ENV 790.30 - Time Series Analysis for Energy Data | Spring 2025

Assignment 2 - Due date 01/23/25

Alex Lopez

Submission Instructions

You should open the .rmd file corresponding to this assignment on RStudio. The file is available on our class repository on Github.

Once you have the file open on your local machine the first thing you will do is rename the file such that it includes your first and last name (e.g., "LuanaLima_TSA_A02_Sp24.Rmd"). Then change "Student Name" on line 4 with your name.

Then you will start working through the assignment by **creating code and output** that answer each question. Be sure to use this assignment document. Your report should contain the answer to each question and any plots/tables you obtained (when applicable).

When you have completed the assignment, **Knit** the text and code into a single PDF file. Submit this pdf using Sakai.

R packages

R packages needed for this assignment:"forecast","tseries", and "dplyr". Install these packages, if you haven't done yet. Do not forget to load them before running your script, since they are NOT default packages.\

```
#Load/install required package here
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(tseries)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(here)
```

```
## here() starts at /home/guest/ENERGY797
```

```r
library(ggplot2)

#load packages to import Excel files
library(readxl)
library(openxlsx)
```

Data set information

Consider the data provided in the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source on our **Data** folder. The data comes from the US Energy Information and Administration and corresponds to the December 2023 Monthly Energy Review. The spreadsheet is ready to be used. You will also find a *.csv* version of the data "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source-Edit.csv". You may use the function *read.table*() to import the *.csv* data in R. Or refer to the file "M2_ImportingData_CSV_XLSX.Rmd" in our Lessons folder for functions that are better suited for importing the *.xlsx*.

```r
#Importing data set
energy.data <-
  read_excel(path = "./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
             skip = 12, sheet = 'Monthly Data', col_names = FALSE)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
```

```r
#extract column names
read_col_names <-
  read_excel(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
             skip = 10,n_max = 1, sheet="Monthly Data",col_names=FALSE)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
```

```
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
```

```
#assign column names
colnames(energy.data) <- read_col_names

#check first few rows of data
head(energy.data)
```

```
## # A tibble: 6 x 14
##   Month               `Wood Energy Production` `Biofuels Production`
##   <dttm>                                 <dbl> <chr>
## 1 1973-01-01 00:00:00                     130. Not Available
## 2 1973-02-01 00:00:00                     117. Not Available
## 3 1973-03-01 00:00:00                     130. Not Available
## 4 1973-04-01 00:00:00                     125. Not Available
## 5 1973-05-01 00:00:00                     130. Not Available
## 6 1973-06-01 00:00:00                     125. Not Available
## # i 11 more variables: `Total Biomass Energy Production` <dbl>,
## #   `Total Renewable Energy Production` <dbl>,
## #   `Hydroelectric Power Consumption` <dbl>,
## #   `Geothermal Energy Consumption` <dbl>, `Solar Energy Consumption` <chr>,
## #   `Wind Energy Consumption` <chr>, `Wood Energy Consumption` <dbl>,
## #   `Waste Energy Consumption` <dbl>, `Biofuels Consumption` <chr>,
## #   `Total Biomass Energy Consumption` <dbl>, ...
```

Question 1

You will work only with the following columns: Total Biomass Energy Production, Total Renewable Energy Production, Hydroelectric Power Consumption. Create a data frame structure with these three time series only. Use the command head() to verify your data.

```
energy.df <- energy.data %>%
  select(4, 5, 6)

head(energy.df)
```

```
## # A tibble: 6 x 3
##   Total Biomass Energy Productio~1 Total Renewable Ener~2 Hydroelectric Power ~3
##                              <dbl>                  <dbl>                   <dbl>
## 1                             130.                   220.                    89.6
## 2                             117.                   197.                    79.5
## 3                             130.                   219.                    88.3
## 4                             126.                   209.                    83.2
## 5                             130.                   216.                    85.6
## 6                             126.                   208.                    82.1
## # i abbreviated names: 1: 'Total Biomass Energy Production',
## #   2: 'Total Renewable Energy Production',
## #   3: 'Hydroelectric Power Consumption'
```

Question 2

Transform your data frame in a time series object and specify the starting point and frequency of the time series using the function ts().

```
ts.energy <- ts(energy.df, start = c(1973,1), frequency = 12)
```

Question 3

Compute mean and standard deviation for these three series.

```
#compute mean for each series
mean.biomass <- mean(ts.energy[,1])
mean.renewable <- mean(ts.energy[,2])
mean.hydro <- mean(ts.energy[,3])

mean.biomass
```

```
## [1] 282.6779
```

```
mean.renewable
```

```
## [1] 402.0167
```

```r
mean.hydro
```

```
## [1] 79.55371
```

```r
#compute standard deviation for each series
sd.biomass <- sd(ts.energy[,1])
sd.renewable <- sd(ts.energy[,2])
sd.hydro <- sd(ts.energy[,3])

sd.biomass
```

```
## [1] 94.05815
```

```r
sd.renewable
```

```
## [1] 143.7927
```

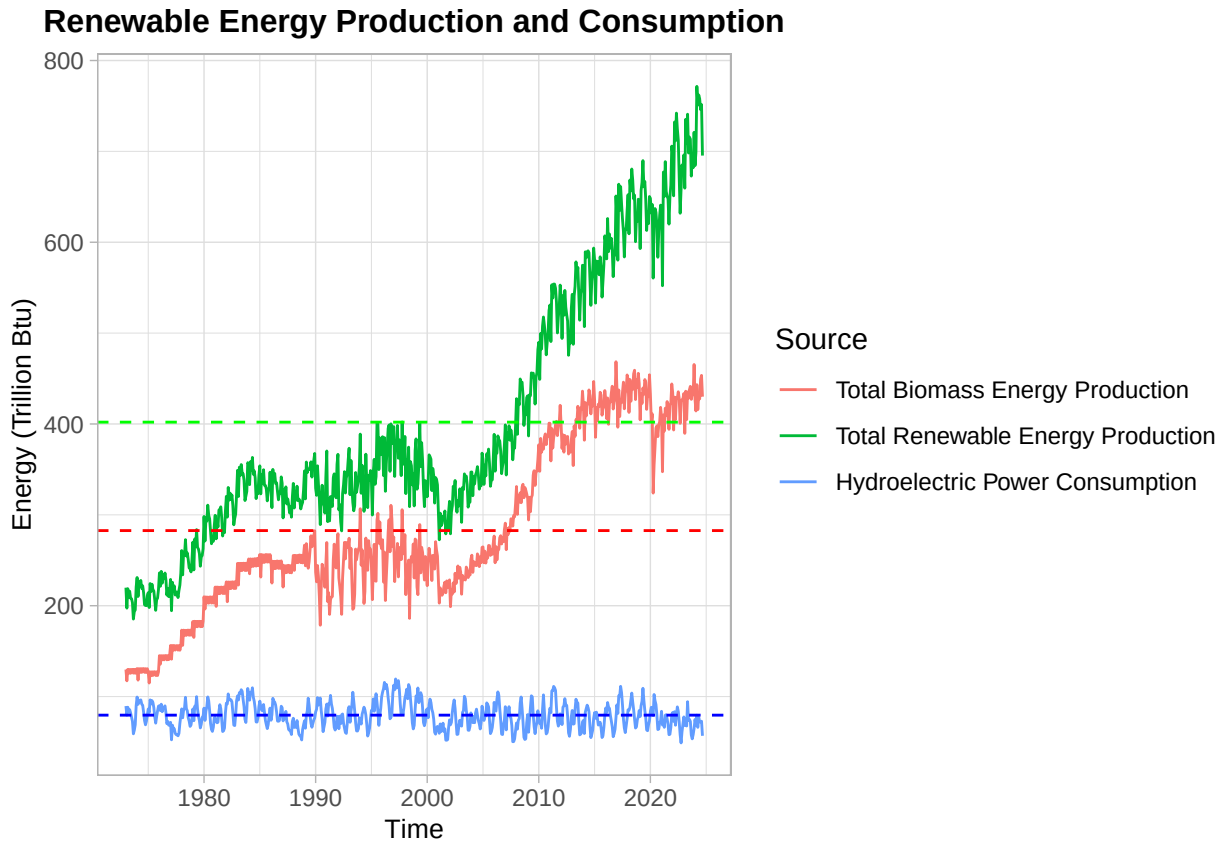
```r
sd.hydro
```

```
## [1] 14.10737
```

Question 4

Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```
autoplot(ts.energy) +
  theme_light() +
  xlab('Time') +
  ylab('Energy (Trillion Btu)') +
  labs(color = 'Source') +
  geom_hline(yintercept = mean.biomass, color = 'red', linetype = 'dashed') +
  geom_hline(yintercept = mean.renewable, color = 'green', linetype = 'dashed') +
  geom_hline(yintercept = mean.hydro, color = 'blue', linetype = 'dashed') +
  ggtitle('Renewable Energy Production and Consumption') +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 12),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size = 10))
```



There appears to be an overall rising trend in energy production from biomass and renewable energy sources in the last 50 years. It is interesting to note that in the period of approximately 10 years in the period from about 1990 to about 2000, energy production from these sources appeared to have constant fluctuations, creating no overall decreasing or increasing trend for this period (just from looking at the plot). Regarding hydroelectric power consumption, however, in the last 50 years there doesn't appear to have been any overall decreasing or increasing trend.

Question 5

Compute the correlation between these three series. Are they significantly correlated? Explain your answer.

```
cor(ts.energy)
```

```
##                                  Total Biomass Energy Production
## Total Biomass Energy Production                        1.0000000
## Total Renewable Energy Production                      0.9678137
## Hydroelectric Power Consumption                       -0.1142927
##                                  Total Renewable Energy Production
## Total Biomass Energy Production                          0.96781371
## Total Renewable Energy Production                        1.00000000
## Hydroelectric Power Consumption                         -0.02916103
##                                  Hydroelectric Power Consumption
## Total Biomass Energy Production                      -0.11429266
## Total Renewable Energy Production                    -0.02916103
## Hydroelectric Power Consumption                       1.00000000
```

```
#Correlation tests between each pair of series
cor.test(ts.energy[, 1], ts.energy[, 2])  #Biomass vs Renewable
```

```
##
##  Pearson's product-moment correlation
##
## data:  ts.energy[, 1] and ts.energy[, 2]
## t = 95.677, df = 619, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9624198 0.9724443
## sample estimates:
##       cor
## 0.9678137
```

```
cor.test(ts.energy[, 1], ts.energy[, 3])  #Biomass vs Hydroelectric
```

```
##
##  Pearson's product-moment correlation
##
## data:  ts.energy[, 1] and ts.energy[, 3]
## t = -2.8623, df = 619, p-value = 0.004348
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.19125123 -0.03593747
## sample estimates:
##        cor
## -0.1142927
```

```
cor.test(ts.energy[, 2], ts.energy[, 3])  #Renewable vs Hydroelectric
```

```
##
```

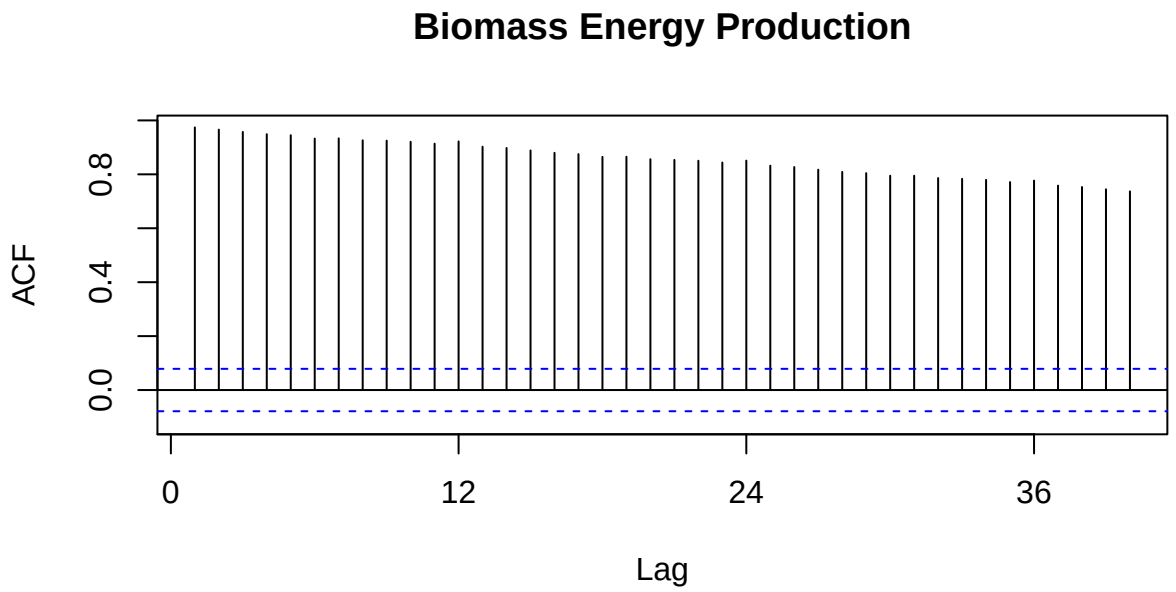
```
##  Pearson's product-moment correlation
##
## data:  ts.energy[, 2] and ts.energy[, 3]
## t = -0.72583, df = 619, p-value = 0.4682
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1075925  0.0496312
## sample estimates:
##         cor
## -0.02916103
```

Because the p-value is less than 0.05 for each of the pair of series, we could say that these three series are significantly correlated with each other.

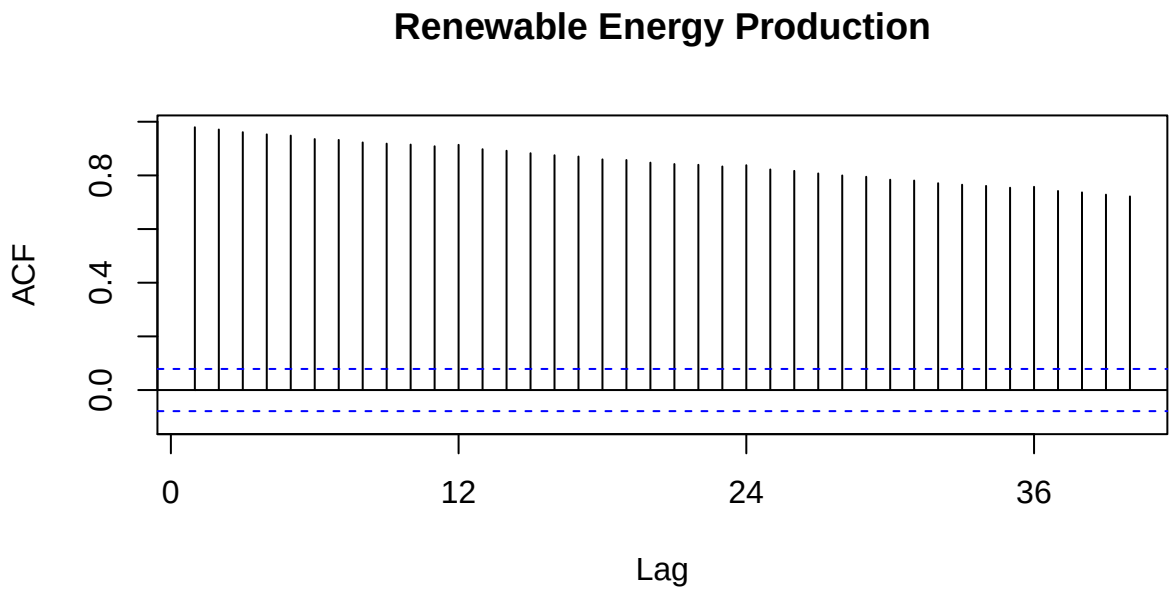
Question 6

Compute the autocorrelation function from lag 1 up to lag 40 for these three variables. What can you say about these plots? Do the three of them have the same behavior?

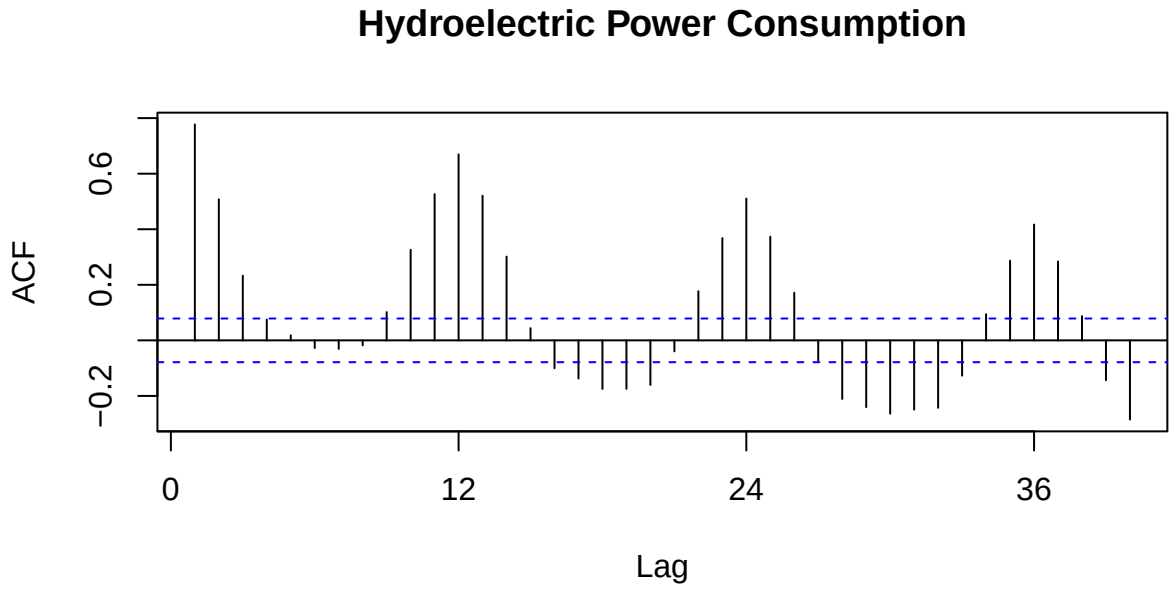
```
biomass.acf = Acf(ts.energy[,1], lag.max = 40, main = 'Biomass Energy Production',
                  type = 'correlation', plot = TRUE)
```




```
renewable.acf = Acf(ts.energy[,2], lag.max = 40, main = 'Renewable Energy Production',
                    plot = TRUE)
```




```
hydro.acf = Acf(ts.energy[,3], lag.max = 40, main = 'Hydroelectric Power Consumption',
                plot = TRUE)
```

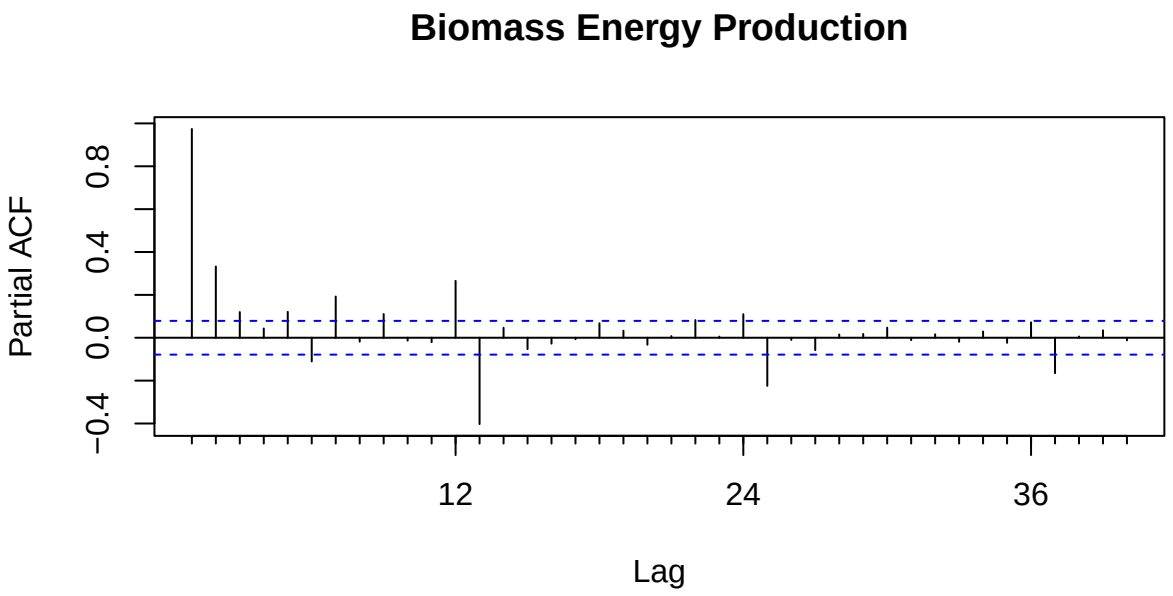


The biomass and renewable plots appear to have the same behavior. A gradual decline in these plots suggests strong correlation over short periods and a decreasing correlation between the data point and its past values as the lag increases. Despite the gradual decline, there still appears to be a strong correlation and significant ACF value at lag 40 for both plots. On the other hand, the hydroelectric plot shows periodic increases and decreases, with peaks approximately every 6 months as the lag gets larger. This sinusoidal patterns suggests seasonal variation.

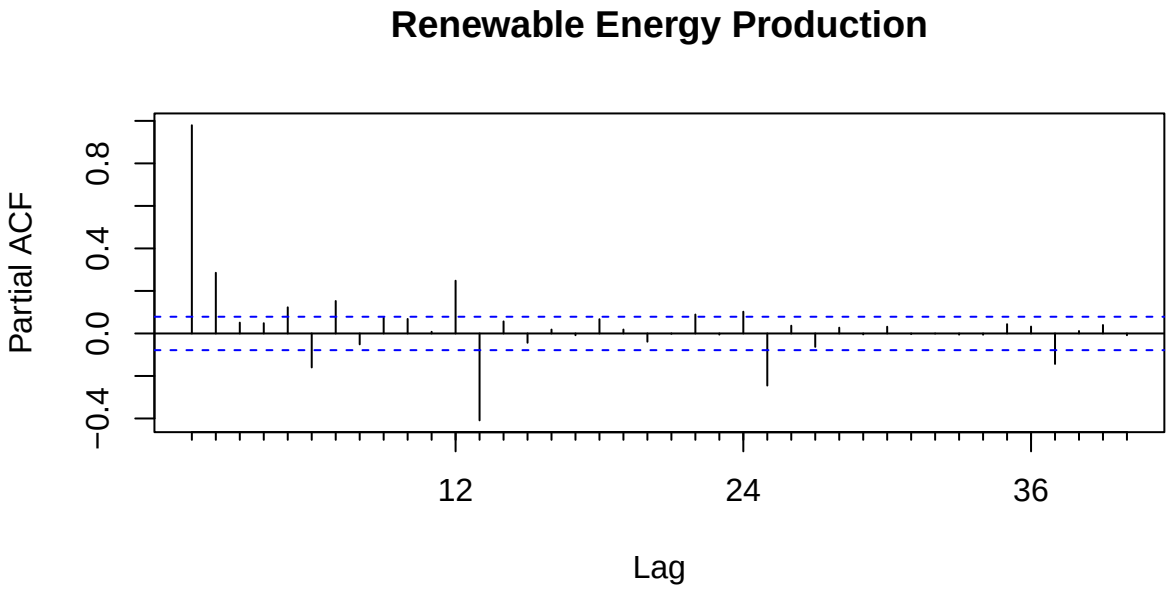
Question 7

Compute the partial autocorrelation function from lag 1 to lag 40 for these three variables. How these plots differ from the ones in Q6?

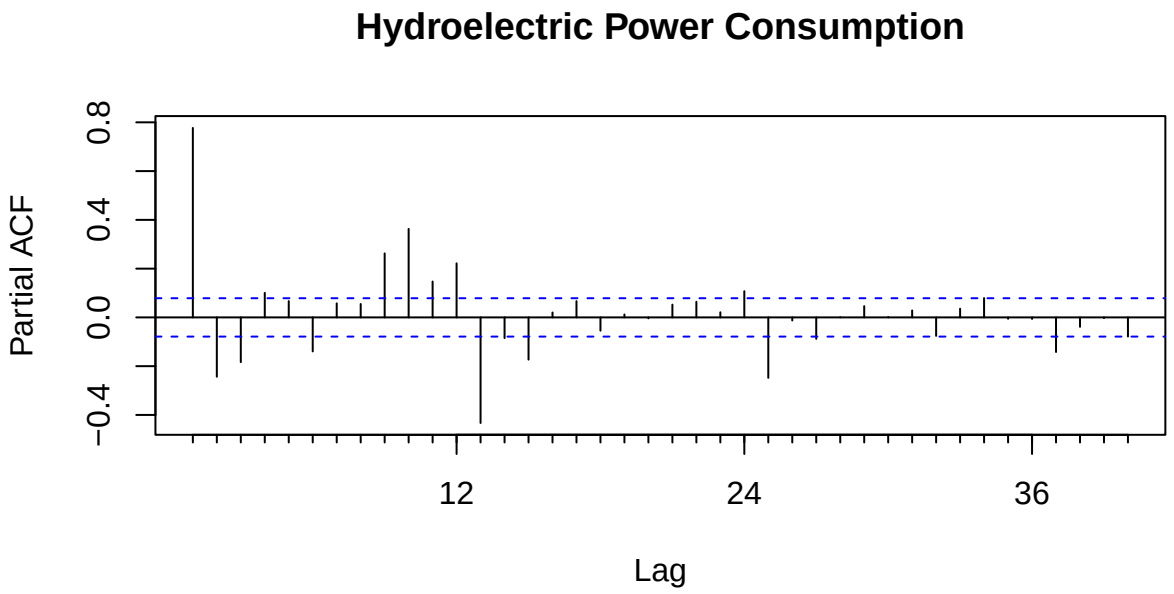
```
biomass.pacf = Pacf(ts.energy[,1], main = 'Biomass Energy Production',
                    lag.max = 40, plot = TRUE)
```




```
renewable.pacf = Pacf(ts.energy[,2], main = 'Renewable Energy Production',
                      lag.max = 40, plot = TRUE)
```



```
hydro.pacf = Pacf(ts.energy[,3], main = 'Hydroelectric Power Consumption',
                  lag.max = 40, plot = TRUE)
```



The PACF plots for all 3 series show significant PACF values at periodic intervals, approximately every 12 months. This seems to suggest seasonality in each of the 3 data series. However, this periodic significance exhibits a gradual decrease every 12 months, with the PACF value at lag 37 (so at the onset of the third 12-month period) barely reaching significance since the PACF value is just outside the blue dotted line.