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

Assignment 4 - Due date 02/11/25

Alex Lopez

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.\

```r
#Load/install required package here
library(readxl)
library(ggplot2)
library(forecast)
library(tseries)
library(Kendall)
library(dplyr)
```

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"**.

```r
#Importing data set - you may copy your code from A3
energy_data <-
  read_excel(path = "./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
             skip = 12, sheet = 'Monthly Data', col_names = FALSE)

#extract column names
read_col_names <-
  read_excel(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
             skip = 10,n_max = 1, sheet="Monthly Data",col_names=FALSE)

#assign column names
colnames(energy_data) <- read_col_names

#select columns
energy_df <- energy_data %>%
  select(5)

#transform data frame to time series object
ts_energy <- ts(energy_df, start = c(1973,1), frequency = 12)
```

Stochastic Trend and Stationarity Tests

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

Q1

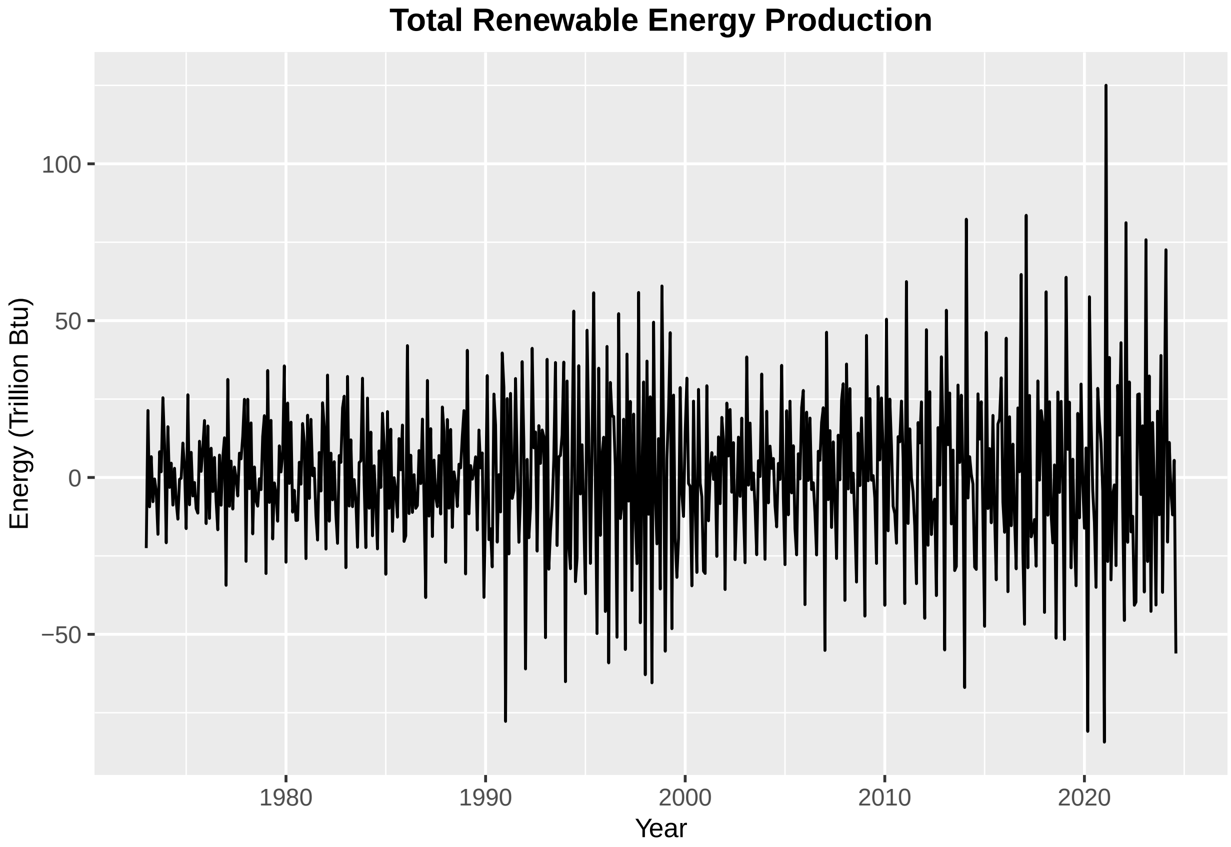
Difference the "Total Renewable Energy Production" series using function diff(). Function diff() is from package base and take three main arguments: * *x* vector containing values to be differenced; * *lag* integer indicating with lag to use; * *differences* integer indicating how many times series should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

```
energy_diff <- diff(ts_energy, lag=1, differences=1)

ts_energy_diff <- ts(energy_diff, start = c(1973,1), frequency=12)

autoplot(ts_energy_diff) +
  ggtitle('Total Renewable Energy Production') +
  xlab('Year') +
  ylab('Energy (Trillion Btu)') +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 12),
          axis.title.x = element_text(size = 10),
          axis.title.y = element_text(size = 10))
```



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.

```r
#create a time vector
nobs <- nrow(ts_energy)
t <- 1:nobs

#fit a linear trend model
linear_trend_m <- lm(ts_energy ~ t)

beta0 <- as.numeric(linear_trend_m$coefficients[1])
beta1 <- as.numeric(linear_trend_m$coefficients[2])

#detrend
linear_trend <- beta0 + beta1 * t
ts_linear <- ts(linear_trend, start=c(1973,1), frequency=12)

ts_detrend <- ts_energy - ts_linear
```

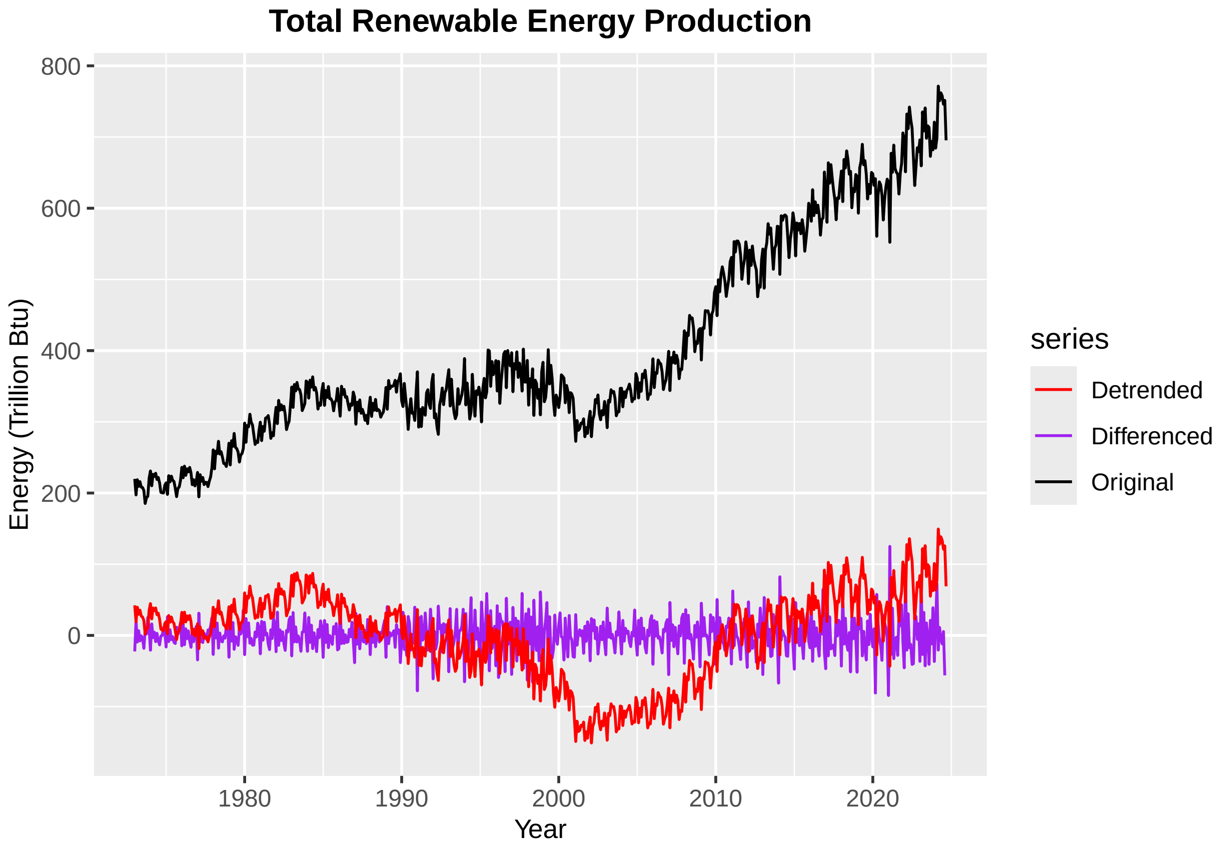
Q3

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?

```r
autoplot(ts_energy, series = 'Original') +
      autolayer(ts_energy_diff, series = 'Differenced') +
      autolayer(ts_detrend, series = 'Detrended') +
      ggtitle('Total Renewable Energy Production') +
      xlab('Year') +
      ylab('Energy (Trillion Btu)') +
      theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 12),
            axis.title.x = element_text(size = 10),
            axis.title.y = element_text(size = 10)) +
            scale_color_manual(values = c('Original' = 'black',
                                          'Differenced' = 'purple',
                                          'Detrended' = 'red'))
```



Answer: The differencing method seems to be more effective in removing the trend. When looking at the plot, there is no trend during any time period of the differenced series. However, the detrended series still shows intervals with clear trends, such as a decreasing trend from around 1995 to around 2000 and an increasing trend from the early 2000s to the early 2010s.

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?

```r
library(cowplot)

#ACF plot for original time series
acf_plot_orig <- autoplot(Acf(ts_energy, lag.max=40, plot=FALSE), ylim = c(-0.5,1)) +
  ggtitle('Original') +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 12),
             axis.title.x = element_text(size = 10),
             axis.title.y = element_text(size = 10))
```

```
## Warning in ggplot2::geom_segment(lineend = "butt", ...): Ignoring unknown
## parameters: ‘ylim‘
```

```r
#ACF plot for differenced series
acf_plot_diff <- autoplot(Acf(ts_energy_diff, lag.max=40, plot=FALSE), ylim = c(-0.5,1)) +
  ggtitle('Differenced') +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 12),
            axis.title.x = element_text(size = 10),
            axis.title.y = element_text(size = 10))
```

```
## Warning in ggplot2::geom_segment(lineend = "butt", ...): Ignoring unknown
## parameters: ‘ylim‘
```

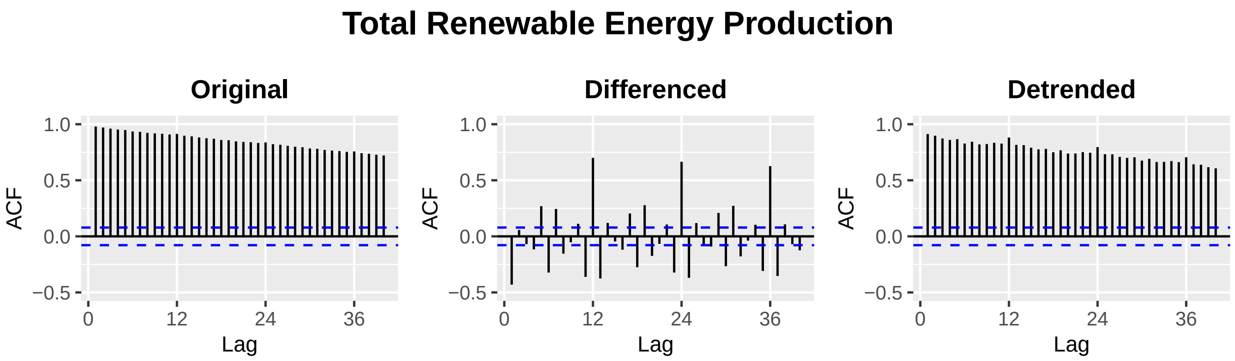
```r
#ACF plot for detrended series
acf_plot_detrend <- autoplot(Acf(ts_detrend, lag.max=40, plot=FALSE), ylim = c(-0.5,1)) +
  ggtitle('Detrended') +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 12),
            axis.title.x = element_text(size = 10),
            axis.title.y = element_text(size = 10))
```

```
## Warning in ggplot2::geom_segment(lineend = "butt", ...): Ignoring unknown
## parameters: ‘ylim‘
```

```r
#arrange the three plots in a row
plot_row <- plot_grid(acf_plot_orig, acf_plot_diff, acf_plot_detrend,
                          nrow = 1, align = "h", rel_widths = c(1, 1, 1))

#title
title <- ggdraw() +
  draw_label("Total Renewable Energy Production", fontface = "bold", size = 15, hjust = 0.5)

#combine title and plot row
print(plot_grid(title, plot_row, ncol = 1, rel_heights = c(0.3, 1)))
```



Answer: Looking at the ACF plots, it looks like differencing was more efficient in eliminating the trend, because the only signifcant correlations visible on the differenced ACF plot occur at regular intervals, with the highest peaks at 12-month intervals. This suggests that, after detrending the series, only the seasonal component of the series remained. In contrast, the detrended ACF plot still shows significant correlations for the entire time period plotted, with all of the lags up to lag 40 having correlations higher than 0.5.

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.

```
#Seasonal Mann-Kendall
summary(SeasonalMannKendall(ts_energy))
```

```
## Score =  12468 , Var(Score) = 190008
## denominator =  15758.5
## tau = 0.791, 2-sided pvalue =< 2.22e-16
```

```
#ADF
print(adf.test(ts_energy, alternative = "stationary"))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_energy
## Dickey-Fuller = -1.0898, Lag order = 8, p-value = 0.9242
## alternative hypothesis: stationary
```

> Answer: According to the Seasonal Mann-Kendall test, since the p-value is less than 0.05, we reject the null hypothesis in favor of the alternative and conclude that the series does follow a trend. According to the ADF test, since the p-value is not less than 0.05, we fail to reject the null hypothesis. We therefore conclude that the series has a unit root, so it has a stochastic trend (it is not stationary). These conclusions match with what we observed in the Q3 plot, because differencing was needed to effectively remove the trend.

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().

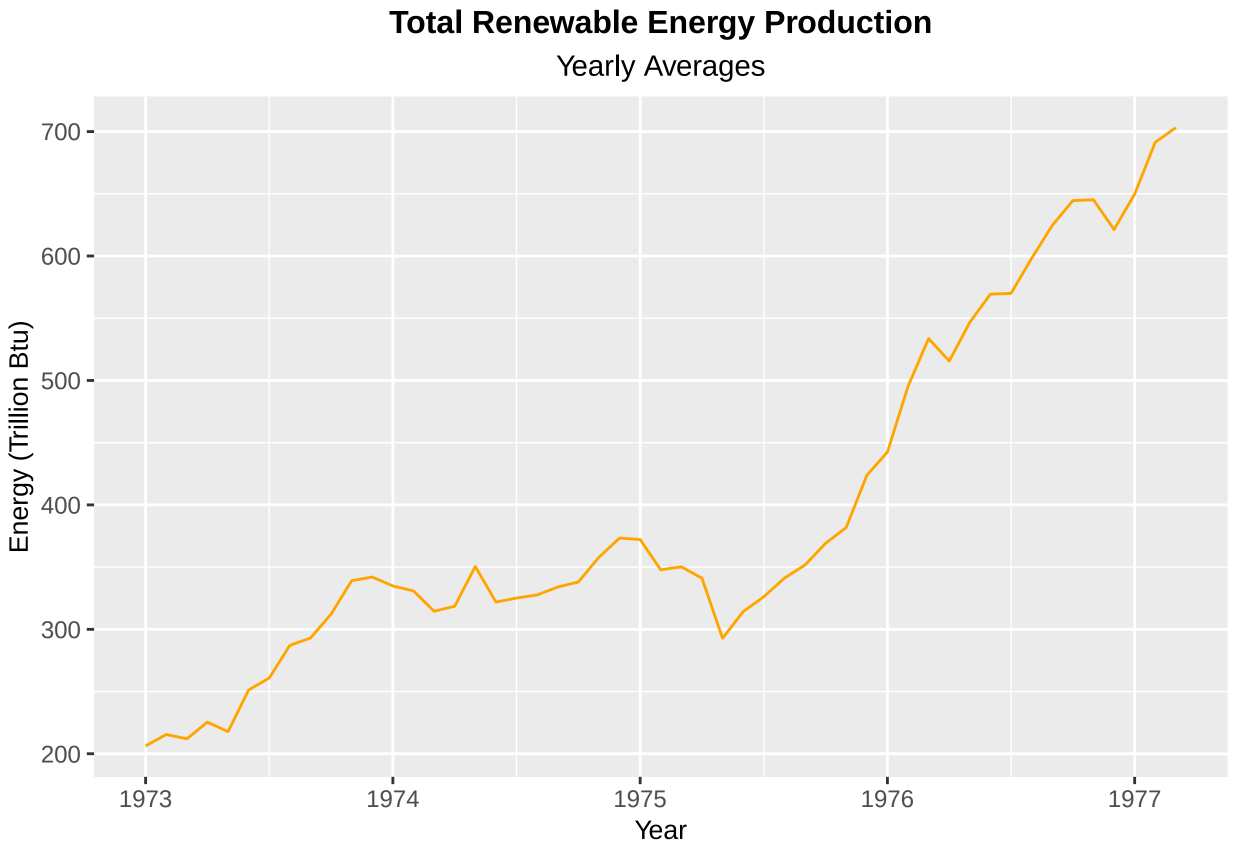
```
#group data in yearly steps instances
energy_matrix <- matrix(ts_energy[1:(12 * (2023 - 1973 + 1))], byrow = FALSE, nrow = 12)
energy_yearly <- colMeans(energy_matrix)

#exclude year 2024, because it's missing monthly figures for Oct, Nov, Dec
num_year <- c(1973:2023)

energy_yearly_df <- data.frame(num_year, 'production' = energy_yearly)

#convert accumulates yearly series into a time series object
ts_energy_yearly <- ts(energy_yearly_df, start = c(1973,1),frequency = 12)
```

```
autoplot(ts_energy_yearly[,2], color = "orange") +
      ggtitle('Total Renewable Energy Production', subtitle = 'Yearly Averages') +
      xlab("Year") +
      ylab("Energy (Trillion Btu)") +
      theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 12),
            plot.subtitle = element_text(hjust = 0.5, size = 11),
            axis.title.x = element_text(size = 10),
            axis.title.y = element_text(size = 10))
```



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?

```
#Mann Kendall
summary(MannKendall(energy_yearly_df$production))
```

```
## Score =  1033 , Var(Score) = 15158.33
## denominator =  1275
## tau = 0.81, 2-sided pvalue =< 2.22e-16
```

```
#Spearman
print(cor.test(energy_yearly_df$production, num_year, method = 'spearman'))
```

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

```
#ADF
print(adf.test(energy_yearly_df$production, alternative = 'stationary'))
```

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

> Answer: According to the Mann Kendall test and the Spearman correlation rank test, the p-values for each of these tests is less than 0.05, so we reject the null hypothesis in favor of the alternative and conclude that the series does follow a trend. Moreover, rho is 0.92 in the Spearman test, which suggests a very high correlation. According to the ADF test, the p-value is not less than 0.05, so we fail to reject the null hypothesis. We therefore conclude that the series has a unit root and a stochastic trend, so it is not stationary. These results are in agreement with those obtained from the monthly series.