# Assignment 8: Time Series Analysis

*Gaby Garcia*

## 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!

## 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.

### Upload Datasets

```
setwd("~/Desktop/Environmental Data Analytics/Environmental_Data_Analytics/Data/Processed")
PeterPaul.nutrients <- read.csv("NTL-LTER_Lake_Nutrients_PeterPaul_Processed.csv")
```

### Change factor to a date format

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

### EPA Air Quality Raw Dataset for PM2.5 in 2018

```
setwd("~/Desktop/Environmental Data Analytics/Environmental_Data_Analytics/Data/Processed")
PM2018<-read.csv("EPAair_PM25_NC2018_raw.csv")
```

### Change factor to a date format

```
PM2018$Date<-as.Date(PM2018$Date, format="%m/%d/%y")
```

### GGPlot Theme

```
library(ggplot2)
gabytheme <- theme_bw(base_size = 14) +
  theme(plot.title=element_text(face="bold", size="14", color="hotpink4", hjust=0.5),
        axis.title=element_text(face="bold.italic", size=11, color="black"),
axis.text = element_text(face="bold", size=7, color = "black"),
panel.background=element_rect(fill="gray88", color="darkblue"),
panel.border = element_rect(color = "black", size = 2),
legend.position = "top", legend.background = element_rect(fill="white", color="black"),
            legend.key = element_rect(fill="transparent", color="NA"))
theme_set(gabytheme)
```

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.

```
library(nlme)
PMTest<- lme(data = PM2018,
                    Daily.Mean.PM2.5.Concentration~ Date,
             #fixed effects ANCOVA model
                    random = ~1|Site.Name)
#specifying a random effect
```

### Mixed Effects Model

```
PMTest
```

```
## Linear mixed-effects model fit by REML
##   Data: PM2018
##   Log-restricted-likelihood: -20297.38
##   Fixed: Daily.Mean.PM2.5.Concentration ~ Date
## (Intercept)        Date
## 20.14183588 -0.00074241
##
## Random effects:
##  Formula: ~1 | Site.Name
##         (Intercept) Residual
## StdDev:    1.841425 3.457061
##
## Number of Observations: 7611
## Number of Groups: 24
```

### Summary of Mixed Effects Model

```
summary(PMTest)
```

```
## Linear mixed-effects model fit by REML
##  Data: PM2018
##        AIC      BIC    logLik
##   40602.76 40630.51 -20297.38
##
## Random effects:
##  Formula: ~1 | Site.Name
##         (Intercept) Residual
## StdDev:    1.841425 3.457061
##
## Fixed effects: Daily.Mean.PM2.5.Concentration ~ Date
##                 Value Std.Error   DF   t-value p-value
## (Intercept) 20.141836  7.382570 7586  2.728296  0.0064
## Date        -0.000742  0.000417 7586 -1.779991  0.0751
##  Correlation:
##      (Intr)
## Date -0.999
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -3.4251256 -0.6846871 -0.1385351  0.5919707  7.9199389
##
## Number of Observations: 7611
## Number of Groups: 24
```

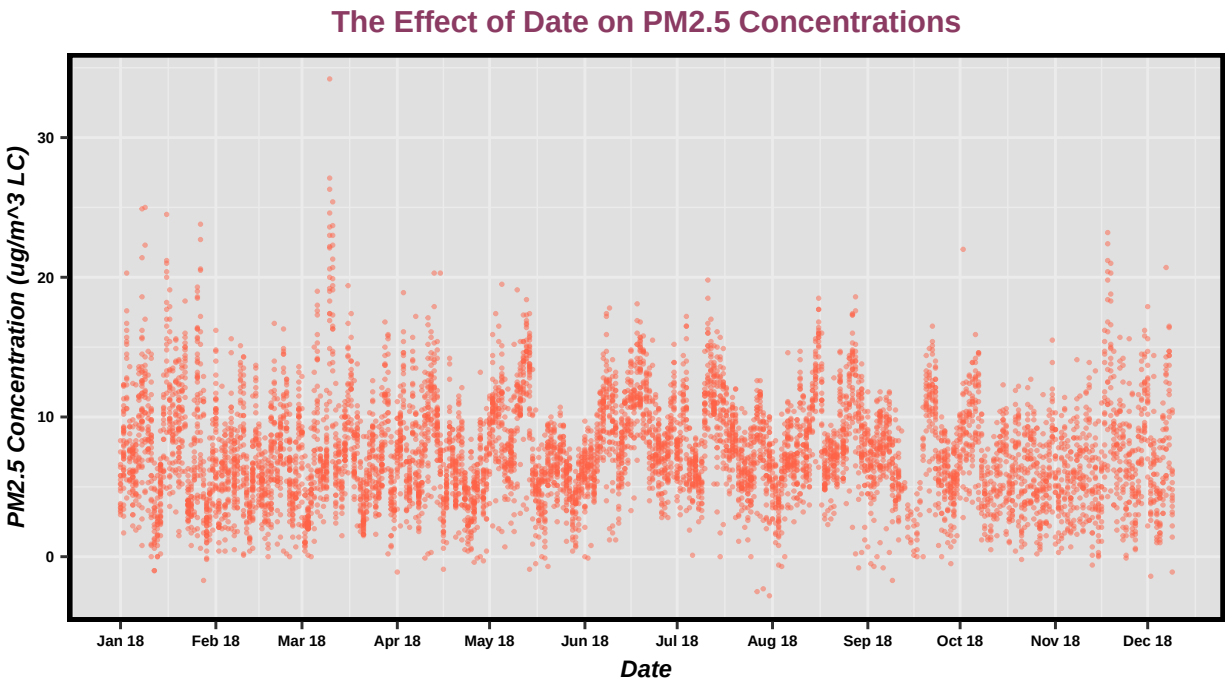
## 3a.

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

```
library(ggplot2)
library(scales)
ggplot(PM2018, aes(x = Date, y =Daily.Mean.PM2.5.Concentration)) +
```

```
  geom_point(size=0.5, alpha=0.5, color="tomato") +

  scale_color_manual(values = c("#7fcdbb"))+
   labs(title="The Effect of Date on PM2.5 Concentrations", x="Date",
       y="PM2.5 Concentration (ug/m^3 LC)") +
  scale_x_date(labels = date_format("%b %y"), breaks = date_breaks("1 month"))
```




## 3b.

Insert the following line of code into your R chunk. This will eliminate duplicate measurements on single dates for each site.

```
PM2018 = PM2018[order(PM2018[,'Date'],-PM2018[,'Site.ID']),]
PM2018 = PM2018[!duplicated(PM2018$Date),]
```

### Creating a new model with cleaned data

```
PMTestClean<-lme(data = PM2018,
                    Daily.Mean.PM2.5.Concentration~ Date, #fixed effects ANCOVA model
                    random = ~1|Site.Name)            #specifying a random effect
```

## 3c.

Determine the temporal autocorrelation in your model.

```
ACF(PMTestClean)
```

```
##    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
```

**2nd value=always falls between 0 and 1(0-100%)**

**In this case, it's 0.51-51% of variability associated with time is autocorrelated from previous dates**

# 3d.

Run a mixed effects model.

```
PMTest.mixed <- lme(data = PM2018, Daily.Mean.PM2.5.Concentration~ Date,
                    random = ~1|Site.Name,
                    correlation = corAR1(form = ~ Date|Site.Name, value = 0.51383),
                   method = "REML")
```

## Summary of Mixed Effects Model

```
summary(PMTest.mixed)
```

```
## Linear mixed-effects model fit by REML
##  Data: PM2018
##        AIC      BIC   logLik
```

```
##   1756.622 1775.781 -873.311
##
## Random effects:
##  Formula: ~1 | Site.Name
##         (Intercept) Residual
## StdDev: 0.001024989 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: According to the summary of our mixed effects model, the coefficient for the parameter of Date is -0.004 (p=0.21, t=-1.24, df=339). However, according to the summary, Date is not a significant predictor of PM2.5 concentration because the p-value for Date is 0.21, which is above 0.05. Thus, there is not a significant trend in PM2.5 concentrations.

## 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.

```
PMTestFixed<- gls(data = PM2018,
                  Daily.Mean.PM2.5.Concentration~ Date, method="REML")
#fixed effects ANCOVA model
```

### Summary of Fixed Effects Model

```
summary(PMTestFixed)
```

```
## Generalized least squares fit by REML
##   Model: Daily.Mean.PM2.5.Concentration ~ Date
##   Data: PM2018
##        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 Comparing Mixed Effects Model and Fixed Effects Model

```
anova(PMTest.mixed, PMTestFixed)
```

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

Which model is better?

> ANSWER: According to the ANOVA test, there is more variability in model structure (error) accounted for by the mixed effects model that includes Site Name as a random effect. We know this because the AIC score of the PMTest.mixed model is 1756.622, compared to the Fixed Effect model's AIC score of 1865.6202. The p-value of <0.0001 indicates that the model fit is significantly different between the two models. Thus, the Mixed Effects model is the best model.

### 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 Data Set to focus on surface values

```
library(tidyverse)
###Choose only surface concentrations at depth=0.
PeterPaul.nutrients.surface2<-
  PeterPaul.nutrients %>%
 ###Remove lakeid, depth_id, and comments.
```

```
  filter(depth == 0) %>%
  filter(!is.na(tn_ug))  ###Remove NA's from tn_ug column
```

### Split Dataset by Lake

```
Peter.nutrients.surface2 <- filter(PeterPaul.nutrients.surface2, lakename == "Peter Lake")
Paul.nutrients.surface2 <- filter(PeterPaul.nutrients.surface2, lakename == "Paul Lake")
```

### Run a Mann Kendall test for Peter Lake

```
library(trend)
Peter.nutrients.surface2$tn_ug<-as.numeric(Peter.nutrients.surface2$tn_ug)
mk.test(Peter.nutrients.surface2$tn_ug)
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.nutrients.surface2$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
```

### Run a Mann Kendall Test for Paul Lake

```
library(trend)
Paul.nutrients.surface2$tn_ug<-as.numeric(Paul.nutrients.surface2$tn_ug)
mk.test(Paul.nutrients.surface2$tn_ug)
```

```
##
##  Mann-Kendall trend test
##
## data:  Paul.nutrients.surface2$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: For Peter Lake, the z-value is 7.29 (p=3.04e-13, n=98), so we see a positive trend in Total Nitrogen surface concentrations over time. The p-value is 3.04e-13,so we reject the null hypothesis that the data come from a population with independent realizations and are identically distributed . For Paul Lake(z=-0.35, n=99, p=0.73), the z-value is -0.35, so we see a negative trend in Total Nitrogen surface concentrations over time. The p-value for Paul Lake is listed as 0.73, so we retain the null hypothesis that the data come from a population with independent realizations and are identically distributed.

# Pettitt's Test

Pettitt's test is also included in the `trend` package. This nonparametric test will determine whether there is a shift in the central tendency of the time series and will tell us at what point the changepoint occurs (if it detects one). Note: Pettitt's Test will only test for one changepoint, and further tests must be run if multiple change points are suspected.

### Pettitt's test for Peter Lake

```
pettitt.test(Peter.nutrients.surface2$tn_ug)
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Peter.nutrients.surface2$tn_ug
## U* = 1884, p-value = 3.744e-10
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              36
```

**Because the p-value is <0.05, the change point is significant. Given 1st change point for Peter Lake is 36, we scroll to observation 36 in data set, so first change point occurred in June 1993.**

# Run separate Mann-Kendall Test for each change point

```
mk.test(Peter.nutrients.surface2$tn_ug[1:35])
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.nutrients.surface2$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.nutrients.surface2$tn_ug[36:98])
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.nutrients.surface2$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
```

### Is there a second change point?

```
pettitt.test(Peter.nutrients.surface2$tn_ug[36:98])
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Peter.nutrients.surface2$tn_ug[36:98]
## U* = 560, p-value = 0.001213
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              21
```

36+20=56, so look at 56th row in datatable to see second change point (because there are only twenty observations in between). It occurred in June 1994.

## Run another Mann-Kendall for the second change point

### Now split dataset into three pieces

```
mk.test(Peter.nutrients.surface2$tn_ug[36:55])
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.nutrients.surface2$tn_ug[36:55]
## z = -1.2004, n = 20, p-value = 0.23
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##     S  varS   tau
## -38.0 950.0  -0.2
```

```
mk.test(Peter.nutrients.surface2$tn_ug[56:98])
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.nutrients.surface2$tn_ug[56:98]
## z = 0.48141, n = 43, p-value = 0.6302
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 4.700000e+01 9.130333e+03 5.204873e-02
```

**If z-score is positive, it's a positive trend. If z-score is negative, it is a negative trend**

**There is not a significant trend in nitrogen concentrations over time at Peter Lake for rows:36-55 because the p-value is 0.23, which is above 0.05. T**

**There is also not a significant trend in nitrogen concentrations over time for rows 56:98 because the p-value is 0.63, which is above 0.05.**

## Pettitt's test for Paul Lake

```
pettitt.test(Paul.nutrients.surface2$tn_ug)
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Paul.nutrients.surface2$tn_ug
## U* = 704, p-value = 0.09624
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              16
```

**Because the p-value is 0.09, which is greater than 0.05, the change point is not significant.**

## Run separate Mann-Kendall Test for each change point

```
mk.test(Paul.nutrients.surface2$tn_ug)
```

```
##
##  Mann-Kendall trend test
##
## data:  Paul.nutrients.surface2$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
```

**We have a non significant trend in Total Nitrogen over time at Paul Lake-p=0.87, z–0.15.**

## 5.

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

## Change Date to Date Object

```
Peter.nutrients.surface2$sampledate <- as.Date(Peter.nutrients.surface2$sampledate,
                                               format = "%Y-%m-%d")
```

```
ggplot(PeterPaul.nutrients.surface2, aes(x = sampledate, y = tn_ug, color = lakename)) +
  geom_point(alpha=0.9) +
  scale_color_manual(values = c("#7fcdbb", "#253494"))+
  geom_vline(xintercept=as.Date("1993-06-02"),color="253494", origin= "1970-01-01", lty=2) +
  geom_vline(xintercept=as.Date("1994-06-22"), color="253494", lty=2)+
  labs(title="The Effect of Sample Date on Total Nitrogen Concentrations",
       x="Sample Date",
       y="Total Nitrogen (mg/L)")+
  scale_x_date(labels = date_format("%m/%Y"), breaks = date_breaks("6 month"))
```