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

Assignment 4 - Due date 02/25/21

Keyang Xue

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_Consumpti
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(tseries)
library(Kendall)
library(forecast)
library(readxl)
library(ggplot2)
library(dplyr)
library(lubridate)
```

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.

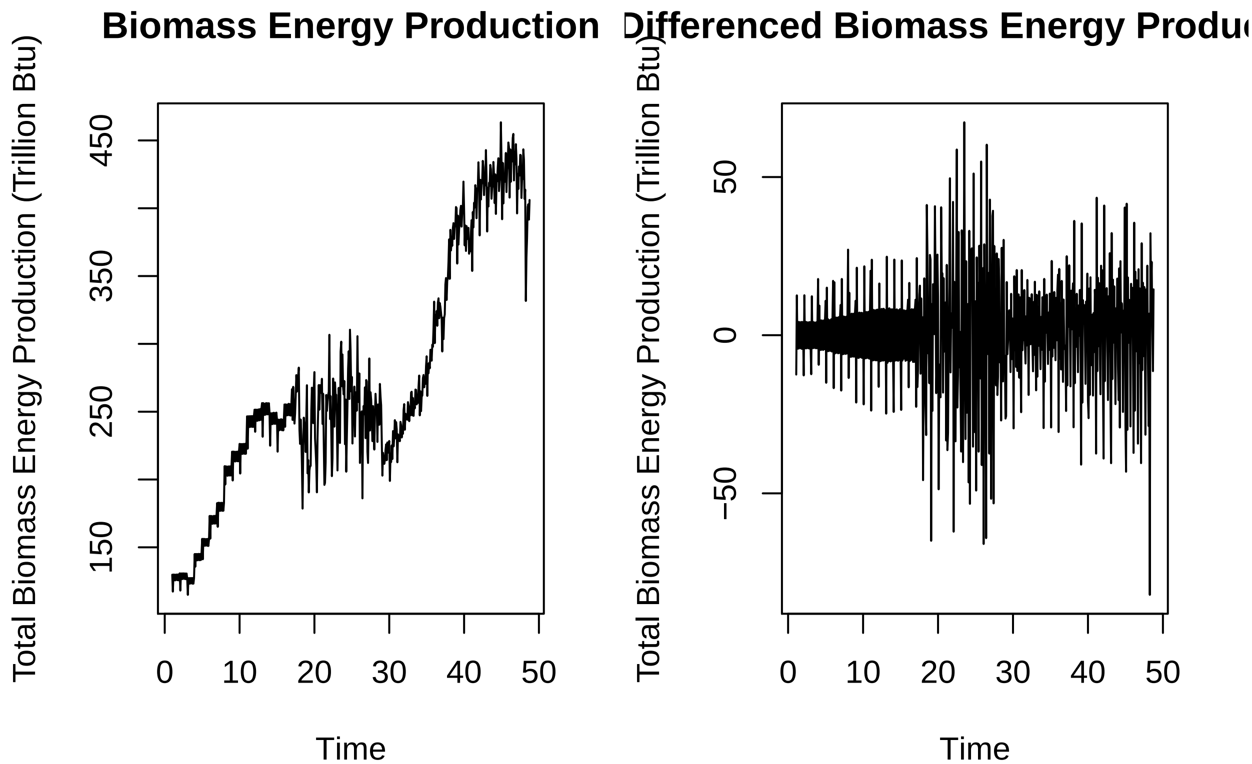
```
#import data
raw_energy <- read_excel("../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx
```

```r
#import column names
col_names <- read_excel("../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx"

#extract required data
energy <- cbind.data.frame(raw_energy[,1],raw_energy[,4:6])
my_date <- as.character(energy[,1])
my_date <- ymd(my_date)
energy <- cbind.data.frame(my_date,raw_energy[,4:6])

#transform data to ts
ts_energy <- ts(energy[2:4], frequency = 12)
```

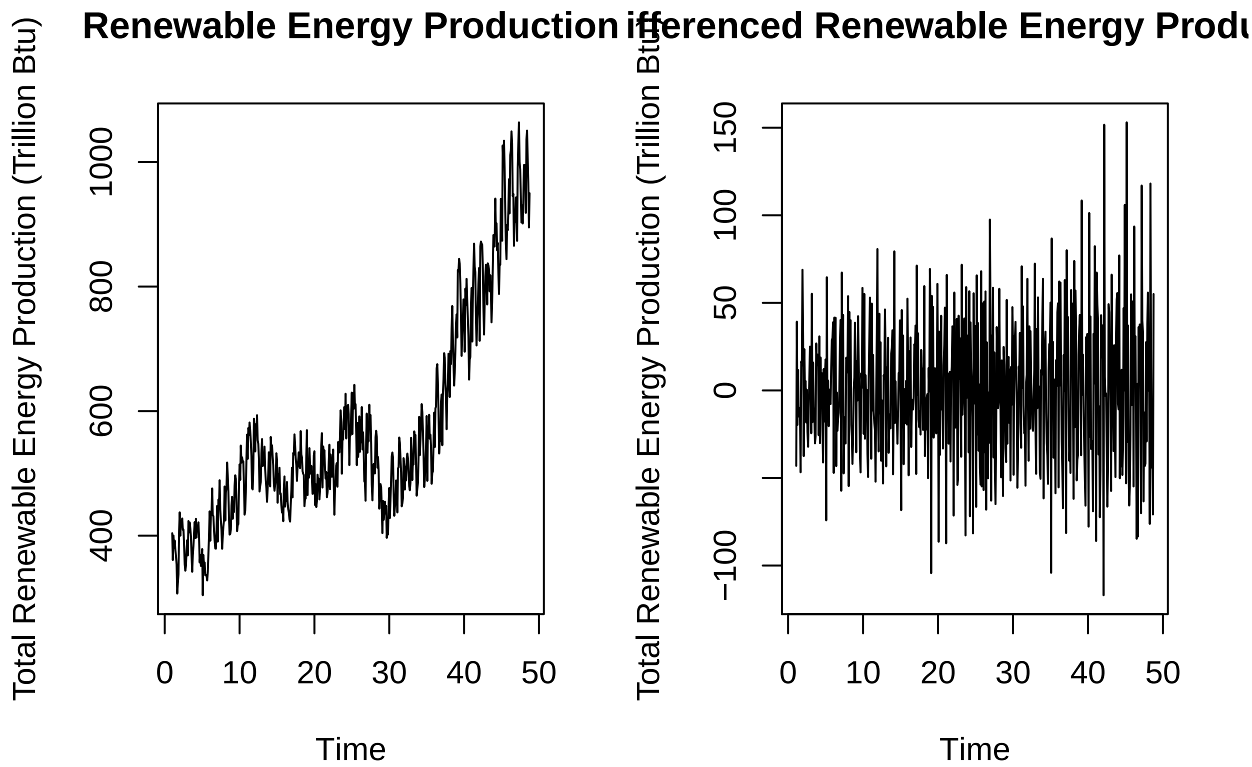
Q1

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?

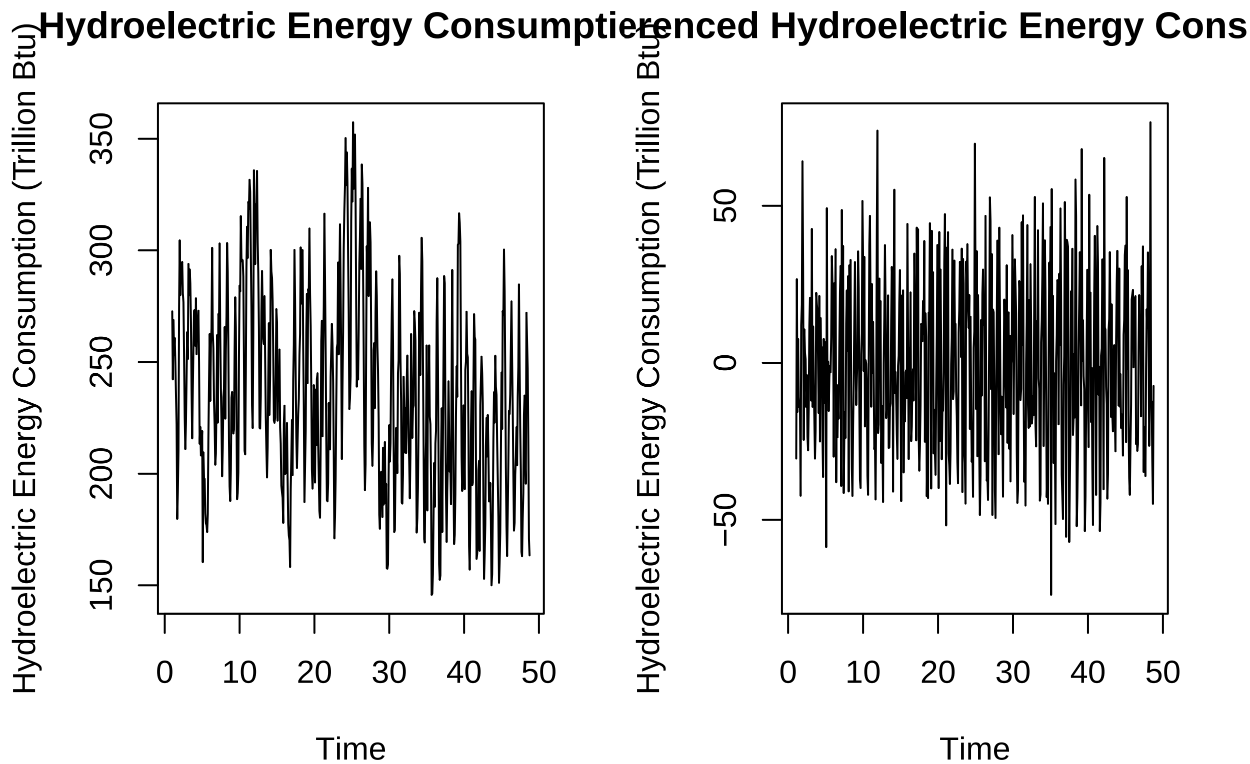
```r
ts_energy_diff <- diff(ts_energy,lag = 1,differences = 1)

par(mfrow=c(1,2))
plot(ts_energy[,1], type="l",
     ylab="Total Biomass Energy Production (Trillion Btu)",
     main='Biomass Energy Production')
# biomass_acf <- Acf(ts_energy[,1],lag.max=40, type="correlation", plot=TRUE,
#                     main = paste(col_names[,4]))
# biomass_pacf <- Pacf(ts_energy[,1],lag.max=40, plot=TRUE,
#                       main = paste(col_names[,4]))

plot(ts_energy_diff[,1], type="l",
     ylab="Total Biomass Energy Production (Trillion Btu)",
     main='Differenced Biomass Energy Production')
```




```
# biomass_acf_diff <- Acf(ts_energy_diff[,1],lag.max=40, type="correlation", plot=TRUE,
#                     main = "Differenced Total Biomass Energy Production")
# biomass_pacf_diff <- Pacf(ts_energy_diff[,1],lag.max=40, plot=TRUE,
#                       main = "Differenced Total Biomass Energy Production")

plot(ts_energy[,2], type="l",
     ylab="Total Renewable Energy Production (Trillion Btu)",
     main='Renewable Energy Production')
# renewable_acf <- Acf(ts_energy[,2],lag.max=40, type="correlation", plot=TRUE,
#                     main = paste(col_names[,5]))
# renewable_pacf <- Pacf(ts_energy[,2],lag.max=40, plot=TRUE,
#                       main = paste(col_names[,5]))

plot(ts_energy_diff[,2], type="l",
     ylab="Total Renewable Energy Production (Trillion Btu)",
     main='Differenced Renewable Energy Production')
```




```r
# renewable_acf_diff <- Acf(ts_energy_diff[,2],lag.max=40, type="correlation", plot=TRUE,
#                       main = "Differenced Total Renewable Energy Production")
# renewable_pacf_diff <- Pacf(ts_energy_diff[,2],lag.max=40, plot=TRUE,
#                         main = "Differenced Total Renewable Energy Production")

plot(ts_energy[,3], type="l",
     ylab="Hydroelectric Energy Consumption (Trillion Btu)",
     main='Hydroelectric Energy Consumption')
# hydro_acf <- Acf(ts_energy[,3],lag.max=40, type="correlation", plot=TRUE,
#                       main = paste(col_names[,6]))
# hydro_pacf <- Pacf(ts_energy[,3],lag.max=40, plot=TRUE,
#                         main = paste(col_names[,6]))

plot(ts_energy_diff[,3], type="l",
     ylab="Hydroelectric Energy Consumption (Trillion Btu)",
     main='Differenced Hydroelectric Energy Consumption')
```




```
# hydro_acf_diff <- Acf(ts_energy_diff[,3],lag.max=40, type="correlation", plot=TRUE,
#                      main = "Differenced Hydroelectric Energy Consumption")
# hydro_pacf_diff <- Pacf(ts_energy_diff[,3],lag.max=40, plot=TRUE,
#                        main = "Differenced Hydroelectric Energy Consumption")
```

The means are constant at 0 after differencing. While the oscillating patterns in the original time series plots are the same, the values after differencing were brought down to oscillate around 0. Hence, the series do not seem to have an increasing or a decreasing trend after differencing.

Q2

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

```
my_year <- c(year(first(my_date)):year(last(my_date)))

biomass_matrix <- matrix(ts_energy[,1],byrow=FALSE,nrow=12)
biomass_yearly <- colMeans(biomass_matrix)
biomass_new_yearly <- data.frame(my_year, biomass_yearly)

renewable_matrix <- matrix(ts_energy[,2],byrow=FALSE,nrow=12)
renewable_yearly <- colMeans(renewable_matrix)
renewable_new_yearly <- data.frame(my_year, renewable_yearly)

hydro_matrix <- matrix(ts_energy[,3],byrow=FALSE,nrow=12)
hydro_yearly <- colMeans(hydro_matrix)
hydro_new_yearly <- data.frame(my_year, hydro_yearly)
```

```r
SMKtest_bio <- SeasonalMannKendall(ts_energy[,1])
print("Results for Seasonal Mann Kendall")
```

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

```r
summary(SMKtest_bio)
```

```
## Score =  9874 , Var(Score) = 150368.7
## denominator =  13442
## tau = 0.735, 2-sided pvalue =< 2.22e-16
```

```r
print("Results of Mann Kendall on average yearly series")
```

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

```r
summary(MannKendall(biomass_yearly))
```

```
## Score =  844 , Var(Score) = 12658.67
## denominator =  1128
## tau = 0.748, 2-sided pvalue =< 2.22e-16
```

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

```
##
##  Spearman's rank correlation rho
##
## data:  biomass_yearly and my_year
## S = 2214, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8798307
```

```r
SMKtest_renewable <- SeasonalMannKendall(ts_energy[,2])
print("Results for Seasonal Mann Kendall")
```

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

```r
summary(SMKtest_renewable)
```

```
## Score =  9476 , Var(Score) = 150368.7
## denominator =  13442
## tau = 0.705, 2-sided pvalue =< 2.22e-16
```

```r
print("Results of Mann Kendall on average yearly series")
```

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

```r
summary(MannKendall(renewable_yearly))
```

```
## Score =  810 , Var(Score) = 12658.67
## denominator =  1128
## tau = 0.718, 2-sided pvalue =< 2.22e-16
```

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

```
##
##  Spearman's rank correlation rho
##
## data:  renewable_yearly and my_year
## S = 2560, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8610508
```

```r
SMKtest_hydro <- SeasonalMannKendall(ts_energy[,3])
print("Results for Seasonal Mann Kendall")
```

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

```r
summary(SMKtest_hydro)
```

```
## Score =  -3880 , Var(Score) = 150368.7
## denominator =  13442
## tau = -0.289, 2-sided pvalue =< 2.22e-16
```

```r
print("Results of Mann Kendall on average yearly series")
```

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

```r
summary(MannKendall(hydro_yearly))
```

```
## Score =  -362 , Var(Score) = 12658.67
## denominator =  1128
## tau = -0.321, 2-sided pvalue =0.0013339
```

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

```
##
##  Spearman's rank correlation rho
##
## data:  hydro_yearly and my_year
## S = 27250, p-value = 0.000661
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.4790491
```

The results of a seasonal Mann Kendall test for biomass series shows a small p value (p<=2.22e-16) and a large score (n=574, score=9874), so we reject the H0 that this time series object is stationary and conclude it has an increasing trend. The Mann Kendall test using yearly biomass series reaches the same conclusion (p<=2.22e-16, n=48, score=844). The Spearman's rank correlation test shows a rho of 0.88 and a small p-value of <2.2e-16, meaning that this series has an increasing trend.

The results of a seasonal Mann Kendall test for renewable energy series shows p<=2.22e-16, score=9476, indicating that we reject the null and conclude that this series is not stationary and has an increasing trend. The Spearman's rank test shows a rho of 0.86 and a p-value <2.2e-16, meaning that we reject the null that the series is stationary and conclude that this series has an increasing trend.

The results of a seasonal Mann Kendall test for hydroelectric series reject the null and conclude that this series is not stationary and has a decreasing trend (p<=2.22e-16, score=-3880, n=574). The Spearman's rank correlation test also rejects the null and concludes that this series has a decreasing trend (rho=-0.48, p=0.000661).

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!

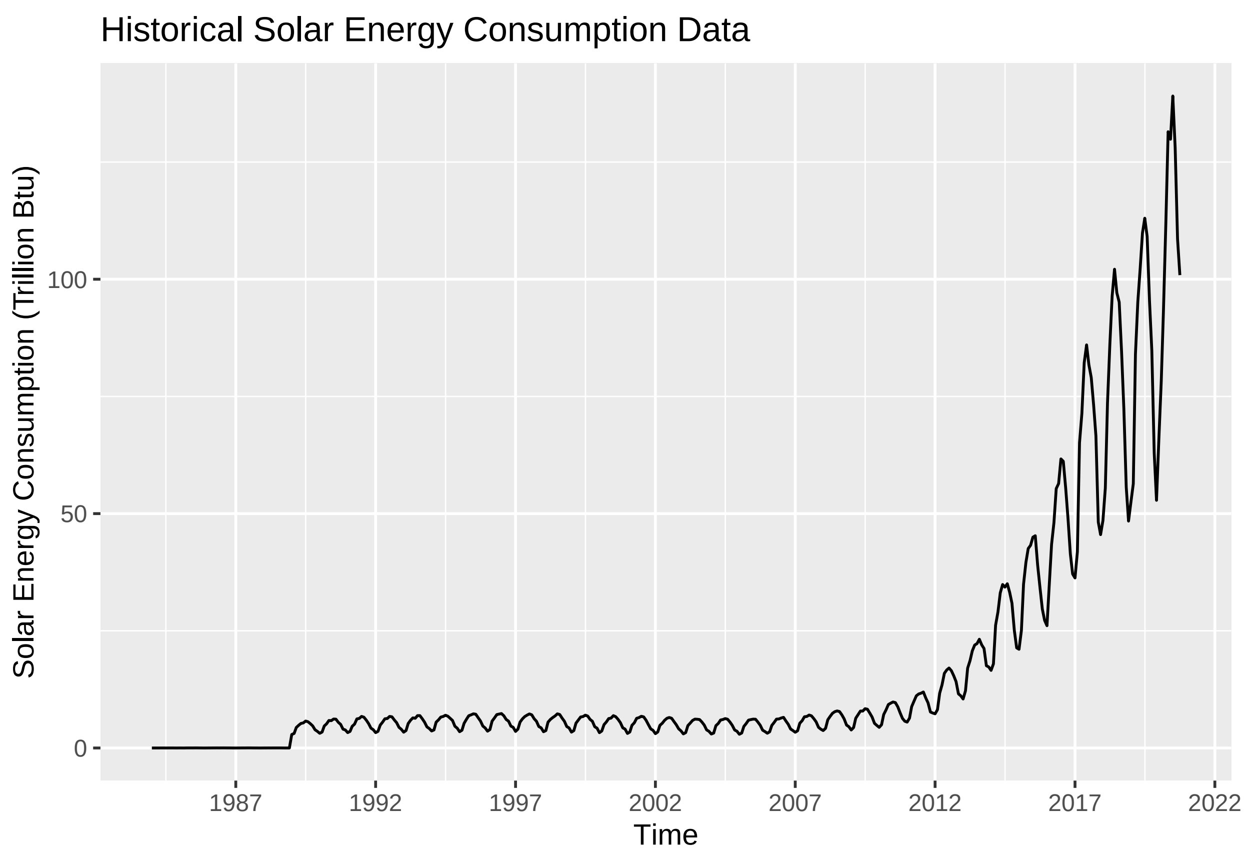
```
SW_energy <- na.omit(data.frame(my_date,raw_energy[,8:9]))
SW_date <- SW_energy[,1]
solar <- as.numeric(SW_energy[,2])
wind <- as.numeric(SW_energy[,3])
SW_energy <- data.frame(SW_date,solar,wind)
SW_ts <-ts(SW_energy[,2:3], frequency = 12)
```

Q4

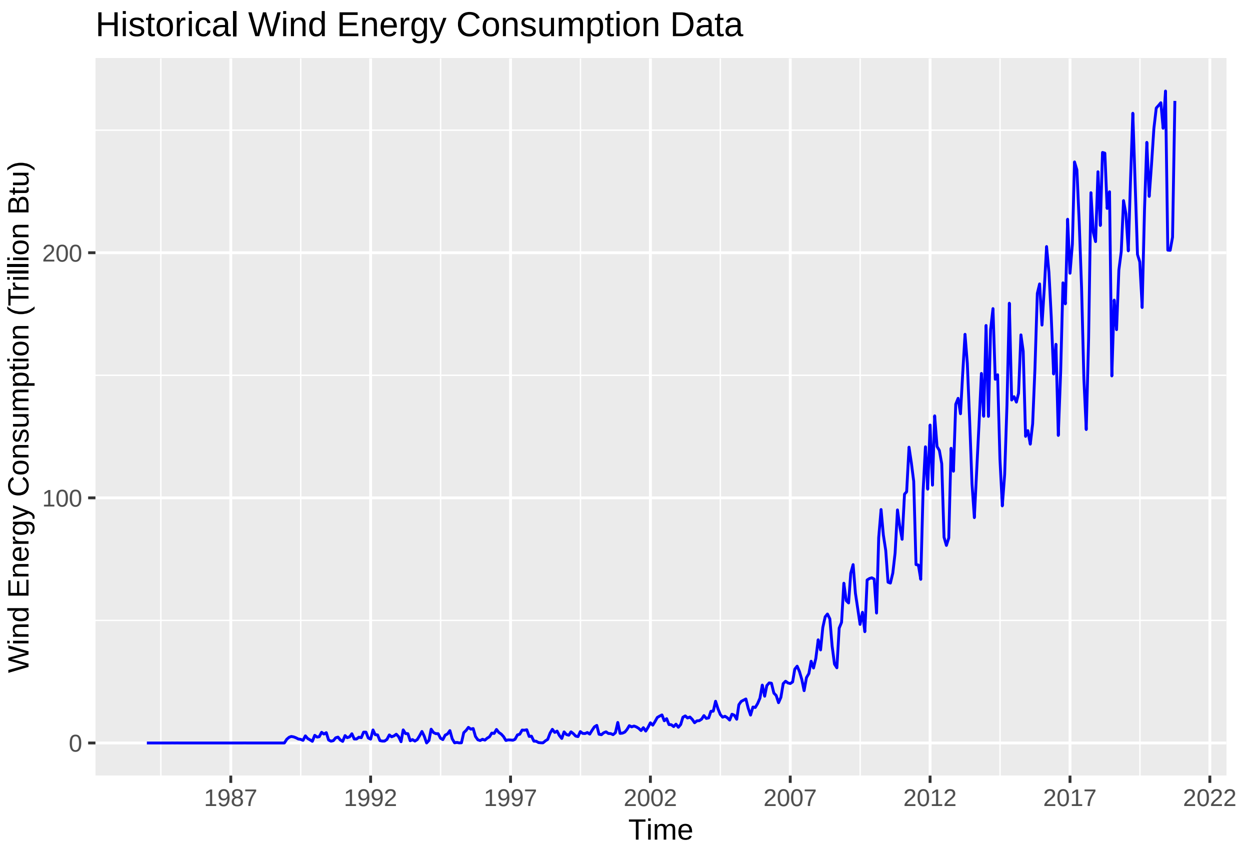
Plot the Solar and Wind energy consumption over time using ggplot. Explore the function scale_x_date() on ggplot and see if you can change the x axis to improve your plot. Hint: use *scale_x_date(date_breaks = "5 years", date_labels = "%Y")")*

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

```
ggplot(SW_energy, aes(x=SW_date,y=solar)) +
      geom_line(color="black")+ xlab("Time") +
  ylab("Solar Energy Consumption (Trillion Btu)") +
  labs(title = 'Historical Solar Energy Consumption Data') +
  scale_x_date(date_breaks = "5 years", date_labels = "%Y")
```



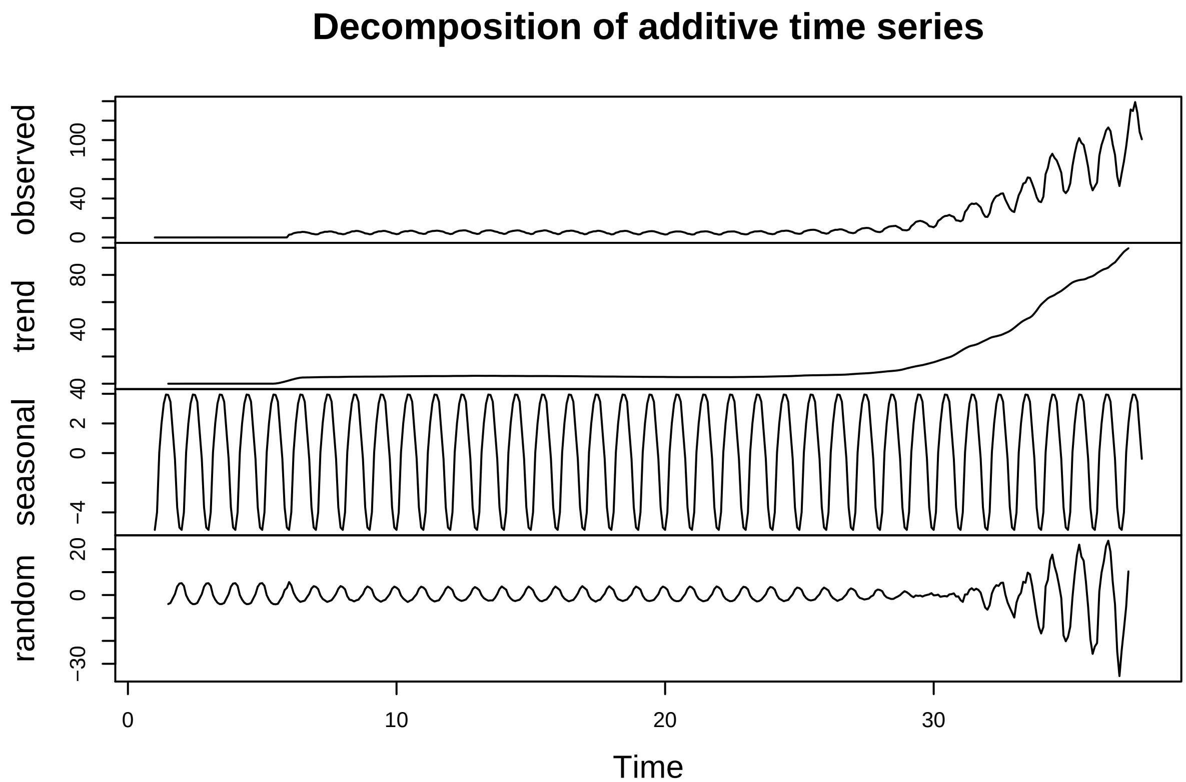

```
ggplot(SW_energy, aes(x=SW_date,y=wind)) +
      geom_line(color="blue")+ xlab("Time") +
  ylab("Wind Energy Consumption (Trillion Btu)") +
  labs(title = 'Historical Wind Energy Consumption Data') +
  scale_x_date(date_breaks = "5 years", date_labels = "%Y")
```



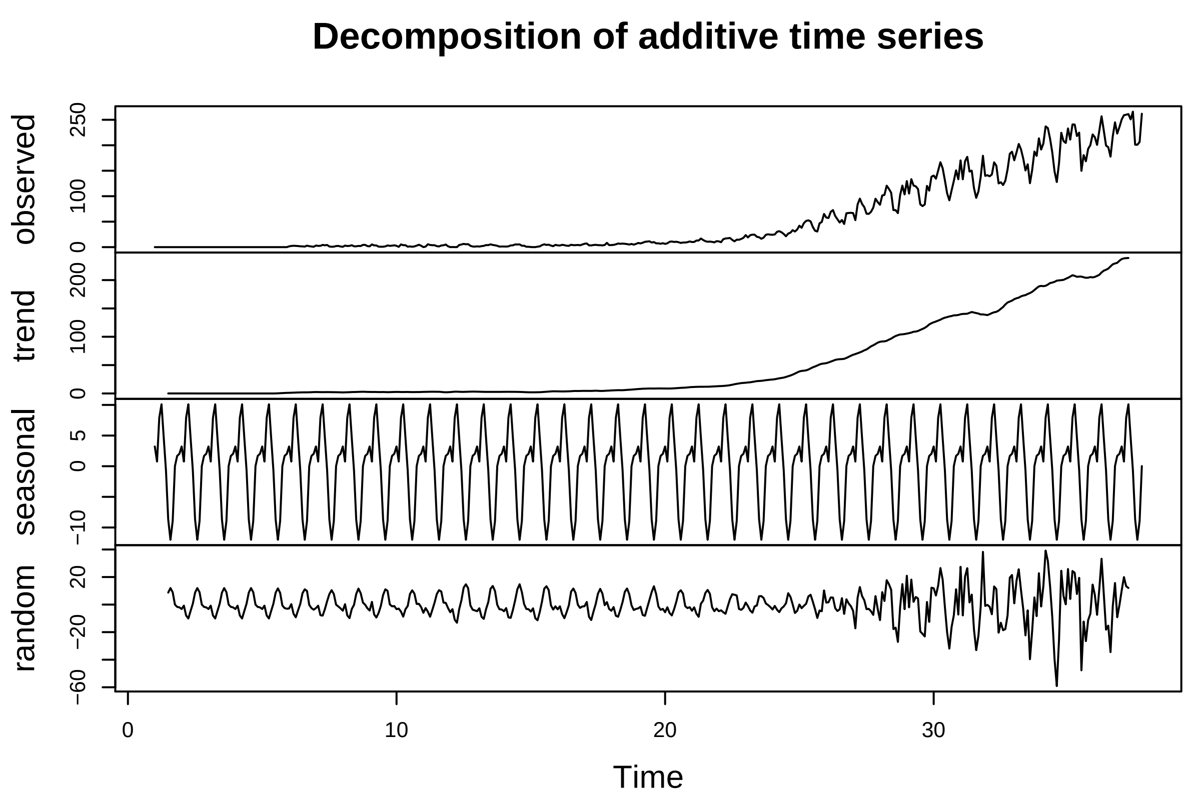
Q5

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?

```
decompose_solar <- decompose(SW_ts[,"solar"],"additive")
plot(decompose_solar)
```




```
decompose_wind <- decompose(SW_ts[,"wind"],"additive")
plot(decompose_wind)
```

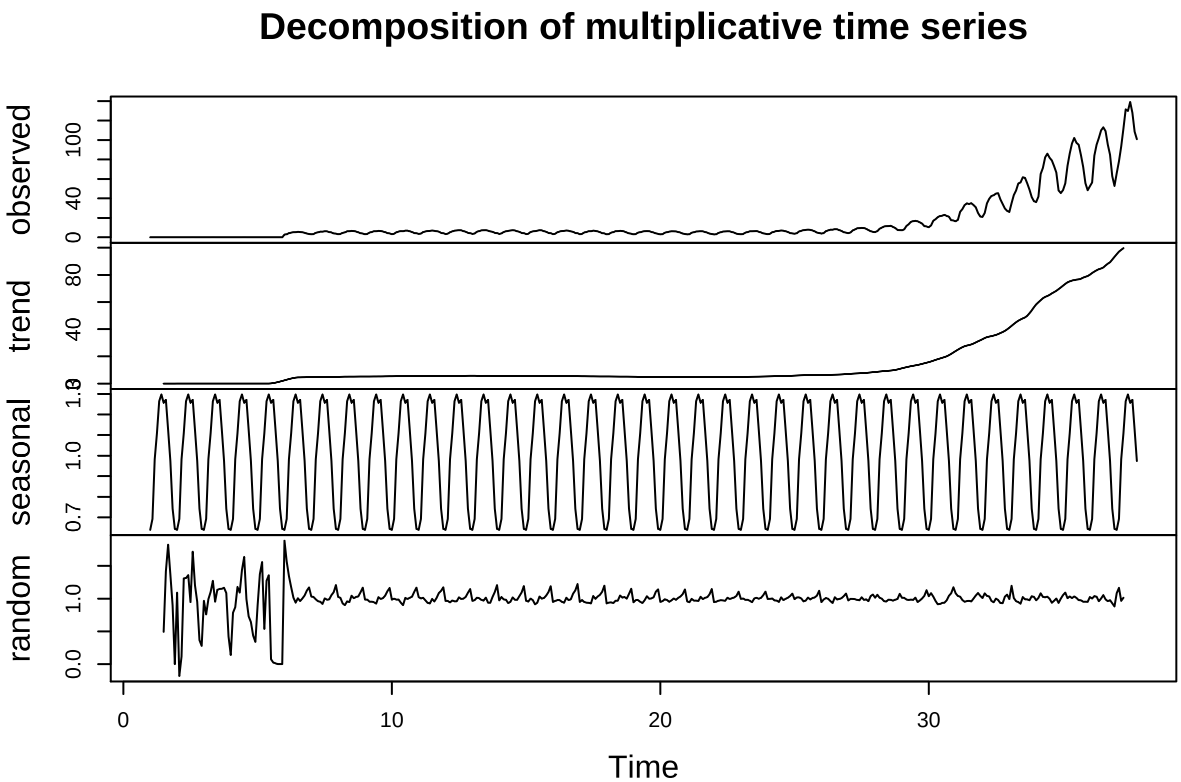


Both time series objects have an increasing trend, and they both show a random effect that does not look random overall, specifically when before lag 29 for solar and lag 22 for wind. Random effect before these time period show seasonality as shown by the identical cyclical pattern in the graph.Random effect of solar and wind after these time periods (lag 29 for solar, lag 24 for wind) seem to be more random.

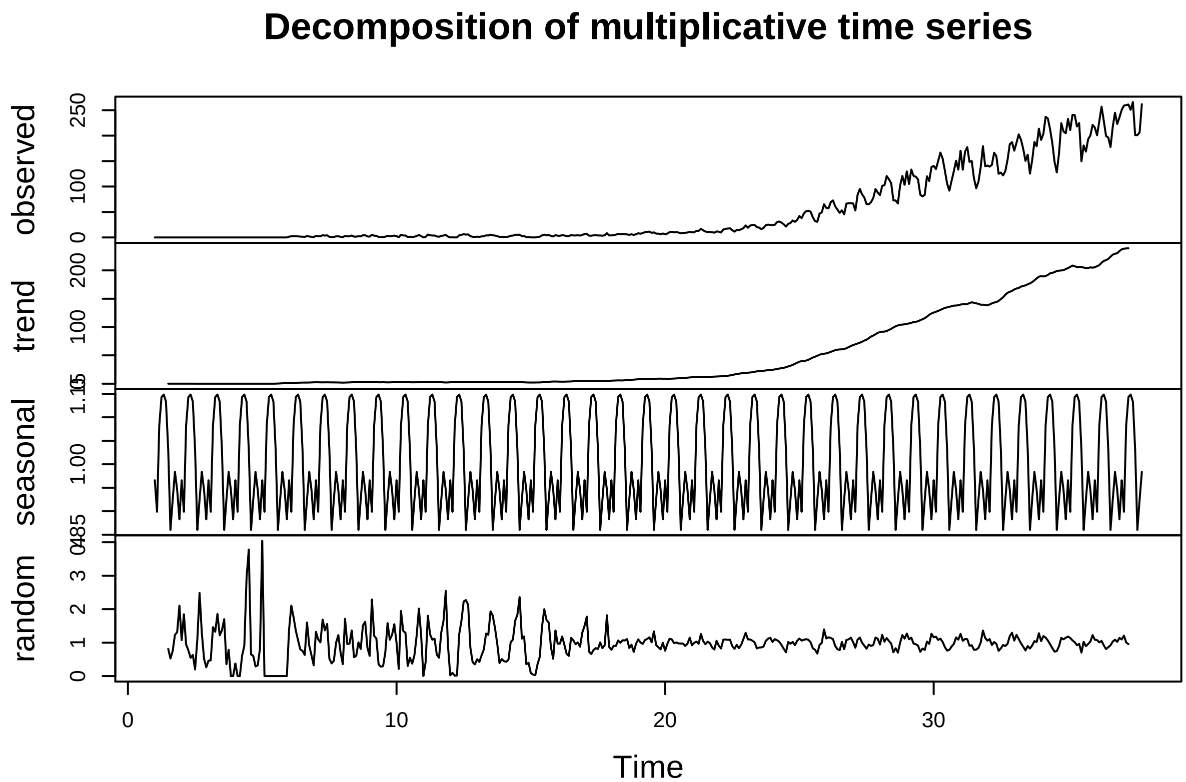
Q6

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?

```
decompose_solar_multi <- decompose(SW_ts[,"solar"],"multiplicative")
plot(decompose_solar_multi)
```




```
decompose_wind_multi <- decompose(SW_ts[,"wind"],"multiplicative")
plot(decompose_wind_multi)
```



Seasonality seems to be weakened in this case, and the random effect is more random in earlier time periods instead of later ones like in the previous question when using the additive method.

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.

Data in the 90s may not be as useful as the data in the early 20s in predicting future consumption in terms of trend and amount because back then the technology was not that advanced, the energy consumption level and the installation of solar energy facilities level are not comparable to the current levels. However, we can examine the seasonal effect in the 90s' data to see how it impacts consumption as there is no increasing trend compounded to the seasonal effect back then as it is in the current era. As for wind data in the early 20s, we can see an increasing trend begins to emerge, and we can use this information to predict the trend in the next six months (increasing trend in solar series begins after 2007).