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

Assignment 4 - Due date 02/17/22

Alex Baad

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.\

```
#Load/install required package here
library(ggplot2)
library(readxl)
library(forecast)
```

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

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

```
##
## Attaching package: 'lubridate'

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

```
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_Consumpti
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
TREP_data <- read_xlsx(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xls
```

```
## New names:
## * `` -> ...1
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * ...
```

```
TREP_data <- TREP_data[,1:6]
TREP_data <- TREP_data[,-c(2:3)]
TREP_data <- TREP_data[,-c(2,4)]
colnames(TREP_data)=c("Date","TREP")
nenergy <- ncol(TREP_data)
nobs <- nrow(TREP_data)
ts_TREP_data <- ts(TREP_data[,2],frequency=12)
```

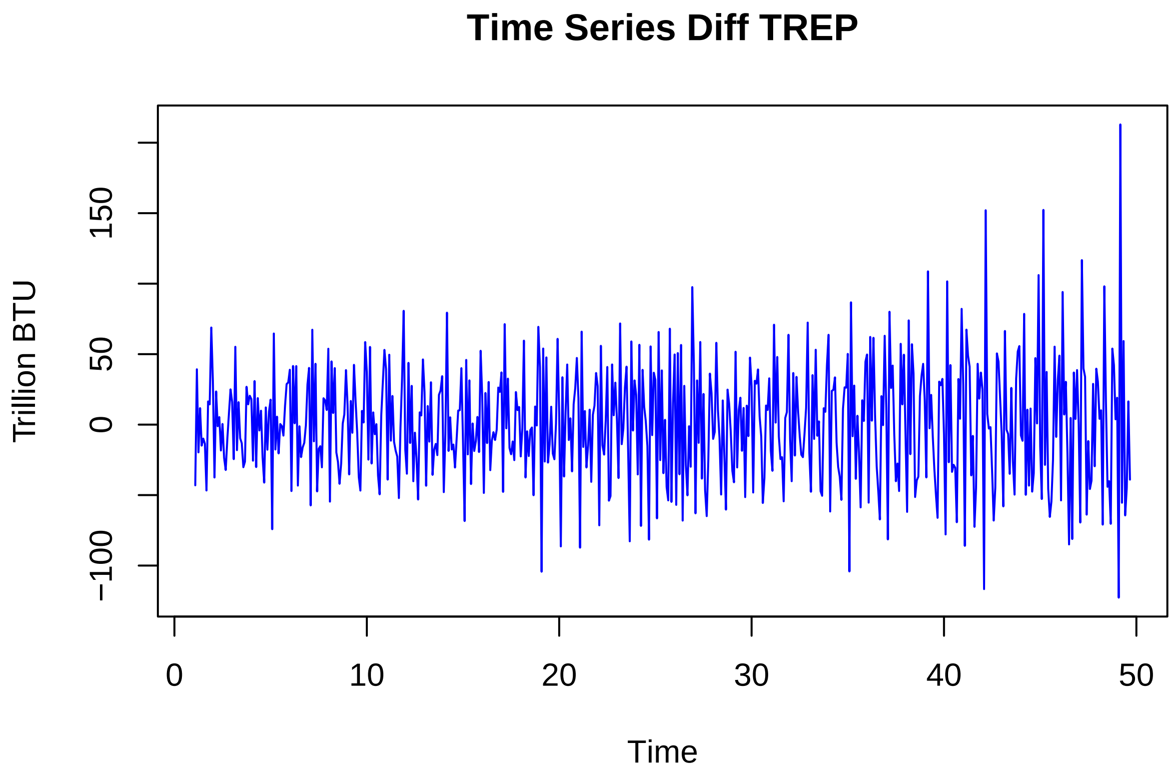
Stochastic Trend and Stationarity Tests

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? Now that the series has been differenced, the TREP data does not seem to have a trend.

```
diff_TREP_data <- diff(ts_TREP_data[,1],lag = 1,difference = 1)
plot(diff_TREP_data,col="blue",ylab="Trillion BTU",main="Time Series Diff TREP")
```



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)

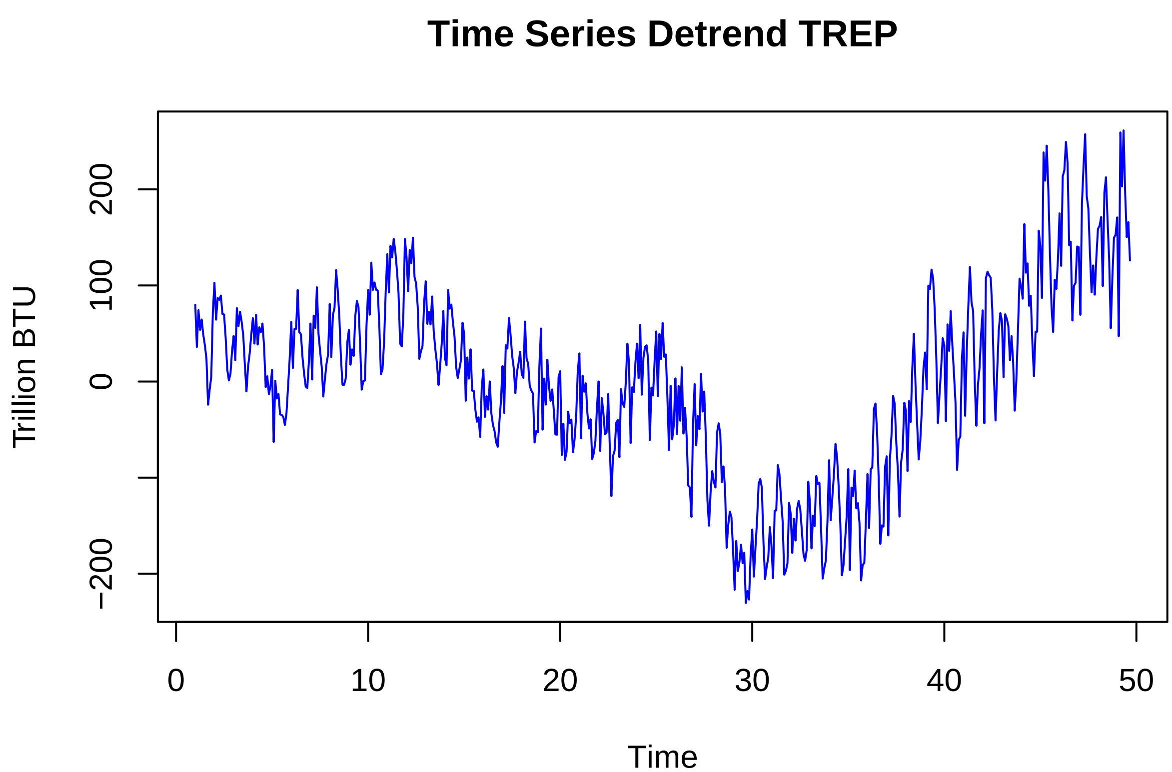
Copy and paste part of your code for A3 where you compute regression for Total Energy Production and the detrended Total Energy Production

```
t <- c(1:nobs)
linear_trend_model_TREP=lm(ts_TREP_data[,1]~t)
summary(linear_trend_model_TREP)
```

```
##
## Call:
## lm(formula = ts_TREP_data[, 1] ~ 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
```

```
beta0_TREP=as.numeric(linear_trend_model_TREP$coefficients[1])
beta1_TREP=as.numeric(linear_trend_model_TREP$coefficients[2])
detrend_energy_data_TREP <- ts_TREP_data[,1]-(beta0_TREP+beta1_TREP*t)
plot(detrend_energy_data_TREP,col="blue",ylab="Trillion BTU",main="Time Series Detrend TREP")
```




```
geom_abline(intercept = beta0_TREP, slope = beta1_TREP, color="red")
```

```
## mapping: intercept = ~intercept, slope = ~slope
## geom_abline: na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

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.

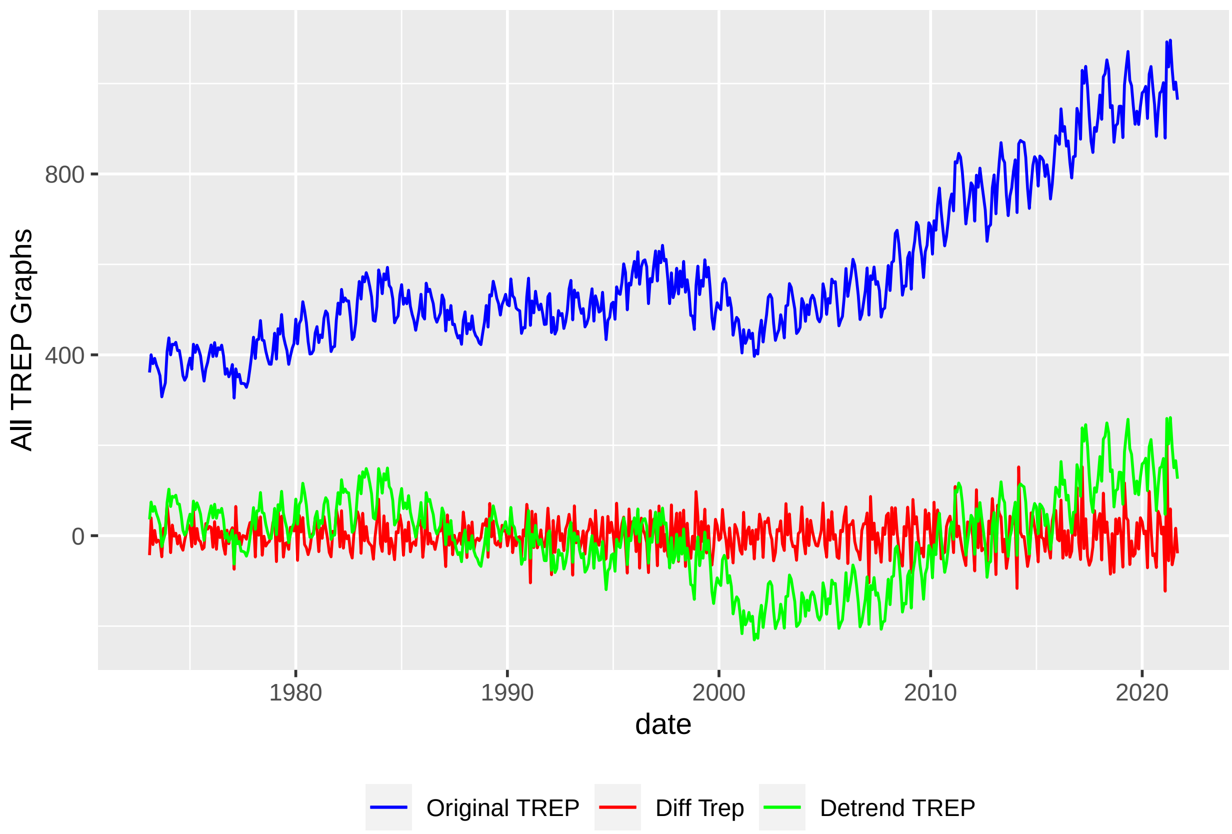
```r
#Data frame - remember to note include January 1973
TREP_data <- TREP_data[-c(1),]
df_detrend_TREP <- as.data.frame(detrend_energy_data_TREP)
detrend_energy_data_TREP <- df_detrend_TREP[-c(1),]
df_detrend_TREP <- as.data.frame(detrend_energy_data_TREP)
diff_TREP_data <- as.data.frame(diff_TREP_data)
colnames(diff_TREP_data)=c("TREP")
TREP_all_data <- data.frame("date"=TREP_data$Date,"original TREP"=TREP_data$TREP,"Diff TREP"=diff_TREP_d
str(TREP_all_data)
```

```
## 'data.frame':    584 obs. of  4 variables:
##  $ date         : POSIXct, format: "1973-02-01" "1973-03-01" ...
##  $ original.TREP: num  361 400 380 392 377 ...
##  $ Diff.TREP    : Time-Series  from 1.08 to 49.7: -43.1 39.3 -19.7 11.7 -14.9 ...
##  $ Detrend.TREP : num  36 74.3 53.8 64.6 48.8 ...
```

Q4

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

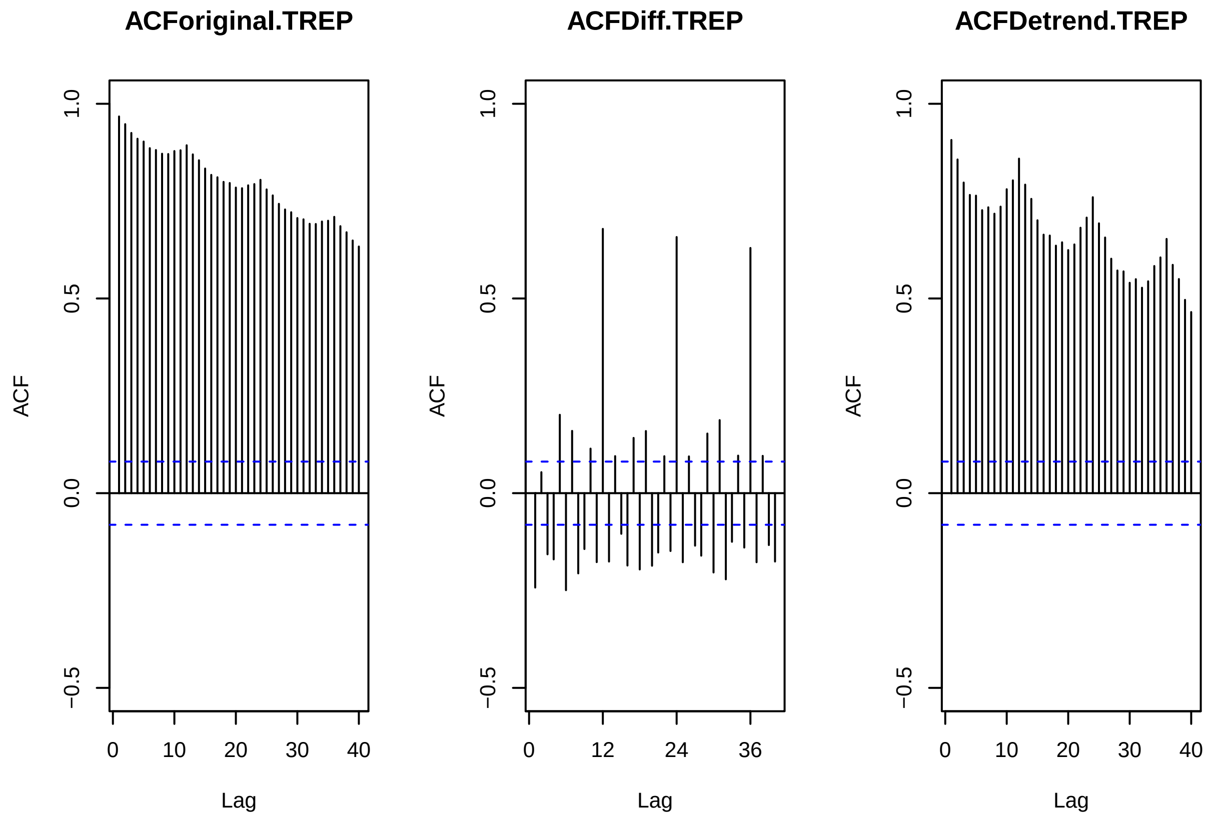
```r
#Use ggplot
nTREP <- ncol(TREP_all_data)

par(mfrow=c(1,3))

ggplot(as.data.frame(TREP_all_data), aes(x=date)) +
  geom_line(aes(y=original.TREP,col="original.TREP")) +
  geom_line(aes(y=Diff.TREP,col="Diff.TREP")) +
  geom_line(aes(y=Detrend.TREP,col="Detrend.TREP")) +
  labs(color="") +
  scale_color_manual(values = c("original.TREP" = "blue", "Diff.TREP" = "red","Detrend.TREP" = "green")
                                labels=c("Original TREP", "Diff Trep", "Detrend TREP")) +
  theme(legend.position = "bottom") +
  ylab(label="All TREP Graphs")
```



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? The differencing process has done a better job at removing the trend as there is still a seasonal trend with the linear regression ACF graph.

```
#Compare ACFs
par(mfrow=c(1,3))
for(i in 2:(nTREP)){
  Acf(TREP_all_data[,i],lag.max=40,main=paste("ACF",colnames(TREP_all_data)[(i)],sep=""),ylim=c(-0.5,1))
}
```



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. For the seasonal Mann-Kendall test, the p-value is below 0.05 therefore we reject the null hypothesis that there is no seasonal trend. For the ADF, the p-value is greater than 0.05 so you do not reject the null hypothesis that the series does include a unit root which means that the series is stochastic and has a trend. This does match what I observed in the original trend that there is a seasonal trend in the original TREP data.

```
SMK_TREP <- SeasonalMannKendall(ts_TREP_data[,1])
print("Results for Seasonal Mann Kendall /n")
```

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

```
print(summary(SMK_TREP))
```

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

```
print("Results for ADF test/n")
```

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

```
print(adf.test(ts_TREP_data[,1]),alternative = "stationary")
```

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

```
TREP_data_matrix <- matrix(ts_TREP_data[,1],byrow=FALSE,nrow=12)
```

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

```
TREP_data_yearly <- colMeans(TREP_data_matrix)
my_year <- c(year(first(TREP_all_data[,1])):year(last(TREP_all_data[,1])))
```

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? For the Mann-Kendall and the ADF tests both still show that there is a seasonal trend even with the yearly aggregated data (p is below 0.05 in MK, and p is above 0.05 in ADF). For the spearman correlation, the correlation is closer to 1, which means that there is a trend.

```
SMK_TREP_yearly <- MannKendall(TREP_data_yearly)
print("Results for Yearly Mann Kendall /n")
```

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

```
print(summary(SMK_TREP_yearly))
```

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

```r
print("Results from Spearman Correlation")
```

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

```r
sp_TREP=cor(TREP_data_yearly,my_year,method="spearman")
print(sp_TREP)
```

```
## [1] 0.8684694
```

```r
print("Results for ADF test on yearly data/n")
```

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

```r
print(adf.test(TREP_data_yearly, alternative = "stationary"))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  TREP_data_yearly
## Dickey-Fuller = -2.2085, Lag order = 3, p-value = 0.4907
## alternative hypothesis: stationary
```