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

Assignment 4 - Due date 02/17/22

Narissa Jimenez-Petchumrus

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.\

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 January 2021 Monthly Energy Review. For this assignment you will work only with the column "Total Renewable Energy Production".

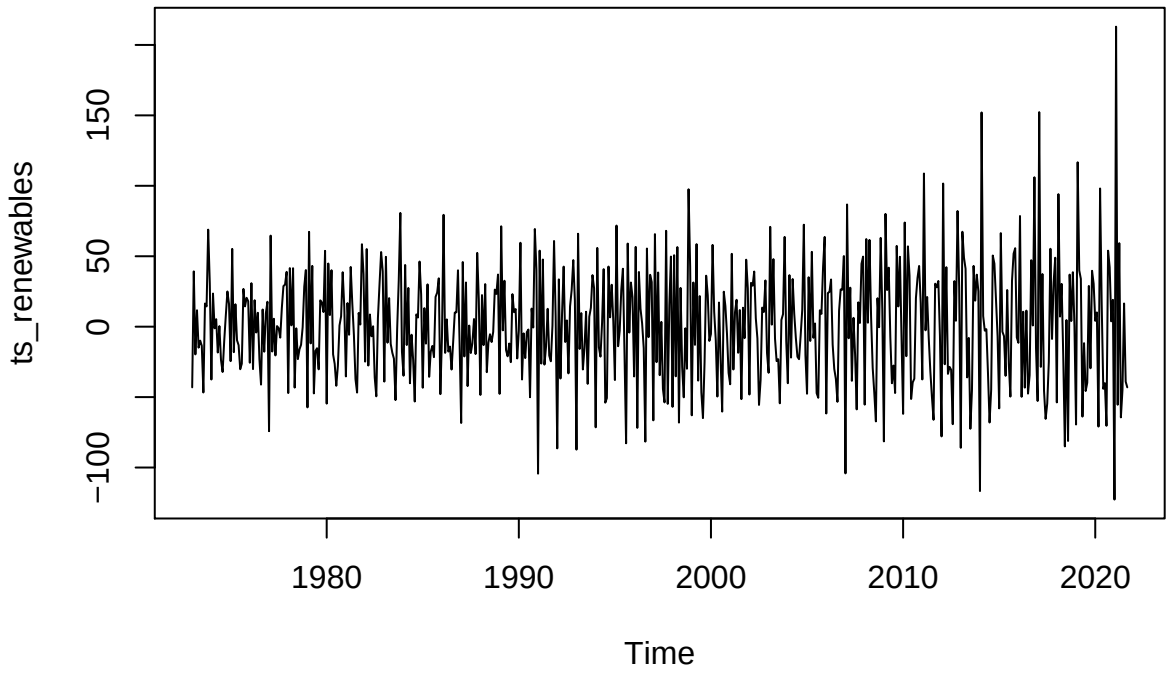
Stochastic Trend and Stationarity Tests

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?

The plot doesn't appear to still have a trend after differencing.



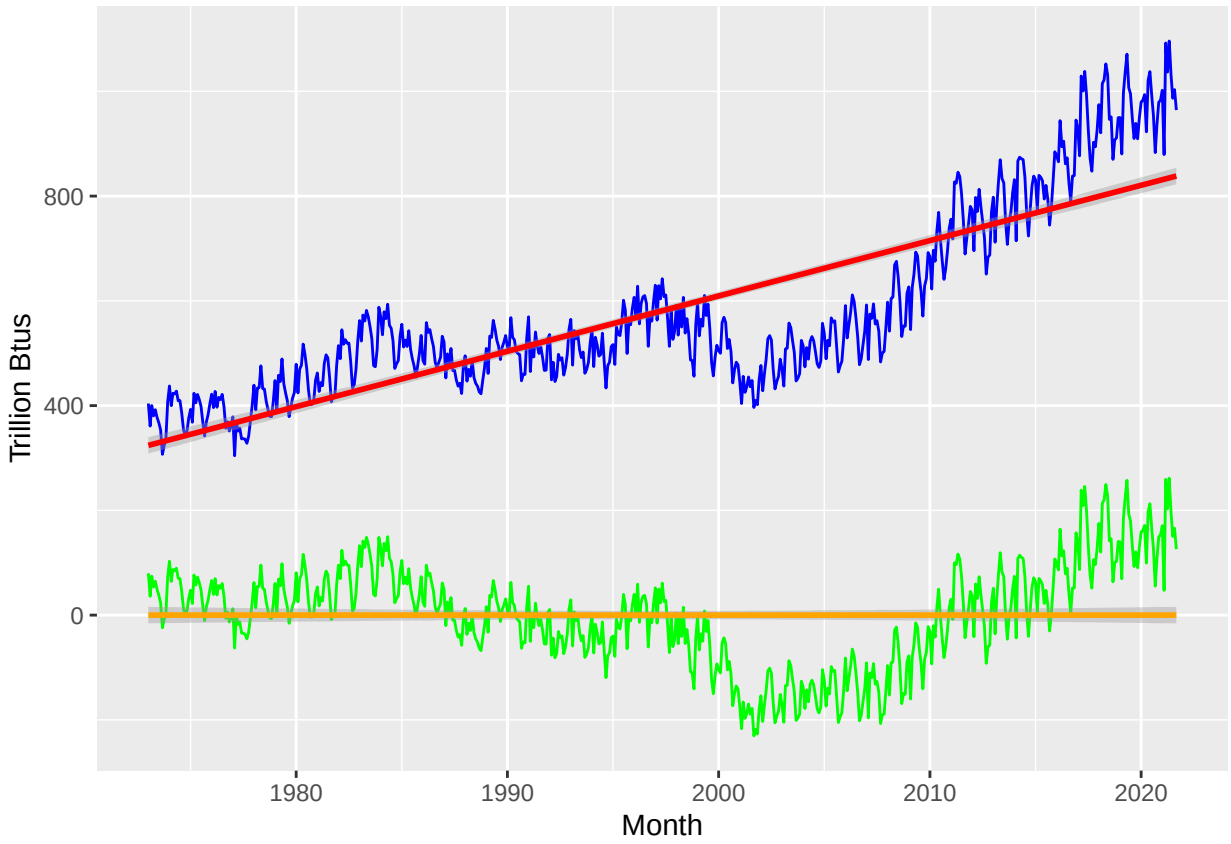
Q2

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

Comparing the plots from Q1 to Q2, the detrended series in green compared to the differenced series from Q1 still displays more of a trend. This could be because using a linear regression isn't the best way to detrend this series as likely the trend isn't deterministic versus stochastic. If the series is stochastic, difference-stationarity is obtained by differencing the series. This was done in the plot in Q1, which removed more trend than the plot in Q2.

```
##
## Call:
## lm(formula = Total_Renewable_Energy_Production ~ t, data = energy_data2)
##
## 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
```

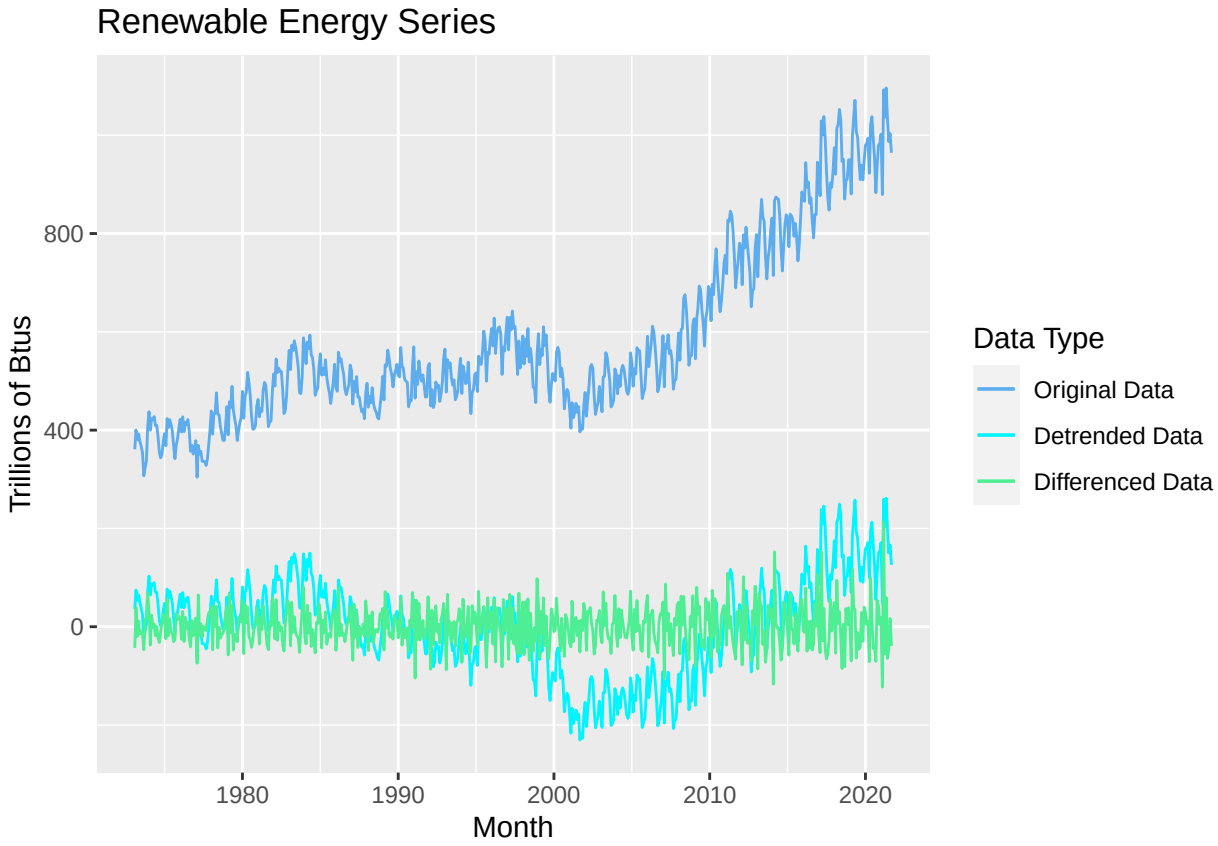


Q3

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.

Q4

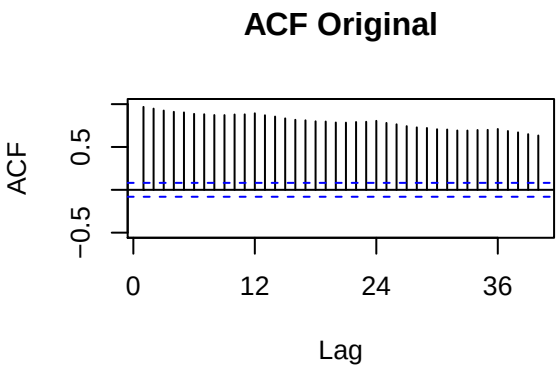
Using ggplot() create a line plot that shows the three series together. Make sure you add a legend to the plot.

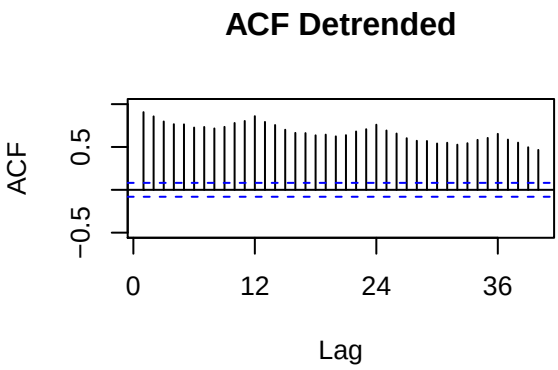


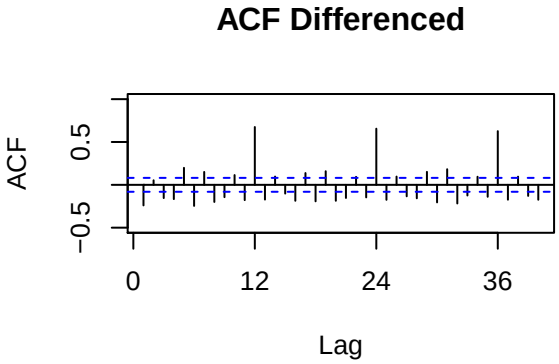
Q5

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?

Looking at the ACFs, the differencing was the best at eliminating the trend as both the original and detrended ACFs we can see the lags decrease while the differenced ACF we don't see a trend. The detrended ACF also displays seasonality while the ACF of the differenced data doesn't show seasonality. In Q4, we can also see that the differenced data's mean is closest to zero compared to the original and detrended data.







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.

According to the Seasonal Mann-Kendall test, the p-value being way below my self-declared critical value of 0.05 means that we can reject the null hypothesis and embrace the alternative hypothesis. This means that there is a trend within the data.

According to the ADF Test, I cannot reject the null hypothesis since my p-value is above the critical value of 0.05 (it's 0.8161). This implies that there's a unit root, thus the series does have a stochastic trend. This does match Q2's plot of the regular data series (the blue line that hasn't been detrended).

```
## tau = 0.715, 2-sided pvalue =< 2.22e-16
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_original
## Dickey-Fuller = -1.4383, Lag order = 8, p-value = 0.8161
## 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.

```
#Group data in yearly steps instances

orig_renewables_drop2021_df<-orig_renewables_df[1:575,]
tail(orig_renewables_drop2021_df)
```

```
##          Month Total_Renewable_Energy_Production
## 571 2020-07-01                           993.568
## 572 2020-08-01                           953.474
## 573 2020-09-01                           883.110
## 574 2020-10-01                           937.063
## 575 2020-11-01                           979.210
## 576 2020-12-01                           982.997
```

```
ts_renewables_drop2021<- as.ts(orig_renewables_drop2021_df[,2])
head(ts_renewables_drop2021)
```

```
## Time Series:
## Start = 1
## End = 6
## Frequency = 1
## [1] 360.900 400.161 380.470 392.141 377.232 367.325
```

```
ts_renewables_matrix<-matrix(ts_renewables_drop2021,byrow=FALSE,nrow=12)
```

```
## Warning in matrix(ts_renewables_drop2021, byrow = FALSE, nrow = 12): data length
## [575] is not a sub-multiple or multiple of the number of rows [12]
```

```
renewable_orig_yearly <- colMeans(ts_renewables_matrix)
renewable_orig_yearly
```

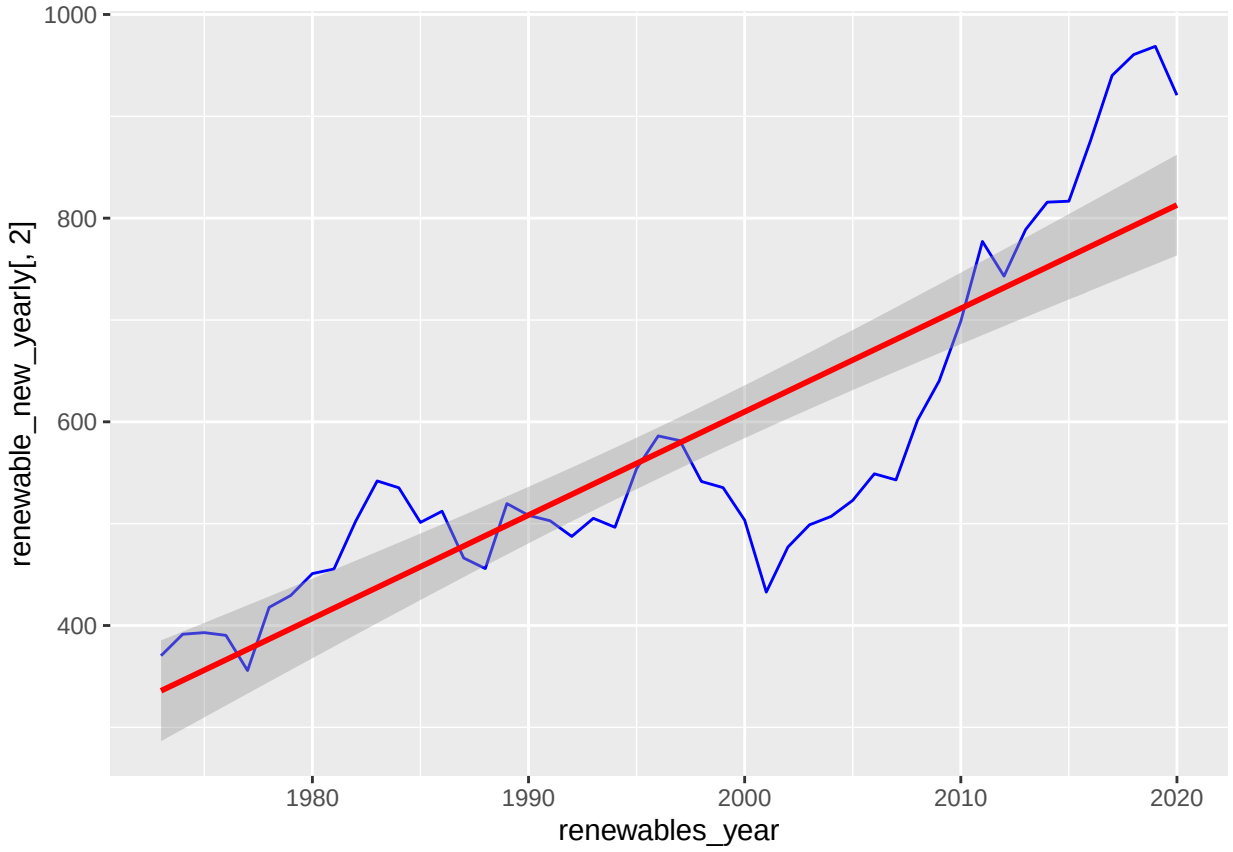
```
##  [1] 370.3686 391.4284 393.0117 390.3247 355.7858 417.8545 429.5001 450.9817
##  [9] 455.4789 502.2897 541.9512 535.2488 501.2395 512.1949 466.3062 455.8974
## [17] 519.7318 508.1857 502.8262 487.4662 505.2767 496.4661 553.9544 586.1697
## [25] 581.5448 541.4457 535.3834 503.4572 432.7787 477.0373 498.8860 507.1209
## [33] 522.8884 548.9600 542.9797 601.7227 640.2355 698.7537 777.1252 743.0914
## [41] 788.8461 815.7195 816.6374 876.0042 940.1299 960.6251 968.7218 920.7433
```

```
renewables_year <- c(year(first(orig_renewables_drop2021_df[,1])):year(last(orig_renewables_drop2021_df

renewable_new_yearly <- data.frame(renewables_year, renewable_orig_yearly)

ggplot(renewable_new_yearly, aes(x=renewables_year, y=renewable_new_yearly[,2])) +
            geom_line(color="blue") +
            geom_smooth(color="red",method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



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?

Comparing the Seasonal Mann-Kendall of the non-aggregated series (Q6) and aggregated series are in agreement as in both cases, we reject the null hypothesis (that the series is stationary) and we embrace the alternative hypothesis that the series follows a trend.

The Spearman Correlation coefficient is clearly above 0 (it's 0.863439) and with a very low p-value below 0.05, which implies that true rho isn't equal to zero or that there's a trend.

The ADF results of the non-aggregated (Q6) and aggregated series are in agreement as in both cases, we accept the null hypothesis as the p-value is above 0.05. This means that the series contains a unit root, thus the series does have a trend.

```
## [1] "Results for Seasonal Mann Kendall /n"
```

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

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

```
## Score =  812 , Var(Score) = 12658.67
## denominator =  1128
## tau = 0.72, 2-sided pvalue =< 2.22e-16
## NULL
```

```
## [1] "Results from Spearman Correlation"
```

```
## [1] 0.863439
```

```
##
##  Spearman's rank correlation rho
##
## data:  renewable_orig_yearly and renewables_year
## S = 2516, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.863439
```

```
## [1] "Results for ADF test/n"
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_original
## Dickey-Fuller = -1.4383, Lag order = 8, p-value = 0.8161
## alternative hypothesis: stationary
```

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  renewable_orig_yearly
## Dickey-Fuller = -1.5617, Lag order = 3, p-value = 0.7491
## alternative hypothesis: stationary
```