MATH1318 Time Series Analysis

Final Project

```
library(TSA)
library(timeDate)
library(tseries)

library(TSA)
library(fUnitRoots)
library(forecast)
library(CombMSC)
# rm(list=ls())


library(lmtest)
# library(FitAR)
library(bestglm)
library(ltsa)
```

Introduction

The majority of energy needs in the United States are met through the use of fossil fuels. In 2020, 35 percent of the country's energy came from petroleum, 10% from coal, and 34% from natural gas. Nuclear power contributed 9%, and renewable energy contributed 12%. Renewable energy primarily was from hydroelectric dams and biomass while wind, geothermal, and solar also had minor contributions.

After China, the United States was the world's second-largest energy consumer in 2010. In the 50 years leading up to 2006, the country's energy consumption had expanded faster than domestic energy production (when they were roughly equal). Imports significantly compensated for the disparity.

Between 1980 and 2008, the global average energy consumption climbed from 63.7 to 75 million BTU (67.2 to 79.1 GJ) per person. From 1980 to 2010, the average energy consumed in the US was around 334 million British thermal units [BTU] (352 GJ) per person. According to one interpretation, as manufacturing of equipment, cars, and other items has been shifted to other countries, transportation of finished products to US has resulted in substantial increase in glasshouse gas emissions and pollution.

BY using Electric production data set time series analysis has been undertaken. For non-stationary data i.e things that are affected by time or constantly fluctuate over time, time series analysis is used. To maintain consistency and dependability, time series analysis often requires a high number of data points.

Time series analysis is used to figure out what's causing trends or systemic patterns across time. Business users can use data visualizations to discover seasonal trends and learn more about causation. These visualizations can now go much beyond line graphs, thanks to new analytics technologies.

Time series forecasting is used to predict the likelihood of future events when they evaluate data at regular intervals. Predictive analytics includes time series forecasting. It can reveal likely data changes such as seasonality or cyclic behavior allowing for better understanding of data factors and better forecasting.

Objective

The aim of this report is to predict electric production for the next 10 months in US. This includes comprehensive analysis like descriptive analysis, proper visualization, model specification, model fitting, model selection, diagnostic checking and interpretations of the results.

Data and Data preprocessing

The Electric Production IP index in US using data from 1985 to 2018 till January. The data set contains 397 observations with 2 columns. The Electric Production data set is sourced from the website https://www.kaggle.com/code/ludovicocuoghi/electric-production-forecast-lstm-sarima-mape-2-5/data

```
#Reading the CSV file
electric_data <- read.csv("Electric_Production.csv", header = TRUE)
head(electric_data)
```

```
##         DATE   Value
## 1 01-01-1985 72.5052
## 2 02-01-1985 70.6720
## 3 03-01-1985 62.4502
## 4 04-01-1985 57.4714
## 5 05-01-1985 55.3151
## 6 06-01-1985 58.0904
```

```
class(electric_data)
```

```
## [1] "data.frame"
```

Summary Statistics

```
# subsetting data
electric_data <- subset (electric_data, select = -DATE)

#Checking Dimensions of the data
dim(electric_data)
```

```
## [1] 397   1
```

```
str(electric_data)
```

```
## 'data.frame':    397 obs. of  1 variable:
##  $ Value: num  72.5 70.7 62.5 57.5 55.3 ...
```

```
#Checking summary of the data
summary(electric_data)
```

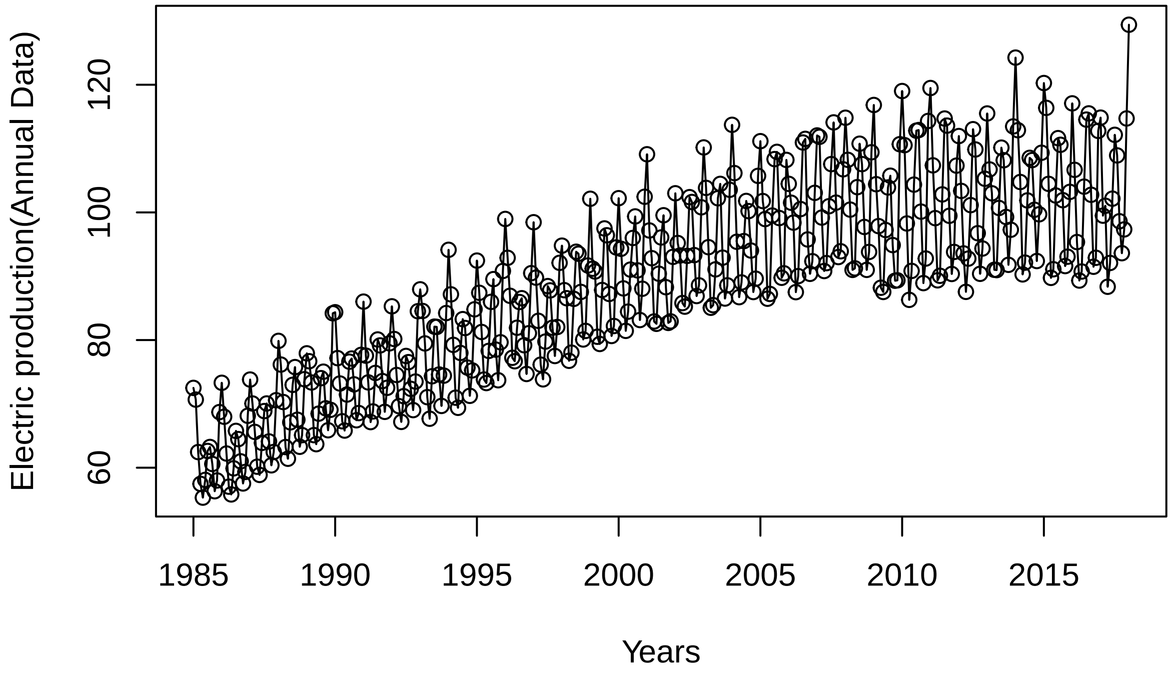
```
##       Value
##  Min.   : 55.32
##  1st Qu.: 77.11
##  Median : 89.78
##  Mean   : 88.85
##  3rd Qu.:100.52
##  Max.   :129.40
```

With a mean of 88.85 and a median of 89.78, the minimum electric production is 55.32 and the maximum electric production is 129.40 (variation is high between min and max values). Because the mean and median are nearly equal, the data is distributed symmetrically.

Time Series Plot for Electric production(Annual Data)

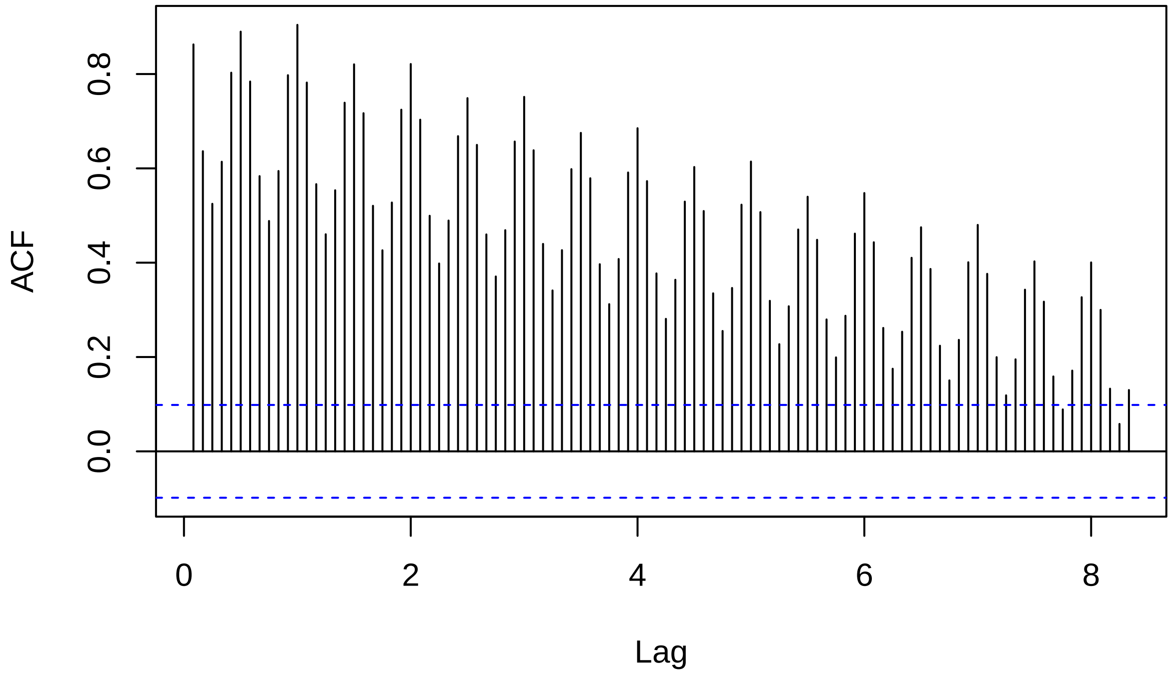
```
par(mfrow=c(1,1))
ts_electric_data <- ts(as.vector(electric_data),start = c(1985,1), end=c(2018,1), frequency = 12)
plot(ts_electric_data,type='o', ylab ='Electric production(Annual Data)', xlab = 'Years', main=" Figure
```

Figure 1: Time series plot for Electric production



```
acf(ts_electric_data,lag.max = 100, main=" Figure 2: ACF plot for Electric production")
```

Figure 2: ACF plot for Electric production



```
ts_electric_data <- ts(as.vector(electric_data),start = c(1,1), frequency = 6)
```

Used ts() function to convert a data frame into a ts object for Time Series analysis with frequency=12 for annual data. After verifying the ACF plot, the frequency is adjusted to 6.

```
par(mfrow=c(1,1))

#Time Series plot for Electric production data
plot(ts_electric_data,type='o', ylab ='Electric production(Annual Data)', xlab = 'Years',main=" Figure
```

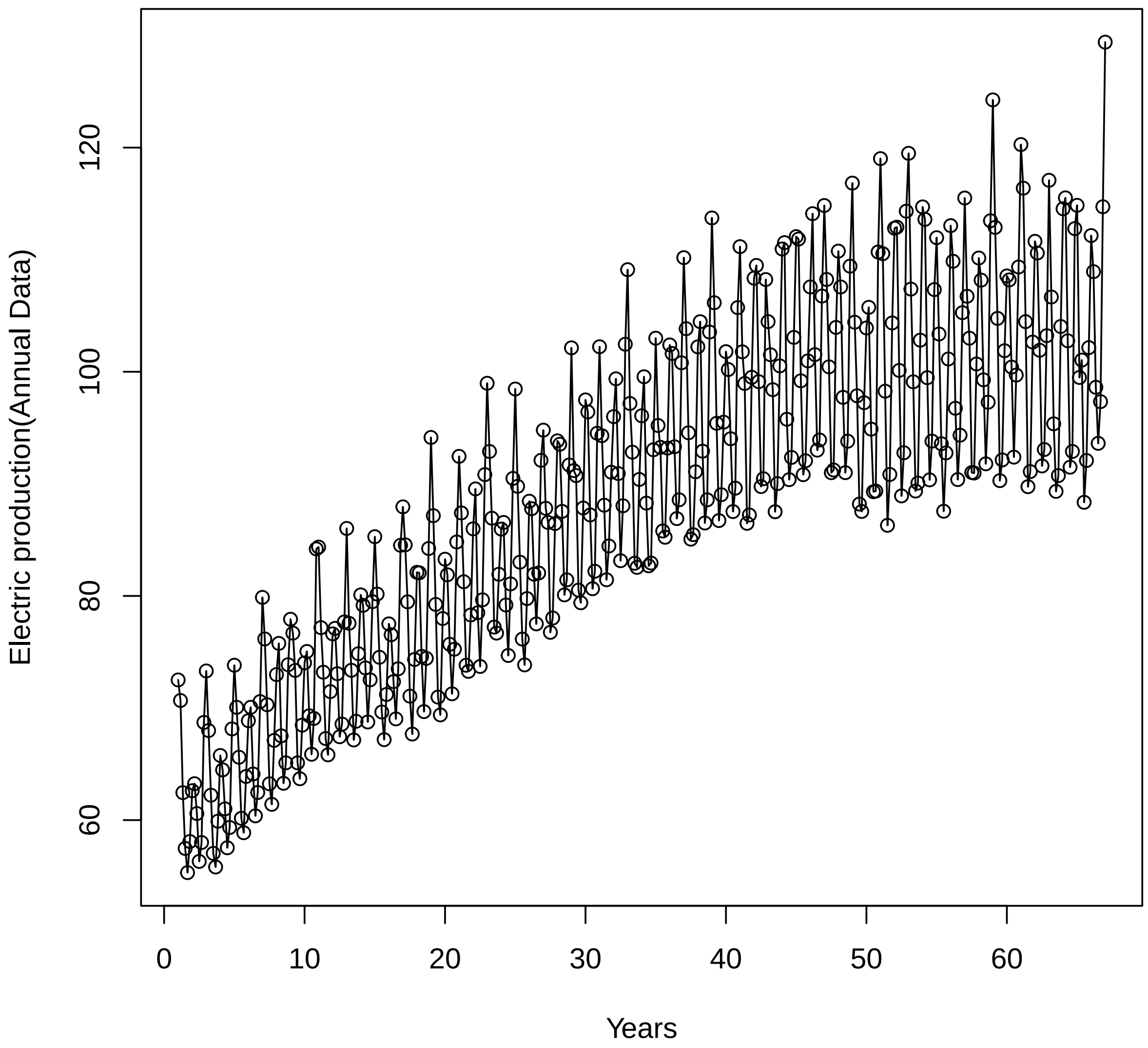


Figure 3: Time series plot for Electric production

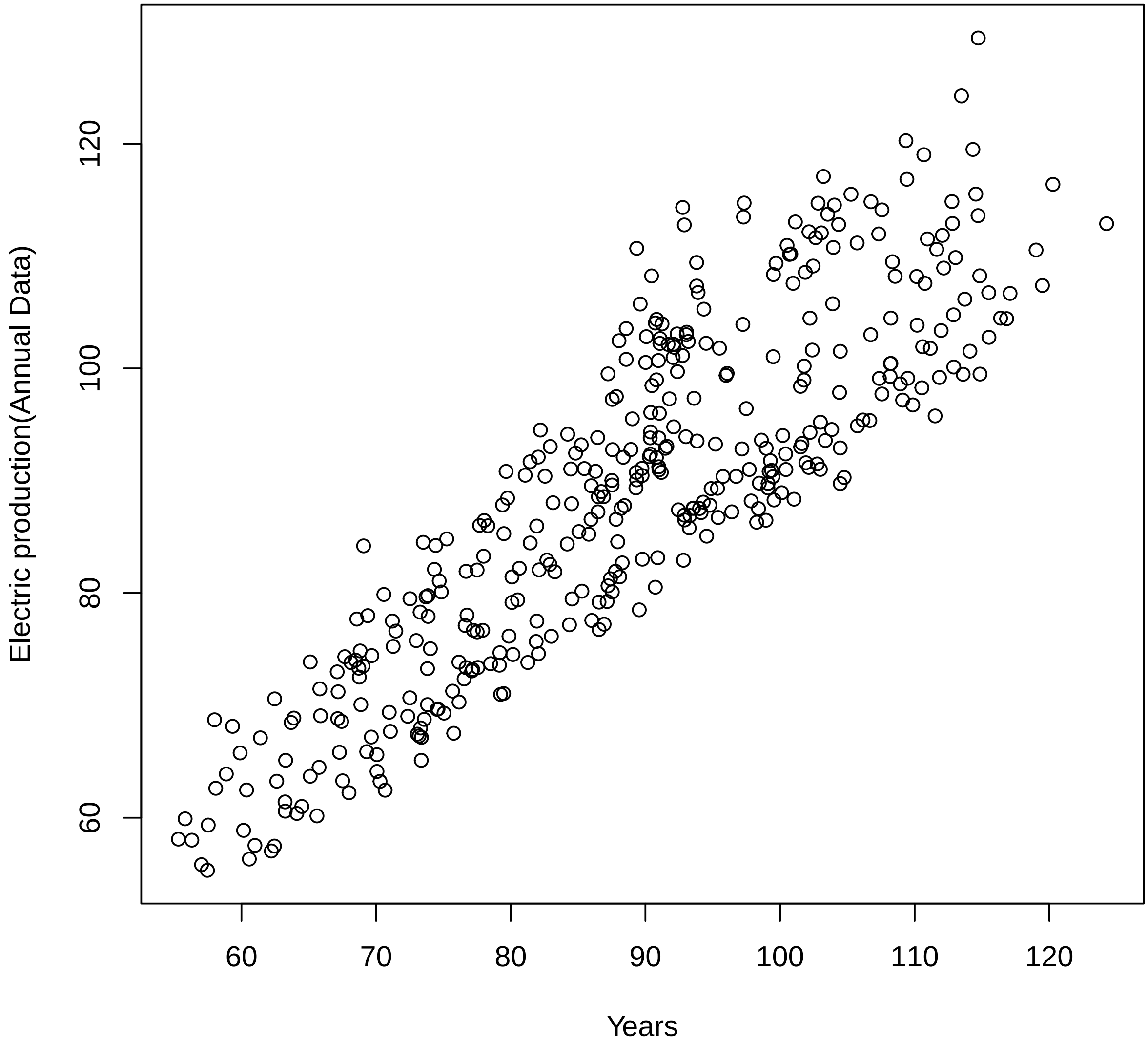
Observations from the Time Series plot

- **Trend** – There is a slightly upward trend and the shape of the graph is linear.
- **Seasonality -** There are multiple seasonality in data and confirmed the same using ACF Plot.
- **Behaviour** – It has both AR and MA behaviour.
- **Variance** – There is no variance.
- **Intervention** – There is no clear intervention in the plot.

Scatter plot of first lag for Electric production(Annual Data)

```
#Scatter plot for first Electric production data
plot(y=ts_electric_data,x=zlag(ts_electric_data),ylab ='Electric production(Annual Data)', xlab = "Year
```

Figure 4:Scatter plot for Electric production(Annual Data)



We can see that there is a positive correlation for initial lag in the graph above.

```
#Checking correlation of first lag
y = ts_electric_data
x = zlag(ts_electric_data)
# Create an index to get rid of the first NA value in x
index = 2:length(x)
```

```
# Calculate correlation between numerical values in x and y
cor(y[index],x[index])
```

```
## [1] 0.8717309
```

The scatter plot of Electric production data (Annual Data) and its first lag reveals a substantial positive link between the lag and the present value. In this scatter plot, the correlation is 87 percent.

Checking Stationary of data

QQ-Plot Electric production(Annual Data)___

The relationship between a particular sample and the normal distribution is depicted by the Q-Q plot. Using QQ-Plot, you can see if the points in a Q-Q plot are on a straight diagonal line and if the data is regularly distributed.

```
#Function1
Function1 <- function(Data, FigureNumber) {

  #QQ Plot
  qqnorm(Data, main = paste( paste("Figure",FigureNumber+1),": QQ plot for Electric production(Annual Da
  qqline(Data, col = 2)

   par(mfrow=c(1,2))

  # Plot the ACF and PACF plots of time series data
  acf(Data, main = paste( paste("Figure",FigureNumber+2),": ACF plot for Electric production(Annual Data
  pacf(Data, main = paste( paste("Figure",FigureNumber+3),": PACF plot for Electric production(Annual Da

  #Dickey-Fuller Test
  k = trunc((length(x)-1)^(1/3))
  adf=adf.test(Data, k=k, alternative = c("stationary"))
  print(adf)

  #Shapiro walk Test
  shapiro=shapiro.test(Data)
  print(shapiro)

  #PP-test
  pp=pp.test(Data)
  print(pp)


}
Function1(ts_electric_data,4)
```

Figure 5 : QQ plot for Electric production(Annual Data)

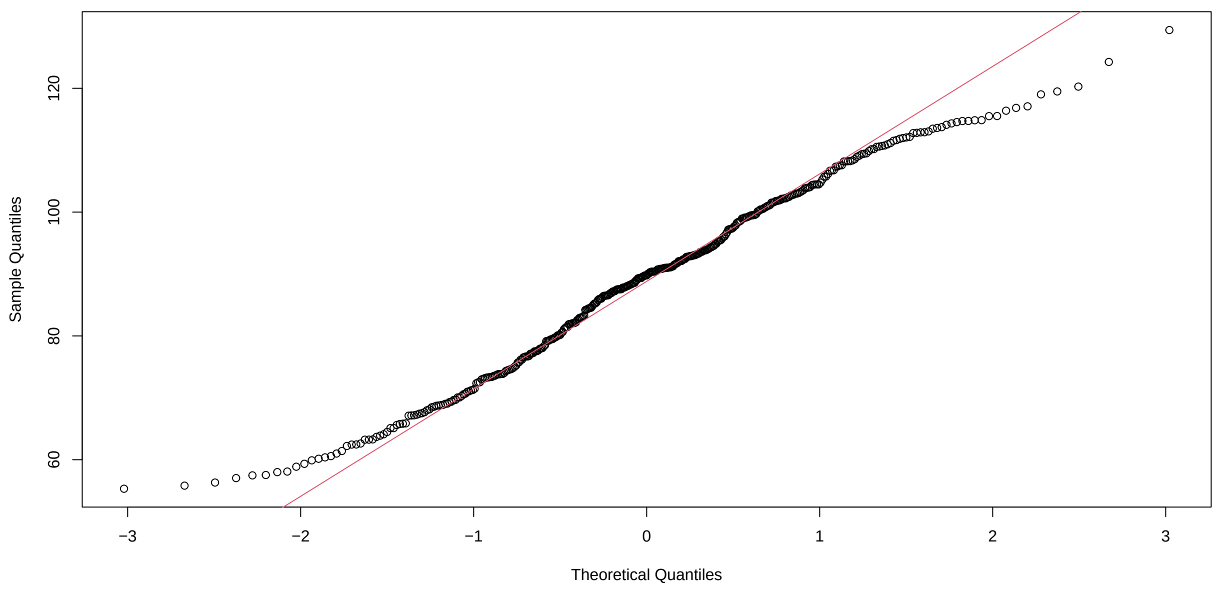


Figure 6 : ACF plot for Electric production(Annual Data)

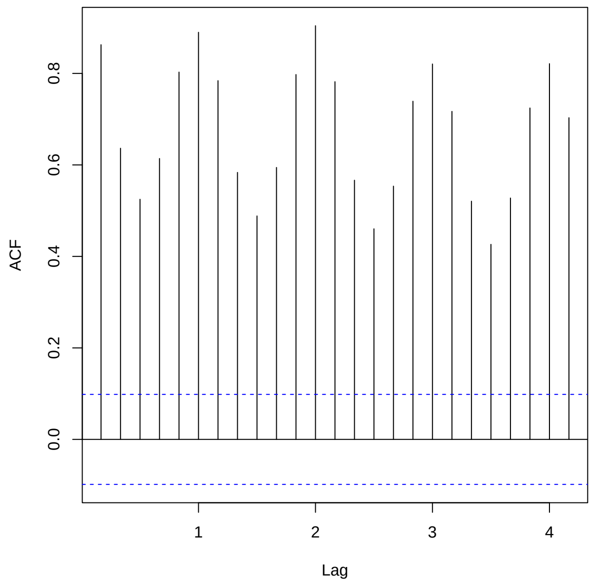
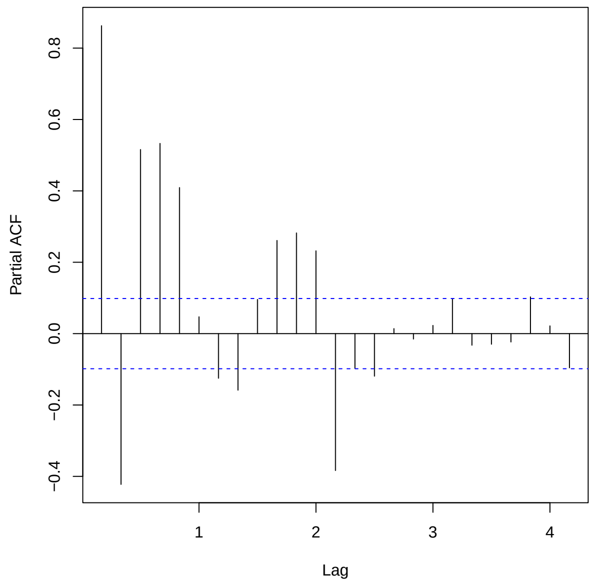


Figure 7 : PACF plot for Electric production(Annual Data)



```
##
##  Augmented Dickey-Fuller Test
##
## data:  Data
## Dickey-Fuller = -5.139, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
##
##
##  Shapiro-Wilk normality test
##
## data:  Data
## W = 0.98707, p-value = 0.001335
##
##
##  Phillips-Perron Unit Root Test
##
## data:  Data
## Dickey-Fuller Z(alpha) = -88.143, Truncation lag parameter = 5, p-value = 0.01
## alternative hypothesis: stationary
```

Figure-5: Observations from the QQ plot

Because the data is not aligned on the line, the QQ Plot does not show normality.

Figure-6 and 7: Observations from the ACF and PACF Plots

ACF describes the relationship between the current value of a time series and its previous values (1-unit past, 2-unit past,…, n-unit past). The partial correlation coefficients between the series and its lags are represented by the PACF graphic.

We can see from the ACF plot that there is a declining trend with seasonality, indicating that the data is non-stationary. PACF likewise has a large significant lag 1, indicating that it is non-stationary. For statistical analysis, the Dickey-Fuller test (unit root test) can be used.

Dickey-Fuller Test for Electric production(Annual Data)

Observations from the Dickey-Fuller Test

The adf test is a popular statistical technique for determining whether or not a time series is stationary. Because the p-value is 0.01 i.e <0.05, the data is steady (reject null hypothesis).

Shapiro Test Walk

Observations from the Shapiro walk test

The Shapiro walk test is a statistical test that compares the sample distribution to a normal distribution to detect substantial deviation from normality in the data. The p-value of 0.001 i.e <0.05 suggests that the model is not regularly distributed (reject null hypothesis). It is necessary to combine visual inspection (QQ Plot) and significance testing for optimal conclusion (Shapiro walk Test).

Phillips–Perron (PP) test

Observations from the Phillips–Perron(PP) Test after first differencing

The Phillips–Perron test is a unit root test that is used to determine whether or not the null hypothesis is true. The augmented Dickey–Fuller test is based on the Dickey–Fuller null hypothesis test because they both address the same issue. The p-value for the Phillips–Perron(PP) test is 0.01 i.e <0.05. (rejecting null hypothesis).

Transformation

Box-Cox transformation for Electric production(Annual Data)

It's used to convert data from a non-normal distribution to a normal distribution (stabilize the variance). The box-cox transformation encompasses both logarithmic and power transformations, depending on the parameter.

The transformation of Y has the form:

y( ) = (y − 1) /   if y   0 y( ) = log(y) if y = 0

It can range from -5 to 5. Power transformation generates a square root transformation that is beneficial with Poisson-like data when 0.5 is used. It is a log transformation when the value is 0, no transformation when the value is 1, and a reciprocal transformation when the value is 1.

We are using the box-cox transformation since the ACF and PACF graphs are non-stationary.

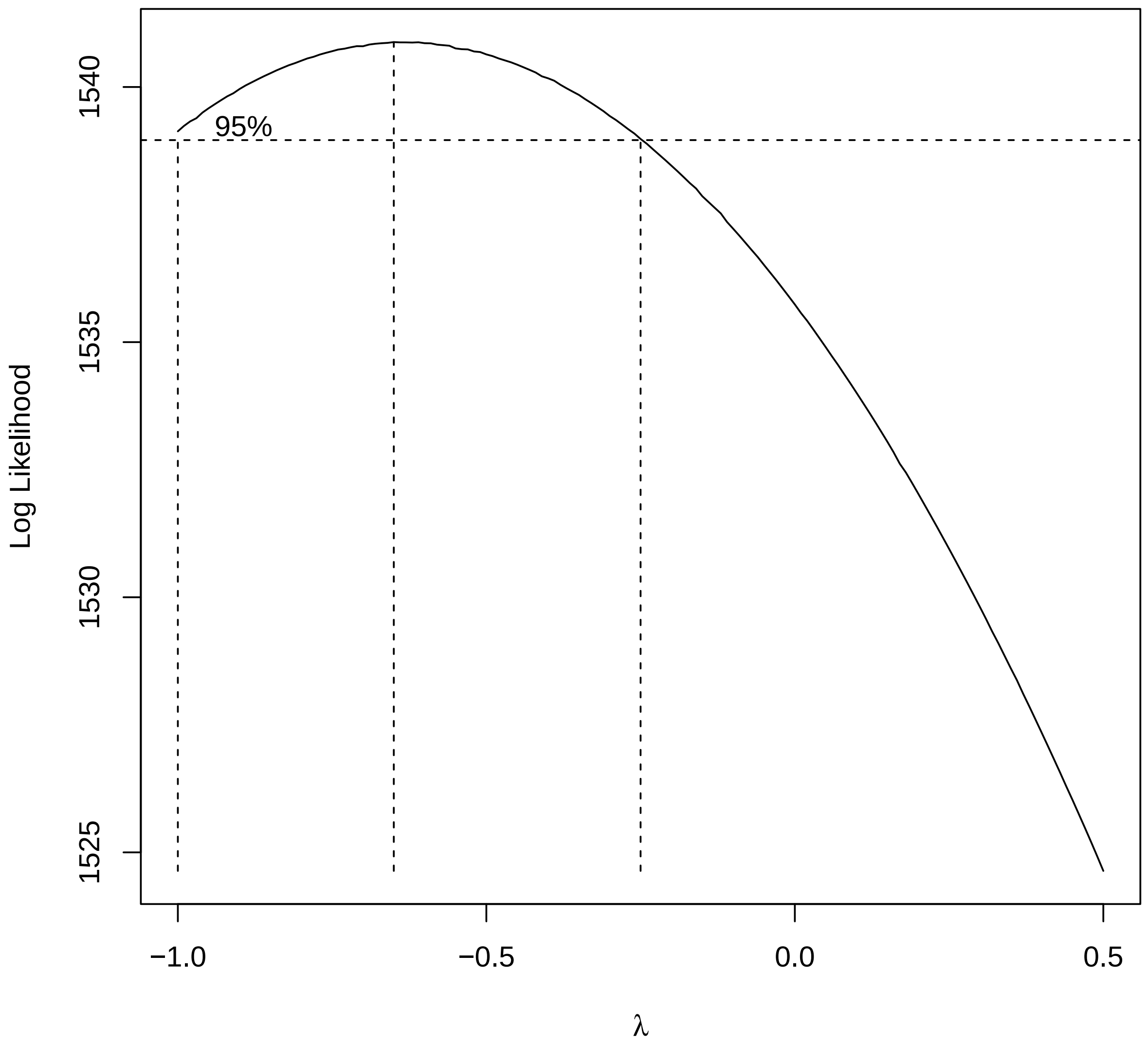
Box-Cox transformation for Electric production(Annual Data)

```
# Perform Box-Cox transformation
BC <- BoxCox.ar(ts_electric_data,lambda = seq(-1, 0.5, 0.01) )
BC$ci
```

```
## [1] -1.00 -0.25
```

```
title(main = "Figure 8: Box-Cox transformation for Electric production(Annual Data)")
```

Figure 8: Box−Cox transformation for Electric production(Annual Data)



Observations from the Box-Cox transformations

From the preceding Box-Cox figure, the confidence interval is -1 (lower limit) to -0.25 (upper bound).

```
#Checking lambda value
lambda <- BC$lambda[which(max(BC$loglike) == BC$loglike)]
lambda
```

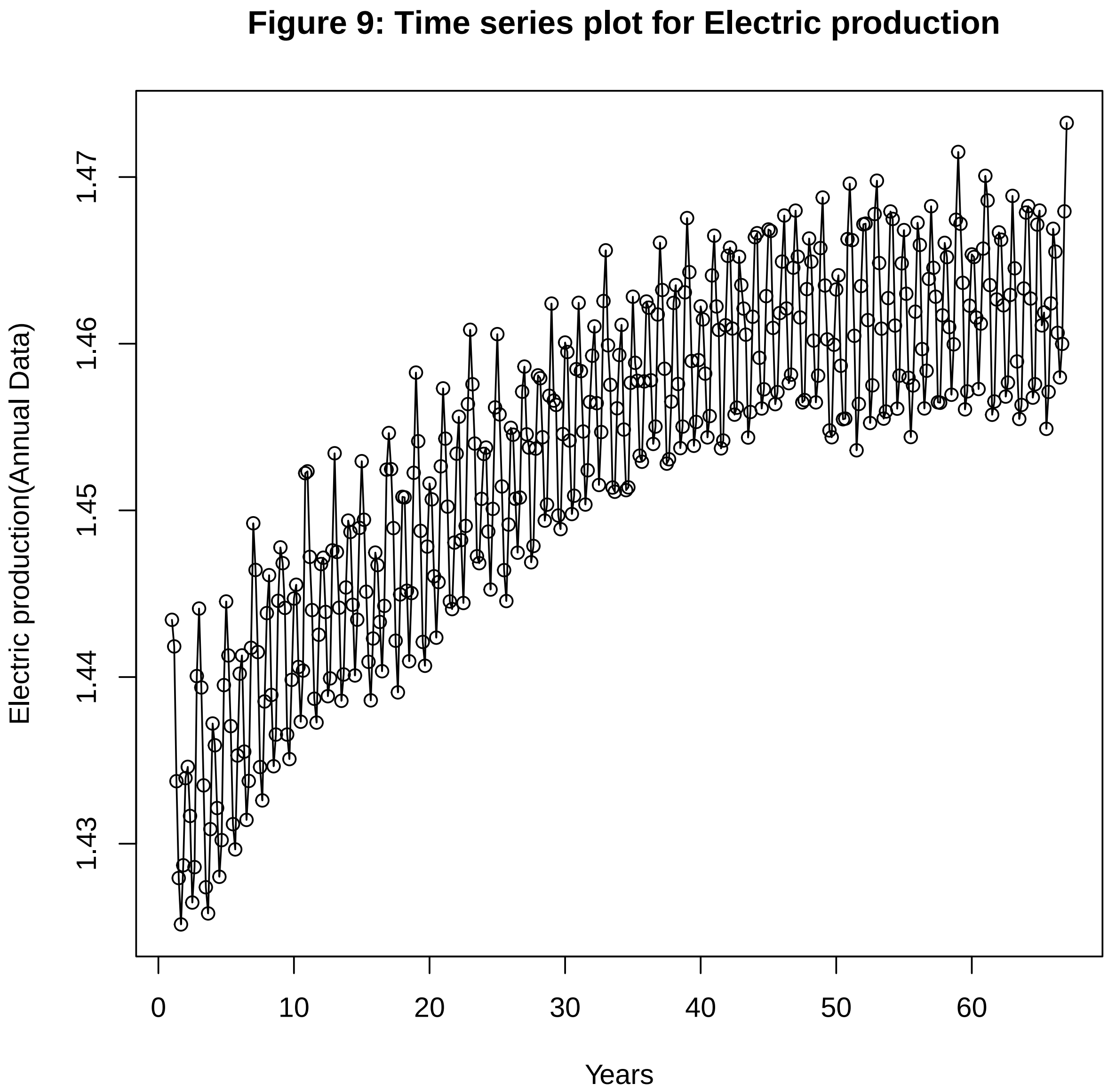
```
## [1] -0.65
```

The box-cox's optimal point is -0.65, which is near to 0.5, indicating that the transformation is reciprocal of square root. To make the time series data steady, the Box-Cox transformation is utilised.

Time series plot of Box-Cox transformation for Electric production(Annual Data)

```
# Apply Box-Cox transformation
BC.electric_ts_data = (ts_electric_data^lambda-1)/lambda

#Time Series Plot for Box-Cox transformation
plot(BC.electric_ts_data,type='o',ylab ='Electric production(Annual Data)', xlab = 'Years',
     main = " Figure 9: Time series plot for Electric production")
```



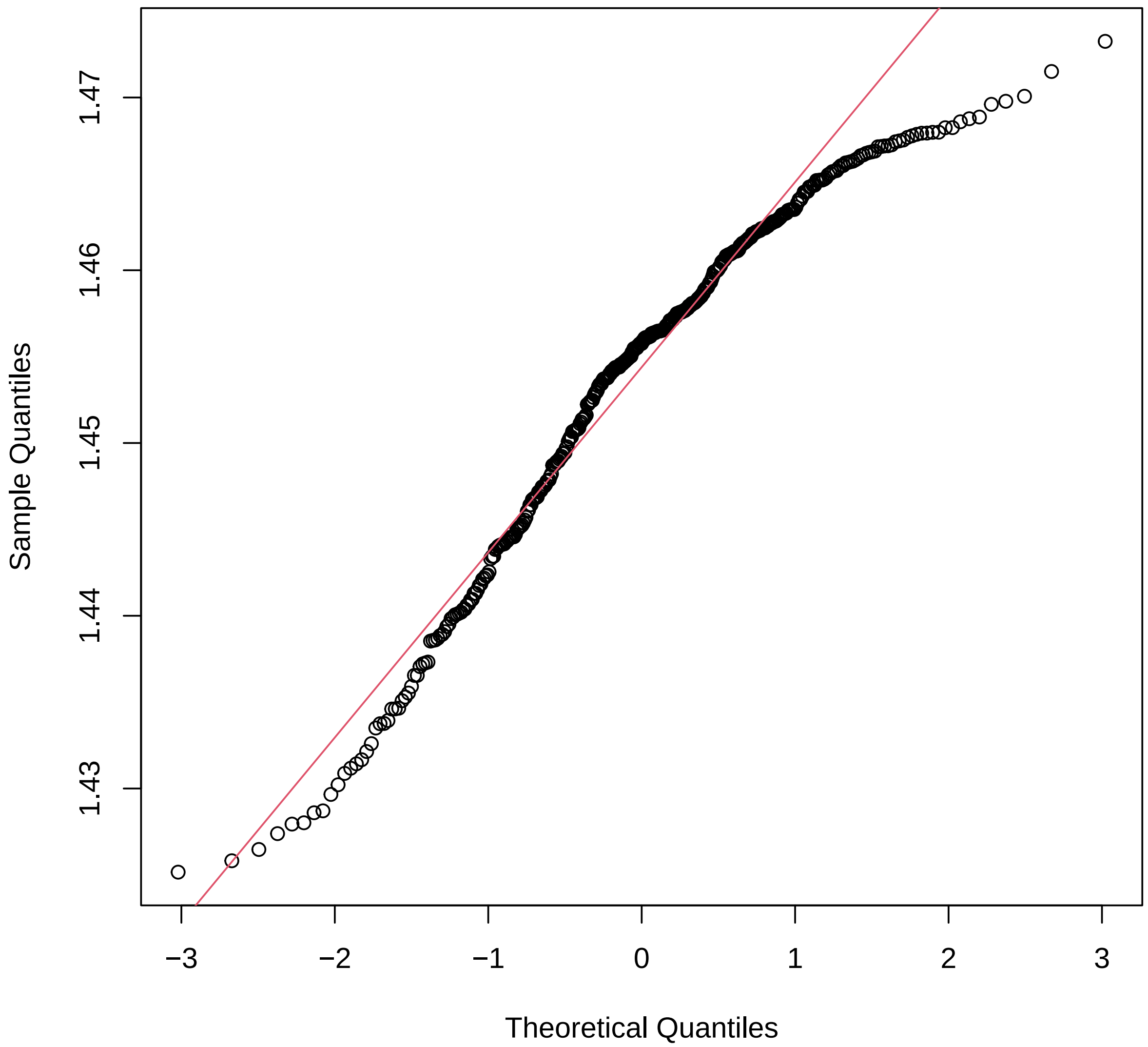
Observations from the Box-Cox transformation Time series plot

There is a change in the y-axis values in the graph above.

```
#Normality Check for Box-Cox data transformation
#QQ Plot
qqnorm(BC.electric_ts_data, main = " Figure 10: QQ plot for Box-Cox transformation")
```

```
qqline(BC.electric_ts_data, col = 2)
```

Figure 10: QQ plot for Box−Cox transformation



```
#Shapiro walk Test
shapiro.test(BC.electric_ts_data)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  BC.electric_ts_data
## W = 0.95746, p-value = 2.659e-09
```

Observations from the Box-Cox transformation QQ plot and Shapiro walk Test

Both the QQ Plot and the Shapiro Walk Test change following transformation (normality).

Seasonal Modeling

We observe that the time series plot has trend as well as seasonality(repeating pattern).We use SARIMA Models to capture time series with seasonality & trend.Seasonal Differencing is used to deal with non-stationary seasonal data. If the mean of time series is not constant over time it is said to be non-stationary.

SARIMA(p,d,q) X (P,D,Q)s

P:AR order

d:The number of ordinary differences

q:MA order

P:Seasonal AR Order

D:The number of seasonal differences

Q:Seasonal MA Order

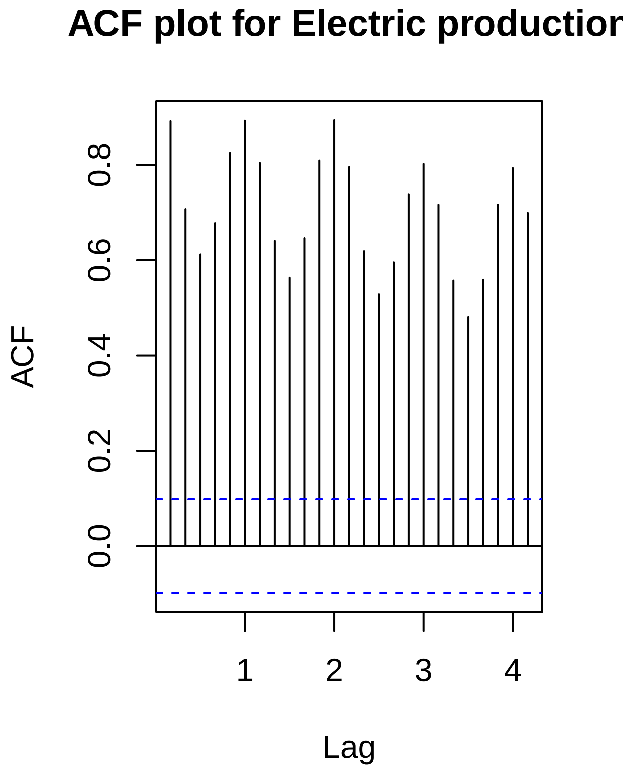
s:Period

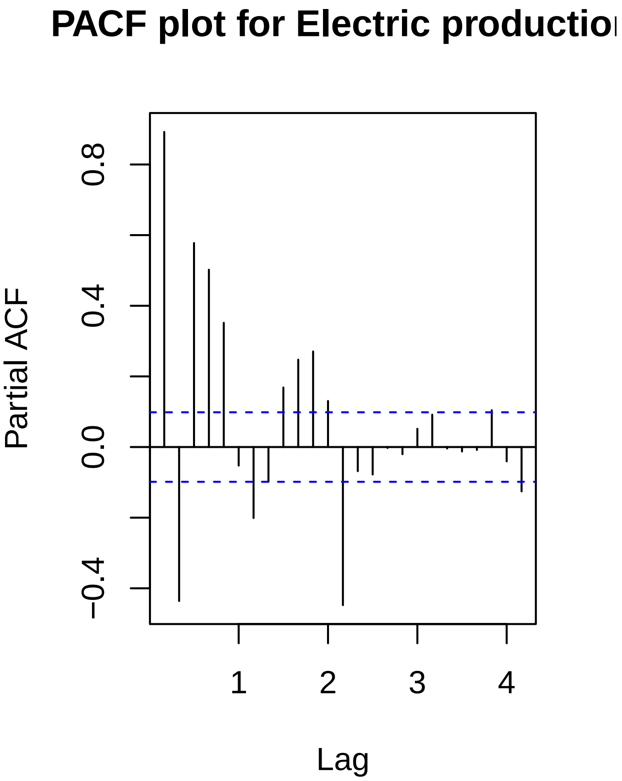
Here, If we look at the Seasonal Lags of PACF & ACF we get P & Q respectively and from the Lags before First Seasonal Lag we get p,q.

Seasonal Difference of Period s for Series Yt is, sYt=Yt-Yt-s

For Seasonality, its better to check ACF & Pacf.

```
par(mfrow=c(1,2))
acf(BC.electric_ts_data, main ="ACF plot for Electric production")
pacf(BC.electric_ts_data, main ="PACF plot for Electric production")
```





From the ACF & PACF plot we observe that there is a wave pattern and a slowing decaying pattern at seasons.We have significant seasonal lags at 1,2,3… of ACF plot and a very high first lag in PACF. Significant lags with slowing decaying pattern shows us Seasonal trend.So we need to deal with seasonality & Seasonal trend.

We would be following the residual approach as we know Whatever is not captured by the Model goes into the residuals. So we look at the residuals and try to improve/tweak the model.

In our seasonal series, we have a variety of elements we observe Ordinary trend, seasonal trend, ordinary Autocorrelation, and seasonal Autocorrelation. Whatever is left after we capture the seasonal trend goes to the residuals. Then we examine the residuals to see what's left, fit the model again, and check what's left, one characteristic at a time. When we have the perfect model, the residuals will be white noise because the models will have captured everything.

Seasonal Differening

In the ACF, the effect of seasonal trend can be observed. As a result, we'll fit a plain model with D = 1 and examine the residuals. We always start with D=1, because we know there is a seasonal tendency, and we set D=1 to remove it. The rest of the models parameters are set to 0.As a result, the seasonal trend is removed, and the rest is captured by the residuals.

```
FitSeasonalModel <- function(TimeSeries,p,d,q,P,D,Q,period,FigureNumber) {
  model = arima(BC.electric_ts_data,order=c(p,d,q),seasonal=list(order=c(P,D,Q), period=period))
  residual = residuals(model);
  # par(mfrow=c(1,1))
  plot(residual,xlab='Time',ylab='Residuals', main =paste( paste("Figure",FigureNumber),":
  Time series plot of the residuals "))
  par(mfrow=c(1,2))
  acf(residual, lag.max = 40, main =paste( paste("Figure",FigureNumber+1),":
  ACF of residuals "))
  pacf(residual, lag.max = 40, main = paste( paste("Figure",FigureNumber+2),":
  PACF of residuals "))
  return(residual)
}
```

```
res.m1=FitSeasonalModel(BC.electric_ts_data,p=0,d=0,q=0,P=0,D=1,Q=0,period = 6,FigureNumber = 1)
```

Fitting Model with D=1

Figure 1 :
Time series plot of the residuals

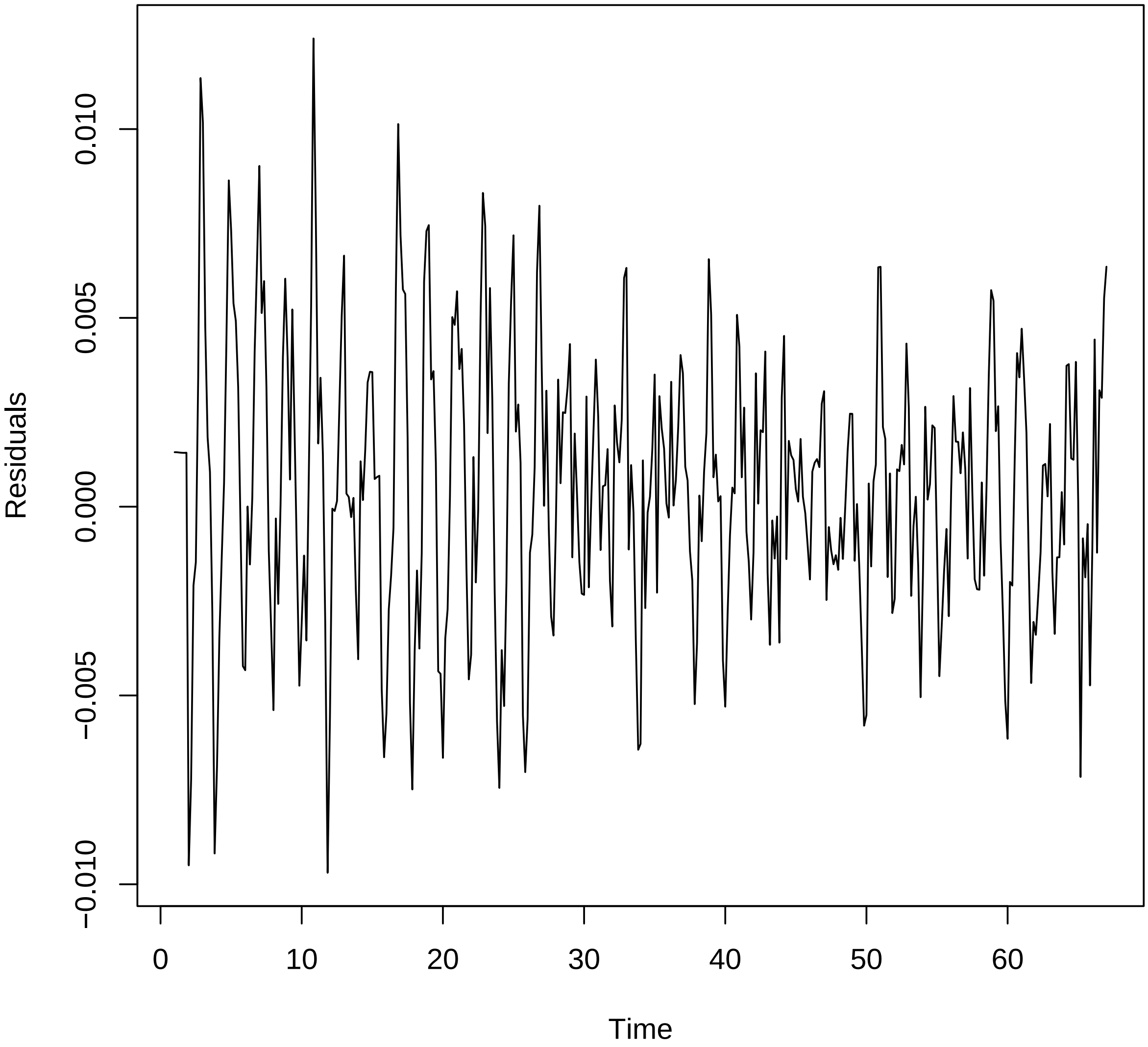


Figure 2 :
ACF of residuals

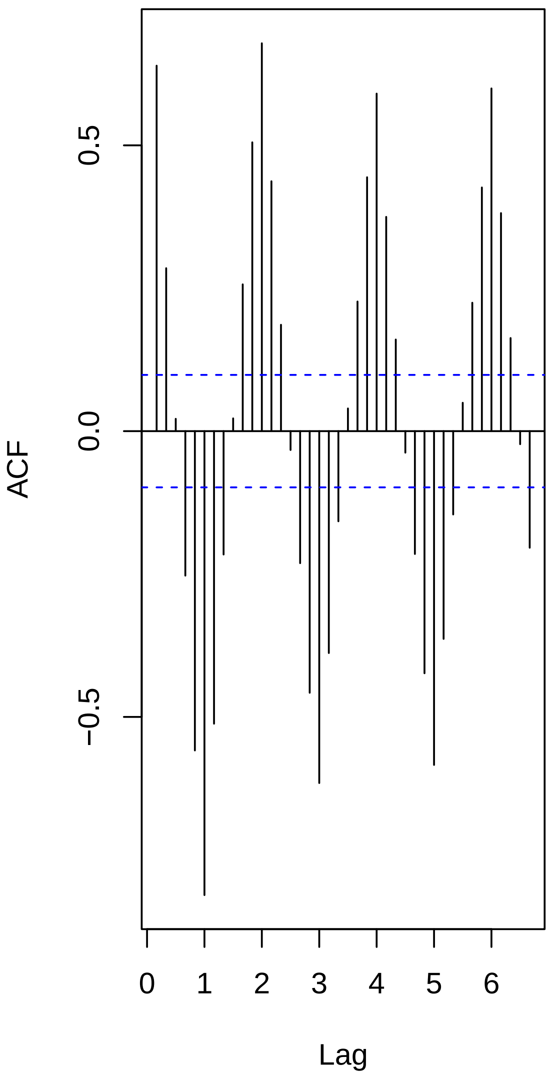
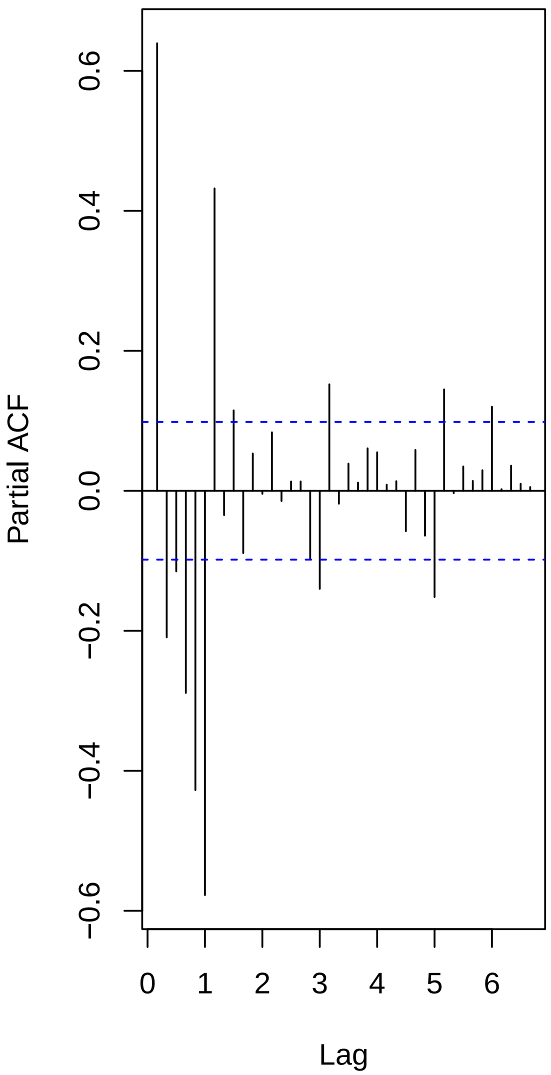


Figure 3 :
PACF of residuals


Partial ACF
Lag