at the r^2 value(i.e. 0.669) which turns out to be partially significant and denoted that 66.9% of the variation in this series is explained by linear time trend. From Fig: 1.4 We can also observe the trend line which is plotted over the time series.

Residual Analysis for Linear trends Model

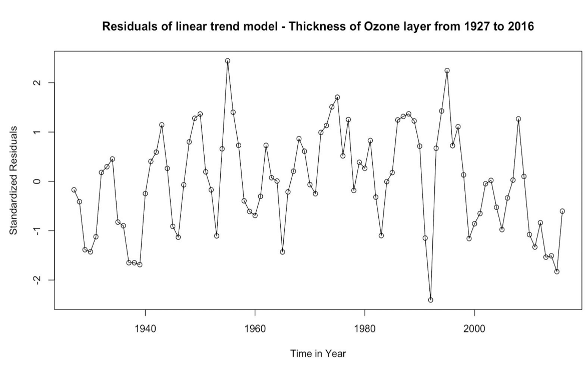

Fig: 1.5

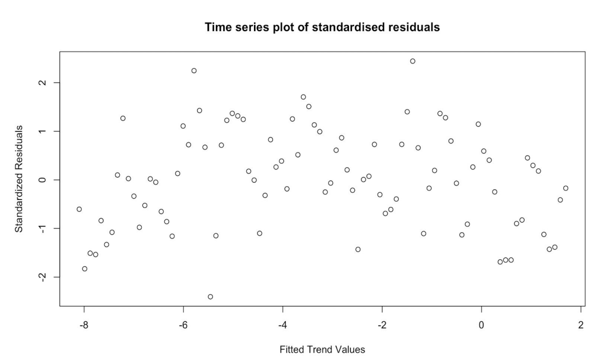

Fig: 1.6

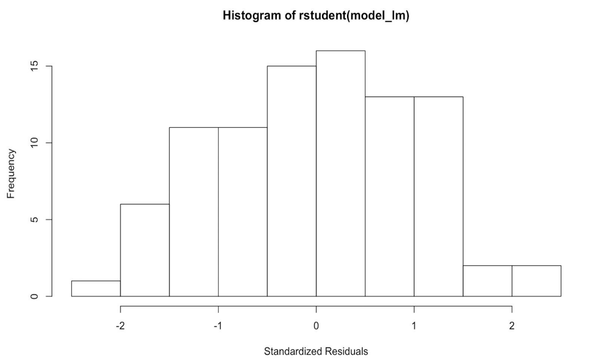

Fig: 1.7

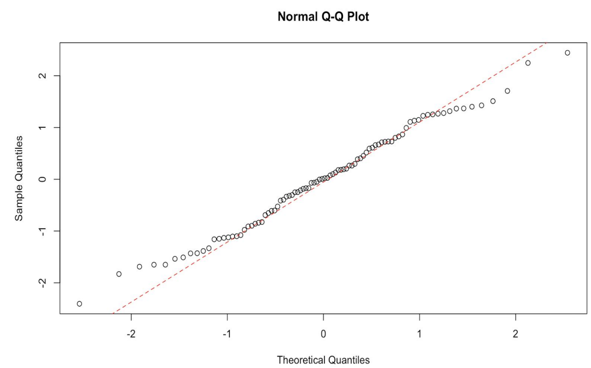

Fig: 1.8

```
Shapiro-Wilk normality test

data:  y
W = 0.98733, p-value = 0.5372
```

Fig: 1.9

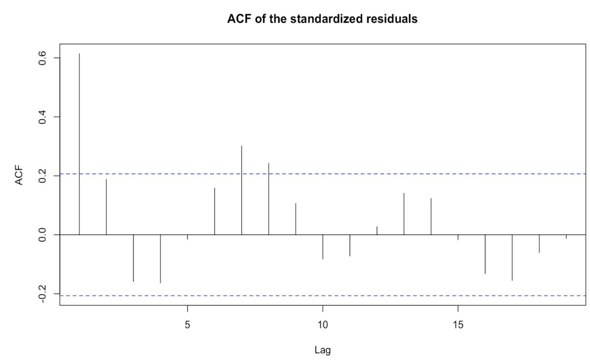

Fig: 1.10

In Fig. 1.5 There is a bit of departure in randomness in the residual plot. From Fig 1.6 we claim that there was no obvious trend in the residuals. Fig 1.7 of histogram appears to be somewhat symmetric as it tails off at both high and low end. From the QQ-Plot figure (fig 1.8) we can observe that the residual points are moving away from the low and high end of the straight line tail and also the points are not lying fully in the middle section of the straight line. So from this we can say that the straight line pattern doesn't capture the assumption of normality distribution of randomness in the model. From the Shapiro Wilk test(Fig: 1.9) we got the p-value of 0.5372, hence we fail to reject the null hypothesis and from that we state that the stochastic component of the model is normally distributed. The ACF values in Fig: 1.10 depicts that there is no white noise because the correlation values are higher than the confidence bounds at several lags.

Quadratic Model

The deterministic quadratic trend model is represented by - μt = β0 + β1t where β0 = intercept and β1 = slope of linear trend and β1 = Quadratic trend in time.

```
lm(formula = ozone_data ~ t + t2)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.1062 -1.2846 -0.0055  1.3379  4.2325 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.733e+03  1.232e+03  -4.654 1.16e-05 ***
t            5.924e+00  1.250e+00   4.739 8.30e-06 ***
t2          -1.530e-03  3.170e-04  -4.827 5.87e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.815 on 87 degrees of freedom
Multiple R-squared:  0.7391,	Adjusted R-squared:  0.7331 
F-statistic: 123.3 on 2 and 87 DF,  p-value: < 2.2e-16
```

Fig: 1.11

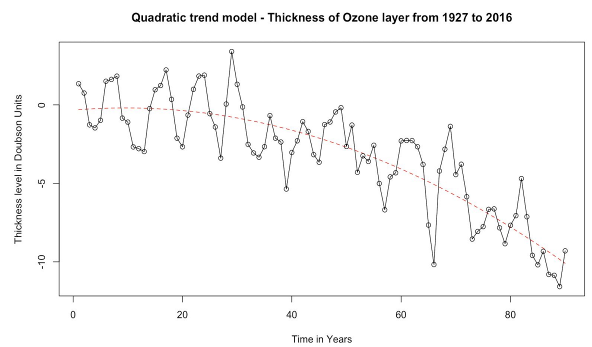

Fig: 1.12

From the summary statistics(Fig: 1.11) we conclude that the p-value in quadratic trend term found to be significant so we reject the null hypothesis and the R-squared term value is 0.7331 which is better than the Linear trend model. As shown in Fig: 1.12 the model is fitted and a quadratic trend line is plotted over the time series.

Residual Analysis for Quadratic trends Model

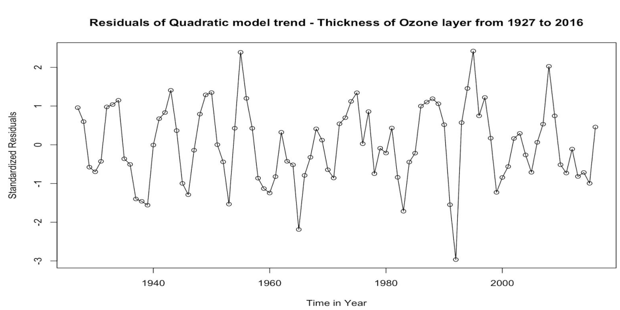

Fig: 1.13

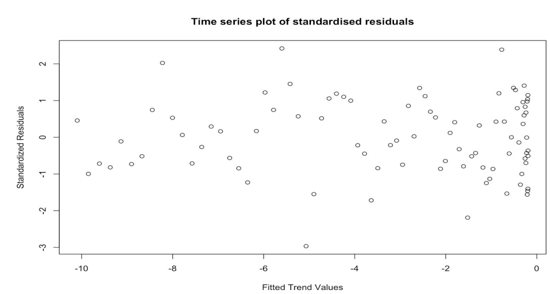

Fig: 1.14

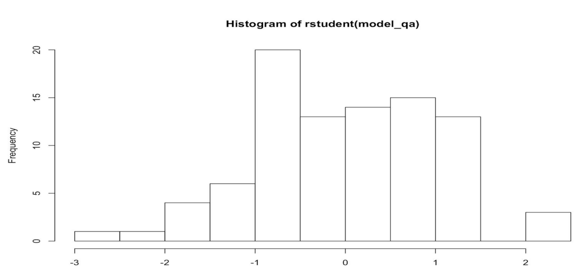

Fig: 1.15

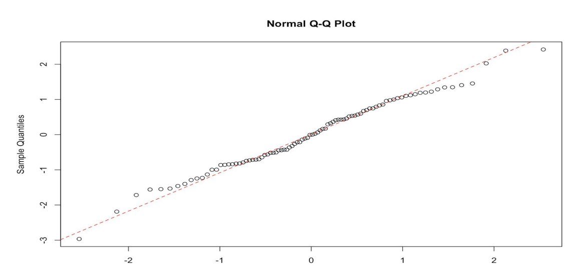

Fig: 1.16

```
        Shapiro-Wilk normality test

data:  y
W = 0.98889, p-value = 0.6493
```

Fig: 1.17

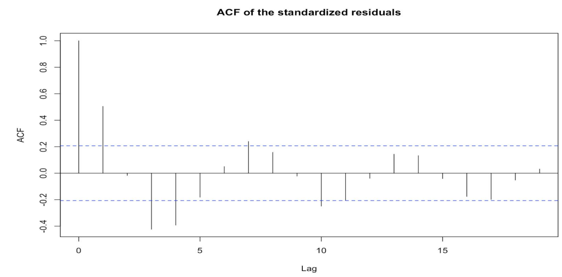

Fig: 1.18

The plot of residual over time does not look random because from the Fig: 1.13 the mean value does not look zero and also the variance is not constant. From the Fig: 1.14 (standardized residual graph) we observed that there are some variations in the residual points and also they don't indicate any pattern. We also observed that the residual points fail to fulfill the normality assumption as the histogram plot(Fig: 1.15) is not symmetric. From the qq-plot (Fig: 1.16) we indicated that the residuals points are almost falling on the straight line. Thus this pattern supports the assumption of normality distributed stochastic components in the model. The Shapiro-Wilk test(Fig: 1.17) gives us the p-value of 0.6493 which tells us not to reject the null hypothesis and conclude that the stochastic component of the mode is normally distributed. From Fig: 1.18 it justifies that there is a smoothness in the plot as well as the correlation values are higher than the confidence bound at certain lags. So it concludes that there is no white noise in the series.

Best Model

In comparison of both the models, the quadratic trend model outperforms the linear trend model based on coefficient of determination(R-Squared) as it better describes the variation in the series. However, both the models fail to suggest the best fit on the given series data due to lack of trend coverage in the series.

Forecasting Model

The forecasting of the time series for the next 5 years using the quadratic model shown in Fig: 1.19

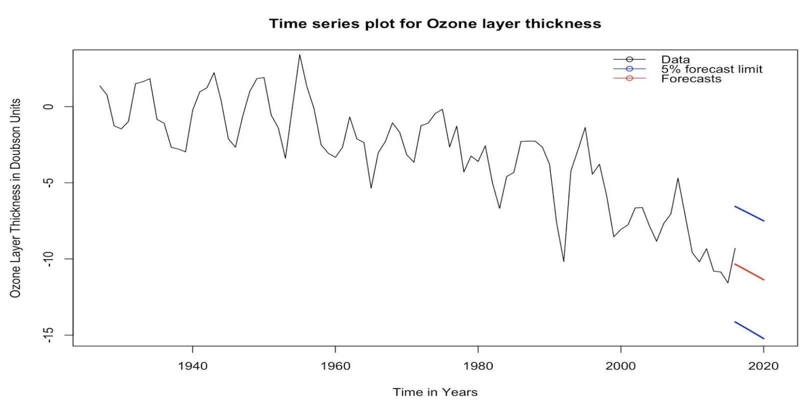

Fig: 1.19

From the forecasting model results, we can predict that there will be a continuous decrease in the thickness of the ozone layer for the next 5 years.

Task - 2

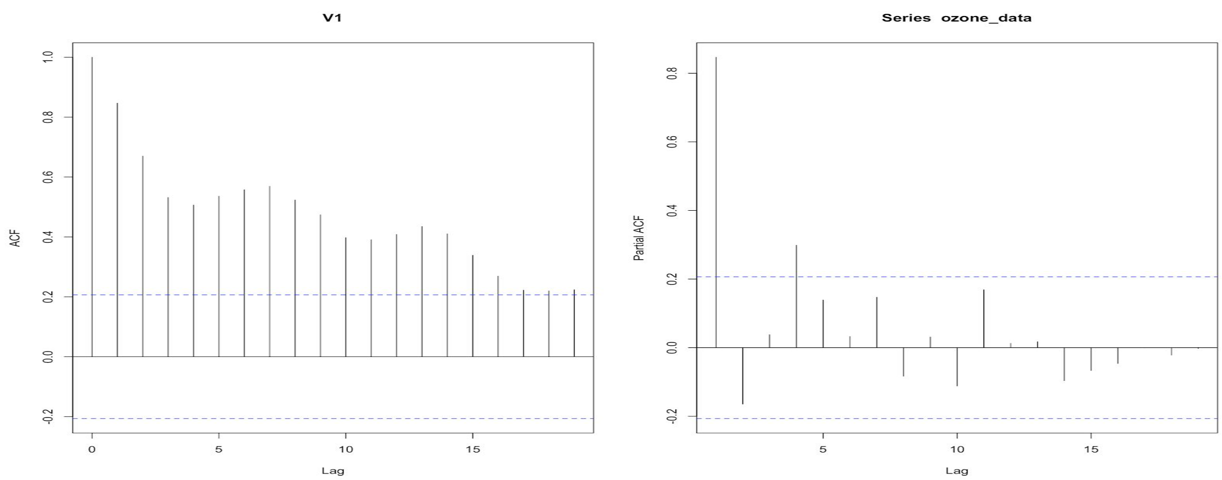

Fig: 1.20

From the above ACF and PACF plots(Fig: 1.20), In the ACF plot, we observed that there was a downward trend and also there might be seasonality in the plot. This denotes the existence of trends in the series and also the series was nonstationary. After applying adf test to the series we found the p-value of 0.08 from which we failed to reject the null hypothesis that states that the series is non-stationary.

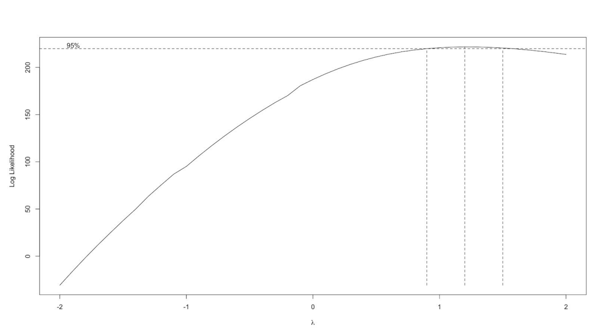

Fig: 1.21

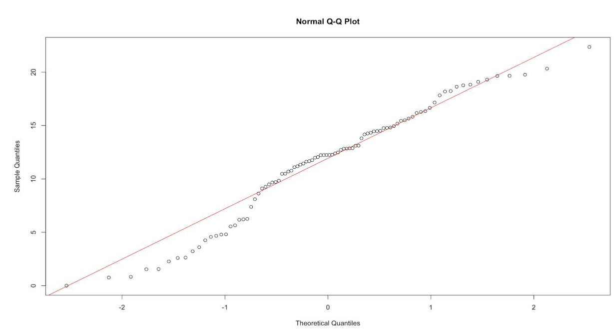

Fig: 1.22

To remove the trend from the data Box-Cox transformation(as shown in fig: 1.21) was applied to the series. But there was not much difference in the time series plot.

From fig 1.22 of qq-plot, we found that the dots were not aligned properly on the straight line and also the p-value from the Shapiro test was not significant (p-value<0.05). After applying box-cox we conclude that there is still a trend in the series.

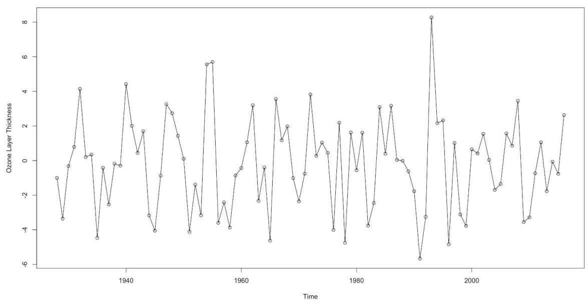

Fig: 1.23

```
Augmented Dickey-Fuller Test

data:  diff_ozone_data_BC
Dickey-Fuller = -7.2426, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

Fig: 1.24

Then we applied the first differentiation to the series and due to that, the series became stationary as shown in Fig: 1.23 and also there was a complete loss in the trend and some change in variance.

After that we again applied adf test(Fig: 1.24) and found a significant p-value of 0.01, So we reject the null hypothesis and state that the series is stationary.

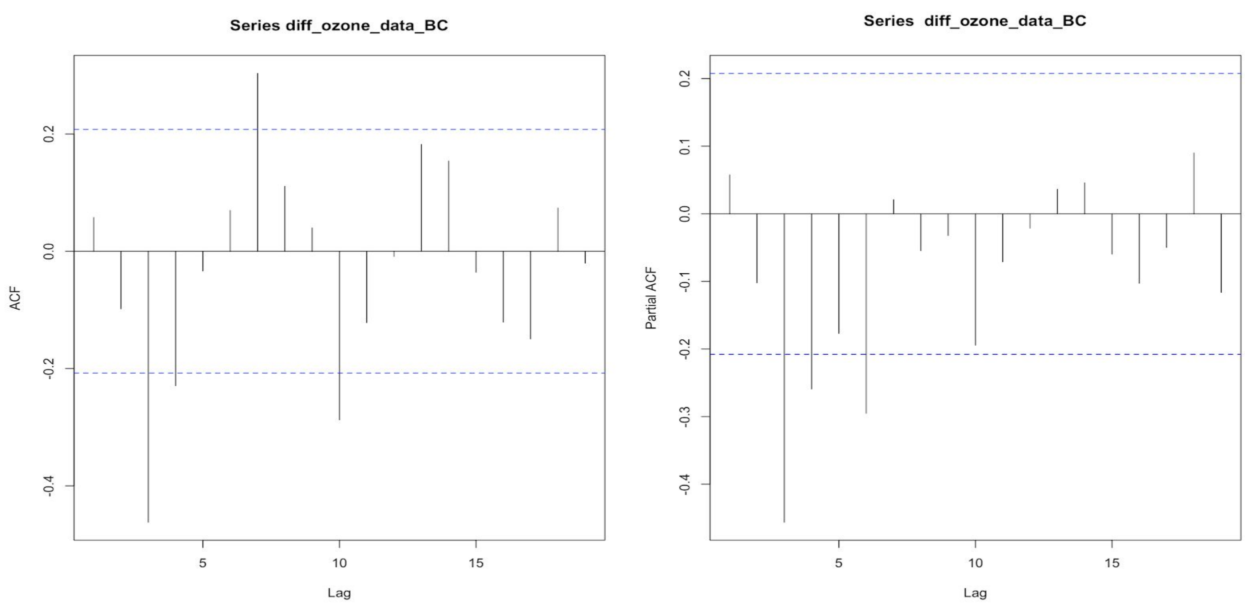

Fig 1.25

After checking again the ACF and PACF we found that there was a slow degradation in the ACF value and there were 3 significant lags in PACF value. So we can consider ARI(3,1) and small models like ARI(1,1), ARI(2,1) and so on.

AR/MA

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

Fig: 1.26

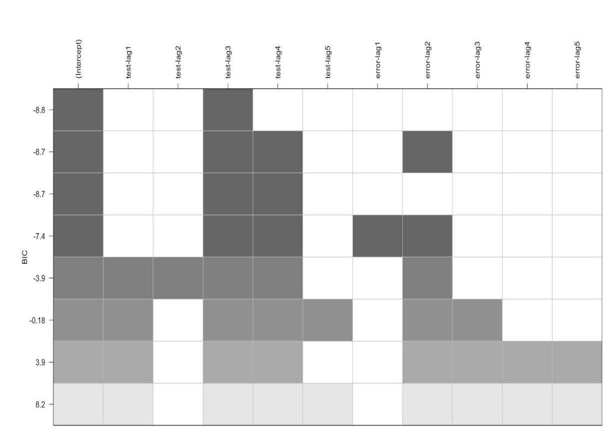

Fig: 1.27

As from the Fig: 1.26 we observed that there is no clear vertex in the EACF model so we had taken the row corresponding to p=4 as the vertex and included ARIMA model sets of ARIMA(4,1,1), ARIMA(3,1, 1), ARIMA(4,1,2) and ARIMA(3,1,3).

From the BIC table (Fig: 1.27) we have considered shaded columns. So columns included are AR(3) and MA(2) so the final ARIMA model would be ARIMA(3,1,2).

Conclusion

In task 1 quadratic trend model performs better than the linear trend model based on the coefficient of determination(R-Squared) as it better describes the variation in the series. In task two, 5 proposed set of ARIMA model based on the EACF and BIC tools are finalized that includes ARIMA(4,1,1), ARIMA(3,1, 1), ARIMA(4,1,2), ARIMA(3,1,3) and ARIMA(3,1,2).

Appendix

```
#importing the libraries
library(TSA)
library(tseries)

##### Task - 1 #####

#Read the data
ozone_data <- read.csv("./Documents/RMIT_Sem_3/Time Series Analysis/Assignment-1/data1.csv", header =
FALSE)
head(ozone_data)
class(ozone_data)

#renaming the rows
rownames(ozone_data) <- seq(from=1927, to=2016)
head(ozone_data)
```

```
#convert the data to timeseries format
ozone_data <- ts(as.vector(ozone_data), start=1927, end=2016)

#To check the thickness of the ozone layer thickness over the year shown using correlation value
x = ozone_data
y = zlag(ozone_data)
index = 2:length(y)
cor(x[index], y[index])

#Plotting the time series plot
plot(ozone_data,type='o',xlab="Time in Years",ylab='Thickness level in Doubson Units',main='Ozone Layer
Thickness from 1927-2016')

#ScatterPlot to check the correlation
plot(y=ozone_data,x=zlag(ozone_data),col=c("red"),xlab = "Previous Year Thickness of Ozone Layer",main =
"Scatter Plot of Ozone Layer Thickness")

#### Linear Model ####

#Applying the linear model to the data
model_lm = lm(ozone_data~time(ozone_data))
summary(model_lm)

#Linear model with regression line
plot(ozone_data,type='o',ylab='Thickness level in Doubson Units',xlab="Time in Years"  ,main ="Linear Trend
model - Thickness of Ozone layer from 1927-2016")
# add the fitted least squares line from model1
abline(model_lm)

# Residuals Analysis for the Linear model trends

# 1. Time series residuals
res.model_lm = rstudent(model_lm)
plot(y = res.model_lm, x = as.vector(time(ozone_data)),xlab = 'Time in Year',type='o', ylab='Standardized
Residuals',main = "Residuals of linear trend model - Thickness of Ozone layer from 1927 to 2016")

# 2. Residuals vs fitted trend
plot(y=rstudent(model_lm),x=as.vector(fitted(model_lm)),xlab='Fitted Trend Values', ylab='Standardized
Residuals',
     type='n', main ="Time series plot of standardised residuals")
points(y=rstudent(model_lm),x=as.vector(fitted(model_lm)))

# 3. Normality of the residual should be checked with histograms
hist(rstudent(model_lm),xlab='Standardized Residuals')

# 4. Normality of the residual should be checked with qq-plots
y = rstudent(model_lm)
qqnorm(y)
```

```
qqline(y, col =2, lwd =1, lty =2)

# 5. Shapiro-Wilk Test for Normality
y = rstudent(model_lm)
shapiro.test(y)

# 6. ACF for standardized residuals
acf(rstudent(model_lm), main ="ACF of the standardized residuals")

#####   Quadratic model  #####

t = time(ozone_data)
t2 = t^2
model_qa = lm(ozone_data~ t + t2)
summary(model_qa)

#Fitting the line
plot(ts(fitted(model_qa)), ylim = c(min(c(fitted(model_qa), as.vector(ozone_data))),
max(c(fitted(model_qa),as.vector(ozone_data)))),
    ylab='Thickness level in Doubson Units' , xlab="Time in Years" ,main = "Quadratic trend model - Thickness of
Ozone layer from 1927 to 2016", type="l",lty=2,col="red")
lines(as.vector(ozone_data),type="o")

# Residuals Analysis for Quadratic model trends
# 1. Time series residuals
res.model_qa = rstudent(model_qa)
plot(y = res.model_qa, x = as.vector(time(ozone_data)),xlab = 'Time in Year',type='o', ylab='Standardized
Residuals',main = "Residuals of Quadratic model trend - Thickness of Ozone layer from 1927 to 2016")

# 2. Residuals vs fitted trend
plot(y=rstudent(model_qa),x=as.vector(fitted(model_qa)),xlab='Fitted Trend Values', ylab='Standardized Residuals',
    type='n', main ="Time series plot of standardised residuals")
points(y=rstudent(model_qa),x=as.vector(fitted(model_qa)))

# 3. Normality of the residual should be checked with histograms
hist(rstudent(model_qa),xlab='Standardized Residuals')

# 4. Normality of the residual should be checked with qq-plots
y = rstudent(model_qa)
qqnorm(y)
qqline(y, col =2, lwd =1, lty =2)

# 5. Shapiro-Wilk Test for Normality
y = rstudent(model_qa)
shapiro.test(y)

# 6. ACF for standardized residuals
acf(rstudent(model_qa), main ="ACF of the standardized residuals")
```

```
### Forecasting...

# create a vector of time for next five years
t = c(2017,2018,2019,2020,2021) # create a time vector for next five years
t2 = t^2
# create a new data frame with t and t2
new_df = data.frame(t,t2)
forecasts = predict(model_qa,new_df, interval = "prediction")
plot(ozone_data,ylab= "Ozone Layer Thickness in Doubson Units", xlab="Time in Years" , ylim =
c(-15,3),xlim=c(1927,2021),main="Time series plot for Ozone layer thickness")
# convert forecasts to time series object starting from the first
# time steps-ahead to be able to use plot function
lines(ts(as.vector(forecasts[,1]), start = 2016), col="red", type="l",lwd=2)
lines(ts(as.vector(forecasts[,2]), start = 2016), col="blue", type="l",lwd=2)
lines(ts(as.vector(forecasts[,3]), start = 2016), col="blue", type="l",lwd=2)

legend("topright", lty=1, pch=1,bty = "n", col=c("black","blue","red"), text.width = 18,
       c("Data","5% forecast limit", "Forecasts"))
```

Task - 2

```
#Read the data
ozone_data <- read.csv("./Documents/RMIT_Sem_3/Time Series Analysis/Assignment-1/data1.csv", header =
FALSE)
head(ozone_data)
class(ozone_data)

#convert the data to timeseries format
ozone_data <- ts(as.vector(ozone_data), start=1927, end=2016)

plot(ozone_data, type="o",ylab="Thickness of Ozone Layer")

#Plotting the series in acf and pacf plot.
par(mfrow=c(1,2))
acf(ozone_data)
pacf(ozone_data)
par(mfrow=c(1,1))

#Checking the p-value using adf test
adf.test(ozone_data)

#apply box-cox transformation
ozone_transform = BoxCox.ar(ozone_data + abs(min(ozone_data))+1)
ozone_transform$ci

#Checking the normality using qq-plot
```

```
lambda = 1.2
ozone_data = ozone_data + abs(min(ozone_data))+1
ozone_data_BC = (ozone_data^lambda-1)/lambda # apply the Box-Cox transformation
par(mfrow=c(1,1))
qqnorm(ozone_data_BC)
qqline(ozone_data_BC, col = 2)
shapiro.test(ozone_data_BC)

#plot the graph after applying box-cox transformation
plot(ozone_data_BC,type='o',ylab='Ozone Layer Thickness')

#Let's calculate the first difference and plot the first differenced series
diff_ozone_data_BC = diff(ozone_data)
plot(diff_ozone_data_BC,type='o',ylab='Ozone Layer Thickness')

#Checking the p-value using adf test
adf.test(diff_ozone_data_BC)

#Plotting the series using acf and pacf plot
par(mfrow=c(1,2))
acf(diff_ozone_data_BC)
pacf(diff_ozone_data_BC)
par(mfrow=c(1,1))

#Plotting the series using eacf plot after differentiation
eacf(diff_ozone_data_BC)

#Plotting the series using BIC table after differentiation
par(mfrow=c(1,1))
res4 = armasubsets(y=diff_ozone_data_BC,nar=6,nma=6,y.name='test',ar.method='ols')
plot(res4)
```

References

1. https://ourworldindata.org/ozone-layer#when-is-the-ozone-layer-expected-to-recover
2. MATH1318 Time Series Analysis notes and tutorials by Dr. Haydar Demirhan.