ENV 790.30 - Time Series Analysis for Energy Data | Spring 2022

Assignment 4 - Due date 02/17/22

Tatiana Sokolova

## Directions

You should open the .rmd file corresponding to this assignment on RStudio. The file is available on our class repository on Github. And to do so you will need to fork our repository and link it to your RStudio.

Once you have the project open the first thing you will do is change “Student Name” on line 3 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\_A04\_Sp21.Rmd”). Submit this pdf using Sakai.

R packages needed for this assignment: “xlsx” or “readxl”, “ggplot2”, “forecast”,“tseries”, and “Kendall”. 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.\

library(xlsx)  
library(readxl)  
library(lubridate)

##   
## Attaching package: 'lubridate'

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

library(ggplot2)  
library(forecast)

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

library(tseries)  
library(Kendall)  
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

## Questions

Consider the same data you used for A3 from the spreadsheet “Table\_10.1\_Renewable\_Energy\_Production\_and\_Consumption\_by\_Source.xlsx”. The data comes from the US Energy Information and Administration and corresponds to the January 2021 Monthly Energy Review. For this assignment you will work only with the column “Total Renewable Energy Production”.

#Importing data set - using xlsx package  
energy\_data <- read.xlsx(file="../Data/Table\_10.1\_Renewable\_Energy\_Production\_and\_Consumption\_by\_Source.xlsx",header=FALSE, startRow=13, sheetIndex=1)  
#extracting column names from row 11  
read\_col\_names <- read.xlsx(file="../Data/Table\_10.1\_Renewable\_Energy\_Production\_and\_Consumption\_by\_Source.xlsx",header=FALSE,startRow=11,endRow=11,sheetIndex=1)  
colnames(energy\_data) <- read\_col\_names  
head(energy\_data)

## Month Wood Energy Production Biofuels Production  
## 1 1973-01-01 129.630 Not Available  
## 2 1973-02-01 117.194 Not Available  
## 3 1973-03-01 129.763 Not Available  
## 4 1973-04-01 125.462 Not Available  
## 5 1973-05-01 129.624 Not Available  
## 6 1973-06-01 125.435 Not Available  
## Total Biomass Energy Production Total Renewable Energy Production  
## 1 129.787 403.981  
## 2 117.338 360.900  
## 3 129.938 400.161  
## 4 125.636 380.470  
## 5 129.834 392.141  
## 6 125.611 377.232  
## Hydroelectric Power Consumption Geothermal Energy Consumption  
## 1 272.703 1.491  
## 2 242.199 1.363  
## 3 268.810 1.412  
## 4 253.185 1.649  
## 5 260.770 1.537  
## 6 249.859 1.763  
## Solar Energy Consumption Wind Energy Consumption Wood Energy Consumption  
## 1 Not Available Not Available 129.630  
## 2 Not Available Not Available 117.194  
## 3 Not Available Not Available 129.763  
## 4 Not Available Not Available 125.462  
## 5 Not Available Not Available 129.624  
## 6 Not Available Not Available 125.435  
## Waste Energy Consumption Biofuels Consumption  
## 1 0.157 Not Available  
## 2 0.144 Not Available  
## 3 0.176 Not Available  
## 4 0.174 Not Available  
## 5 0.210 Not Available  
## 6 0.176 Not Available  
## Total Biomass Energy Consumption Total Renewable Energy Consumption  
## 1 129.787 403.981  
## 2 117.338 360.900  
## 3 129.938 400.161  
## 4 125.636 380.470  
## 5 129.834 392.141  
## 6 125.611 377.232

#creating df structure for column of interest and   
df <- energy\_data[,c('Month','Total Renewable Energy Production')]  
head(df)

## Month Total Renewable Energy Production  
## 1 1973-01-01 403.981  
## 2 1973-02-01 360.900  
## 3 1973-03-01 400.161  
## 4 1973-04-01 380.470  
## 5 1973-05-01 392.141  
## 6 1973-06-01 377.232

#removing January 1973 to compare with differenced df  
df\_584<-df[-c(1),]  
head(df\_584)

## Month Total Renewable Energy Production  
## 2 1973-02-01 360.900  
## 3 1973-03-01 400.161  
## 4 1973-04-01 380.470  
## 5 1973-05-01 392.141  
## 6 1973-06-01 377.232  
## 7 1973-07-01 367.325

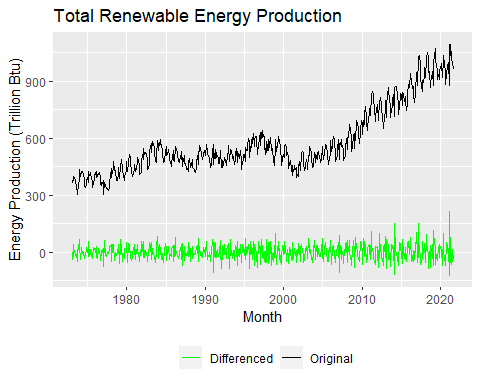
## Stochastic Trend and Stationarity Tests

### Q1

Difference the “Total Renewable Energy Production” series using function diff(). Function diff() is from package base and take three main arguments: \* *x* vector containing values to be differenced; \* *lag* integer indicating with lag to use; \* *differences* integer indicating how many times series should be differenced.

Try differencing at lag 1 only once, i.e., make lag=1 and differences=1. Plot the differenced series. Do the series still seem to have trend?

energy\_data\_diff <- diff(df[,"Total Renewable Energy Production"],lag=1, differences=1)  
  
ggplot(df\_584, aes(x=df\_584[,1], y=df\_584[,"Total Renewable Energy Production"])) +  
 geom\_line(aes(x=df\_584[,1], y = energy\_data\_diff, color = "Differenced")) +  
 geom\_line(aes(x=df\_584[,1], y=df\_584[,"Total Renewable Energy Production"], color = "Original")) +  
 labs(color="") +  
 scale\_color\_manual(values = c("Differenced" = "green", "Original" = "black"),  
 labels=c("Differenced", "Original")) +  
 theme(legend.position = "bottom") +  
 ggtitle("Total Renewable Energy Production")+  
 ylab(paste0("Energy Production (Trillion Btu)")) +  
 xlab(paste0("Month"))

 The series appears to be detrended.

### Q2

Now let’s compare the differenced series with the detrended series you calculated on A3. In other words, for the “Total Renewable Energy Production” compare the differenced series from Q1 with the series you detrended in A3 using linear regression. (Hint: Just copy and paste part of your code for A3)

Copy and paste part of your code for A3 where you compute regression for Total Energy Production and the detrended Total Energy Production

#Create vector t  
nobs=nrow(df)  
t <- c(1:nobs)  
  
#Fit a linear trend to TS of Total Renewable Energy Production  
linear\_trend\_model\_renew=lm(df[,2]~t)   
summary(linear\_trend\_model\_renew)

##   
## Call:  
## lm(formula = df[, 2] ~ t)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -230.488 -57.869 5.595 62.090 261.349   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 323.18243 8.02555 40.27 <2e-16 \*\*\*  
## t 0.88051 0.02373 37.10 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 96.93 on 583 degrees of freedom  
## Multiple R-squared: 0.7025, Adjusted R-squared: 0.702   
## F-statistic: 1377 on 1 and 583 DF, p-value: < 2.2e-16

beta0r=as.numeric(linear\_trend\_model\_renew$coefficients[1]) #first coefficient is the intercept term or beta0  
beta1r=as.numeric(linear\_trend\_model\_renew$coefficients[2]) #second coefficient is the slope or beta1  
  
#remove the trend from TS of Total Renewable Energy Production  
detrend\_renew\_data <- df[,2]-(beta0r+beta1r\*t)

### Q3

Create a data frame with 4 columns: month, original series, detrended by Regression Series and differenced series. Make sure you properly name all columns. Also note that the differenced series will have only 584 rows because you lose the first observation when differencing. Therefore, you need to remove the first observations for the original series and the detrended by regression series to build the new data frame.

#removing Jan 1973 from detrended data  
detrend\_renew\_data <- detrend\_renew\_data[-1]  
head(detrend\_renew\_data)

## [1] 35.95655 74.33705 53.76554 64.55603 48.76653 37.97902

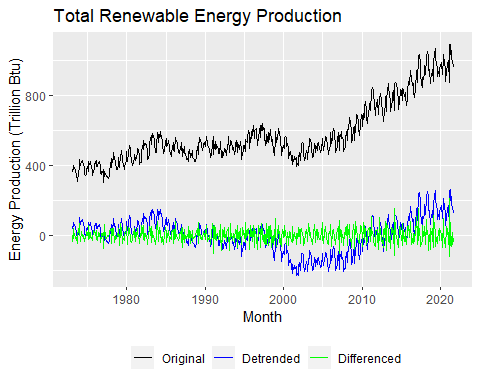
new\_df <- data.frame(Month = df\_584$Month,  
 Original = df\_584[,2],  
 Detrended = detrend\_renew\_data,   
 Differenced = energy\_data\_diff)  
head(new\_df)

## Month Original Detrended Differenced  
## 1 1973-02-01 360.900 35.95655 -43.081  
## 2 1973-03-01 400.161 74.33705 39.261  
## 3 1973-04-01 380.470 53.76554 -19.691  
## 4 1973-05-01 392.141 64.55603 11.671  
## 5 1973-06-01 377.232 48.76653 -14.909  
## 6 1973-07-01 367.325 37.97902 -9.907

### Q4

Using ggplot() create a line plot that shows the three series together. Make sure you add a legend to the plot.

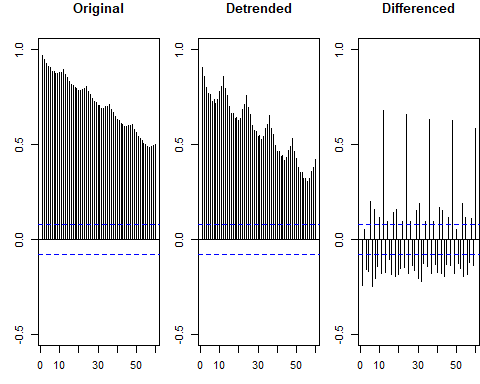
#Use ggplot  
ggplot(new\_df, aes(x=new\_df[,1], y=new\_df[,2])) +  
 geom\_line(aes(x=new\_df[,1], y= new\_df[,2], color = "Original")) +  
 geom\_line(aes(x=new\_df[,1], y = new\_df[,3], color = "Detrended")) +  
 geom\_line(aes(x=new\_df[,1], y= new\_df[,4], color = "Differenced")) +  
 labs(color="") +  
 scale\_color\_manual(values = c("Original" = "black","Detrended" = "blue","Differenced" = "green"),  
 labels=c("Original", "Detrended", "Differenced")) +  
 theme(legend.position = "bottom") +  
 ggtitle("Total Renewable Energy Production")+  
 ylab(paste0("Energy Production (Trillion Btu)")) +  
 xlab(paste0("Month"))



### Q5

Plot the ACF for the three series and compare the plots. Add the argument ylim=c(-0.5,1) to the Acf() function to make sure all three y axis have the same limits. Which method do you think was more efficient in eliminating the trend? The linear regression or differencing?

#Comparing ACFs  
par(mar=c(3,3,3,0));par(mfrow=c(1,3))  
Acf(new\_df$Original,lag.max=60,main="Original",ylim=c(-0.5,1))  
Acf(new\_df$Detrended,lag.max=60,main="Detrended",ylim=c(-0.5,1))  
Acf(new\_df$Differenced,lag.max=60,main="Differenced",ylim=c(-0.5,1))



The differencing was more efficient in eliminating the trend.

### Q6

Compute the Seasonal Mann-Kendall and ADF Test for the original “Total Renewable Energy Production” series. Ask R to print the results. Interpret the results for both tests. What’s the conclusion from the Seasonal Mann Kendall test? What’s the conclusion for the ADF test? Do they match what you observed in Q2? Recall that having a unit root means the series has a stochastic trend. And when a series has stochastic trend we need to use a different procedure to remove the trend.

#converting to time series  
ts\_renew\_energy <- ts(df[,2],frequency=12)  
  
#Seasonal Mann-Kendall  
SMKtest <- SeasonalMannKendall(ts\_renew\_energy)  
print("Results for Seasonal Mann Kendall")

## [1] "Results for Seasonal Mann Kendall"

print(summary(SMKtest))

## Score = 9984 , Var(Score) = 159104  
## denominator = 13968  
## tau = 0.715, 2-sided pvalue =< 2.22e-16  
## NULL

The p value is sharing how significant the tau value is and it definitely statistically different from 0 so the null hypothesis that there is no trend present in the data should be rejected. The larger score indicates that it is a positive trend. This matches what I observed in Q2.

#Null hypothesis is that data has a unit root  
print("Results for ADF test")

## [1] "Results for ADF test"

print(adf.test(ts\_renew\_energy,alternative = "stationary")) #stationary over a unit root but could be non stationary

##   
## Augmented Dickey-Fuller Test  
##   
## data: ts\_renew\_energy  
## Dickey-Fuller = -1.4383, Lag order = 8, p-value = 0.8161  
## alternative hypothesis: stationary

From the ADF test, since the p-value is greater than 0.05, we can conclude that we fail to reject the null hypothesis that the time series is non-stationary and can therefore conclude the time series has some time-dependent structure and does not have a constant variance over time. This matches what I observed in Q2.

### Q7

Aggregate the original “Total Renewable Energy Production” series by year. You can use the same procedure we used in class. Store series in a matrix where rows represent months and columns represent years. And then take the columns mean using function colMeans(). Recall the goal is the remove the seasonal variation from the series to check for trend.

energy\_data\_matrix <- matrix(df[,2],byrow=FALSE,nrow=12)

## Warning in matrix(df[, 2], byrow = FALSE, nrow = 12): data length [585] is not a  
## sub-multiple or multiple of the number of rows [12]

energy\_data\_yearly <- colMeans(energy\_data\_matrix)  
  
#for Spearman test  
my\_year <- c(year(first(df[,1])):year(last(df[,1])))  
head(my\_year)

## [1] 1973 1974 1975 1976 1977 1978

### Q8

Apply the Mann-Kendall, Spearman correlation rank test, and ADF. Are the results from the test in agreement with the test results for the non-aggregated series, i.e., results for Q6?

#Mann-Kendall  
print("Results of Mann Kendall on average yearly series")

## [1] "Results of Mann Kendall on average yearly series"

print(summary(MannKendall(energy\_data\_yearly)))

## Score = 854 , Var(Score) = 13458.67  
## denominator = 1176  
## tau = 0.726, 2-sided pvalue =< 2.22e-16  
## NULL

The p-value is still statistically different from 0 so the rejection of the null hypothesis (there is no trend present in the data) still holds. Tau is slightly bigger than that of the non-aggregated series, implying that there is an even stronger positive correlation within this aggregated series.

#Spearman Correlation Test  
print("Results from Spearman Correlation")

## [1] "Results from Spearman Correlation"

sp\_rho=cor.test(energy\_data\_yearly,my\_year,method="spearman")   
print(sp\_rho)

##   
## Spearman's rank correlation rho  
##   
## data: energy\_data\_yearly and my\_year  
## S = 2578, p-value < 2.2e-16  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.8684694

Question 6 did not require the Spearman test because it cannot handle seasonality. The p-value for the test on the yearly data indicates that there is very strong evidence for rejecting the null hypothesis that there is no monotonic association in the series. The rho indicates that there is a strong positive correlation within this aggregated series.

#Augmented Dickey-Fuller Test  
print("Results for ADF test on yearly data")

## [1] "Results for ADF test on yearly data"

print(adf.test(energy\_data\_yearly, alternative = "stationary"))

##   
## Augmented Dickey-Fuller Test  
##   
## data: energy\_data\_yearly  
## Dickey-Fuller = -2.2085, Lag order = 3, p-value = 0.4907  
## alternative hypothesis: stationary

The p-value is still greater than 0.05 and therefore we can continue to conclude that the time series is stationary. The Dickey-Fuller value is more negative than that of the non-aggregated series, implying that there is an even stronger rejection of the null hypothesis.