Assignment 8: Time Series Analysis

Yifei Zhang

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.

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.

```
getwd()
```

```
## [1] "/Users/yifeizhang/R/Environmental Data Analytics"
```

```
library(tidyverse)
```

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

## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

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

```
library(nlme)
```

```
##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
library(trend)
EPAair <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv")
PeterPaul <- read.csv("./Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaul_Processed.csv")

yifeitheme <- theme_light(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "right")
theme_set(yifeitheme)

PeterPaul$sampledate <- as.Date(PeterPaul$sampledate, format = "%Y-%m-%d")
EPAair$Date <- as.Date(EPAair$Date, format = "%m/%d/%y")
```

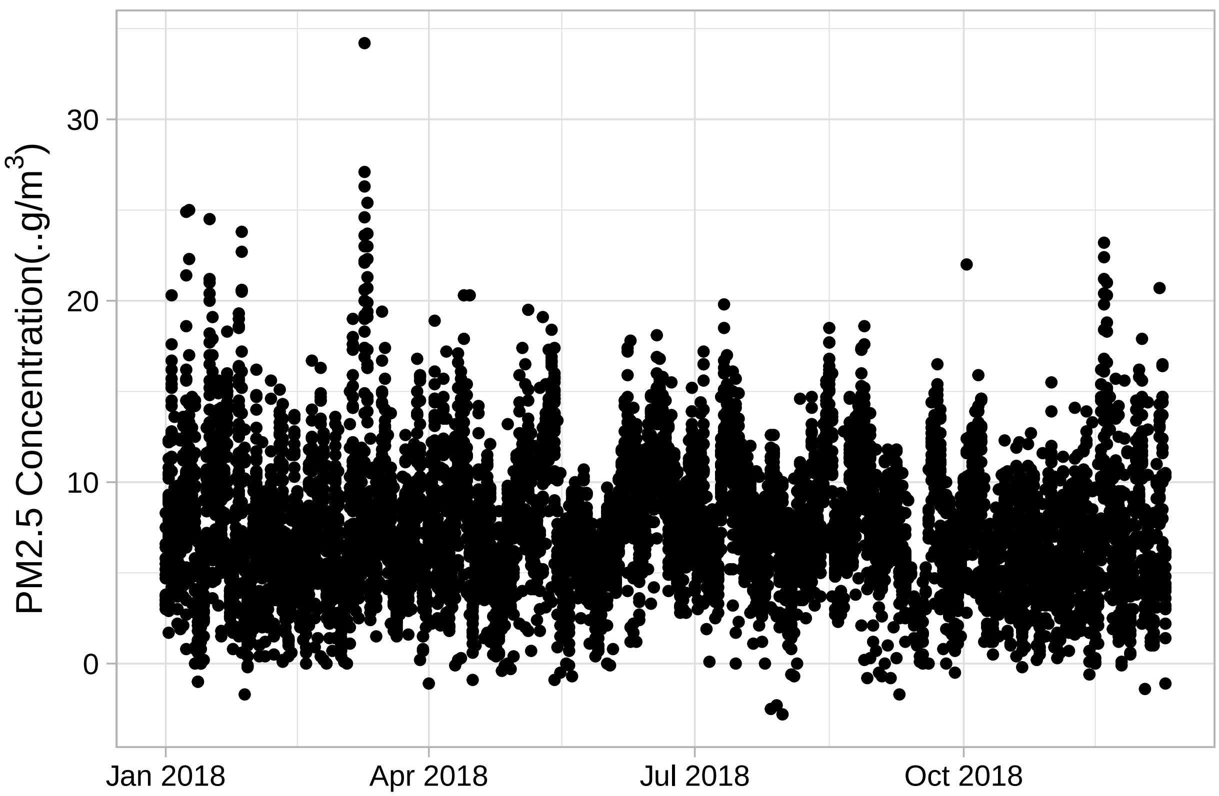
Run a hierarchical (mixed-effects) model

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(EPAair, aes(x = Date, y = Daily.Mean.PM2.5.Concentration))+
  geom_point()+
  labs(x="")+
  ylab(expression(paste("PM2.5 Concentration(\U003BCg/m"^3,")")))
```



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.

```
EPAair = EPAair[order(EPAair[,'Date'],-EPAair[,'Site.ID']),]
EPAair = EPAair[!duplicated(EPAair$Date),]

PM2.5Test_auto <- lme(data = EPAair,
                      Daily.Mean.PM2.5.Concentration ~ Date,
                     random = ~1|Site.Name)
PM2.5Test_auto
```

```
## Linear mixed-effects model fit by REML
##   Data: EPAair
##   Log-restricted-likelihood: -928.6076
##   Fixed: Daily.Mean.PM2.5.Concentration ~ Date
##  (Intercept)         Date
## 90.465022634 -0.004727976
##
## Random effects:
##  Formula: ~1 | Site.Name
##         (Intercept) Residual
## StdDev:    1.650184 3.559209
##
```

```
## Number of Observations: 343
## Number of Groups: 3
```

```r
ACF(PM2.5Test_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
PM2.5Test_mixed <- lme(data = EPAair,
                    Daily.Mean.PM2.5.Concentration ~ Date,
                    random = ~1|Site.Name,
                    #specify autocorrelation structure of order 1
                    correlation = corAR1(form = ~ Date|Site.Name, value = 0.513),
                    #define method as restricted maximum likelihood
                    method = "REML")
summary(PM2.5Test_mixed)
```

```
## Linear mixed-effects model fit by REML
##  Data: EPAair
##        AIC      BIC   logLik
##   1756.622 1775.781 -873.311
##
## Random effects:
##  Formula: ~1 | Site.Name
##         (Intercept) Residual
## StdDev: 0.001019731 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: The p-value is 0.21 > 0.05, which indicates there is not a significant trend in PM2.5 concentration 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.

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

```
## Generalized least squares fit by REML
##   Model: Daily.Mean.PM2.5.Concentration ~ Date
##   Data: EPAair
##        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(PM2.5Test_mixed, PM2.5Test_fixed)
```

```
##                 Model df      AIC      BIC    logLik   Test  L.Ratio
## PM2.5Test_mixed     1  5 1756.622 1775.781 -873.3110
## PM2.5Test_fixed     2  3 1865.202 1876.698 -929.6011 1 vs 2 112.5802
```

```
##                    p-value
## PM2.5Test_mixed
## PM2.5Test_fixed  <.0001
```

Which model is better?

> ANSWER: The mixed effect model is better due to the lower AIC

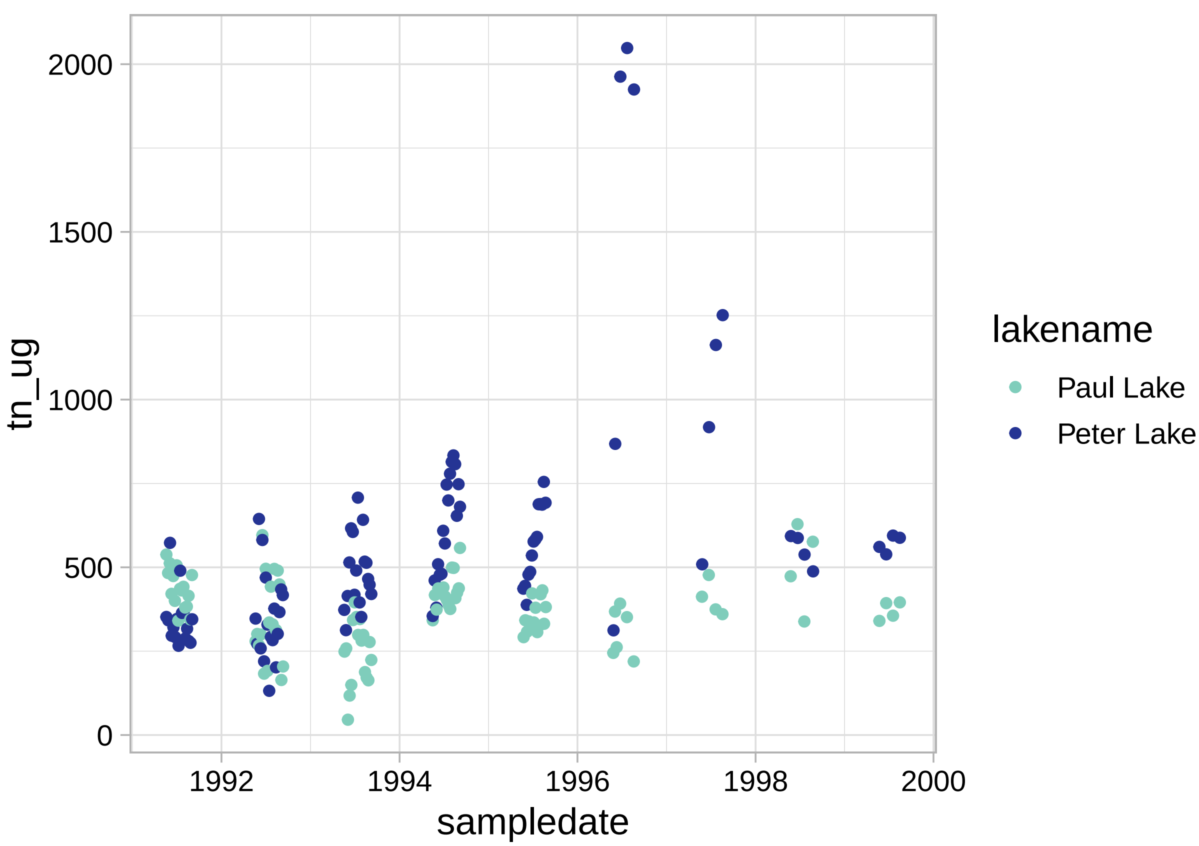
Run a Mann-Kendall test

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).

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

# Initial visualization of data
ggplot(PeterPaul.surface, aes(x = sampledate, y = tn_ug, color = lakename)) +
  geom_point() +
  scale_color_manual(values = c("#7fcdbb", "#253494"))
```



```r
# Split dataset by lake
Peter.surface <- filter(PeterPaul.surface, lakename == "Peter Lake")
Paul.surface <- filter(PeterPaul.surface, lakename == "Paul Lake")

mk.test(Peter.surface$tn_ug)
```

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

```r
mk.test(Paul.surface$tn_ug)
```

```
##
##  Mann-Kendall trend test
##
## data:  Paul.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
#Alternative hypothesis: there is a trend happenning over time

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

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

```r
# Run separate Mann-Kendall for each change point
mk.test(Peter.surface$tn_ug[1:35])
```

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

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

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

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

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

```r
# Run another Mann-Kendall for the second change point
mk.test(Peter.surface$tn_ug[36:56])
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.surface$tn_ug[36:56]
## z = -1.0569, n = 21, p-value = 0.2906
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
##  -36.0000000 1096.6666667   -0.1714286
```

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

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

What are the results of this test?

> ANSWER: There is not a trend over time in total N surface concentration in Paul lake.(Mann-Kendall test, p-value = 0.7) For Peter lake, there is a trend over time in total N surface concentrations, and the two change points are at 1993-06-05 and 1994-06-30.(pettitt test, p-value <0.0001)

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

```
ggplot(PeterPaul.surface, aes(x = sampledate, y = tn_ug, color = lakename)) +
  geom_point() +
  geom_vline(xintercept = as.Date("1993-06-02"),color="yellow", lty = 2)+
  geom_vline(xintercept = as.Date("1994-06-29"),color="yellow", lty = 2)+
  scale_color_manual(values = c("#7fcdbb", "#253494"))+
  labs(x = "", y = "total N(\U003BCg/L)")
```