ENV 797 - Time Series Analysis for Energy Data | Spring 2024

Assignment 2 - Due date 02/25/24

Faustin Kambale

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.\

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

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*.

```
hw_data <- read.csv(file
="/Users/faustinkambale/Library/CloudStorage/OneDrive-DukeUniversity/Spring 2024 classes/Time Series 4 
                    header=TRUE,skip=0)
hw_data
```

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.

```
hw_data <- hw_data[, c('Month', 'Total.Biomass.Energy.Production',
                       'Total.Renewable.Energy.Production',
                       'Hydroelectric.Power.Consumption')]
head(hw_data)
str(hw_data)
```

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_renerg <- ts(hw_data$Total.Renewable.Energy.Production, start=c(1980,1),
                frequency=12) # ts for Energy production
```

```
ts_bioenerg <- ts(hw_data$Total.Biomass.Energy.Production, start=c(1980,1),
                  frequency=12) # ts for biomass energy production
```

```
ts_hydro <- ts(hw_data$Hydroelectric.Power.Consumption, start=c(1980,1),
               frequency=12) # ts for hydroelectric power consumption
```

Question 3

Compute mean and standard deviation for these three series.

```
avgbioenerg <-mean(hw_data$Total.Biomass.Energy.Production)
avgbioenerg # the mean for biomass energy
```

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

```
avgrenerg <-mean(hw_data$Total.Renewable.Energy.Production)
avgrenerg # the mean for renewable energy
```

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

```
avghydro <-mean(hw_data$Hydroelectric.Power.Consumption)
avghydro # the mean for electricity consumption
```

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

```
sdbioenerg <-sd(hw_data$Total.Biomass.Energy.Production)
sdbioenerg # the standard deviation for biomass energy
```

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

```
sdrenerg <-sd(hw_data$Total.Renewable.Energy.Production)
sdrenerg # the standard deviation for renewable energy
```

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

```
sdhydro <-sd(hw_data$Hydroelectric.Power.Consumption)
sdhydro # the standard deviation for electricity consumption
```

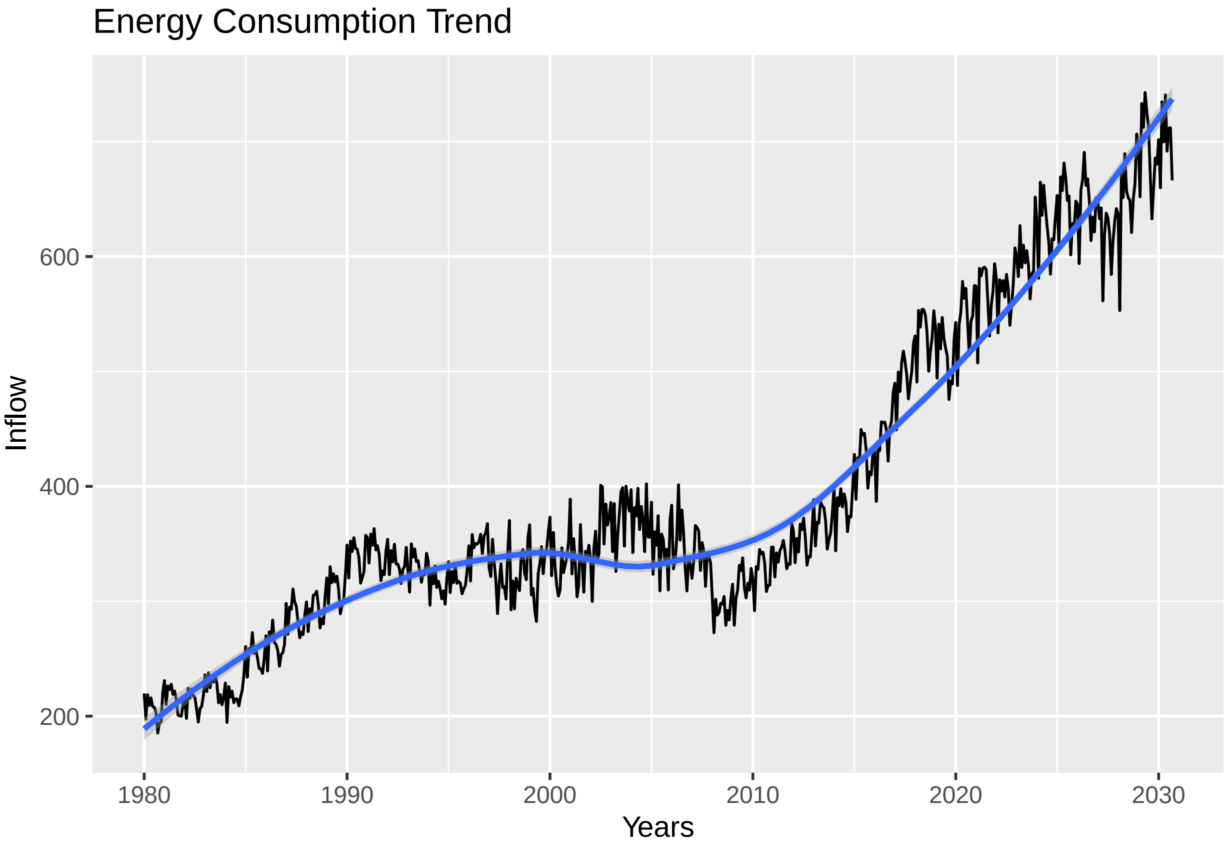
```
## [1] 14.14734
```

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.

```
autoplot(ts_renerg, main="Energy Consumption Trend") +
  # autoplot for energy consumption
  geom_smooth()+
  xlab("Years") +
  ylab("Inflow") +
  labs(color="Reservoir")
```

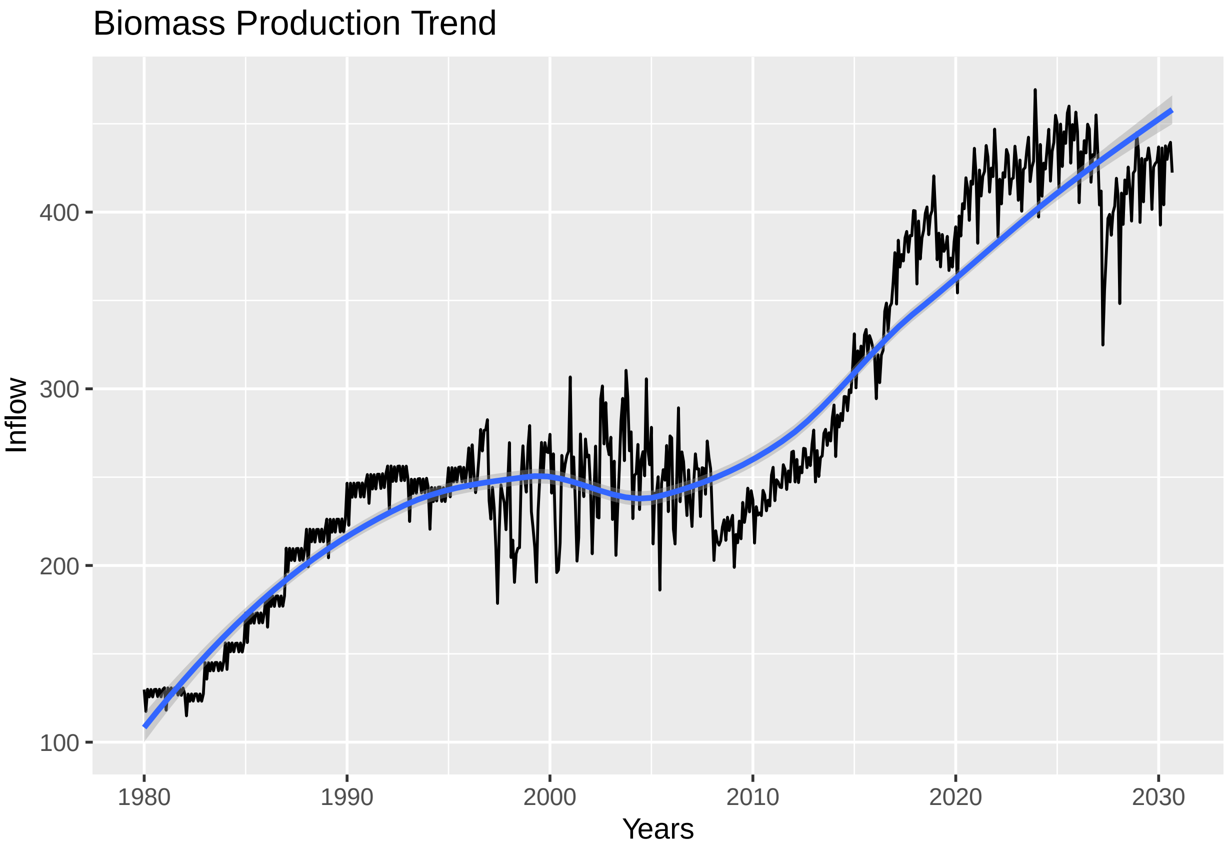
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



The graph shows the trend in energy consumption from 1980 to 2023. From the trend line, we can see a steady increasing in energy consumption trend over that period.

```
autoplot(ts_bioenerg, main = "Biomass Production Trend") +
  # autoplot for biomass energy production
  geom_smooth()+
  xlab("Years") +
  ylab("Inflow") +
  labs(color="Reservoir")
```

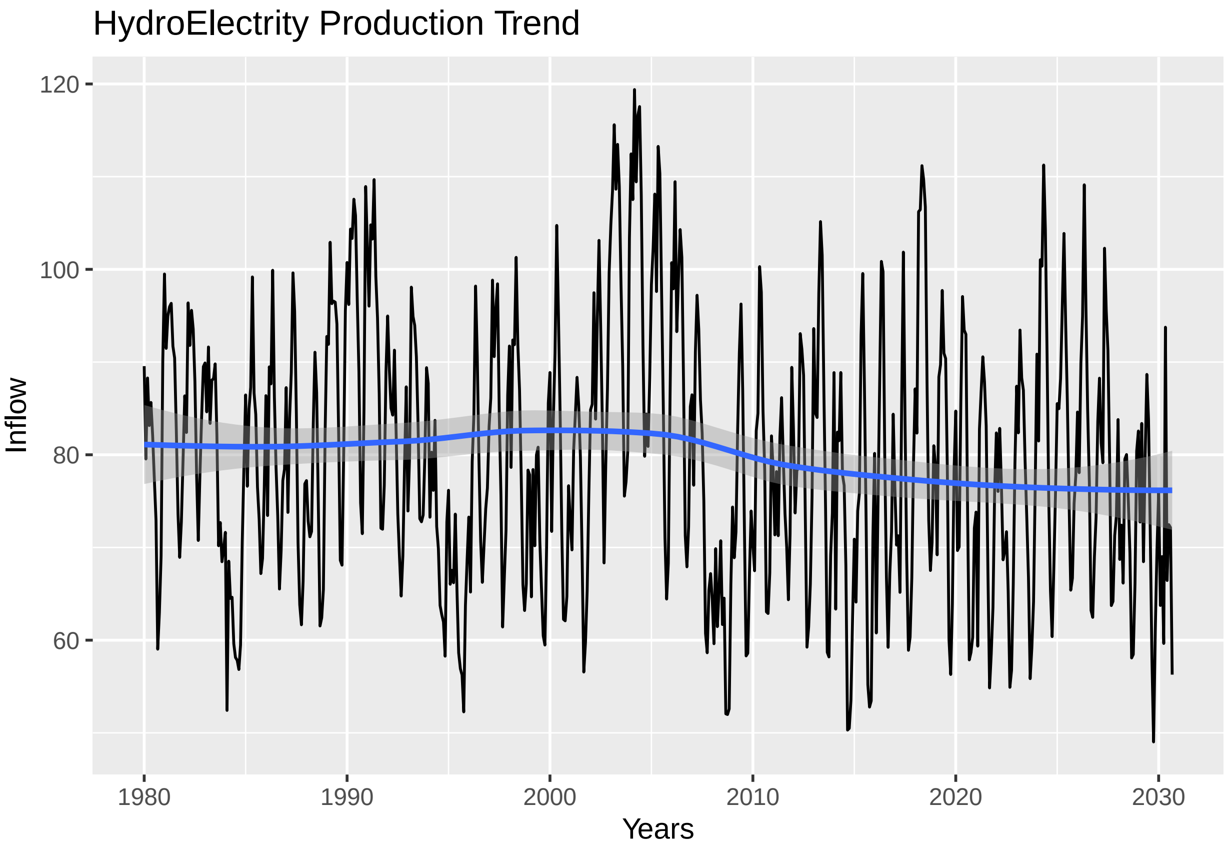
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



This graph shows an increasing biomas production trend during the period 1980 to 2023 (with a projection until 2030). Though flustuating, this trend is increasing over time.

```
autoplot(ts_hydro, main = "HydroElectrity Production Trend") +
  # autoplot for hydro electricity production
  geom_smooth()+
  xlab("Years") +
  ylab("Inflow") +
  labs(color="Reservoir")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



This graph displays some variability in the hydroelectric production during the period between 1980 and 2023. However, the trend is not clearly identifiable and would necessitate additional analysis to determine the direction.

Question 5

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

```r
correlation1 <- cor(hw_data$Total.Biomass.Energy.Production,
                    hw_data$Total.Renewable.Energy.Production)
print(correlation1)
```

```
## [1] 0.9707462
```

```r
correlation2 <- cor(hw_data$Total.Biomass.Energy.Production,
                    hw_data$Hydroelectric.Power.Consumption)
print(correlation2)
```

```
## [1] -0.09656318
```

```r
correlation3 <- cor(hw_data$Total.Renewable.Energy.Production,
                    hw_data$Hydroelectric.Power.Consumption)
print(correlation3)
```

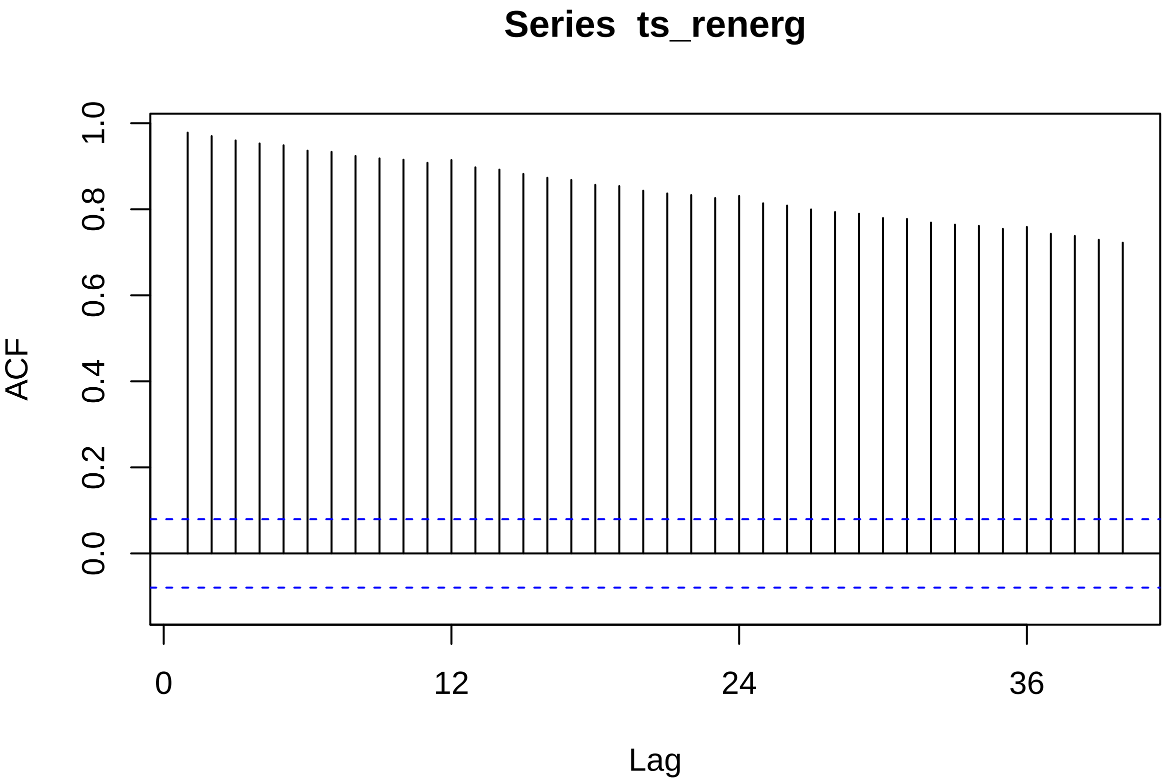
```
## [1] -0.001768629
```

This analysis show a strong and positive correlation between Biomass Energy production and Renewable energy production variables (correlation1) and negative correlations between other variables (correlation 2 and 3).

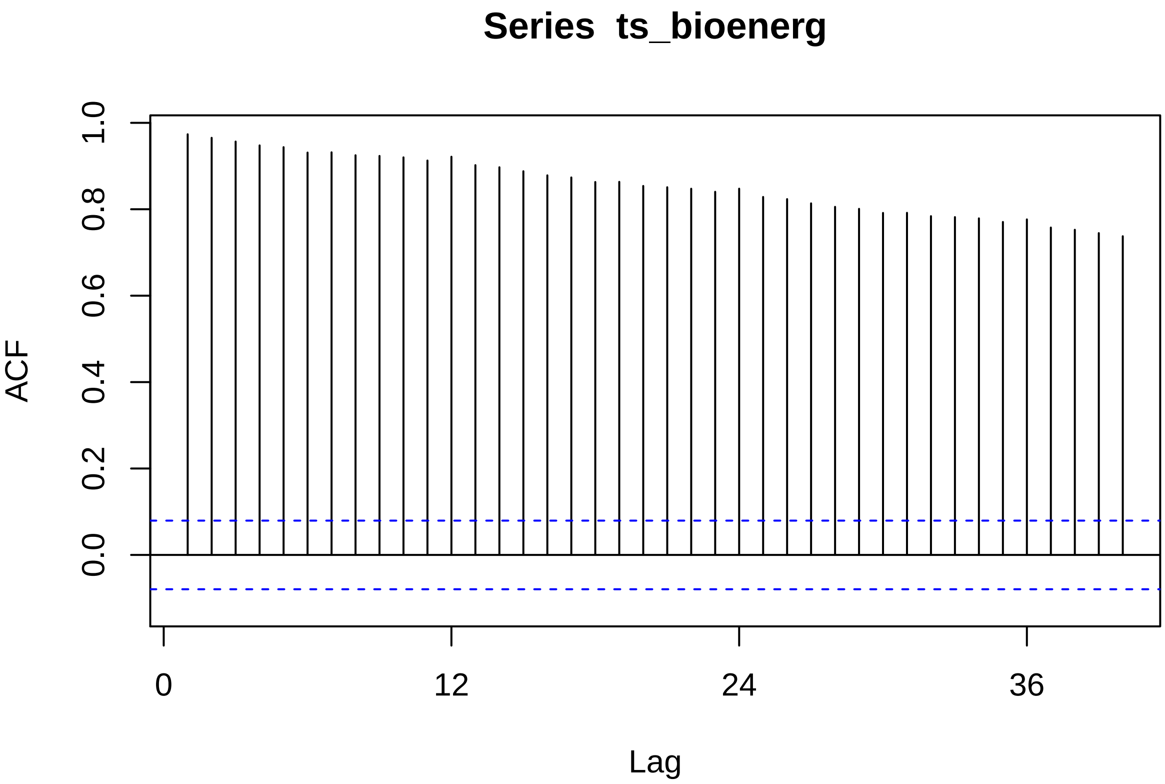
Question 6

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?

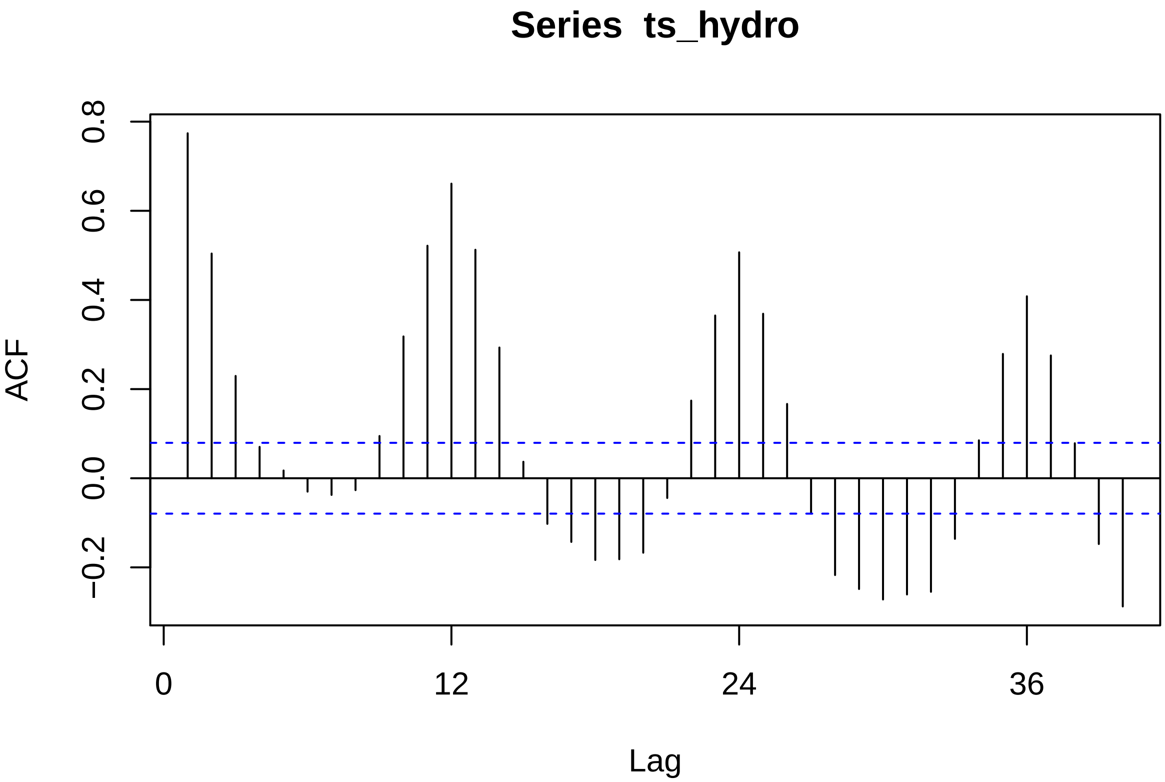
```r
Energy_acf <- Acf(ts_renerg, lag.max=40)
```



```r
Biom_acf <- Acf(ts_bioenerg, lag.max = 40)
```



```r
Hydro_acf <- Acf(ts_hydro, lag.max = 40)
```

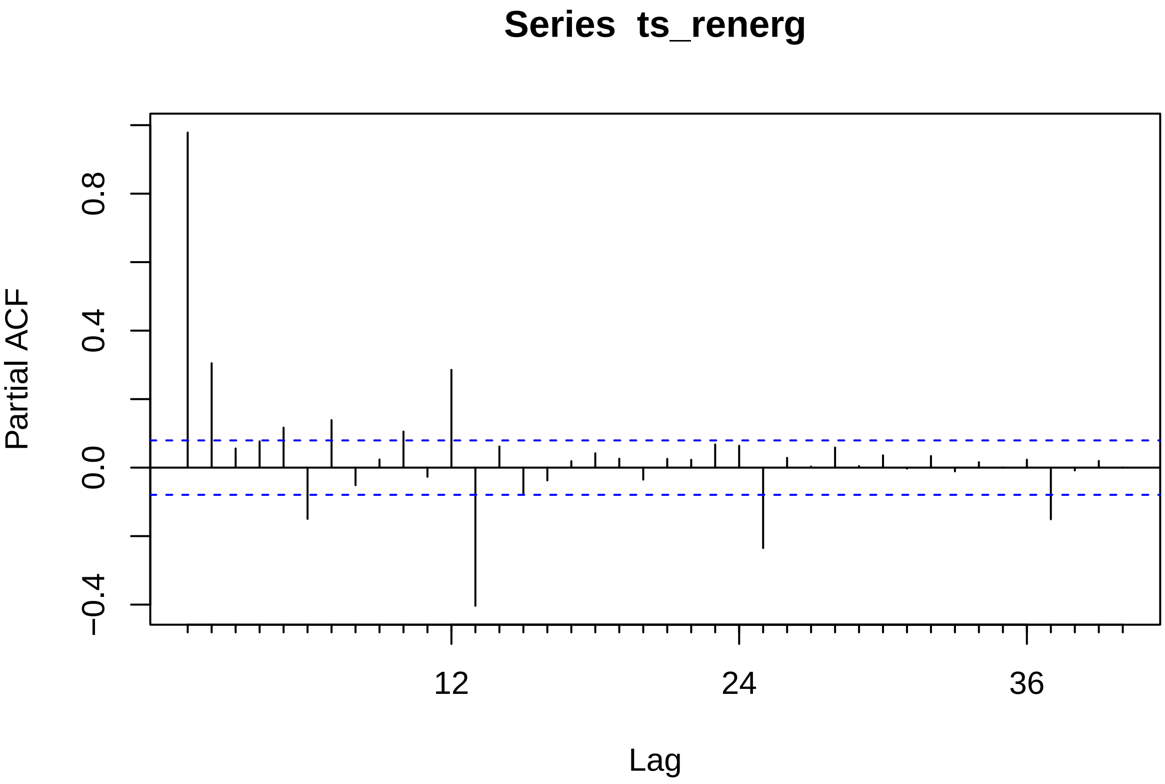


These plots show a common positive autocorrelation among them. However, the Hydroelectricity (plot3) shows some negative auto correlations while plot 1 and 2 only show positive auto correlations.

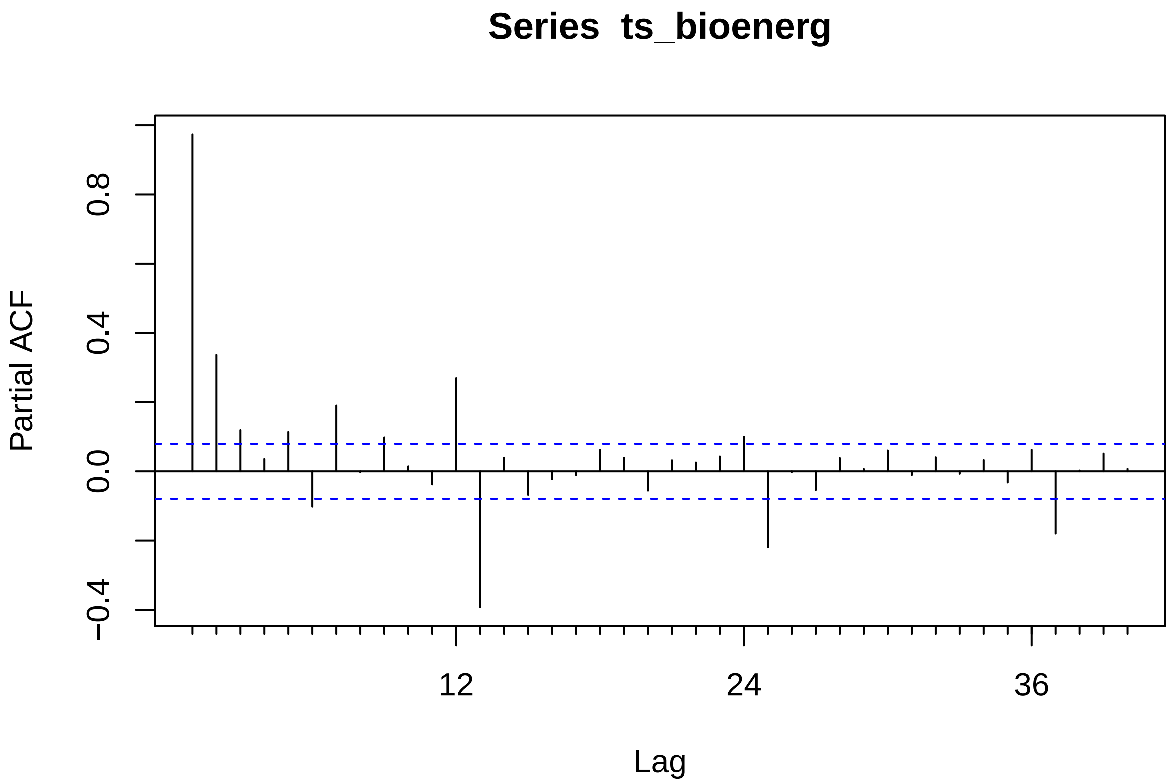
Question 7

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

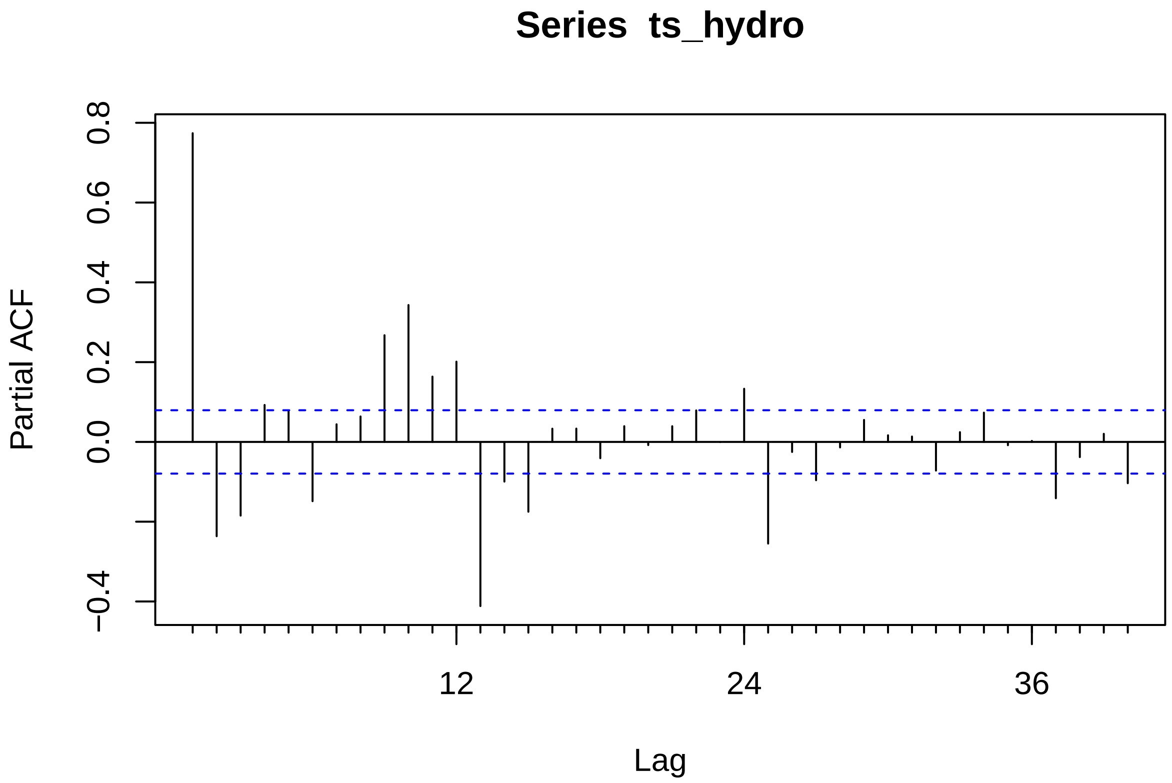
```
Energy_pacf <- Pacf(ts_renerg, lag.max=40)
```



```
Biom_pacf <- Pacf(ts_bioenerg, lag.max = 40)
```



```r
Hydro_pacf <- Pacf(ts_hydro, lag.max = 40)
```



For me, the main difference between previous autocorrelation (acf) and these partial autocorrelation (pacf) is the negative values. In the previous auto correlations, only the *Hydro_acf* was displaying a negative autocorrelation, while all partial auto correlations display negative values.

Also, one can easily note that the acf clearly shows trend in the data (declining trend of energy and biomass production), making interpretation straightforward.