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

Assignment 2 - Due date 01/26/22

Yared S. Asfaw

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_Sp22.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.

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 January 2022 Monthly Energy Review. The spreadsheet is ready to be used. Use the command *read.table*() to import the data in R or *panda.read_excel*() in Python (note that you will need to import pandas package). }

```
#Importing the data set
library(openxlsx)
USEne.data <- read.xlsx("./Data/Raw/Table_10.1_Ren._Energy_PC._by_Source.xlsx", startRow=11)
```

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.

```
#Create data frame with the three variables
selected_var.df <- data.frame(USEne.data [,c(1,4,5,6)])

is.data.frame(selected_var.df)
```

```
## [1] TRUE
```

```
head(selected_var.df)
```

```
##   Month Total.Biomass.Energy.Production Total.Renewable.Energy.Production
## 1    NA                  (Trillion Btu)                    (Trillion Btu)
## 2 26665                         129.787                           403.981
## 3 26696                         117.338                             360.9
## 4 26724                         129.938                           400.161
## 5 26755                         125.636                            380.47
## 6 26785                         129.834                           392.141
##   Hydroelectric.Power.Consumption
## 1                  (Trillion Btu)
## 2                         272.703
## 3                         242.199
## 4                          268.81
## 5                         253.185
## 6                          260.77
```

```
selected_var.df.data <- selected_var.df[-1,]
#To remove the row that contains the unit of measurement
is.data.frame(selected_var.df.data)
```

```
## [1] TRUE
```

```
head(selected_var.df.data)
```

```
##   Month Total.Biomass.Energy.Production Total.Renewable.Energy.Production
## 2 26665                         129.787                           403.981
## 3 26696                         117.338                             360.9
## 4 26724                         129.938                           400.161
## 5 26755                         125.636                            380.47
## 6 26785                         129.834                           392.141
## 7 26816                         125.611                           377.232
##   Hydroelectric.Power.Consumption
## 2                         272.703
## 3                         242.199
## 4                          268.81
## 5                         253.185
## 6                          260.77
## 7                         249.859
```

```
#changing the class or type of the selected variables from character to numeric
#For simplicity let us rename the columns
colnames(selected_var.df.data) <- c("month", "biomass","renewable", "hydro")

selected_var.df.data$biomass <- as.numeric(selected_var.df.data$biomass)

selected_var.df.data$renewable <- as.numeric(selected_var.df.data$renewable)

selected_var.df.data$hydro <- as.numeric(selected_var.df.data$hydro)
```

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 object of Total Biomass Energy Production
ts.biomass <- ts(selected_var.df.data$biomass, start = c(1973, 1), frequency = 12)

#Time series object of Total Renewable Energy Production
ts.Re_energy <- ts(selected_var.df.data$renewable, start = c(1973, 1), frequency = 12)

#Time series object of Hydroelectric Power Consumption
ts.hyd_usage <- ts(selected_var.df.data$hydro, start = c(1973, 1), frequency = 12)
```

Question 3

Compute mean and standard deviation for these three series.

```
mean(ts.biomass)
```

```
## [1] 273.7839
```

The mean of the total biomass energy production for the given time series is 273.78 Trillion Btu.

```
sd(ts.biomass)
```

```
## [1] 89.42852
```

The standard deviation of the total biomass energy production for the given time series is 89.43 Trillion Btu.

```
mean(ts.Re_energy)
```

```
## [1] 581.1708
```

The mean of the total renewable energy production for the given time series is 581.17 Trillion Btu.

```
sd(ts.Re_energy)
```

```
## [1] 177.5607
```

The standard deviation of the total renewable energy production for the given time series is 177.56 Trillion Btu.

```
mean(ts.hyd_usage)
```

```
## [1] 235.9653
```

The mean of the hydroelectric power consumption for the given time series is 235.97 Trillion Btu.

```
sd(ts.hyd_usage)
```

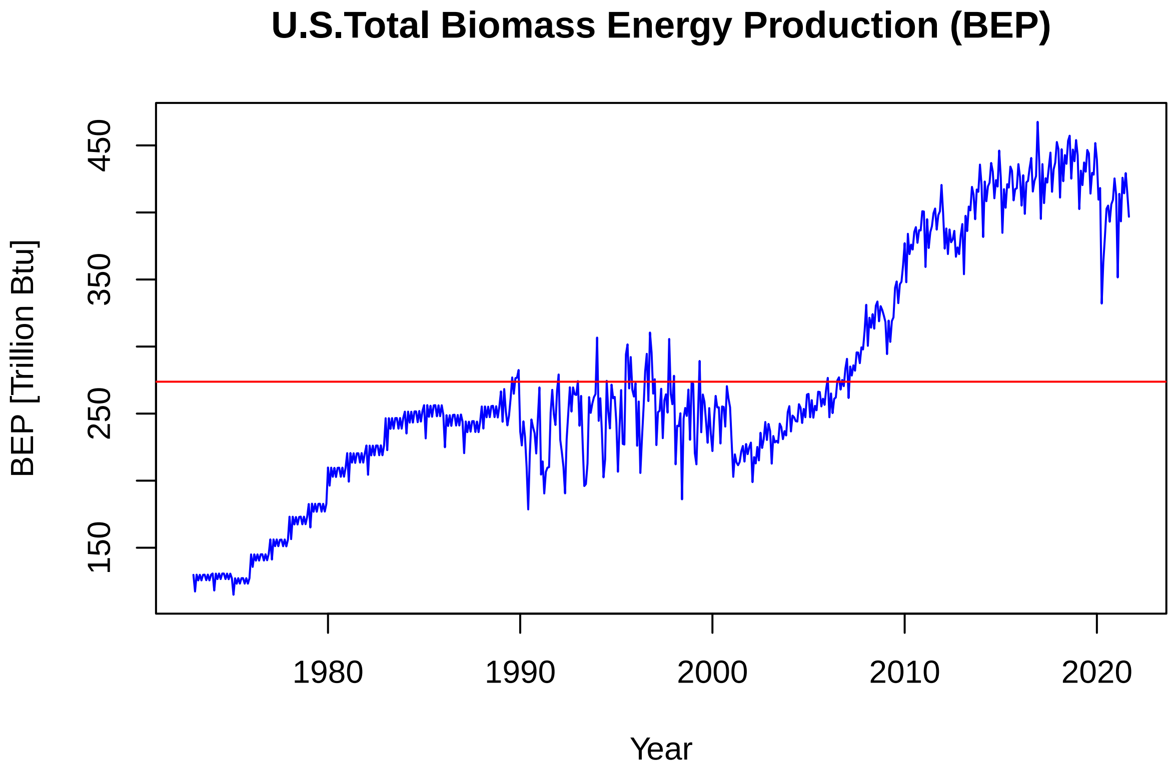
```
## [1] 44.01749
```

The standard deviation of the total biomass energy production for the given time series is 44.02 Trillion Btu.

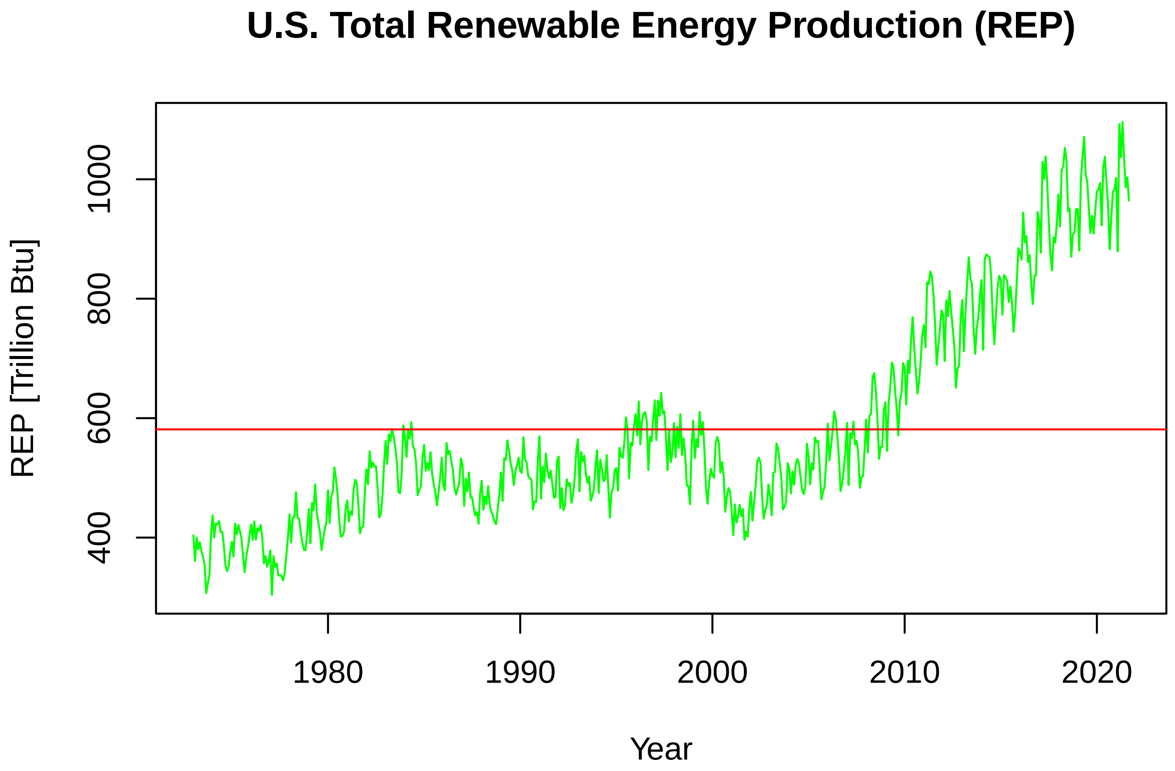
Question 4

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.

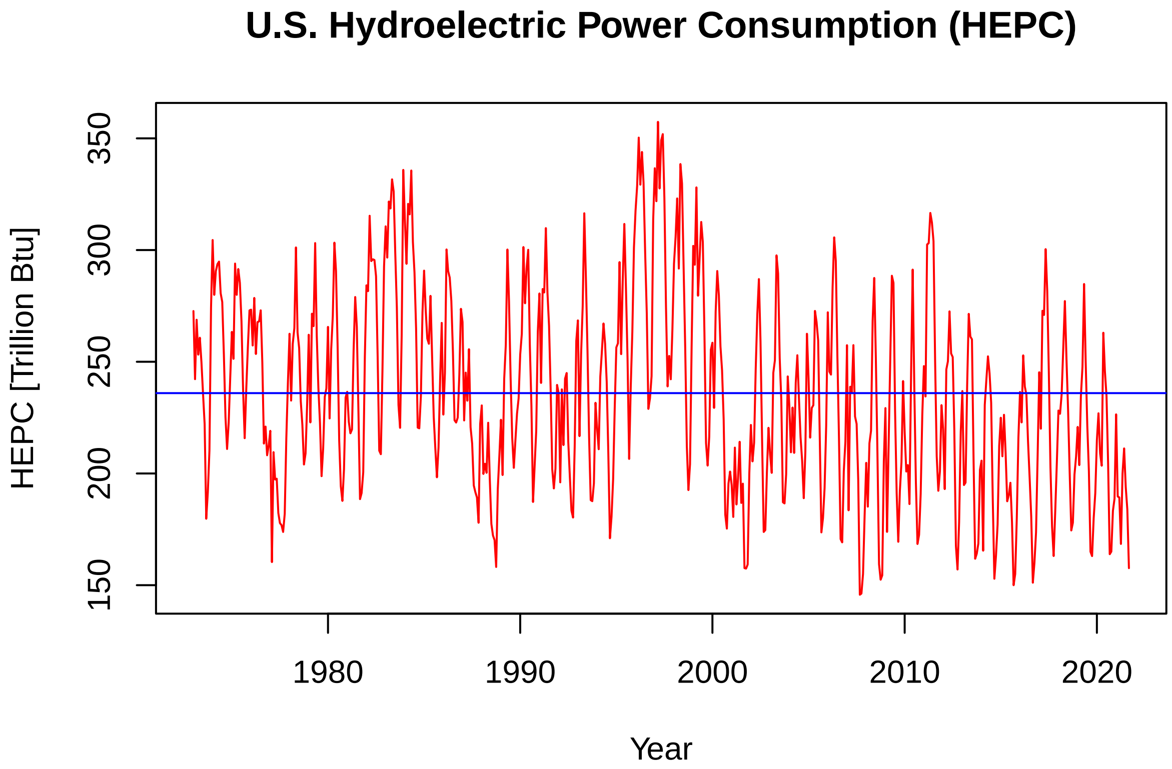
```
# Time series plot of Total Biomass Energy Production
plot(ts.biomass,type="l",col="blue",xlab="Year",ylab="BEP [Trillion Btu]",
main="U.S.Total Biomass Energy Production (BEP)")
abline(h=mean(ts.biomass),col="red")
```



```
# Time series plot of Total Renewable Energy Production
plot(ts.Re_energy,type="l",col="green",xlab="Year",ylab="REP [Trillion Btu]",
main="U.S. Total Renewable Energy Production (REP)")
abline(h=mean(ts.Re_energy),col="red")
```



```
# Time series plot of Hydroelectric Power Consumption
plot(ts.hyd_usage,type="l",col="red",xlab="Year",ylab="HEPC [Trillion Btu]",
main="U.S. Hydroelectric Power Consumption (HEPC)")
abline(h=mean(ts.hyd_usage),col="blue")
```



Question 5

Are they significantly correlated? Explain your answer.

```
cor(selected_var.df.data [,2:4])
```

```
##               biomass   renewable       hydro
## biomass     1.0000000  0.92328377 -0.28049970
## renewable   0.9232838  1.00000000 -0.05680651
## hydro      -0.2804997 -0.05680651  1.00000000
```

Correlation table of the three variables. Another option is shown below, separately doing the correlation between pairs of variables.

```
cor(selected_var.df.data$biomass, selected_var.df.data$renewable)
```

```
## [1] 0.9232838
```

The total biomass energy production and the total renewable energy production are highly significantly correlated with a correlation coefficient of 0.92.

```
cor(selected_var.df.data$biomass, selected_var.df.data$hydro)
```

```
## [1] -0.2804997
```

The total biomass energy production and the hydroelectric power consumption are slightly negatively correlated with a correlation coefficient of -0.2805. That means an increase in biomass energy production will decrease the hydroelectric power consumption by 28.05%.

```
cor(selected_var.df.data$renewable, selected_var.df.data$hydro)
```

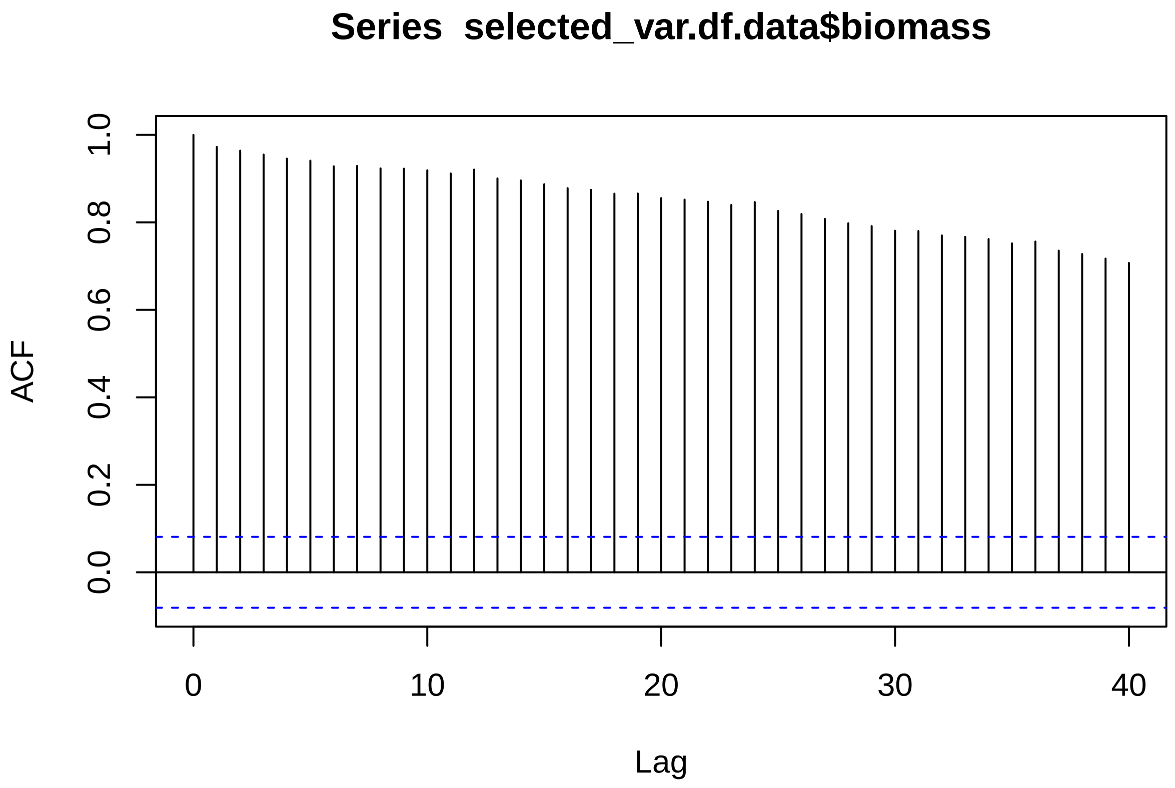
```
## [1] -0.05680651
```

The total renewable energy production and hydroelectric power consumption show a very small negative or almost no correlation with a correlation coefficient of -0.057.

Question 6

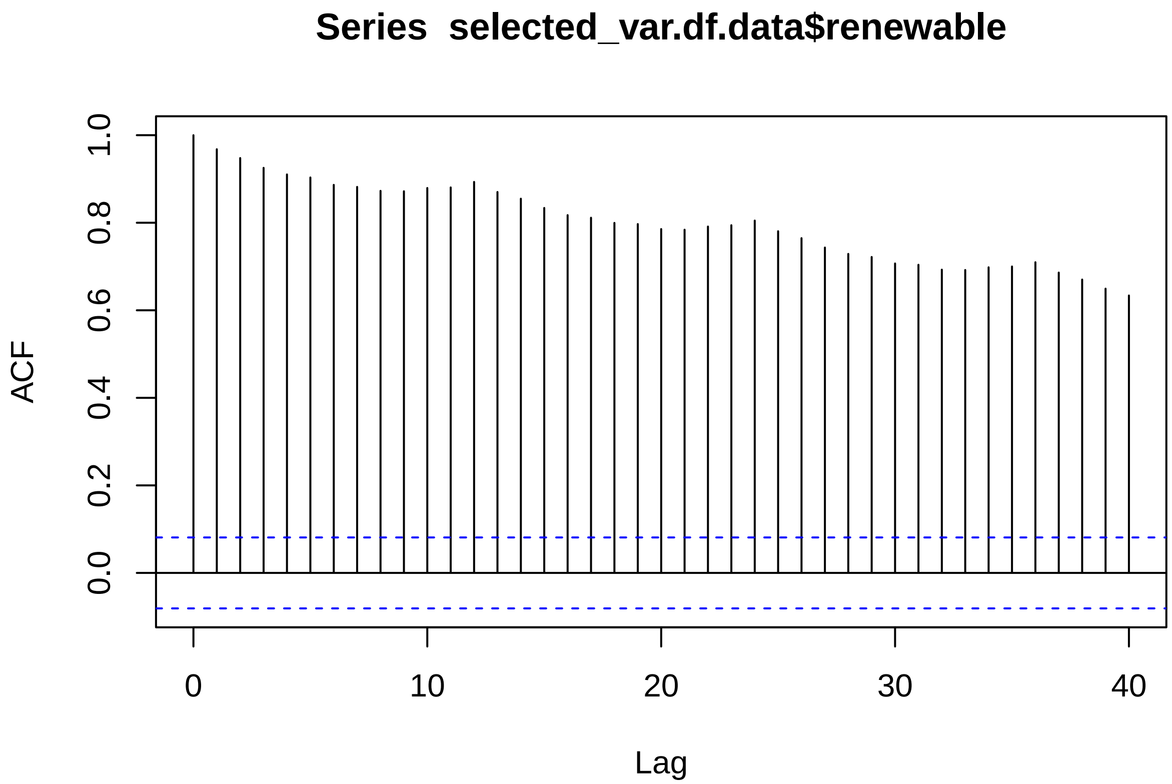
Compute the autocorrelation function from lag 1 up to lag 40 for these three variables. What can you say about these plots? Do the three of them have the same behavior?

```
acf_ts.biomass <- acf(selected_var.df.data$biomass, lag.max = 40, plot=TRUE)
```



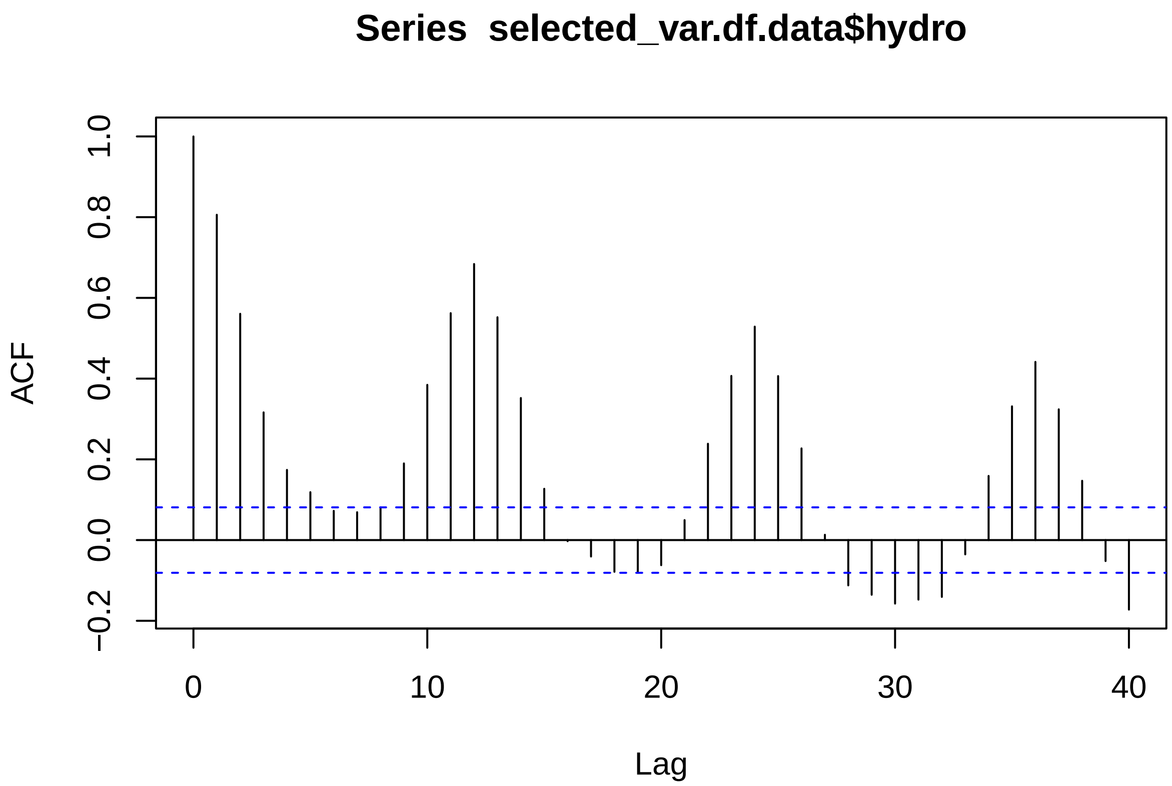
There is a significant correlation among the observations of the total biomass energy production at different time series. However, this correlation show a decreasing trend as the time lag increases.

```
acf_ts.Re_energy <- acf(selected_var.df.data$renewable, lag.max = 40, plot=TRUE)
```



Overall the correlation between the observations of the total renewable energy production at different time series is significant but show a decreasing trend as the time lag increase.

```
acf_ts.hyd_usage <- acf(selected_var.df.data$hydro, lag.max = 40, plot=TRUE)
```



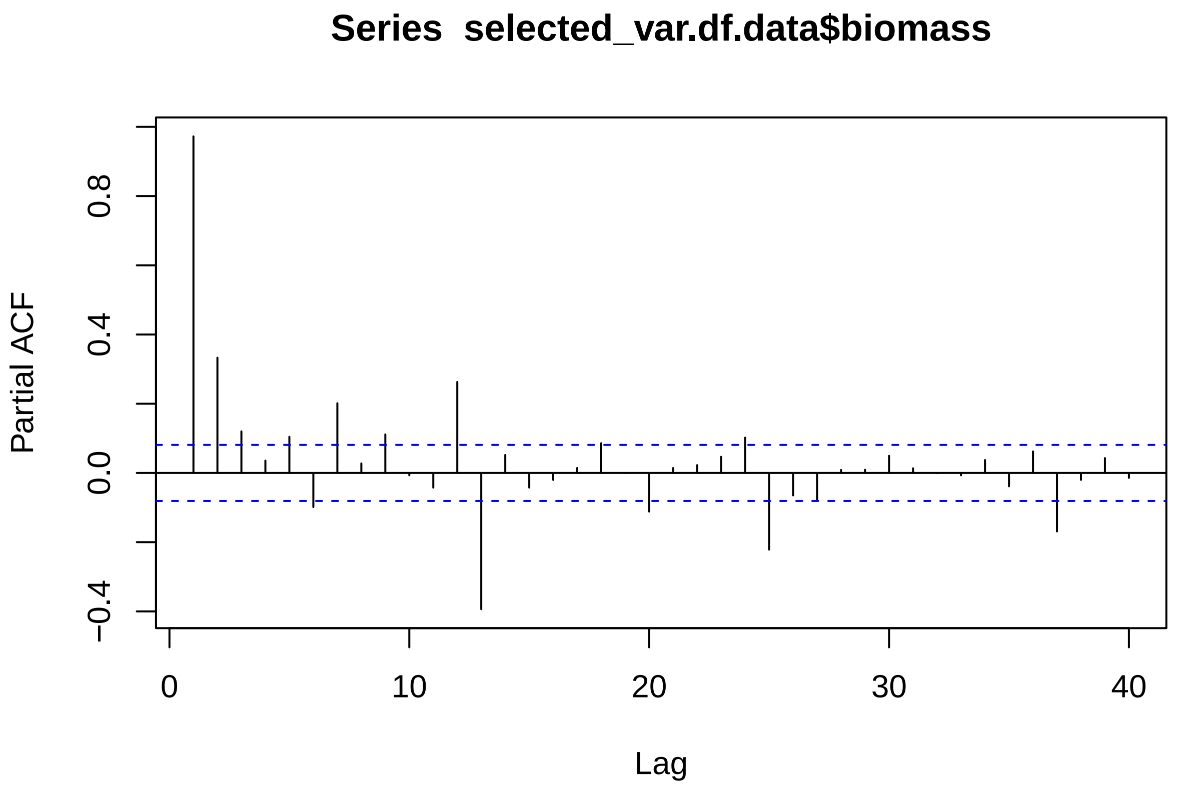
Overall the correlation between the observations of the the hydroelectric power consumption show a noticeable decreasing trend for the first 10 lag and then starts to increase following a kind of normal distribution (slightly increasing, reach at peak and then decline again slightly) for the times the observations are available.

The first two plots (the total biomass energy production and the total renewable energy production) nearly have the same trend with slightly different behaviours. However, the third plot (the hydroelectric power plot) has completely different behaviour from the other two plots.

Question 7

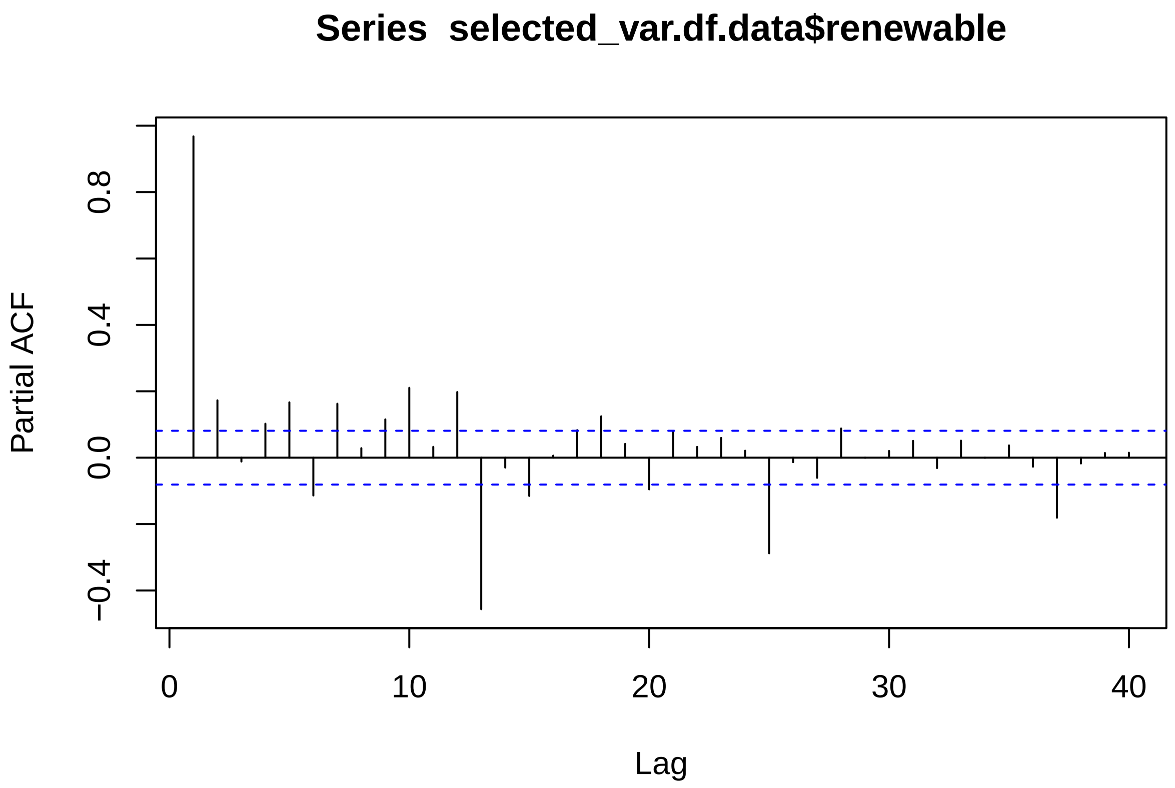
Compute the partial autocorrelation function from lag 1 to lag 40 for these three variables. How these plots differ from the ones in Q6?

```
pacf_ts.biomass <- pacf(selected_var.df.data$biomass, lag.max = 40, plot=TRUE)
```



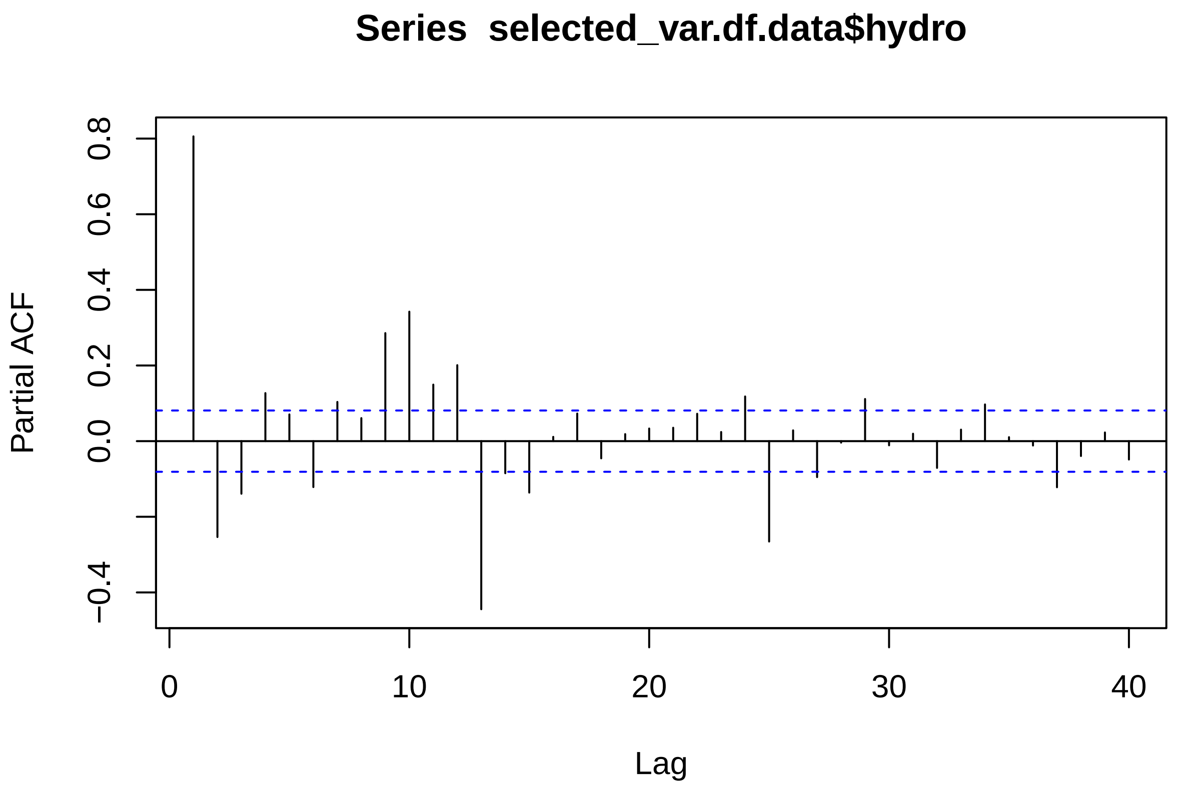
Looking at partial autocorrelation of the observations of the total biomass energy energy production from lag 1 to lag 40, there is significant relationship among some observations until lag 11 and the relationship continue to fade upon moving to lag 40.

```
pacf_ts.Re.energy <- pacf(selected_var.df.data$renewable, lag.max = 40, plot=TRUE)
```



Looking at the partial aoutcorrelation of the total renewable energy production observations from lag 1 to lag 40, there is significant relationship among some observations until lag 11 and the relationship continue to fade upon moving to lag 40.

```
pacf_ts.hyd_usage <- pacf(selected_var.df.data$hydro, lag.max = 40, plot=TRUE)
```



Looking at the partial aoutcorrelation of the hydroelectric power consumption observations from lag 1 to lag 40, there is significant relationship among some observations around lag 9 and 10 only.

The partial autocorrelation plots differ from the one under question number 6 (autocorrelation plots), the partial autocorrelation plots use the first and 40th lags by ommitting and striping out any infulence between the two data points.