ENV 790.30 - Time Series Analysis for Energy Data | Spring 2023

Assignment 4 - Due date 02/17/23

Maggie O'Shea

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 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_A04_Sp23.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 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.\

```
#Load/install required package here
library(lubridate)
library(ggplot2)
library(forecast)
library(Kendall)
library(tseries)
library(base)
library(outliers)
library(tidyverse)
library(dplyr)
```

Questions

Consider the same data you used for A3 from the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumpti
The data comes from the US Energy Information and Administration and corresponds to the December 2022 Monthly Energy Review. For this assignment you will work only with the column "Total Renewable Energy Production".

```
#Importing data set - using xlsx package
#Note: xlsx was resulting in an error due to the java version I have on my machine.
#I thus used readxl instead.
#install.packages("readxl")
library(readxl)
```

```r
energy_data<- read_excel("./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx"
read_col_names <- read_xlsx(path="Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source
colnames(energy_data) <- read_col_names

energy_clean <- energy_data%>%
  select("Month", "Total Renewable Energy Production")

ts_energy <- ts(energy_clean, frequency = 12, start = c(1973, 1))
```

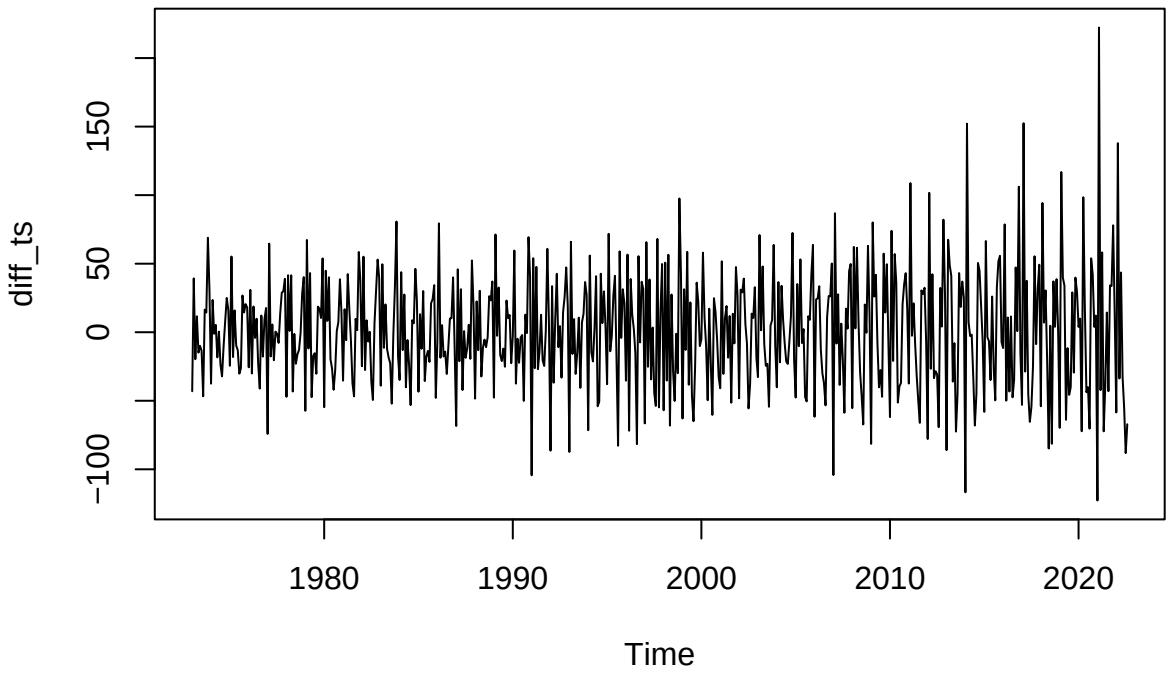
Stochastic Trend and Stationarity Tests

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?

```r
diffseries <- diff(energy_clean$`Total Renewable Energy Production`, lag = 1, differences = 1)
diff_ts <- ts(diffseries, frequency = 12, start = c(1973, 1, 1))
plot(diff_ts)
```



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)

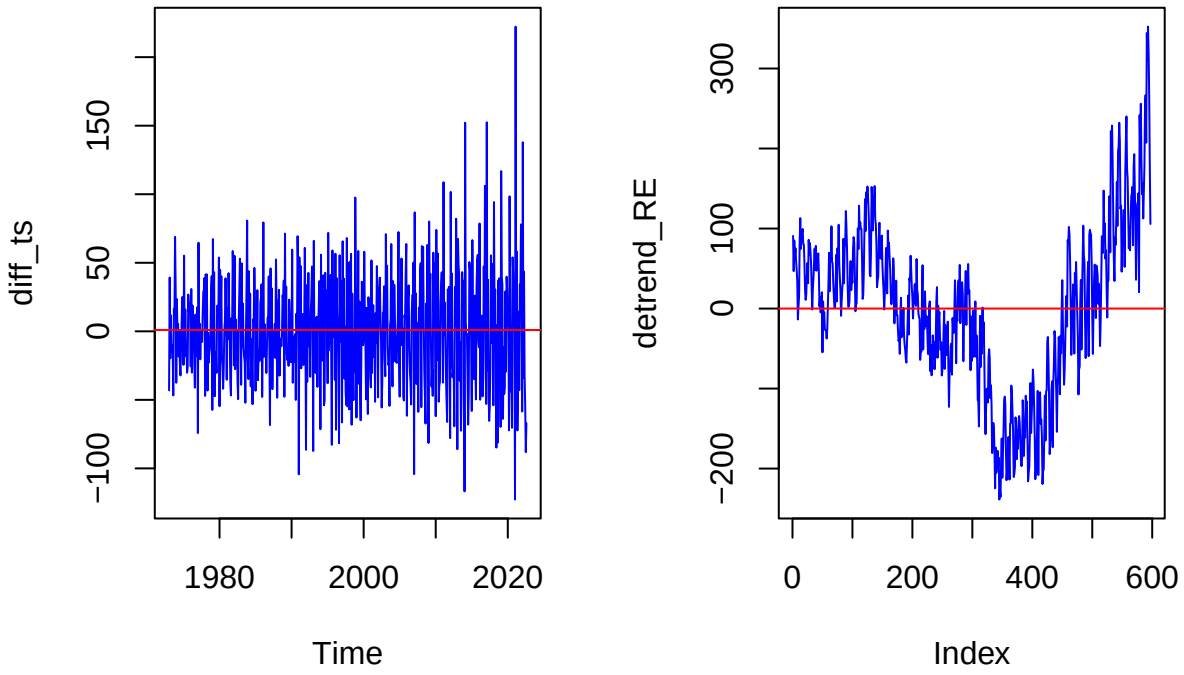
Copy and paste part of your code for A3 where you compute regression for Total Energy Production and the detrended Total Energy Production

```
#Detrend Renewable Energy
nobs <- nrow(energy_clean)
t <- 1:nobs
linearTrend_renew<- lm(energy_clean$`Total Renewable Energy Production` ~ t)
summary(linearTrend_renew)
```

```
##
## Call:
## lm(formula = energy_clean$`Total Renewable Energy Production` ~
##     t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -238.75  -61.85    8.59   64.48  352.27
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 312.2475     8.4902   36.78   <2e-16 ***
## t             0.9362     0.0246   38.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103.6 on 595 degrees of freedom
## Multiple R-squared:  0.7088, Adjusted R-squared:  0.7083
## F-statistic:  1448 on 1 and 595 DF,  p-value: < 2.2e-16
```

```
beta0_renew <- linearTrend_renew$coefficients[1]
beta1_renew <- linearTrend_renew$coefficients[2]
detrend_RE <- energy_clean$`Total Renewable Energy Production` - (beta0_renew+beta1_renew*t)

#Compare Plots
par(mfrow = c(1,2))
  plot(diff_ts,type="l",col="blue")
  abline(h=mean(diff_ts),col="red")
  plot(detrend_RE,type="l",col="blue")
  abline(h=mean(detrend_RE),col="red")
```



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 loose 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.

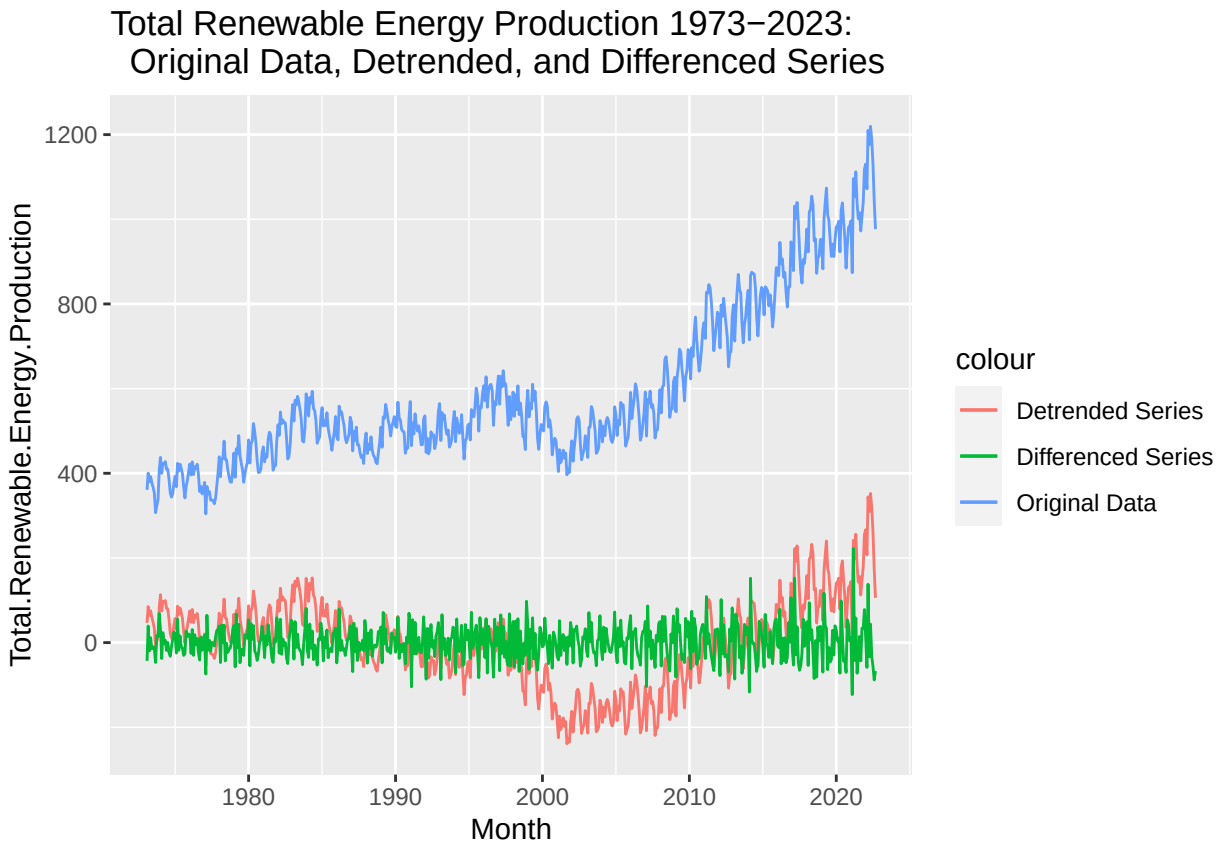
```
#Data frame - remember to not include January 1973
#Removing first observation (January 1973)
detrend_RE_v2 <- detrend_RE[-1];
orig_Data <- energy_clean[-1,]

detrendanddiff_v1 <- data.frame(orig_Data, detrend_RE_v2, diffseries)
fulldataframe <- detrendanddiff_v1%>%
  rename(
         "Detrended_Series" = detrend_RE_v2,
         "Differenced_Series" = diffseries)
```

Q4

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

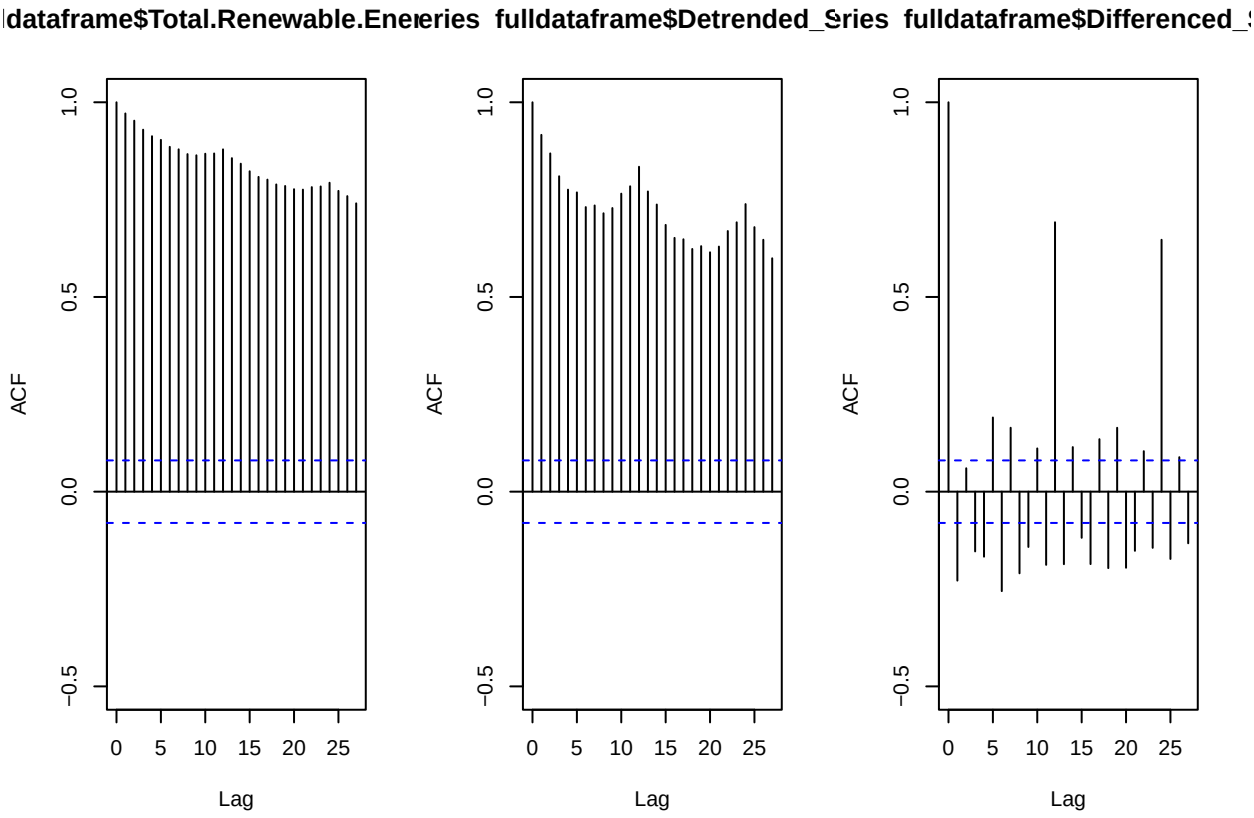
```
#Use ggplot
ggplot(fulldataframe, aes(x=Month))+
  geom_line(aes(y=Total.Renewable.Energy.Production, color="Original Data"))+
  geom_line(aes(y=Detrended_Series, color = "Detrended Series"))+
  geom_line(aes(y=Differenced_Series, color ="Differenced Series"))+
  guides(fill = guide_legend(title = "Series"))+
  labs(title="Total Renewable Energy Production 1973-2023:
  Original Data, Detrended, and Differenced Series")
```



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?

```
#Compare ACFs
par(mfrow = c(1,3))
  acf(fulldataframe$Total.Renewable.Energy.Production, ylim=c(-0.5,1))
  acf(fulldataframe$Detrended_Series, ylim=c(-0.5,1))
  acf(fulldataframe$Differenced_Series, ylim=c(-0.5,1))
```



Q5 Answer:

The differenced series appears to have removed the trend and shows instead correlations that vary from positive to negative and a few high correlations at lag 12 and 24. In contrast, the detrended series still has a trend, though does show more seasonality than the original ACF. Thus, the differenced series was more effective in eliminating the trend altogether.

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 test. Whats 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.

```
#Null hypothesis is that data has a unit root
print(adf.test(ts_energy[,1],alternative = "stationary"))
```

```
## Warning in adf.test(ts_energy[, 1], alternative = "stationary"): p-value smaller
## than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  ts_energy[, 1]
```

```
## Dickey-Fuller = -8.945, Lag order = 8, p-value = 0.01
## alternative hypothesis: stationary
```

```
#SMK:
SMKtest <- SeasonalMannKendall(ts_energy[,1])
print(SMKtest)
```

```
## tau = 1, 2-sided pvalue =< 2.22e-16
```

Q6 Answer The conclusion from this ADF tests ($p<0.01$) suggests that we should reject the null and thus that the series is stationary.This does not align with what the ACF plots visually show, and examining the plots from Q2, provides additional evidence that the detrending did not remove the trend. The Seasonal Mann Kendall, in contrast, suggests that there is a trend (because $p<0.01$ and rejects the null that there is not a trend). This confirms that there is a trend and thus the plots in Q2 show that, given that there is statistical evidence of a trend, based on the plots in Q2 suggests that differencing was the better strategy to remove the trend.

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.

```
#Group data in yearly steps instances
energy_matrix <- matrix(ts_energy[,1],byrow=FALSE,nrow=12)
```

```
## Warning in matrix(ts_energy[, 1], byrow = FALSE, nrow = 12): data length [597]
## is not a sub-multiple or multiple of the number of rows [12]
```

```
energy_year <- colMeans(energy_matrix)
years<- 1973:2022

energy_yearly <- data.frame(years, energy_year)%>%
  rename("MeanYearlyRenewableProduction" = energy_year,
         "Year" = years)
```

Q8

Apply the Mann Kendal, 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?

```
print(adf.test(energy_yearly[,2],alternative = "stationary"))
```

```
## Warning in adf.test(energy_yearly[, 2], alternative = "stationary"): p-value
## greater than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
```

```
##
## data:  energy_yearly[, 2]
## Dickey-Fuller = 0.23771, Lag order = 3, p-value = 0.99
## alternative hypothesis: stationary
```

```
summary(MannKendall(energy_yearly[,2]))
```

```
## Score =  1201 , Var(Score) = 14291.67
## denominator =  1225
## tau = 0.98, 2-sided pvalue =< 2.22e-16
```

```
sp_rho=cor.test(energy_yearly[,2],years, method="spearman")
print(sp_rho)
```

```
##
##  Spearman's rank correlation rho
##
## data:  energy_yearly[, 2] and years
## S = 156, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.992509
```

Q8 Answer: The results from the ADF test on the yearly means of renewable energy production shows that, with a p-value of 0.99, we cannot reject the null and thus provides statistical evidence that the series is not stationary. This is the opposite of the Q6 ADF test which had a p-value of <0.01. The results of the Mann Kendall test align with the seasonal mann kendall results ($p<0.01$) and suggest, as the Q6 SMKtest suggested, that there is a trend. The spearman's rank correlation had a p-value of <0.01 which suggests that the correlation is not zero and there is, therefore, a statistically significant correlation between renewable energy production and time.