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

Assignment 2 - Due date 02/25/24

Julia Kagiliery

## 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.

#Load/install required package here  
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(tseries)

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

library(forecast)   
library(lubridate)

##   
## Attaching package: 'lubridate'

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

library(ggplot2)  
library(tinytex)

## Data set information

Consider the data provided in the spreadsheet “Table\_10.1\_Renewable\_Energy\_Production\_and\_Consumption\_by\_Source.xlsx” 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 version of the data “Table\_10.1\_Renewable\_Energy\_Production\_and\_Consumption\_by\_Source-Edit.csv”. You may use the function to import the 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 .

library(readxl)  
energy\_data <- read\_excel(path="~/Julia\_Kagiliery\_TSA\_Sp24/Data/Table\_10.1\_Renewable\_Energy\_Production\_and\_Consumption\_by\_Source.xlsx",skip = 12, sheet="Monthly Data",col\_names=FALSE)

## New names:  
## • `` -> `...1`  
## • `` -> `...2`  
## • `` -> `...3`  
## • `` -> `...4`  
## • `` -> `...5`  
## • `` -> `...6`  
## • `` -> `...7`  
## • `` -> `...8`  
## • `` -> `...9`  
## • `` -> `...10`  
## • `` -> `...11`  
## • `` -> `...12`  
## • `` -> `...13`  
## • `` -> `...14`

energy\_data

## # A tibble: 609 × 14  
## ...1 ...2 ...3 ...4 ...5 ...6 ...7 ...8 ...9 ...10  
## <dttm> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <chr> <chr> <dbl>  
## 1 1973-01-01 00:00:00 130. Not Avai… 130. 220. 89.6 0.49 Not … Not … 130.  
## 2 1973-02-01 00:00:00 117. Not Avai… 117. 197. 79.5 0.448 Not … Not … 117.  
## 3 1973-03-01 00:00:00 130. Not Avai… 130. 219. 88.3 0.464 Not … Not … 130.  
## 4 1973-04-01 00:00:00 125. Not Avai… 126. 209. 83.2 0.542 Not … Not … 125.  
## 5 1973-05-01 00:00:00 130. Not Avai… 130. 216. 85.6 0.505 Not … Not … 130.  
## 6 1973-06-01 00:00:00 125. Not Avai… 126. 208. 82.1 0.579 Not … Not … 125.  
## 7 1973-07-01 00:00:00 130. Not Avai… 130. 208. 77.4 0.614 Not … Not … 130.  
## 8 1973-08-01 00:00:00 130. Not Avai… 130. 203. 72.9 0.579 Not … Not … 130.  
## 9 1973-09-01 00:00:00 126. Not Avai… 126. 185. 59.0 0.49 Not … Not … 126.  
## 10 1973-10-01 00:00:00 130. Not Avai… 130. 194. 63.0 0.578 Not … Not … 130.  
## # ℹ 599 more rows  
## # ℹ 4 more variables: ...11 <dbl>, ...12 <chr>, ...13 <dbl>, ...14 <dbl>

## 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\_data1 <- energy\_data[,4:6]

colnames(energy\_data1) <- c('Total Biomass Energy Production', 'Total Renewable Energy Production', 'Hydroelectric Power Consumption')  
head(energy\_data1)

## # A tibble: 6 × 3  
## Total Biomass Energy Productio…¹ Total Renewable Ener…² Hydroelectric Power …³  
## <dbl> <dbl> <dbl>  
## 1 130. 220. 89.6  
## 2 117. 197. 79.5  
## 3 130. 219. 88.3  
## 4 126. 209. 83.2  
## 5 130. 216. 85.6  
## 6 126. 208. 82.1  
## # ℹ abbreviated names: ¹​`Total Biomass Energy Production`,  
## # ²​`Total Renewable Energy Production`, ³​`Hydroelectric Power Consumption`

## 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().

tsenergy <- energy\_data1 |>  
 ts(start = c(1973,1), frequency = 12)

## Question 3

Compute mean and standard deviation for these three series.

library(skimr)  
skim(tsenergy)

Data summary

|  |  |
| --- | --- |
| Name | tsenergy |
| Number of rows | 609 |
| Number of columns | 3 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 3 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

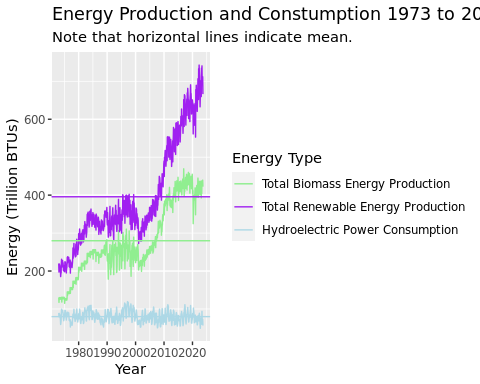
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Total Biomass Energy Production | 0 | 1 | 279.80 | 92.67 | 114.94 | 222.74 | 254.18 | 373.48 | 469.36 | ▃▇▃▂▃ |
| Total Renewable Energy Production | 0 | 1 | 395.72 | 137.80 | 185.30 | 309.92 | 346.51 | 499.56 | 742.75 | ▃▇▁▂▂ |
| Hydroelectric Power Consumption | 0 | 1 | 79.73 | 14.15 | 49.02 | 69.02 | 78.99 | 89.40 | 119.40 | ▃▇▇▃▁ |

The mean for total biomass energy production is 279.80457 and the standard deviation is 92.66504. The mean for total rennewable energy production is 395.72134 and the standard deviation is 137.79520. The mean for hydroelectric power consumption is 79.73071 and the standard deviation is 49.022

## 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.

autoplot(tsenergy) +  
 xlab("Year") +  
 ylab("Energy (Trillion BTUs)") +  
 labs(color="Energy Type", title="Energy Production and Constumption 1973 to 2023", subtitle ="Note that horizontal lines indicate mean.") +  
 scale\_color\_manual(values = c("Total Biomass Energy Production" = "lightgreen", "Total Renewable Energy Production"= "purple", "Hydroelectric Power Consumption" = "lightblue")) +  
 geom\_hline(yintercept=279.80457, col="lightgreen") +  
 geom\_hline(yintercept= 395.72134, col="purple") +  
 geom\_hline(yintercept=79.73071, col="lightblue")

 The hydroelectric power consumption appears to oscilate around it’s own mean which leads me to belive that the data is possibly seasonal. Both the total biomass production and the renewable energy production are increasing over time so it is a touch more difficult to conjecture based on these plots alone. No matter what, plots should not be used in isolation to draw conclusions, further statistical analysis is warranted.

## Question 5

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

tsenergy |>  
 cor() |>  
 print()

## Total Biomass Energy Production  
## Total Biomass Energy Production 1.00000000  
## Total Renewable Energy Production 0.97074621  
## Hydroelectric Power Consumption -0.09656318  
## Total Renewable Energy Production  
## Total Biomass Energy Production 0.970746212  
## Total Renewable Energy Production 1.000000000  
## Hydroelectric Power Consumption -0.001768629  
## Hydroelectric Power Consumption  
## Total Biomass Energy Production -0.096563177  
## Total Renewable Energy Production -0.001768629  
## Hydroelectric Power Consumption 1.000000000

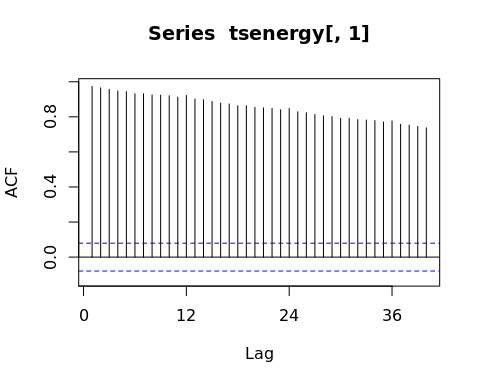
It appears as though total biomass energy production and total renewable energy production are strongly positively corelated as they have a correlation coefficient of 0.97075. However, these two are not correlated to hydroelectric power consumption as those coefficents are (in magnitude) lower than 0.7 (about where I would draw the line at taking strong conclusions about correlation).

## Question 6

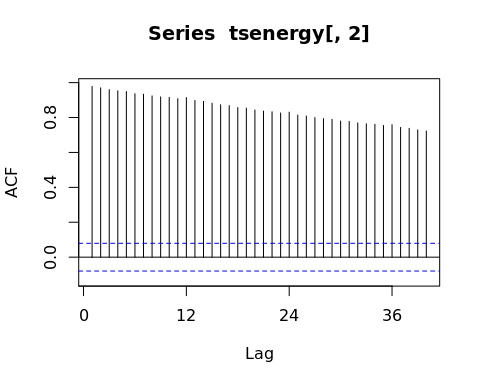
Compute the autocorrelation function from lag 1 up to lag 40 for these three variables. What can you say about these plots? Do the three of them have the same behavior?

These plot reflect something about the trends in the data between the relatively long time lags (that it the lag is cumulative in that more than just the proevious time point are accounted for). The first two plots seem to reflect a similar behavior in that these larger lag times do not best fit the data because there is some fundamental change within that lag (the upward trend). Hence, it seems like we might need to remove a trend before evaluating seasonality. The third plot looks a little different in that there seems to be a sinosoidal-type behavior, perhaps indicating a different relationship with the longer lag time, that is maybe there is less of a trend and more seasonaility behind that particular time series.

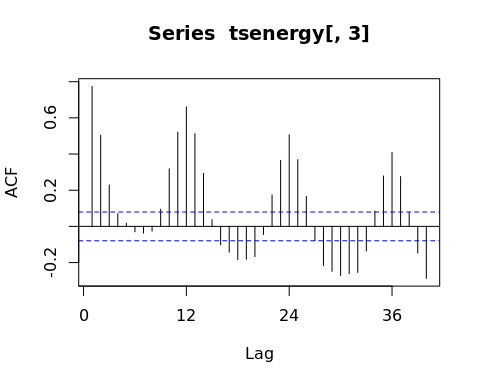
tsenergy[,1] |>  
 Acf(lag.max = 40)



tsenergy[,2] |>  
 Acf(lag.max = 40)



tsenergy[,3] |>  
 Acf(lag.max = 40)

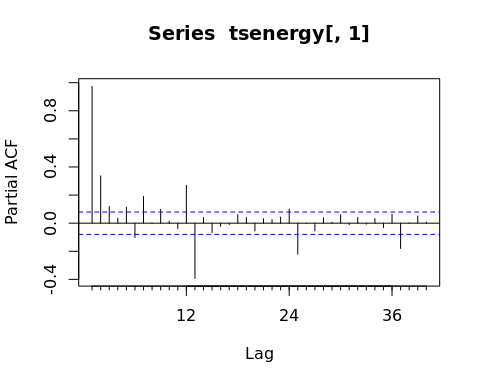


## Question 7

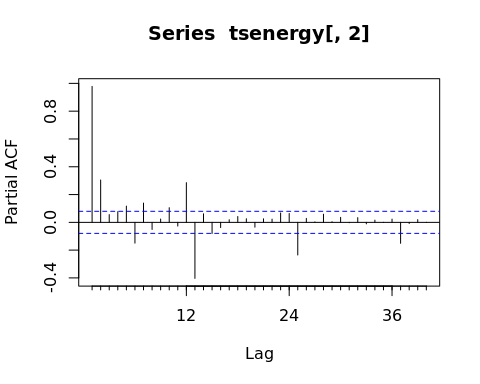
Compute the partial autocorrelation function from lag 1 to lag 40 for these three variables. How these plots differ from the ones in Q6?

The PACF looks different here because it accounts for a different lag. That is, in PACF lags before the previous one are ignored so the next time point is viewed only by the one immediately prior. These plots look far more similar because they ignore the long cumulative trends and hence tell us a bit more about exclusively the seasonaility. All three plots seem to demonstarte a similar oscillating seasonaility.

tsenergy[,1] |>  
 Pacf(lag.max = 40)



tsenergy[,2] |>  
 Pacf(lag.max = 40)



tsenergy[,3] |>  
 Pacf(lag.max = 40)

