Assignment 7: Time Series Analysis

Rachel Schoenecker

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on time series analysis.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Fay_A07_TimeSeries.Rmd") prior to submission.

The completed exercise is due on Monday, March 14 at 7:00 pm.

Set up

1. Set up your session:

- Check your working directory
- Load the tidyverse, lubridate, zoo, and trend packages
- Set your ggplot theme

```
#1
getwd()
```

```
## [1] "C:/Users/rscho/OneDrive/Documents/MEM Courses/Sem2/ENV 872/Environmental_Data_Analytics_2022/As
library(tidyverse); library(lubridate); library(zoo); library(trend)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.1.1

## Warning: package 'readr' was built under R version 4.1.1

## Warning: package 'purrr' was built under R version 4.1.1

## Warning: package 'forcats' was built under R version 4.1.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## Warning: package 'lubridate' was built under R version 4.1.2

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union

## Warning: package 'zoo' was built under R version 4.1.1

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Warning: package 'trend' was built under R version 4.1.2
```

```r
mytheme <- theme_bw(base_size = 10) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "right")
theme_set(mytheme)
```

2. Import the ten datasets from the Ozone_TimeSeries folder in the Raw data folder. These contain ozone concentrations at Garinger High School in North Carolina from 2010-2019 (the EPA air database only allows downloads for one year at a time). Import these either individually or in bulk and then combine them into a single dataframe named `GaringerOzone` of 3589 observation and 20 variables.

```r
#2
oz_2010 <-
  read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2010_raw.csv")
oz_2011 <-
  read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2011_raw.csv")
oz_2012 <-
  read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2012_raw.csv")
oz_2013 <-
  read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2013_raw.csv")
oz_2014 <-
  read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2014_raw.csv")
oz_2015 <-
  read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2015_raw.csv")
oz_2016 <-
  read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2016_raw.csv")
oz_2017 <-
  read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2017_raw.csv")
oz_2018 <-
  read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2018_raw.csv")
oz_2019 <-
  read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2019_raw.csv")

oz.total <- rbind(oz_2010, oz_2011, oz_2012, oz_2013, oz_2014, oz_2015,
                  oz_2016, oz_2017, oz_2018, oz_2019)
```

Wrangle

3. Set your date column as a date class.

4. Wrangle your dataset so that it only contains the columns Date, Daily.Max.8.hour.Ozone.Concentration, and DAILY_AQI_VALUE.

5. Notice there are a few days in each year that are missing ozone concentrations. We want to generate a daily dataset, so we will need to fill in any missing days with NA. Create a new data frame that contains a sequence of dates from 2010-01-01 to 2019-12-31 (hint: `as.data.frame(seq())`). Call this new data frame Days. Rename the column name in Days to "Date".

6. Use a `left_join` to combine the data frames. Specify the correct order of data frames within this function so that the final dimensions are 3652 rows and 3 columns. Call your combined data frame GaringerOzone.

```
# 3
oz.total$Date <- as.Date(oz.total$Date, format = "%m/%d/%Y")

# 4
oz.total.sub <- select(oz.total, Date, Daily.Max.8.hour.Ozone.Concentration,
                       DAILY_AQI_VALUE)

# 5
Days <- as.data.frame(seq(as.Date("2010-01-01"), as.Date("2019-12-31"), "days"))
colnames(Days) <- "Date"

# 6
GaringerOzone <- left_join(Days, oz.total.sub)
```

```
## Joining, by = "Date"
```

Visualize

7. Create a line plot depicting ozone concentrations over time. In this case, we will plot actual concentrations in ppm, not AQI values. Format your axes accordingly. Add a smoothed line showing any linear trend of your data. Does your plot suggest a trend in ozone concentration over time?

```
#7
plot1 <- ggplot(GaringerOzone, aes(x = Date,
                                   y = Daily.Max.8.hour.Ozone.Concentration)) +
  geom_line() + geom_smooth(method = lm, se = F) +
  labs(x = "", y = "Ozone Concentration, ppm")
plot1
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 63 rows containing non-finite values (stat_smooth).
```
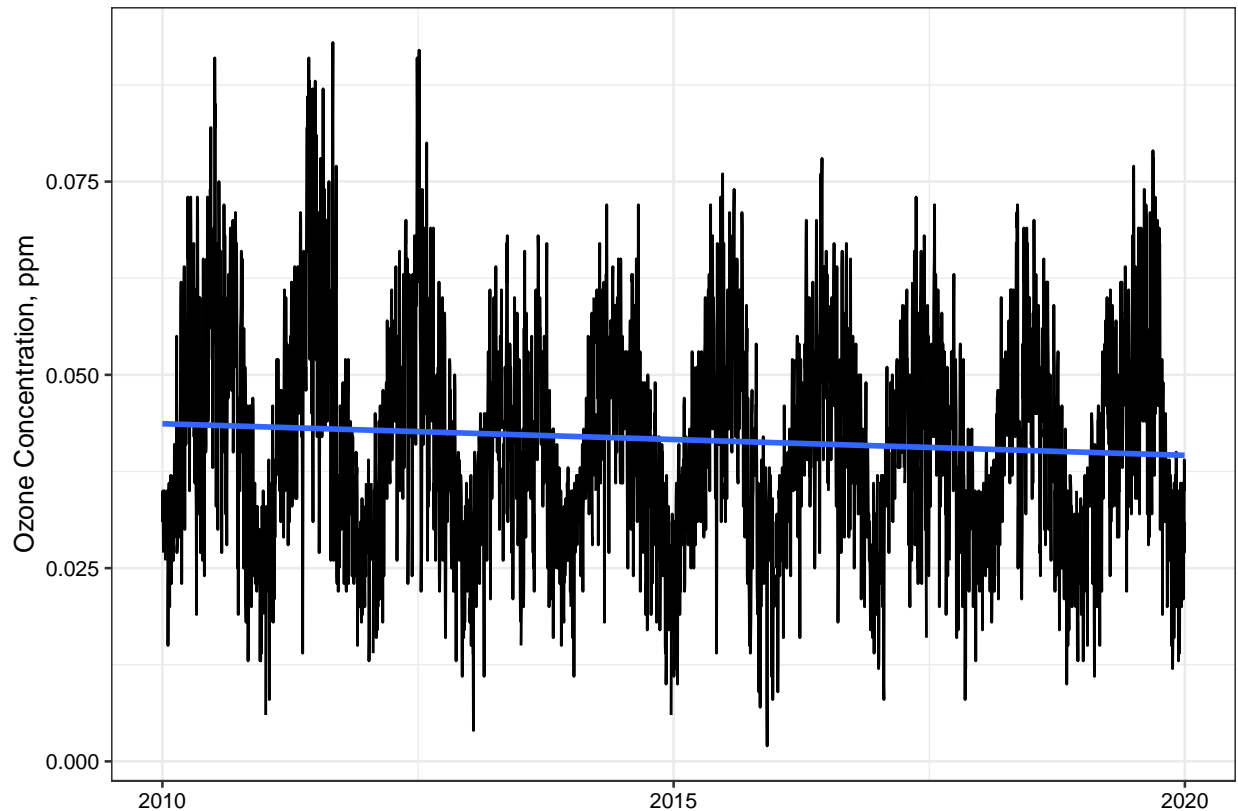
Answer: The plot suggests a slightly decreasing trend in ozone concentration over time.

Time Series Analysis

Study question: Have ozone concentrations changed over the 2010s at this station?

8. Use a linear interpolation to fill in missing daily data for ozone concentration. Why didn't we use a piecewise constant or spline interpolation?

```
#8
GaringerOzone$Daily.Max.8.hour.Ozone.Concentration <-
  na.approx(GaringerOzone$Daily.Max.8.hour.Ozone.Concentration)
GaringerOzone$DAILY_AQI_VALUE <-
  na.approx(GaringerOzone$DAILY_AQI_VALUE)
```

Answer: We used a linear interpolation so that the data fell in between the surrounding data and the new values "connected the dots", as opposed to simply replicating the nearest observation or interpolating with a higher order equation.

9. Create a new data frame called `GaringerOzone.monthly` that contains aggregated data: mean ozone concentrations for each month. In your pipe, you will need to first add columns for year and month to form the groupings. In a separate line of code, create a new Date column with each month-year combination being set as the first day of the month (this is for graphing purposes only)

```
#9
GaringerOzone.monthly <- GaringerOzone %>%
  mutate(Month = month(Date)) %>%
  mutate(Year = year(Date)) %>%
  mutate(Month.Year = my(paste0(Month,"/",Year))) %>%
```

```
    group_by(Month.Year) %>%
    summarise(MeanOzone = mean(Daily.Max.8.hour.Ozone.Concentration))
```

10. Generate two time series objects. Name the first `GaringerOzone.daily.ts` and base it on the dataframe of daily observations. Name the second `GaringerOzone.monthly.ts` and base it on the monthly average ozone values. Be sure that each specifies the correct start and end dates and the frequency of the time series.

```
#10
GaringerOzone.daily.ts <- ts(GaringerOzone$Daily.Max.8.hour.Ozone.Concentration,
                             start = 2010-01-01, frequency = 365)
GaringerOzone.monthly.ts <- ts(GaringerOzone.monthly$MeanOzone,
                               start = 2010-01-01, frequency = 12)
```

11. Decompose the daily and the monthly time series objects and plot the components using the `plot()` function.

```
#11
daily.decomp <- stl(GaringerOzone.daily.ts, s.window = "periodic")
plot(daily.decomp)
```
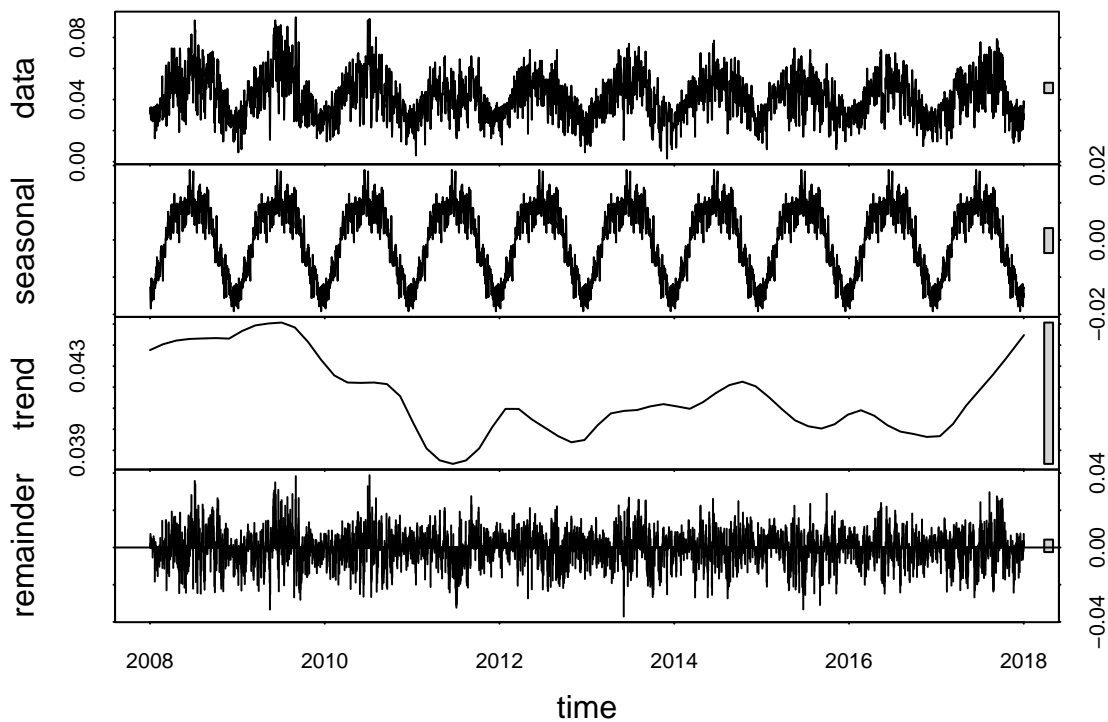


```
monthly.decomp <- stl(GaringerOzone.monthly.ts, s.window = "periodic")
plot(monthly.decomp)
```
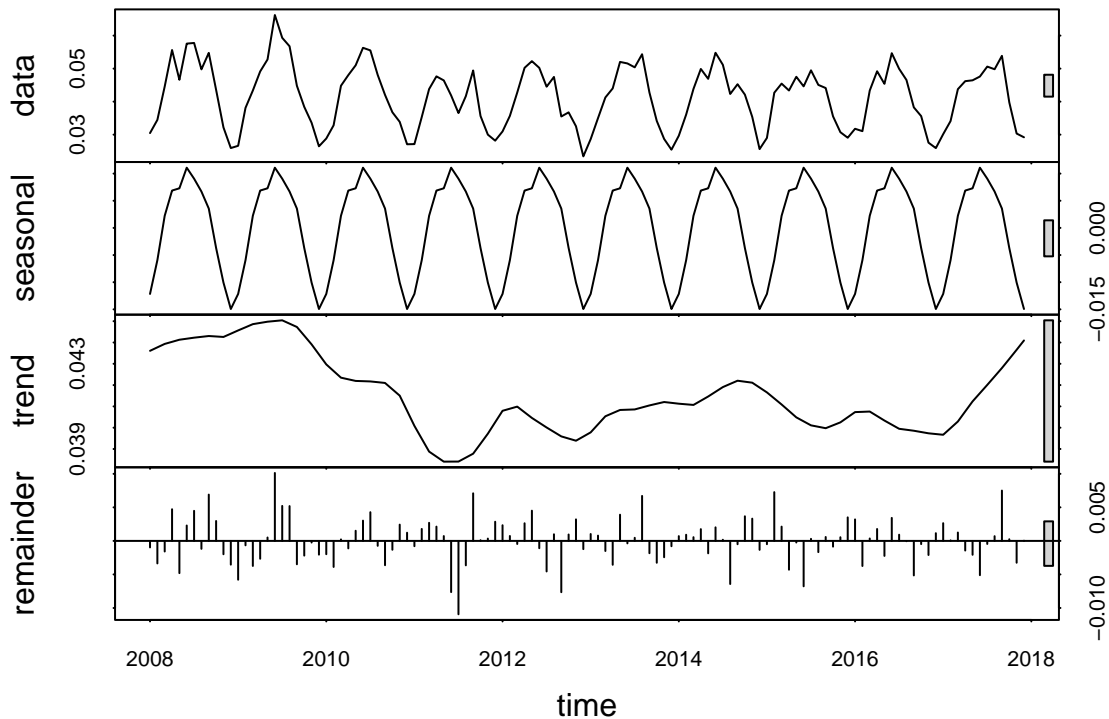
12. Run a monotonic trend analysis for the monthly Ozone series. In this case the seasonal Mann-Kendall is most appropriate; why is this?

```
#12
monthly.trend <- Kendall::SeasonalMannKendall(GaringerOzone.monthly.ts)
summary(monthly.trend)
```

```
## Score =  -77 , Var(Score) = 1499
## denominator =  539.4972
## tau = -0.143, 2-sided pvalue =0.046724
```

Answer: The seasonal Mann-Kendall test is most appropriate because it is the only trend analysis that allows for seasonality, which is clearly present in our data.

13. Create a plot depicting mean monthly ozone concentrations over time, with both a geom_point and a geom_line layer. Edit your axis labels accordingly.

```
# 13
plot2 <- ggplot(GaringerOzone.monthly, aes(x = Month.Year, y = MeanOzone)) +
  geom_point(col = "darkblue") + geom_line(col = "blue") +
  labs(x = "", y = "Mean Ozone Concentration, ppm")
plot2
```
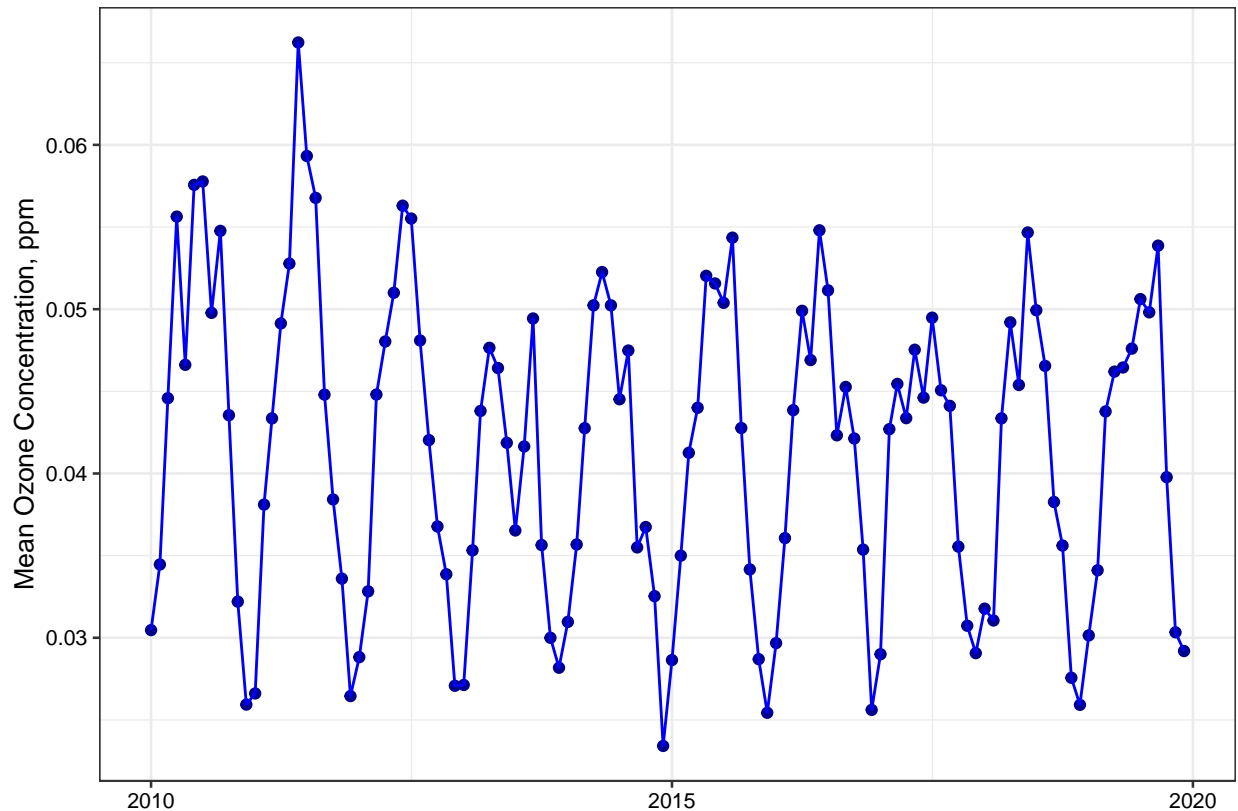
14. To accompany your graph, summarize your results in context of the research question. Include output from the statistical test in parentheses at the end of your sentence. Feel free to use multiple sentences in your interpretation.

    Answer: A Seasonal Mann-Kendall test shows that the overall trend in mean ozone concentration has significantly changed over the 2010s (tau=-0.143,p=0.047).

15. Subtract the seasonal component from the `GaringerOzone.monthly.ts`. Hint: Look at how we extracted the series components for the EnoDischarge on the lesson Rmd file.

16. Run the Mann Kendall test on the non-seasonal Ozone monthly series. Compare the results with the ones obtained with the Seasonal Mann Kendall on the complete series.

```
#15
Month.Comps <- as.data.frame(monthly.decomp$time.series[,1:3])
Month.Comps <- mutate(Month.Comps, Observed = GaringerOzone.monthly$MeanOzone,
                    Date = GaringerOzone.monthly$Month.Year)
Monthly.ts.nonseas <- GaringerOzone.monthly.ts - Month.Comps$seasonal

#16
nonseas.trend <- Kendall::MannKendall(Monthly.ts.nonseas)
summary(nonseas.trend)
```

```
## Score =  -1179 , Var(Score) = 194365.7
## denominator =  7139.5
## tau = -0.165, 2-sided pvalue =0.0075402
```

    Answer: The Mann-Kendall test of the nonseasonal data had an even more significant result than the season Mann-Kendall test, with the p-value being less than 0.01.

7