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

Assignment 2 - Due date 02/25/24

Ina Liao

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.\

```
#install.packages("lubridate")
#install.packages("ggplot2")
#install.packages("forecast")
#install.packages("here")
#install.packages("patchwork")
#install.packages("tidyr")
#install.packages("knitr")
#install.packages("kableExtra")
library(lubridate)
library(ggplot2)
library(forecast) #added for Acf and Pacf functions
library(here)
library(patchwork)
library(tidyr)
library(knitr)
library(kableExtra)
```

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*.

```
#check working directory
here()

#import data
raw_energy<-read.csv(here("Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.csv"),s
raw_energy

#transform date format
Date<-ym(raw_energy$Month)
raw_energy<-cbind(Date,raw_energy[,2:14])
raw_energy
```

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.

```
df_energy<-raw_energy[,c(1,4,5,6)]
head(df_energy)
```

```
##         Date 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
```

```
#check if there is any missing data
missing_data<-any(is.na(df_energy))
missing_data #there is no missing data in the dataframe
```

```
## [1] FALSE
```

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().

```
#check the column names
colnames(df_energy)
```

```
## [1] "Date"                                "Total.Biomass.Energy.Production"
## [3] "Total.Renewable.Energy.Production" "Hydroelectric.Power.Consumption"
```

Table 1: Summary Statistics Table

	Biomass	Renewable	Hydroelectric
Min.	114.9420	185.3000	49.02200
1st Qu.	222.7380	309.9160	69.01700
Median	254.1820	346.5110	78.99300
Mean	279.8046	395.7213	79.73071
3rd Qu.	373.4810	499.5600	89.39700
Max.	469.3600	742.7530	119.39700

```
#create time series objects
ts_biomass<-ts(df_energy$Total.Biomass.Energy.Production,start=c(1973,1),frequency=12)
ts_renewable<-ts(df_energy$Total.Renewable.Energy.Production,start=c(1973,1),frequency=12)
ts_hydro<-ts(df_energy$Hydroelectric.Power.Consumption,start=c(1973,1),frequency=12)
```

Question 3

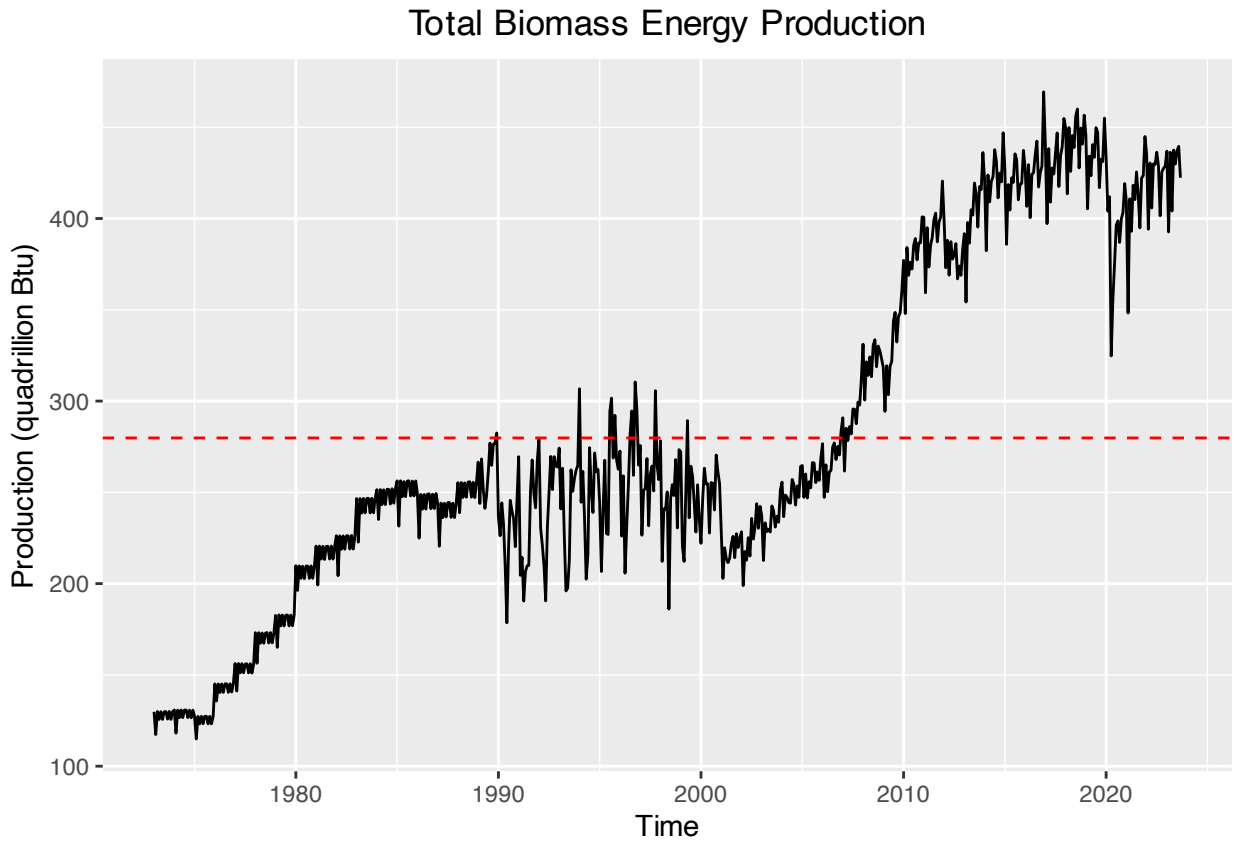
Compute mean and standard deviation for these three series.

```
combined_data<-cbind(Biomass=summary(ts_biomass),
                     Renewable=summary(ts_renewable),
                     Hydroelectric=summary(ts_hydro))
#summary table
summary_table<-kable(combined_data,caption="Summary Statistics Table") %>%
  kable_styling(bootstrap_options = "striped", full_width = FALSE)
summary_table
```

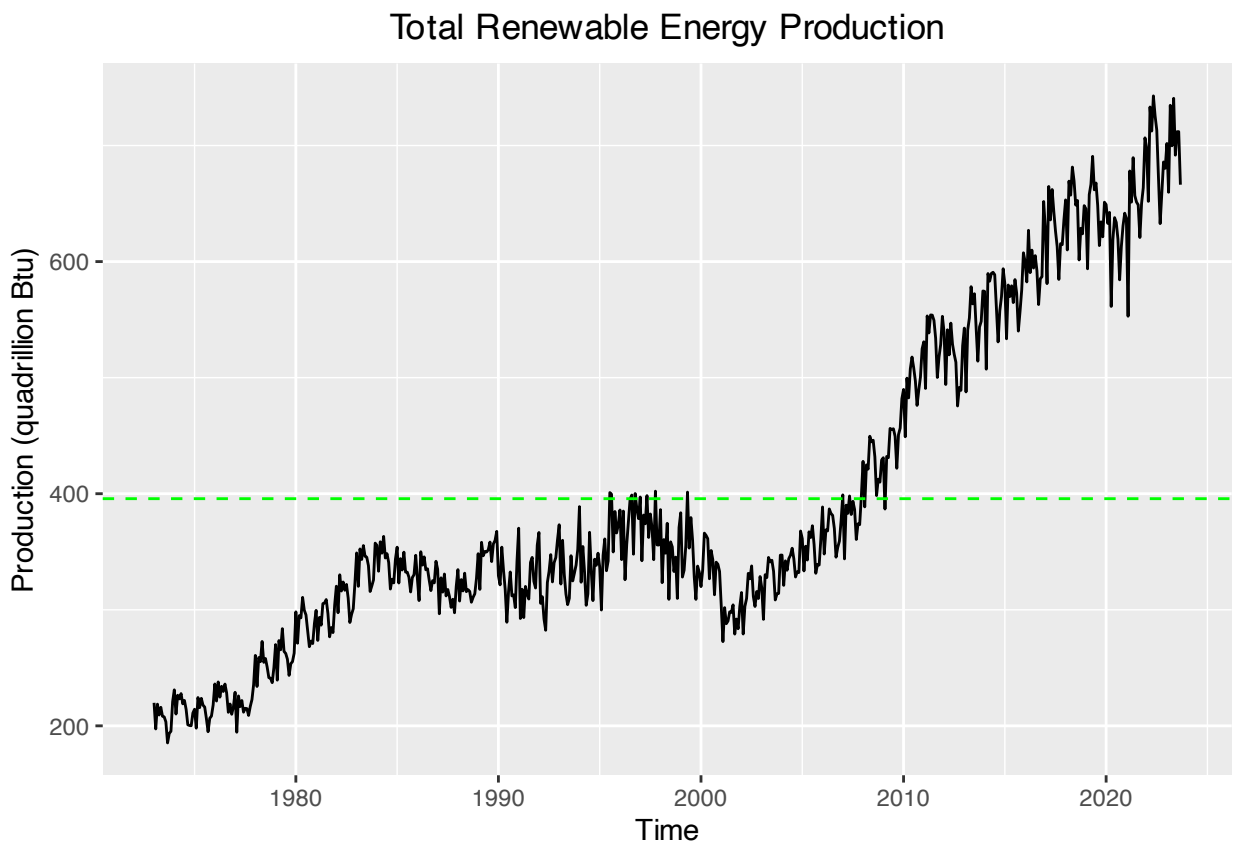
Question 4

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_biomass<-autoplot(ts_biomass) +
  labs(title="Total Biomass Energy Production",x="Time",y="Production (quadrillion Btu)")+
  theme(plot.title=element_text(hjust = 0.5))+
  geom_hline(yintercept= mean(ts_biomass), linetype="dashed", color = "red")
plot_biomass
```



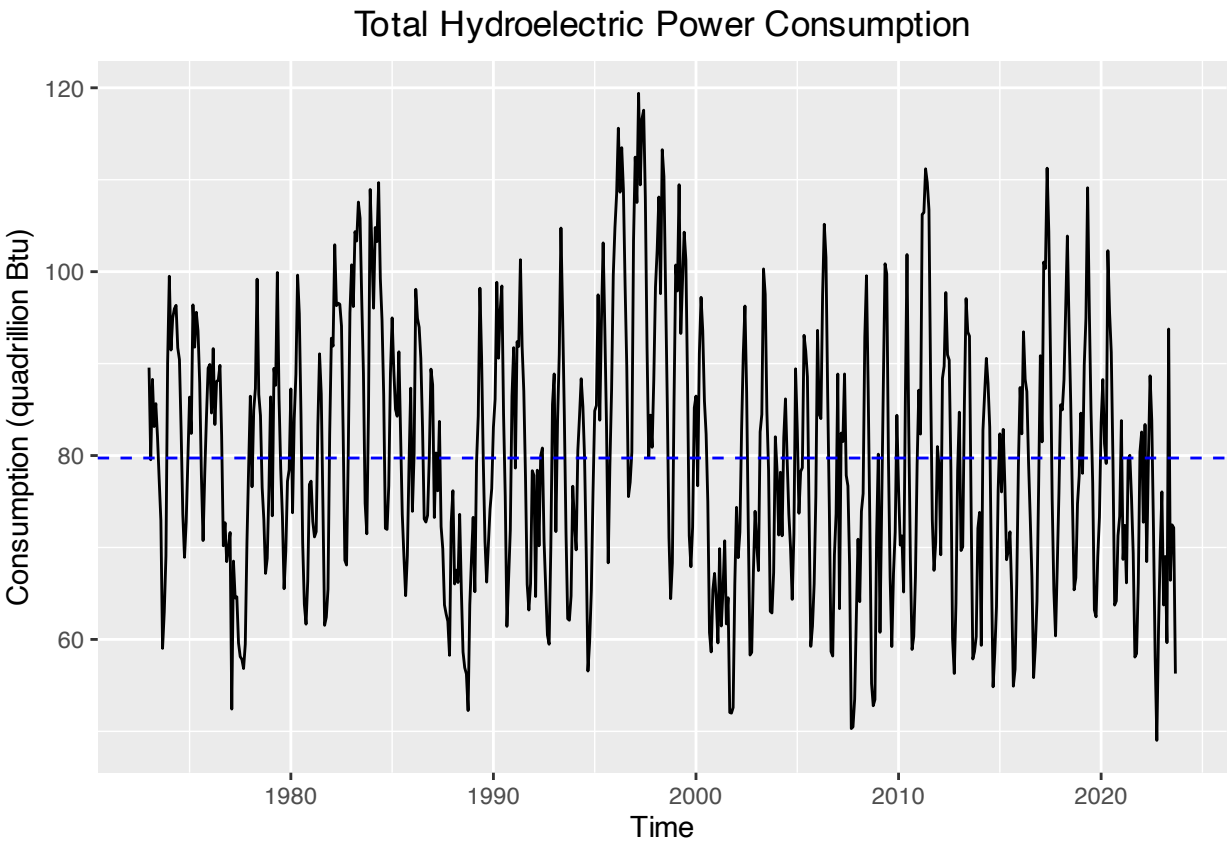

```
plot_renewable<-autoplot(ts_renewable) +
  labs(title="Total Renewable Energy Production",x="Time",y="Production (quadrillion Btu)")+
  theme(plot.title=element_text(hjust = 0.5))+
  geom_hline(yintercept= mean(ts_renewable), linetype="dashed", color = "green")
plot_renewable
```




```
plot_hydro<-autoplot(ts_hydro) +
  labs(title="Total Hydroelectric Power Consumption",x="Time",y="Consumption (quadrillion Btu)")+
  theme(plot.title=element_text(hjust = 0.5))+
  geom_hline(yintercept= mean(ts_hydro), linetype="dashed", color = "blue")
plot_hydro
```

Table 2: Correlation between the Three Time Seres

Renewable_Biomass	Renewable_Hydro	Biomass_Hydro
0.8475	0.0496	-0.0647



Question 5

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

```
cor_renew_bio<-cor(ts_renewable,ts_biomass,method = "kendall")
cor_renew_hydro<-cor(ts_renewable,ts_hydro,method = "kendall")
cor_bio_hydro<-cor(ts_biomass,ts_hydro,method = "kendall")


correlation<-cbind(Renewable_Biomass=round(cor_renew_bio,4),
                   Renewable_Hydro=round(cor_renew_hydro,4),
                   Biomass_Hydro=round(cor_bio_hydro,4))

cor_table<-kable(correlation,caption="Correlation between the Three Time Seres")%>%
  kable_styling(bootstrap_options = "striped", full_width = FALSE)
cor_table
```

Assuming that there are no linear relationships between the variables, I used Kendall's Rank Correlation,which is more robust to outliers compared to Spearman's. The results of the test indicate a strong positive correlation (tau=0.875) between renewable and biomass production. However, there is no correlation between hydroelectric power consumption and renewable production (tau=-0.0647) or biomass production (tau=0.0496).

Question 6

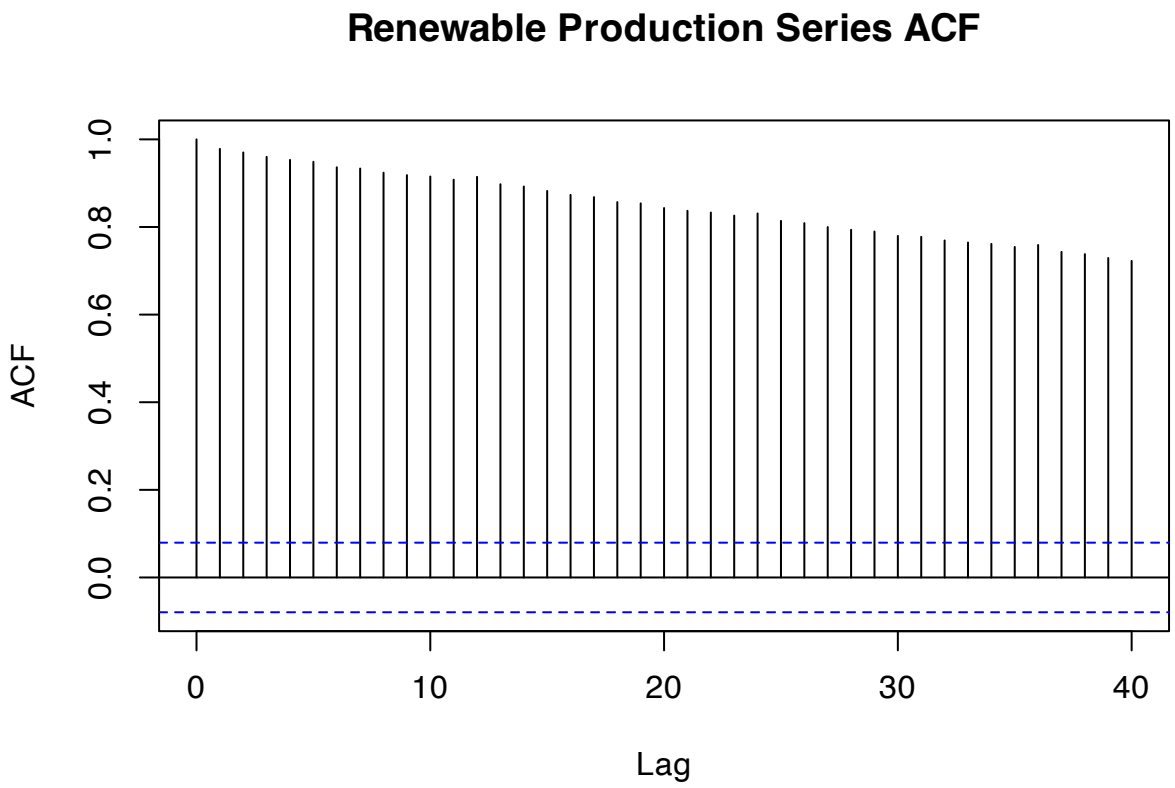
Compute the autocorrelation function from lag 1 up to lag 40 for these three variables. What can you say about these plots? Do the three of them have the same behavior? #use ACF

```
acf_renew<-Acf(ts_renewable,lag.max=40)
```

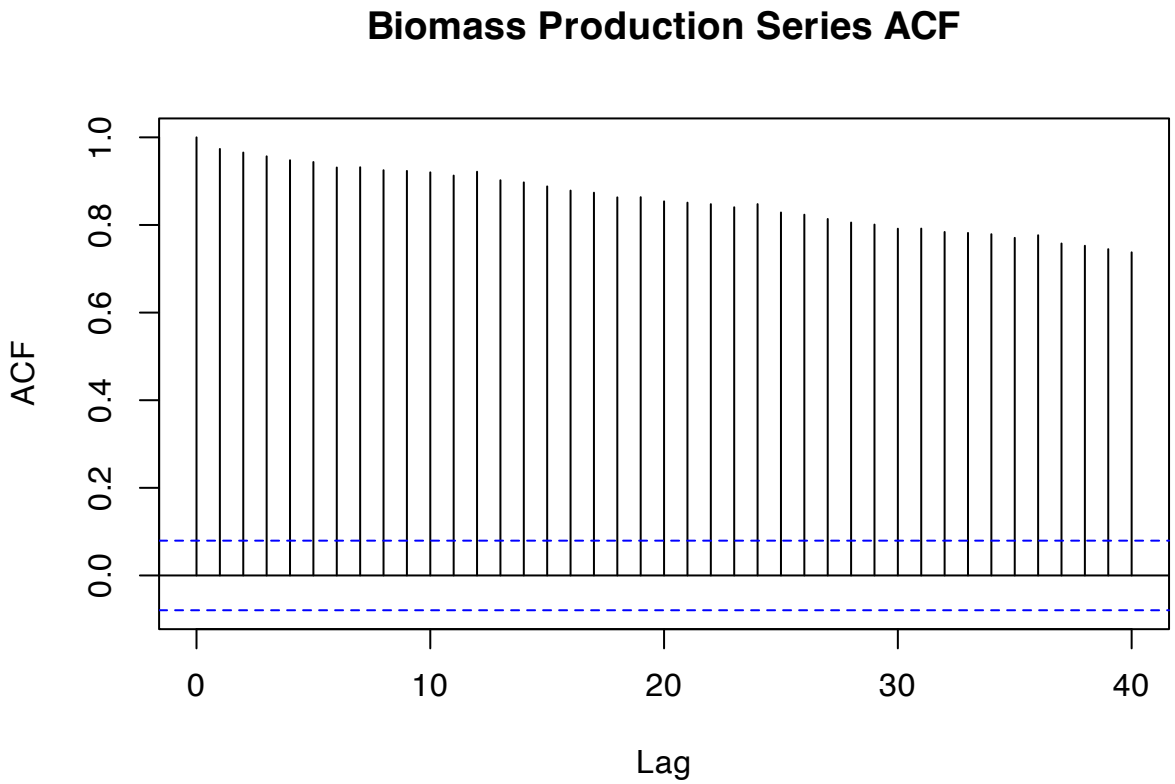
```
acf_biomass<-Acf(ts_biomass,lag.max=40)
```

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

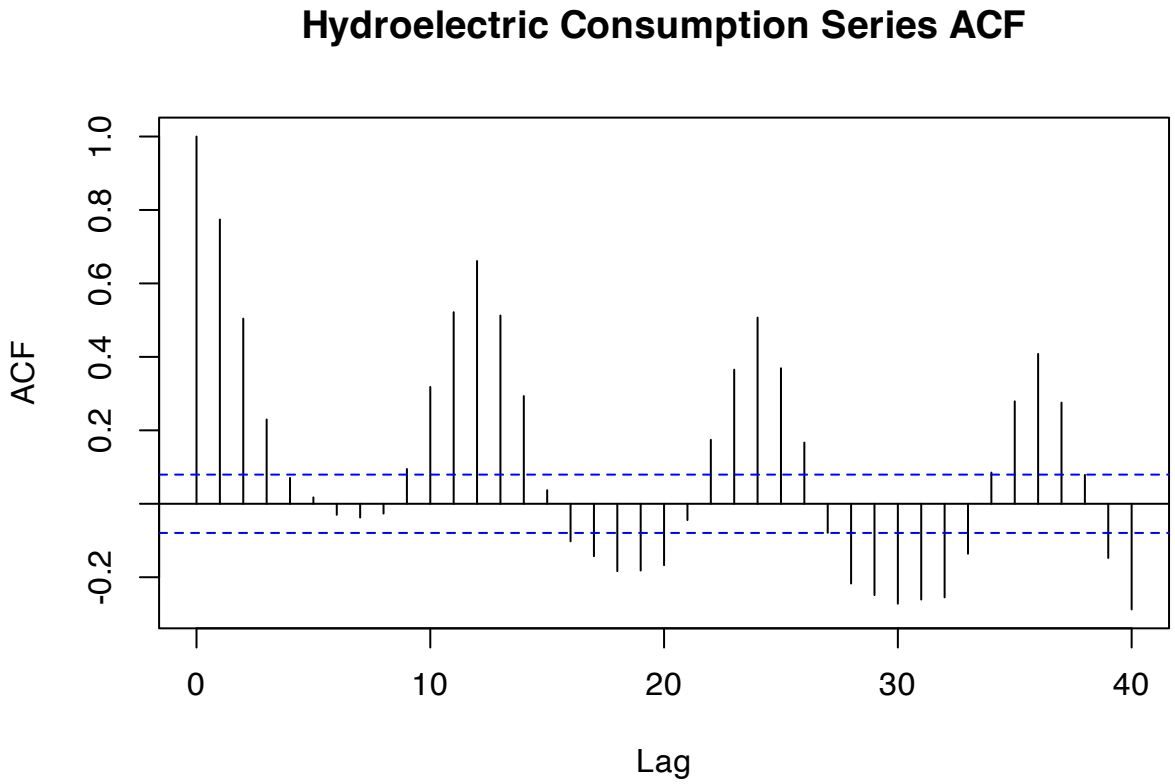
```
plot_renew_acf<-plot(acf_renew, main = "Renewable Production Series ACF")
```




```
plot_biomass_acf<-plot(acf_biomass, main = "Biomass Production Series ACF")
```




```
plot_hydro_acf<-plot(acf_hydro, main = "Hydroelectric Consumption Series ACF")
```



(1) The renewable and biomass production does not have seasonality, but their lagged data shows a strong positive correlation. The plots also show a slowly decreasing trend in the data.

(2) Hydroelectric power consumption exhibits seasonality. The consumption data shows positive autocorrelation during winter and negative autocorrelation during summer.

Question 7

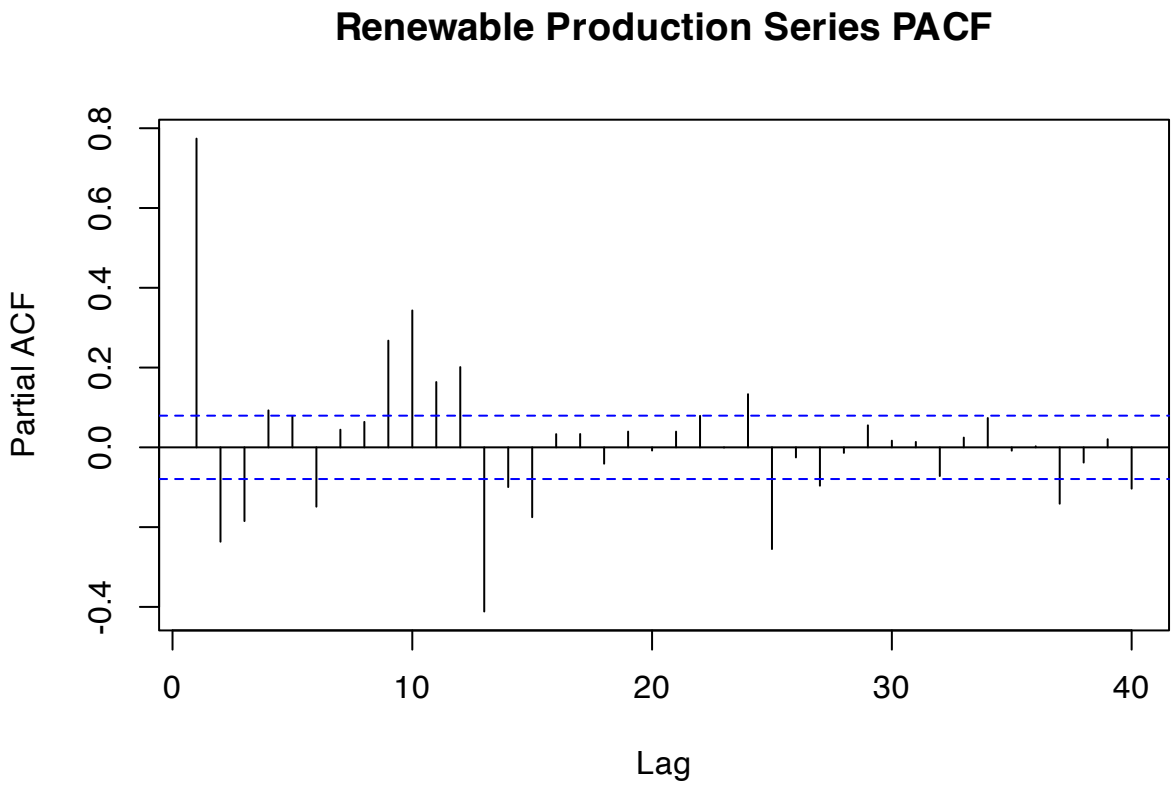
Compute the partial autocorrelation function from lag 1 to lag 40 for these three variables. How these plots differ from the ones in Q6?

```
pacf_renew<-Pacf(ts_renewable,lag.max=40)
```

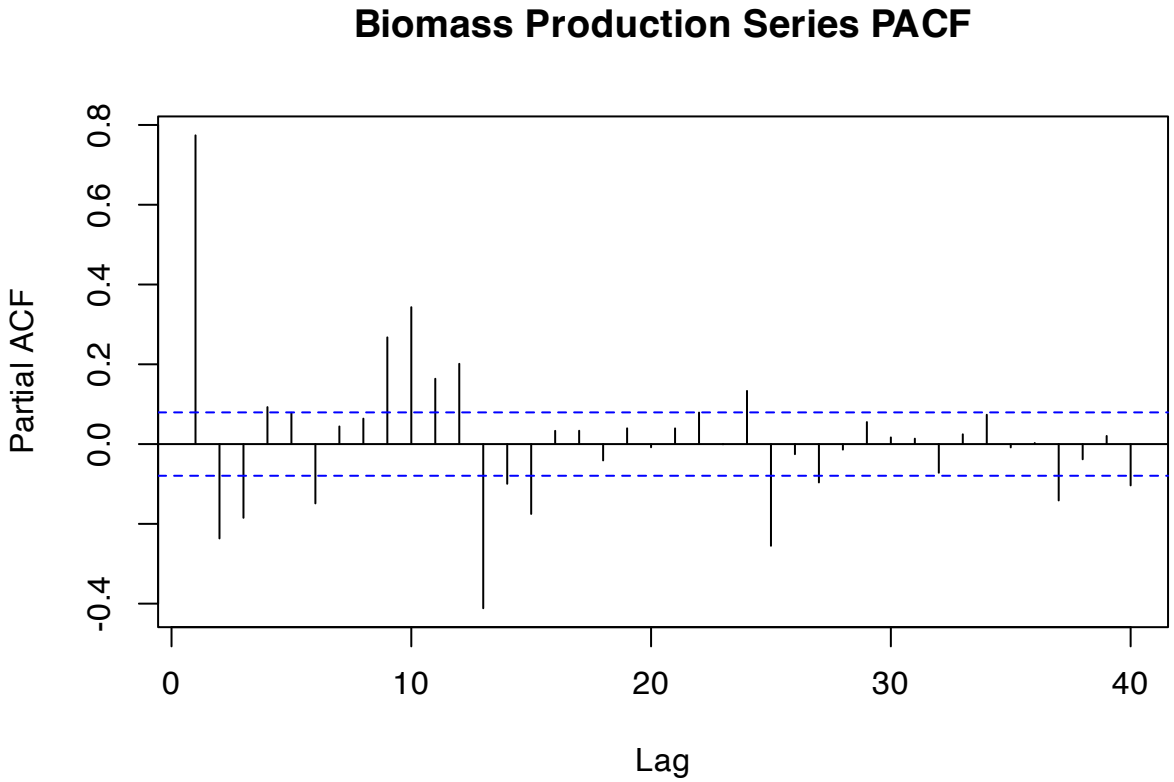
```
pacf_biomass<-Pacf(ts_biomass,lag.max=40)
```

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

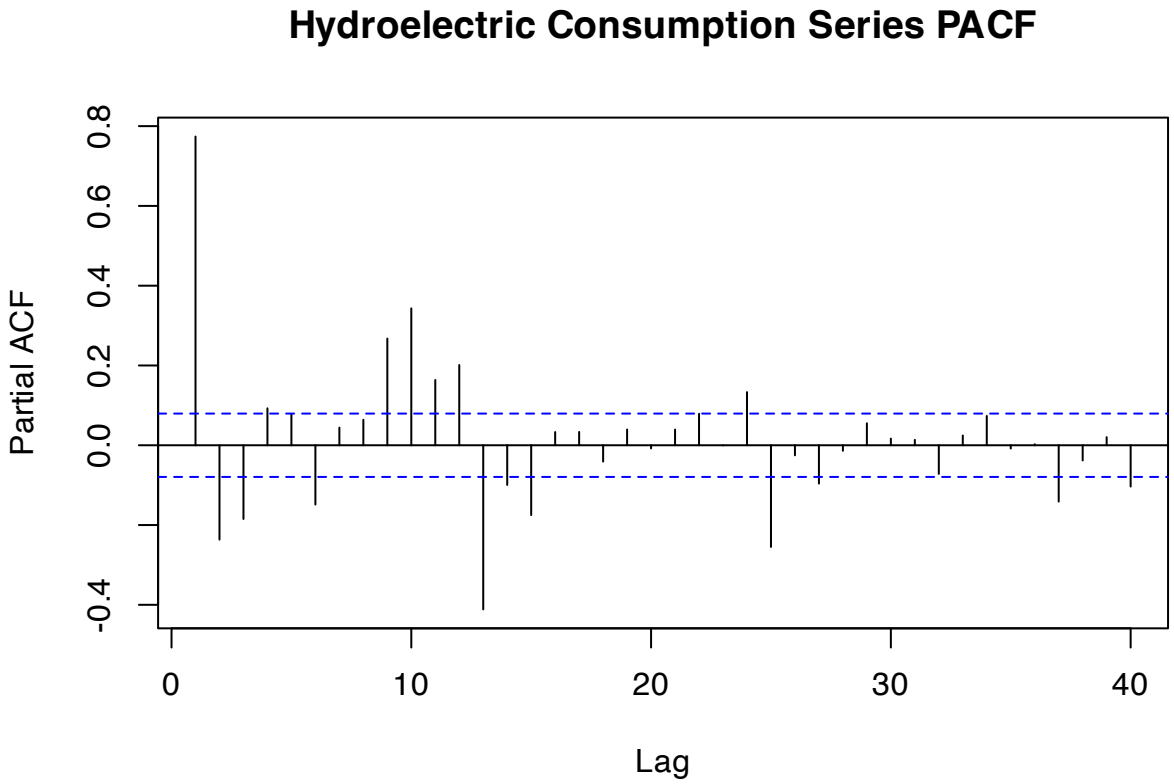
```
plot_renew_pacf<-plot(pacf_hydro, main = "Renewable Production Series PACF")
```




```
plot_biomass_pacf<-plot(pacf_hydro, main = "Biomass Production Series PACF")
```




```
plot_hydro_pacf<-plot(pacf_hydro, main = "Hydroelectric Consumption Series PACF")
```



After removing the effect of intermediate lags, we can see that:

(1) The production of renewable and biomass do not display a strong positive correlation between the time

series. However, there is a strong cutoff at lag 13, indicating that there might be seasonal pattern on the data.

(2) The seasonality of hydroelectric power consumption is less significant in the partial ACF plot. However, similar to renewable and biomass production, there are spikes at lag 13, suggesting a possible seasonal pattern in the data. We can further build a time series model that includes lags 1 and 13 to better understand the trend and seasonality in the data.