

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

## Assignment 2 - Due date 01/27/26

Bella Huang

### 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\_Sp26.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
```

```
library(readxl)
library(openxlsx)
library(ggplot2)
```

## Data set information

Consider the data provided in the spreadsheet “Table\_10.1\_Renewable\_Energy\_Production\_and\_Consumption\_by\_Source.xlsx” on our **Data** folder. The data comes from the US Energy Information and Administration and corresponds to the December 2025 Monthly Energy Review. The spreadsheet is ready to be used. Refer to the file “M2 ImportingData\_XLSX.Rmd” in our Lessons folder for instructions on how to read .xlsx files.

```
#Importing data set
energy_data1 <- read_excel(path="../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx")

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


read_col_names <- read_excel(path="../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx")

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


colnames(energy_data1) <- read_col_names
```

## 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_ts <- energy_data1 %>%
  select(
    `Month`,
    `Total Biomass Energy Production`,
    `Total Renewable Energy Production`,
    `Hydroelectric Power Consumption`)

head(energy_ts)

## # A tibble: 6 x 4
##   Month           `Total Biomass Energy Production` `Total Renewable Energy` `1
##   <dttm>                      <dbl>                  <dbl>
## 1 1973-01-01 00:00:00          130.                 220.
## 2 1973-02-01 00:00:00          117.                 197.
## 3 1973-03-01 00:00:00          130.                 219.
## 4 1973-04-01 00:00:00          126.                 209.
## 5 1973-05-01 00:00:00          130.                 216.
## 6 1973-06-01 00:00:00          126.                 208.
## # i abbreviated name: 1: `Total Renewable Energy Production`
## # i 1 more variable: `Hydroelectric Power Consumption` <dbl>
```

## 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().

```
energy_ts_ts <- ts(
  energy_ts[, c(
    "Total Biomass Energy Production",
    "Total Renewable Energy Production",
    "Hydroelectric Power Consumption"
  )],
  start = c(1973, 1),
  frequency = 12)
```

## Question 3

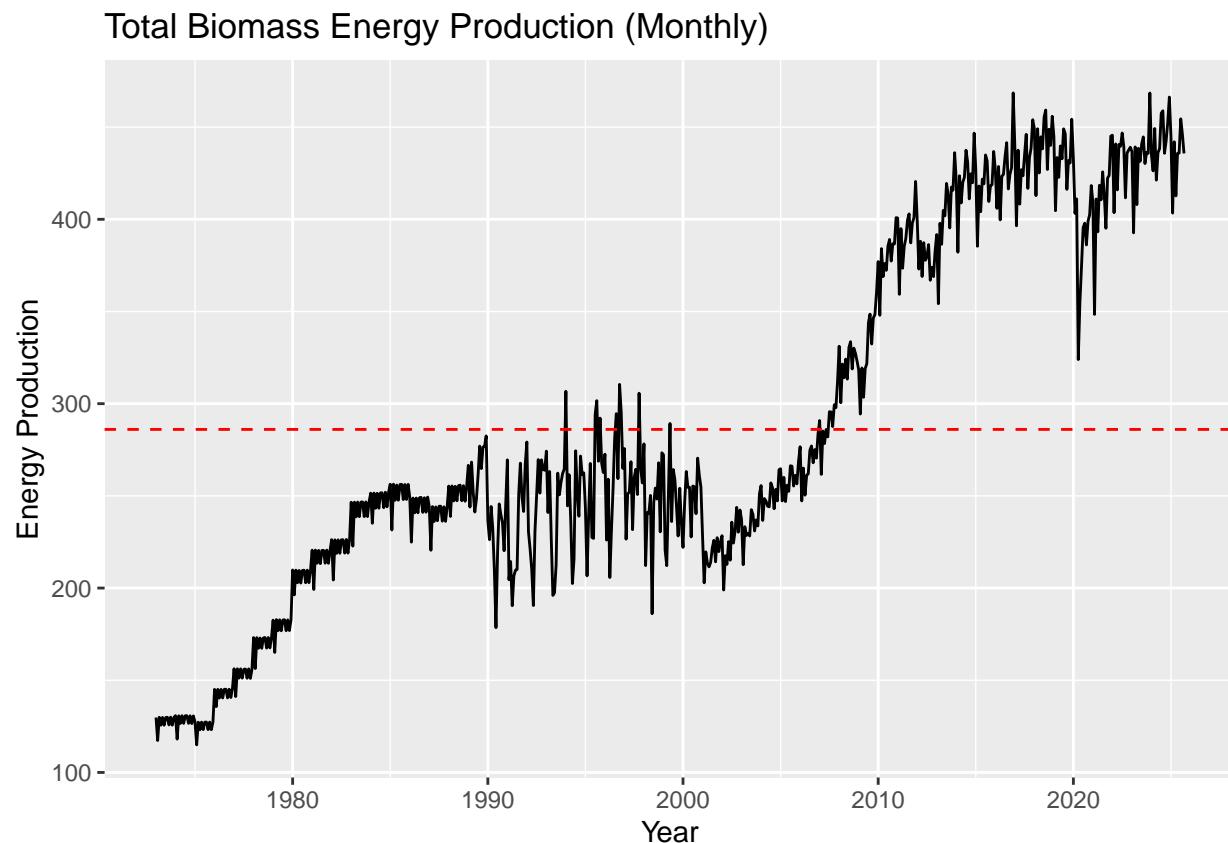
Compute mean and standard deviation for these three series.

```
mean <- apply(energy_ts_ts, 2, mean, na.rm = TRUE)
sd   <- apply(energy_ts_ts, 2, sd, na.rm = TRUE)
```

## 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.

```
# Total Biomass Energy Production
autoplot(energy_ts_ts[, "Total Biomass Energy Production"]) +
  geom_hline(
    yintercept = mean(energy_ts_ts[, "Total Biomass Energy Production"], na.rm = TRUE),
    color = "red",
    linetype = "dashed"
  ) +
  labs(
    title = "Total Biomass Energy Production (Monthly)",
    x = "Year",
    y = "Energy Production")
```

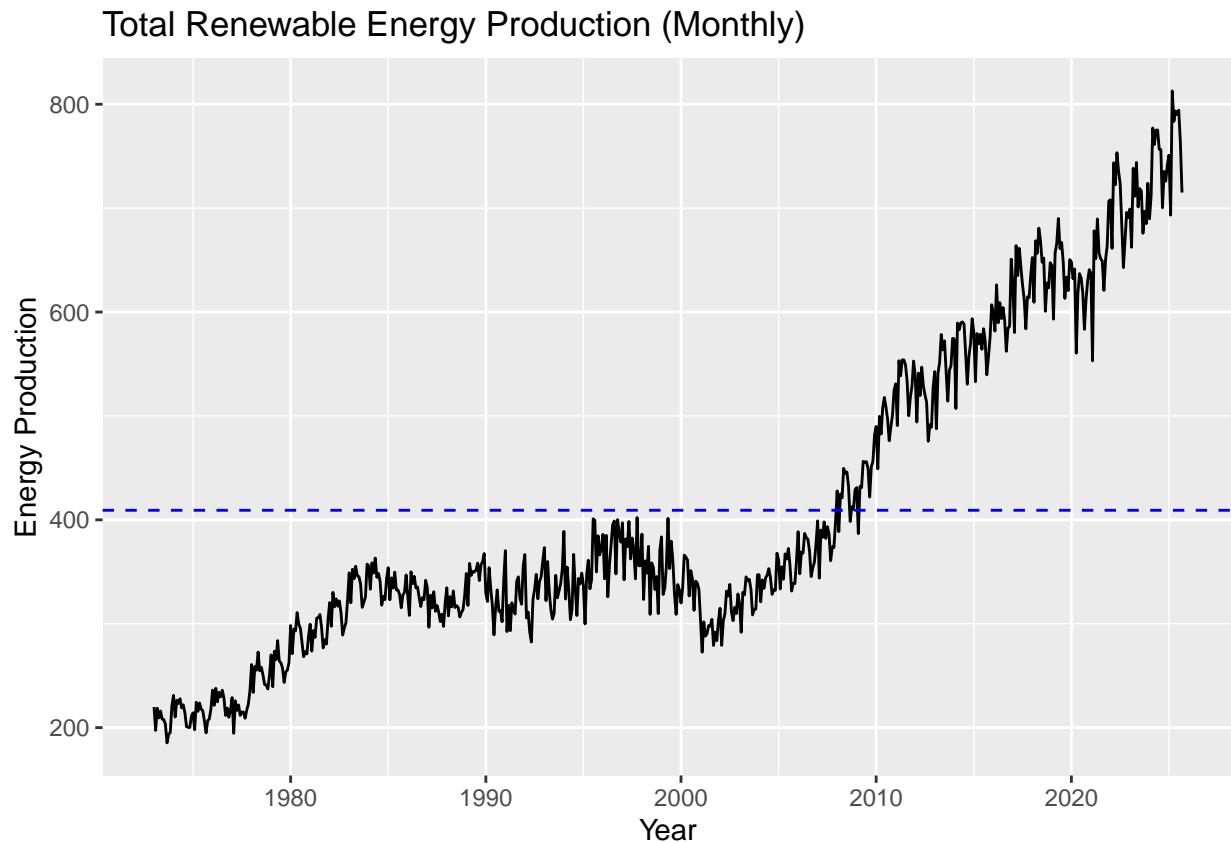


```
# Total Renewable Energy Production
autoplot(energy_ts_ts[, "Total Renewable Energy Production"]) +
  geom_hline(
    yintercept = mean(energy_ts_ts[, "Total Renewable Energy Production"], na.rm = TRUE),
    color = "blue",
    linetype = "dashed"
  ) +
```

```

  labs(
    title = "Total Renewable Energy Production (Monthly)",
    x = "Year",
    y = "Energy Production")

```

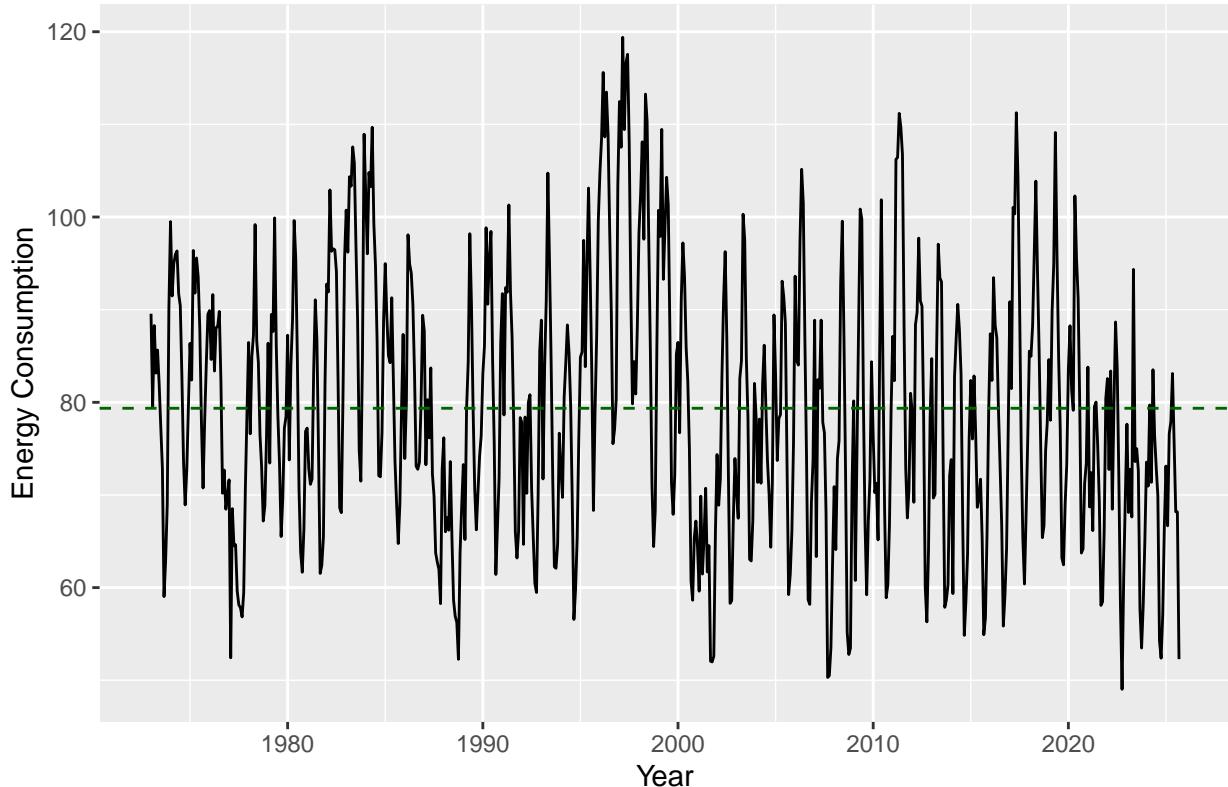


```

# Hydroelectric Power Consumption
autoplot(energy_ts_ts[, "Hydroelectric Power Consumption"]) +
  geom_hline(
    yintercept = mean(energy_ts_ts[, "Hydroelectric Power Consumption"], na.rm = TRUE),
    color = "darkgreen",
    linetype = "dashed"
  ) +
  labs(
    title = "Hydroelectric Power Consumption (Monthly)",
    x = "Year",
    y = "Energy Consumption")

```

## Hydroelectric Power Consumption (Monthly)



### Question 5

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

```
cor.test(  
  energy_ts_ts[, "Total Biomass Energy Production"],  
  energy_ts_ts[, "Total Renewable Energy Production"])  
  
##  
## Pearson's product-moment correlation  
##  
## data: energy_ts_ts[, "Total Biomass Energy Production"] and energy_ts_ts[, "Total Renewable Energy Production"]  
## t = 92.851, df = 631, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
##  0.9595516 0.9702413  
## sample estimates:  
##       cor  
## 0.9652985  
  
cor.test(  
  energy_ts_ts[, "Total Biomass Energy Production"],  
  energy_ts_ts[, "Hydroelectric Power Consumption"])
```

```

## 
## Pearson's product-moment correlation
##
## data: energy_ts_ts[, "Total Biomass Energy Production"] and energy_ts_ts[, "Hydroelectric Power Con
## t = -3.4157, df = 631, p-value = 0.000677
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.21045616 -0.05741173
## sample estimates:
##          cor
## -0.1347374

cor.test(
  energy_ts_ts[, "Total Renewable Energy Production"],
  energy_ts_ts[, "Hydroelectric Power Consumption"])

## 
## Pearson's product-moment correlation
##
## data: energy_ts_ts[, "Total Renewable Energy Production"] and energy_ts_ts[, "Hydroelectric Power C
## t = -1.4701, df = 631, p-value = 0.142
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.13573488 0.01959335
## sample estimates:
##          cor
## -0.05842436

```

The correlation coefficient between total biomass energy production and total renewable energy production is 0.965, which means that they have a very strong positive linear relationship. Also, the p-value < 2.2e-16, which means the two series are significantly and strongly correlated.

The correlation between total biomass energy production and hydroelectric power consumption is  $-0.135$ , which means that they have a weak negative linear relationship. Also, the p-value = 0.0006, meaning that despite being weak in magnitude, the relationship between the two series is statistically different from zero.

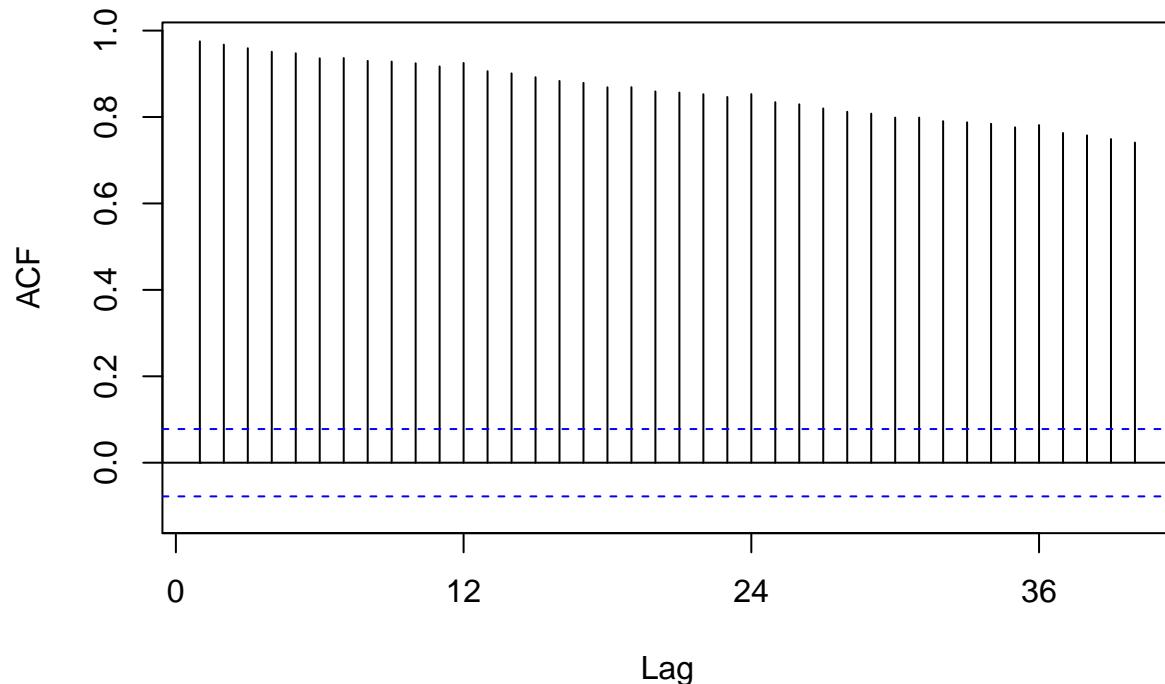
The correlation between total renewable energy production and hydroelectric power consumption is  $-0.058$ , which is very weak. What's more, p-value = 0.142 so the hypothesis test fails to reject the null hypothesis and there is no statistically significant linear correlation between these two series.

## 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?

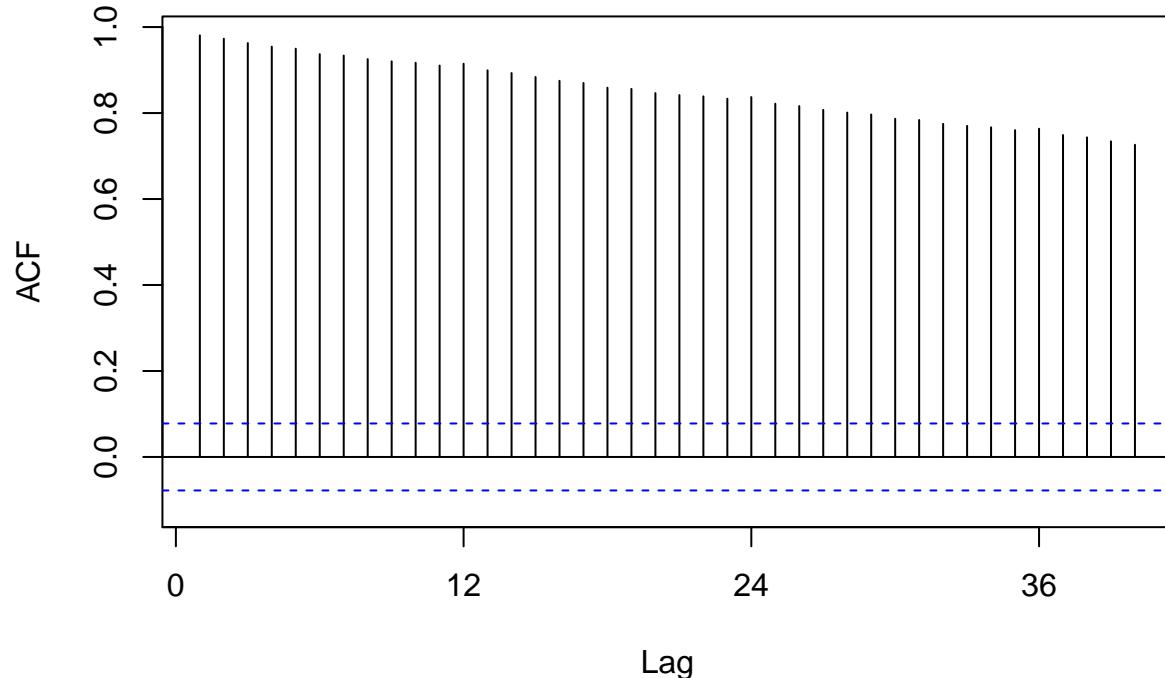
```
BE_acf <- Acf(energy_ts_ts[, "Total Biomass Energy Production"], lag.max = 40, plot = TRUE, main = "ACF
```

## ACF: Total Biomass Energy Production



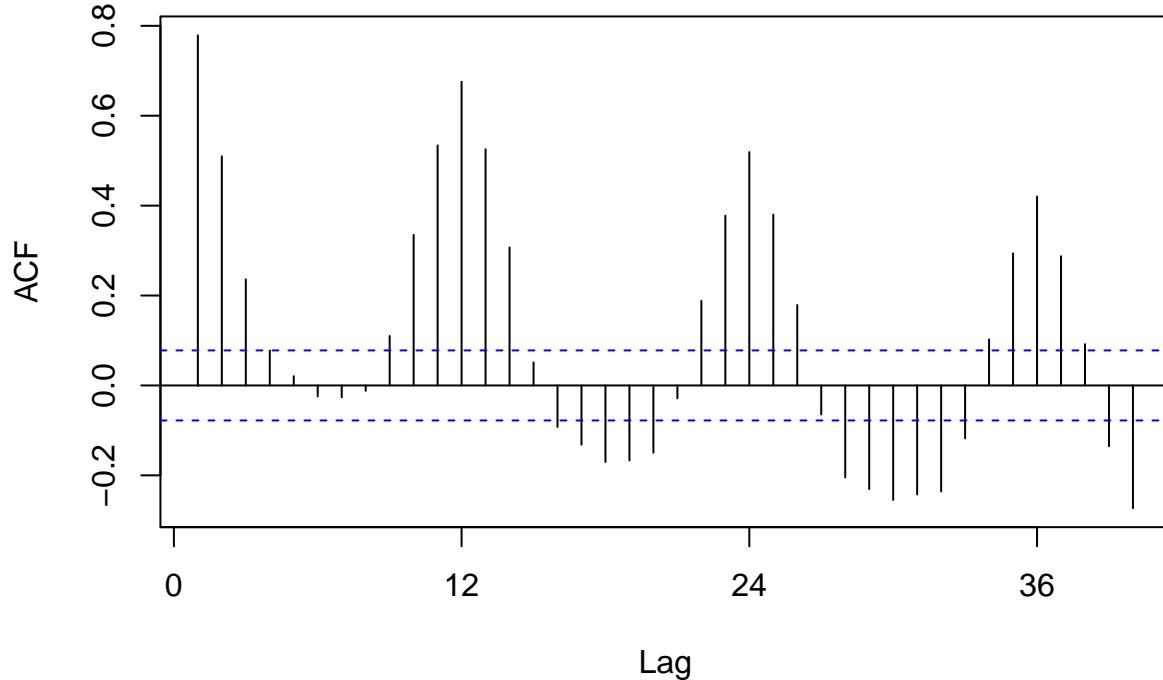
```
RE_acf <- Acf(energy_ts_ts[, "Total Renewable Energy Production"], lag.max = 40, plot = TRUE, main = "A")
```

## ACF: Total Renewable Energy Production



```
HP_acf <- Acf(energy_ts_ts[, "Hydroelectric Power Consumption"], lag.max = 40, plot = TRUE, main = "ACF:
```

## ACF: Hydroelectric Power Consumption



```
BE_acf
```

```
##  
## Autocorrelations of series 'energy_ts_ts[, "Total Biomass Energy Production"]', by lag  
##  
##   0    1    2    3    4    5    6    7    8    9    10   11   12  
## 1.000 0.975 0.968 0.959 0.951 0.948 0.936 0.937 0.930 0.928 0.924 0.917 0.925  
## 13   14   15   16   17   18   19   20   21   22   23   24   25  
## 0.906 0.901 0.892 0.883 0.879 0.869 0.869 0.859 0.857 0.853 0.846 0.853 0.834  
## 26   27   28   29   30   31   32   33   34   35   36   37   38  
## 0.829 0.820 0.812 0.808 0.799 0.799 0.790 0.788 0.784 0.776 0.781 0.763 0.758  
## 39   40  
## 0.749 0.741
```

```
RE_acf
```

```
##  
## Autocorrelations of series 'energy_ts_ts[, "Total Renewable Energy Production"]', by lag  
##  
##   0    1    2    3    4    5    6    7    8    9    10   11   12  
## 1.000 0.981 0.973 0.963 0.955 0.950 0.937 0.934 0.925 0.920 0.917 0.910 0.915  
## 13   14   15   16   17   18   19   20   21   22   23   24   25  
## 0.899 0.893 0.884 0.875 0.870 0.859 0.856 0.847 0.842 0.839 0.833 0.837 0.822  
## 26   27   28   29   30   31   32   33   34   35   36   37   38  
## 0.816 0.807 0.801 0.797 0.787 0.784 0.775 0.770 0.767 0.760 0.764 0.749 0.743
```

```
##      39     40  
## 0.734 0.726
```

```
HP_acf
```

```
##  
## Autocorrelations of series 'energy_ts_ts[, "Hydroelectric Power Consumption"]', by lag  
##  
##      0      1      2      3      4      5      6      7      8      9      10  
## 1.000  0.779  0.510  0.236  0.078  0.021 -0.024 -0.027 -0.012  0.110  0.335  
##   11     12     13     14     15     16     17     18     19     20     21  
##  0.534  0.676  0.526  0.307  0.051 -0.092 -0.132 -0.170 -0.167 -0.150 -0.029  
##   22     23     24     25     26     27     28     29     30     31     32  
##  0.188  0.378  0.519  0.380  0.179 -0.065 -0.205 -0.230 -0.255 -0.243 -0.236  
##   33     34     35     36     37     38     39     40  
## -0.118  0.102  0.294  0.420  0.288  0.092 -0.135 -0.273
```

The autocorrelation function for total biomass energy production shows very high positive auto-correlations at all lags, with a very slow decay as the lag increases. Almost all the auto-correlation coefficients are significant, which means that this time series has a strong and non-stationary pattern.

The ACF of total renewable energy production remains large and positive across all lags with a slow decline. It is very similar to that of total biomass energy production, indicating a long-term trend.

Unlike the other two functions, the hydroelectric power consumption series shows a different autocorrelation structure with a cyclical behavior. The auto-correlations decay more rapidly than in the other two series, suggesting weaker long-term persistence with a strong seasonal trend.

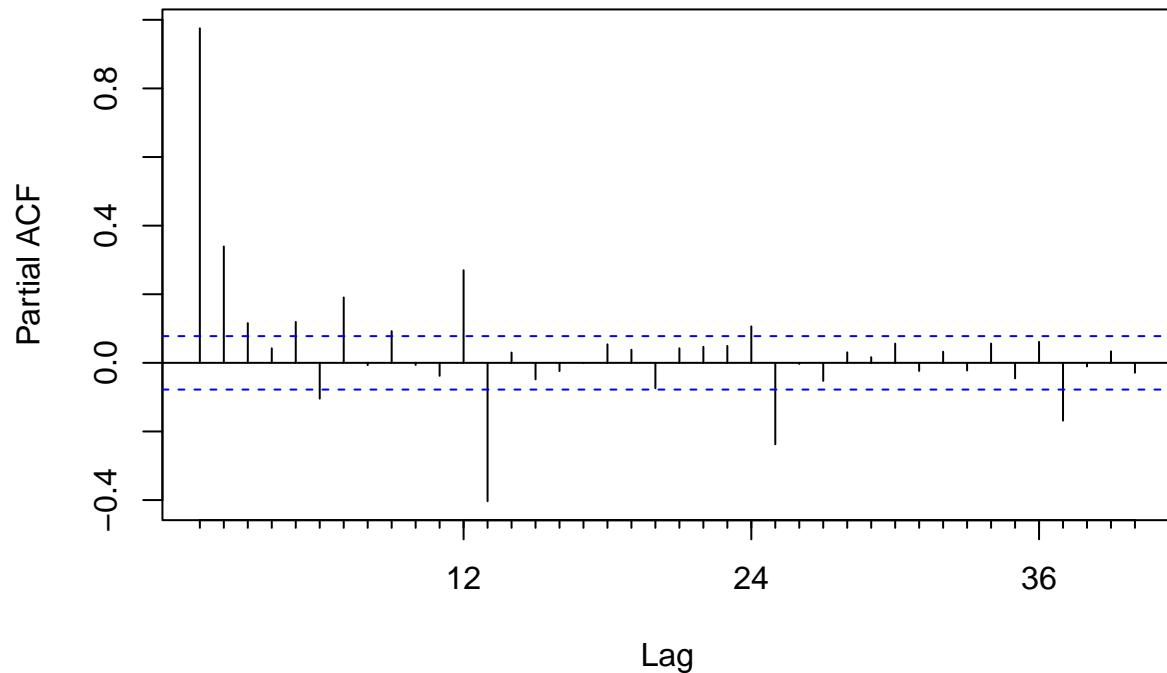
Compare: No, the three series do not show similar trend. Although the first two graphs are showing a similar strong positive auto-correlation, the last one is different with a seasonal trend.

## 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?

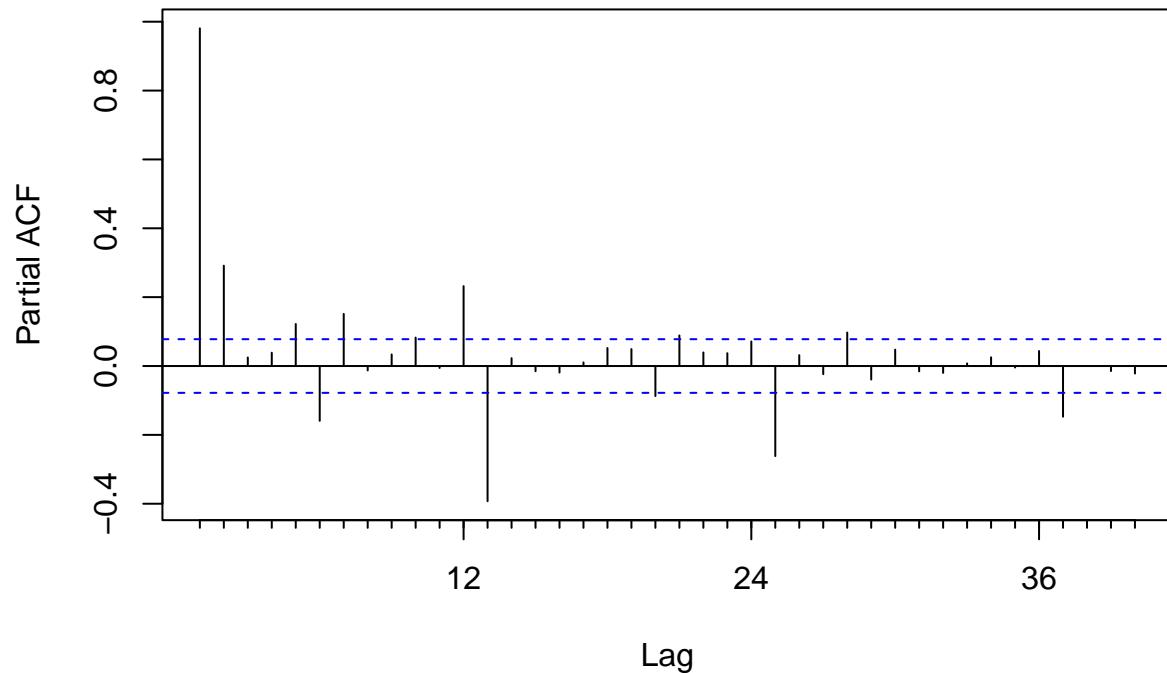
```
BE_pacf <- Pacf(energy_ts_ts[, "Total Biomass Energy Production"], lag.max = 40, plot = TRUE, main = "A")
```

## ACF: Total Biomass Energy Production



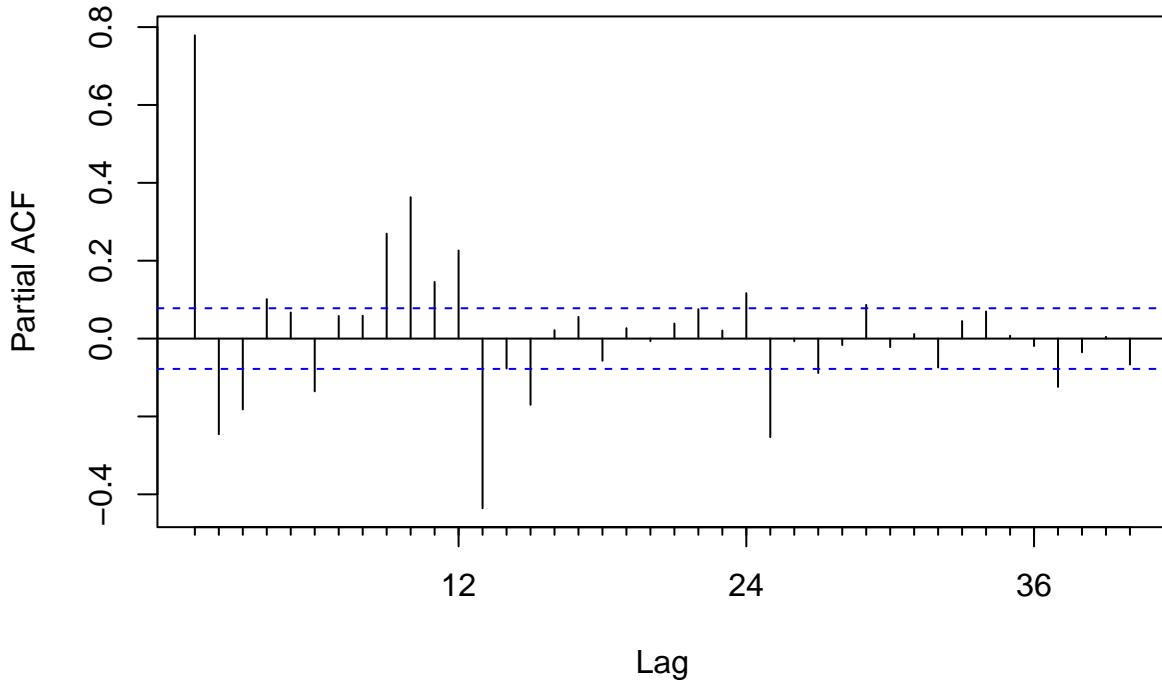
```
RE_pacf <- Pacf(energy_ts_ts[, "Total Renewable Energy Production"], lag.max = 40, plot = TRUE, main =
```

## ACF: Total Renewable Energy Production



```
HP_pacf <- Pacf(energy_ts_ts[, "Hydroelectric Power Consumption"], lag.max = 40, plot = TRUE, main = "ACF")
```

## ACF: Hydroelectric Power Consumption



```
BE_pacf
```

```
##
## Partial autocorrelations of series 'energy_ts_ts[, "Total Biomass Energy Production"]', by lag
##
##          1      2      3      4      5      6      7      8      9      10     11
##  0.975  0.339  0.116  0.042  0.119 -0.104  0.191 -0.007  0.092 -0.006 -0.038
##          12     13     14     15     16     17     18     19     20     21     22
##  0.270 -0.403  0.030 -0.048 -0.024 -0.001  0.054  0.038 -0.074  0.042  0.047
##          23     24     25     26     27     28     29     30     31     32     33
##  0.049  0.106 -0.237 -0.003 -0.053  0.031  0.017  0.056 -0.024  0.032 -0.022
##          34     35     36     37     38     39     40
##  0.056 -0.045  0.061 -0.168 -0.010  0.034 -0.029
```

```
RE_pacf
```

```
##
## Partial autocorrelations of series 'energy_ts_ts[, "Total Renewable Energy Production"]', by lag
##
##          1      2      3      4      5      6      7      8      9      10     11
##  0.981  0.291  0.025  0.039  0.122 -0.159  0.151 -0.013  0.034  0.083 -0.006
##          12     13     14     15     16     17     18     19     20     21     22
##  0.232 -0.393  0.023 -0.015 -0.019  0.010  0.052  0.049 -0.087  0.089  0.039
##          23     24     25     26     27     28     29     30     31     32     33
##  0.037  0.071 -0.262  0.032 -0.024  0.097 -0.039  0.048 -0.016 -0.020  0.008
```

```
##      34      35      36      37      38      39      40
##  0.025 -0.004  0.044 -0.147  0.000 -0.015 -0.022
```

```
HP_pacf
```

```
##
## Partial autocorrelations of series 'energy_ts_ts[, "Hydroelectric Power Consumption"]', by lag
##
##      1      2      3      4      5      6      7      8      9      10     11
##  0.779 -0.246 -0.182  0.101  0.067 -0.135  0.058  0.059  0.269  0.363  0.146
##      12     13     14     15     16     17     18     19     20     21     22
##  0.226 -0.436 -0.077 -0.170  0.021  0.056 -0.057  0.027 -0.007  0.039  0.076
##      23     24     25     26     27     28     29     30     31     32     33
##  0.021  0.117 -0.253 -0.007 -0.088 -0.017  0.087 -0.022  0.012 -0.074  0.045
##      34     35     36     37     38     39     40
##  0.069  0.007 -0.019 -0.124 -0.035  0.004 -0.066
```

For biomass and total renewable energy production, their long-time memory patterns are dominated by short-lag dependence, especially lag 1, since their PACF shows a large and significant spike at lag 1 following be much smaller partial auto-correlations at higher lags.

For hydroelectric power consumption, the strong long-term trend is dominated by seasonal dynamics, since its PACF shows significant spikes at seasonal lags.