ENV 790.30 - Time Series Analysis for Energy Data | Spring 2024

Assignment 2 - Due date 02/25/24

Samantha Pace

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.\

```r
#Load/install required package here

#install.packages("forecast")
#install.packages("tseries")
#install.packages("dplyr")
#install.packages(ggplot2)


library(forecast)
library(dplyr)
library(tseries)
library(ggplot2)

# working directory
getwd()
```

```
## [1] "C:/Users/saman/OneDrive/Desktop/Duke Spring 24/GITHUB/TSA_Sp24"
```

Data set information

Consider the data provided in the spreadsheet “Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sourc 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*.

```
#Importing data set

library(readxl)

#Importing data using read.xlsx
energy_data <- read_excel(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source
                          skip = 12,
                          sheet = "Monthly Data", col_names = FALSE)
```

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

```
# Getting column names from row 11
read_col_names <-read_excel(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sourc
```

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

```
energy_data <- read_excel(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source
## New names:
## * ‘‘ -> ‘...1‘
## * ‘‘ -> ‘...2‘
## * ‘‘ -> ‘...3‘
## * ‘‘ -> ‘...4‘
## * ‘‘ -> ‘...5‘
## * ‘‘ -> ‘...6‘
## * ‘‘ -> ‘...7‘
## * ‘‘ -> ‘...8‘
## * ‘‘ -> ‘...9‘
## * ‘‘ -> ‘...10‘
## * ‘‘ -> ‘...11‘
## * ‘‘ -> ‘...12‘
## * ‘‘ -> ‘...13‘
## * ‘‘ -> ‘...14‘
```

```r
# inputting correct column names
colnames(energy_data) <- read_col_names
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.

```r
# selecting the 3 columns needed
energy_data_3 <- energy_data[,4:6]

# checking with head() function
head(energy_data_3)
```

```
## # 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().

```
# transforming data frame in a time series object
ts_energy_data_3 <- ts(energy_data_3, start = c(1973,1), frequency = 12)
```

Question 3

Compute mean and standard deviation for these three series.

```
# mean for ts of Total Biomass Energy Production column
mean(ts_energy_data_3[,1])
```

```
## [1] 279.8046
```

```
# standard deviation for ts of Total Biomass Energy Production column
sd(ts_energy_data_3[,1])
```

```
## [1] 92.66504
```

```
# mean for ts of Total Renewable Energy Production column
mean(ts_energy_data_3[,2])
```

```
## [1] 395.7213
```

```
# standard deviation for ts of Total Renewable Energy Production column
sd(ts_energy_data_3[,2])
```

```
## [1] 137.7952
```

```
# mean for ts of Hydroelectric Power Consumption column
mean(ts_energy_data_3[,3])
```

```
## [1] 79.73071
```

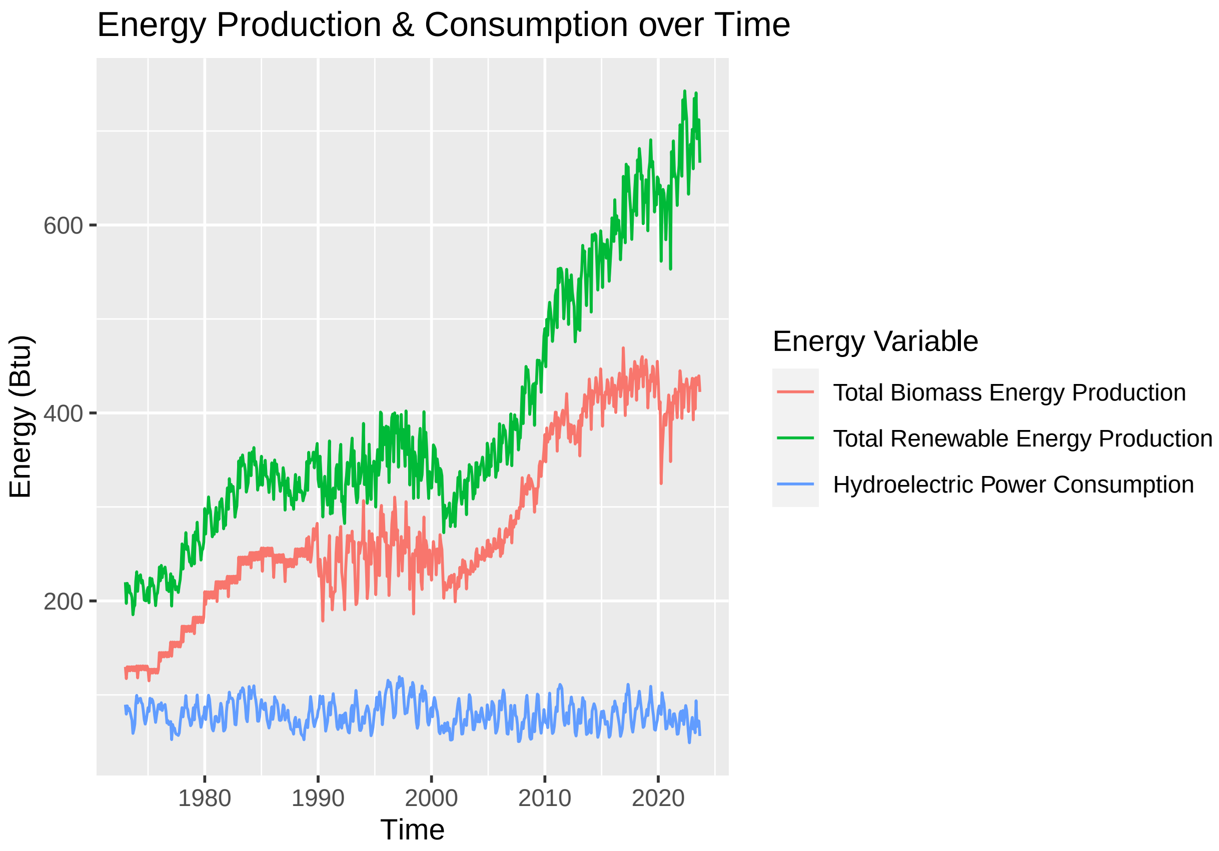
```
# standard deviation for ts of Hydroelectric Power Consumption column
sd(ts_energy_data_3[,3])
```

```
## [1] 14.14734
```

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.

```
# using autoplot, can't add the mean lines
autoplot(ts_energy_data_3) +
  ylab("Energy (Btu)") +
  xlab("Time") +
  labs(color="Energy Variable") +
  ggtitle("Energy Production & Consumption over Time")
```



```
  #abline(h=mean(ts_energy_data_3[,1]), col = 'red')
```

Question 5

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

```r
#correlation is covariance/standard deviation
#cov(ts_energy_data_3)/sd(ts_energy_data_3)

# this one works with coefficient but no p value
cor(ts_energy_data_3)
```

```
##                                   Total Biomass Energy Production
## Total Biomass Energy Production                        1.00000000
## Total Renewable Energy Production                      0.97074621
## Hydroelectric Power Consumption                       -0.09656318
##                                   Total Renewable Energy Production
## Total Biomass Energy Production                         0.970746212
## Total Renewable Energy Production                       1.000000000
## Hydroelectric Power Consumption                        -0.001768629
##                                   Hydroelectric Power Consumption
## Total Biomass Energy Production                      -0.096563177
## Total Renewable Energy Production                    -0.001768629
## Hydroelectric Power Consumption                       1.000000000
```

```r
# correlation between Biomass and Renewable energy; small p value and 0.97 correlation value
cor.test(energy_data_3$`Total Biomass Energy Production`,
         energy_data_3$`Total Renewable Energy Production`,
         method='pearson')
```

```
##
##  Pearson’s product-moment correlation
##
## data:  energy_data_3$‘Total Biomass Energy Production‘ and energy_data_3$‘Total Renewable Energy Pro
## t = 99.608, df = 607, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9657830 0.9749987
## sample estimates:
##       cor
## 0.9707462
```

```r
# correlation between Biomass and Hydroelectric energy, p value of 0.017 and correlation coefficient of
cor.test(energy_data_3$`Total Biomass Energy Production`,
         energy_data_3$`Hydroelectric Power Consumption`)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  energy_data_3$‘Total Biomass Energy Production‘ and energy_data_3$‘Hydroelectric Power Consum
## t = -2.3902, df = 607, p-value = 0.01714
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1746734 -0.0172452
## sample estimates:
##         cor
## -0.09656318
```

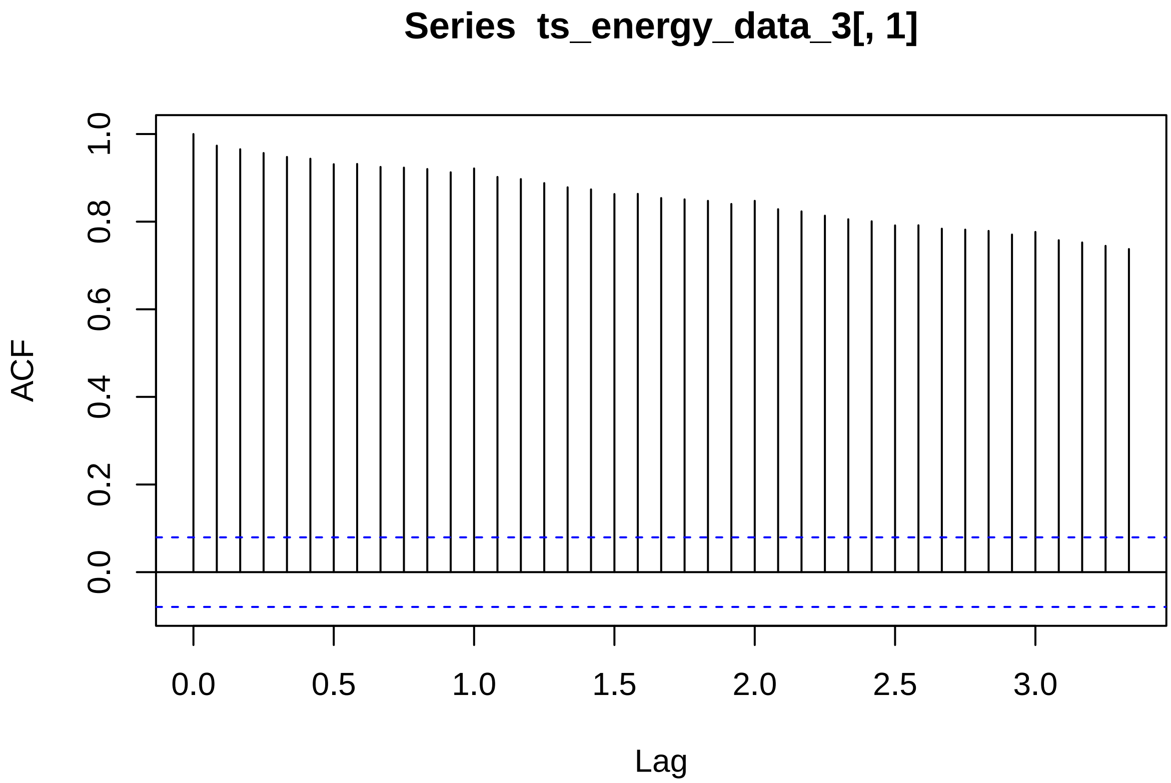
```r
#correlation between Hydroelectric and Renewable, p value of 0.96 and cor of -0.0018
cor.test(energy_data_3$`Total Renewable Energy Production`,
         energy_data_3$`Hydroelectric Power Consumption`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  energy_data_3$‘Total Renewable Energy Production‘ and energy_data_3$‘Hydroelectric Power Consu
## t = -0.043574, df = 607, p-value = 0.9653
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08120750  0.07769257
## sample estimates:
##          cor
## -0.001768629
```

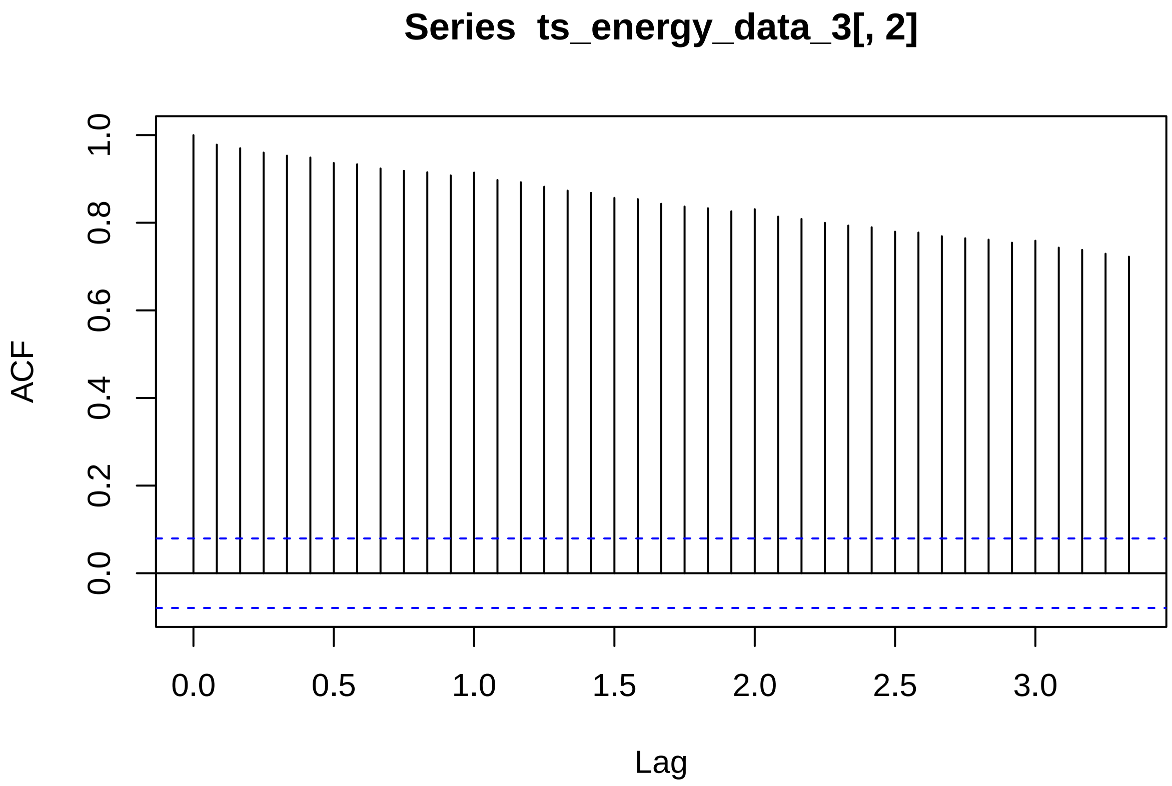
Question 6

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?

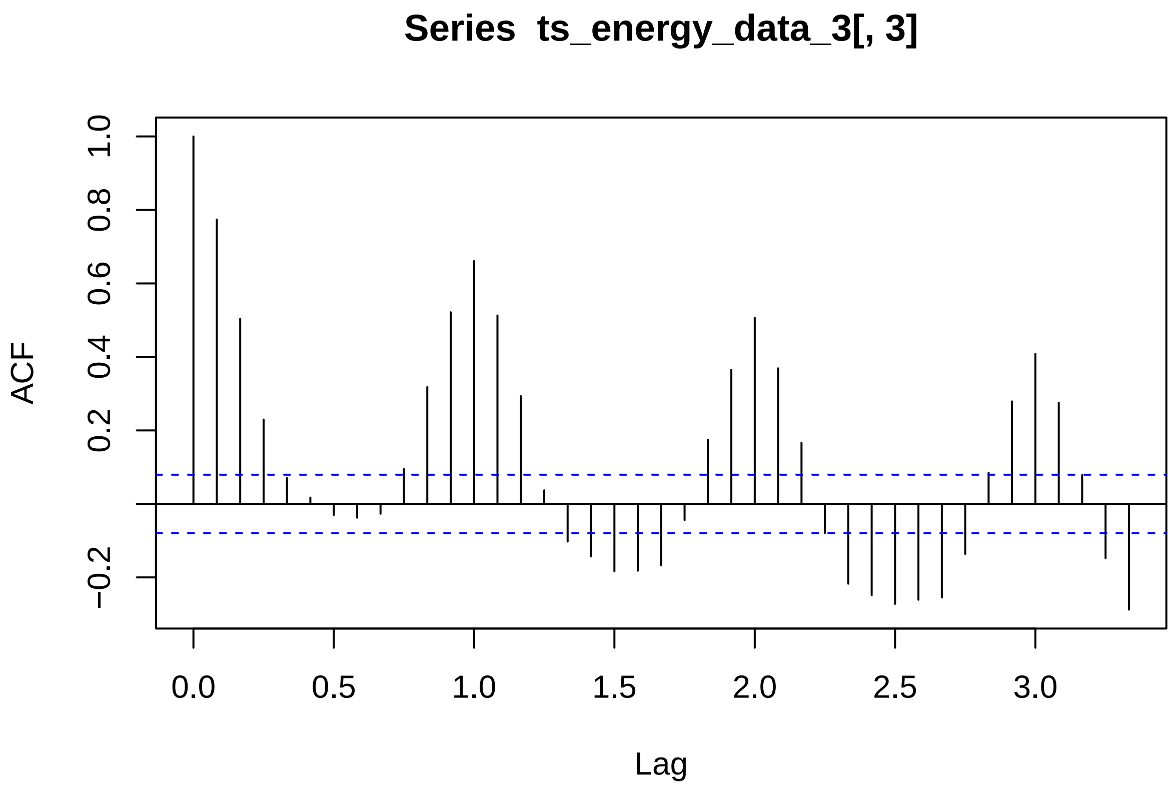
```r
# computing the autocorrelation function from lag 1 to 40
acf(ts_energy_data_3[,1], lag.max = 40, type = "correlation", plot = TRUE)
```



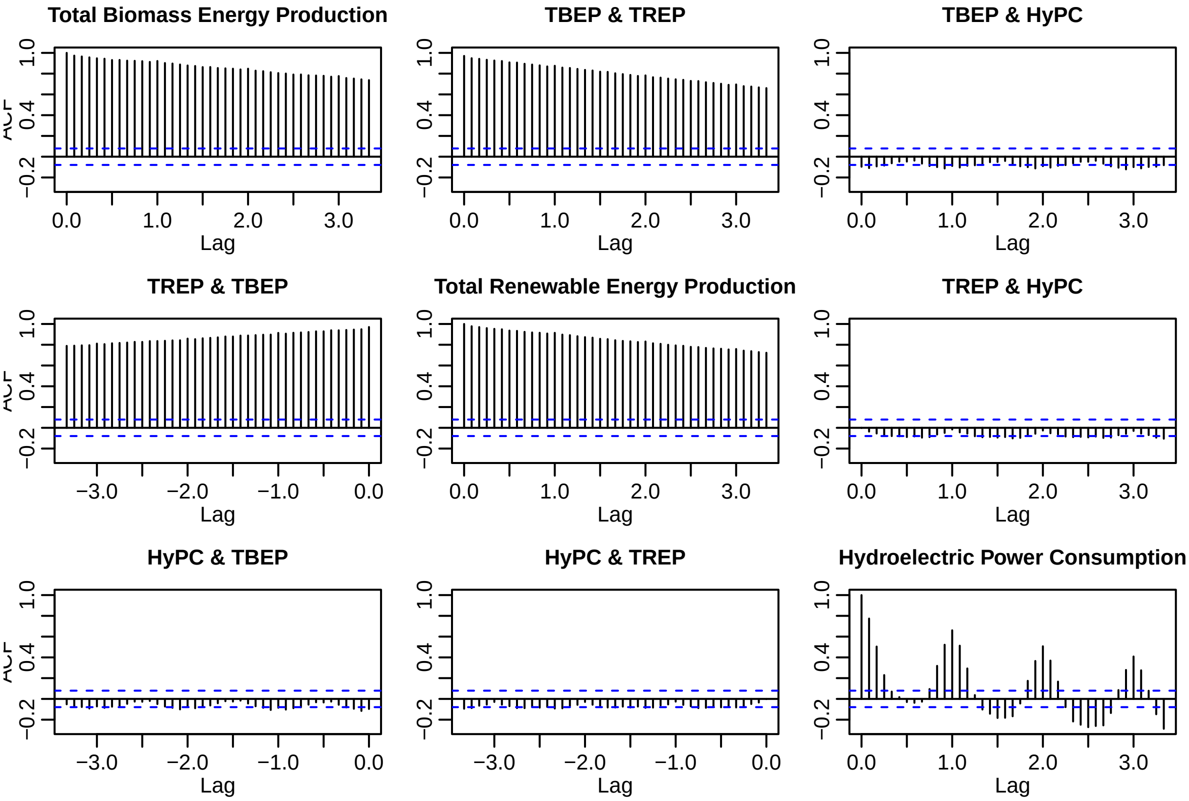
```
acf(ts_energy_data_3[,2], lag.max = 40, type = "correlation", plot = TRUE)
```



```
acf(ts_energy_data_3[,3], lag.max = 40, type = "correlation", plot = TRUE)
```



```
# all three together
acf(ts_energy_data_3, lag.max = 40, type = "correlation", plot = TRUE)
```

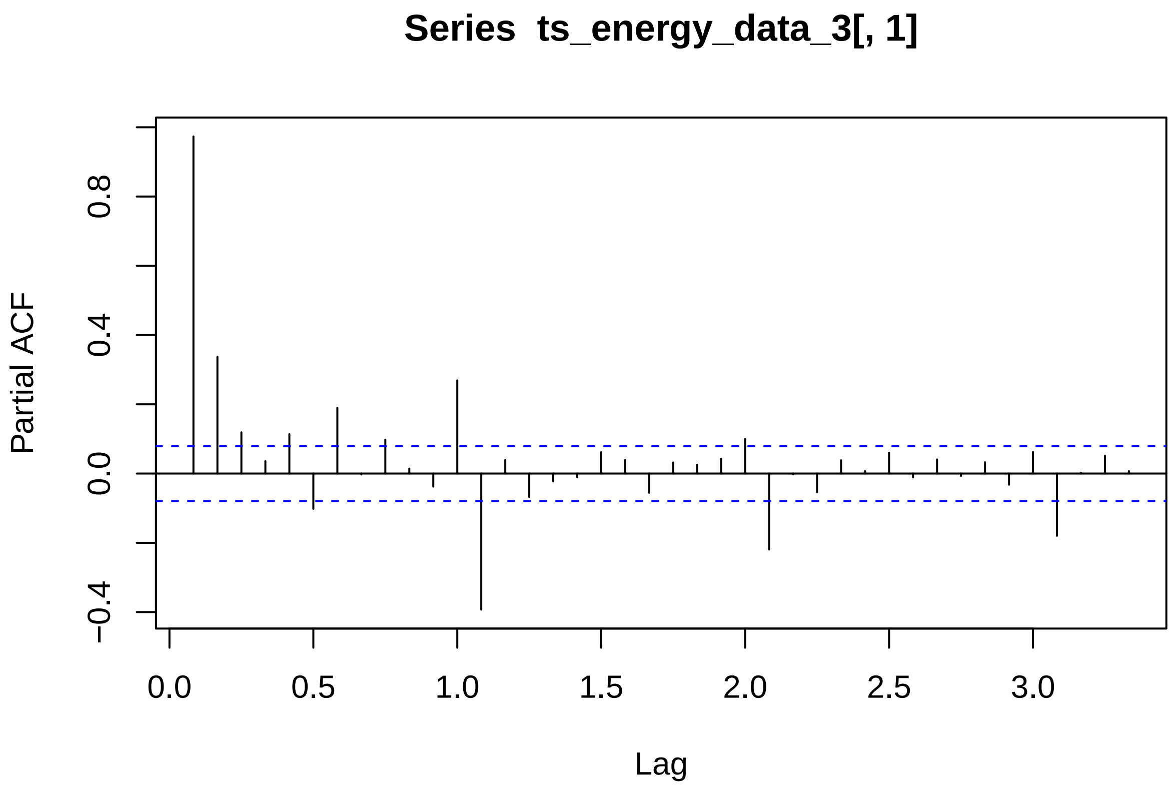


Answer: There is strong correlation in Total Biomass Energy Production and Total Renewable Energy Production; they behave similarly. The Hydroelectric Power Comsumption behaves differently from the other two and has a wave-like cyclical pattern behavior.

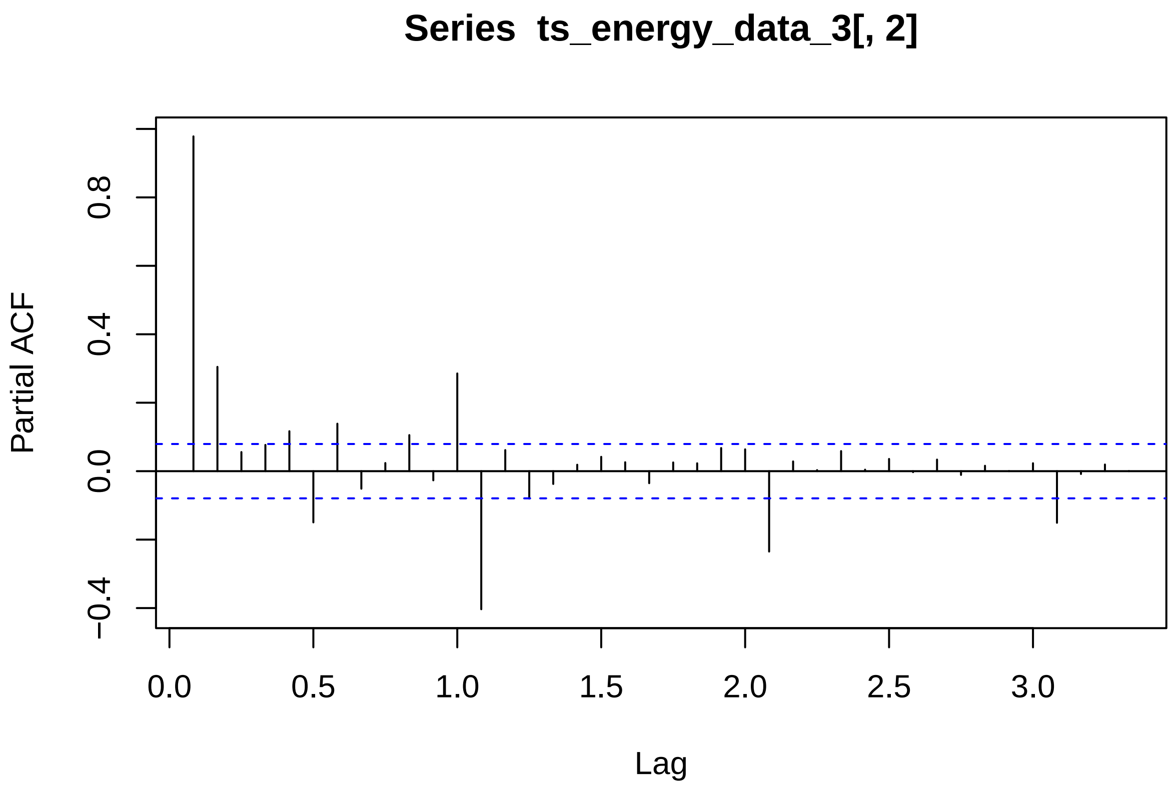
Question 7

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

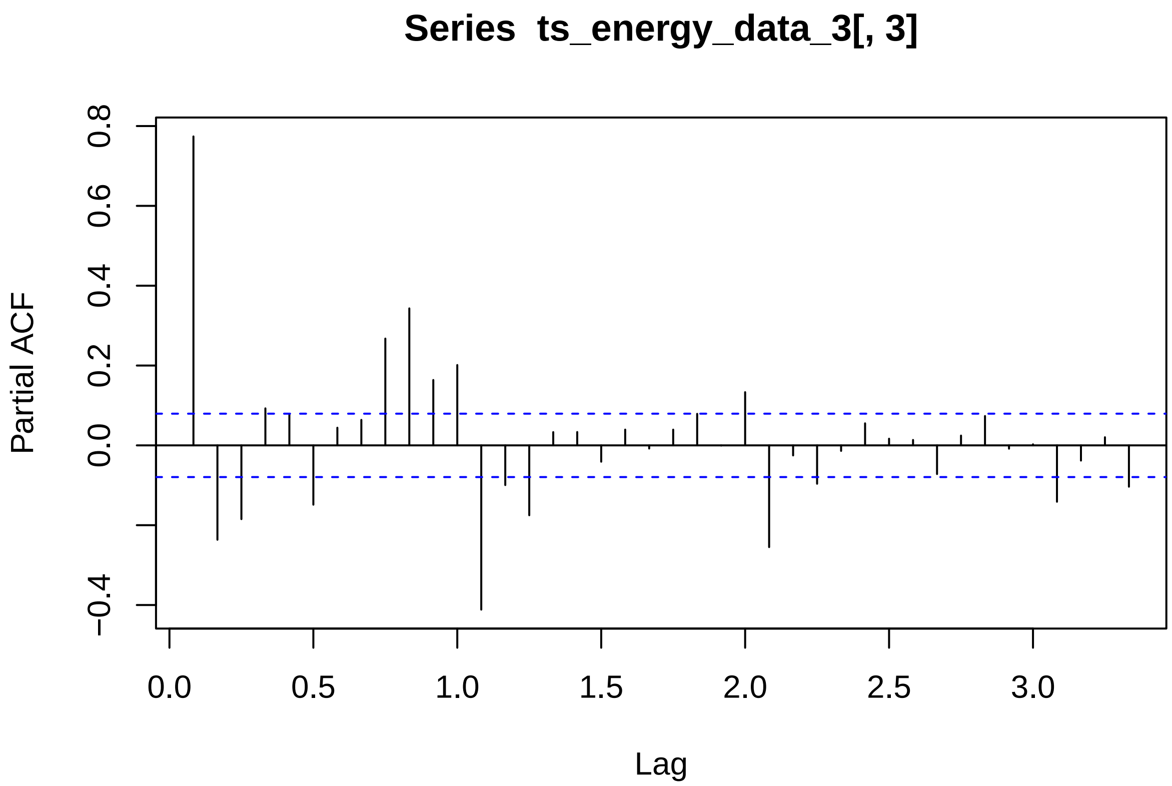
```r
# computing the pacf from lag 1 to 40
pacf(ts_energy_data_3[,1], lag.max = 40, plot = TRUE)
```



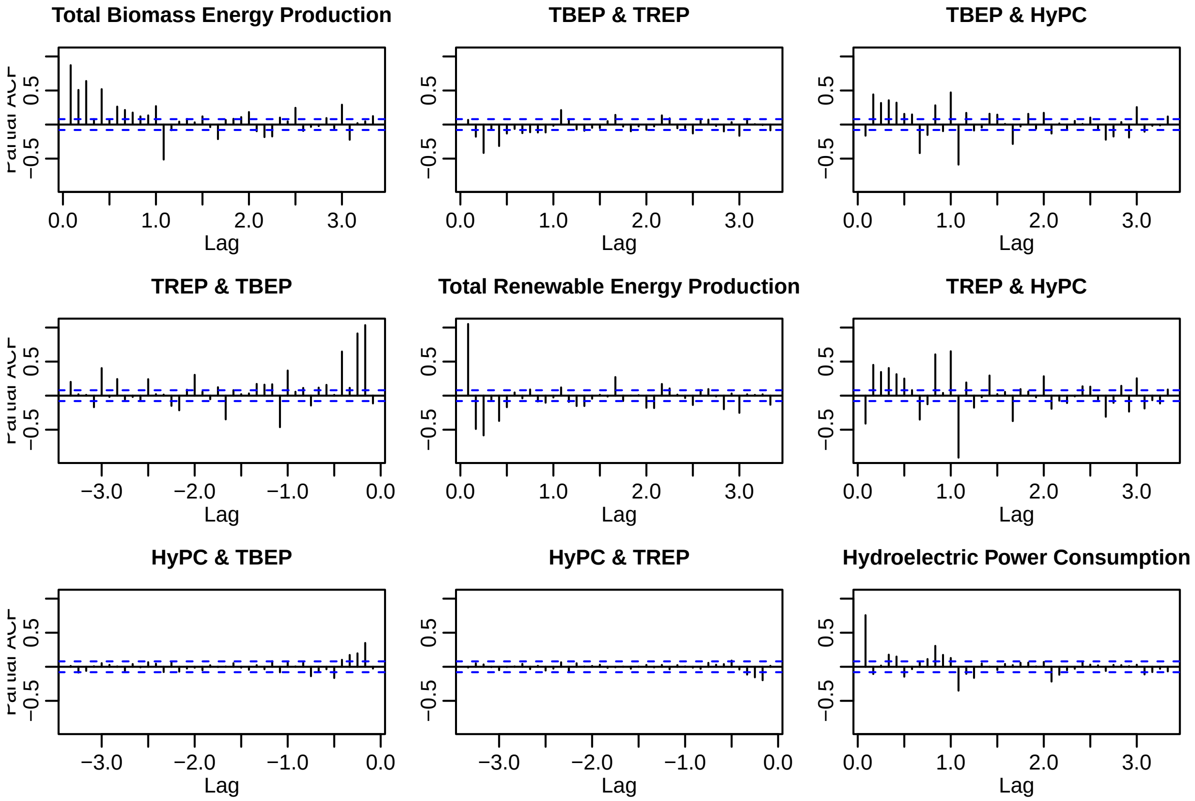
```
pacf(ts_energy_data_3[,2], lag.max = 40, plot = TRUE)
```



```
pacf(ts_energy_data_3[,3], lag.max = 40, plot = TRUE)
```



```
# all three together
pacf(ts_energy_data_3, lag.max = 40, plot = TRUE)
```


Total Biomass Energy Production
TBEP & TREP
TBEP & HyPC
TREP & TBEP
Total Renewable Energy Production
TREP & HyPC
HyPC & TBEP
HyPC & TREP
Hydroelectric Power Consumption
Partial ACF
−0.5
0.5
0.0
1.0
2.0
3.0
−3.0
−2.0
−1.0
Lag