Assignment 8: Time Series Analysis

Jiaqi Li

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics (ENV872L) on time series analysis.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Use the lesson as a guide. It contains code that can be modified to complete the assignment.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document. Space for your answers is provided in this document and is indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file. You will need to have the correct software installed to do this (see Software Installation Guide) Press the `Knit` button in the RStudio scripting panel. This will save the PDF output in your Assignments folder.
6. After Knitting, please submit the completed exercise (PDF file) to the dropbox in Sakai. Please add your last name into the file name (e.g., "Salk_A08_TimeSeries.pdf") prior to submission.

The completed exercise is due on Tuesday, 19 March, 2019 before class begins.

Brainstorm a project topic

1. Spend 15 minutes brainstorming ideas for a project topic, and look for a dataset if you are choosing your own rather than using a class dataset. Remember your topic choices are due by the end of March, and you should post your choice ASAP to the forum on Sakai.

Question: Did you do this?

> ANSWER: Yes. I will use a dataset containing chlororprene concentrations and meteorological factors in LaPlace, LA. Details on the project have been discussed with Prof. Salk via email.

Set up your session

2. Set up your session. Upload the EPA air quality raw dataset for PM2.5 in 2018, and the processed NTL-LTER dataset for nutrients in Peter and Paul lakes. Build a ggplot theme and set it as your default theme. Make sure date variables are set to a date format.

```
# Set up
setwd("/Users/ljq/Desktop/Blue Devils/Data Analysis/ENV872_02")
library(tidyverse)
air <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv")
nutri <- read.csv("./Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaul_Processed.csv")
air$Date <- as.Date(air$Date, format = "%m/%d/%y")
nutri$sampledate <- as.Date(nutri$sampledate, format = "%Y-%m-%d")

# Set a default theme
light.theme <- theme_light() +
  theme(axis.text = element_text(color = "black"), legend.position = "top",
```

```
        plot.margin = unit(c(0, 0.5, 0, 0.5), "cm"))
theme_set(light.theme)
```

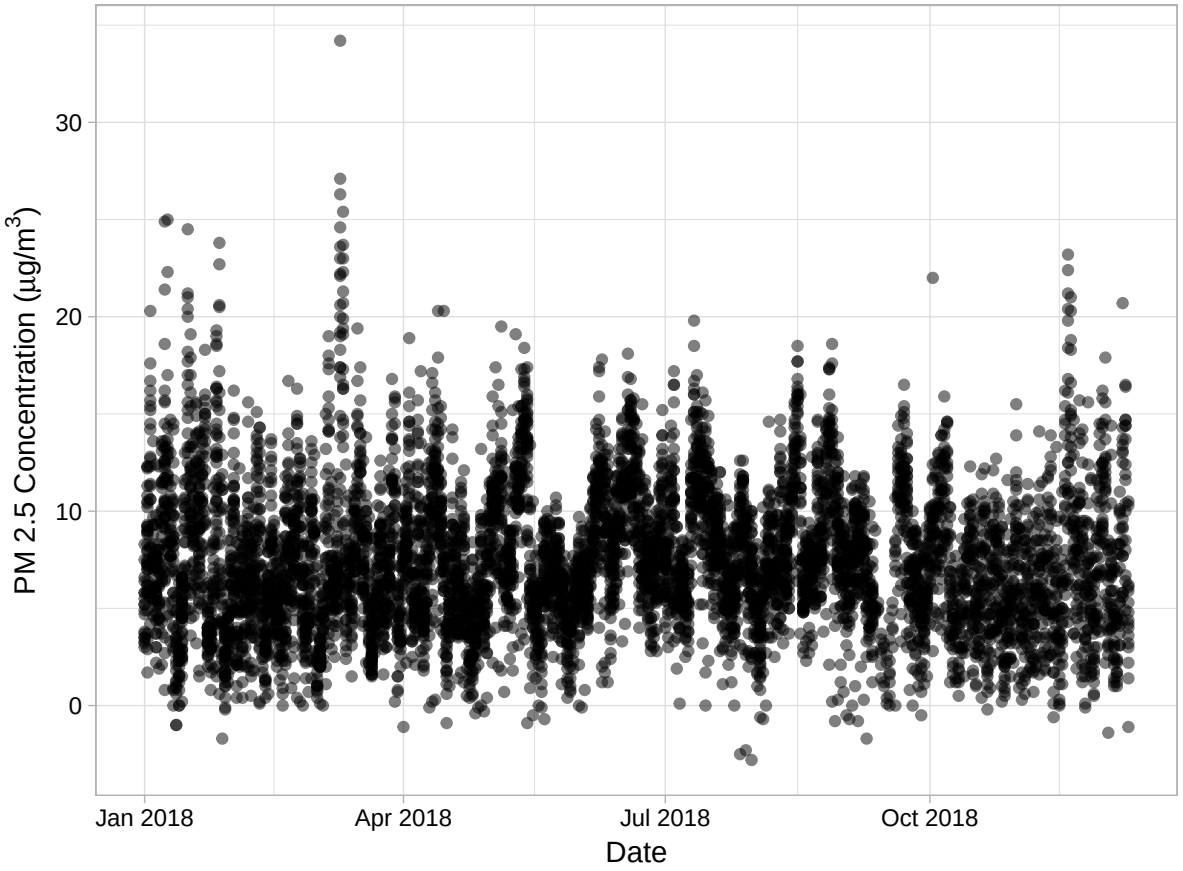
Run a hierarchical (mixed-effects) model

Research question: Do PM2.5 concentrations have a significant trend in 2018?

3. Run a repeated measures ANOVA, with PM2.5 concentrations as the response, Date as a fixed effect, and Site.Name as a random effect. This will allow us to extrapolate PM2.5 concentrations across North Carolina.

3a. Illustrate PM2.5 concentrations by date. Do not split aesthetics by site.

```
ggplot(air, aes(x = Date, y = Daily.Mean.PM2.5.Concentration)) +
  geom_point(alpha = 0.5) +
  ylab(expression(paste("PM 2.5 Concentration (", mu, "g/",m^3,")", sep = "")))
```



3b. Insert the following line of code into your R chunk. This will eliminate duplicate measurements on single dates for each site. PM2.5 = PM2.5[order(PM2.5[,'Date'],-PM2.5[,'Site.ID']),] PM2.5 = PM2.5[!duplicated(PM2.5$Date),]

3c. Determine the temporal autocorrelation in your model.

3d. Run a mixed effects model.

```
# Eliminate duplicate measurements
library(nlme)
air.tidy = air[order(air[,'Date'],-air[,'Site.ID']),]
```

```r
air.tidy = air.tidy[!duplicated(air.tidy$Date),]
airTest.auto <- lme(data = air.tidy, Daily.Mean.PM2.5.Concentration ~ Date,
                    random = ~1|Site.Name)
ACF(airTest.auto)
```

```
##    lag          ACF
## 1    0  1.000000000
## 2    1  0.513829909
## 3    2  0.194512680
## 4    3  0.117925187
## 5    4  0.126462863
## 6    5  0.100699787
## 7    6  0.058215891
## 8    7 -0.053090104
## 9    8  0.017671857
## 10   9  0.012177847
## 11  10 -0.003699721
## 12  11 -0.020305291
## 13  12 -0.044621086
## 14  13 -0.055602646
## 15  14 -0.065787345
## 16  15 -0.123987593
## 17  16 -0.055414056
## 18  17  0.002911218
## 19  18  0.025133456
## 20  19 -0.015306468
## 21  20 -0.143472007
## 22  21 -0.155495492
## 23  22 -0.060369985
## 24  23  0.003954231
## 25  24  0.042295682
## 26  25  0.001320007
```

```r
airTest.mixed <- lme(data = air.tidy, Daily.Mean.PM2.5.Concentration ~ Date, random = ~1|Site.Name,
                     correlation = corAR1(form = ~ 1|Site.Name, value = 0.514), method = "REML")
summary(airTest.mixed)
```

```
## Linear mixed-effects model fit by REML
##  Data: air.tidy
##        AIC      BIC    logLik
##   1758.509 1777.668 -874.2544
##
## Random effects:
##  Formula: ~1 | Site.Name
##         (Intercept) Residual
## StdDev:   0.8955988 3.605739
##
## Correlation Structure: AR(1)
##  Formula: ~1 | Site.Name
##  Parameter estimate(s):
##       Phi
## 0.5319392
## Fixed effects: Daily.Mean.PM2.5.Concentration ~ Date
##                Value Std.Error  DF   t-value p-value
```

```
## (Intercept) 76.74206  61.34912 339  1.250907  0.2118
## Date        -0.00392   0.00346 339 -1.132908  0.2581
##  Correlation:
##      (Intr)
## Date -1
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.3179497 -0.6207613 -0.1187907  0.6089835  3.4109280
##
## Number of Observations: 343
## Number of Groups: 3
```

Is there a significant increasing or decreasing trend in PM2.5 concentrations in 2018?

> ANSWER: There isn't a significant trend in PM2.5 concentrations in 2018 based on the mixed effects model (t-value = -1.133, df = 339, p-value = 0.2581).

3e. Run a fixed effects model with Date as the only explanatory variable. Then test whether the mixed effects model is a better fit than the fixed effect model.

```
# Fixed effects model
airTest.fixed <- gls(data = air.tidy, Daily.Mean.PM2.5.Concentration ~ Date, method = "REML")
summary(airTest.fixed)
```

```
## Generalized least squares fit by REML
##   Model: Daily.Mean.PM2.5.Concentration ~ Date
##   Data: air.tidy
##        AIC      BIC    logLik
##   1865.202 1876.698 -929.6011
##
## Coefficients:
##                Value Std.Error   t-value p-value
## (Intercept) 98.57796  34.60285  2.848840  0.0047
## Date        -0.00513   0.00195 -2.624999  0.0091
##
##  Correlation:
##      (Intr)
## Date -1
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.3531000 -0.6348100 -0.1153454  0.6383004  3.4063068
##
## Residual standard error: 3.584321
## Degrees of freedom: 343 total; 341 residual
```

```
anova(airTest.mixed, airTest.fixed)
```

```
##               Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## airTest.mixed     1  5 1758.509 1777.668 -874.2544
## airTest.fixed     2  3 1865.202 1876.698 -929.6011 1 vs 2 110.6934  <.0001
```

Which model is better?

> ANSWER: According to the results of ANOVA test between the mixed effects model and the fixed effects model, the mixed effects model which includes site names as a random effect is better. It has a lower AIC score of 1758.509, and is significantly different than the fixed effect model

(ANOVA; p-value < 0.0001).

Run a Mann-Kendall test

Research question: Is there a trend in total N surface concentrations in Peter and Paul lakes?

4. Duplicate the Mann-Kendall test we ran for total P in class, this time with total N for both lakes. Make sure to run a test for changepoints in the datasets (and run a second one if a second change point is likely).

```
# Mann-Kendall Test
library(trend)

nutri.surface <-
  nutri %>%
  select(-lakeid, -depth_id, -comments) %>%
  filter(depth == 0) %>%
  filter(!is.na(tn_ug))

Peter.nutri.surface <- filter(nutri.surface, lakename == "Peter Lake")
Paul.nutri.surface <- filter(nutri.surface, lakename == "Paul Lake")

#Peter Lake
# Mann-Kendall Test
mk.test(Peter.nutri.surface$tn_ug)
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.nutri.surface$tn_ug
## z = 7.2927, n = 98, p-value = 3.039e-13
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 2.377000e+03 1.061503e+05 5.001052e-01
```

```
# Test for change point
pettitt.test(Peter.nutri.surface$tn_ug)
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Peter.nutri.surface$tn_ug
## U* = 1884, p-value = 3.744e-10
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              36
```

```
# Seperate Mann-Kendall Test
mk.test(Peter.nutri.surface$tn_ug[1:35])
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.nutri.surface$tn_ug[1:35]
```

```
## z = -0.22722, n = 35, p-value = 0.8203
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
##  -17.00000000 4958.33333333   -0.02857143
```

```
mk.test(Peter.nutri.surface$tn_ug[36:98])
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.nutri.surface$tn_ug[36:98]
## z = 3.1909, n = 63, p-value = 0.001418
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 5.390000e+02 2.842700e+04 2.759857e-01
```

```
# Test for change point
pettitt.test(Peter.nutri.surface$tn_ug[36:98])
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Peter.nutri.surface$tn_ug[36:98]
## U* = 560, p-value = 0.001213
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              21
```

```
# Seperate Mann-Kendall Test
mk.test(Peter.nutri.surface$tn_ug[36:57])
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.nutri.surface$tn_ug[36:57]
## z = -0.56396, n = 22, p-value = 0.5728
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
##  -21.00000000 1257.66666667   -0.09090909
```

```
mk.test(Peter.nutri.surface$tn_ug[57:98])
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.nutri.surface$tn_ug[57:98]
## z = 0.15172, n = 42, p-value = 0.8794
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
##   15.0000000 8514.3333333    0.0174216
```

```r
#Paul Lake
# Mann-Kendall Test
mk.test(Paul.nutri.surface$tn_ug)
```

```
##
##  Mann-Kendall trend test
##
## data:  Paul.nutri.surface$tn_ug
## z = -0.35068, n = 99, p-value = 0.7258
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
## -1.170000e+02  1.094170e+05 -2.411874e-02
```

```r
# Test for change point
pettitt.test(Paul.nutri.surface$tn_ug)
```

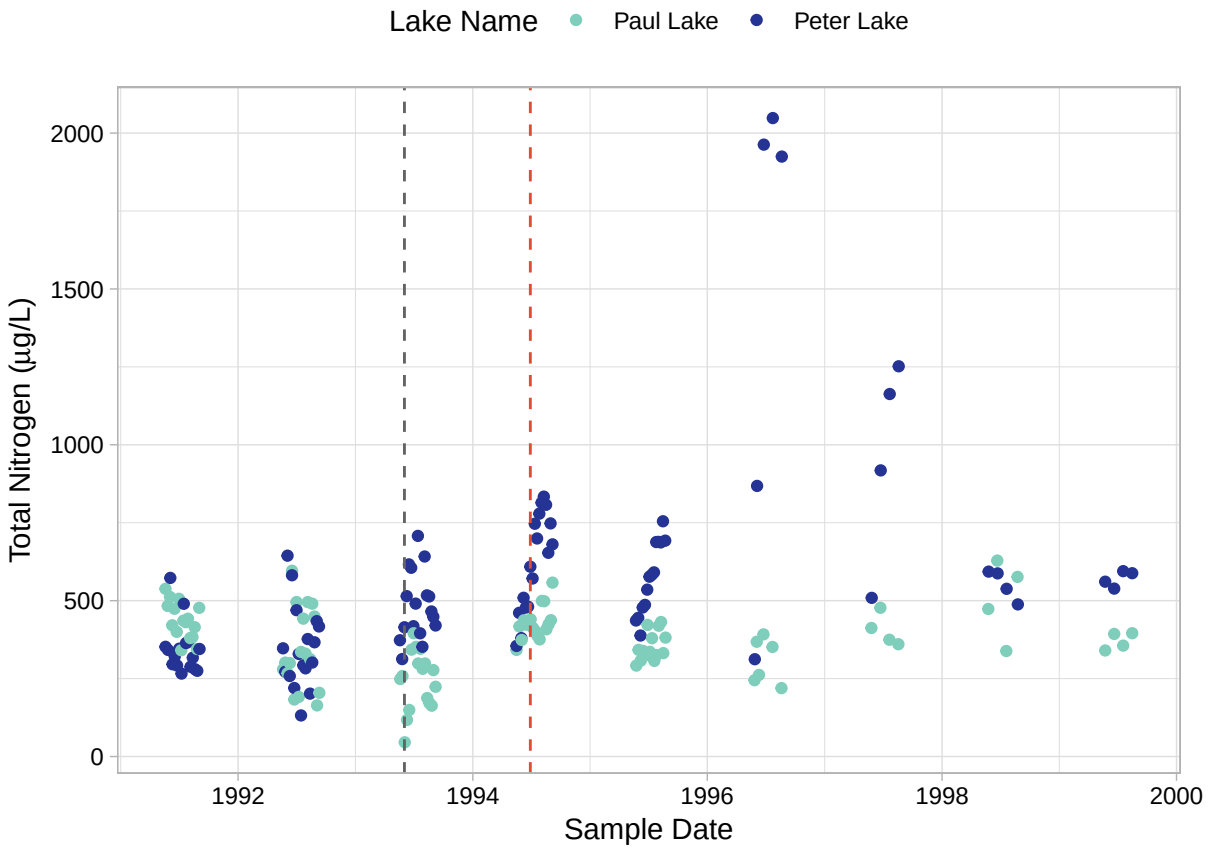
```
##
##  Pettitt's test for single change-point detection
##
## data:  Paul.nutri.surface$tn_ug
## U* = 704, p-value = 0.09624
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              16
```

What are the results of this test?

> ANSWER: The results suggest that there is a positive monotoic trend in total N surface concentrations over time in Peter Lake (Mann-Kendall Test; z = 7.2927, n = 98, p-value < 0.0001). However, no trend (Mann-Kendall Test; z = -0.35068, n = 99, p-value = 0.7258) nor changepoint (Pettitt's Test; p-value = 0.0962) is detected in Paul Lake.
>
> In Peter Lake, the first changepoint is on 1993/06/02 (Pettitt's Test; p-value < 0.0001). There is no trend in Peter Lake before 1993/06/02 (Mann-Kendall Test; z = -0.22722, n = 35, p-value = 0.8203), but there is a positive trend after 1993/06/02 (Mann-Kendall Test; z = 3.1909, n = 63, p-value = 0.00142). And the result for the second Pettitt's test suggests that there is another changepoint on 1994/06/29. There is no trend between 1993/06/02 and 1994/06/29 (Mann-Kendall Test; z = -0.56396, n = 22, p-value = 0.5728) nor after 1994/06/29 in Peter Lake (Mann-Kendall Test; z = 0.15172, n = 42, p-value = 0.8794).

5. Generate a graph that illustrates the TN concentrations over time, coloring by lake and adding vertical line(s) representing changepoint(s).

```r
# Plot for TN
ggplot(nutri.surface, aes(x = sampledate, y = tn_ug, color = lakename)) +
  geom_point() +
  geom_vline(xintercept=as.Date('1993-06-02'), lty = 2, color = "#636363") +
  geom_vline(xintercept=as.Date('1994-06-29'), lty = 2, color = "#e34a33") +
  scale_color_manual(values = c("#7fcdbb", "#253494"), name = "Lake Name") +
  xlab("Sample Date") +
  ylab(expression(paste("Total Nitrogen (", mu, "g/L)", sep = "")))
```


Lake Name
Paul Lake
Peter Lake
Total Nitrogen (µg/L)
2000
1500
1000
500
0
1992
1994
1996
1998
2000
Sample Date