ENV 797 - Time Series Analysis for Energy and Environment Applications | Spring 2025

Assignment 4 - Due date 02/11/25

Justin Maynard

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_Sp25.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)
library(tseries)
library(Kendall)
library(trend)
library(cowplot)
```

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 - you may copy your code from A3
rm(list=ls())

data <- read_excel("Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
                   skip = 12, col_names = FALSE)
```

```r
names <- read_excel("Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
                    skip = 10, n_max = 1, col_names = F)
colnames(data) <- names

data <- data[,c("Month", "Total Renewable Energy Production")]

ts <- ts(data[,2], start = c(1973,1,1), end = c(2023,12,1), frequency = 12)
```

Stochastic Trend and Stationarity Tests

For this part you will work only with the column Total Renewable Energy Production.

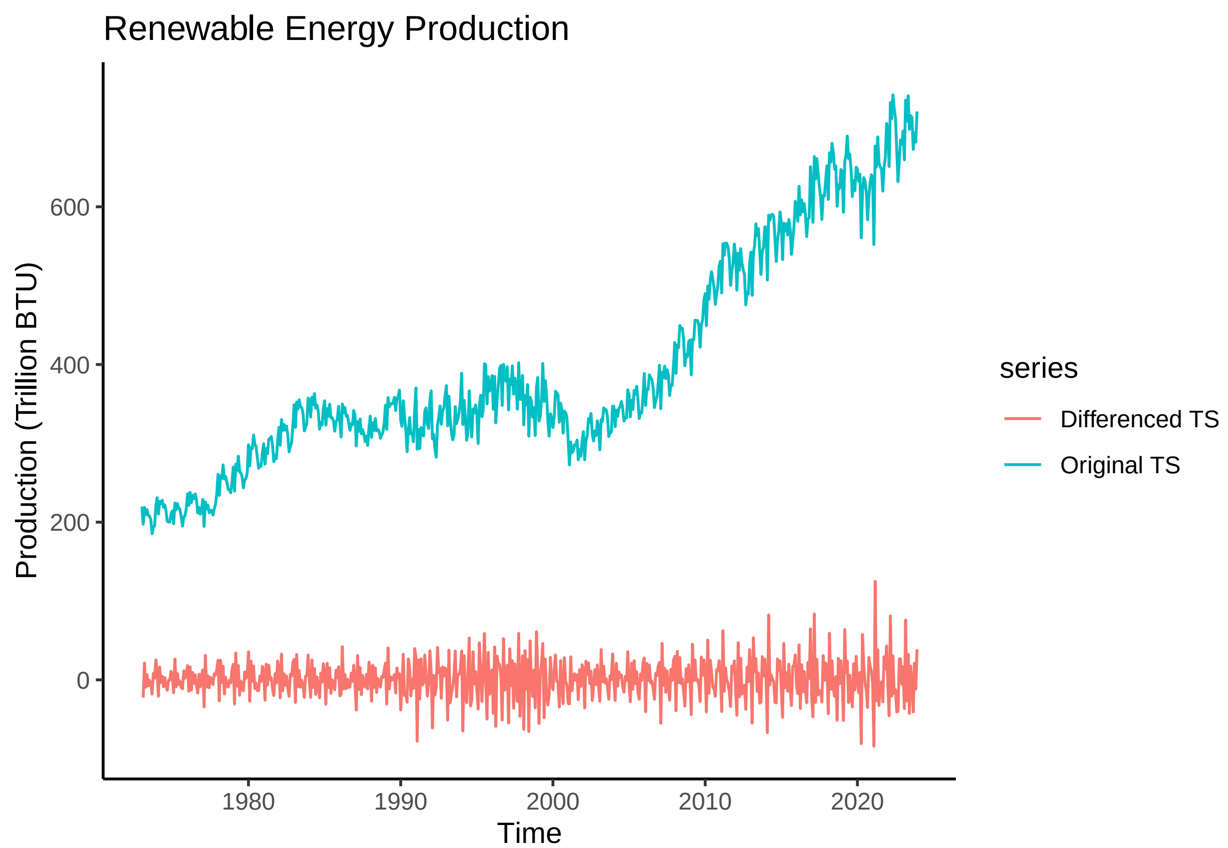
Q1

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
diff <- diff(x = ts, lag = 1, differences = 1)

autoplot(ts, series = 'Original TS') +
autolayer(diff, series = 'Differenced TS') + theme_classic() +
  ylab("Production (Trillion BTU)") +
  ggtitle("Renewable Energy Production") +
  theme_classic()
```



Q2

Copy and paste part of your code for A3 where you run the regression for Total Renewable Energy Production and subtract that from the original series. This should be the code for Q3 and Q4. make sure you use the same name for you time series object that you had in A3, otherwise the code will not work.

```
#Getting rid of 2024 data as it is not complete
nobs <- 612 #nrow(ts)
t <- c(1:nobs)

linear_trend_model_renew <- lm(ts ~ t)

print(summary(linear_trend_model_renew))
```

```
##
## Call:
## lm(formula = ts ~ t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -148.76  -36.24   12.25   41.49  142.27
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 180.24534    4.89680   36.81   <2e-16 ***
```

```
## t              0.70757    0.01384   51.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.5 on 610 degrees of freedom
## Multiple R-squared:  0.8107, Adjusted R-squared:  0.8104
## F-statistic:  2613 on 1 and 610 DF,  p-value: < 2.2e-16
```

```
intercept_renew <- linear_trend_model_renew$coefficients[1]
t_renew <- linear_trend_model_renew$coefficients[2]

linear_trend_renew <- intercept_renew + t_renew * t
ts_linear_renew <- ts(linear_trend_renew,start=c(1973,1,1),frequency=12)

detrend_renew <- ts - linear_trend_renew
ts_detrend_renew <- ts(detrend_renew, start = c(1973,1,1),frequency = 12)
```

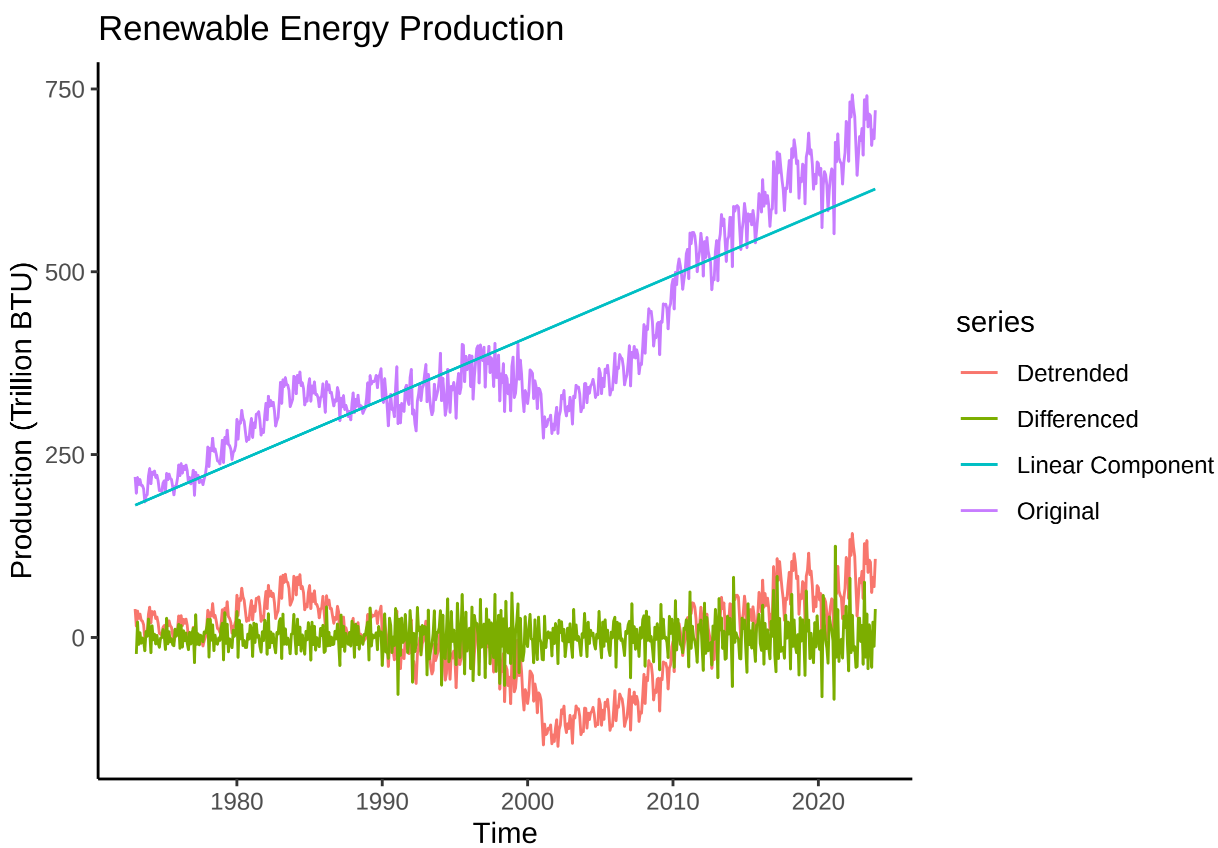
Q3

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 Q2 using linear regression.

Using autoplot() + autolayer() create a plot that shows the three series together. Make sure your plot has a legend. The easiest way to do it is by adding the `series=` argument to each autoplot and autolayer function. Look at the key for A03 for an example on how to use autoplot() and autolayer().

What can you tell from this plot? Which method seems to have been more efficient in removing the trend?

```
autoplot(ts, series = "Original")+
  autolayer(ts_detrend_renew,series="Detrended")+
  autolayer(ts_linear_renew,series="Linear Component") +
  autolayer(diff, series = "Differenced") +
  ylab("Production (Trillion BTU)") +
  ggtitle("Renewable Energy Production") +
  theme_classic()
```

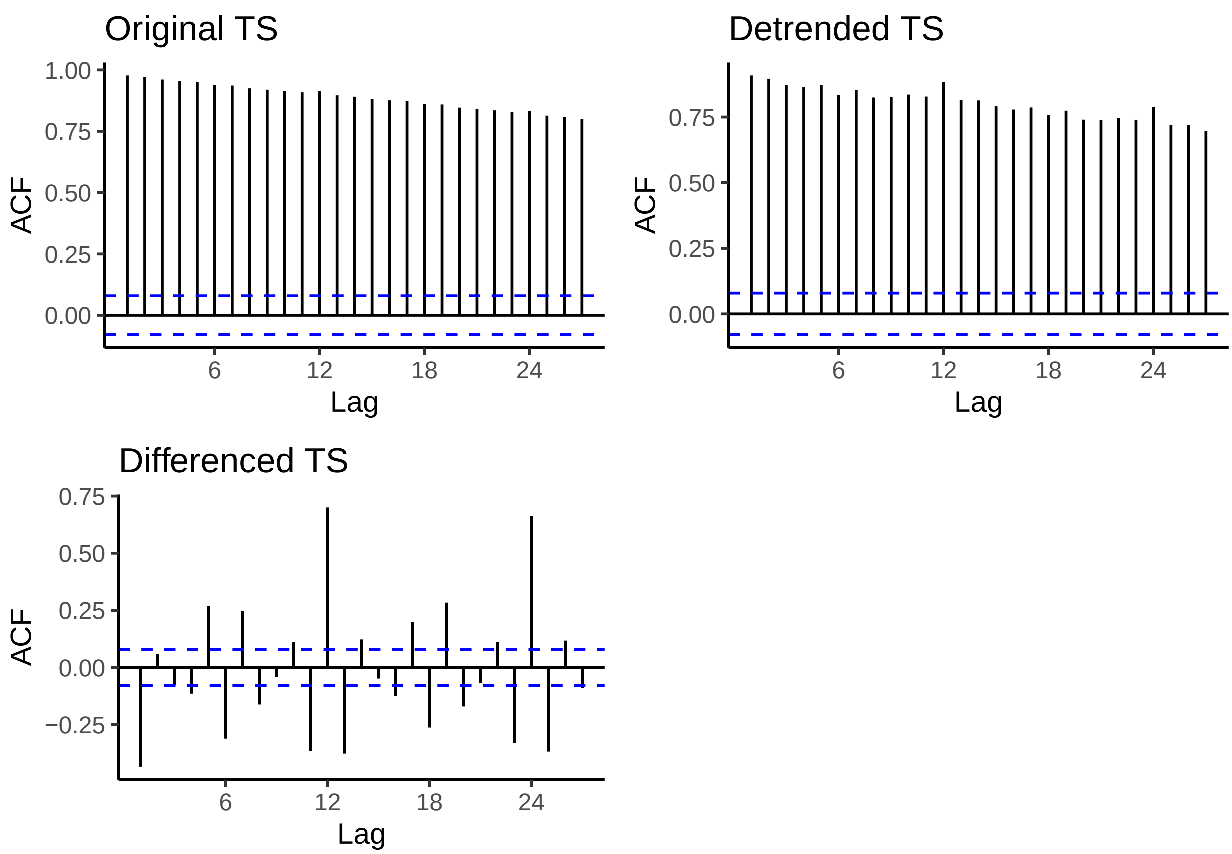


Answer: Based on the plot, the differencing method appears to be better at removing the trend, as the time series remains centered around zero.

Q4

Plot the ACF for the three series and compare the plots. Add the argument `ylim=c(-0.5,1)` to the autoplot() or Acf() function - whichever you are using to generate the plots - to make sure all three y axis have the same limits. Looking at the ACF which method do you think was more efficient in eliminating the trend? The linear regression or differencing?

```
plot_grid(ts_graph, ts_detrend_graph, ts_diff_graph)
```



Answer: The differencing was the most efficient in eliminating the trend, as the ACF values are far lower. The detrended ACF values are not much of an improvement compared to the original time series.

Q5

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. What is the conclusion from the Seasonal Mann Kendall test? What's the conclusion for the ADF test? Do they match what you observed in Q3 plot? Recall that having a unit root means the series has a stochastic trend. And when a series has stochastic trend we need to use differencing to remove the trend.

```r
print(summary(smk.test(ts)))
```

```
##
##  Seasonal Mann-Kendall trend test (Hirsch-Slack test)
##
## data: ts
## alternative hypothesis: two.sided
##
## Statistics for individual seasons
##
## H0
##                       S    varS   tau     z   Pr(>|z|)
```

```
## Season 1:   S = 0   989 15158.3 0.776 8.025 1.0174e-15 ***
## Season 2:   S = 0   987 15158.3 0.774 8.009 1.1612e-15 ***
## Season 3:   S = 0   973 15158.3 0.763 7.895 2.9081e-15 ***
## Season 4:   S = 0   975 15158.3 0.765 7.911 2.5526e-15 ***
## Season 5:   S = 0   975 15158.3 0.765 7.911 2.5526e-15 ***
## Season 6:   S = 0   991 15158.3 0.777 8.041 8.9116e-16 ***
## Season 7:   S = 0  1021 15158.3 0.801 8.285 < 2.22e-16 ***
## Season 8:   S = 0  1039 15158.3 0.815 8.431 < 2.22e-16 ***
## Season 9:   S = 0  1015 15158.3 0.796 8.236 < 2.22e-16 ***
## Season 10:   S = 0 1023 15158.3 0.802 8.301 < 2.22e-16 ***
## Season 11:   S = 0 1016 15157.3 0.797 8.244 < 2.22e-16 ***
## Season 12:   S = 0 1009 15158.3 0.791 8.187 2.6740e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
##  Seasonal Mann-Kendall trend test (Hirsch-Slack test)
##
## data:  ts
## z = 28.164, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##      S   varS
##  12013 181899
```

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts
## Dickey-Fuller = -1.2251, Lag order = 8, p-value = 0.9024
## alternative hypothesis: stationary
```

> Answer: The results from the Seasonal Mann Kendall test tell us that all seasons are significant, as their P value is <.05. This confirms that seasons do have an influence on the values of the time series. The P value of the ADF test (p value > .05) tell us that we fail to reject the null hypothesis, indicating a unit root is present.

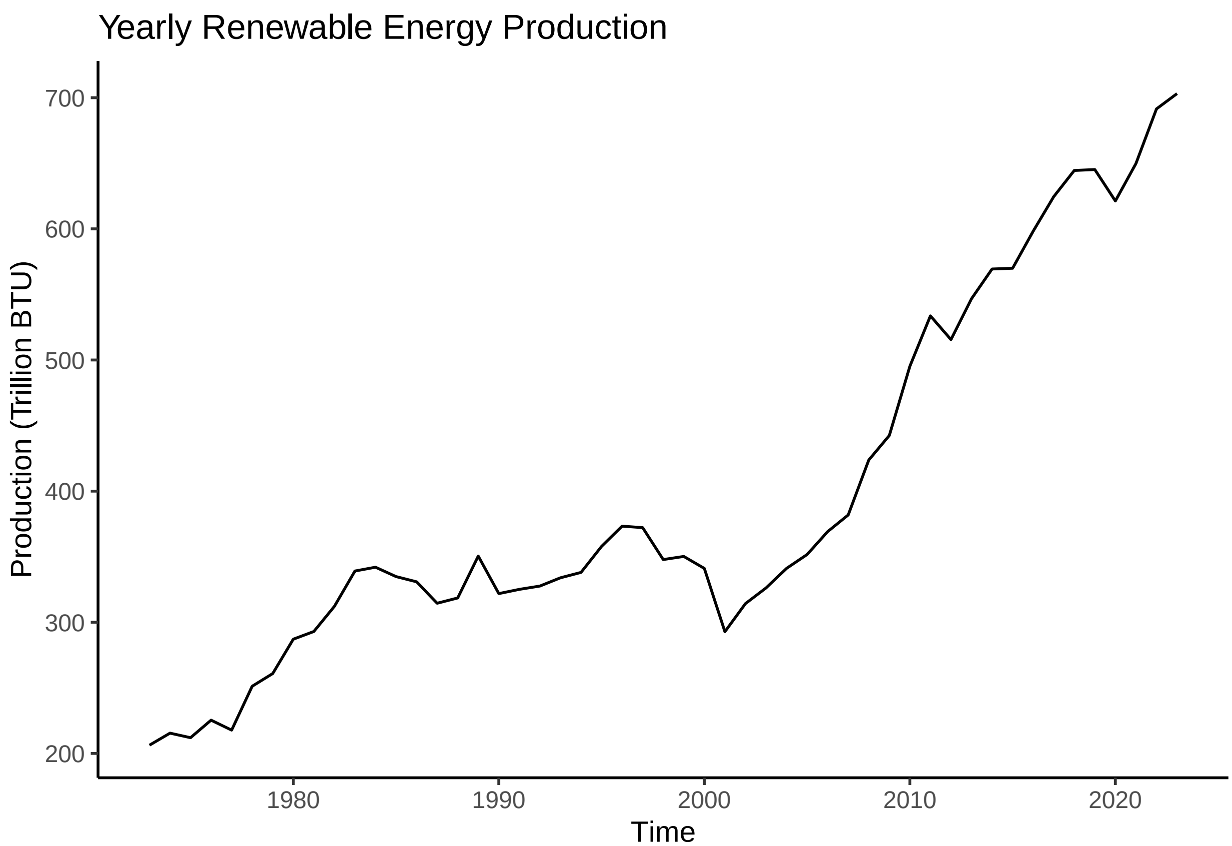
Q6

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. Convert the accumulates yearly series into a time series object and plot the series using autoplot().

```
agg_by_year_matrix <- matrix(ts,byrow=FALSE,nrow=12)
agg_by_year <- colMeans(agg_by_year_matrix)
my_year <- c(1973:2023)

agg_by_year_ts <- ts(agg_by_year, start = 1973, frequency = 1)
```

```r
autoplot(agg_by_year_ts) + theme_classic() + ggtitle("Yearly Renewable Energy Production") +
ylab("Production (Trillion BTU)")
```



Q7

Apply the Mann Kendall, Spearman correlation rank test and ADF. Are the results from the test in agreement with the test results for the monthly series, i.e., results for Q6?

```r
print(cor.test(agg_by_year_ts,my_year,method="spearman"))
```

```
##
##  Spearman's rank correlation rho
##
## data:  agg_by_year_ts and my_year
## S = 1852, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9161991
```

```r
print(summary(MannKendall(agg_by_year_ts)))
```

```
## Score =  1033 , Var(Score) = 15158.33
```

```
## denominator =  1275
## tau = 0.81, 2-sided pvalue =< 2.22e-16
## NULL
```

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  agg_by_year_ts
## Dickey-Fuller = -1.0646, Lag order = 3, p-value = 0.9196
## alternative hypothesis: stationary
```

> Answer: The results of the Spearman correlation test (p value $<$ .05) tell us that the time series is not stationary. This is in agreement from failing to reject the null hypothesis from the ADF test in the previous question. The ADF test for both monthly and yearly have p values $>$ .05, and we fail to reject the null hypothesis, thus indicating a unit root is present. The p value from the MK test is $<$ .05, meaning we reject the null hypothesis, and the time series follows a trend. This also aligns with our grapical analysis, as the differencing method better removed the trend. This is important as the differencing method must be used when dealing with a stochastic trend, proving that our trend is stochastic and not stationary.