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

Assignment 4 - Due date 02/17/23

Kexin(Kassie) Huang

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

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

```
library(tseries)
library(Kendall)
```

Questions

Consider the same data you used for A3 from the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption 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

raw_data <- read_excel("./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",sl
```

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

```
colnames(raw_data) <- read_excel("./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sour
```

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

```
#import date column
library(lubridate)
```

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

```
my_date <- ymd(raw_data$Month)

#extract the total renewable energy production column
renewable <- raw_data$`Total Renewable Energy Production`
```

Stochastic Trend and Stationarity Tests

Q1

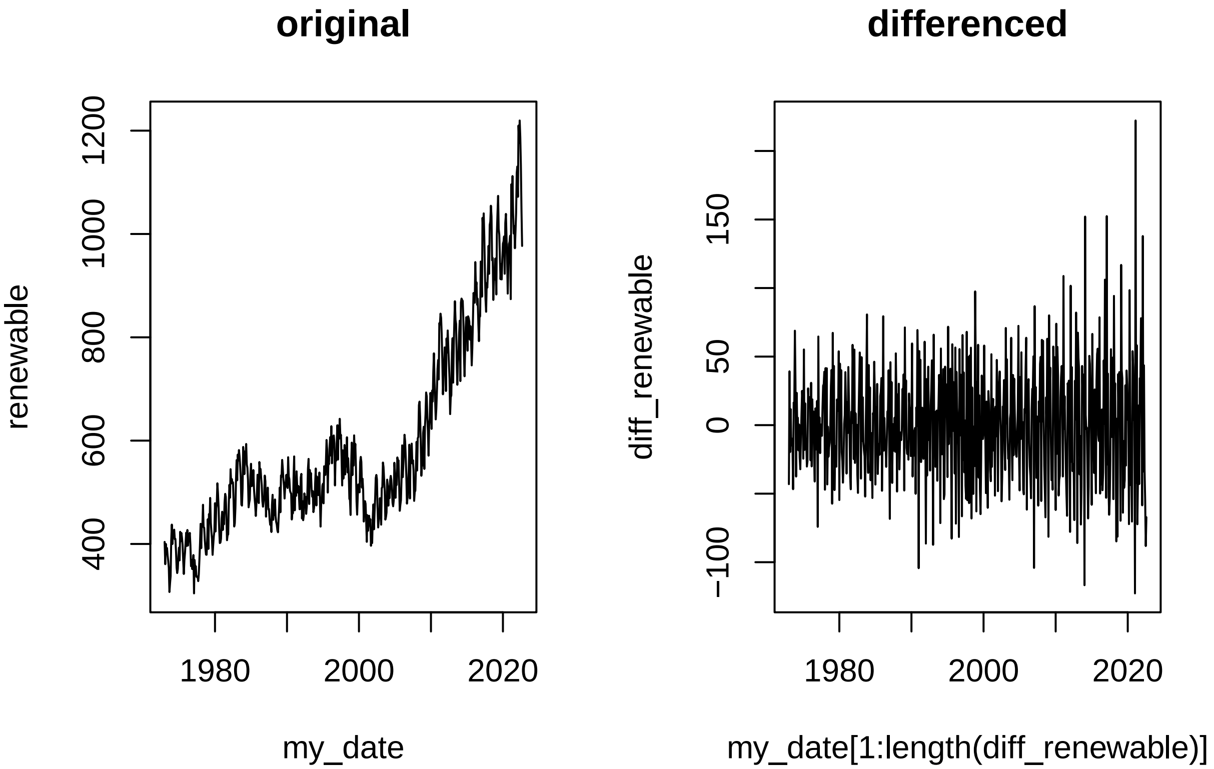
Difference the “Total Renewable Energy Production” series using function diff(). Function diff() is from package base and take three main arguments: * *x* vector containing values to be differenced; * *lag* integer indicating with lag to use; * *differences* integer indicating how many times series should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

Answer:

```
#create the differenced series
diff_renewable <- diff(x = renewable,lag = 1,differences = 1)

#plot the trend
par(mfrow=c(1,2))
plot(x=my_date,y=renewable,main='original',type='l')
plot(x=my_date[1:length(diff_renewable)],y=diff_renewable,main='differenced',type='l')
```



No. The differenced series doesn't seem to have trend any more.

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)

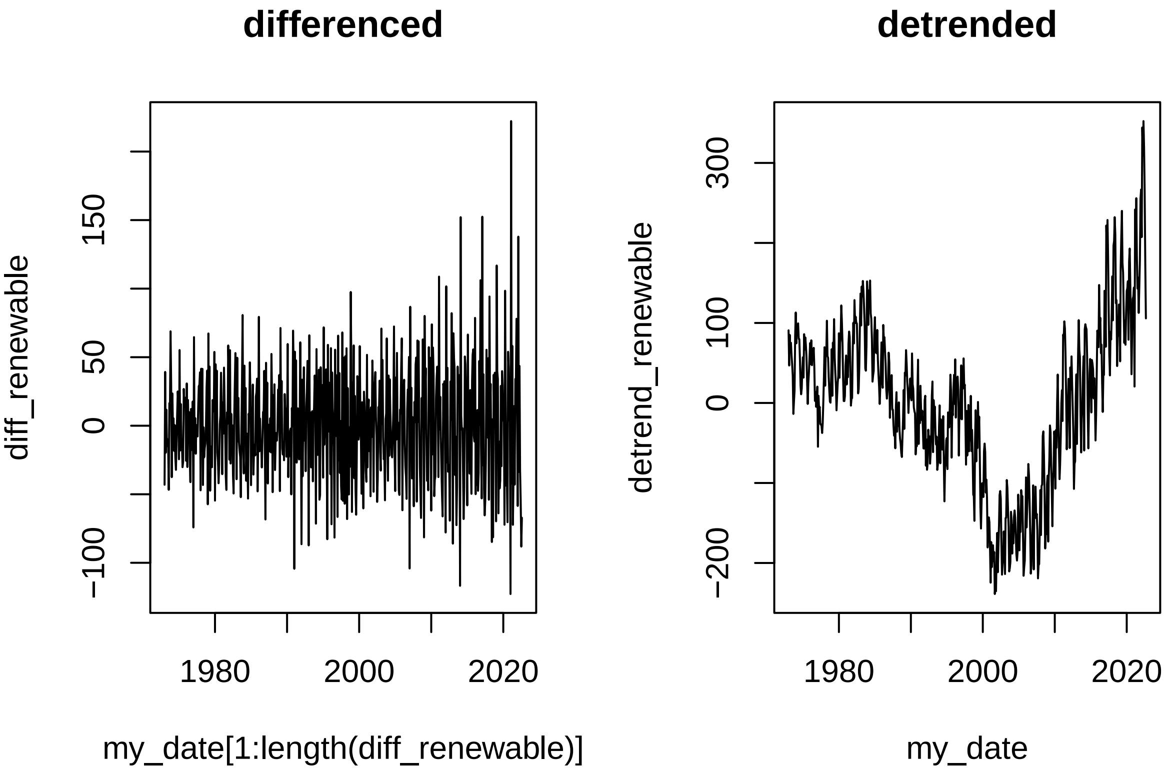
Copy and paste part of your code for A3 where you compute regression for Total Energy Production and the detrended Total Energy Production

```
#linear regression
t <- 1:length(renewable)
linear_trend_renewable <- lm(renewable~t)
summary(linear_trend_renewable)
```

```
##
## Call:
## lm(formula = renewable ~ 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_renewable <- linear_trend_renewable$coefficients[1]
beta1_renewable <- linear_trend_renewable$coefficients[2]
detrend_renewable <- renewable-(beta0_renewable + beta1_renewable*t)

#plot the series
par(mfrow=c(1,2))
plot(x=my_date[1:length(diff_renewable)],y=diff_renewable,main='differenced',type='l')
plot(x=my_date,y=detrend_renewable,main='detrended',type='l')
```



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.

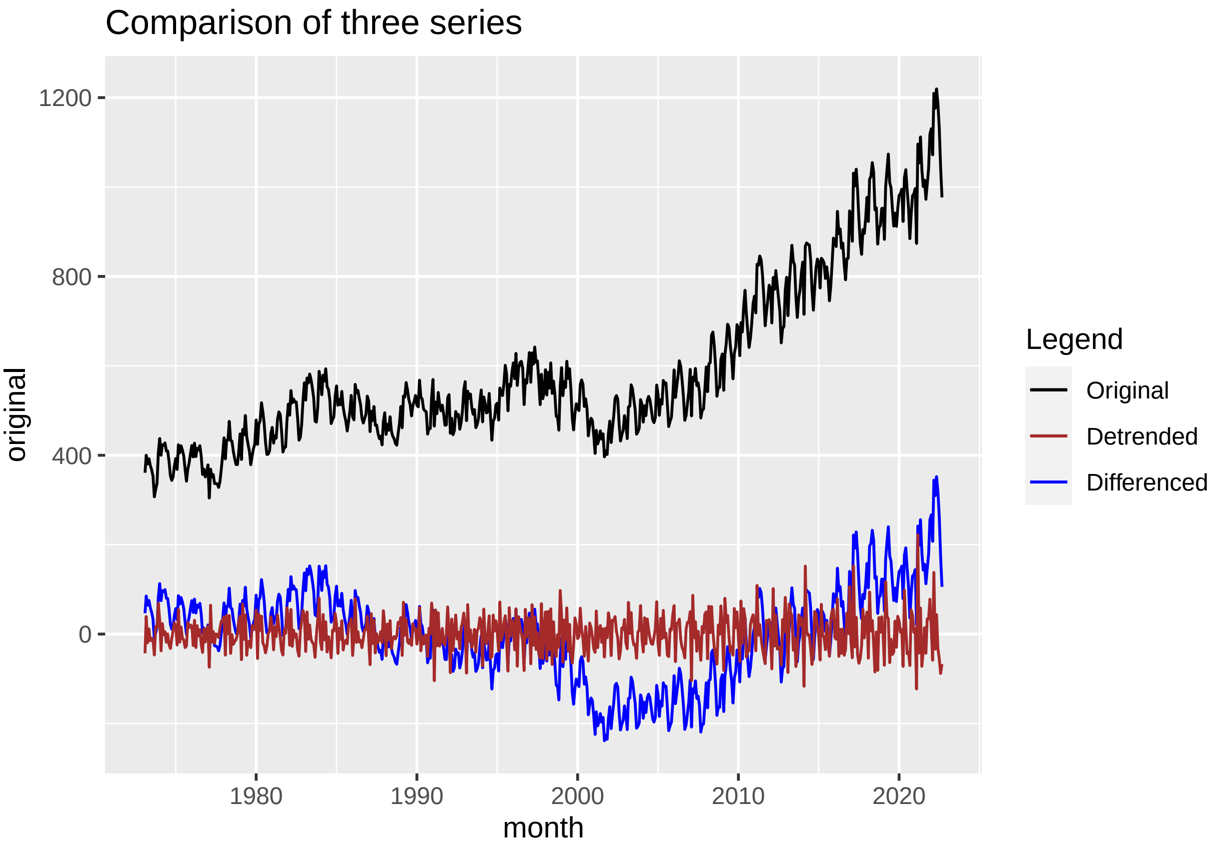
```
#Data frame - remember to not include January 1973
df_renewable <- data.frame(month=my_date[-1],original=renewable[-1],detrended=detrend_renewable[-1],dif
```

Q4

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

```
#Use ggplot
library(ggplot2)

# create ggplot with three lines
ggplot(data = df_renewable, aes(x = month)) +
  geom_line(aes(y = original, color = "black")) +
  geom_line(aes(y = detrended, color = "brown")) +
  geom_line(aes(y = differenced, color = "blue")) +
  scale_color_manual(name = "Legend",
                     values = c("black", "brown", "blue"),
                     labels = c("Original", "Detrended", "Differenced")) +
  ggtitle("Comparison of three series")
```

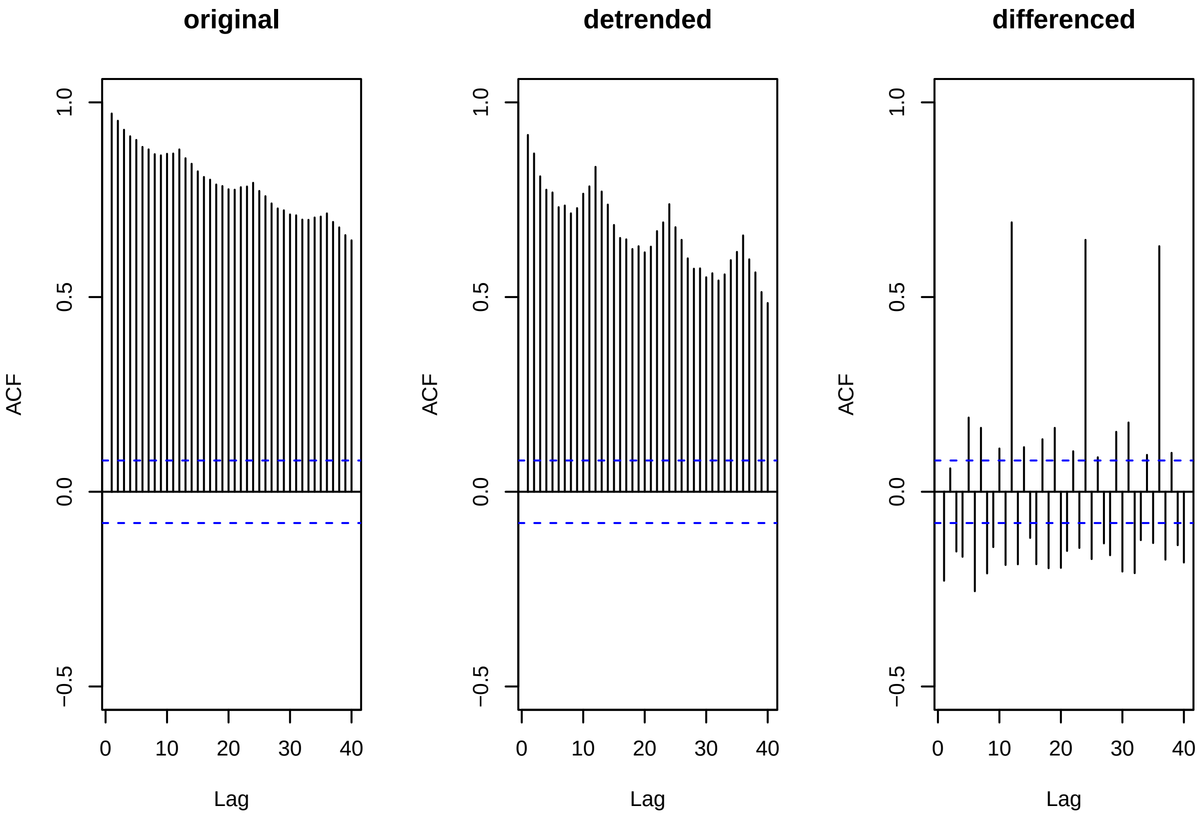


Q5

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))
```

```
for (i in 1:3) {
  Acf(ts(df_renewable[,i+1]),lag.max = 40,main=colnames(df_renewable)[i+1],ylim=c(-0.5,1))
}
```



Differencing is more efficient in eliminating the trend. The detrended one shows a combination of a general downward trend and a wave-like pattern. The differenced one, on the other hand, shows no trend, only some spikes.

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.

```
ts_renewable <- ts(renewable,start=my_date[1],frequency=12)
SeasonalMannKendall(ts_renewable)
```

```
## tau = 0.727, 2-sided pvalue =< 2.22e-16
```

```
adf.test(ts_renewable)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_renewable
## Dickey-Fuller = -1.2055, Lag order = 8, p-value = 0.9056
## alternative hypothesis: stationary
```

Seasonal MannKendall: The p-value is 2.22e-16, less than 0.05, so we can reject the null hypothesis that "the original time series is stationary." ADF: The p-value is 0.9056, greater than 0.05. Therefore, we cannot reject the null hypothesis that "the original time series contains a unit root." In other words, the original series has a stochastic 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.

```
renewable_data_matrix <- matrix(ts_renewable[1:588],byrow=FALSE,nrow=12)
#the incomplete 2022 data was removed
renewable_data_yearly <- colMeans(renewable_data_matrix)
```

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?

```
ts_renewable_yearly <- ts(renewable_data_yearly)
SeasonalMannKendall(ts_renewable_yearly)
```

```
## tau = 0.735, 2-sided pvalue =9.5035e-14
```

```
adf.test(ts_renewable_yearly)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_renewable_yearly
## Dickey-Fuller = -0.68508, Lag order = 3, p-value = 0.9654
## alternative hypothesis: stationary
```

Yes, the results are in agreement with the non-aggregated series. The p-value of seasonal MannKendall test is still less than 0.05 and p-value of the ADF test is greater than 0.05. The aggregated series can still not be regarded as stationary and we cannot reject the hypothesis that it has a unit root.