Assignment 8: Time Series Analysis

Xin Wang

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 did.

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.

```r
AQPM25 <- read.csv("../Data/Raw/EPAair_PM25_NC2018_raw.csv")
PPnutrient <- read.csv("../Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaul_Processed.csv")
AQPM25$Date <- as.Date(AQPM25$Date,"%m/%d/%y")
PPnutrient$sampledate <- as.Date(PPnutrient$sampledate,"%Y-%m-%d")
wang <- theme_classic() +
  theme(plot.title=element_text(size = 18,hjust = 0.5),
        panel.background=element_rect(fill="white",color="grey30"),
        axis.title = element_text(size = 15),
        legend.title = element_text(size = 15), legend.text = element_text(size = 12),
        legend.margin=margin(6,6,6,6))
```

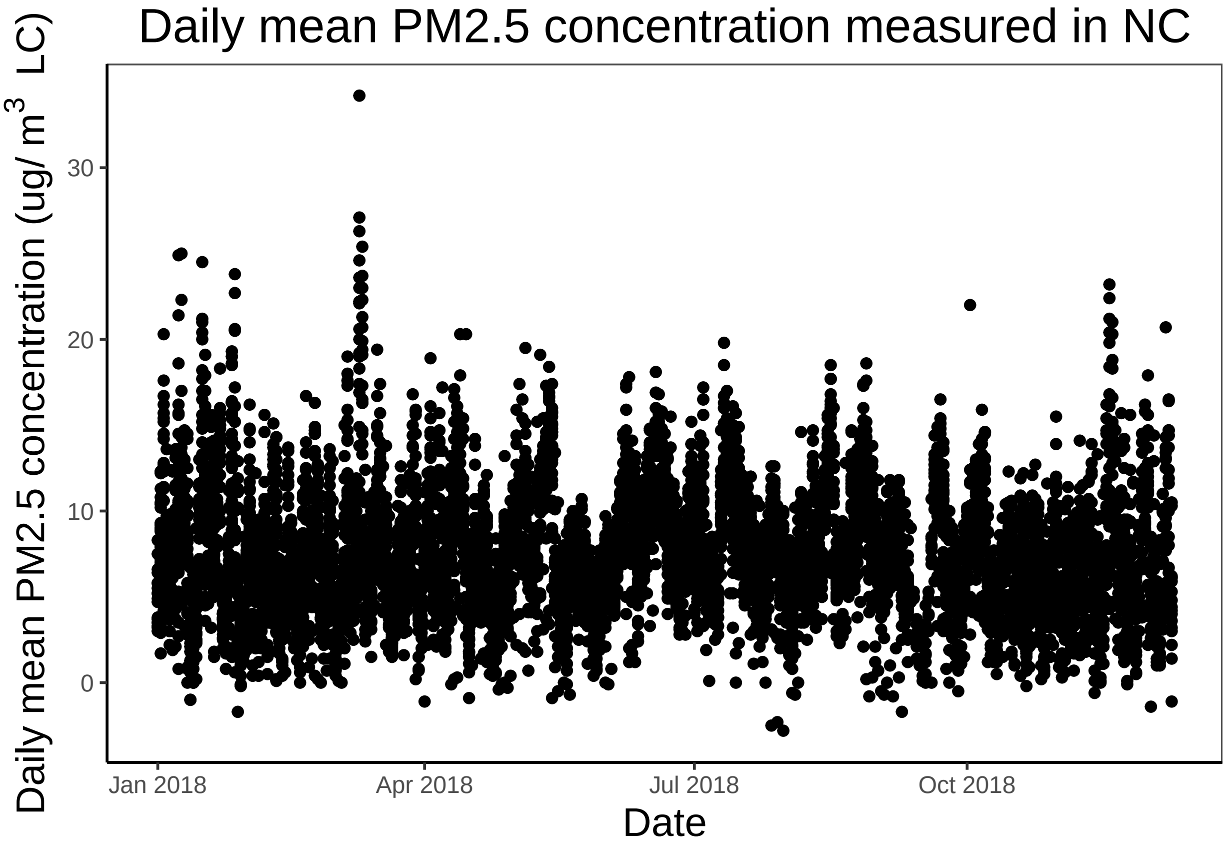
Run a hierarchical (mixed-effects) model

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(AQPM25,aes(x=Date, y=Daily.Mean.PM2.5.Concentration)) +
  geom_point() + wang +
  labs(title = "Daily mean PM2.5 concentration measured in NC",
       y="Daily mean PM2.5 concentration (ug/"~m^3~" LC)")
```



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.

```
PM2.5 = AQPM25[order(AQPM25[,'Date'],-AQPM25[,'Site.ID']),]
PM2.5 = PM2.5[!duplicated(PM2.5$Date),]
#3c
temp.auto <- lme(data = PM2.5,
                 Daily.Mean.PM2.5.Concentration ~ Date * Site.Name,
                 random = ~1|Site.Name)
ACF(temp.auto)
```

```
##    lag          ACF
## 1    0  1.000000000
## 2    1  0.515034682
## 3    2  0.196472686
## 4    3  0.117571590
## 5    4  0.124699761
## 6    5  0.098751072
## 7    6  0.057376429
## 8    7 -0.055302895
## 9    8  0.016831748
## 10   9  0.015394253
## 11  10 -0.000129952
## 12  11 -0.020530515
## 13  12 -0.045028734
## 14  13 -0.056146236
## 15  14 -0.066456910
## 16  15 -0.125049312
## 17  16 -0.055982362
## 18  17  0.002738986
## 19  18  0.025107752
## 20  19 -0.015544877
## 21  20 -0.144517752
## 22  21 -0.156619351
## 23  22 -0.060828900
## 24  23  0.003863609
## 25  24  0.042413623
## 26  25  0.001161753
```

```
#3d
mixed <- lme(data = PM2.5,
                  Daily.Mean.PM2.5.Concentration ~ Date * Site.Name,
                  random = ~1|Site.Name,
                  correlation = corAR1(value = 0.515),
                  method = "REML")
summary(mixed)
```

```
## Warning in pt(-abs(tVal), fDF): NaNs produced
```

```
## Linear mixed-effects model fit by REML
##  Data: PM2.5
##        AIC      BIC    logLik
##   1760.112 1794.492 -871.0558
##
## Random effects:
##  Formula: ~1 | Site.Name
##         (Intercept) Residual
## StdDev:   0.9006687 3.611323
##
## Correlation Structure: AR(1)
##  Formula: ~1 | Site.Name
##  Parameter estimate(s):
##       Phi
## 0.5340768
## Fixed effects: Daily.Mean.PM2.5.Concentration ~ Date * Site.Name
##                                     Value Std.Error  DF    t-value p-value
## (Intercept)                     10160.528 17547.173 337  0.5790407  0.5629
## Date                               -0.571     0.986 337 -0.5788390  0.5631
## Site.NameMillbrook School      -10379.545 17548.887   0 -0.5914646     NaN
## Site.NameTriple Oak            -10074.878 17547.290   0 -0.5741558     NaN
## Date:Site.NameMillbrook School      0.584     0.987 337  0.5917378  0.5544
## Date:Site.NameTriple Oak            0.567     0.986 337  0.5743778  0.5661
##  Correlation:
##                                (Intr) Date St.NMS St.NTO D:S.NS
## Date                           -1
## Site.NameMillbrook School      -1      1
## Site.NameTriple Oak            -1      1    1
## Date:Site.NameMillbrook School  1     -1   -1     -1
## Date:Site.NameTriple Oak        1     -1   -1     -1      1
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.34196751 -0.62767202 -0.08684712  0.59118660  3.37661054
##
## Number of Observations: 343
## Number of Groups: 3
```

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

> ANSWER: There is no significant increasing or decreasing trend in PM2.5 concentrations in 2018 (p value of Date is 0.56).

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

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

```
## Warning in anova.lme(mixed, fixed): fitted objects with different fixed
## effects. REML comparisons are not meaningful.
```

```
##       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## mixed     1  9 1760.111 1794.492 -871.0558
## fixed     2  3 1865.202 1876.698 -929.6011 1 vs 2 117.0906  <.0001
```

Which model is better?

> ANSWER: The mixed effect model has lower AIC than the fixed effect one and is therefore better than the latter.

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).

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

```
## Warning: package 'bindrcpp' was built under R version 3.5.2
```

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

```
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
```

```
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
```

```
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
```

```
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
```

```
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
```

```
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
```

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

```
## 
##  Mann-Kendall trend test
```

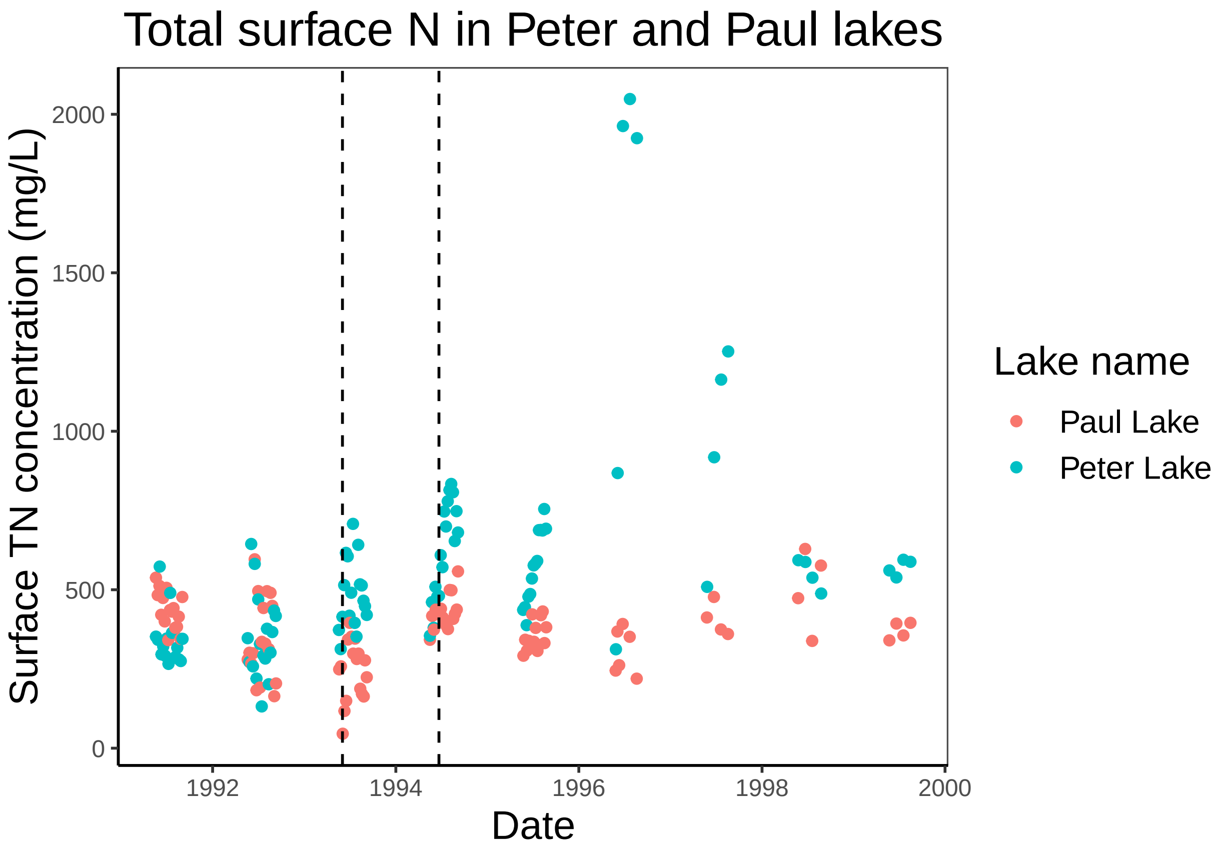
```
##
## data:  Peter.surface$tn_ug[57:98]
## z = 0.15172, n = 42, p-value = 0.8794
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
##   15.0000000 8514.3333333    0.0174216
```

What are the results of this test?

> ANSWER: There is a significant trend in total N surface concentration in Peter Lake (Mann-Kendall test, p=3.039e-13) over time, but not a significant trend in Paul Lake (p=0.7258). There is a significant changepoint in surface TN in Peter Lake on 1993-06-02 (Pettitt's Test, p=3.744e-10) and a second changepoint on 1994-6-22 (Pettitt's Test, p=0.001213).

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,color=lakename)) +
  geom_point() + wang +
  geom_vline(xintercept = as.Date("1993/06/02"),lty=2) +
  geom_vline(xintercept = as.Date("1994/06/22"),lty=2) +
  labs(title="Total surface N in Peter and Paul lakes",
      x="Date",y="Surface TN concentration (mg/L)",color="Lake name")
```


Total surface N in Peter and Paul lakes
Surface TN concentration (mg/L)
Date
Lake name
● Paul Lake
● Peter Lake