ENV 790.30 - Time Series Analysis for Energy Data | Spring 2021

Assignment 4 - Due date 02/25/21

Thomas Hancock

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.

Questions

Consider the same data you used for A2 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.

R packages needed for this assignment:"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(lubridate)
library(tidyverse)
library(forecast)
library(tseries)
library(Kendall)
```

Stochastic Trend and Stationarity Test

For this part you will once again work only with the following columns: Total Biomass Energy Production, Total Renewable Energy Production, Hydroelectric Power Consumption. Create a data frame structure with these three time series and the Date column. Don't forget to format the date object.

```
raw_energy <- read_excel("../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx
raw_energy <- raw_energy[-1,] #Remove row that has units (they are all the same - Trillion BTU)
```

```
energy_small <- raw_energy[,c(4:6)] #Create data frame with subset of data

energy_small <-  data.frame(as.Date(raw_energy$Month), lapply(energy_small, as.numeric))  # Set data to
```

```r
energy_small <- rename(energy_small, Month = as.Date.raw_energy.Month.)

head(energy_small) #Check subset
```

```
##        Month Total.Biomass.Energy.Production Total.Renewable.Energy.Production
## 1 1973-01-01                         129.787                           403.981
## 2 1973-02-01                         117.338                           360.900
## 3 1973-03-01                         129.938                           400.161
## 4 1973-04-01                         125.636                           380.470
## 5 1973-05-01                         129.834                           392.141
## 6 1973-06-01                         125.611                           377.232
##   Hydroelectric.Power.Consumption
## 1                         272.703
## 2                         242.199
## 3                         268.810
## 4                         253.185
## 5                         260.770
## 6                         249.859
```

```r
numobs <- ncol(energy_small)-1

energy_ts <- ts(energy_small[,-1],
                start = c(year(energy_small$Month[1]), month(energy_small$Month[1])),
                frequency = 12)
```

Q1

Now let's try to difference these three series using function diff(). Start with the original data from part (b). Try differencing first at lag 1 and plot the remaining series. Did anything change? Do the series still seem to have trend?

```r
energy_diff <- diff(energy_ts, differences = 1)

head(energy_diff)
```

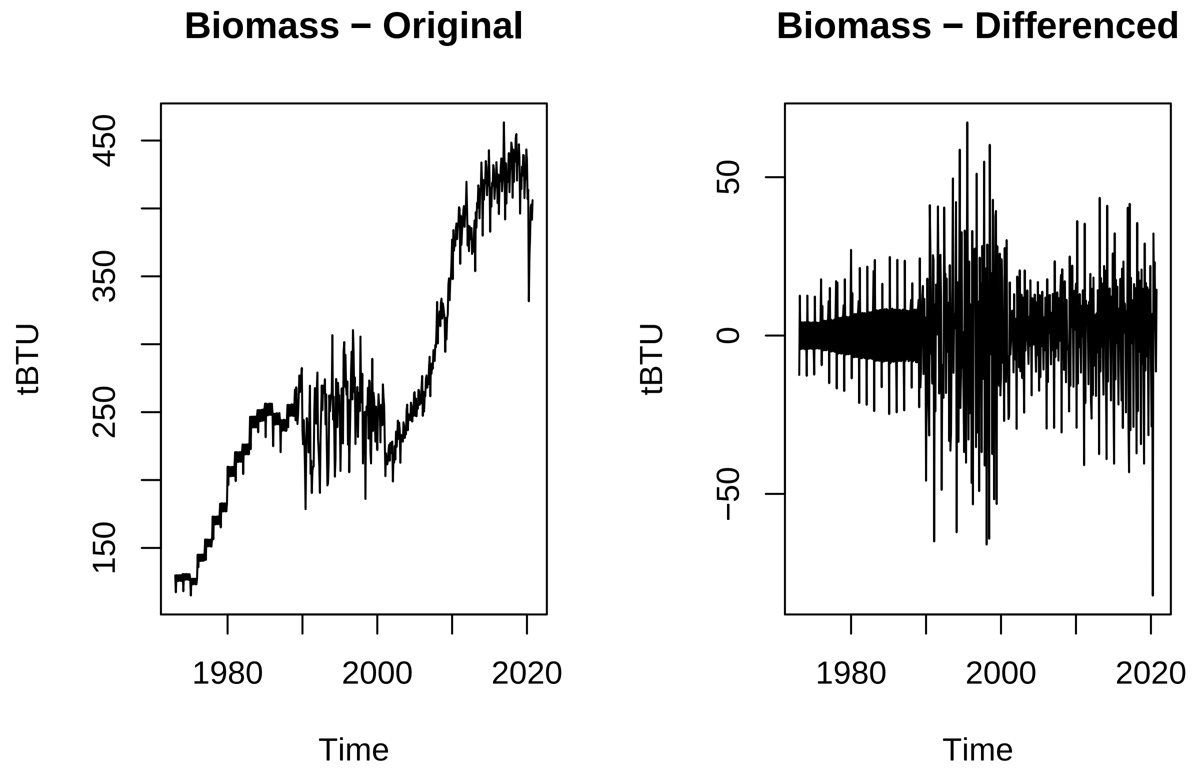
```
##          Total.Biomass.Energy.Production Total.Renewable.Energy.Production
## Feb 1973                         -12.449                           -43.081
## Mar 1973                          12.600                            39.261
## Apr 1973                          -4.302                           -19.691
## May 1973                           4.198                            11.671
## Jun 1973                          -4.223                           -14.909
## Jul 1973                           4.176                            -9.907
##          Hydroelectric.Power.Consumption
## Feb 1973                         -30.504
## Mar 1973                          26.611
## Apr 1973                         -15.625
## May 1973                           7.585
## Jun 1973                         -10.911
## Jul 1973                         -14.189
```

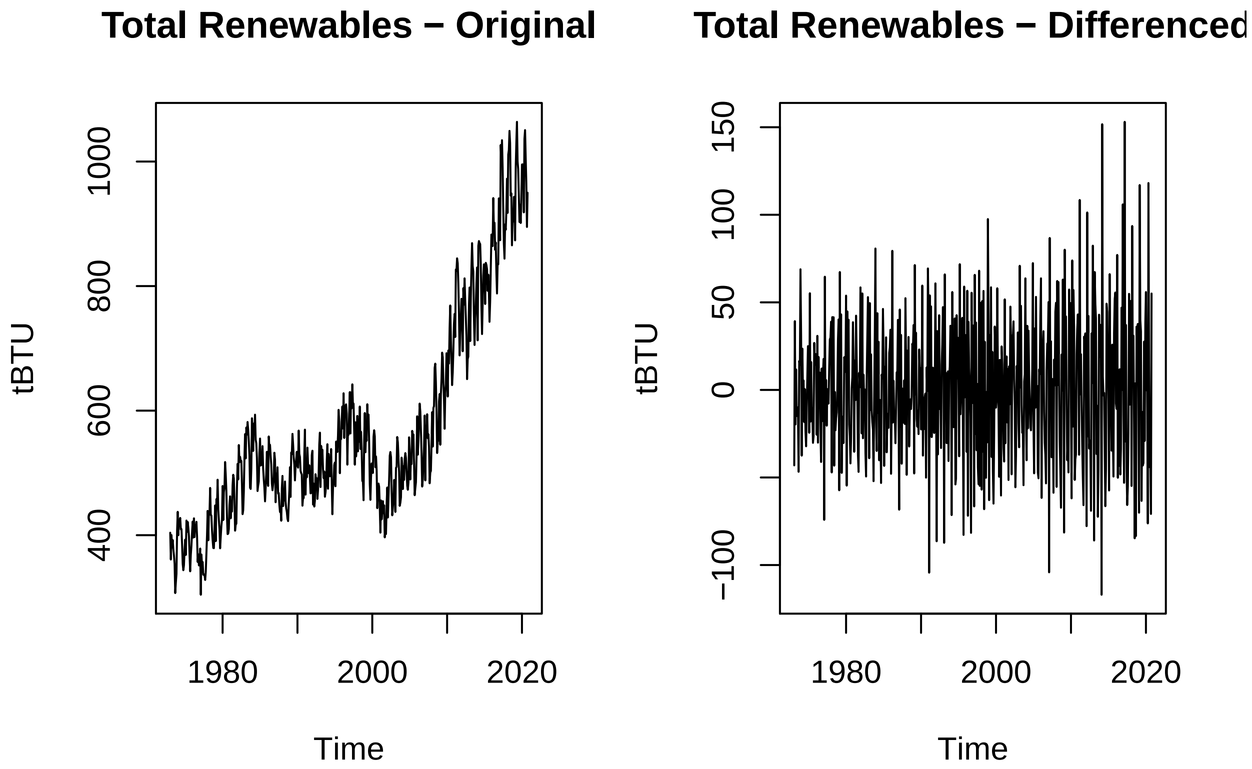
```r
tail(energy_diff)
```

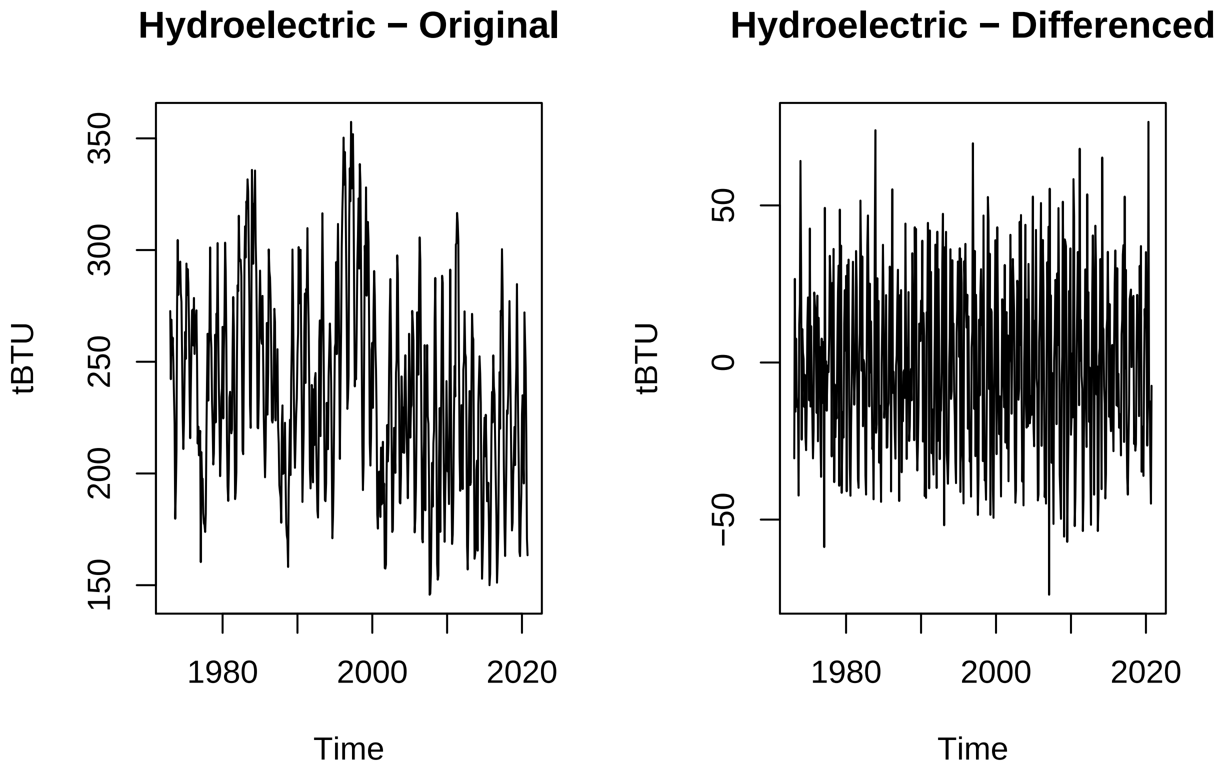
```
##          Total.Biomass.Energy.Production Total.Renewable.Energy.Production
## May 2020                          32.232                           118.120
## Jun 2020                          13.965                            14.027
## Jul 2020                          23.155                           -44.154
## Aug 2020                           1.969                           -40.603
## Sep 2020                         -11.365                           -70.828
## Oct 2020                          14.497                            55.033
##          Hydroelectric.Power.Consumption
## May 2020                          76.590
## Jun 2020                         -12.653
## Jul 2020                         -12.331
## Aug 2020                         -31.389
## Sep 2020                         -44.927
## Oct 2020                          -7.406
```

```
name_list <- c("Biomass", "Total Renewables", "Hydroelectric")

for (i in 1:numobs) {
  par(mfrow = c(1,2))
  plot.ts(energy_ts[,i], main = paste0(name_list[i], " - Original"), ylab = "tBTU")
  plot.ts(energy_diff[,i], main = paste0(name_list[i], " - Differenced"), ylab = "tBTU")
}
```




Total Renewables − Original
tBTU
400
600
800
1000
1980
2000
2020
Time
Total Renewables − Differenced
tBTU
−100
0
50
100
150
1980
2000
2020
Time



> The trends that can be seen in the original plots are not present in the differenced series (the differenced series averages seem to be 0 across time, even though there is still scatter).

Q2

Compute Mann-Kendall and Spearman's Correlation Rank Test for each time series. Ask R to print the results. Interpret the results.

```
energy_yearly <- energy_small %>%
  group_by(year(Month)) %>%
  summarise(Year_Biomass= mean(Total.Biomass.Energy.Production),
            Year_Renewable = mean(Total.Renewable.Energy.Production),
            Year_Hydro = mean(Hydroelectric.Power.Consumption))
```

```
my_year <- c(year(first(energy_small$Month)):year(last(energy_small$Month)))

for (i in 1:numobs) {
  SMKtest <- SeasonalMannKendall(energy_ts[,i])
  print(paste0(name_list[i], " Results for Seasonal Mann Kendall"))
  print(summary(SMKtest))
  cat("\n")

  sp_rho=cor(energy_yearly[,(i+1)],my_year,method="spearman")
  print(paste0(name_list[i], " Results from Spearman Correlation"))
  print(sp_rho)
  cat("\n\n")
}
```

```
## [1] "Biomass Results for Seasonal Mann Kendall"
## Score =  9874 , Var(Score) = 150368.7
## denominator =  13442
## tau = 0.735, 2-sided pvalue =< 2.22e-16
## NULL
##
## [1] "Biomass Results from Spearman Correlation"
##                   [,1]
## Year_Biomass 0.8827616
##
##
## [1] "Total Renewables Results for Seasonal Mann Kendall"
## Score =  9476 , Var(Score) = 150368.7
## denominator =  13442
## tau = 0.705, 2-sided pvalue =< 2.22e-16
## NULL
##
## [1] "Total Renewables Results from Spearman Correlation"
##                     [,1]
## Year_Renewable 0.8617021
##
##
## [1] "Hydroelectric Results for Seasonal Mann Kendall"
## Score =  -3880 , Var(Score) = 150368.7
## denominator =  13442
## tau = -0.289, 2-sided pvalue =< 2.22e-16
## NULL
##
## [1] "Hydroelectric Results from Spearman Correlation"
##                  [,1]
## Year_Hydro -0.4921841
```

> It looks like there is a strong upward trend across the years for the biomass and total renewable datasets, as evidenced by the high (close to 1) Spearman correlation value and the Seasonal Mann Kendall test with p < 0.01 (statistically significant). For they hdyroelectric series, there is evidence of a downward/negative trend based on the SMK p value < 0.01 and a moderately negative Spearman correlation (about -0.5). This smaller magnitude correlation implies that there is less of a trend in the hydroelectric series than the other two, but it is still statistically significant. This result matches what we have seen in the plots and earlier assignments.

Decomposing the series

For this part you will work only with the following columns: Solar Energy Consumption and Wind Energy Consumption.

Q3

Create a data frame structure with these two time series only and the Date column. Drop the rows with *Not Available* and convert the columns to numeric. You can use filtering to eliminate the initial rows or conver to numeric and then use the drop_na() function. If you are familiar with pipes for data wrangling, try using it!

```r
solwin <- data.frame(as.Date(raw_energy$Month),
                     lapply(raw_energy[,c("Solar Energy Consumption", "Wind Energy Consumption")],
                            as.numeric))
```

```
## Warning in lapply(raw_energy[, c("Solar Energy Consumption", "Wind Energy
## Consumption")], : NAs introduced by coercion

## Warning in lapply(raw_energy[, c("Solar Energy Consumption", "Wind Energy
## Consumption")], : NAs introduced by coercion
```

```
solwin <- solwin %>%
  rename(., Month = as.Date.raw_energy.Month.) %>%
  drop_na()

head(solwin)
```

```
##        Month Solar.Energy.Consumption Wind.Energy.Consumption
## 1 1984-01-01                   -0.001                   0.000
## 2 1984-02-01                    0.001                   0.002
## 3 1984-03-01                    0.002                   0.002
## 4 1984-04-01                    0.003                   0.006
## 5 1984-05-01                    0.007                   0.008
## 6 1984-06-01                    0.010                   0.006
```

```
tail(solwin)
```

```
##          Month Solar.Energy.Consumption Wind.Energy.Consumption
## 437 2020-05-01                  131.479                 250.766
## 438 2020-06-01                  129.862                 265.978
## 439 2020-07-01                  139.094                 201.045
## 440 2020-08-01                  128.030                 200.970
## 441 2020-09-01                  108.597                 206.359
## 442 2020-10-01                  100.881                 261.944
```

Q4

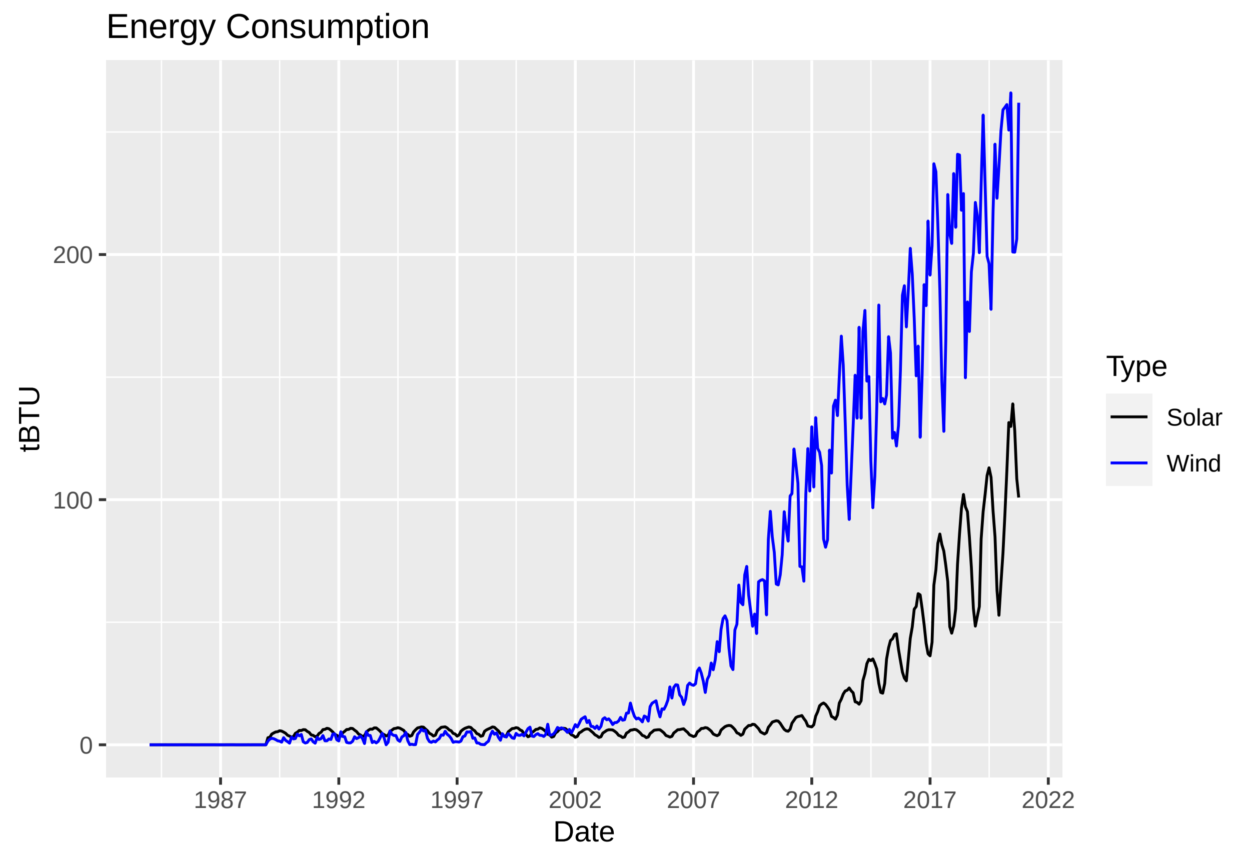
Plot the Solar and Wind energy consumption over time using ggplot. Explore the function scale_x_date() on ggplot and see if you can change the x axis to improve your plot. Hint: use *scale_x_date(date_breaks = “5 years”, date_labels = “%Y”)")*

Try changing the color of the wind series to blue. Hint: use *color = “blue”*

```
# To get both on the same plot with a legend (not sure if that's what is intended), we need
# to change the structure of the data frame
solwin_long <- pivot_longer(solwin, c(2:3), names_to = "Type", values_to = "Consumption")

# Plot with two geom_line features (I don't know how to add a legend to this)
# ggplot(solwin) +
#   geom_line(aes(x = Month, y = Solar.Energy.Consumption)) +
#   geom_line(aes(x = Month, y = Wind.Energy.Consumption), color = "blue") +
#   scale_x_date(date_breaks = "5 years", date_labels = "%Y") +
#   labs(x = "Date", y = "tBTU", title = "Energy Consumption")

# Plot using the long data format to allow for a legend
ggplot(solwin_long) +
  geom_line(aes(x = Month, y = Consumption, color = Type)) +
  scale_x_date(date_breaks = "5 years", date_labels = "%Y") +
  labs(x = "Date", y = "tBTU", title = "Energy Consumption") +
  scale_color_manual(values = c("black", "blue"),labels = c("Solar", "Wind"))
```



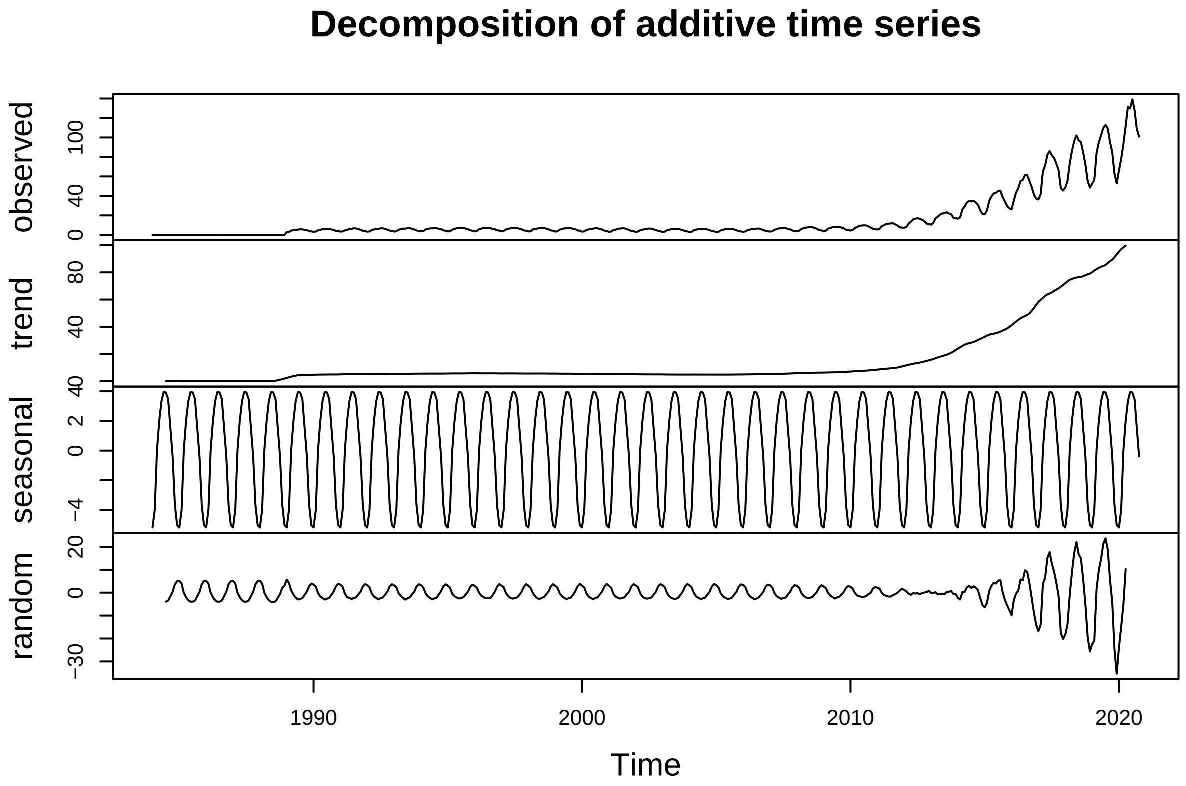
Q5

Transform wind and solar series into a time series object and apply the decompose function on them using the additive option. What can you say about the trend component? What about the random component? Does the random component look random? Or does it appear to still have some seasonality on it?

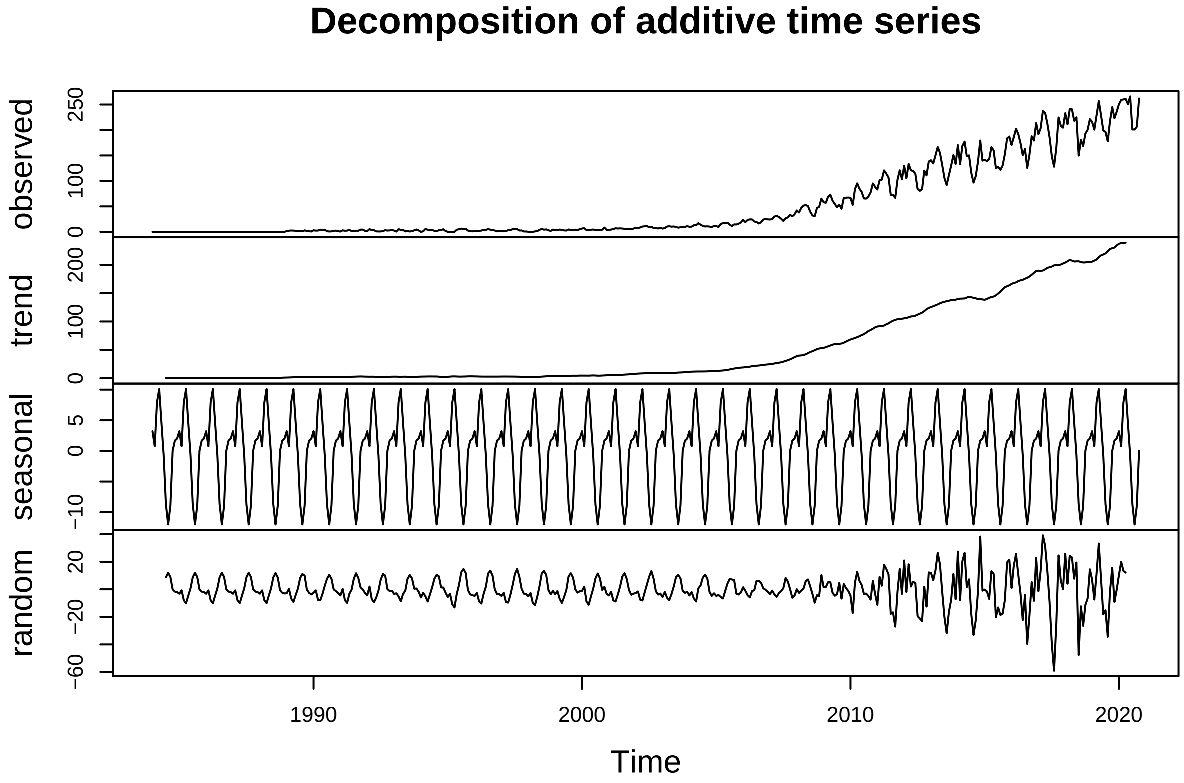
```
solwin_ts <- ts(solwin[,-1], start = c(year(solwin$Month[1]), month(solwin$Month[1])), frequency = 12)

decomp_sol <- decompose(solwin_ts[,1], "additive")
decomp_wind <- decompose(solwin_ts[,2], "additive")

plot(decomp_sol)
```



```
plot(decomp_wind)
```



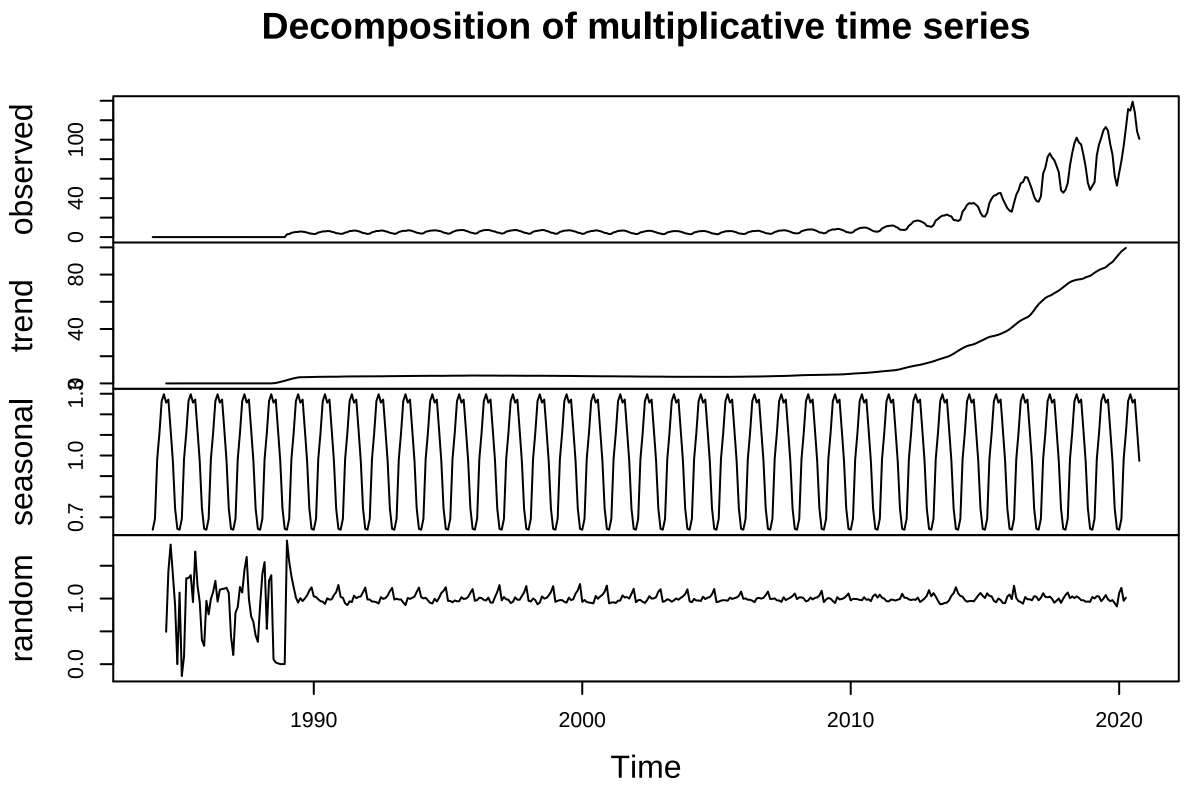
> The trend component for both series appears to be flat and very small up until the early 2000s. Then both series increase significantly (almost exponentially). Both random components have repeating peaks and troughs that look cyclical and seasonal. The repetitions are fairly equal in magnitude within each series until about 2010, at which point they get smaller for a few years, then start growing in magnitude up through the end (2021). This randomness definitely does not appear to be truly random, including what appears to be some seasonality and time-dependent change in magnitude.

Q6

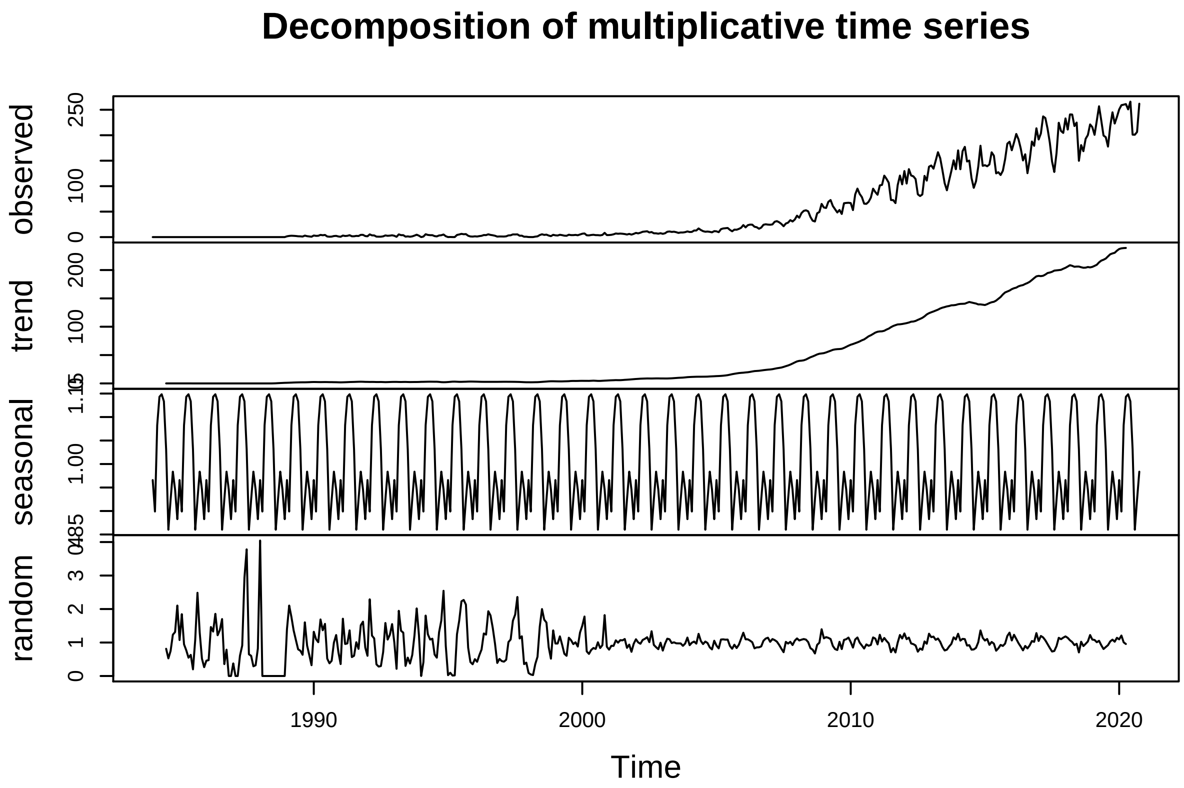
Use the decompose function again but now change the type of the seasonal component from additive to multiplicative. What happened to the random component this time?

```
decomp_sol_mult <- decompose(solwin_ts[,1], "multiplicative")
decomp_wind_mult <- decompose(solwin_ts[,2], "multiplicative")

plot(decomp_sol_mult)
```



```
plot(decomp_wind_mult)
```



> The random component is now larger in magnitude at the beginning of each time series up until about 1990 for solar and 2000 for wind. Then, the random component decreases to much smaller variations. There seems to be less repetition in the random component, but there may still be a small seasonal element.

Q7

When fitting a model to this data, do you think you need all the historical data? Think about the date from 90s and early 20s. Are there any information from those year we might need to forecast the next six months of Solar and/or Wind consumption. Explain your response.

> The earliest data (from the 1980s and early 1990s) probably are not very helpful for creating a forecast of solar or wind power consumption. Since there was little to no growth and very little overall usage of these energy sources, those historical data do not seem to provide much information to help predict what the future consumption will be. Especially since the “shape” of the data has changed (i.e., going from a fairly flat line to a near-exponential growth), it may be better to not include the earlier data or at least use a model that does not weight those data heavily.