ENV 790.30 - Time Series Analysis for Energy Data | Spring 2021

Assignment 2 - Due date 02/05/21

Traian Nirca

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 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. Rename the pdf file such that it includes your first and last name (e.g., "LuanaLima_TSA_A02_Sp21.Rmd"). 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
#install.packages("xlsx")
library(xlsx)
```

```
## Warning: package 'xlsx' was built under R version 4.0.3
```

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 January 2021 Monthly Energy Review. The spreadsheet is ready to be used. Modified: You may use the command *read.xlsx*() from package *xlsx* to import the data in R. You will need to install the package if you haven't done it yet. Since this is a excel file, you need to specify which sheet to import. Here I am doing it with argument *sheetIndex=1*. You could also use *sheetName="Monthly Data"*. Since after the header you have a row with units, I am skipping the header on the first read.xlsx command. And then I call the function again just to get row 11 which has the column names. Keep in mind that there are other way to import this file. You could save it as *.csv* and then use the *read.table()* or *read.csv()*, but I wanted to share a way to read it as it is.

```
#Importing data set without change the original file using read.xlsx
energy_data <- read.xlsx(file="../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source

#Now let's extract the column names from row 11
read_col_names <- read.xlsx(file="../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sou
colnames(energy_data) <- read_col_names
head(energy_data)
```

```
##        Month Wood Energy Production Biofuels Production
## 1 1973-01-01                129.630       Not Available
## 2 1973-02-01                117.194       Not Available
## 3 1973-03-01                129.763       Not Available
## 4 1973-04-01                125.462       Not Available
## 5 1973-05-01                129.624       Not Available
## 6 1973-06-01                125.435       Not Available
##   Total Biomass Energy Production Total Renewable Energy Production
## 1                         129.787                           403.981
## 2                         117.338                           360.900
## 3                         129.938                           400.161
## 4                         125.636                           380.470
## 5                         129.834                           392.141
## 6                         125.611                           377.232
##   Hydroelectric Power Consumption Geothermal Energy Consumption
## 1                         272.703                         1.491
## 2                         242.199                         1.363
## 3                         268.810                         1.412
## 4                         253.185                         1.649
## 5                         260.770                         1.537
## 6                         249.859                         1.763
##   Solar Energy Consumption Wind Energy Consumption Wood Energy Consumption
## 1            Not Available           Not Available                 129.630
## 2            Not Available           Not Available                 117.194
## 3            Not Available           Not Available                 129.763
## 4            Not Available           Not Available                 125.462
## 5            Not Available           Not Available                 129.624
## 6            Not Available           Not Available                 125.435
##   Waste Energy Consumption Biofuels Consumption
## 1                    0.157        Not Available
## 2                    0.144        Not Available
## 3                    0.176        Not Available
## 4                    0.174        Not Available
## 5                    0.210        Not Available
## 6                    0.176        Not Available
##   Total Biomass Energy Consumption Total Renewable Energy Consumption
## 1                          129.787                            403.981
## 2                          117.338                            360.900
## 3                          129.938                            400.161
## 4                          125.636                            380.470
## 5                          129.834                            392.141
## 6                          125.611                            377.232
```

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.

```
work_data <- data.frame("Total Biomass Energy Production"=energy_data[,4], "Total Renewable Energy Produ
head(work_data)
```

```
##   Total.Biomass.Energy.Production Total.Renewable.Energy.Production
## 1                         129.787                           403.981
## 2                         117.338                           360.900
## 3                         129.938                           400.161
## 4                         125.636                           380.470
## 5                         129.834                           392.141
## 6                         125.611                           377.232
##   Hydroelectric.Power.Consumption
## 1                         272.703
## 2                         242.199
## 3                         268.810
## 4                         253.185
## 5                         260.770
## 6                         249.859
```

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().

```
time_series <- ts(work_data, frequency = 12, start = c(1973,1))
#It is a monthly data so the frequency is 12, while the first point is January 1973
head(time_series)
```

```
##      Total.Biomass.Energy.Production Total.Renewable.Energy.Production
## [1,]                         129.787                           403.981
## [2,]                         117.338                           360.900
## [3,]                         129.938                           400.161
## [4,]                         125.636                           380.470
## [5,]                         129.834                           392.141
## [6,]                         125.611                           377.232
##      Hydroelectric.Power.Consumption
## [1,]                         272.703
## [2,]                         242.199
## [3,]                         268.810
## [4,]                         253.185
## [5,]                         260.770
## [6,]                         249.859
```

Question 3

Compute mean and standard deviation for these three series.

```
cat("For Total Biomass Energy Production:\n")
```

```
## For Total Biomass Energy Production:
```

```
cat("Mean_bio =", mean(time_series[,1]), "\nSd_bio =", sd(time_series[,1]))
```

```
## Mean_bio = 270.6961
## Sd_bio = 87.36311
```

```
cat("\nFor Total Renewable Energy Production:\n")
```

```
##
## For Total Renewable Energy Production:
```

```
cat("Mean_ren =", mean(time_series[,2]), "\nSd_ren =", sd(time_series[,2]))
```

```
## Mean_ren = 572.7321
## Sd_ren = 168.4588
```

```
cat("\nHydroelectric Power Consumption:\n")
```

```
##
## Hydroelectric Power Consumption:
```

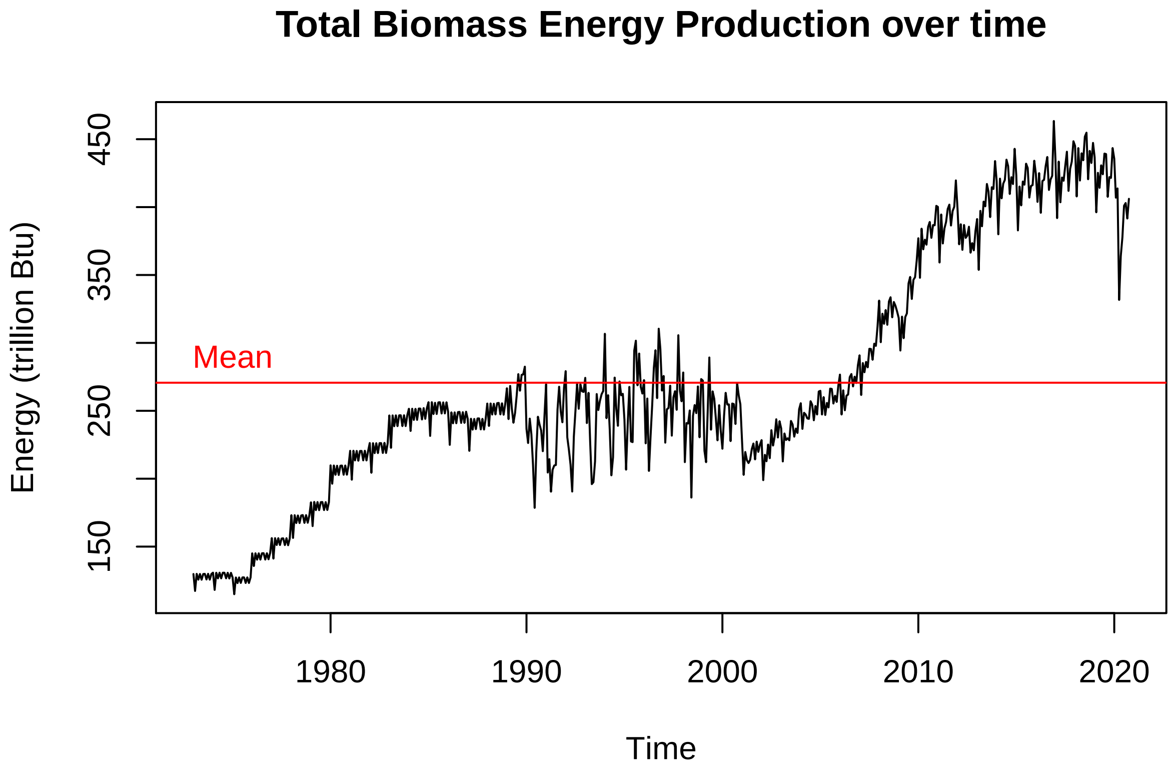
```
cat("Mean_hydro =", mean(time_series[,3]), "\nSd_hydro =", sd(time_series[,3]))
```

```
## Mean_hydro = 236.9515
## Sd_hydro = 43.90392
```

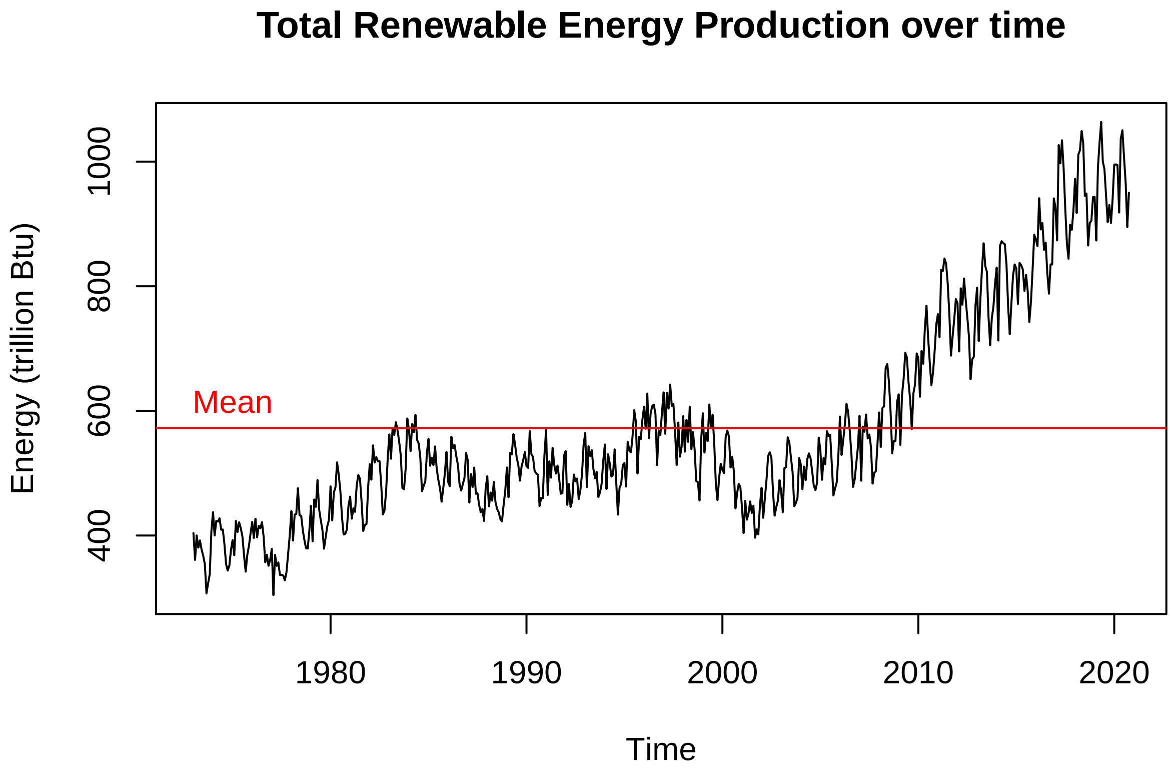
Question 4

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.

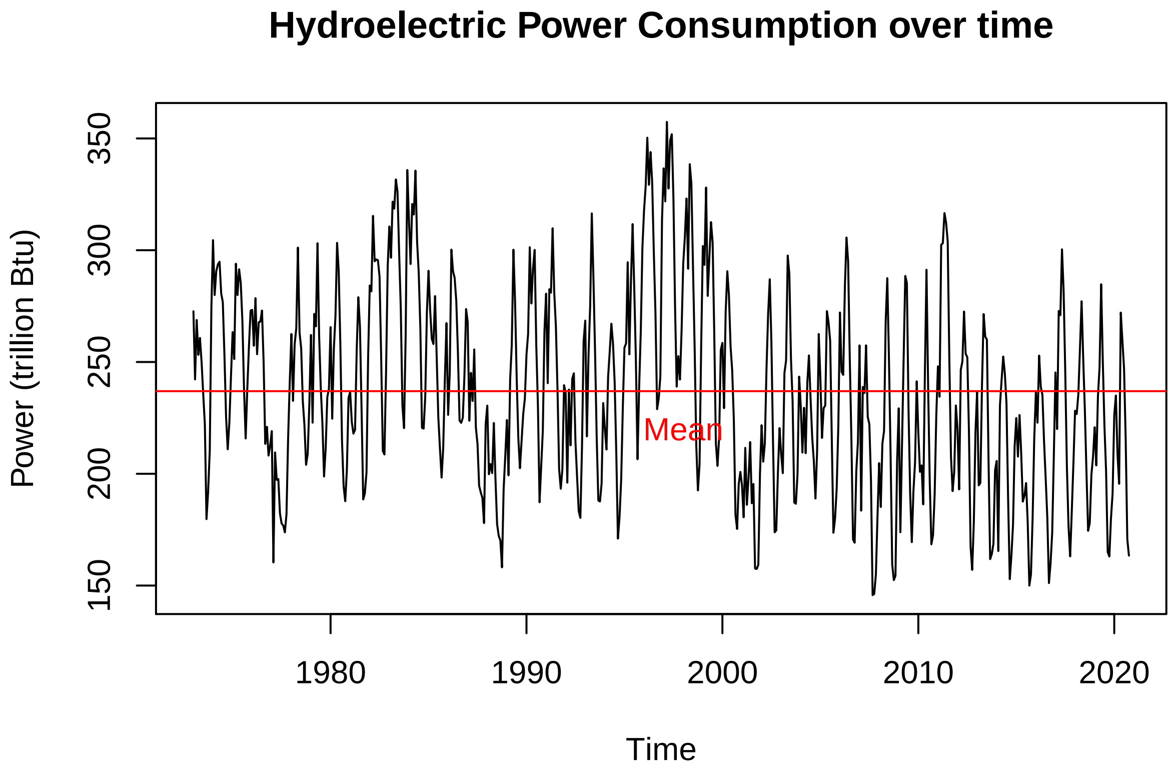
```
plot(time_series[,1], ylab="Energy (trillion Btu)", main="Total Biomass Energy Production over time")
abline(h=270.69, col="red")
text(1975, 290, "Mean", col="red")
```



```
plot(time_series[,2], ylab="Energy (trillion Btu)", main="Total Renewable Energy Production over time")
abline(h=572.73, col="red")
text(1975, 615, "Mean", col="red")
```



```
plot(time_series[,3], ylab="Power (trillion Btu)", main="Hydroelectric Power Consumption over time")
abline(h=236.95, col="red")
text(1998, 220, "Mean", col="red")
```



Both biomass and renewable energy production has been increasing significantly over time since the seventies. In 2020 we can see a drop in biomass energy production, most likely due to the fact that 2020 has been an outlier. hydroelectric power consumption has seen a small decrease since the early 2000.

Question 5

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

```
cat("CORRELATION between Total Biomass Energy Production and Total Renewable Energy Production:\n")
```

```
## CORRELATION between Total Biomass Energy Production and Total Renewable Energy Production:
```

```
cat("Cor =",cor(time_series[,1], time_series[,2]) )
```

```
## Cor = 0.9234609
```

```
cat("\nCORRELATION between Total Biomass Energy Production and Hydroelectric Power Consumption:\n")
```

```
##
## CORRELATION between Total Biomass Energy Production and Hydroelectric Power Consumption:
```

```
cat("Cor =",cor(time_series[,1], time_series[,3]) )
```

```
## Cor = -0.2555675
```

```
cat("\nCORRELATION between Total Renewable Energy Production and Hydroelectric Power Consumption:\n")
```

```
##
## CORRELATION between Total Renewable Energy Production and Hydroelectric Power Consumption:
```

```
cat("Cor =",cor(time_series[,2], time_series[,3]) )
```

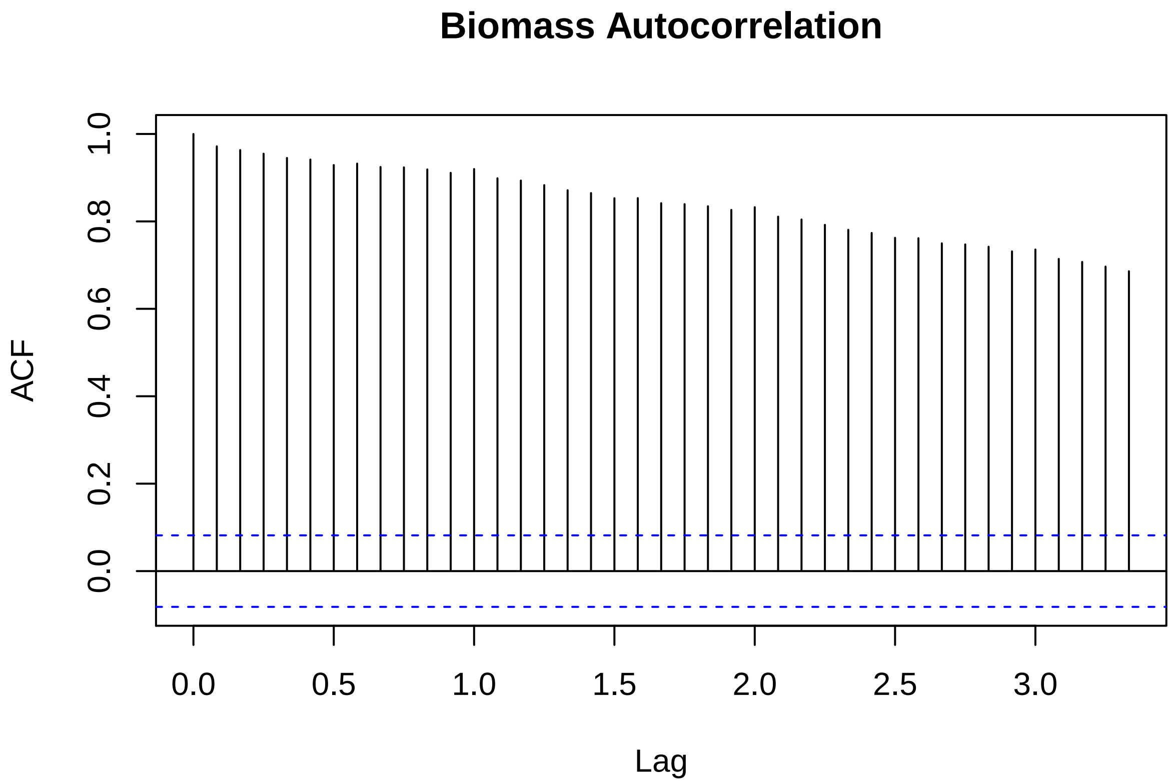
```
## Cor = -0.002756852
```

There is a very strong correlation between biomass and renewable energy production, we can even see it graphically when we compare the two plots. There is basically no linear correlation between the renewable energy production and the hydroelectric consumption, Cor~0. While there is some negative correlation between the biomass and hydroelectric energy, it is very weak. It is possible it is linked only to a few variables.

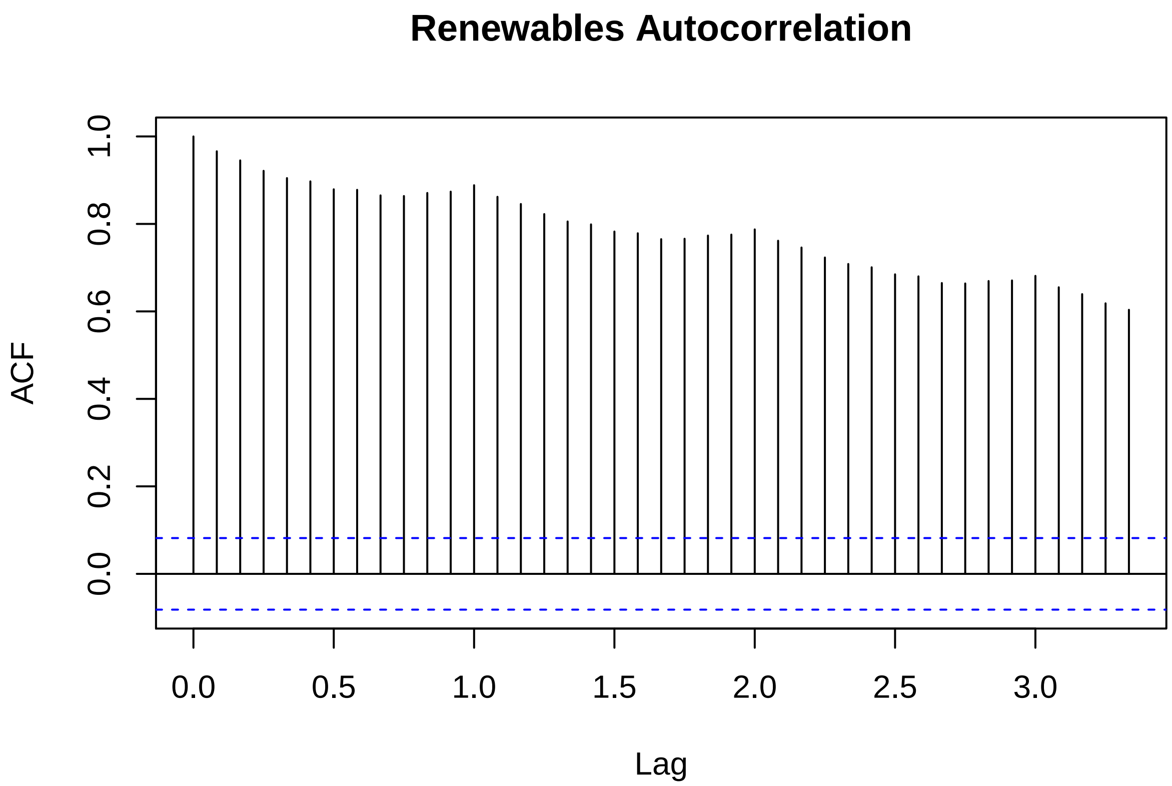
Question 6

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?

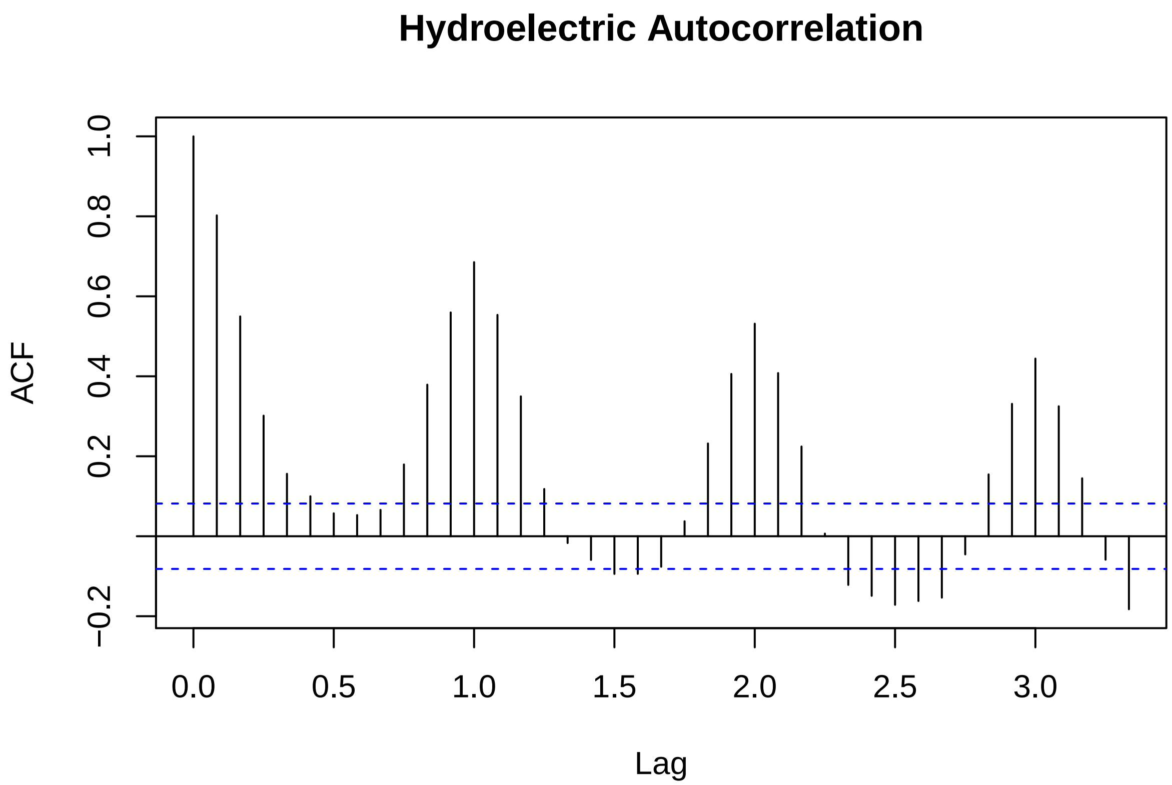
```
#cat("AUTOCORRELATION Function:\n")

acf(time_series[,1], lag.max = 40, type = c("correlation"), main="Biomass Autocorrelation")
```



```
acf(time_series[,2], lag.max = 40, type = c("correlation"), main="Renewables Autocorrelation")
```



```
acf(time_series[,3], lag.max = 40, type = c("correlation"), main="Hydroelectric Autocorrelation")
```

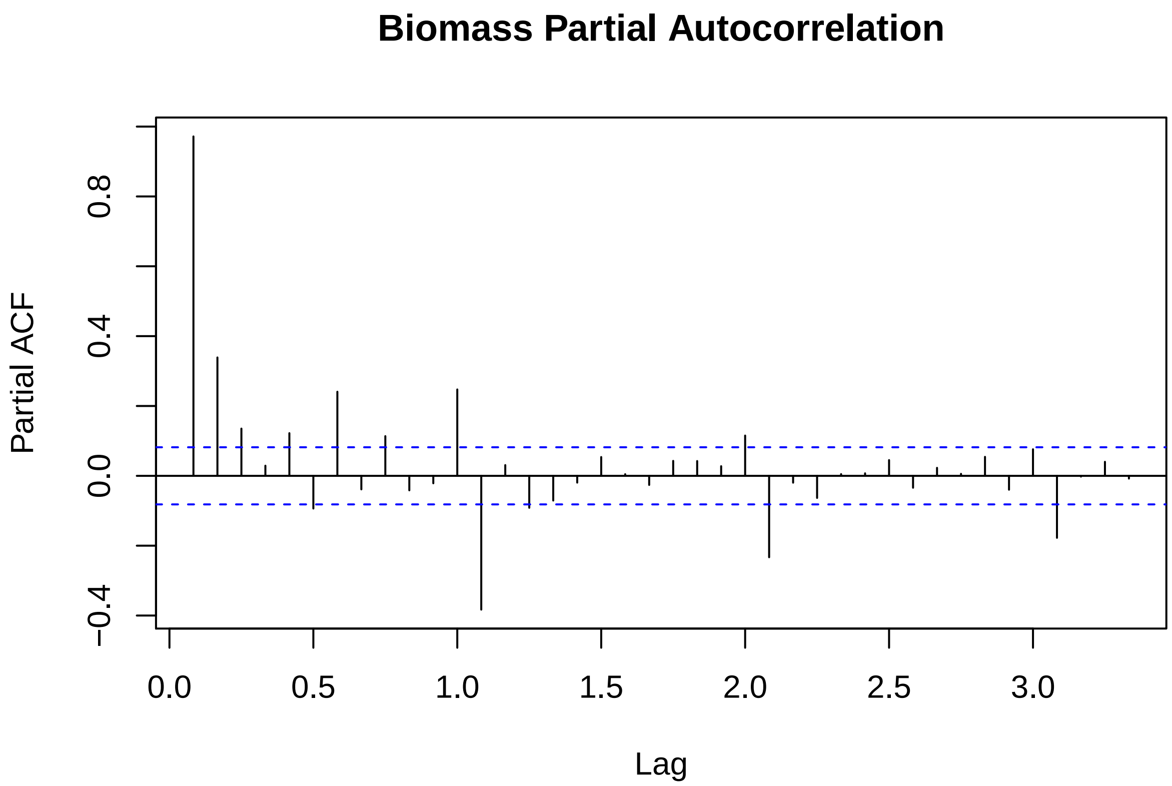


The three plots do not behave in the same way. The first two are decreasing, while the third is oscillating: this might be related to seasonality. In the second plot we also observe a very small oscillation in the decreasing trend.

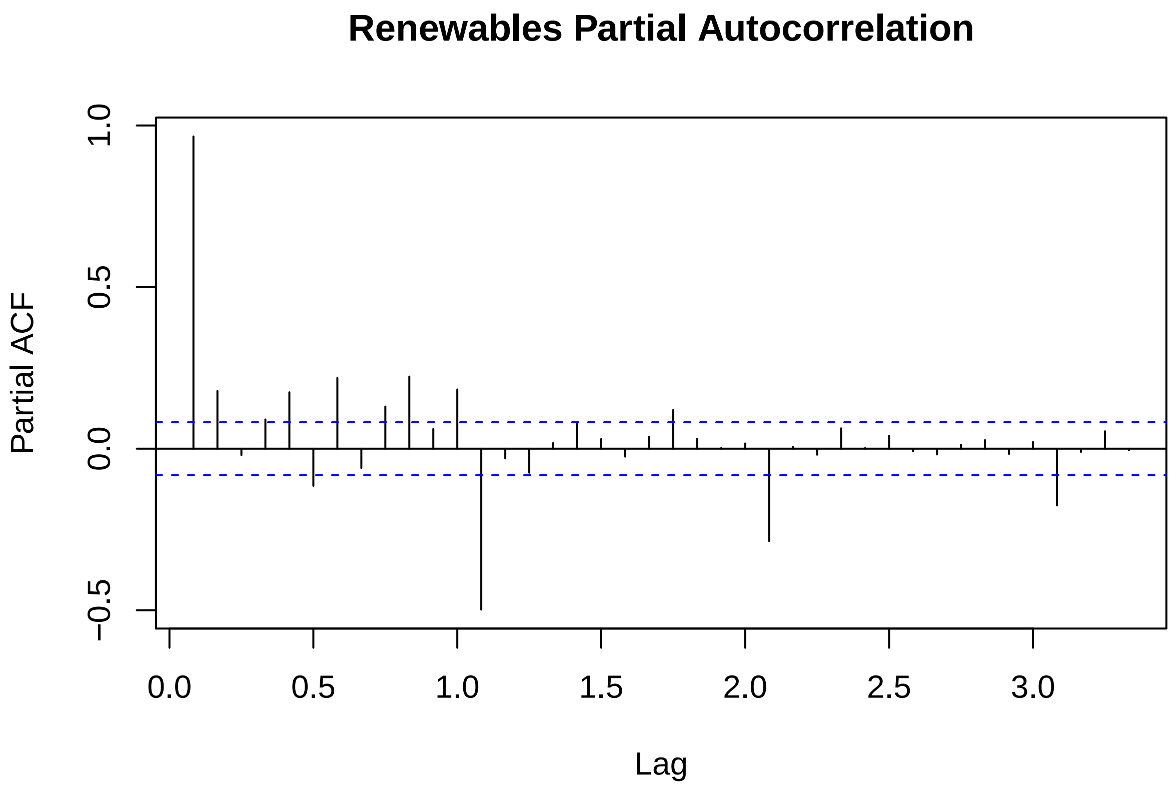
Question 7

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

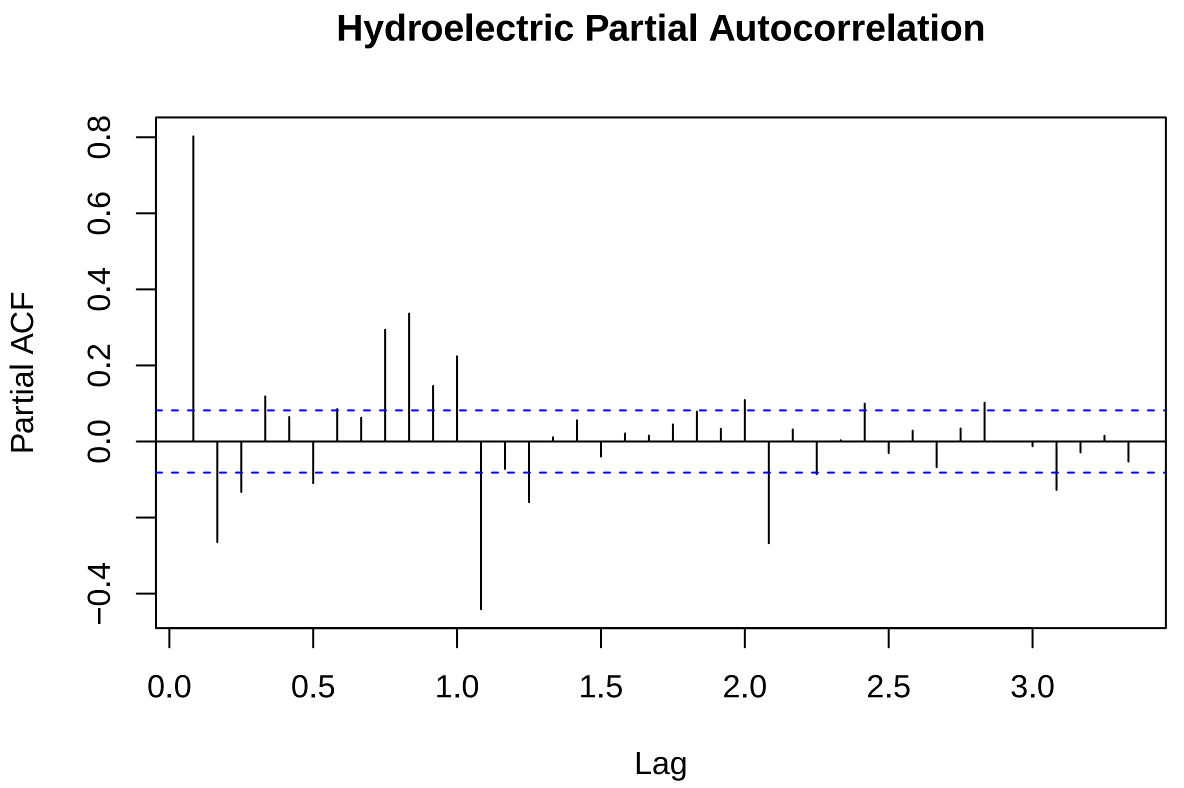
```
pacf(time_series[,1], lag.max = 40,  main="Biomass Partial Autocorrelation")
```



```
pacf(time_series[,2], lag.max = 40,  main="Renewables Partial Autocorrelation")
```



```
pacf(time_series[,3], lag.max = 40, main="Hydroelectric Partial Autocorrelation")
```



Not really sure how to comment these. There seems to be an absence of a pattern, compared to the ACF plots. There are significant correlations at the first lag, but small or insignificant afterwards: this indicates an autoregressive term in the data.