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

Assignment 4 - Due date 02/17/23

Kelsey Husted

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
#install.packages("openxlsx")
library(readxl)
library(openxlsx)
library(lubridate)
library(ggplot2)
library(forecast)
library(tseries)
library(Kendall)
library(tidyverse)
```

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
df <- read.xlsx("./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx", startRo
df$Month <- convertToDate(df$Month)

df <- df[-c(1), ] #remove the row with the units

#Selecting for Renewable Energy Production column
```

```r
energy_data <- df %>%
  select(Month, Total.Renewable.Energy.Production)
#transform renewable energy data column from character to numeric
energy_data$Total.Renewable.Energy.Production = as.numeric(energy_data$Total.Renewable.Energy.Productio
```

Stochastic Trend and Stationarity Tests

Q1

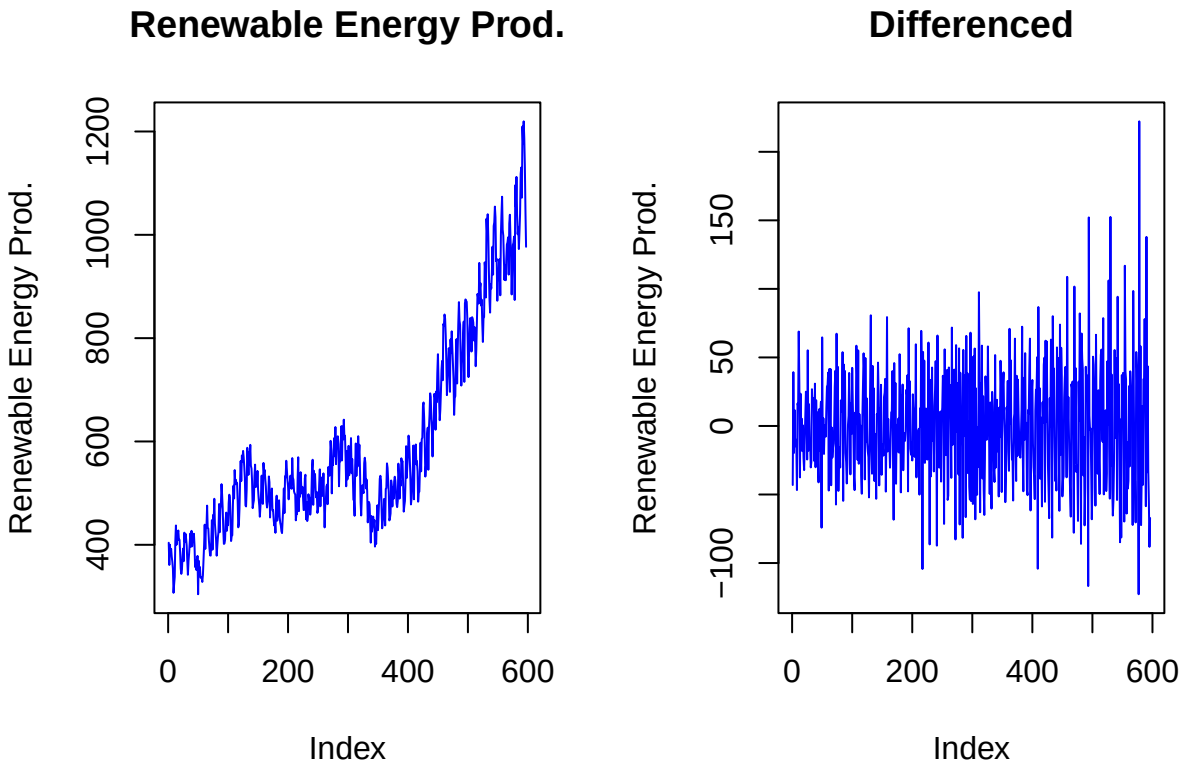
Difference the "Total Renewable Energy Production" series using function diff(). Function diff() is from package base and take three main arguments: * *x* vector containing values to be differenced; * *lag* integer indicating with lag to use; * *differences* integer indicating how many times series should 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: The trend was removed after differencing the time series as displayed in the plots below.

```r
#differencing
diff <- diff(energy_data[,2], lag = 1, differences = 1)

par(mfrow = c(1,2))
plot(energy_data[,2], type="l", col = "blue", ylab = "Renewable Energy Prod.", main = "Renewable Energy
plot(diff, type = "l", col="blue", ylab = "Renewable Energy Prod.", main = "Differenced")
```



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

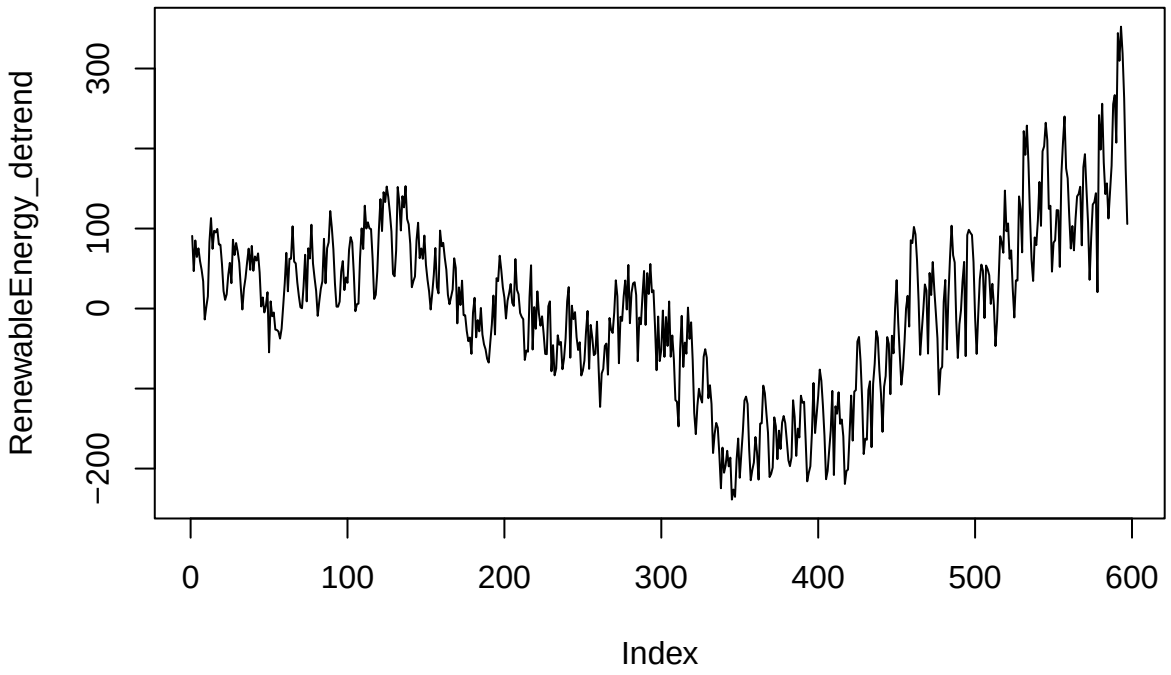
```
num.obs <- nrow(energy_data)
observ <- 1:num.obs

#run linear model
linear_renewable<- lm(energy_data[,2] ~ observ)
summary(linear_renewable)
```

```
##
## Call:
## lm(formula = energy_data[, 2] ~ observ)
##
## 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 ***
## observ        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
```

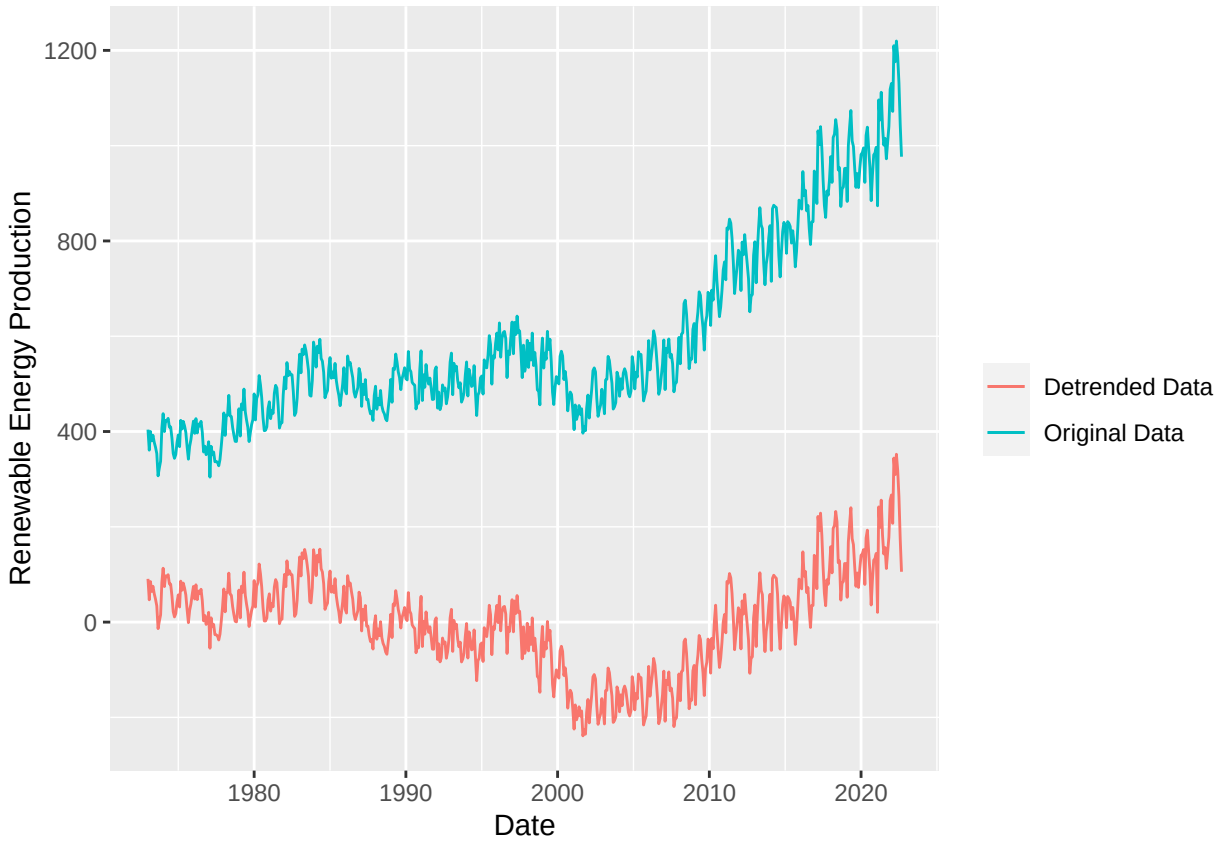
```
#save regression coefficients
beta0_renewable=as.numeric(linear_renewable$coefficients[1])  #first coefficient is the intercept term
beta1_renewable=as.numeric(linear_renewable$coefficients[2])

#Detrend
RenewableEnergy_detrend <- energy_data[,2]-(beta0_renewable+beta1_renewable*observ)
plot(RenewableEnergy_detrend, type = "l")
```




```
ggplot(energy_data, aes(x = Month, y = energy_data[,2], colour = "Original Data", col="blue")) +
  geom_line() +
  ylab(paste0("Detrended Renewable Energy")) +
  geom_line(aes(y=RenewableEnergy_detrend, colour = "Detrended Data"))+
  labs(x = "Date", y = "Renewable Energy Production", color = ' ')
```

```
## Warning: Duplicated aesthetics after name standardisation: colour
```



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.

```
#Remove first row (i.e., January 1973)
energy_data<- energy_data[-c(1), ]
Detrend <- RenewableEnergy_detrend[2:597]

#Data frame - remember to not include January 1973
date <- energy_data[,1]
RenewableEnergyProduction <- energy_data[,2]
energy_df <- data.frame(date ,RenewableEnergyProduction, Detrend, diff)
```

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(energy_df, aes(x = date, y = RenewableEnergyProduction, colour = "Original Data", col="blue")) +
  geom_line() +
  geom_line(aes(y=Detrend, colour = "Detrended Data", col = "green")) +
  geom_line(aes(y=diff, colour = "Differenced Data", col = "red")) +
  labs(x = "Date", y = "Renewable Energy Production", color = ' ')
```

```
## Warning: Duplicated aesthetics after name standardisation: colour
```

```
## Duplicated aesthetics after name standardisation: colour
## Duplicated aesthetics after name standardisation: colour
```

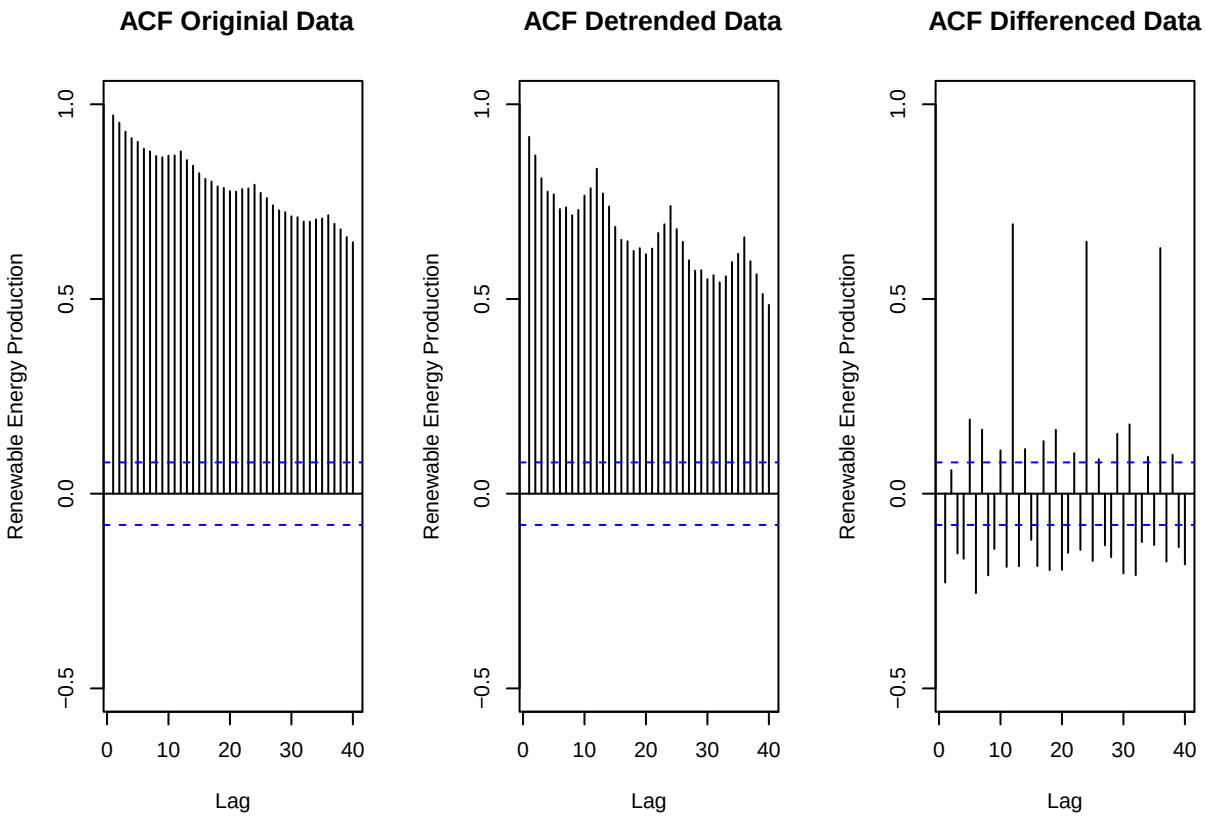
Q5

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?

> Answer: According to the ACF plots displayed below, differencing the original data seems to be a more efficient approach to removing the trend. Small lags that are small and positive and diminish over time are typically associated with a trend which is not seen in the differenced ACF plot compared to the other plots.

```
#Compare ACFs


par(mfrow=c(1,3))
Acf(energy_df[,2], lag.max = 40, ylim=c(-0.5,1), ylab="Renewable Energy Production", main=" ACF Originia
Acf(energy_df[,3], lag.max = 40, ylim=c(-0.5,1), ylab="Renewable Energy Production", main=" ACF Detrende
Acf(energy_df[,4], lag.max = 40, ylim=c(-0.5,1),ylab="Renewable Energy Production", main="ACF Differenc
```



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.

> Answer: A Seasonal Mann-Kendall Test is used to determine whether or not a trend exists in a time series data. Since the results of the Seasonal Mann-Kendall Test had a p-value less than 0.05, the null hypothesis is rejected which indicates a trend present in the data. The ADF test had a p-value almost equal to 1 which indicates a stoichastic trend and differencing (rather than using regression) removed the trend more efficiently as seen in the plots from Q4.

```
#Create a ts variable
ts_energy_data <- ts(df[,5], frequency = 12, start = c(1973,1))
#Mann-Kendall Test
SMKtest <- SeasonalMannKendall(ts_energy_data)
print(summary(SMKtest)) # p-value=<2.22e-16 so we reject the null hypothesis (i.e., no trend)
```

```
## Score =  10577 , Var(Score) = 169001
## denominator =  14553
## tau = 0.727, 2-sided pvalue =< 2.22e-16
## NULL
```

```
#ADF Test
print(adf.test(ts_energy_data,alternative = "stationary")) #p-value = 0.9056 so we accept the null hypo
```

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

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

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.

```
#Create a Year Column
annual_energy_data <- energy_data %>%
  mutate(Year = lubridate::year(energy_data$Month))%>%
  select('Year', 'Total.Renewable.Energy.Production')
#Group by Year for annual data
annual_energy_data <- annual_energy_data %>%
  group_by(Year) %>%
  summarise(RenewableEnergy = mean(Total.Renewable.Energy.Production))
```

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?

> Answer: Both the Mann-Kendall test and the ADF Test display results that are in agreement with Q6 results. Both tests indicate the presence of a trend.

```
#Create a ts variable
ts_annual_energy_data <- ts(annual_energy_data[,2], frequency = 1, start = c(1973))
#Mann-Kendall Test
MKtest <- MannKendall(ts_annual_energy_data)
print(summary(MKtest)) # p-value=<2.22e-16 so we reject the null hypothesis (i.e., no trend)
```

```
## Score =  913 , Var(Score) = 14291.67
## denominator =  1225
## tau = 0.745, 2-sided pvalue =< 2.22e-16
## NULL
```

```
#Spearman Correlation Test
cor.test(annual_energy_data$RenewableEnergy,annual_energy_data$Year, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  annual_energy_data$RenewableEnergy and annual_energy_data$Year
## S = 2548, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8776471
```

```
#ADF Test
print(adf.test(ts_annual_energy_data,alternative = "stationary"))  #p-value = 0.99 so we accept the nul
```

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_annual_energy_data
## Dickey-Fuller = 0.066004, Lag order = 3, p-value = 0.99
## alternative hypothesis: stationary
```