Assignment 8: Time Series Analysis

Ying Wei Jong

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: Thanks for the reminder! I would like to work on the Neonicotinoid dataset. My research question will be to find out which toxin has highest mortality effect on each organisms within the dataset, and whether there are any changes in mortality patterns over time.

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.

```
#Reading files
PM25_2018 <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv", header=T)
PPNutrient <- read.csv("./Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaul_Processed.csv", header=T)

#Set up default theme
library(ggplot2)
my.theme <- theme_bw(base_size = 12) +
  theme(axis.text=element_text(color="gray0"), legend.position = "right")
theme_set(my.theme)
```

```
#Date formatting
PM25_2018$Date <- as.Date(PM25_2018$Date, format="%m/%d/%y")
PPNutrient$sampledate <- as.Date(PPNutrient$sampledate, format = "%Y-%m-%d")
```

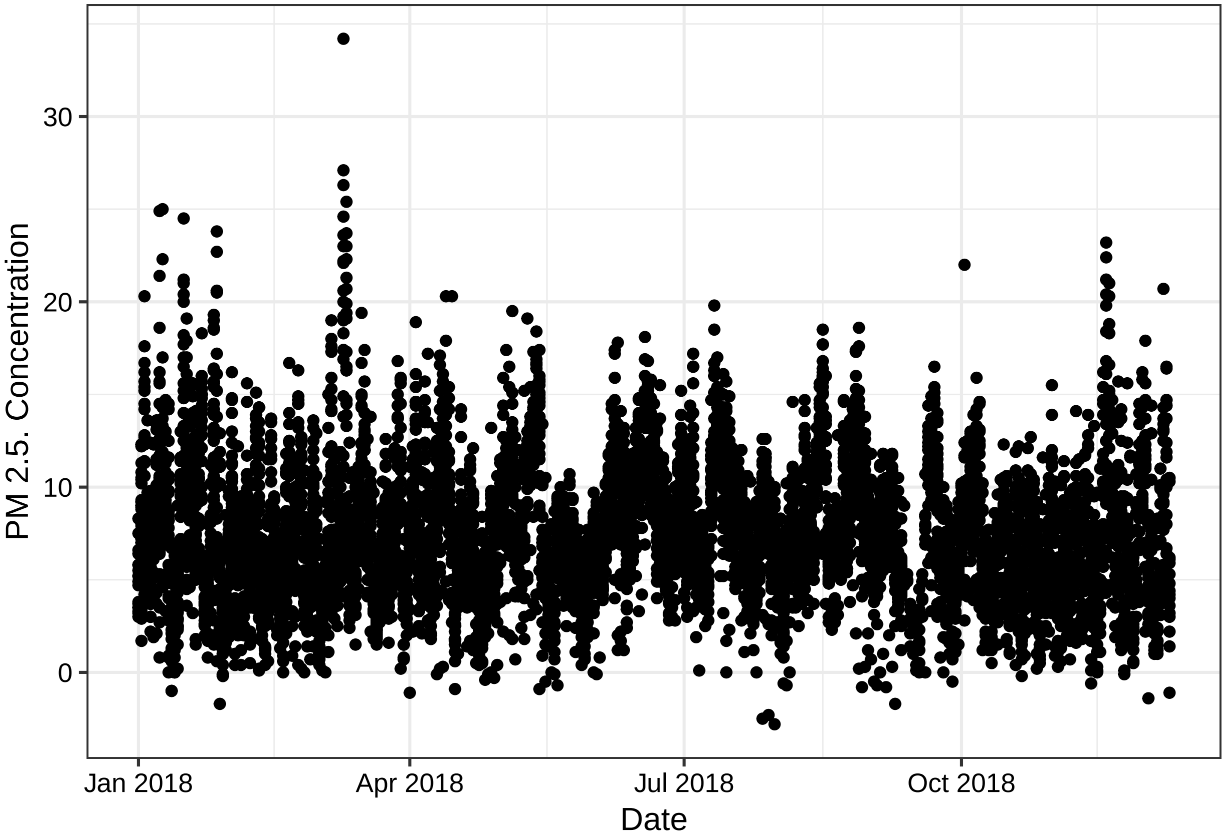
Run a hierarchical (mixed-effects) model

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.

```
library(nlme)
#3a
ggplot(PM25_2018, aes(x = Date, y = Daily.Mean.PM2.5.Concentration)) +
  geom_point() +
  ylab("PM 2.5. Concentration")
```



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.

```r
#3b
PM25_2018 = PM25_2018[order(PM25_2018[,'Date'],-PM25_2018[,'Site.ID']),]
PM25_2018 = PM25_2018[!duplicated(PM25_2018$Date),]

#3c
Temp.auto <- lme(data=PM25_2018, Daily.Mean.PM2.5.Concentration ~ Date, random= ~1|Site.Name)
summary(Temp.auto)
```

```
## Linear mixed-effects model fit by REML
##  Data: PM25_2018
##        AIC      BIC    logLik
##   1865.215 1880.543 -928.6076
##
## Random effects:
##  Formula: ~1 | Site.Name
##         (Intercept) Residual
## StdDev:    1.650184 3.559209
##
## Fixed effects: Daily.Mean.PM2.5.Concentration ~ Date
##                Value Std.Error  DF   t-value p-value
## (Intercept) 90.46502  34.57133 339  2.616764  0.0093
## Date        -0.00473   0.00195 339 -2.425102  0.0158
##  Correlation:
##      (Intr)
## Date -0.999
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.38072443 -0.63365107 -0.09616694  0.61426094  3.42056220
##
## Number of Observations: 343
## Number of Groups: 3
```

```r
ACF(Temp.auto) #Lag1: 0.513829909
```

```
##    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
```

```
#3d
Temp.mixed <- lme(data=PM25_2018, Daily.Mean.PM2.5.Concentration ~ Date,
                  random= ~1|Site.Name,
                  correlation = corAR1(form = ~ Date|Site.Name, value = 0.5138), method = "REML")
summary(Temp.mixed)
```

```
## Linear mixed-effects model fit by REML
##  Data: PM25_2018
##        AIC      BIC   logLik
##   1756.622 1775.781 -873.311
##
## Random effects:
##  Formula: ~1 | Site.Name
##         (Intercept) Residual
## StdDev:  0.00103013 3.597269
##
## Correlation Structure: ARMA(1,0)
##  Formula: ~Date | Site.Name
##  Parameter estimate(s):
##      Phi1
## 0.5384349
## Fixed effects: Daily.Mean.PM2.5.Concentration ~ Date
##                Value Std.Error  DF   t-value p-value
## (Intercept) 83.14801  60.63585 339  1.371268  0.1712
## Date        -0.00426   0.00342 339 -1.244145  0.2143
##  Correlation:
##      (Intr)
## Date -1
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.3220745 -0.6187194 -0.1116751  0.6164257  3.4192603
##
## Number of Observations: 343
## Number of Groups: 3
```

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

> ANSWER: There was a non-significant decreasing trend in PM2.5 concentrations in 2018.

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.

```
Temp.fixed <- gls(data=PM25_2018, Daily.Mean.PM2.5.Concentration ~ Date, method = "REML")
summary(Temp.fixed)
```

```
## Generalized least squares fit by REML
##   Model: Daily.Mean.PM2.5.Concentration ~ Date
##   Data: PM25_2018
##        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(Temp.mixed, Temp.fixed)
```

```
##            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## Temp.mixed     1  5 1756.622 1775.781 -873.3110
## Temp.fixed     2  3 1865.202 1876.698 -929.6011 1 vs 2 112.5802  <.0001
```

Which model is better?

> ANSWER: Mixed model has lower AIC values, so the mixed model is better

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).

```
library(trend)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------- tidyverse 1.2.1 --

## v tibble  2.0.1     v purrr   0.2.5
## v tidyr   0.8.2     v dplyr   0.7.8
## v readr   1.3.1     v stringr 1.3.1
## v tibble  2.0.1     v forcats 0.3.0

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

```
PPNutrient.surface <-
  PPNutrient %>%
  select(-lakeid, -depth_id, -comments) %>%
  filter(depth == 0) %>%
  filter(!is.na(tn_ug)) #filter out NAs in total nitrogen column
```

```
Peter.nutrients.surface <- filter(PPNutrient.surface, lakename == "Peter Lake")
Paul.nutrients.surface <- filter(PPNutrient.surface, lakename == "Paul Lake")

mk.test(Peter.nutrients.surface$tn_ug) #Significant, positive trend
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.nutrients.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
```

```
pettitt.test(Peter.nutrients.surface$tn_ug) #The first change point occurs at 36th data, some time betw
```

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

```
mk.test(Peter.nutrients.surface$tn_ug[1:36])
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.nutrients.surface$tn_ug[1:36]
## z = 0.040863, n = 36, p-value = 0.9674
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 4.000000e+00 5.390000e+03 6.349206e-03
```

```
mk.test(Peter.nutrients.surface$tn_ug[37:98])
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.nutrients.surface$tn_ug[37:98]
## z = 2.9642, n = 62, p-value = 0.003035
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 4.890000e+02 2.710433e+04 2.585933e-01
```

```
pettitt.test(Peter.nutrients.surface$tn_ug[37:98]) #The second change point occurs at 36+20 = 56th data
```

```
##
##  Pettitt's test for single change-point detection
##
```

```
## data:  Peter.nutrients.surface$tn_ug[37:98]
## U* = 522, p-value = 0.002339
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              20
```

```
mk.test(Peter.nutrients.surface$tn_ug[57:98]) #No trend detected after second changing point
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.nutrients.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
```

```
mk.test(Paul.nutrients.surface$tn_ug) #No significant trend
```

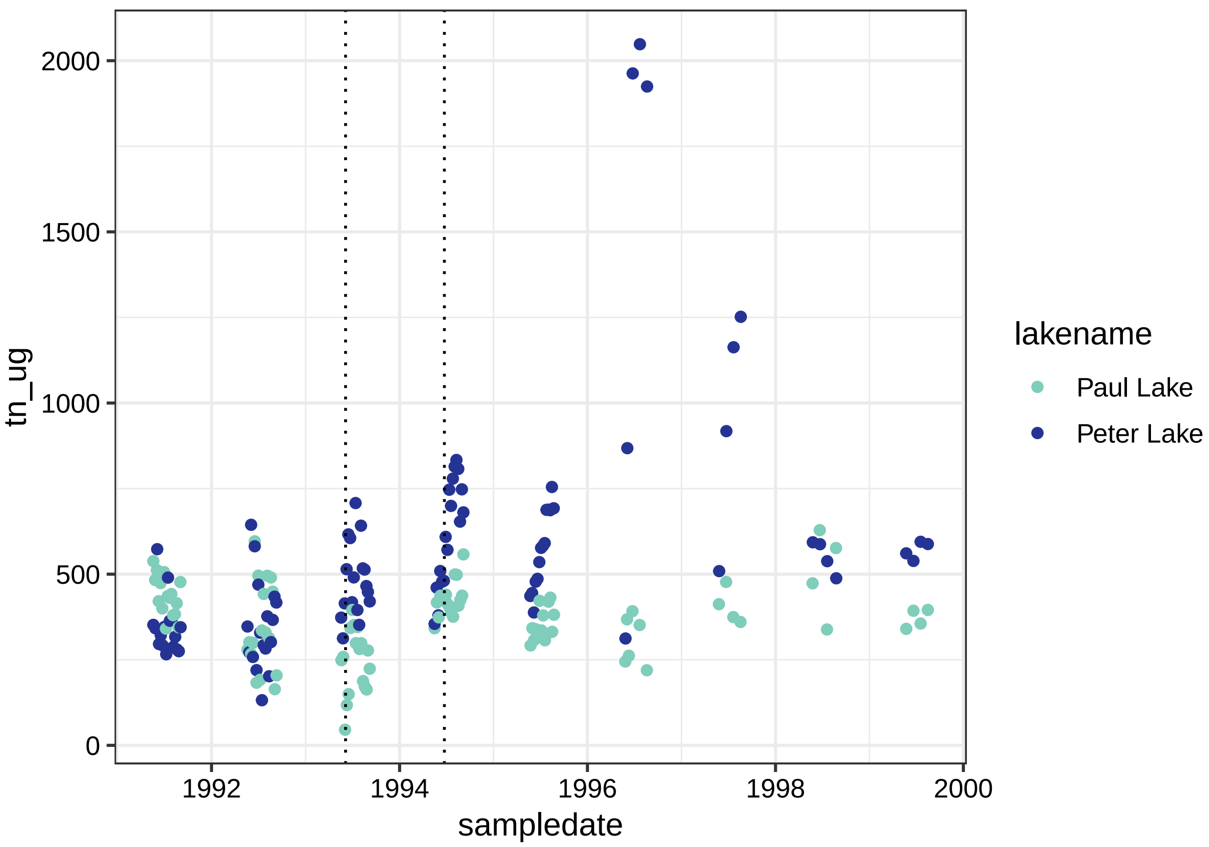
```
##
##  Mann-Kendall trend test
##
## data:  Paul.nutrients.surface$tn_ug
## z = -0.35068, n = 99, p-value = 0.7258
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S           varS            tau
## -1.170000e+02  1.094170e+05 -2.411874e-02
```

What are the results of this test?

> ANSWER: While there are no significant trends observed at Paul Lake, positive significant trends were observed at Peter Lake with two changing points. The first changing point occurs at 36th data, some time between 1993-06-02 and 1993-06-09. The second point occurs at 56th data, some time between 1994-06-22 and 1994-06-29

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

```
ggplot(PPNutrient.surface, aes(x=sampledate, y=tn_ug, col=lakename)) +
  geom_point() +
  geom_vline(xintercept = as.Date("1994-06-24"), color="black", linetype="dotted") +
  geom_vline(xintercept = as.Date("1993-06-05"), color="black", linetype="dotted") +
  scale_color_manual(values = c("#7fcdbb", "#253494"))
```


2000
1500
1000
500
0
tn_ug
1992
1994
1996
1998
2000
sampledate
lakename
● Paul Lake
● Peter Lake