MATH1318 Time Series Analysis Assignment 1

Time Series Analysis of the thickness of Ozone

Milind Shaileshkumar Parvatia s3806853

21/04/2021

Index

Introduction

Background

The ozone layer is the common term for the high concentration of ozone molecules that are found in the stratosphere around 15–30km above the earth's surface. It covers the entire planet and protects life on earth by absorbing harmful ultraviolet-B (UV-B) radiation from the sun (Department of Agriculture, Water and the Environment). Due to the industrial revolution carbon emission has exponentially increased from the early 19th century. This led to the creation of chlorofluorocarbons (CFCs) which is blamed for erasing the ozone layer and later named 'the ozone hole'.

Report Aim

This study is to investigate this trend and provide a detailed report on this matter. The report is divided into two parts. Part 1 is to discuss the main trend which can present the thickness of the ozone layer over time and find the best fitting model and used that model to predict the next 5-year trend of the thickness of the ozone layer. Part 2 is to find the set of model parameter for the ARIMA(p, d, q) model using model specification tools like ACF-PACF, EACF, BIC table clearly and write clear and correct comments to back up your choices of (p, d, q) orders.

Methodology

Data Collection

The data is composed of a single series of factorial values which are in Dobson units and taken each year from 1927 to 2016. The data is provided by the client.

Data Description

- Single series in Dobson units which represent the thickness of the ozone layer
- Each data points in series represent one year from 1927 to 2016

Data Preprocessing (Modules 1-5)

Rstudio will be used for this task and to represent the given dataset addition column representing years from 1927 to 2016 will be added and the whole dataset will be converted into ts (time-series object). An additional correlation will be checked after that to see the correlation behaviour of years vs. thickness measures.

Task 1

This is a time-series problem and to solve this task we will follow a multistage model-building strategy of time series analysis with the following three steps:

- **Model specification**, (or identification): This step follows understanding and determines which models we should consider for model fitting from linear, quadratic, cosine and seasonalities models.
- **Model fitting**: These steps will use different model fitting techniques to determine the difference in the model.
- **Model diagnostics**: This is the final step and in this step, we will diagnose using different model fitting parameters and residual analysis to determine the best fitting model.

To perform model diagnostics we will do a residual analysis of histogram distribution comparison, normal QQ plots of the standardised residuals, Shapiro Wilk test and ACF test.

What is Residual Analysis?

The *estimator* or *predictor* of unobserved stochastic component {Xt},

$$X\hat{}t=Yt-\mu\hat{}t$$

is called **residual** corresponding to the *t*th observation.

After computation of residuals or standardised residual, we examine various residual plots. The first plot to examine is the plot of the residuals over time.

We have to check whether the trend model is reasonably correct. In terms of the best model, the residuals should behave roughly like the true stochastic component and not like white noise, we can achieve it by making various assumptions about the stochastic component that can be assessed by looking at the residuals.

Task 2

This task is focused on proposing ARIMA(p,d,q) models which are different specification tools like **ACF-PACF, EACF, BIC** table and from these tools we will try to find the best p and q values for the ARIMA model.

What is ACF-PACF?

ACF-PACF are autocorrelation function and passive autocorrelation function which are extremely helpful to identify parameters of p and q. Here ACF will help us identify the true pattern of series like if there's no significant value the series has white noise and can't be used.

The ACF of the AR(p) series tails off as a mixture of exponential decays or damped sine waves depending on the roots of the autocorrelation equation. A damped sine wave appears if some of the roots are complex and all autocorrelations after lag q will be insignificant in the sample ACF of a MA(q) series. For the general ARMA(p,q) processes, the sample ACF will tail off after lag q like an AR(p) process.

In the sample PACF of an AR(p) series, the estimates of partial autocorrelations at the lags 1,2,…,q will be significant and then they will vanish after lag q. The PACF of MA(q) series tails off as a mixture of exponential decays or damped sine waves depending on the roots of the autocorrelation equation.

What is EACF?

The sample ACF and PACF provide effective tools for identifying pure AR(p) or MA(q) models. However, for a mixed ARMA model, its theoretical ACF and PACF have infinitely many nonzero values, making it difficult to identify mixed models from the sample ACF and PACF. Many graphical tools have been proposed to make it easier to identify the ARMA orders. The extended autocorrelation (EACF) method is one of those methods to identify the order of autoregressive and moving average components of ARMA models. It is supposed to have good sampling properties for moderately large sample sizes.

The EACF method uses the fact that if the AR part of a mixed ARMA model is known, "filtering out" the autoregression from the observed time series results in a pure MA process that enjoys the cutoff property in its ACF.

What is BIC?

BIC Is known as Bayesian Information Criterion and it's mostly used in model selection and specification.

$$BIC=-2\log(\text{maximum likelihood})+k\log(n)$$

Thus, an ARMA(k,j) model can be approximately estimated by regressing the time series on its lags 1 to k together with the lags 1 to j of the residuals from the high order autoregression; the BIC of this autoregressive model is an estimate of the BIC obtained with maximum likelihood estimation.

Tools

- Rstudio

Libraries

- readr
- TSA
- xts

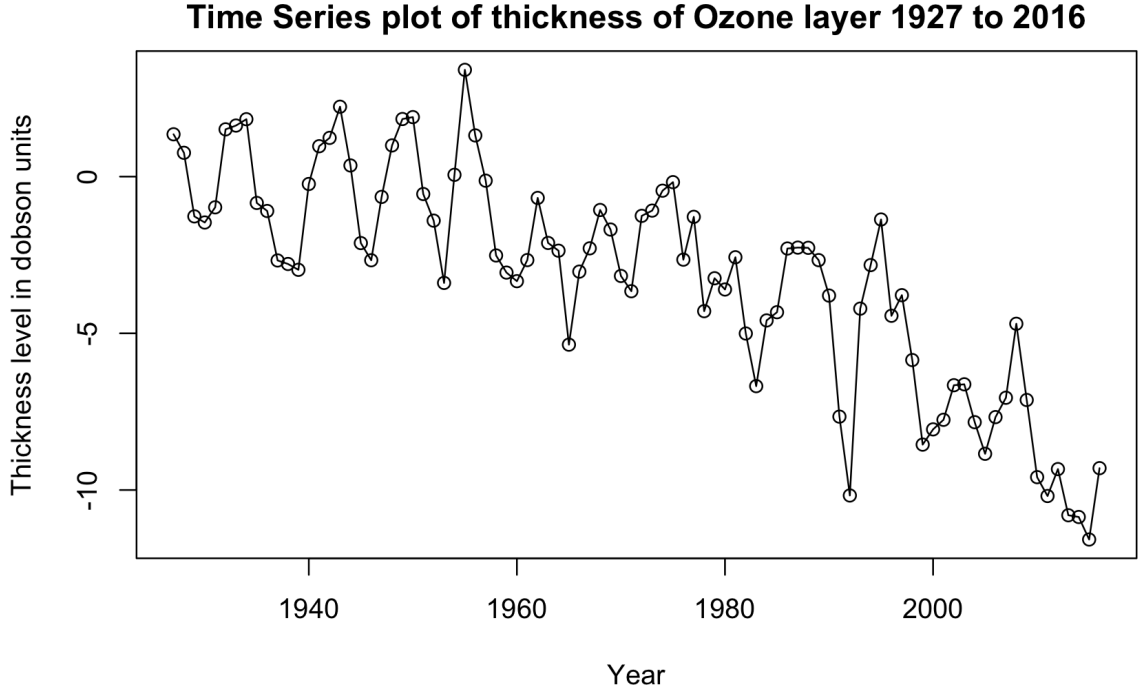
Part 1

Descriptive analysis

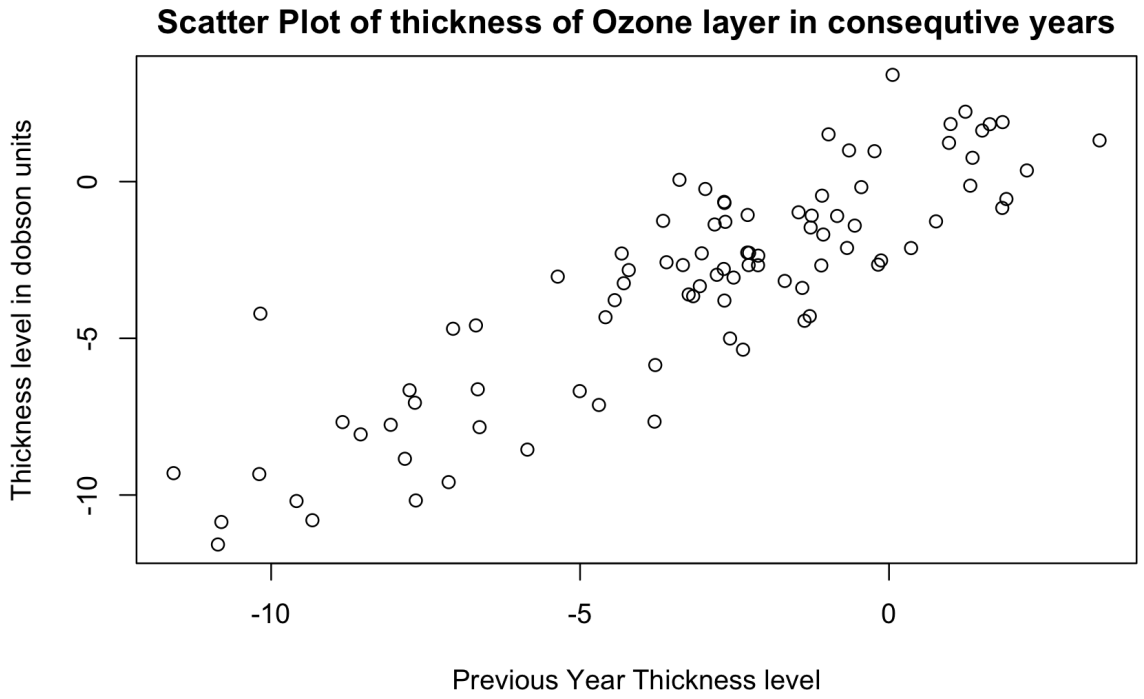
Descriptive analysis is the basic interpretation of any time series plot which includes the trend of series over time, changepoint which can be blamed for any changes in the behaviour of series due to some external or internal reason, seasonalities which are repeating patterns over time, the behaviour of series which can be determined from moving average and Autoregressive behaviour and lastly changing variance which can determine the amount of fluctuation from series.

To support our assumptions of all these behaviours of series we will use different specification tools described below.

- **Time series plot**: this plot helps identify what our data will look like over a given period and make time series more representable.



- **Scatter plot of time series lag**: this plot helps represent the correlation of data points to each other and provides us strong visual evidence of correlated data.

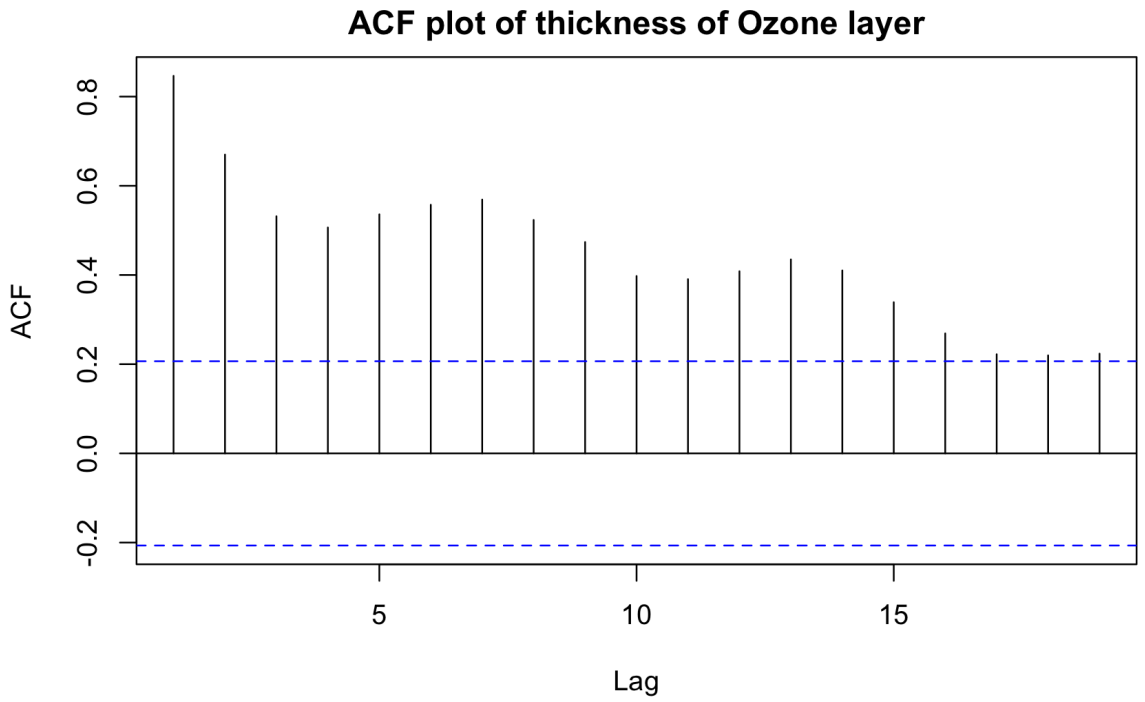


- **Correlation factor**: this is the correlation function of Rstudio which helps us calculate the actual value of correlation from this year against the previous year.

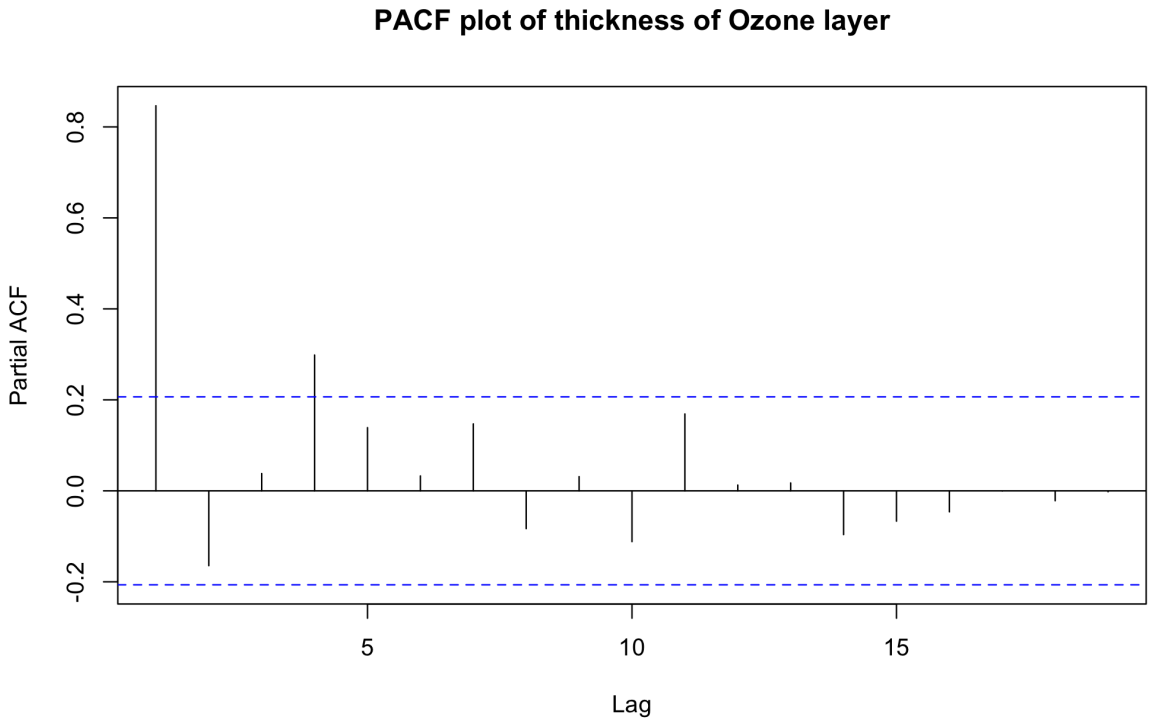
  ```
  [1] 0.8700381
  ```

  The correlation factor is providing strong support for high correlation which is as expected due to the thickness of the ozone layer is dependent on the previous year's level of Ozone molecule in atmospheres.

- **ACF**: Autocorrelation function (ACF) is a visual way to show a serial correlation in time series data, this plot also helps us identify trend, seasonality, stationarity of series.



- **PACF**: Partial autocorrelation function (PACF) gives the partial correlation of a stationary time series with its own lagged values, regressing the values of the time series at all shorter lags. It contrasts with the autocorrelation function, which does not control for other lags (Autocorrelation).



From figure 1 to figure 4, we can discuss the following descriptive points from the series.

Trend: We can see the negative trend from figure 1 and figure 3, which provides strong evidence of reduction of thickness level from the Ozone layer.

Seasonalities: There seem to be no direct repeating patterns in figure 1 but from figure 3 we can see wave-like patterns, which indicate hidden seasonal behaviour in this plot.

Changepoint: one possible changing point can be in 1962 where we can see the stationarity of series is affected and the trend of series has become downwards compared to earlier years.

Behaviour: From figure 1 we can say that series is fluctuating over time and in scatter plot of the lag figure it has successive points related to one another and data points are highly affecting next year's data, from this, we can say that series have both moving average and autoregressive behaviour.

Changing variance: There is an effect of changing variance after the year 1962 where series can be seen low fluctuating compared to earlier steady moving average behaviour in series.

Finding best model

Since it's a non-stationary trend we can consider it as a nonconstant mean trend and classical regression analysis can be used to model a nonconstant mean trend. We will try to fit linear, quadratic, seasonal means, and cosine trends and see which one can work well with detail analysis to identify the best fitting model for the series.

Linear trend

linear trend model is expressed as follows:

$$\mu t = \beta 0 + \beta 1 t$$

Where β0 represents intercept and β1 corresponds to the slope of the linear trend. We can see the summary of this model below. (modules)

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

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(timeseries)  -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
```

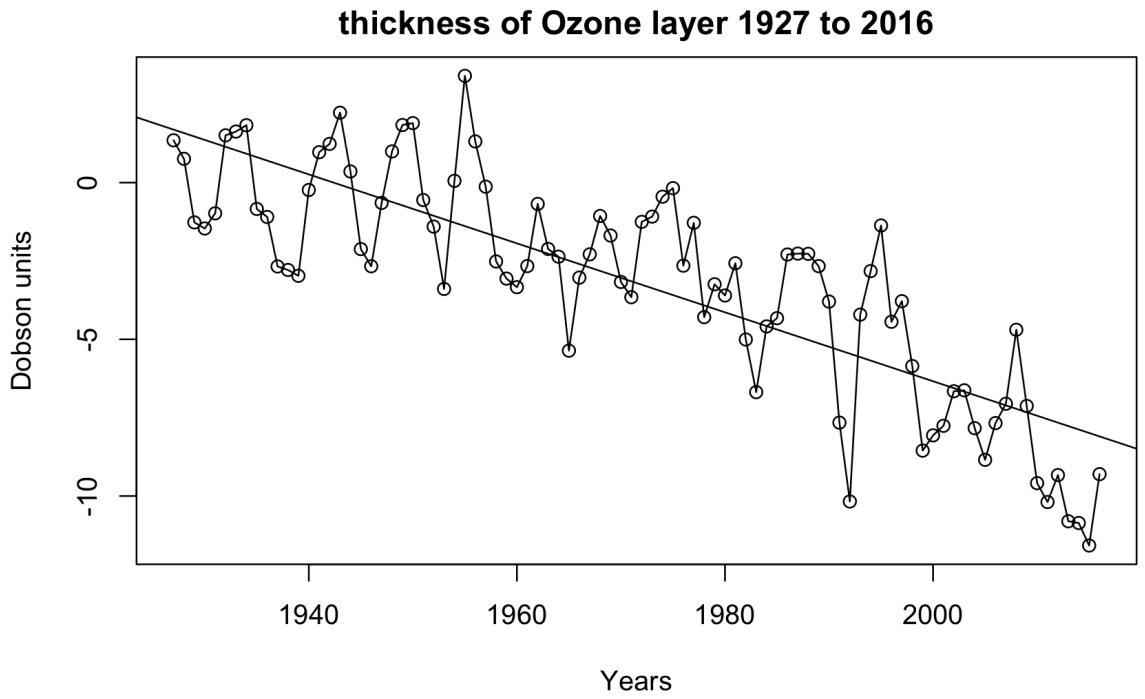
Interpretation of Summary

Estimates of slope and intercept are β^1=-0.110 and β^0=213.720, respectively. Here the slope is statistically significant at a 5% significance level.

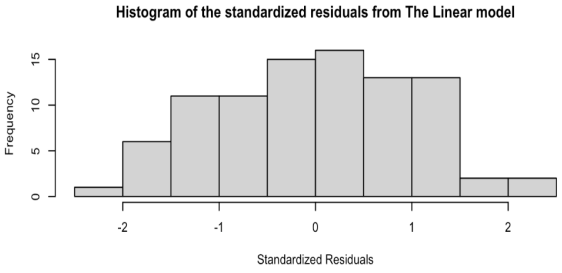
We can use R-square as a measure of goodness of fitted trend and from the above summary. According to multiple R-square, about 66.93% of the residuals in the series is explained by the linear trend. The adjusted R-square provides an approximately unbiased estimate of true R-square which is 66.55% here.

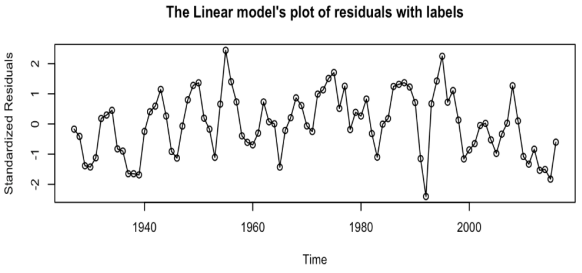
We can also determine that both β^1 and β^0 have passed the null hypothesis test with `Pr(>|t|) <2e-16`.

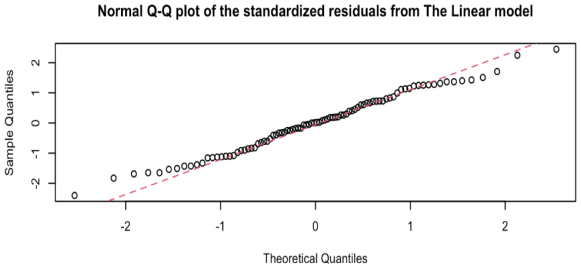
Model fitting plot

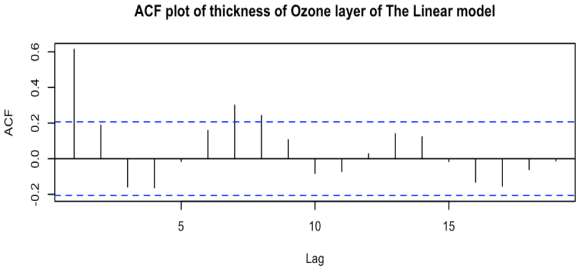


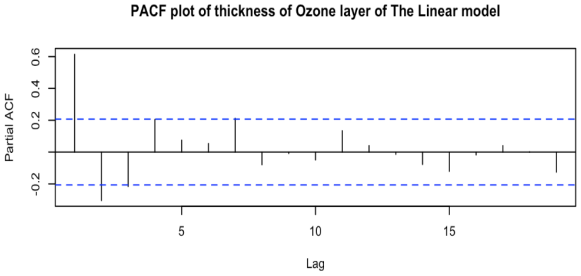
Model diagnostic with residual analysis











Histogram of residual: Histogram of residual represents normality of residual plots and we can say this model has really good normalization of residual as we can see it's the pretty symmetric distribution of data points.

Residual plot: Residual plot is an indication of Moving Average behaviour and its fitting between (-2,2) which is accepted as variation limits.

Normal Q-Q plot: Normal QQ plot also indicates the normality of residuals and from this figure, we can see clearly that quantities are only fitting between ranges(-1,1).

ACF plot: ACF plot is an autocorrelation function plot that confirms the smoothness of the time series plot as we have correlation values higher than the confidence bound at several lags till 5 lag. This is not what we expect from a white noise process.

PACF plot: PACF plot is a Partial autocorrelation function plot that indicates the size of the moving average window and here it indicates correlation values are higher than bounds.

Shapiro-Wilk normality test

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

This test is to prove the normality of data and here p-value is greater than 0.05, which means residuals are normally distributed.

Quadratic trend

The deterministic quadratic trend model is expressed as follows

$$\mu t=\beta 0+\beta 1t+\beta 2t2$$

Where β0 represents intercept, β1 corresponds to the linear trend, and β2 corresponds to quadratic trend in time. (modules)

```
Call:

lm(formula = timeseries ~ 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
```

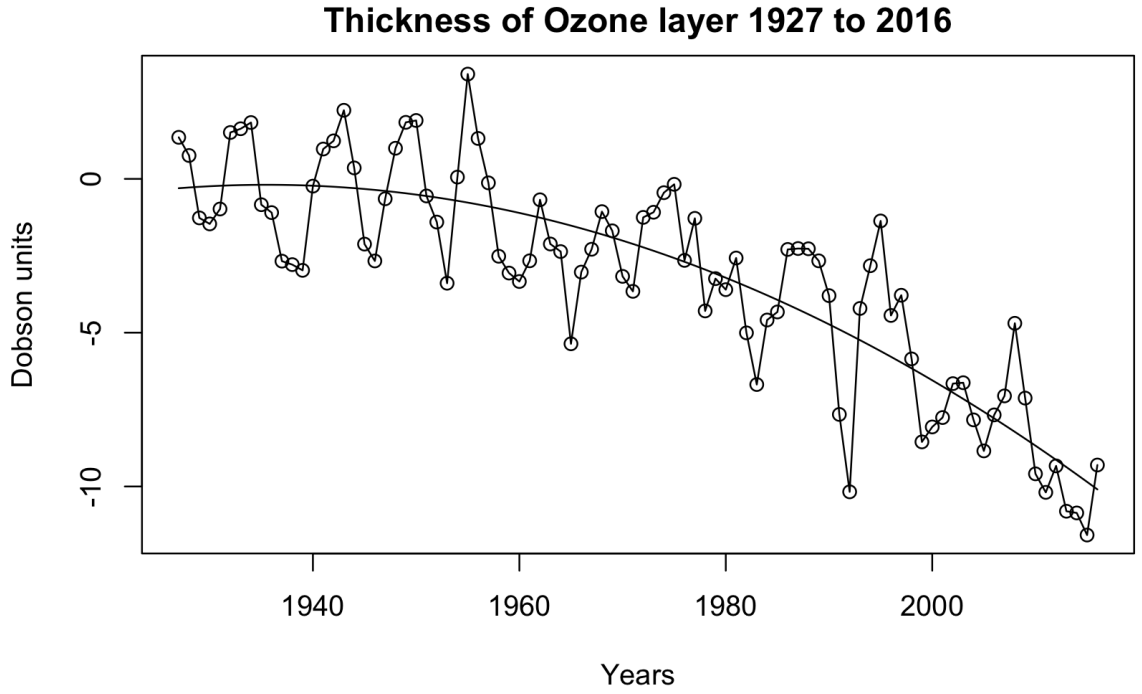
Interpretation of Summary

Estimates of slope and intercept are β2=-1.530e-03, β1=5.924e+00 and β0=-5.733e+03, respectively. Here the slope is statistically significant at a 5% significance level.

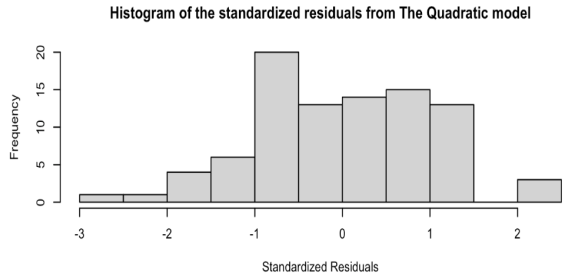
According to multiple R-square, about 73.91% of the residuals in the series is explained by the quadratic trend. The adjusted R-square provides an approximately unbiased estimate of true R-square. Hence the value is 73.31% in this case.

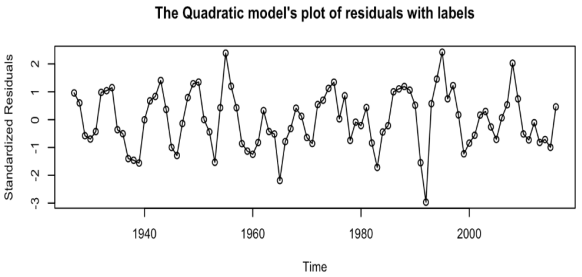
We can also determine that all β2, β1 and β0 have passed the null hypothesis test with `Pr(>|t|)` values are `< 0.05`.

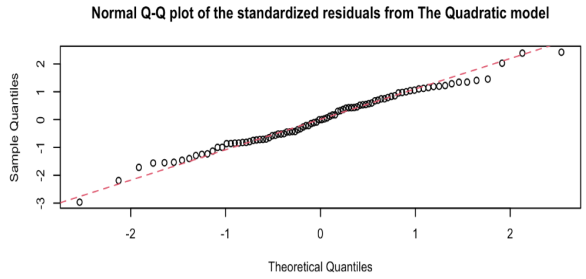
Model fitting plot

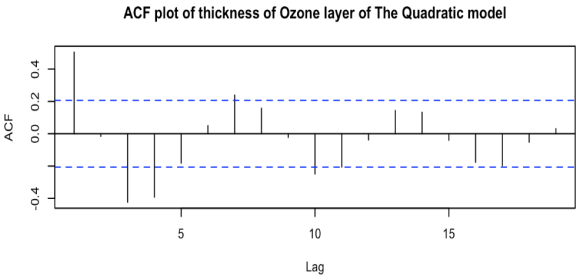


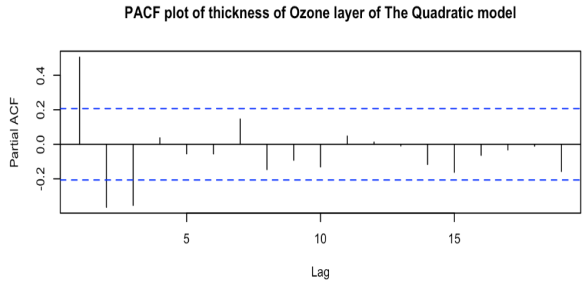
Model diagnostic











Histogram of residual: Histogram of residual represents normality of residual plots and we can say this model has somewhat normalization of residual as we can see it's one bar at -1 is quite higher and a gap in normally distributed data points.

Residual plot: Residual plot is an indication of Moving Average behaviour and its fitting between (-2,2) which is accepted as variation limits.

Normal Q-Q plot: Normal QQ plot also indicates the normality of residuals and from this figure, we can see clearly that quantities are only fitting between ranges(-2,2) which is good.

ACF plot: ACF plot is an autocorrelation function plot that confirms the smoothness of the time series plot as we have correlation values higher than the confidence bound at several lags till 5 lag. This is not what we expect from a white noise process.

PACF plot: PACF plot is a Partial autocorrelation function plot that indicates the size of the moving average window and here it indicates correlation values are higher than bounds.

Shapiro-Wilk normality test

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

This test is to prove the normality of data and here p-value is greater than 0.05, which means residuals are normally distributed.

Seasonal trend

Here we assume that the observed series can be represented as

$$Yt=\mu t+Xt$$

where E(Xt)=0 for all t. The most general assumption for μt with monthly seasonal data is that there are 12 parameters, β1,β2,…,β12, giving the expected average temperature for each of the 12 months. To represent seasonality, we may write a **seasonal model** such that

$$\mu_t = \begin{cases} \beta_1 & \text{for } t = 1, 13, 25, \ldots \\ \beta_2 & \text{for } t = 2, 14, 26, \ldots \\ \vdots & \\ \beta_{12} & \text{for } t = 12, 24, 36, \ldots \end{cases}$$

We need to set up indicator variables (sometimes called dummy variables) that indicate the month to which each of the data points pertains before going on with the estimation of parameters. We can also include an intercept term β0 in the model but here we chose to remove it from the model. (module)

```
Call:
lm(formula = timeseries1 ~ year. - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1118 -1.8902  0.2771  2.4230  6.8748

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
year.Season-1    -2.086      1.221  -1.709  0.09140 .
year.Season-2    -2.414      1.221  -1.977  0.05149 .
year.Season-3    -3.193      1.221  -2.615  0.01065 *
year.Season-4    -3.484      1.221  -2.854  0.00550 **
year.Season-5    -3.896      1.221  -3.191  0.00202 **
year.Season-6    -3.529      1.221  -2.890  0.00495 **
year.Season-7    -3.451      1.221  -2.827  0.00593 **
year.Season-8    -2.964      1.221  -2.427  0.01745 *
year.Season-9    -3.468      1.221  -2.840  0.00572 **
year.Season-10   -3.538      1.221  -2.898  0.00485 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.663 on 80 degrees of freedom
Multiple R-squared:  0.4691,     Adjusted R-squared:  0.4028
F-statistic:  7.07 on 10 and 80 DF,  p-value: 7.353e-08
```

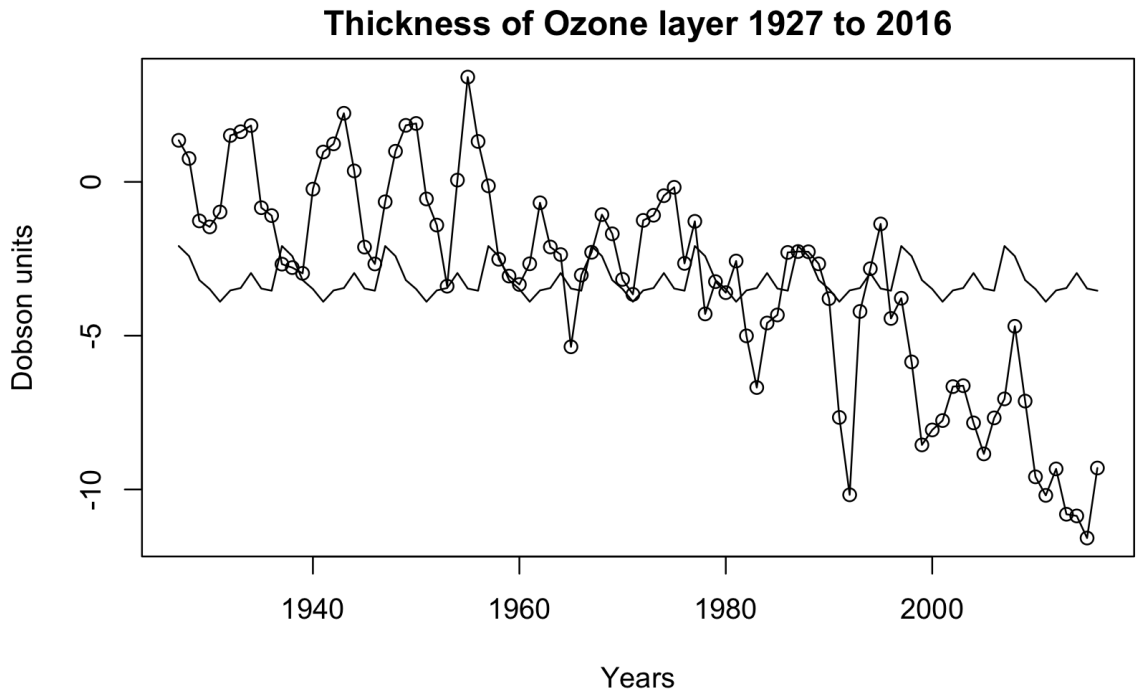
Interpretation of Summary

Estimates of the slope are ranging from year.Season-1 to year.Season-10 and we have removed the interceptor to make all the slope’s value compare to year.Season-1. Here the slope is statistically significant at a 5% significance level.

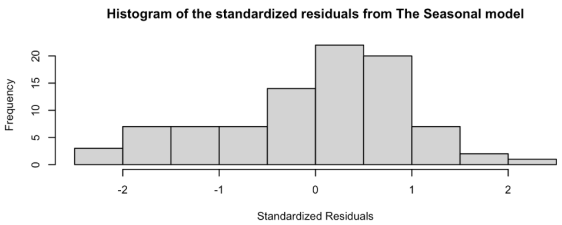
We can use R-square as a measure of goodness of fitted trend and from the above summary. According to multiple R-square, about 46.91% of the residuals in the series is explained by the linear trend. The adjusted R-square provides an approximately unbiased estimate of true R-square which is 40.28% here.

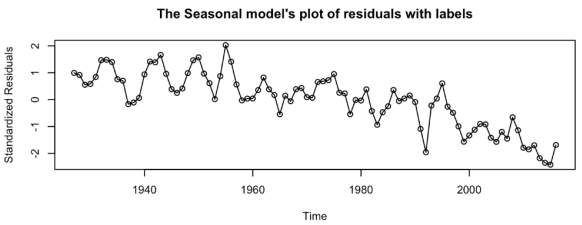
We can also determine that none of the slopes has significant values of `Pr(>|t|) < 0.05`

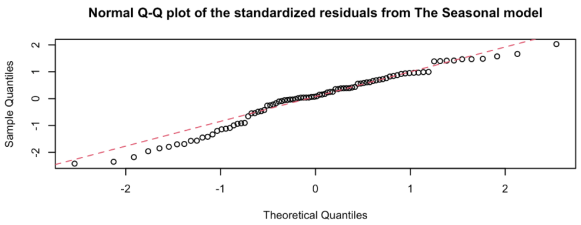
Model fitting plot

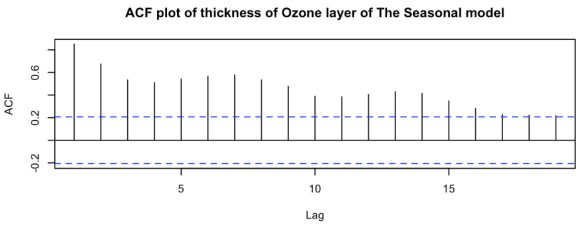


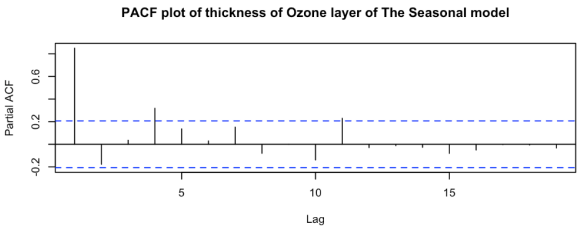
Model diagnostic











Histogram of residual: Histogram of residual represents normality of residual plots and we can say this model has not normalization of residual as we can see it's one side has more residual data points than other and it's not well normalized.

Residual plot: Residual plot is an indication of Moving Average behaviour and its fitting between (-2,2) which is accepted as variation limits.

Normal Q-Q plot: Normal QQ plot also indicates normality of residuals and from this figure, we can see almost half of residuals are not fitting on the dashed line and thus this gives us strong evidence of non-normality in residuals.

ACF plot: ACF plot is an autocorrelation function plot and from this plot, we can say that that there is still a trend in this plot and we can't determine Autocorrelation values, thus we can't use this model

PACF plot: PACF plot is a Partial autocorrelation function plot and here since we reject the ACF plot we will not go through PACF to check its moving average behaviour.

Shapiro-Wilk normality test

```
data:  y
W = 0.96653, p-value = 0.02025
```

This test is to prove the normality of data and here p-value is less than 0.05, which means residuals significantly deviate from the normal distribution.

Harmonic trend

We can include the information on the shape of the seasonal trend in the model by assigning a cosine curve as the mean function μt:

$$\mu t = \beta \cos(2\pi f t + \Phi)$$

Here,

β(>0) f and Φ are called the amplitude, frequency, and phase of the curve. As t varies, the curve oscillates within the interval [−β,β]. Since the curve repeats itself exactly every 1/f time units, 1/f is called the period of the cosine wave. When we set f=1/12, a cosine wave will repeat itself every 12 months. So we say that the period is 12.

For estimation purposes, we need to make the above cosine trend model linear in terms of its parameters. With the following misinterpretation, we get

$$\beta \cos(2\pi f t + \Phi) = \beta 1 \cos(2\pi f t) + \beta 2 \sin(2\pi f t)$$

where

$$\beta = \beta 12 + \beta 22 \text{ and } \Phi = \operatorname{atan}(-\beta 2 / \beta 1)$$

and, conversely,

$$\beta 1 = \beta \cos(\Phi) \text{ and } \beta 2 = \beta \sin(\Phi).$$

Consequently, we will use

ß2, respectively. The simplest such model for the trend would be expressed as

$$\mu t=ß0+ß1\cos(2\pi ft)+ß2\sin(2\pi ft)$$

Here the constant term ß0 represents a cosine with frequency zero.(modules)

```
Call:
lm(formula = timeseries ~ har. - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3520 -1.8906  0.4837  2.3643  6.4248

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
har.cos(2*pi*t) -2.970e+00  4.790e-01  -6.199 1.79e-08 ***
har.sin(2*pi*t)  5.462e+11  7.105e+11   0.769    0.444
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.522 on 88 degrees of freedom
Multiple R-squared:  0.4601,     Adjusted R-squared:  0.4479
F-statistic:  37.5 on 2 and 88 DF,  p-value: 1.663e-12
```

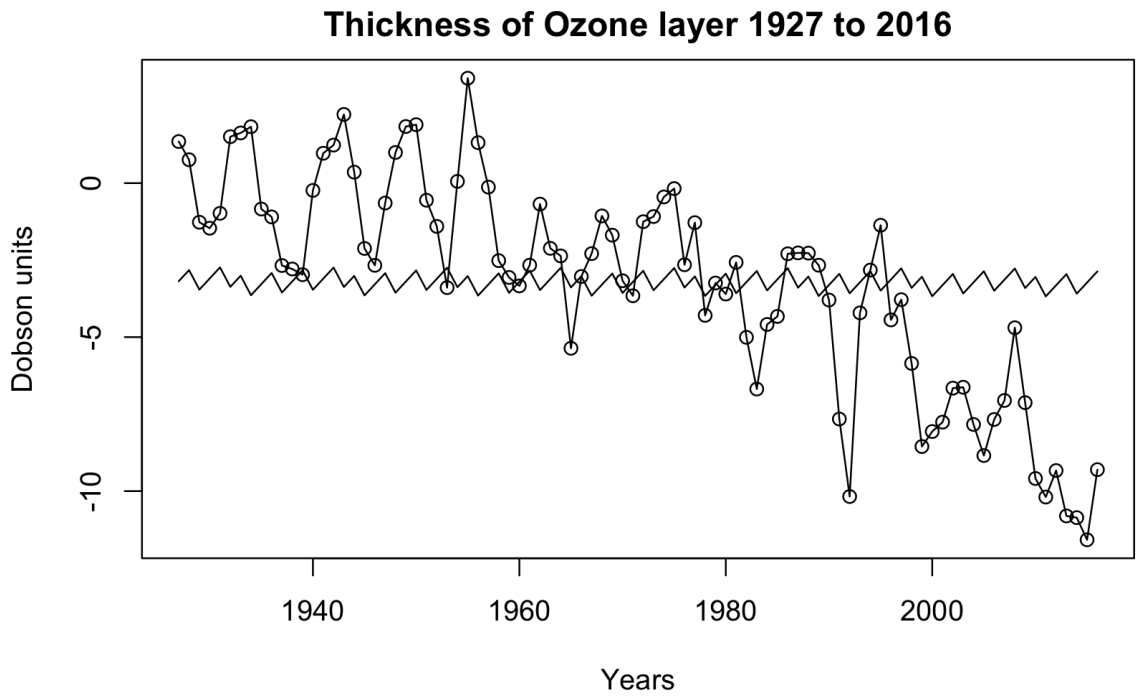
Interpretation of Summary

Estimates of the slope are harmonic sin and cosine functions and we have removed the interceptor. Here the slope is statistically significant at a 5% significance level.

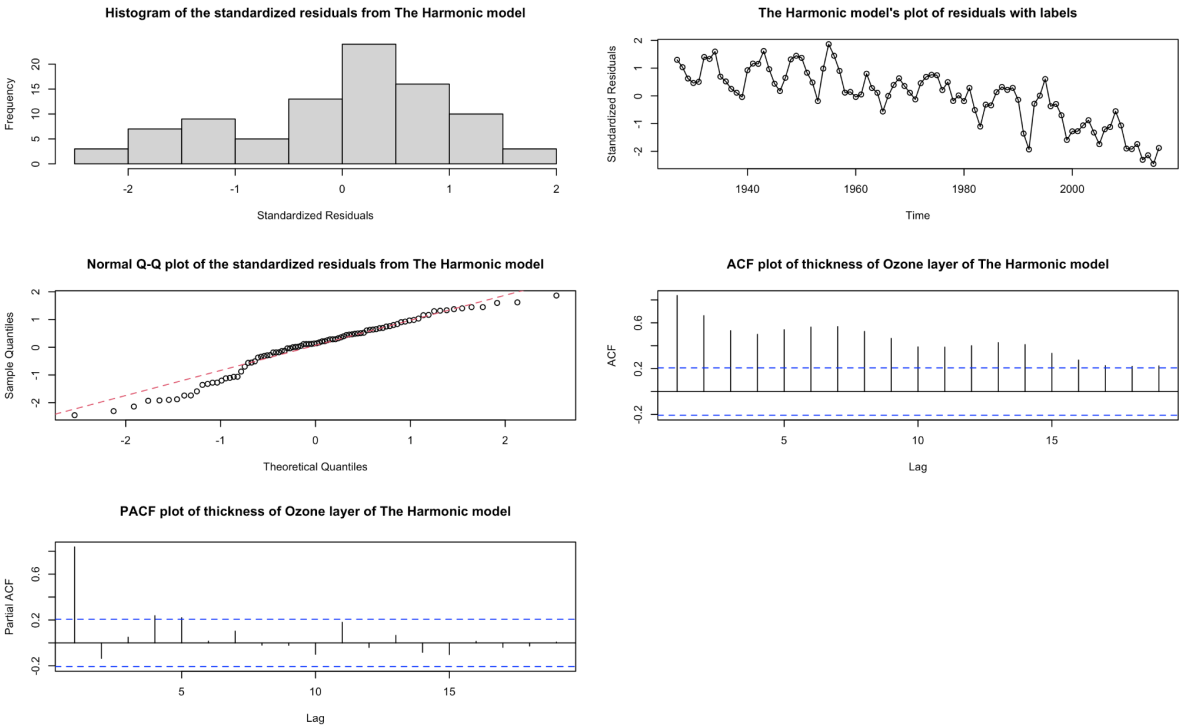
We can use R-square as a measure of goodness of fitted trend and from the above summary. According to multiple R-square, about 46.01% of the residuals in the series is explained by the linear trend. The adjusted R-square provides an approximately unbiased estimate of true R-square which is 44.79% here.

We can also determine that cos have passed the t-test with `Pr(>|t|) 1.79e-08 < 0.05`, while sin function fails to do so.

Model fitting plot



Model diagnostic



Histogram of residual: Histogram of residual represents normality of residual plots and we can say this model has not normalization of residual as we can see it's one side has more residual data points than other and it's not well normalized.

Residual plot: Residual plot is an indication of Moving Average behaviour and its fitting between (-2,2) which is accepted as variation limits.

Normal Q-Q plot: Normal QQ plot also indicates normality of residuals and this figure indicates that some of the residual points are fitting on dash lines but we can see clearly on the left side < -1 residuals are not on dash line thus we can say this model has major non normalised residual points.

ACF plot: ACF plot is an autocorrelation function plot and from this plot, we can say that there is still a trend in this plot and we can't determine Autocorrelation values, thus we can't use this model.

PACF plot: PACF plot is a Partial autocorrelation function plot and here since we reject the ACF plot we will not go through PACF to check its moving average behaviour.

Shapiro-Wilk normality test

```
data:  y
W = 0.95856, p-value = 0.005875
```

This test is to prove the normality of data and here p-value is less than 0.05, which means residuals significantly deviate from the normal distribution.

Comparison of models

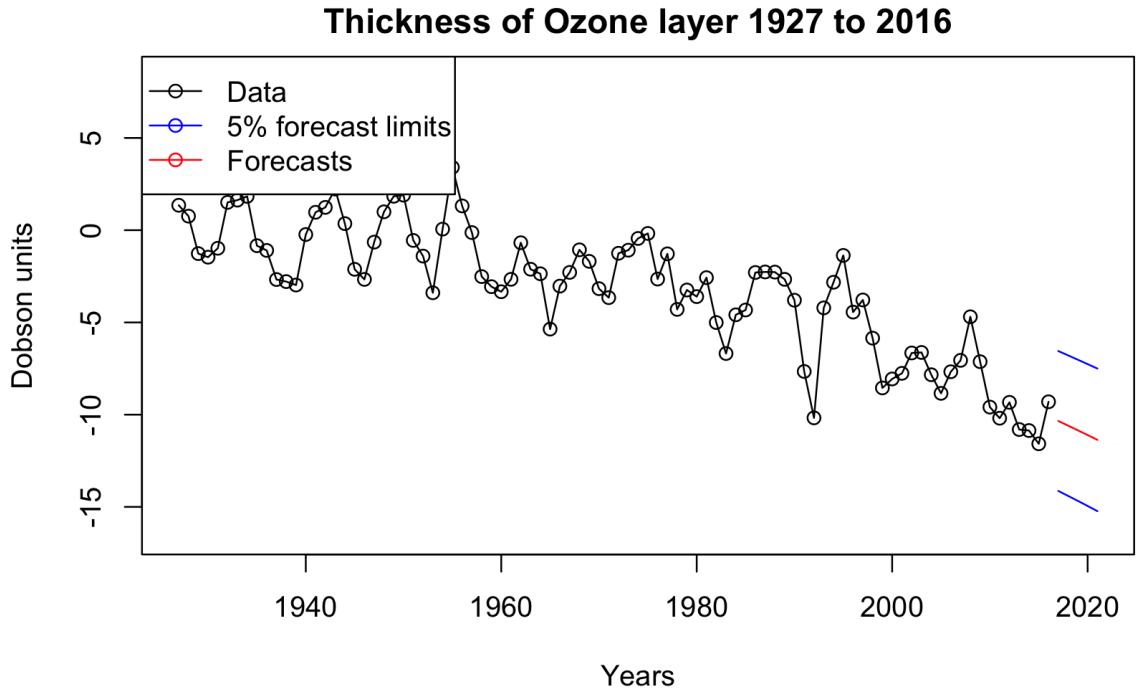
From the above residual analysis, we can see that the quadratic model has the highest R^2 values and can easily represent 73.71% residuals in the data. Which is by far the highest compared to other models. Also in detailed model diagnostics, we can see that it has highly normalised residuals from the histogram and normal QQ plot. From the residual plot, we can see it fits (-2,2) variation. Also, it has great ACF and PACF plots. Finally, the quadratic model also passes the Shapiro-Wilk normality test with a p-value of 0.6493 indicating strong proof of normality in residuals. From all this consideration we can select quadratic models as the best fitting model compared to all the other models.

Forecasting

To test the best-fitted model is working well we will forecast the next 5 year of data using the Quadratic model. We will use simple regression to predict new future values of time by passing it to the Quadratic model. To explain the forecast we will use h = 5 (next 5 year).

1. Get the last date from the series and generate new series from the last date to the next h dates by adding default frequency.
2. Once that's done create a model and perform model fitting to get newly generated values.

We have to use predict() function with the fitted model object and the sequence created at step 1 as inputs.



Forecast Interpretation

We can see that since the model is a quadratic line, it follows the curve for the next h = 5 year and provides us with 5% forecast limits which are to ensure the possible range for the new values.

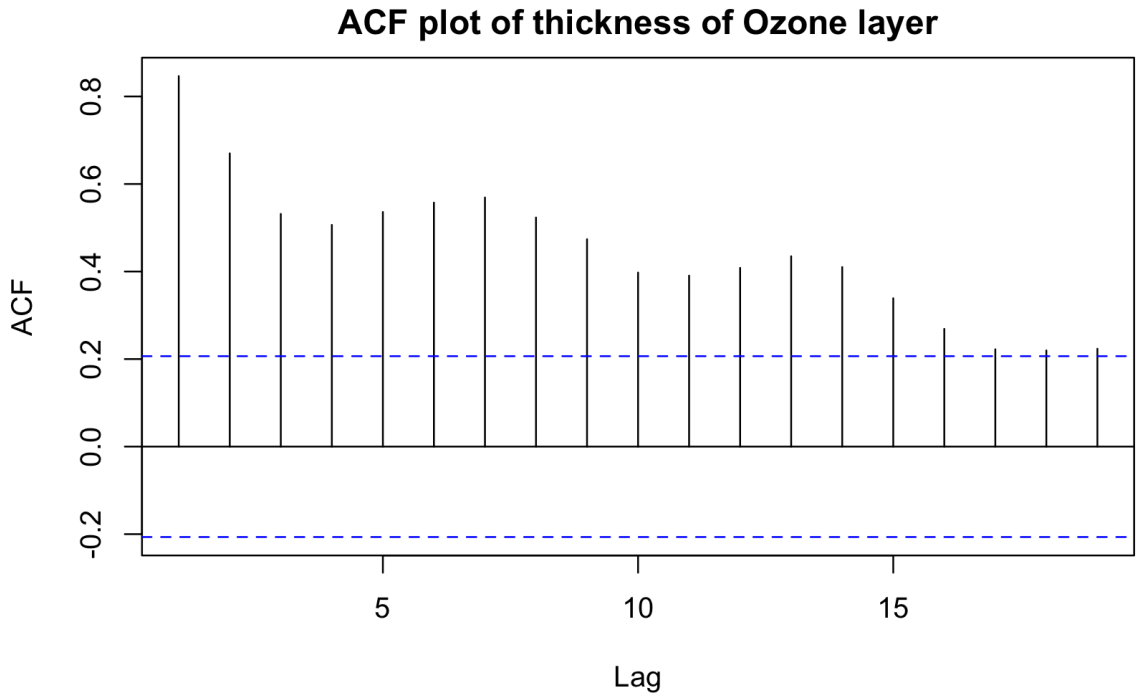
Part 2

Propose a set of possible ARIMA(p, d, q) models

We can find parameters of the ARIMA model from different types of plots like ACF-PACF, EACF, BIC table. To find parameters of the ARIMA model we will try to explore these specification tools one by one to get possible parameters for the ARIMA model.

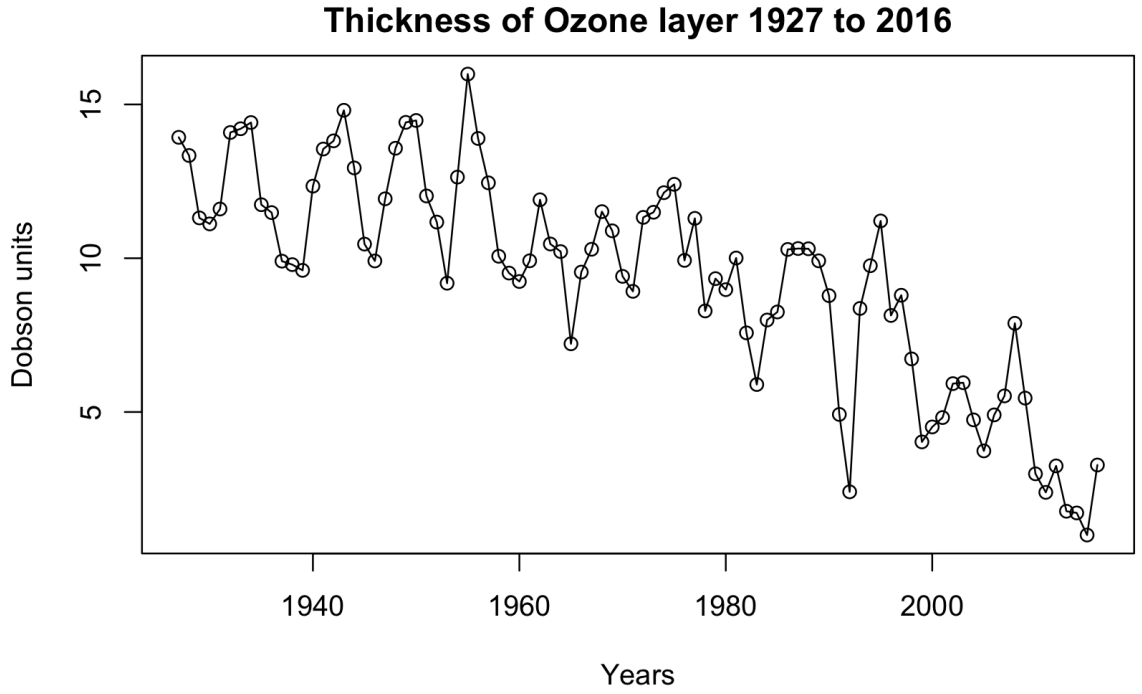
ACF-PACF

From the ACF plot, we can see the decreasing trend with some wave-like behaviour which indicates that there is a trend in data and stochastic trend in series does not allow us to get values of p and q. To get these values we have to use the transformation of series to remove this downward trend and create a non-stationary model.

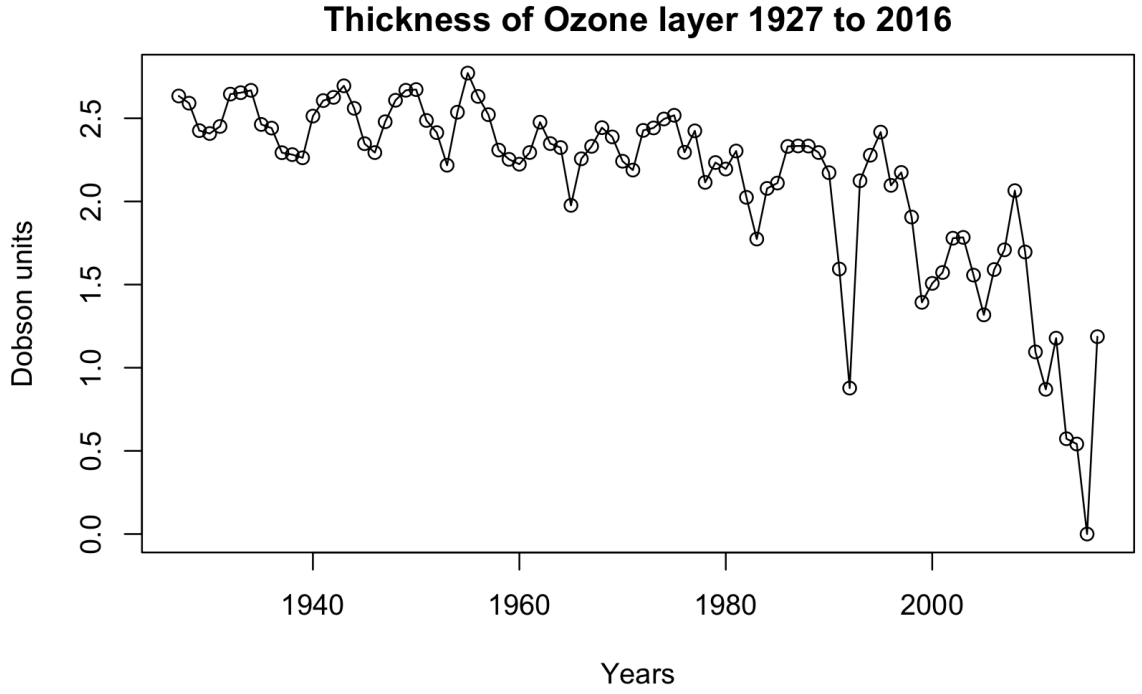


To remove the trend from the series we will use a logarithmic function on the dataset to convert it into logarithmic series. Since the trend is negative and the dataset has a lot of negative values we have to convert data of time series to positive. To do that we will find the minimum value from data and add the absolute value of data with small integers like 1 to the whole series. This will convert the whole series to positive series and makes getting logs of the series very easy.

Positive TimeSeries



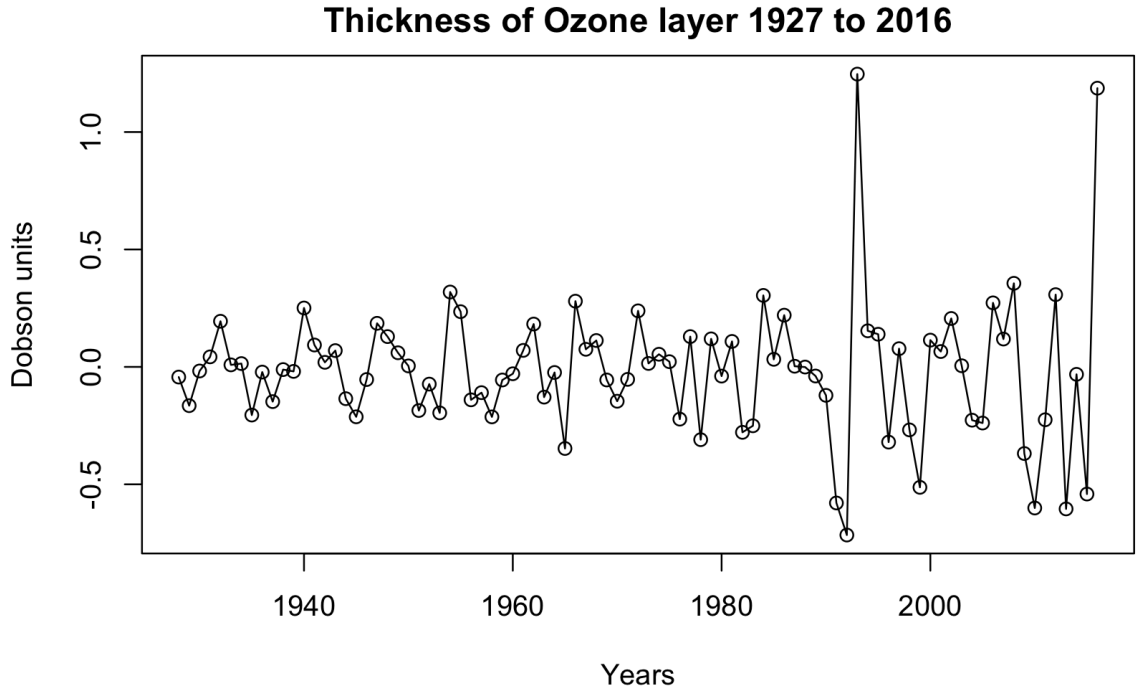
Log of Time series



Once we get the log of positive time-series data we can further try the transformation to get the difference of time series and convert it to the first difference series. We can try the N difference of series until a satisfying series is generated.

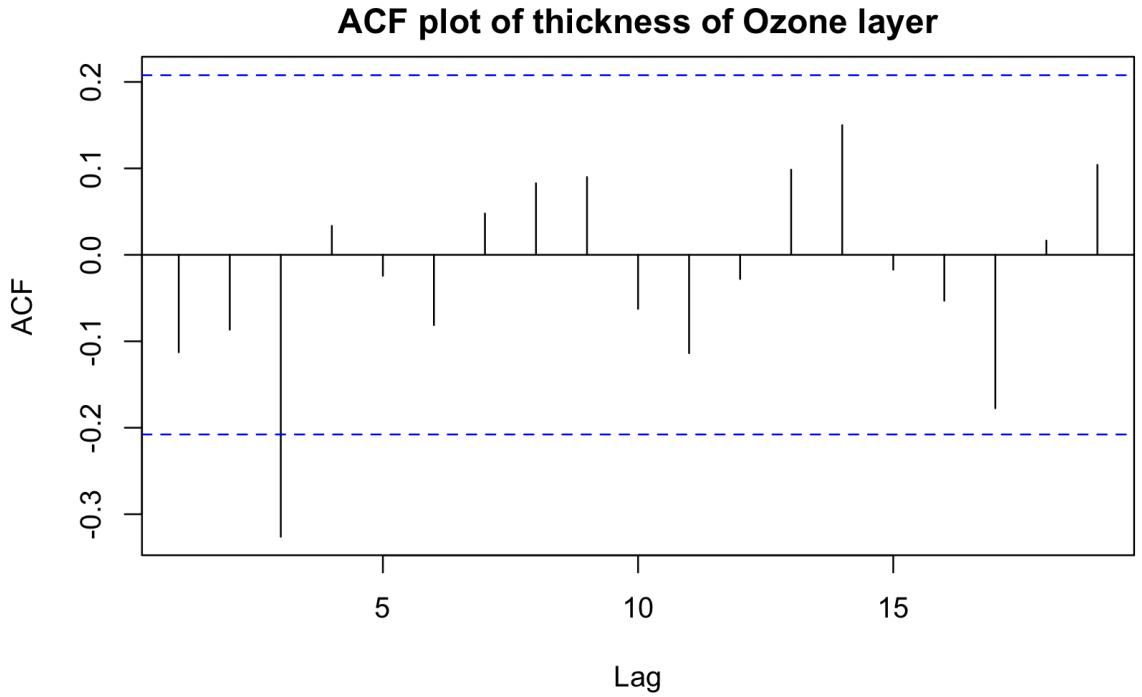
Since even after the log, our series has a trend we will try further transformation by doing the first difference of series.

First Difference Series



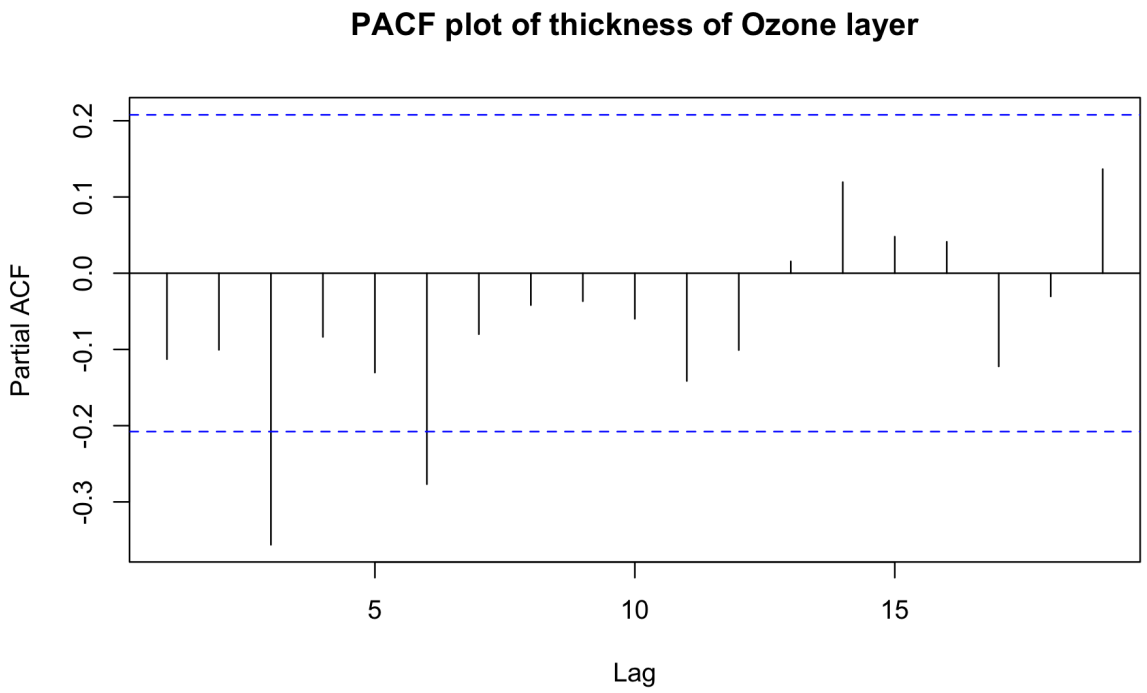
Finally, we can see this series has no trend with some variation but we can use this to get p and q values. **We will use this series in all the specification tools to find ARIMA models.**

ACF model of first difference series



From this plot, we can see one significant value crossing the boundary.

PACF model of first difference series



From the PACF plot, we can say there is one significant value crossing the boundary.

EACF

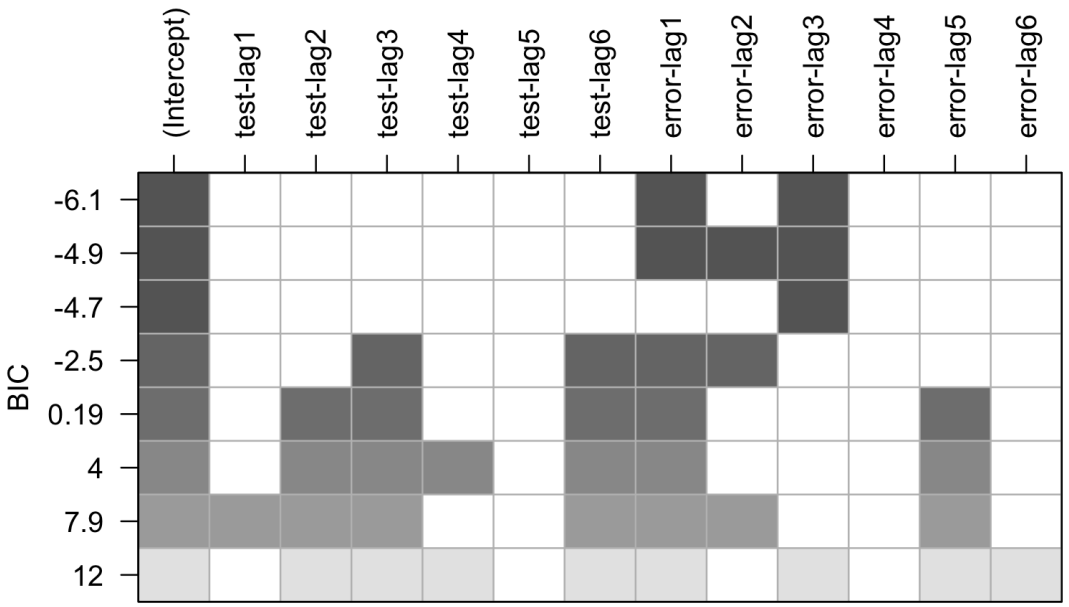
Below we can see the Extended Autocorrelation Function's output. From this table, we can easily identify the p and q model for the ARIMA model.

AR/MA

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

From this we get p = 0,1 and q = 3,4.

BIC



From the BIC Image, we can see values of p are 0 and values of q are 1, 3 and we can also consider 2 just to be safe.

Final ARIMA Models

From all the above calculation we can say our ARIMA models are p = 0, 1, q = 0, 1, 3 and d =1.

ARIMA(p,d,q) = {

From the ACF-PACF model
ARIMA(0,1,0), ARIMA(1,1,0), ARIMA(0,1,1)

From the EACF model
ARIMA(0,1,3), ARIMA(0,1,4), ARIMA(1,1,3), ARIMA(1,1,4),

From the BIC model
ARIMA(0,1,1), ARIMA(0,1,3), ARIMA(0,1,2),
}

Final models

ARIMA(p,d,q) = {
ARIMA(0,1,0), ARIMA(0,1,1), ARIMA(0,1,3), ARIMA(1,1,0), ARIMA(1,1,1), ARIMA(1,1,3),
}

Conclusion

To conclude this report in Task 1 we found the quadratic model to be the best fitting compared to other models with almost reaching 73.91% in R^2. The residual analysis provides us strong support for the quadratic model with residuals normalised, and ACF and PACF plots are indicating good autocorrelation lags. After getting the best fit model, the forecasting of the future 5 years was performed successfully which follows the downward trend of the Ozone layer thickness. In Task 2 we removed the trend behaviour from the data and used specification tools like ACF-PACF, EACF, BIC table to find the best possible parameter for ARIMA(p,d,q) model. We found that the series required at least the first difference to calculate ARIMA p and q. After working with specification we found a total of 6 possible ARIMA models.

References

Department of Agriculture, Water and the Environment. 2018. *Department of Agriculture, Water and the Environment*. [online] Available at: <https://www.environment.gov.au/protection/ozone/publications/ozone-layer-factsheet#:~:text=The%20ozone%20layer%20is%20the,B)%20radiation%20from%20the%20sun.> [Accessed 18 April 2018].

En.wikipedia.org. n.d. *Autocorrelation - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/Autocorrelation> [Accessed 21 April 2021].

Demirhan, D., n.d. In: Time Series Analysis with R, Springer.

Appendix

```
##############################################################################################

#Load all the required libraries and remove all the parameters used previously.
rm(list = ls())
library(readr)
library(xts)
library(TSA)

##############################################################################################
######################################### Task 1 ############################################
##############################################################################################

# Load csv file as dataset
dataset <- read.csv("data1.csv", header = FALSE)

# add new column to support the data visualisation from 1927 to 2016
rownames(dataset) <- seq(from=1927, to=2016)

##############################################################################################

# created ts object from dataset with frequency of 1 representing Annual time series
timeseries <- ts(dataset$V1, start = 1927, end = 2016, frequency = 1)

# Created plot of ts object
plot(timeseries,ylab='Thickness level in dobson units', xlab='Year',type='o', main = "Time Series plot
of thickness of Ozone layer 1927 to 2016")

# Created plot of ACF and PACF plot
acf(dataset, main = "PACF plot of thickness of Ozone layer")
pacf(dataset, main = "ACF plot of thickness of Ozone layer")

# created scatter plot of timeseires with one lag to compare correlation between data points
plot(y=timeseries, x=zlag(timeseries), ylab='Thickness level in dobson units',xlab='Previous Year
Thickness level', main = "Scatter Plot of thickness of Ozone layer in consequtive years")

#used cor() function in r to display correlation values
y = timeseries              # Read the color data into y
x = zlag(timeseries)        # Generate  of the color series
index = 2:length(x)    # Create an index to get rid of the first NA value in x
cor(y[index],x[index]) # Calculate correlation between numerical values in x and y```

##############################################################################################

# Created function to implement model plotting every time
modelPlotting <- function(timeseries, model, modelname){

  #Timeseries plot
  plot(timeseries,
     ylim = c(min(c(fitted(model), as.vector(timeseries))),
```

```
            max(c(fitted(model), as.vector(timeseries)))),
    ylab='Dobson units',
    xlab='Years',
    type='o',
    main = "Thickness of Ozone layer 1927 to 2016")

  #fitting model to the timeseries plot
  lines(ts(fitted(model), start = 1927))

  par(mfrow=c(3,2))

  #Histogram plot of Standardized Residuals
  hist(rstudent(model),xlab='Standardized Residuals', main = c(paste0("Histogram of the
standardized residuals from ", toString(modelname))))

  #Timeseries plot of residuals of model
plot(y=rstudent(model), x=as.vector(time(timeseries)), xlab='Time', ylab='Standardized Residuals',
type='o', main = c(paste0(toString(modelname), "'s plot of residuals with labels")))

  # Normal qq-plot of residuals of model
  y = rstudent(model)

qqnorm(y, main = c(paste0("Normal Q-Q plot of the standardized residuals from ",
toString(modelname))))
  qqline(y, col = 2, lwd = 1, lty = 2)

  # ACF plot
  acf(y, main = c(paste0("ACF plot of thickness of Ozone layer of ", toString(modelname))))

  # PACF plot
  pacf(y, main = c(paste0("PACF plot of thickness of Ozone layer of ", toString(modelname))))

  par(mfrow=c(1,1))

  #Shapiro Wilk test
  shapiro.test(y)
}

######################################## Model 1 ########################################

# Linear Model 1 summary and plots

model1 = lm(timeseries ~ time(timeseries)) # label the model as model1
summary(model1)

modelPlotting(timeseries, model1, "The Linear model")

######################################## Model 2 ########################################

# Quadratic Model 1 summary and plots

t = time(timeseries)
```

```
t2 = t^2
model2 = lm(timeseries~t+t2)
summary(model2)

modelPlotting(timeseries, model2, "The Quadratic model")

##################################### Model 3 ########################################

# Seasonal Model 1 summary and plots

timeseries1 <- ts(dataset$V1, start = 1927, frequency = 10)

year.=season(timeseries1) # period added to improve table display and this line sets up indicators
model3 = lm(timeseries1~ year.-1) # -1 removes the intercept term
summary(model3)

modelPlotting(timeseries, model3, "The Seasonal model")

##################################### Model 4 ########################################

# Harmonic Model 4 summary and plots

har.=harmonic(timeseries, 0.4) # calculate cos(2*pi*t) and sin(2*pi*t)
model4.har = lm(timeseries ~ har.-1)
summary(model4.har)

modelPlotting(timeseries, model4.har, "The Harmonic model")

##################################### Forecasting ####################################

h = 5 # 5 steps ahead forecasts
# Now we will implement the two-step algorithm

t.start = time(timeseries)[1]
t.end = time(timeseries)[length(timeseries)]

t = seq((t.end+1), (t.end+h), 1)
t2 = t^2

modelT = model2 # label the model as model1

new = data.frame(t,t2)

forecasts = predict(modelT, new, interval = "prediction")
print(forecasts)


mergedForcastsSeries <- c(timeseries, forecasts[,1])

plot(timeseries,
   ylim = c(min(c(as.vector(mergedForcastsSeries)-5)), max(c(as.vector(mergedForcastsSeries)+5))),
    xlim = c(t.start, t.end+5),
```

```
    ylab = "Dobson units",
    xlab='Years',
    type = 'o',
    main = "Thickness of Ozone layer 1927 to 2016")
# We need to convert forecasts to time series object starting from the first
# time steps-ahead to be able to use plot function.
# We do this for all columns of forecasts
lines(ts(as.vector(forecasts[,1]), start = 2017), col="red", type="l")
lines(ts(as.vector(forecasts[,2]), start = 2017), col="blue", type="l")
lines(ts(as.vector(forecasts[,3]), start = 2017), col="blue", type="l")
legend("topleft", lty=1, pch=1, col=c("black","blue","red"), text.width = 22,c("Data","5% forecast limits", "Forecasts"))


##############################################################################################
######################################### Task 2 ##############################################
##############################################################################################

# ACF and PACF model of timeseries
acf(timeseries , main = "ACF plot of thickness of Ozone layer")
pacf(timeseries , main = "PACF plot of thickness of Ozone layer")

# Converting data to positive numbers by adding absolute minimum value with small integer
positiveData <- timeseries + abs(min(timeseries)) +1
plot(positiveData,
    ylab = "Dobson units",
    xlab='Years',
    type = 'o',
    main = "Thickness of Ozone layer 1927 to 2016")

# Add log to the time series
OzoneLog <- log(positiveData)
plot(OzoneLog,
    ylab = "Dobson units",
    xlab='Years',
    type = 'o',
    main = "Thickness of Ozone layer 1927 to 2016")

# Find first difference of log data
OzoneLogDiff <- diff(log(positiveData), differences = 1)
plot(OzoneLogDiff,
    ylab = "Dobson units",
    xlab='Years',
    type = 'o',
    main = "Thickness of Ozone layer 1927 to 2016")

# ACF and PACF model of timeseries
acf(OzoneLogDiff , main = "ACF plot of thickness of Ozone layer")
pacf(OzoneLogDiff , main = "PACF plot of thickness of Ozone layer")

# EACF Model
eacf(diff(log(abc), differences = 1))
```

```
# BIC modelling
BICModel = armasubsets(y=diff(log(abc), differences = 1), nar=6, nma=6, y.name='test',
ar.method='ols')
plot(BICModel)
```