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

Assignment 2 - Due date 01/26/22

Kristen Pulley

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.\

```
#Load/install required package here
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 4.1.2
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(tseries)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readxl)
library(ggplot2)
```

Data set information

Consider the data provided in the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source 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
raw_consump_data<-read_excel("Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx
```

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.

```
consump_data <- raw_consump_data[,c("Month","Total Biomass Energy Production","Total Renewable Energy Pr
head(raw_consump_data)
```

```
## # A tibble: 6 x 14
##   Month               `Wood Energy Production` `Biofuels Produ~ `Total Biomass ~
##   <dttm>                                 <dbl> <chr>                        <dbl>
## 1 1973-01-01 00:00:00                     130. Not Available                 130.
## 2 1973-02-01 00:00:00                     117. Not Available                 117.
## 3 1973-03-01 00:00:00                     130. Not Available                 130.
## 4 1973-04-01 00:00:00                     125. Not Available                 126.
## 5 1973-05-01 00:00:00                     130. Not Available                 130.
## 6 1973-06-01 00:00:00                     125. Not Available                 126.
## # ... with 10 more variables: Total Renewable Energy Production <dbl>,
## #   Hydroelectric Power Consumption <dbl>, Geothermal Energy Consumption <dbl>,
## #   Solar Energy Consumption <chr>, Wind Energy Consumption <chr>,
## #   Wood Energy Consumption <dbl>, Waste Energy Consumption <dbl>,
## #   Biofuels Consumption <chr>, Total Biomass Energy Consumption <dbl>,
## #   Total Renewable Energy Consumption <dbl>
```

```
str(raw_consump_data)
```

```
## tibble [585 x 14] (S3: tbl_df/tbl/data.frame)
##  $ Month                             : POSIXct[1:585], format: "1973-01-01" "1973-02-01" ...
##  $ Wood Energy Production            : num [1:585] 130 117 130 125 130 ...
##  $ Biofuels Production               : chr [1:585] "Not Available" "Not Available" "Not Available" "
##  $ Total Biomass Energy Production   : num [1:585] 130 117 130 126 130 ...
##  $ Total Renewable Energy Production : num [1:585] 404 361 400 380 392 ...
##  $ Hydroelectric Power Consumption   : num [1:585] 273 242 269 253 261 ...
##  $ Geothermal Energy Consumption     : num [1:585] 1.49 1.36 1.41 1.65 1.54 ...
##  $ Solar Energy Consumption          : chr [1:585] "Not Available" "Not Available" "Not Available" "
```

```
##  $ Wind Energy Consumption           : chr [1:585] "Not Available" "Not Available" "Not Available" "
##  $ Wood Energy Consumption           : num [1:585] 130 117 130 125 130 ...
##  $ Waste Energy Consumption          : num [1:585] 0.157 0.144 0.176 0.174 0.21 0.176 0.17 0.184 0.1
##  $ Biofuels Consumption              : chr [1:585] "Not Available" "Not Available" "Not Available" "
##  $ Total Biomass Energy Consumption  : num [1:585] 130 117 130 126 130 ...
##  $ Total Renewable Energy Consumption: num [1:585] 404 361 400 380 392 ...
```

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_consump <- ts(consump_data[,2:4]) #note that we are only transforming columns with inflow data, not
```

Question 3

Compute mean and standard deviation for these three series.

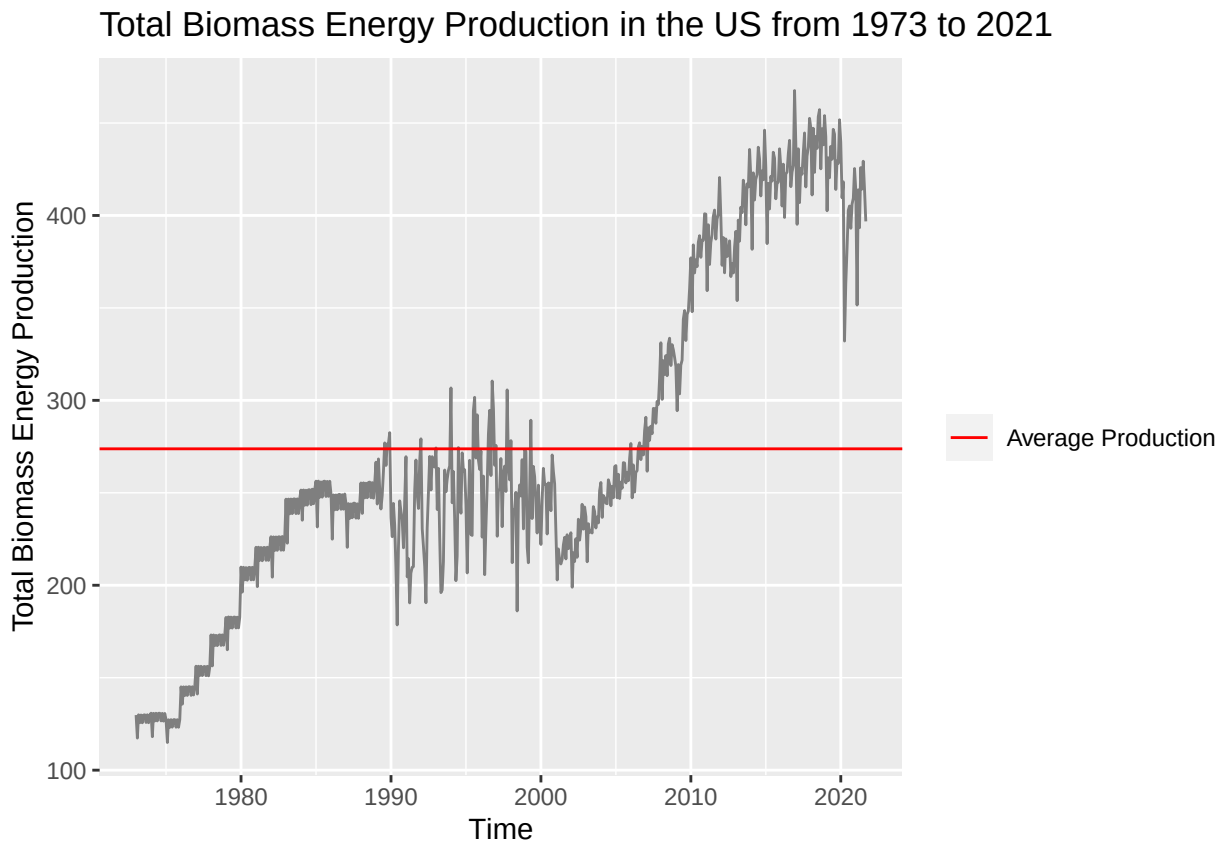
```
bio_mean<-mean(ts_consump[,1])
bio_sd<-sd(ts_consump[,1])
renew_mean<-mean(ts_consump[,2])
renew_sd<-sd(ts_consump[,2])
hydro_mean<-mean(ts_consump[,3])
hydro_sd<-sd(ts_consump[,3])
```

Question 4

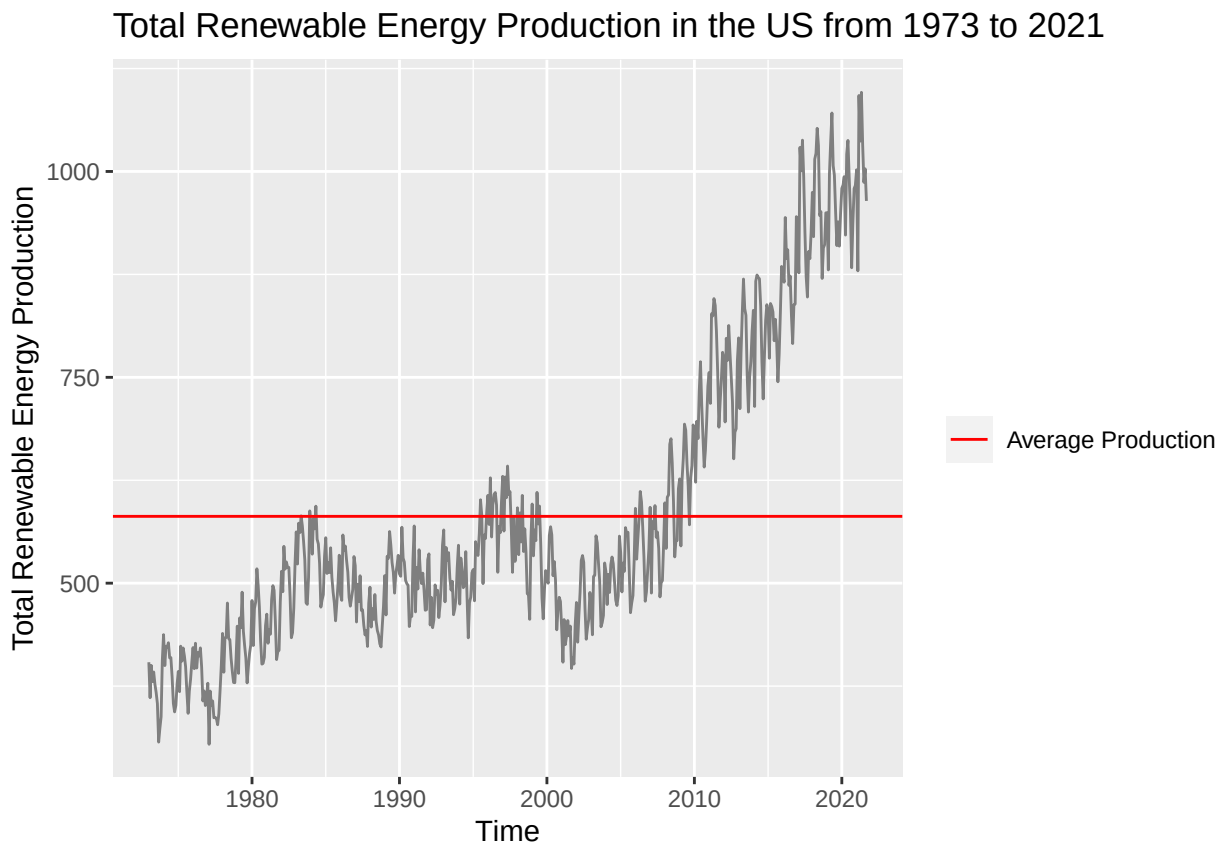
Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```
names(raw_consump_data)[4] <- 'bio'
bioplot<-ggplot(raw_consump_data, aes(x=Month,))+
  geom_line(aes(y = bio, color = "green"))+
  xlab("Time") +
  ylab("Total Biomass Energy Production")+
  geom_hline(yintercept = bio_mean,color="red")+
  scale_color_manual("",
                     breaks = c("Average Production"),
                     values = c("Average Production"="red"))+
  labs(title = "Total Biomass Energy Production in the US from 1973 to 2021")

names(raw_consump_data)[5] <- 'renew'
renewlot<-ggplot(raw_consump_data, aes(x=Month,))+
  geom_line(aes(y = renew, color = "blue"))+
  xlab("Time") +
  ylab("Total Renewable Energy Production")+
  geom_hline(yintercept = renew_mean,color="red")+
  scale_color_manual("",
                     breaks = c("Average Production"),
```

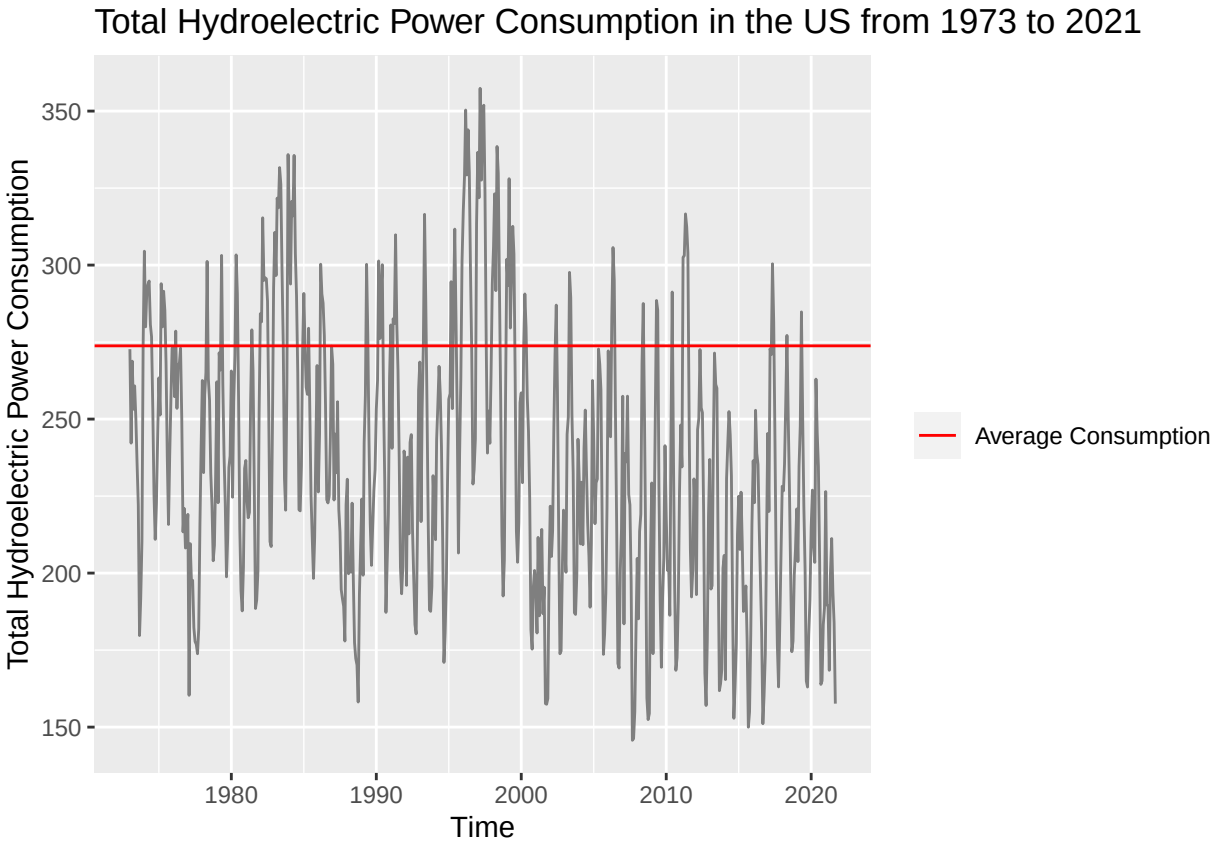
```
                      values = c("Average Production"="red"))+
  labs(title = "Total Renewable Energy Production in the US from 1973 to 2021")

names(raw_consump_data)[6] <- 'hydro'
hydroplot<-ggplot(raw_consump_data, aes(x=Month,))+
  geom_line(aes(y = hydro, color = "purple"))+
  xlab("Time") +
  ylab("Total Hydroelectric Power Consumption")+
  geom_hline(yintercept = bio_mean,color="red")+
  scale_color_manual("",
                      breaks = c("Average Consumption"),
                      values = c("Average Consumption"="red"))+
  labs(title = "Total Hydroelectric Power Consumption in the US from 1973 to 2021")
bioplot
```




```
renewlot
```


Total Renewable Energy Production in the US from 1973 to 2021
Total Renewable Energy Production
1000
750
500
1980
1990
2000
2010
2020
Time
Average Production

```
hydroplot
```



Question 5

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

Total biomass energy production is significantly positively correlated to to total renewable energy production and has an weak negative correlation to hydroelectric power consumption. Total renewable energy production has an even weaker negative correlation to hydroelectric power consumption.

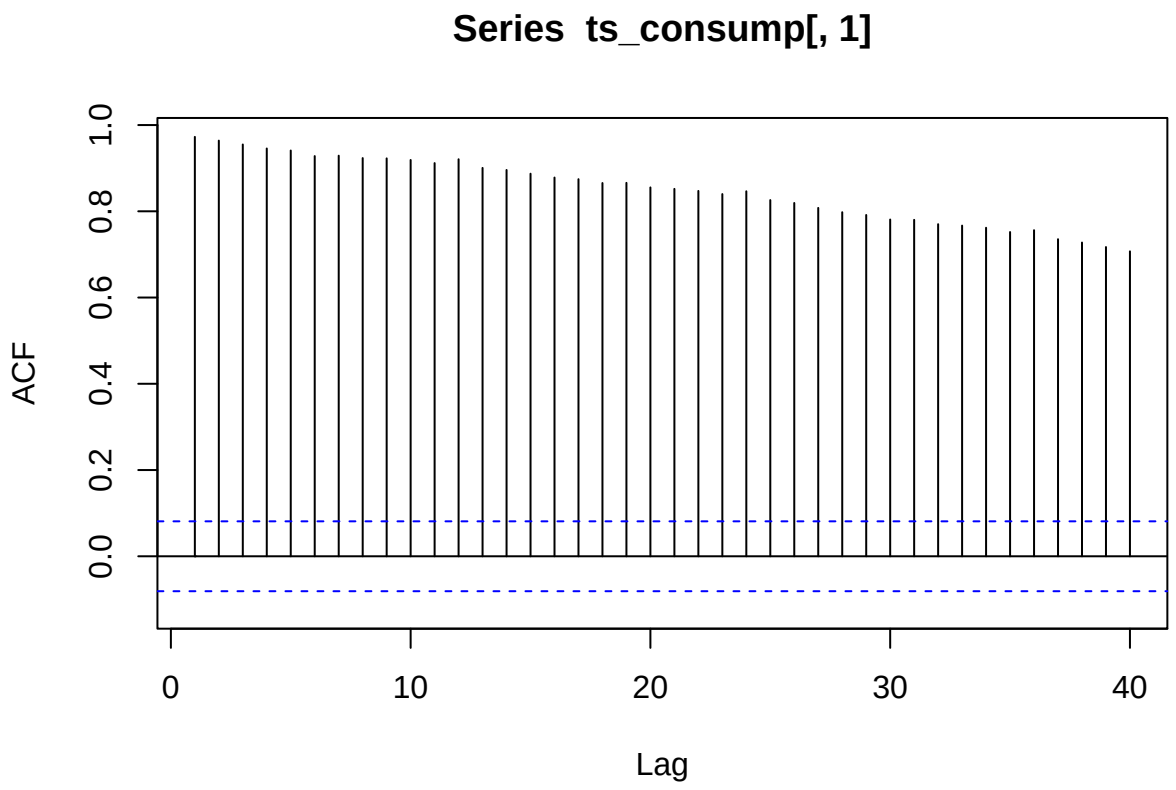
```
cor(ts_consump)
```

```
##                                  Total Biomass Energy Production
## Total Biomass Energy Production                        1.0000000
## Total Renewable Energy Production                      0.9232838
## Hydroelectric Power Consumption                       -0.2804997
##                                  Total Renewable Energy Production
## Total Biomass Energy Production                         0.92328377
## Total Renewable Energy Production                       1.00000000
## Hydroelectric Power Consumption                        -0.05680651
##                                  Hydroelectric Power Consumption
## Total Biomass Energy Production                       -0.28049970
## Total Renewable Energy Production                     -0.05680651
## Hydroelectric Power Consumption                        1.00000000
```

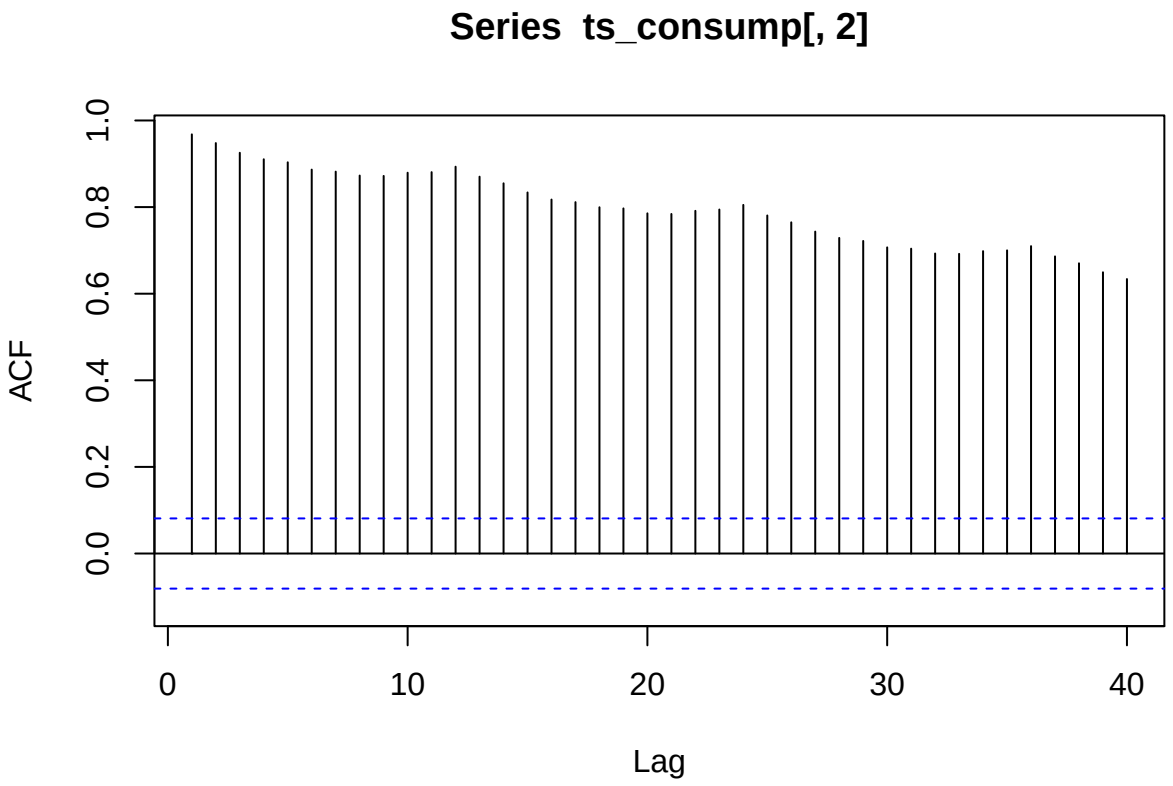
Question 6

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? Both biomass and renewable energy production show a relative constantly mean overtime (stationary), while the hydroelectric energy production shows a seasonal lag (non stationary).

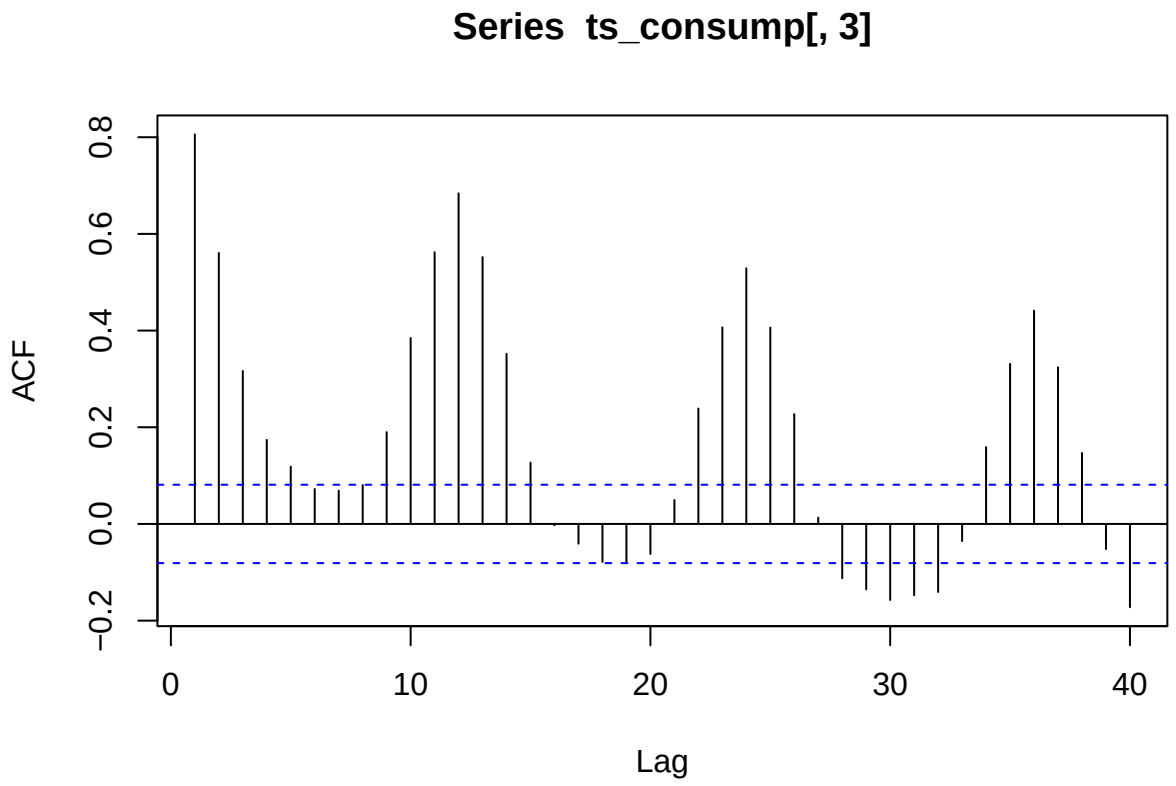
```
bio.acf<-Acf(ts_consump[,1], lag=40)
```




```
renw.acf<-Acf(ts_consump[,2], lag=40)
```




```
hydro.acf<-Acf(ts_consump[,3], lag=40)
```

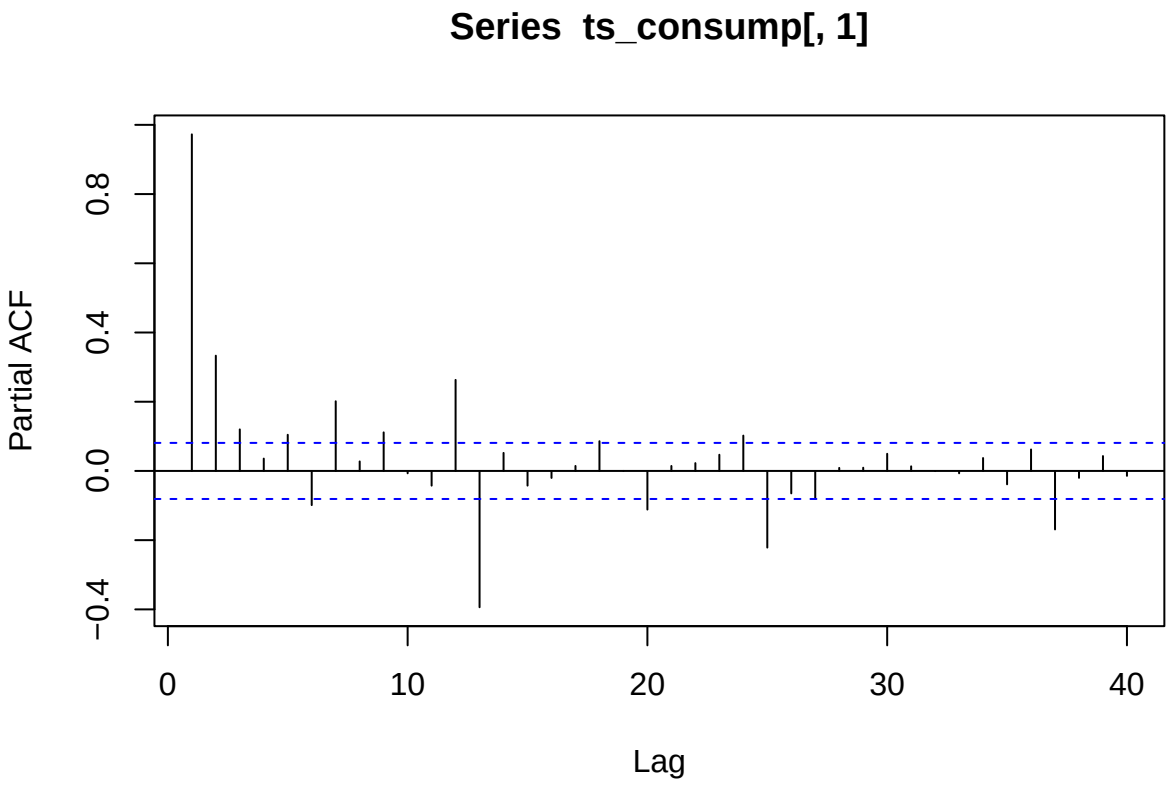


Question 7

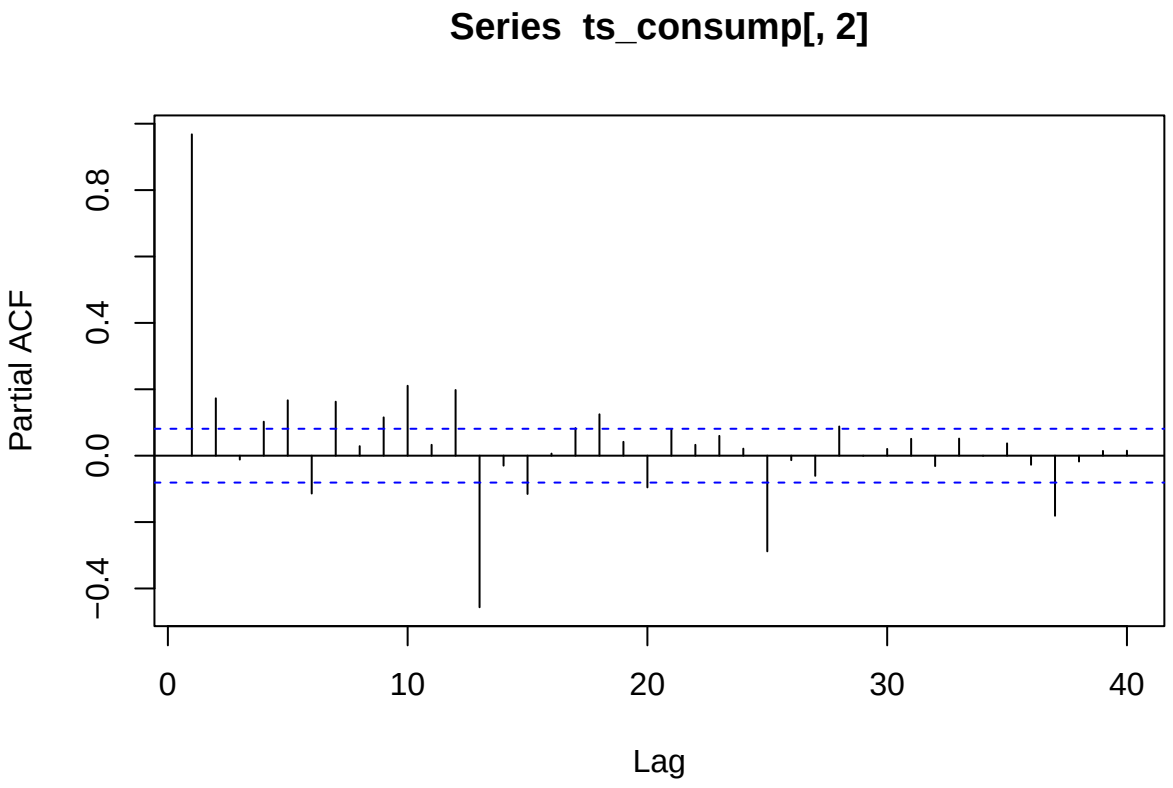
Compute the partial autocorrelation function from lag 1 to lag 40 for these three variables. How these plots differ from the ones in Q6?

After removing the influence of intermediate variables, all three plots show a slow decrease into a seasonal trend. This is different than the plots in Q6 because biomass and renewable production did not show the seasonal trend.

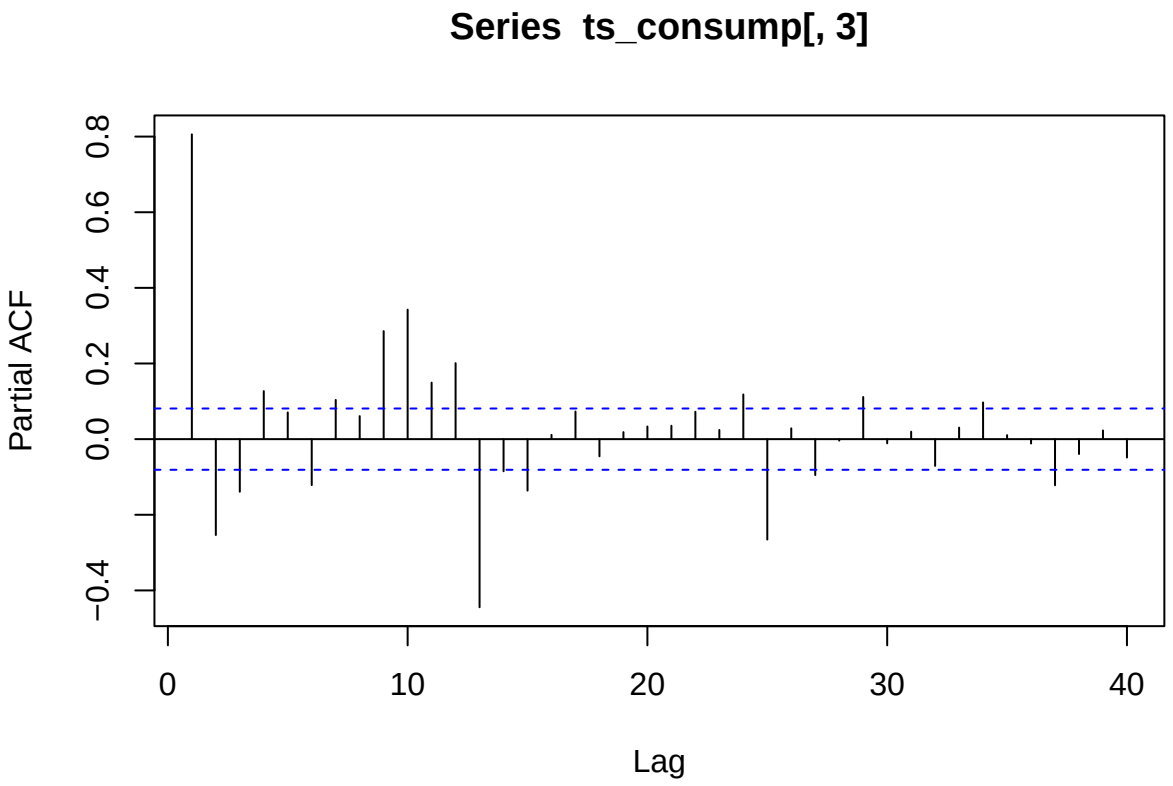
```
bio.pacf<-pacf(ts_consump[,1], lag=40)
```




```
renw.pacf<-pacf(ts_consump[,2], lag=40)
```




```
hydro.pacf<-pacf(ts_consump[,3], lag=40)
```


Series ts_consump[, 3]
Partial ACF
0.8
0.6
0.4
0.2
0.0
−0.4
0
10
20
30
40
Lag