Assignment 8: Time Series Analysis

Siying Chen

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'll use the NTL_LTER nutrient and chemical/physical dataset along with the phytoplankton dataset I found online and perform time series analysis to find out the relationship between phytoplankton and nutrients/D.O./irradiance 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.

```
getwd()
```

```
## [1] "/Users/Sylvia/Downloads/ENV872/ENV872"
```

```
library(tidyverse)
```

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

## 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_confli
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(nlme)
```

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

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

```
library(lsmeans)
```

```
## Loading required package: emmeans
```

```
## The 'lsmeans' package is now basically a front end for 'emmeans'.
## Users are encouraged to switch the rest of the way.
## See help('transition') for more information, including how to
## convert old 'lsmeans' objects and scripts to work with 'emmeans'.
```

```
library(multcompView)
library(trend)

PM2.5 <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv")
PM2.5$Date <- as.Date(PM2.5$Date, format = "%m/%d/%y")

PeterPaul.nutrients <- read.csv("./Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaul_Processed.csv")
PeterPaul.nutrients$sampledate <- as.Date(PeterPaul.nutrients$sampledate, format = "%Y-%m-%d")

mytheme <- theme_bw(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "bottom",
        panel.grid.major = element_line(size = 0.5, linetype = 'solid'),
        panel.grid.minor = element_line(size = 0.25, linetype = 'dashed'),
        title = element_text(face = "bold"))
theme_set(mytheme)
```

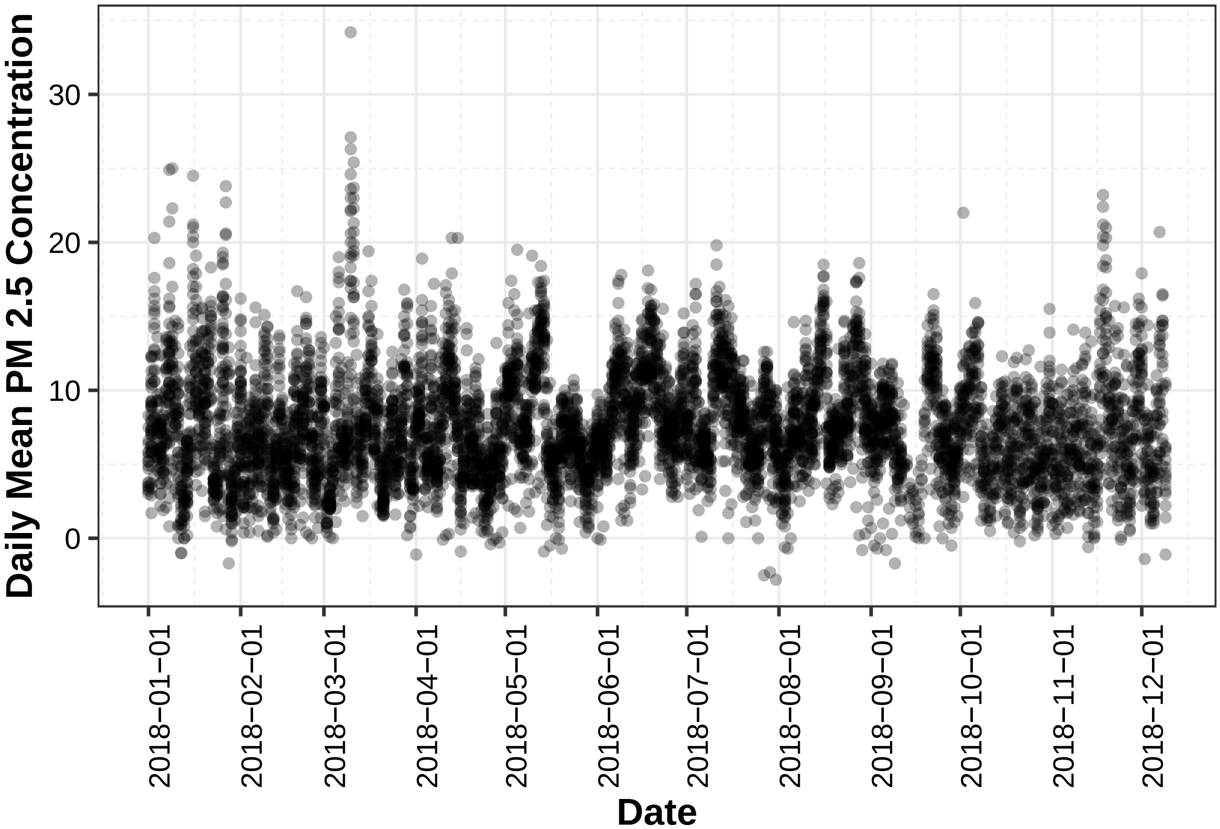
Run a hierarchical (mixed-effects) model

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.

```
PM2.5.plot <- ggplot(PM2.5, aes(x = Date, y = Daily.Mean.PM2.5.Concentration)) +
  geom_point(alpha = 0.3) +
  ylab("Daily Mean PM 2.5 Concentration") +
  scale_x_date(date_breaks = "1 month") +
  theme(axis.text.x = element_text(angle=90))
print(PM2.5.plot)
```



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.

```
# Remove duplicate data
PM2.5 = PM2.5[order(PM2.5[,'Date'],-PM2.5[,'Site.ID']),]
PM2.5 = PM2.5[!duplicated(PM2.5$Date),]

# Temporal autocorrelation test
PM2.5.auto <- lme(data = PM2.5,
                  Daily.Mean.PM2.5.Concentration ~ Date,
                  random = ~1|Site.Name)
PM2.5.auto
```

```
## Linear mixed-effects model fit by REML
##   Data: PM2.5
##   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(PM2.5.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
# Mixed effect model
PM2.5.mixed <- lme(data = PM2.5,
                   Daily.Mean.PM2.5.Concentration ~ Date,
                   random = ~1|Site.Name,
                   correlation = corAR1(form = ~ Date|Site.Name, value = 0.514),
                   method = "REML")
summary(PM2.5.mixed)
```

```
## Linear mixed-effects model fit by REML
##  Data: PM2.5
##        AIC      BIC   logLik
##   1756.622 1775.781 -873.311
##
## Random effects:
```

```
##  Formula: ~1 | Site.Name
##         (Intercept) Residual
## StdDev: 0.001028133 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 is no significant trend detected in PM2.5 concentrations in 2018 since p-value is greater than 0.05.

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.

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

```
## Generalized least squares fit by REML
##   Model: Daily.Mean.PM2.5.Concentration ~ Date * Site.Name
##   Data: PM2.5
##        AIC      BIC    logLik
##   1865.812 1892.552 -925.9059
##
## Coefficients:
##                                    Value Std.Error    t-value p-value
## (Intercept)                    11540.630 15142.989  0.7621104  0.4465
## Date                              -0.649     0.851 -0.7618782  0.4467
## Site.NameMillbrook School     -11622.193 15144.205 -0.7674350  0.4434
## Site.NameTriple Oak           -11446.924 15143.029 -0.7559203  0.4502
## Date:Site.NameMillbrook School     0.654     0.851  0.7677302  0.4432
## Date:Site.NameTriple Oak           0.644     0.851  0.7561773  0.4501
##
##  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 residuals:
##         Min          Q1         Med          Q3         Max
## -2.38548587 -0.63305418 -0.09196293  0.58568909  3.41224448
##
## Residual standard error: 3.561158
## Degrees of freedom: 343 total; 337 residual
```

```
anova(PM2.5.mixed, PM2.5.fixed)
```

```
## Warning in anova.lme(PM2.5.mixed, PM2.5.fixed): fitted objects with
## different fixed effects. REML comparisons are not meaningful.
```

```
##             Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## PM2.5.mixed     1  5 1756.622 1775.781 -873.3110
## PM2.5.fixed     2  7 1865.812 1892.552 -925.9059 1 vs 2 105.1899  <.0001
```

Which model is better?

> ANSWER: The mixed effects model is better since it has a lower AIC. P-value is <.0001, which means the two models have a significantly different fit.

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).

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

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

# Run a Mann-Kendall test
mk.test(Peter.nutrients.surface$tn_ug)
```

```
##
##  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
```

```r
# Test for change point
pettitt.test(Peter.nutrients.surface$tn_ug) # change point at k=36
```

```
##
##  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
```

```r
# Run separate Mann-Kendall for each change point
mk.test(Peter.nutrients.surface$tn_ug[1:35]) # no trend detected
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.nutrients.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.nutrients.surface$tn_ug[36:98]) # trend detected
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.nutrients.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
# Is there a second change point?
pettitt.test(Peter.nutrients.surface$tn_ug[36:98]) #second change point at 36+21=57
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Peter.nutrients.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.nutrients.surface$tn_ug[36:56]) # no significatn trend
```

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

```
## data:  Peter.nutrients.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.nutrients.surface$tn_ug[57:98]) # no significatn trend
```

```
##
##  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
```

```
# Run the same test for Paul Lake.
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
```

```
pettitt.test(Paul.nutrients.surface$tn_ug)
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Paul.nutrients.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: For Peter lake, there's a significant trend detected at time location 36 (1993-06-02), the second Mann-Kendall test reveals that there's a second change point from time location 36 to 98, the second change point it at 36+21=57 (1994-06-29), there is no more trend detected in the rest of the time segments. For Paul lake, there's no significant trend detected.

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

```
TN_mktest <- ggplot(PeterPaul.nutrients.surface, aes(x = sampledate, y = tn_ug, color = lakename)) +
  geom_point() +
  scale_color_manual(values = c("#7fcdbb", "#253494")) +
  geom_vline(xintercept = as.Date("1993-06-02"),
```

```
              color = "purple", lty = 2) +
  geom_vline(xintercept = as.Date("1994-06-29"),
              color = "red", lty = 2) +
  labs(x = "Sample Date", y = "TN (µg/L)")
print(TN_mktest)
```