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

Assignment 4 - Due date 02/17/22

Rob Kravec

Directions

```
# Load / install required package here
library(tidyverse)
library(forecast)
library(tseries)
library(Kendall)
library(readxl)
library(patchwork)
library(lubridate)
```

Questions

Consider the same data you used for A3 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. For this assignment you will work only with the column "Total Renewable Energy
Production".

```
# Read in data
file_path = paste0('../Data/Table_10.1_Renewable_Energy_Production_and',
                   '_Consumption_by_Source.xlsx')
data <- read_excel(path = file_path, sheet = "Monthly Data", skip = 10,
                   na = "Not Available")

# Remove first row, which contains units for each column
data <- data[-1, ]

# Rename Total Renewable Energy Production
data <- data %>%
  rename(Renewable_prod = 'Total Renewable Energy Production')

# Create smaller dataset with columns of interest
data_small <- data %>%
  select(Month, Renewable_prod) %>%
  mutate(Renewable_prod = as.numeric(Renewable_prod))

# Create time series object
data_ts <- ts(data = data_small$Renewable_prod, start = c(1973, 1),
              end = c(2021, 9), frequency = 12)
```

Stochastic Trend and Stationarity Tests

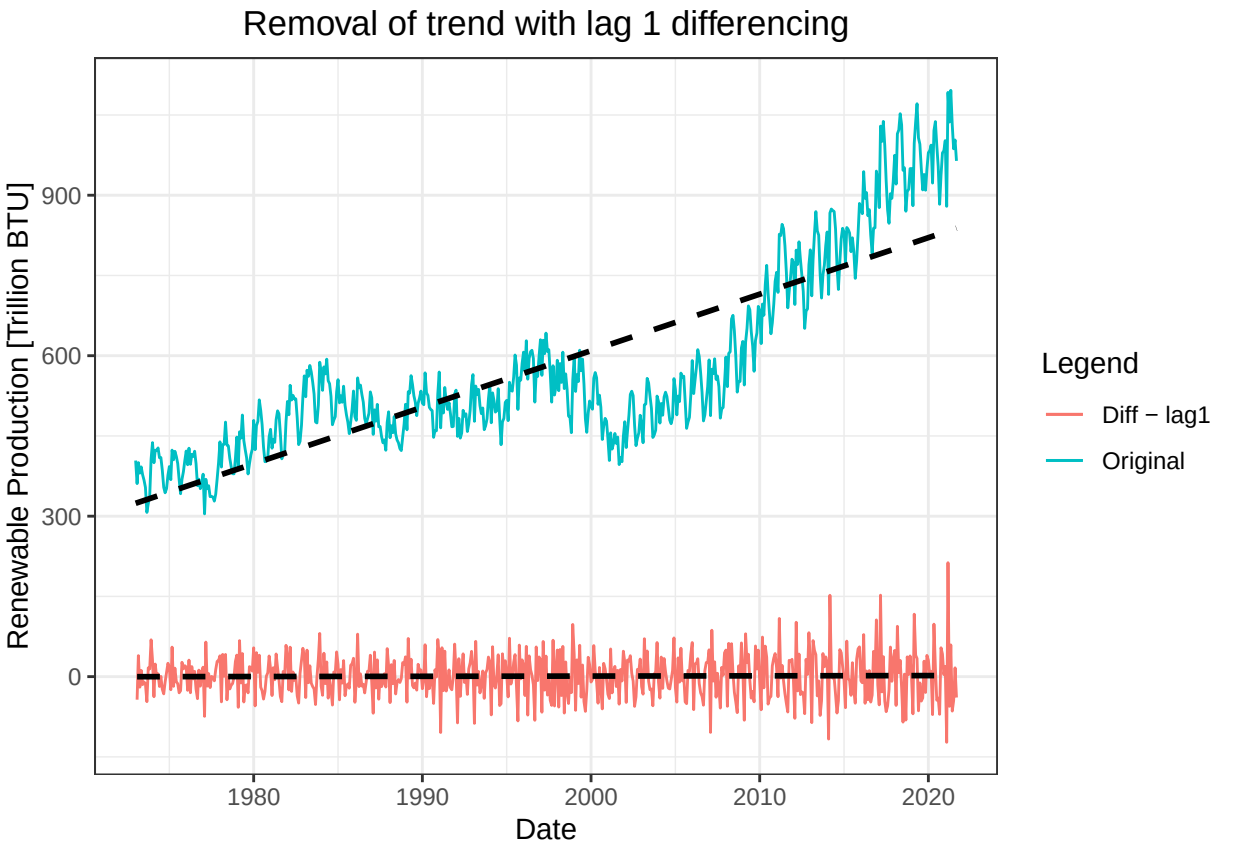
Q1

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?

```
# Create new column with requested lag. Manually fill first value with NA
data_small <- data_small %>%
  mutate(q1 = c(NA, diff(x = Renewable_prod, lag = 1, differences = 1)))

# Generate plot
ggplot(data = data_small, mapping = aes(x = Month)) +
  geom_line(mapping = aes(y = Renewable_prod, color = "Original")) +
  geom_smooth(mapping = aes(y = Renewable_prod), color = "black",
              linetype = 2, method = "lm", se = F) +
  geom_line(mapping = aes(y = q1, color = "Diff - lag1")) +
  geom_smooth(mapping = aes(y = q1), color = "black",
              linetype = 2, method = "lm", se = F) +
  labs(y = "Renewable Production [Trillion BTU]",
       x = "Date",
       title = "Removal of trend with lag 1 differencing",
       color = "Legend") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```



In the original series, we see a clear upward trend over time. In the differenced series, we see no such trend. The trend line for the differenced series is approximately horizontal near y = 0.

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

```
# Define time vector
t <- 1:nrow(data)

# Perform regression
lm_q2 <- lm(data_small$Renewable_prod ~ t)

# Create detrended series
detrend_q2 <- data_small$Renewable_prod -
  (t * summary(lm_q2)$coefficients[2] + summary(lm_q2)$coefficients[1])

# Display first few rows of detrended series
head(detrend_q2)
```

```
## [1] 79.91806 35.95655 74.33705 53.76554 64.55603 48.76653
```

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.

Notes: - Rather than remove the first row (and throw away data), I just added an NA value to the differenced series - Let's refer to the "detrended by regression" series simply as `Detrended`

```
# Create dataframe
q3_df <- cbind(data_small, Detrended = detrend_q2) %>%
  rename(Original = Renewable_prod,
         Differenced = q1)

# Display first 6 rows
head(q3_df)
```

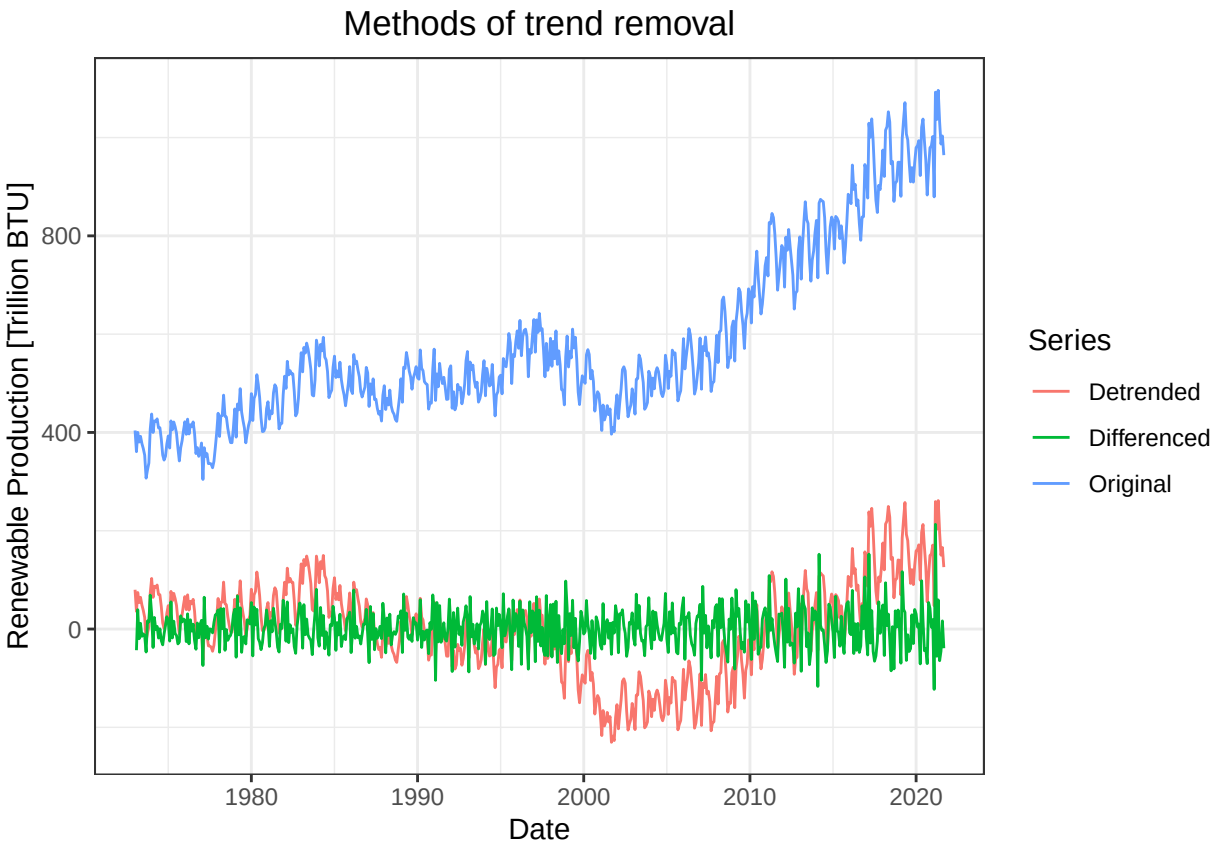
```
##        Month Original Differenced Detrended
## 1 1973-01-01  403.981          NA  79.91806
## 2 1973-02-01  360.900     -43.081  35.95655
## 3 1973-03-01  400.161      39.261  74.33705
## 4 1973-04-01  380.470     -19.691  53.76554
## 5 1973-05-01  392.141      11.671  64.55603
## 6 1973-06-01  377.232     -14.909  48.76653
```

Q4

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

```
# Pivot dataset longer
df_long <- pivot_longer(data = q3_df, cols = !Month)

# Use ggplot
ggplot(data = df_long, mapping = aes(x = Month, y = value, color = name)) +
  geom_line() +
  labs(x = "Date", y = "Renewable Production [Trillion BTU]",
       title = "Methods of trend removal",
       color = "Series") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```



This comparison is quite interesting! Both of our detrending methods remove the upward trend from the original series. However, our previous method of detrending (i.e., linear regression) produces a series with much larger variation around its mean.

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?

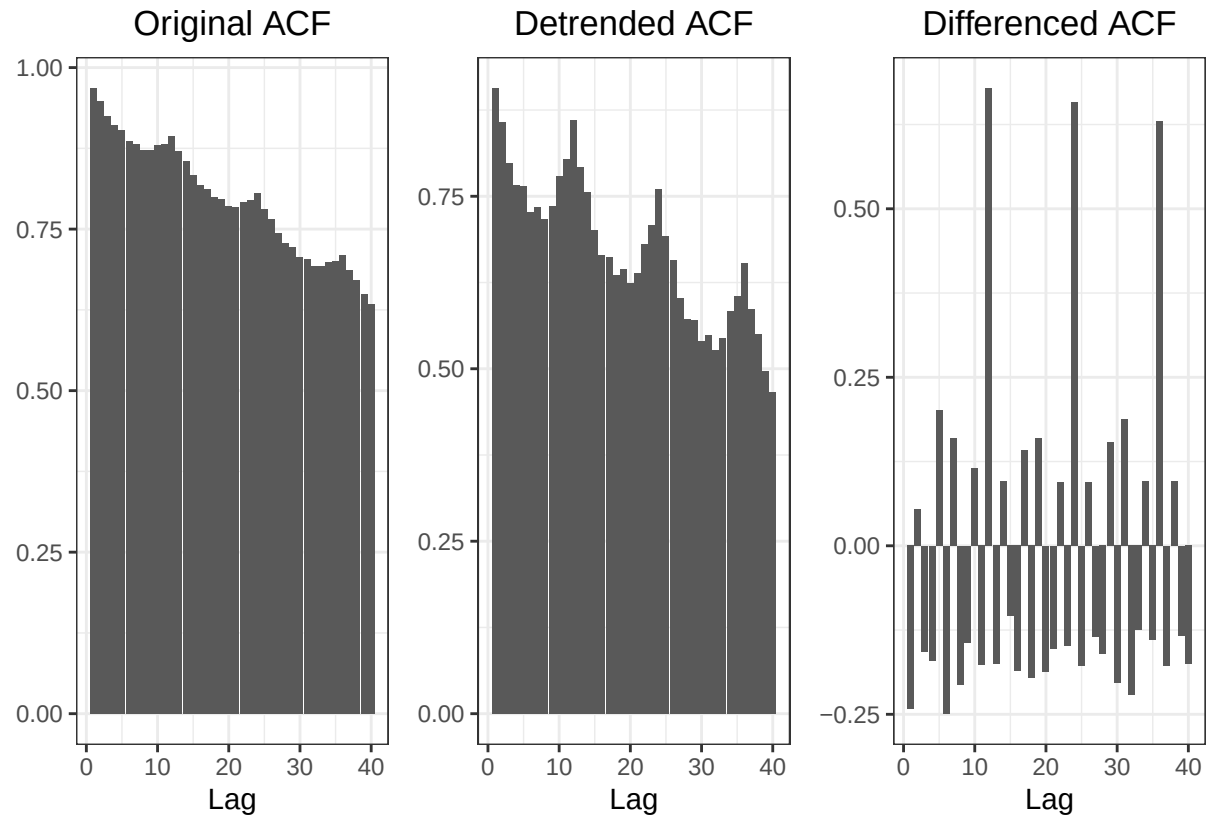
```
# Define function to generate ACF plots
plot_acf <- function(ts, lag_amt, title = "ACF") {
  # Prepare data
  acf_data <- data.frame(lag = 1:lag_amt,
                         acf = Acf(ts, lag.max = lag_amt,
                                   plot = F)$acf[2:(lag_amt + 1)])

  # Create plot
  acf_plt <- ggplot(data = acf_data, mapping = aes(x = lag, y = acf)) +
    geom_bar(stat = 'identity') +
    labs(x = 'Lag', y = '', title = title) +
    theme_bw() +
    theme(plot.title = element_text(hjust = 0.5),
          axis.title.y = element_blank())

  # Return plot
  return(acf_plt)
```

```
}

# Compare ACFs
original_acf <- plot_acf(q3_df$Original, lag_amt = 40, title = "Original ACF")
detrend_acf <- plot_acf(q3_df$Detrended, lag_amt = 40, title = "Detrended ACF")
differenced_acf <- plot_acf(q3_df$Differenced, lag_amt = 40,
                            title = "Differenced ACF")
original_acf + detrend_acf + differenced_acf
```



The differencing method was much more effective at eliminating the trend. This result is evident in the much smaller ACF values for the differenced series relative to the other two. We should note, however, that the differenced series still shows signs of a seasonal trend, as evidenced by the relatively large ACF values at lags 12, 24, and 36 (corresponding to the same month in future years).

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.

```
# Run Seasonal Mann Kendall, and print results, borrowing some code from Lab 4
SMKtest <- SeasonalMannKendall(data_ts)
print("Results for Seasonal Mann Kendall /n",)
```

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

```
print(summary(SMKtest))
```

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

```
# Run ADF test, and print results, borrowing some code from Lab 4
adftest <- adf.test(data_ts, alternative = "stationary")
print(adftest)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  data_ts
## Dickey-Fuller = -1.4383, Lag order = 8, p-value = 0.8161
## alternative hypothesis: stationary
```

The Seasonal Mann-Kendall test suggests that we should reject the null hypothesis that our time series is stationary, suggesting the presence of a deterministic trend.

The Augmented Dickey-Full test fails to provide evidence to reject the null hypothesis that the time series contains a unit root, suggesting the presence of a stochastic trend.

The presence of the stochastic trend aligns with our observations about the time series detrended by linear regression:

- The series still exhibits significant variation from the mean (for long periods of time)
- ACF values are still quite high (though not as high as values seen in the original series)

The differenced time series displays more desirable properties (as described in previous questions), which makes sense, given that differencing can be effective at removing stochastic trends.

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.

```
# Create yearly matrix
yearly_matrix <- matrix(data_ts, byrow = F, nrow = 12)

# Since our time series has a number of observations that is not a multiple
# of 12, the matrix function "wraps around" to fill in the last 3 values.
# This behavior is undesirable, and we'll replace those values with NA
yearly_matrix[10:12, 49] <- c(NA, NA, NA)

# Lastly, aggregate the series by taking column means
yearly_means <- colMeans(yearly_matrix, na.rm = T)
```

Q8

Apply the Mann Kendall, 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?

```
# Mann-Kendall
print("Mann-Kendall")
```

```
## [1] "Mann-Kendall"
```

```
print(summary(MannKendall(yearly_means)))
```

```
## Score =  864 , Var(Score) = 13458.67
## denominator =  1176
## tau = 0.735, 2-sided pvalue =< 2.22e-16
## NULL
```

```
# Spearman correlation rank test
year <- year(q3_df$Month[1]):year(q3_df$Month[nrow(q3_df)])
print("Spearman correlation rank test")
```

```
## [1] "Spearman correlation rank test"
```

```
print(cor.test(yearly_means, year, method="spearman"))
```

```
##
##  Spearman's rank correlation rho
##
## data:  yearly_means and year
## S = 2548, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##  rho
## 0.87
```

```
# ADF test
adftest <- adf.test(yearly_means, alternative = "stationary")
print(adftest)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  yearly_means
## Dickey-Fuller = -0.84991, Lag order = 3, p-value = 0.9512
## alternative hypothesis: stationary
```

Yes, the results from these tests are in agreement with the test results for the non-aggregated results. The Mann-Kendall and Spearman correlation tests suggest the presence of a deterministic trend, while the Augmented Dickey-Fuller test suggests the presence of a stochastic trend.