# Assignment 8: Time Series Analysis

*Walker Grimshaw*

## 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: I did look on the forum and brainstorm ideas, but I did not select an idea and post it to the Sakai forum.

### 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
suppressMessages(library(tidyverse))
suppressMessages(library(lubridate)) # easy date manipulation
suppressMessages(library(gridExtra)) # multiple plots in a figure
suppressMessages(library(RColorBrewer))
suppressMessages(library(FSA)) # dunn.test after Kruskal Wallace
suppressMessages(library(nlme))
suppressMessages(library(lsmeans))
suppressMessages(library(multcompView))
suppressMessages(library(trend))

## Import Data
```

```
EPA_PM25_Raw <- read.csv("../Data/Raw/EPAair_PM25_NC2018_raw.csv")
PP_Nutrients <- read.csv("../Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaul_Processed.csv")

## ggplot theme
WalkersTheme <- theme_bw(base_size = 12) +
  theme(legend.position = "top")

theme_set(WalkersTheme)

## set date formats
EPA_PM25_Raw$Date <- as.Date(EPA_PM25_Raw$Date, format = "%m/%d/%y")
PP_Nutrients$sampledate <- as.Date(PP_Nutrients$sampledate, format = "%Y-%m-%d")
```

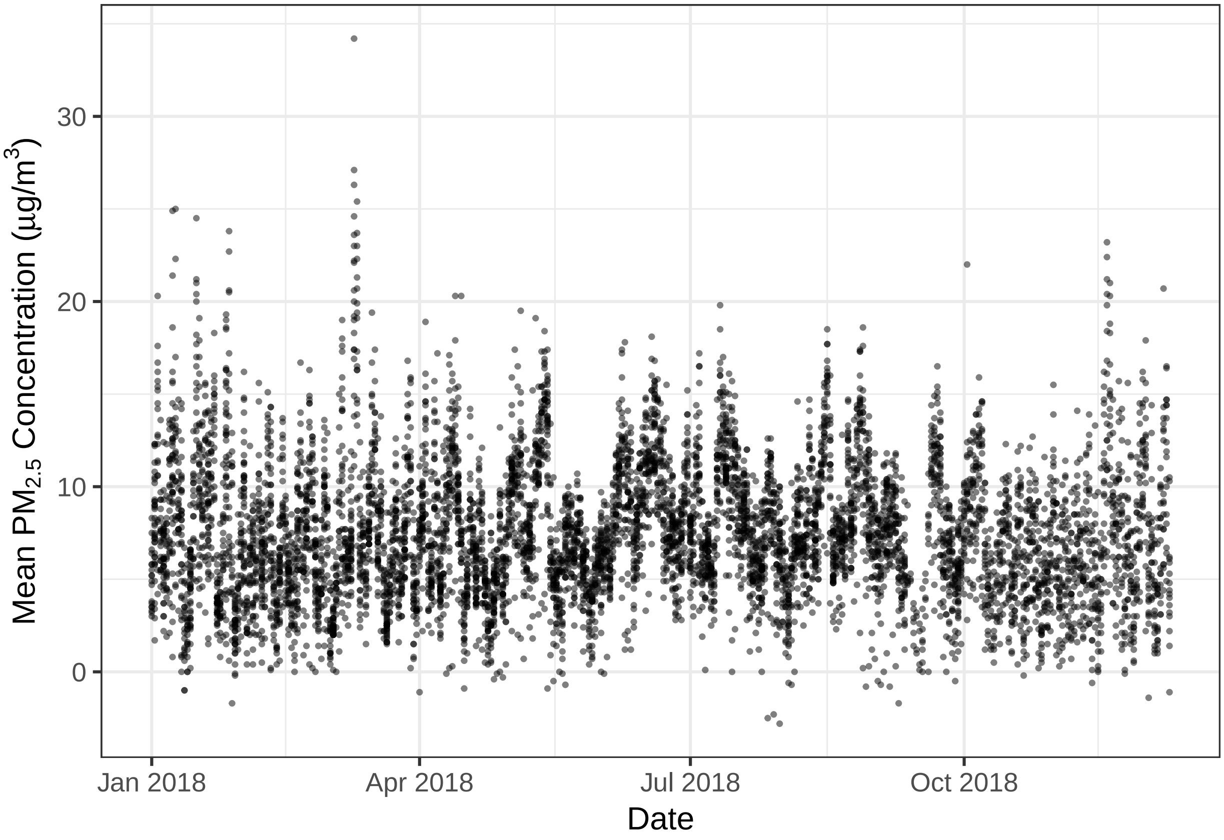
### Run a hierarchical (mixed-effects) model

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.

```
## plot of PM2.5 by date aggregated for all sites
ggplot(EPA_PM25_Raw, aes(x = Date, y = Daily.Mean.PM2.5.Concentration)) +
  geom_point(size = 0.5, alpha = 0.5) +
  labs(y = expression(paste("Mean PM"[2.5]* " Concentration (", mu, "g/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),]

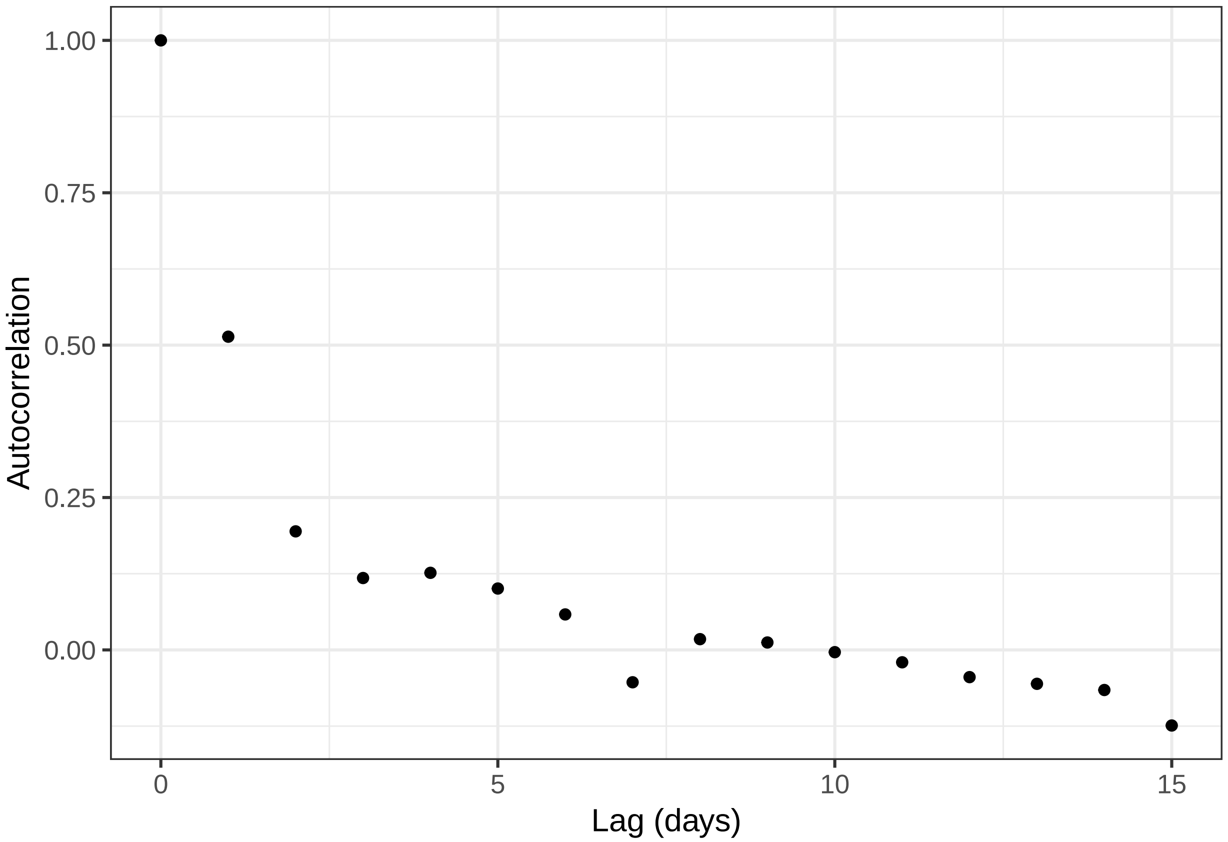
3c. Determine the temporal autocorrelation in your model.

3d. Run a mixed effects model.

```
## 3b
## run code to get only one data point per date

EPA_PM25_Unique = EPA_PM25_Raw[order(EPA_PM25_Raw[,'Date'],-EPA_PM25_Raw[,'Site.ID']),]
EPA_PM25_Unique = EPA_PM25_Unique[!duplicated(EPA_PM25_Unique$Date),]

## 3c Temporal Autocorrelation
## first create mixed effects model for PM2.5
PM25_trend_test <- lme(data = EPA_PM25_Unique,
                       Daily.Mean.PM2.5.Concentration ~ Date,
                       random = ~1|Site.Name)

## autocorrelation
PM25_auto <- ACF(PM25_trend_test, maxLag = 15); PM25_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
```

```
ggplot(PM25_auto, aes(x=lag, y=ACF)) +
  geom_point() +
  labs(x = "Lag (days)", y = "Autocorrelation")
```



```
## 3d mixed effects model
## repeated measures ANOVA
PM25_Trend <- lme(data = EPA_PM25_Raw,
                  Daily.Mean.PM2.5.Concentration ~ Date,
                  random = ~1|Site.Name)
PM25_Trend
```

```
## Linear mixed-effects model fit by REML
##   Data: EPA_PM25_Raw
##   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
```

```
## mixed effects model of PM2.5 by date, with site as a random effect
PM25_trend_mixed <- lme(data = EPA_PM25_Unique,
                      Daily.Mean.PM2.5.Concentration ~ Date,
                      #specify autocorrelation structure of order 1
                      random = ~1|Site.Name,
                      correlation = corAR1(form = ~ Date|Site.Name,
                                           value = 0.514),
                      method = "REML")
summary(PM25_trend_mixed)
```

```
## Linear mixed-effects model fit by REML
##  Data: EPA_PM25_Unique
##        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 a decreasing trend in PM2.5 concentrations in 2018, but it is not statistically significant (p=0.21).

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 effects model
PM25_trend_fixed <- gls(data = EPA_PM25_Unique,
                  Daily.Mean.PM2.5.Concentration ~ Date,
                  method = "REML")

## ANOVA to compare fixed effects and mixed effects model
anova(PM25_trend_mixed, PM25_trend_fixed)
```

```
##                  Model df      AIC      BIC    logLik   Test  L.Ratio
## PM25_trend_mixed     1  5 1756.622 1775.781 -873.3110
## PM25_trend_fixed     2  3 1865.202 1876.698 -929.6011 1 vs 2 112.5802
##                  p-value
## PM25_trend_mixed
## PM25_trend_fixed  <.0001
```

Which model is better?

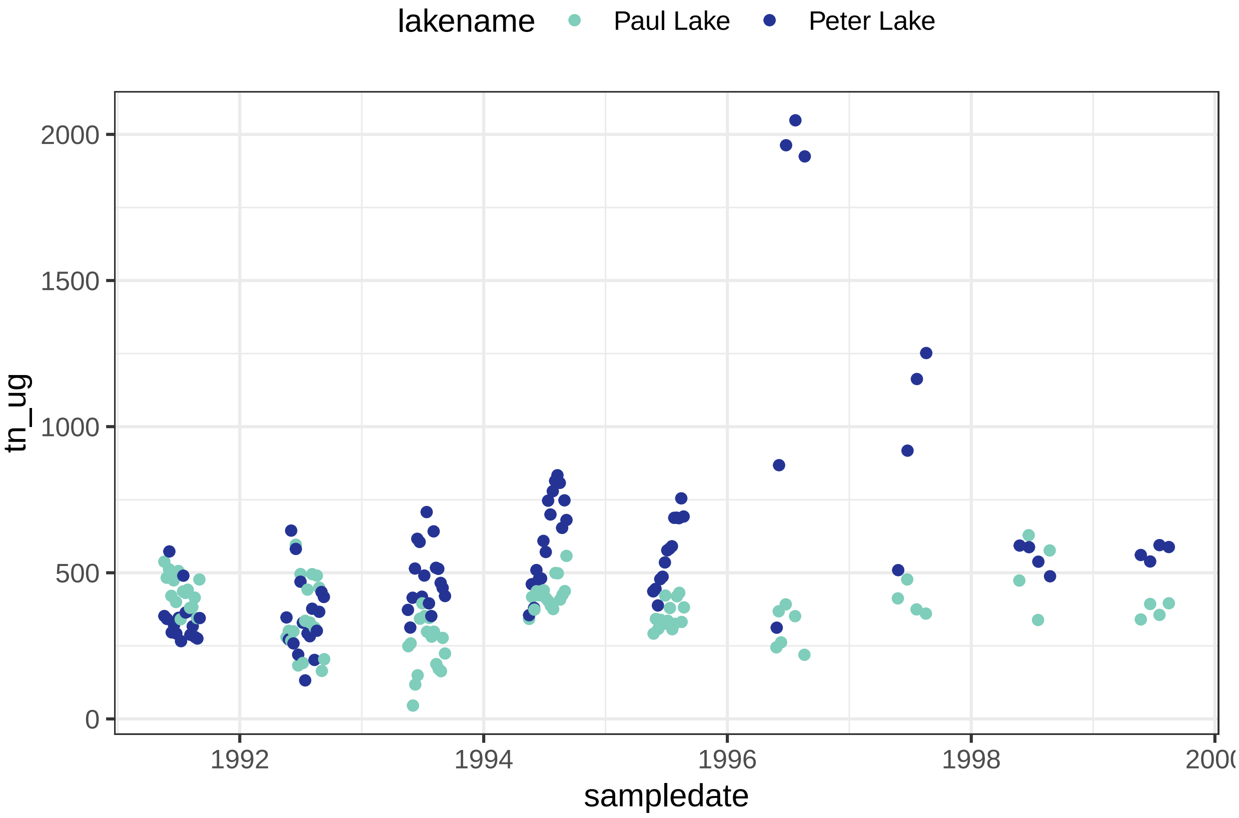
> ANSWER: The models are statistically significantly different, and the mixed effects model is better, as indicated by the lower AIC value.

### 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
PP_Nutrients_Surface <-
  PP_Nutrients %>%
  select(-lakeid, -depth_id, -comments) %>% # remove ones with minus sign
  filter(depth == 0) %>%
  filter(!is.na(tn_ug)) # remove nas associated with total phosphorous

# Initial visualization of data
## Paul lake acts as a control
ggplot(PP_Nutrients_Surface, aes(x = sampledate, y = tn_ug, color = lakename)) +
  geom_point() +
  scale_color_manual(values = c("#7fcdbb", "#253494"))
```



```
# Split dataset by lake
Peter_Nutrients_Surface <- filter(PP_Nutrients_Surface, lakename == "Peter Lake")
Paul_Nutrients_Surface <- filter(PP_Nutrients_Surface, lakename == "Paul Lake")
```

```r
# Run a Mann-Kendall test
# trend over time
# z indicates direction of trend and magnitude of confidence (p-value)
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
# there appears to be at least one break point for Peter Lake

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

```
##
##  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
# there is a change point at data point 36, 1993-06-02

# Test for second change point
pettitt.test(Peter_Nutrients_Surface$tn_ug[36:98])
```

```
##
##  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
# second change point at data point 36+21=57, 1994-06-29

pettitt.test(Peter_Nutrients_Surface$tn_ug[57:98])
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Peter_Nutrients_Surface$tn_ug[57:98]
## U* = 127, p-value = 0.5584
## alternative hypothesis: two.sided
## sample estimates:
```

```
## probable change point at time K
##                               11
```

```
# unlikely third change point

## run Mann-Kendall Test for each of the three intervals for Peter Lake
mk.test(Peter_Nutrients_Surface$tn_ug[1:35]) # non-significant decrease
```

```
##
##  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
```

```
mk.test(Peter_Nutrients_Surface$tn_ug[36:57]) # non-significant decrease
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter_Nutrients_Surface$tn_ug[36:57]
## z = -0.56396, n = 22, p-value = 0.5728
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
##  -21.00000000 1257.66666667   -0.09090909
```

```
mk.test(Peter_Nutrients_Surface$tn_ug[58:98]) # non-significant increase
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter_Nutrients_Surface$tn_ug[58:98]
## z = 0.14602, n = 41, p-value = 0.8839
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 1.400000e+01 7.926667e+03 1.707317e-02
```

```
## Mann-Kendall for Paul Lake
mk.test(Paul_Nutrients_Surface$tn_ug)
```

```
##
##  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 for Paul Lake
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
```

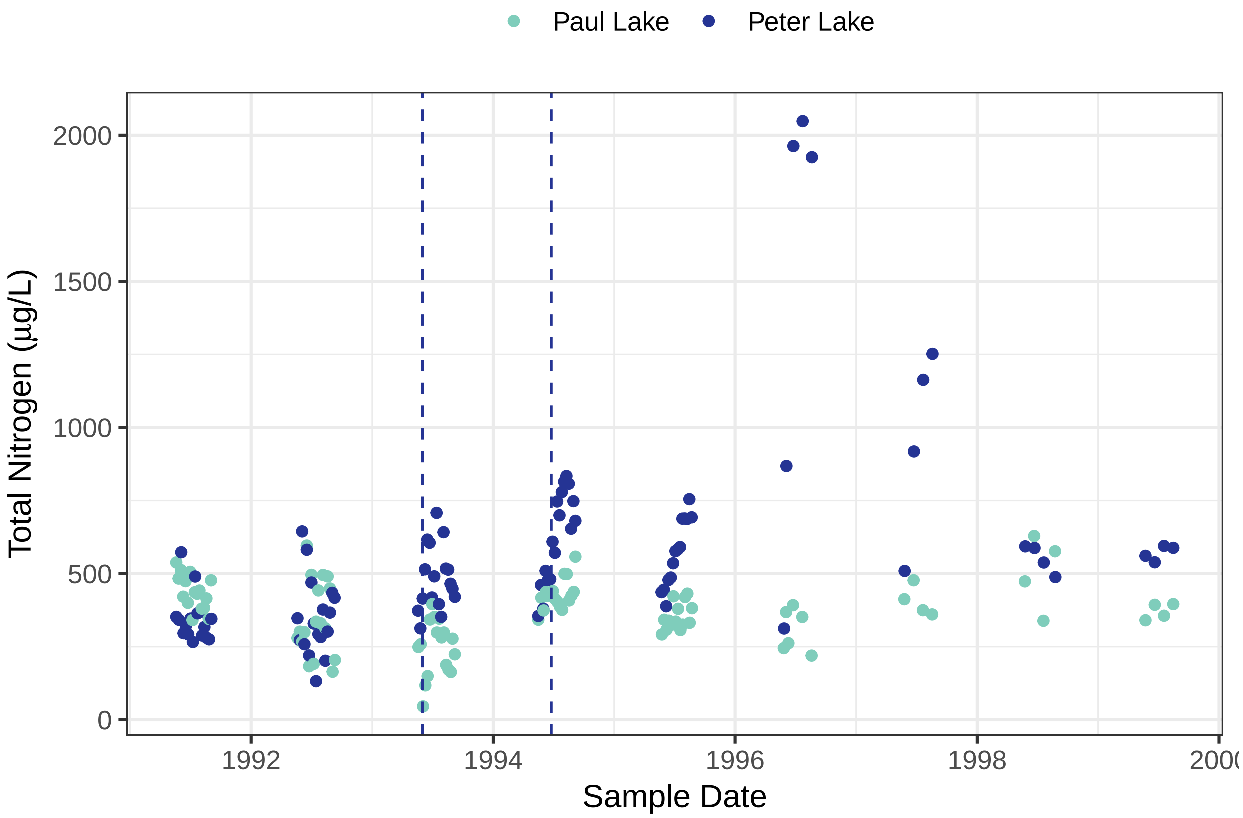
```
# no breaks for Paul Lake
```

What are the results of this test?

> ANSWER: There are two break points in the total nitrogen data for Paul Lake, one in the summer of 1993 and the other in the summer of 1994. Within each of the three time intervals, there is no statistically significant trend in the level of total nitrogen in Peter Lake. There is no significant trend for surface total nitrogen levels in Paul Lake.

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

```
ggplot(PP_Nutrients_Surface, aes(x = sampledate, y = tn_ug, color = lakename)) +
  geom_point() +
  scale_color_manual(values = c("#7fcdbb", "#253494")) +
  geom_vline(xintercept = as.Date("1993-06-01"),
             linetype = "dashed", color = "#253494") +
  geom_vline(xintercept=as.Date("1994-06-25"),
             linetype= "dashed", color = "#253494") +
  labs(x = "Sample Date",
       y = expression(paste("Total Nitrogen (", mu, "g/L)")),
       color = NULL)
```


Paul Lake
Peter Lake
Total Nitrogen (µg/L)
2000
1500
1000
500
0
1992
1994
1996
1998
200
Sample Date