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

Assignment 2 - Due date 02/25/24

Cynthia Zhou

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

Data set information

Consider the data provided in the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.x 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)
raw_data <- read_excel(path="../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.x

read_col_names <- read_excel(path="../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_So

colnames(raw_data) <- read_col_names
head(raw_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.

```
new_data<-raw_data[,4:6]
new_data <- cbind(raw_data[,1],new_data)
head(new_data)
```

```
##        Month Total Biomass Energy Production Total Renewable Energy Production
## 1 1973-01-01                         129.787                             219.839
## 2 1973-02-01                         117.338                             197.330
## 3 1973-03-01                         129.938                             218.686
## 4 1973-04-01                         125.636                             209.330
## 5 1973-05-01                         129.834                             215.982
## 6 1973-06-01                         125.611                             208.249
##   Hydroelectric Power Consumption
## 1                          89.562
## 2                          79.544
## 3                          88.284
## 4                          83.152
## 5                          85.643
## 6                          82.060
```

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_data <- ts(new_data[,2:4],start=c(1973,1),frequency=12)
head(ts_data)
```

```
##          Total Biomass Energy Production Total Renewable Energy Production
## Jan 1973                         129.787                             219.839
## Feb 1973                         117.338                             197.330
## Mar 1973                         129.938                             218.686
## Apr 1973                         125.636                             209.330
## May 1973                         129.834                             215.982
```

```
## Jun 1973                          125.611                                208.249
##          Hydroelectric Power Consumption
## Jan 1973                           89.562
## Feb 1973                           79.544
## Mar 1973                           88.284
## Apr 1973                           83.152
## May 1973                           85.643
## Jun 1973                           82.060
```

Question 3

Compute mean and standard deviation for these three series.

```
mean(ts_data[,"Total Biomass Energy Production"])
```

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

```
mean(ts_data[,"Total Renewable Energy Production"])
```

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

```
mean(ts_data[,"Hydroelectric Power Consumption"])
```

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

```
sd(ts_data[,"Total Biomass Energy Production"])
```

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

```
sd(ts_data[,"Total Renewable Energy Production"])
```

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

```
sd(ts_data[,"Hydroelectric Power Consumption"])
```

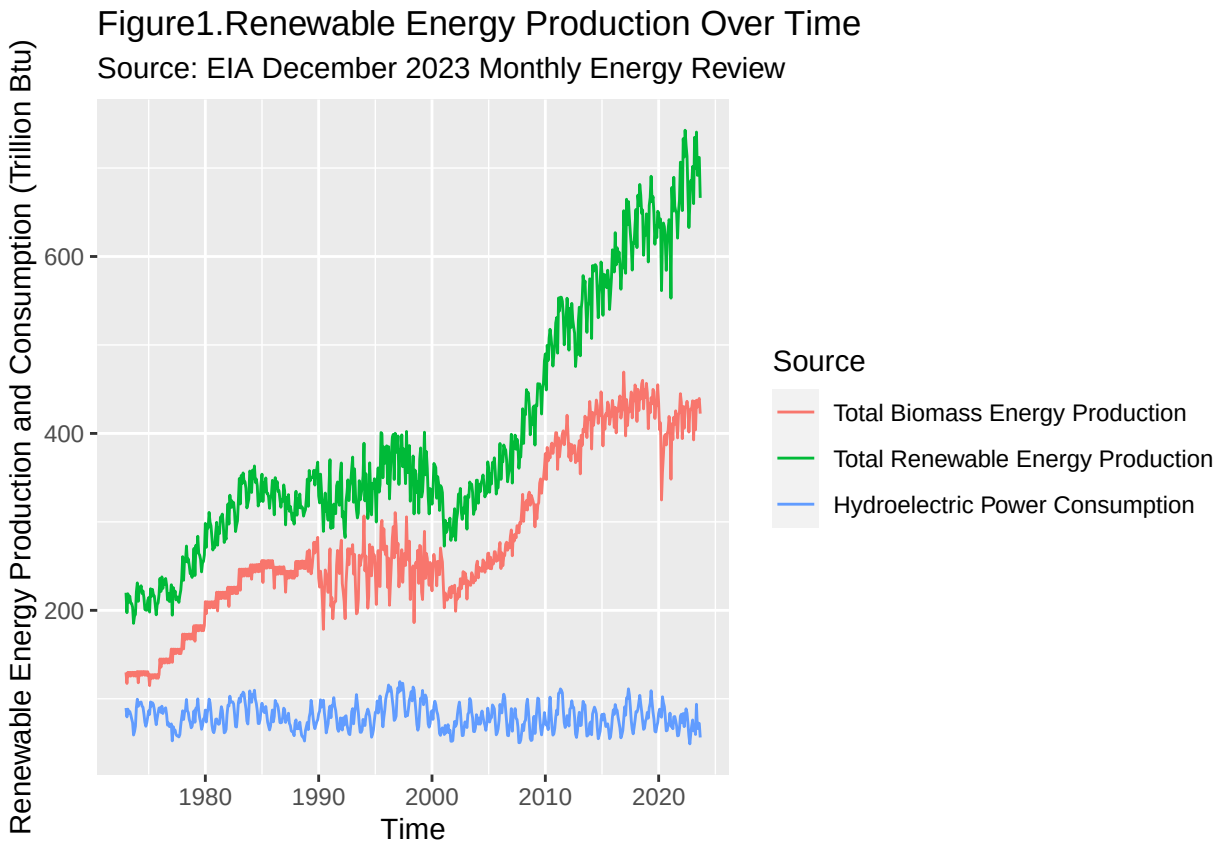
```
## [1] 14.14734
```

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.

In Figure 1, we can find that both total biomass and renewable energy production have gone up over time, and hydroelectric power consumption kept a flat trend. To be more specific, total renewable energy production lays on the dominant position among these three variables, followed by total biomass energy production. Hydroelectric power consumption is the lowest.

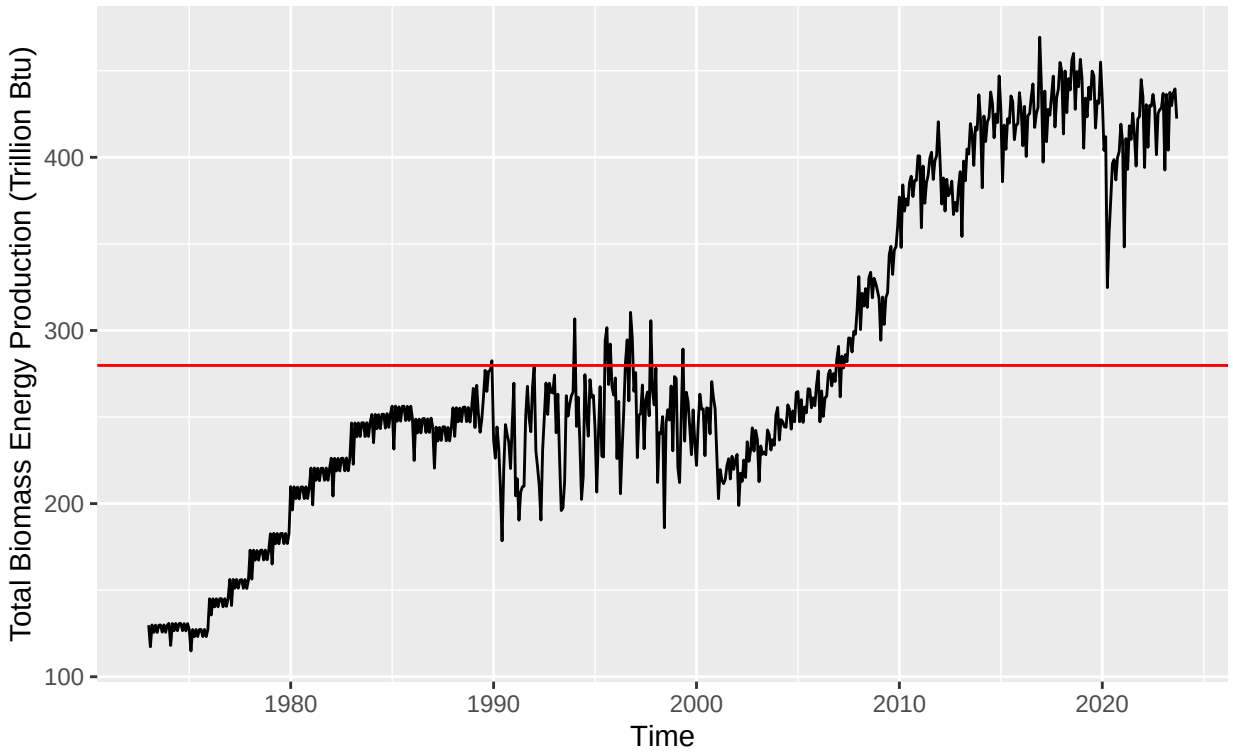
```
autoplot(ts_data) +
  xlab("Time") +
  ylab("Renewable Energy Production and Consumption (Trillion Btu)") +
  labs(color="Source", title="Figure1.Renewable Energy Production Over Time",subtitle = "Source: EIA De
```




```
autoplot(ts_data[,1]) +
  xlab("Time") +
  ylab("Total Biomass Energy Production (Trillion Btu)") +
  labs(color="Source", title="Figure2.Total Biomass Energy Production Over Time",subtitle = "Source: EI
  geom_hline(aes(yintercept = mean(ts_data[,1])), color="red")
```

Figure2.Total Biomass Energy Production Over Time

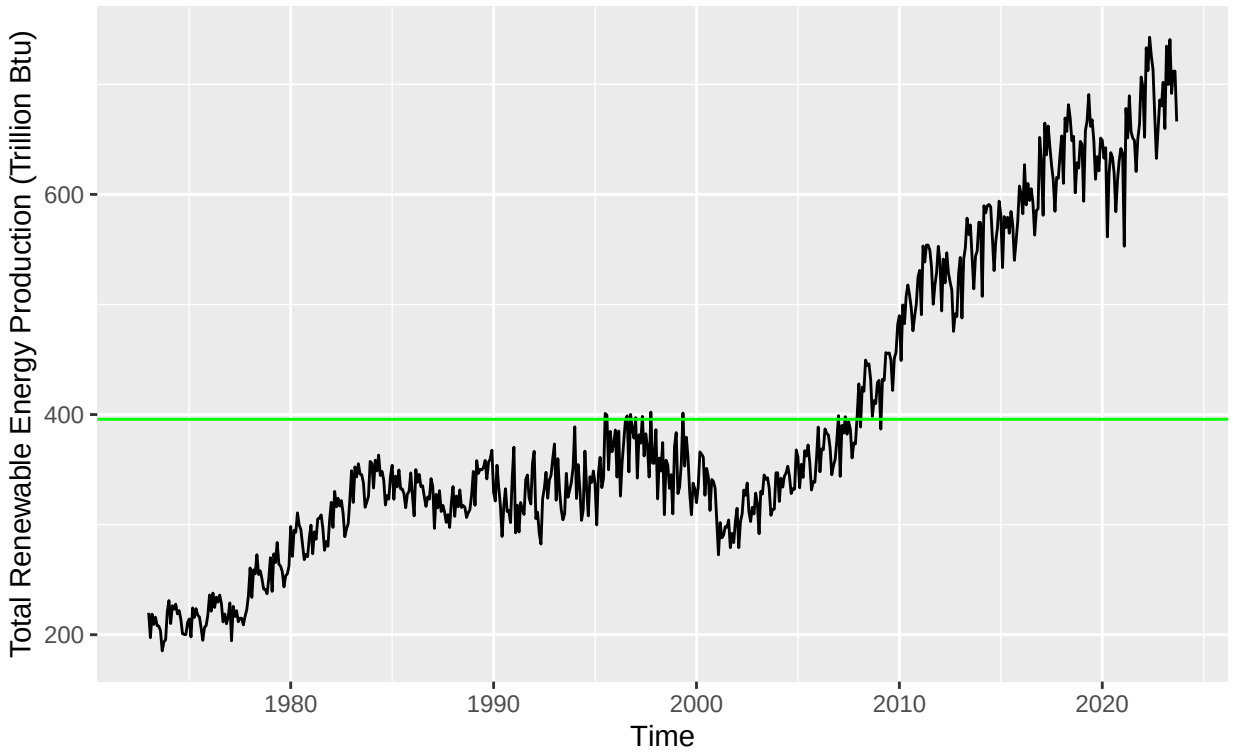
Source: EIA December 2023 Monthly Energy Review




```
autoplot(ts_data[,2]) +
  xlab("Time") +
  ylab("Total Renewable Energy Production (Trillion Btu)") +
  labs(color="Source", title="Figure3.Total Renewable Energy Production Over Time",subtitle = "Source: I
  geom_hline(aes(yintercept = mean(ts_data[,2])), color="green")
```

Figure3.Total Renewable Energy Production Over Time

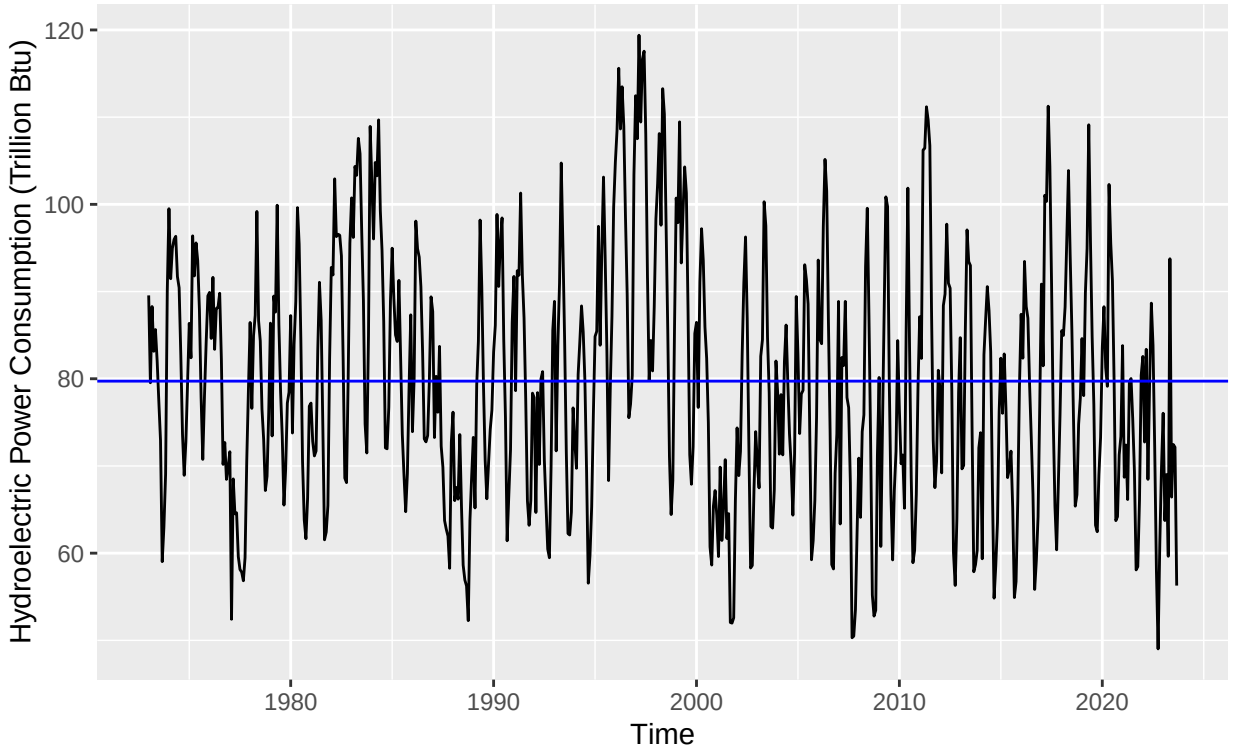
Source: EIA December 2023 Monthly Energy Review




```
autoplot(ts_data[,3]) +
  xlab("Time") +
  ylab("Hydroelectric Power Consumption (Trillion Btu)") +
  labs(color="Source", title="Figure4.Hydroelectric Power Consumption Over Time",subtitle = "Source: EI
  geom_hline(aes(yintercept = mean(ts_data[,3])), color="blue")
```

Figure4.Hydroelectric Power Consumption Over Time

Source: EIA December 2023 Monthly Energy Review



Question 5

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

Total Biomass Energy Production is significantly correlated with Total Renewable Energy Production (cor>0.9) , and there is no strong correlation between Hydroelectric Power Consumption and the other two variables as values are very small.

```
cor(ts_data)
```

```
##                                  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
```

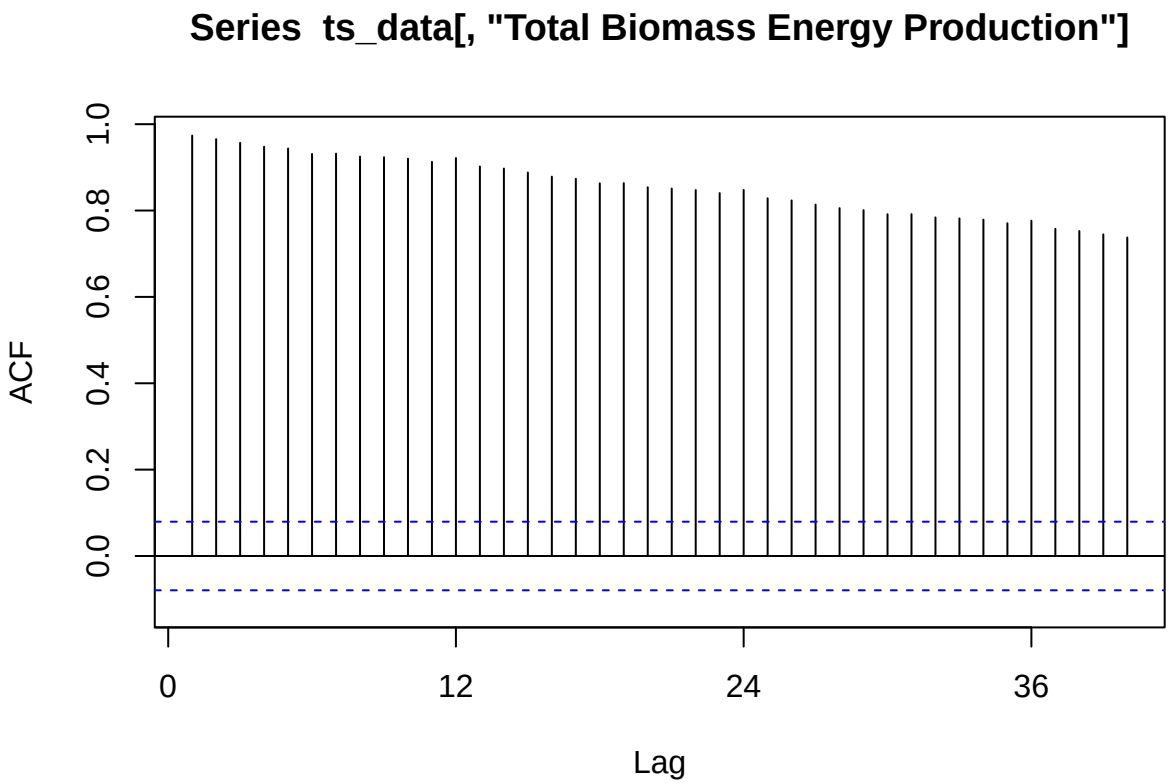
Question 6

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?

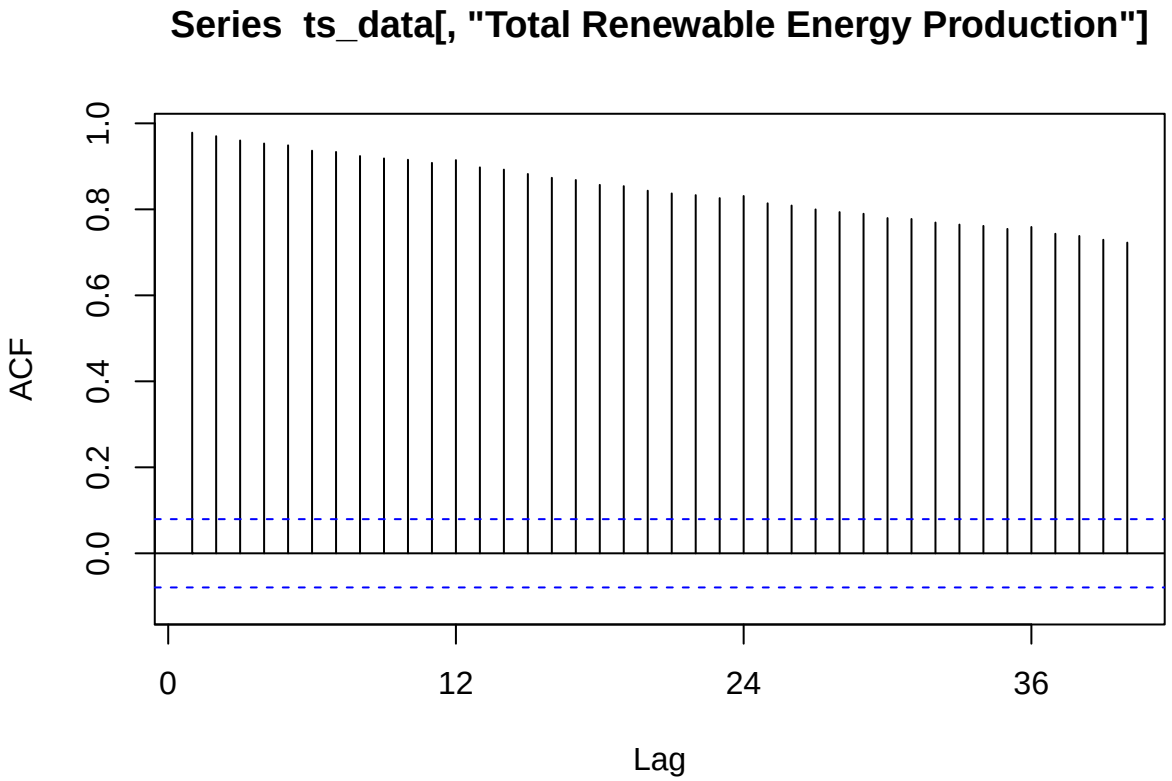
For biomass plot, the autocorrelation decreases very slowly over time, indicating the stability

of this energy production in time series. As for total renewable energy production, there is a similar behavior to the biomass one but it has a little quicker decrease trend that refers to less persistence. Both plots suggest a decreasing dependence over time. In hydroelectric plot, there is a periodic trend, which may indicate seasonality of this kind of energy. These three plots show different behaviors.

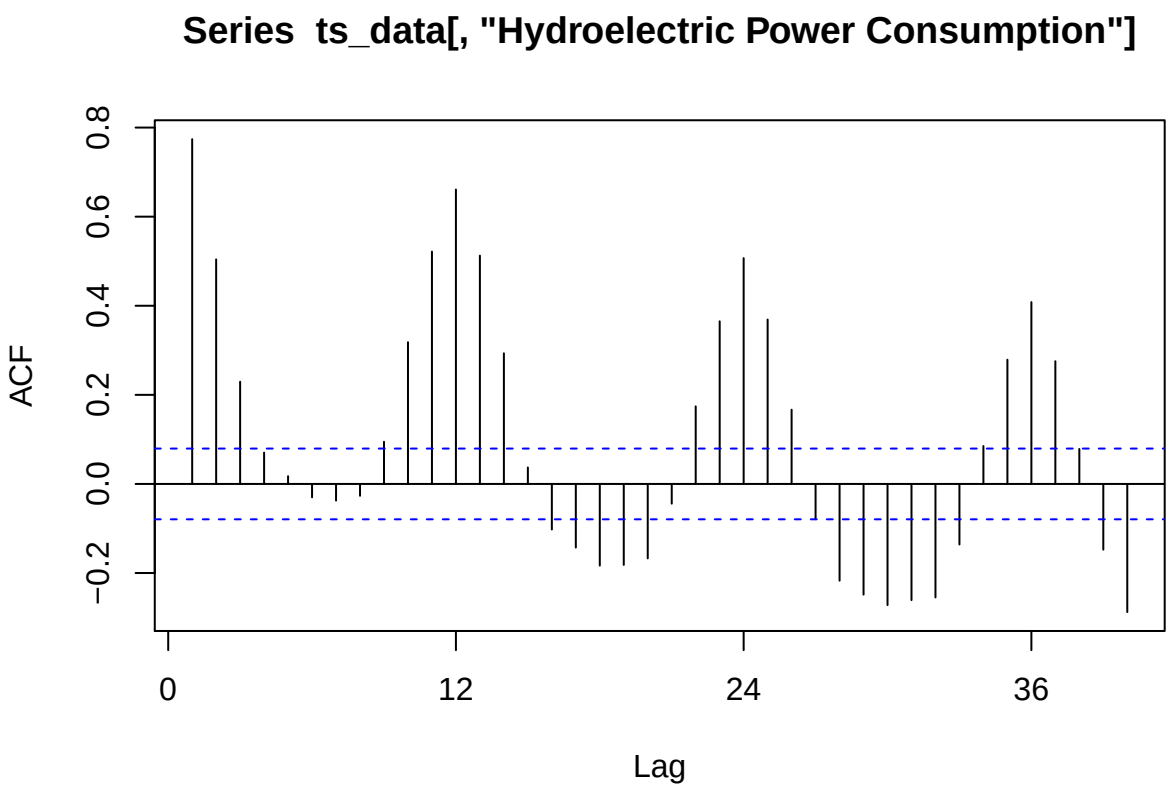
```
biomass_acf=Acf(ts_data[,"Total Biomass Energy Production"],lag.max = 40)
```




```
renew_acf=Acf(ts_data[, "Total Renewable Energy Production"],lag.max = 40)
```




```
hy_acf=Acf(ts_data[, "Hydroelectric Power Consumption"],lag.max = 40)
```

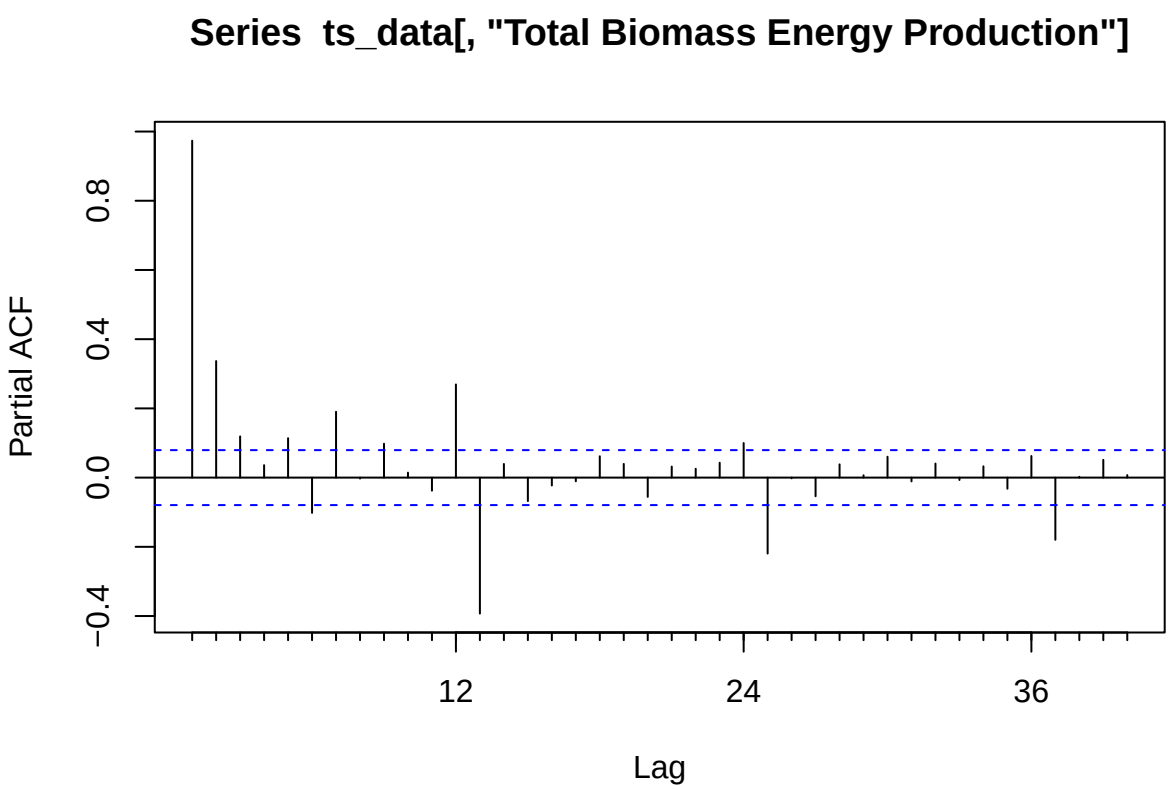


Question 7

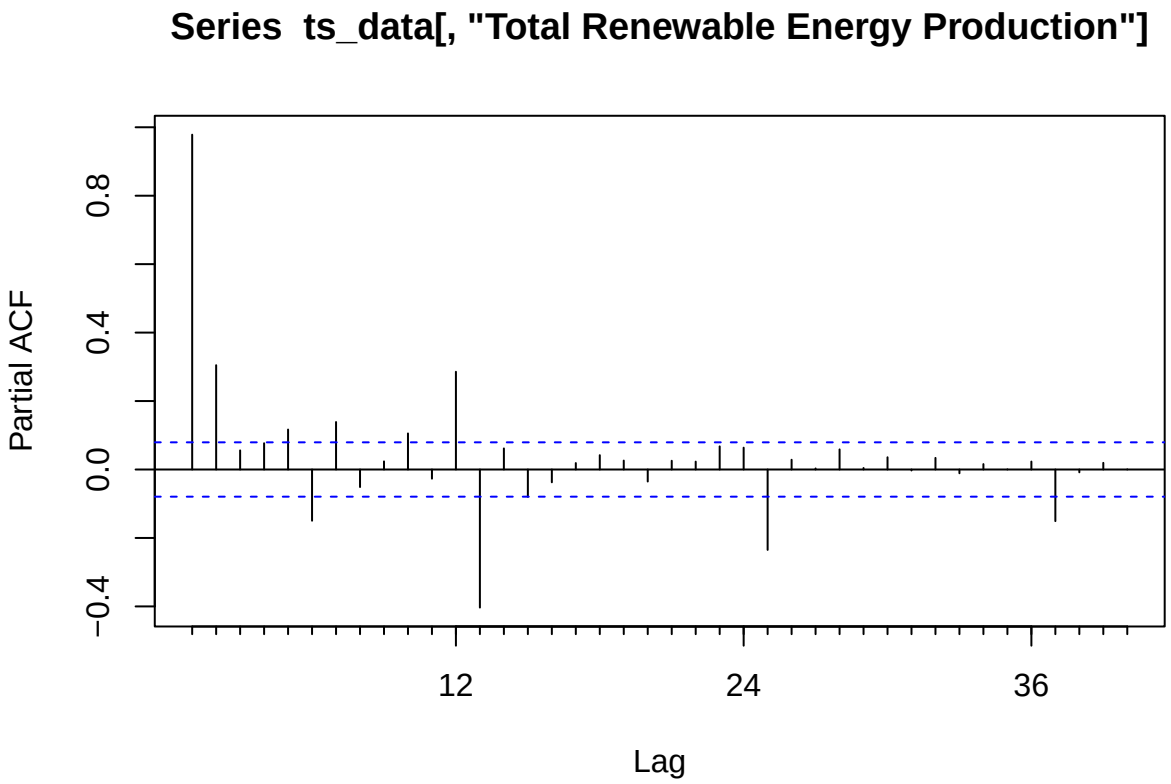
Compute the partial autocorrelation function from lag 1 to lag 40 for these three variables. How these plots differ from the ones in Q6?

PACF removes the influence of all these intermediate variables. The plots below show very random patterns, while the ones in Q6 follow some trends.

```
biomass_pacf=Pacf(ts_data[,"Total Biomass Energy Production"],lag.max = 40)
```




```
renew_pacf=Pacf(ts_data[, "Total Renewable Energy Production"],lag.max = 40)
```




```
hy_pacf=Pacf(ts_data[, "Hydroelectric Power Consumption"],lag.max = 40)
```