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

Assignment 2 - Due date 02/05/21

Chao Ouyang

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

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. 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 data set
energy_data <- as.data.frame(read_excel(path = "../Data/Table_10.1_Renewable_Energy_Production_and_Cons
                                       skip = 10, sheet = 1, col_names = TRUE))
energy_data <- energy_data[-1, ]
rownames(energy_data) <- NULL
# head(energy_data, 10)
```

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_data_sub <- energy_data[, c(1, 4, 5, 6)]
str(energy_data_sub)
```

```
## 'data.frame':    574 obs. of  4 variables:
##  $ Month                             : POSIXct, format: "1973-01-01" "1973-02-01" ...
##  $ Total Biomass Energy Production  : chr  "129.787" "117.338" "129.938" "125.636" ...
##  $ Total Renewable Energy Production: chr  "403.981" "360.9" "400.161" "380.47" ...
##  $ Hydroelectric Power Consumption  : chr  "272.703" "242.199" "268.81" "253.185" ...
```

```
energy_data_sub[, 2:4] <- sapply(energy_data_sub[, 2:4], as.numeric)
str(energy_data_sub)
```

```
## 'data.frame':    574 obs. of  4 variables:
##  $ Month                             : POSIXct, format: "1973-01-01" "1973-02-01" ...
##  $ Total Biomass Energy Production  : num  130 117 130 126 130 ...
##  $ Total Renewable Energy Production: num  404 361 400 380 392 ...
##  $ Hydroelectric Power Consumption  : num  273 242 269 253 261 ...
```

```
head(energy_data_sub, 10)
```

```
##         Month Total Biomass Energy Production Total Renewable Energy Production
## 1  1973-01-01                         129.787                           403.981
## 2  1973-02-01                         117.338                           360.900
## 3  1973-03-01                         129.938                           400.161
## 4  1973-04-01                         125.636                           380.470
## 5  1973-05-01                         129.834                           392.141
## 6  1973-06-01                         125.611                           377.232
## 7  1973-07-01                         129.787                           367.325
## 8  1973-08-01                         129.918                           353.757
## 9  1973-09-01                         125.782                           307.006
## 10 1973-10-01                         129.970                           323.453
##    Hydroelectric Power Consumption
## 1                          272.703
## 2                          242.199
## 3                          268.810
## 4                          253.185
## 5                          260.770
## 6                          249.859
## 7                          235.670
## 8                          222.077
## 9                          179.733
## 10                         191.723
```

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_data_sub$Month[1]
```

```
## [1] "1973-01-01 UTC"
```

```
energy_data_sub$Month[nrow(energy_data_sub)]
```

```
## [1] "2020-10-01 UTC"
```

```r
ts_data <- ts(energy_data_sub[, 2:4], start = c(1973, 1), end = c(2020, 10), frequency = 12)
head(ts_data, 10)
```

```
##          Total Biomass Energy Production Total Renewable Energy Production
## Jan 1973                         129.787                           403.981
## Feb 1973                         117.338                           360.900
## Mar 1973                         129.938                           400.161
## Apr 1973                         125.636                           380.470
## May 1973                         129.834                           392.141
## Jun 1973                         125.611                           377.232
## Jul 1973                         129.787                           367.325
## Aug 1973                         129.918                           353.757
## Sep 1973                         125.782                           307.006
## Oct 1973                         129.970                           323.453
##          Hydroelectric Power Consumption
## Jan 1973                         272.703
## Feb 1973                         242.199
## Mar 1973                         268.810
## Apr 1973                         253.185
## May 1973                         260.770
## Jun 1973                         249.859
## Jul 1973                         235.670
## Aug 1973                         222.077
## Sep 1973                         179.733
## Oct 1973                         191.723
```

Question 3

Compute mean and standard deviation for these three series.

```r
# means of the three time series
ts_means <- colMeans(ts_data)
ts_means
```

```
##   Total Biomass Energy Production Total Renewable Energy Production
##                          270.6961                          572.7321
##   Hydroelectric Power Consumption
##                          236.9515
```

```r
# sds of the three time series
ts_sds <- sapply(ts_data, sd)
ts_sds
```

```
##   Total Biomass Energy Production Total Renewable Energy Production
##                          87.36311                         168.45877
##   Hydroelectric Power Consumption
##                          43.90392
```

Question 4

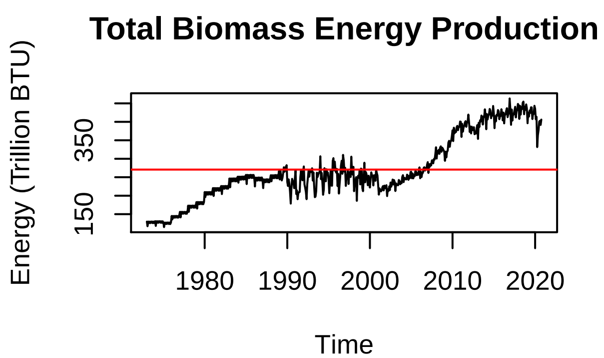
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

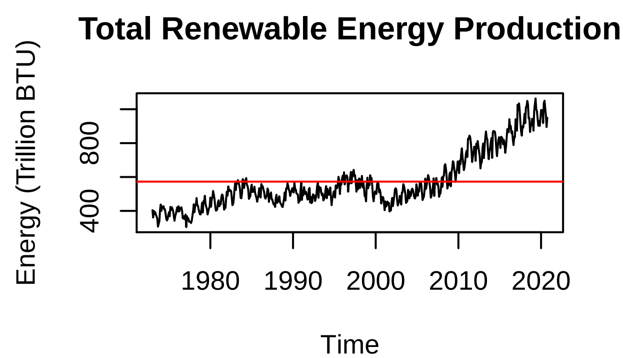
```r
par(mfrow=c(2,2))
plot(ts_data[, 1], type = "l", ylab = "Energy (Trillion BTU)",
```

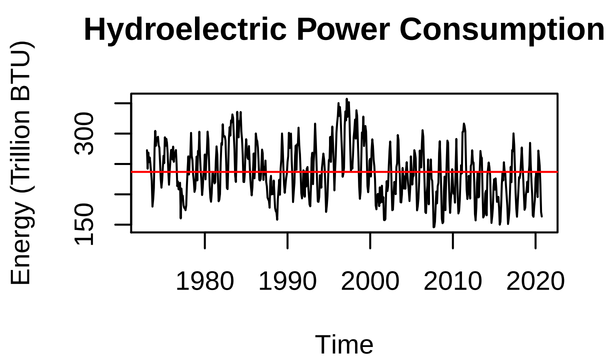
```
     main = "Total Biomass Energy Production")
abline(h = ts_means[1], col = "red")

plot(ts_data[, 2], type = "l", ylab = "Energy (Trillion BTU)",
     main = "Total Renewable Energy Production")
abline(h = ts_means[2], col = "red")

plot(ts_data[, 3], type = "l", ylab = "Energy (Trillion BTU)",
     main = "Hydroelectric Power Consumption")
abline(h = ts_means[3], col = "red")
```







Total biomass energy production experienced two periods of rapid increase roughly before and after the last decade of the 20th century (1990 - 2000), which recorded noticeable fluctuations between years at around 200 - 300 trillion BTU.

Total renewable energy production showed a moderate increase and some fluctuations before the 21th century, and then expanded in a rapid rate. The overall trend is similar to the total biomass energy production time series.

Hydroelectric power consumption was relatively stable over the years, despite showing significant seasonal and yearly variations.

Question 5

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

```r
cor(ts_data)
```

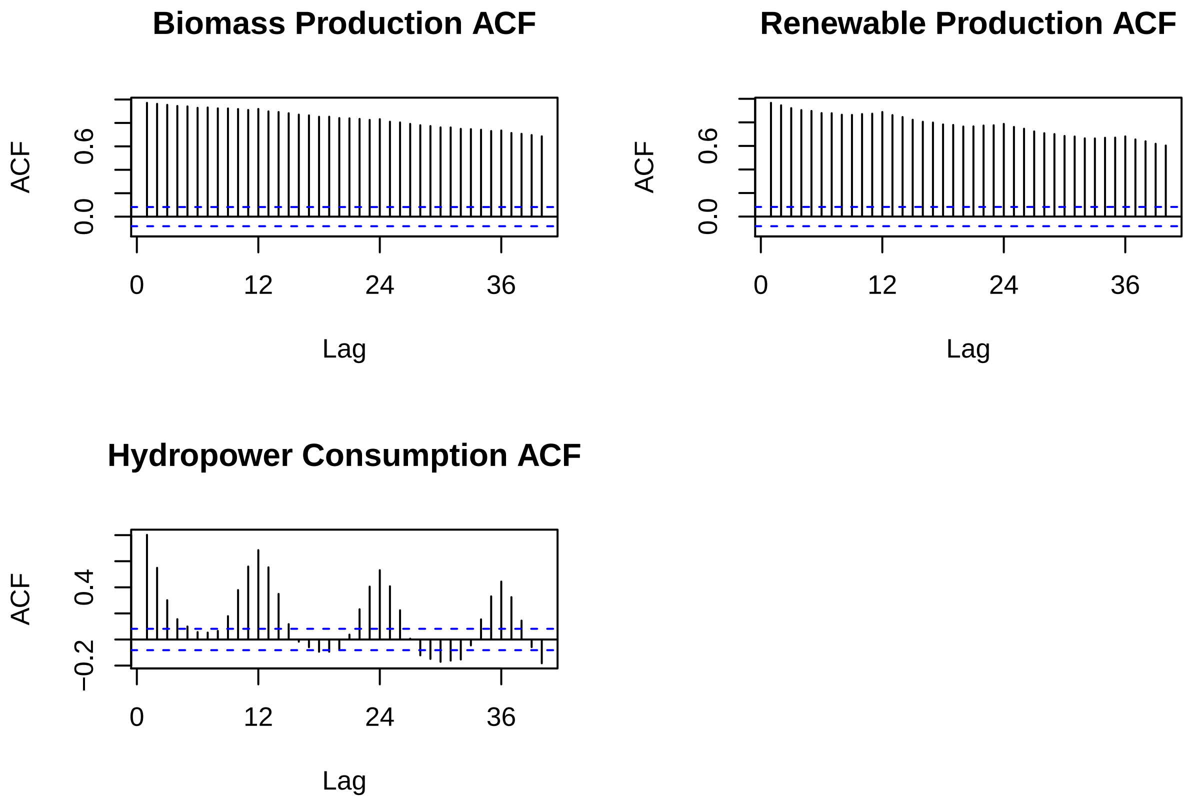
```
##                                     Total Biomass Energy Production
## Total Biomass Energy Production                           1.0000000
## Total Renewable Energy Production                         0.9234609
## Hydroelectric Power Consumption                          -0.2555675
##                                     Total Renewable Energy Production
## Total Biomass Energy Production                           0.923460855
## Total Renewable Energy Production                         1.000000000
## Hydroelectric Power Consumption                          -0.002756852
##                                     Hydroelectric Power Consumption
## Total Biomass Energy Production                        -0.255567465
## Total Renewable Energy Production                      -0.002756852
## Hydroelectric Power Consumption                         1.000000000
```

Based on the output results above, renewable energy production and biomass energy production are significantly positively correlated (0.923). In addition, hydroelectric power consumption is negatively correlated with both biomass energy production and renewable energy production, but the correlations are much weaker especially with renewable energy production (-0.003).

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
par(mfrow=c(2,2))
biomass_acf <- Acf(ts_data[, 1], lag.max = 40, type = "correlation",
                   plot = T, main = "Biomass Production ACF")
renewable_acf <- Acf(ts_data[, 2], lag.max = 40, type = "correlation",
                   plot = T, main = "Renewable Production ACF")
hydro_acf <- Acf(ts_data[, 3], lag.max = 40, type = "correlation",
                   plot = T, main = "Hydropower Consumption ACF")
```

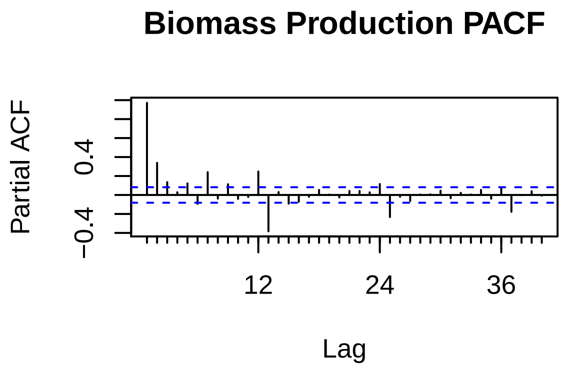


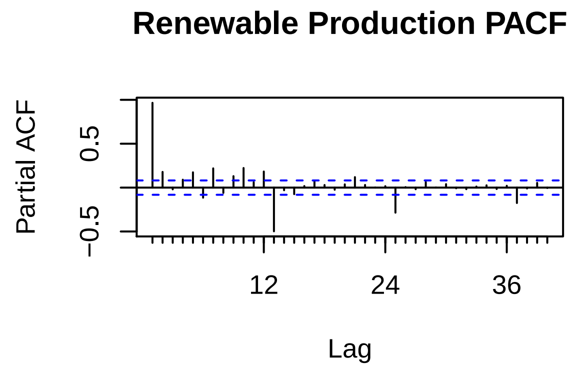
Biomass production and renewable production show similar trend based on the calculated ACF. Observations that are further apart in time are less correlated, even though there seems to be a slight seasonal pattern. On the other hand, the seasonal pattern in hydropower consumption ACF is quite noticeable, as observations that are 11 - 13 months apart seem to be more strongly correlated.

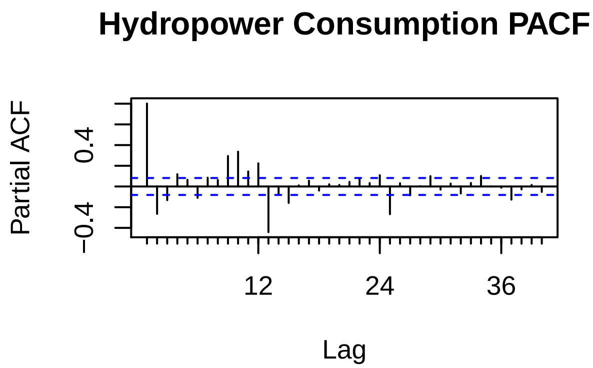
Question 7

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

```
par(mfrow=c(2,2))
biomass_pacf <- Pacf(ts_data[, 1], lag.max = 40,
                    plot = T, main = "Biomass Production PACF")
renewable_pacf <- Pacf(ts_data[, 2], lag.max = 40,
                    plot = T, main = "Renewable Production PACF")
hydro_pacf <- Pacf(ts_data[, 3], lag.max = 40,
                    plot = T, main = "Hydropower Consumption PACF")
```







Overall the PACF values are much smaller than the ACF values found in Q6, with many staying within the two blue dashed lines (statistically insignificant). Unlike in the ACF plots, we observe a more obvious seasonality in the PACF plots of biomass production and renewable production.