Assignment 8: Time Series Analysis

Laurie Muzzy

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'm going to look at the Pb datasets from EPA Outdoor Air Quality from Detroit, MI, from 1987-2017, to determine what sites have decreased in lead exposure 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/laurie/Desktop/Envtl_Data_Analytics/MuzzyGitFile"
```

```
library(nlme)
```

```
## Warning: package 'nlme' was built under R version 3.4.4
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.4.4
```

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

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

```r
library(multcompView)
library(lsmeans)
```

```
## Warning: package 'lsmeans' was built under R version 3.4.4

## Loading required package: emmeans

## Warning: package 'emmeans' was built under R version 3.4.4

## 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'.
```

```r
library(trend)
```

```
## Warning: package 'trend' was built under R version 3.4.4
```

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.4.2

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

## v ggplot2 3.1.0       v purrr   0.3.0
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.2       v stringr 1.3.1
## v readr   1.3.1       v forcats 0.3.0

## Warning: package 'ggplot2' was built under R version 3.4.4

## Warning: package 'tibble' was built under R version 3.4.4

## Warning: package 'tidyr' was built under R version 3.4.4

## Warning: package 'readr' was built under R version 3.4.4

## Warning: package 'purrr' was built under R version 3.4.4

## Warning: package 'dplyr' was built under R version 3.4.4

## Warning: package 'stringr' was built under R version 3.4.4

## Warning: package 'forcats' was built under R version 3.4.3

## -- Conflicts ------------------------------------------------------------- tidyverse_conflicts() --
## x lubridate::as.difftime() masks base::as.difftime()
## x dplyr::collapse()        masks nlme::collapse()
## x lubridate::date()        masks base::date()
## x dplyr::filter()          masks stats::filter()
## x lubridate::intersect()   masks base::intersect()
## x dplyr::lag()             masks stats::lag()
## x lubridate::setdiff()     masks base::setdiff()
## x lubridate::union()       masks base::union()
```

```r
library(tidyr)
```

```r
EPAair_PM25_NC2018_raw <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv")
#View(EPAair_PM25_NC2018_raw)

EPAair_PM25_NC2018_raw$Date <- as.Date(EPAair_PM25_NC2018_raw$Date,
                                        format = "%m/%d/%y")
```

```
## Warning in strptime(x, format, tz = "GMT"): unknown timezone 'zone/tz/
## 2018i.1.0/zoneinfo/America/New_York'
```

```r
class(EPAair_PM25_NC2018_raw$Date) #Date
```

```
## [1] "Date"
```

```r
EPAair_PM25_NC2018_raw$AQS_PARAMETER_DESC <- "PM2.5"

PeterPaul.chem <- read.csv("./Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaul_Processed.csv")
#View(PeterPaul.chem)

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

```
## [1] "Date"
```

```r
LFM8theme <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "bottom")
theme_set(LFM8theme)
```

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.

```r
#3
#PM2.5 = response
#Date = fixed effect    `correlation = structure(form = ~ time | subjvar)'
#Site.Name = random effect
EPAair_PM25_NC2018_raw = EPAair_PM25_NC2018_raw[order(EPAair_PM25_NC2018_raw[,'Date'],-EPAair_PM25_NC20
EPAair_PM25_NC2018_raw = EPAair_PM25_NC2018_raw[!duplicated(EPAair_PM25_NC2018_raw$Date),]

PM2.5mixed <- lme(data = EPAair_PM25_NC2018_raw,
                  Daily.Mean.PM2.5.Concentration ~ Date,  # response ~ explan
                  random = ~1|Site.Name, #random
                  correlation = corAR1(value = 0.513, form = ~ Date|Site.Name),
                  method = "REML")

summary(PM2.5mixed) #AIC 1756.622; pval is kinda high, so it says DATE not a sig predictor
```

```
## Linear mixed-effects model fit by REML
##  Data: EPAair_PM25_NC2018_raw
```

```
##        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
```

```
PM2.5fixed <- gls(data = EPAair_PM25_NC2018_raw,
                  Daily.Mean.PM2.5.Concentration ~ Date,
                  method = "REML")
summary(PM2.5fixed)
```

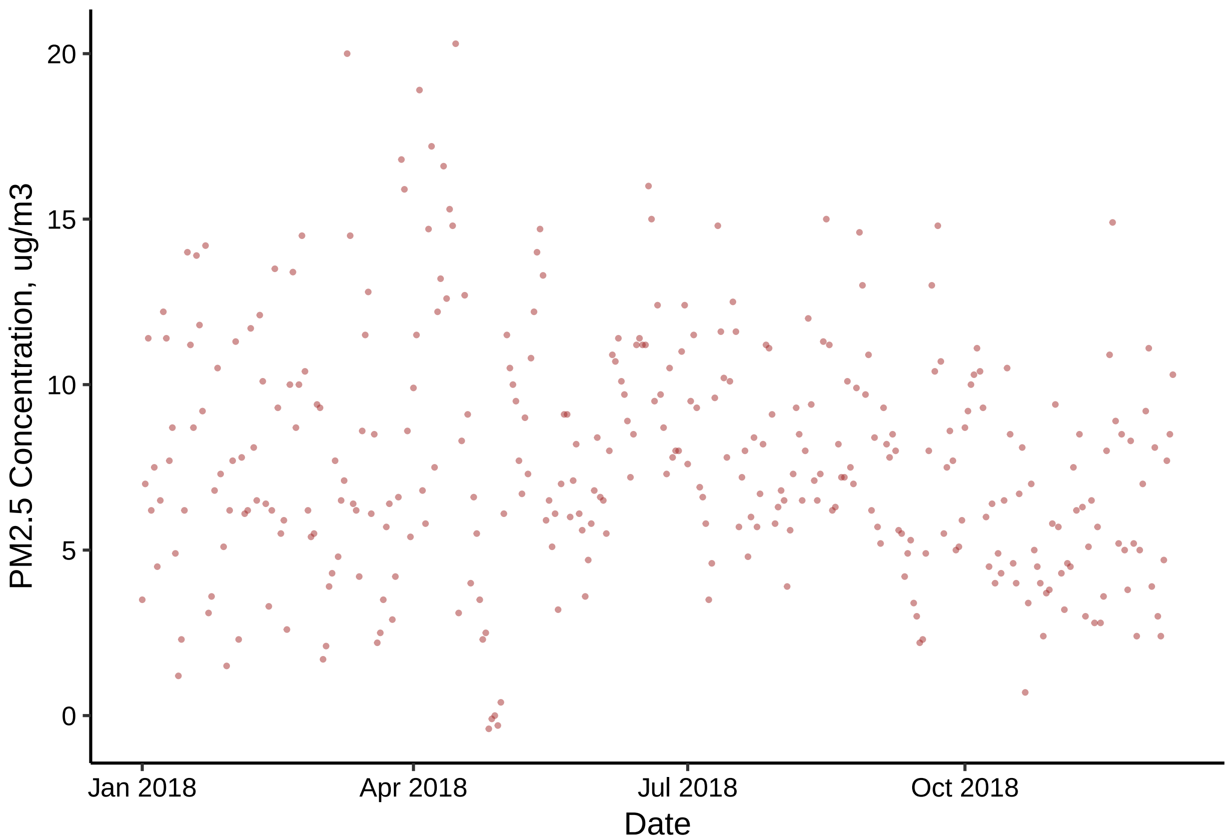
```
## Generalized least squares fit by REML
##   Model: Daily.Mean.PM2.5.Concentration ~ Date
##   Data: EPAair_PM25_NC2018_raw
##        AIC      BIC    logLik
##   1865.202 1876.698 -929.6011
##
## Coefficients:
##                Value Std.Error   t-value p-value
## (Intercept) 98.57796  34.60285  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
```

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

```
##            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## PM2.5mixed     1  5 1756.622 1775.781 -873.3110                        
## PM2.5fixed     2  3 1865.202 1876.698 -929.6011 1 vs 2 112.5802  <.0001
```

```r
#3a
PM2.5Site <- EPAair_PM25_NC2018_raw %>%
  select(Date, Daily.Mean.PM2.5.Concentration, Site.Name) %>%
  na.exclude()
#View(PM2.5Site)

PM2.5inNC <- ggplot(PM2.5Site, aes(x = Date, y = Daily.Mean.PM2.5.Concentration)) +
  geom_point(size = 0.5, alpha = 0.5, color = "brown") +
  labs(x = "Date", y = "PM2.5 Concentration, ug/m3")
print(PM2.5inNC)
```



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.

```r
#3c temporal autocorrelation
PM2.5corr <- lme(data = EPAair_PM25_NC2018_raw,
                 Daily.Mean.PM2.5.Concentration ~ Date,
                 random = ~1|Site.Name)
```

```r
PM2.5corr
```

```
## Linear mixed-effects model fit by REML
##   Data: EPAair_PM25_NC2018_raw
##   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(PM2.5corr) # ACF = 0.513
```

```
##    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
#3d mixed effects model
PM2.5mixed <- lme(data = EPAair_PM25_NC2018_raw,
                  Daily.Mean.PM2.5.Concentration ~ Date,  # response ~ explan
                  random = ~1|Site.Name, #random
                  correlation = corAR1(value = 0.513, form = ~ Date|Site.Name),
                  method = "REML")
PM2.5mixed
```

```
## Linear mixed-effects model fit by REML
```

```
##   Data: EPAair_PM25_NC2018_raw
##   Log-restricted-likelihood: -873.311
##   Fixed: Daily.Mean.PM2.5.Concentration ~ Date
##  (Intercept)         Date
## 83.148009025 -0.004261058
##
## 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
## Number of Observations: 343
## Number of Groups: 3
```

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

> ANSWER: There isn't a significant trend in PM2.5 concentrations over the course of the year, evidenced from the ACF value of 0.51 (about 50% of the concentrations are correlated to the values of the day before or after, which makes sense).

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.

```
PM2.5fixed <- gls(data = EPAair_PM25_NC2018_raw,
                  Daily.Mean.PM2.5.Concentration ~ Date)
summary(PM2.5fixed)
```

```
## Generalized least squares fit by REML
##   Model: Daily.Mean.PM2.5.Concentration ~ Date
##   Data: EPAair_PM25_NC2018_raw
##        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(PM2.5mixed, PM2.5fixed)
```

```
##            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
```

```
## PM2.5mixed     1  5 1756.622 1775.781 -873.3110
## PM2.5fixed     2  3 1865.202 1876.698 -929.6011 1 vs 2 112.5802  <.0001
```

```
#           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
#PM2.5mixed     1  5 1756.622 1775.781 -873.3110
#PM2.5fixed     2  3 1865.202 1876.698 -929.6011 1 vs 2 112.5802  <.0001
```

Which model is better?

> ANSWER: The AIC is lower in the mixed effects model, so MIXED is better.

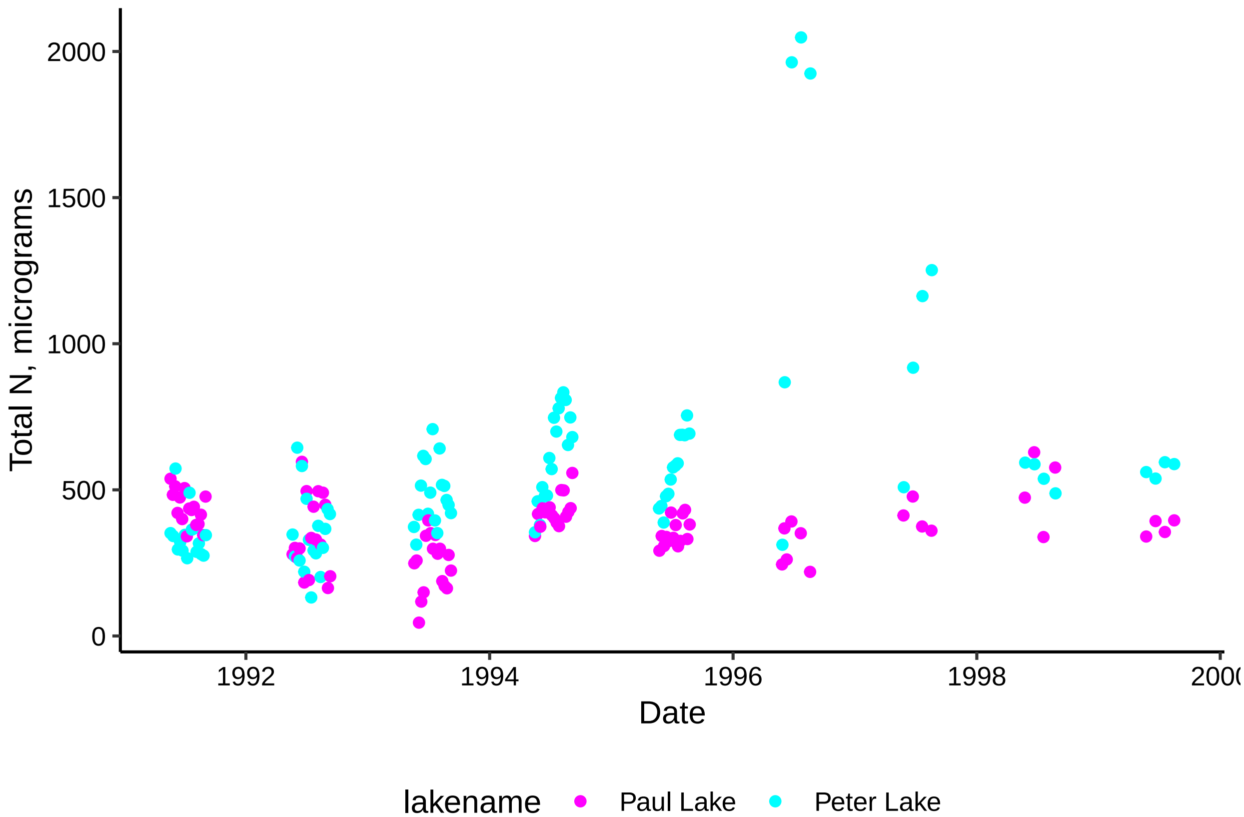
Run a Mann-Kendall test

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

ggplot(PeterPaul.N.surface, aes(x = sampledate,y = tn_ug, color = lakename)) +
  geom_point() +
  scale_color_manual(values = c("magenta", "cyan")) +
  labs(x = "Date", y = "Total N, micrograms")
```



```r
Peter.N.surface <- filter(PeterPaul.N.surface, lakename == "Peter Lake")
Paul.N.surface <- filter(PeterPaul.N.surface, lakename == "Paul Lake")

#Peter Lake
mk.test(Peter.N.surface$tn_ug) #pval v low, z = 7.29, a significant positive trend
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.N.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
pettitt.test(Peter.N.surface$tn_ug) #low pval, significant change point at 36, from 1993-05-26
```

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

```r
mk.test(Peter.N.surface$tp_ug[1:35]) #pval 0.589 , z = 0.53 so no trend
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.N.surface$tp_ug[1:35]
## z = 0.53998, n = 35, p-value = 0.5892
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 3.900000e+01 4.952333e+03 6.587922e-02
```

```r
mk.test(Peter.N.surface$tp_ug[36:98]) #pval 0.00531, z = -2.78 means a bit of a negative trend, but ins
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.N.surface$tp_ug[36:98]
## z = -2.7876, n = 63, p-value = 0.00531
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
##  -471.0000000 28427.0000000    -0.2411674
```

```r
pettitt.test(Peter.N.surface$tn_ug[36:98]) #36+21=57
```

```
##
##  Pettitt's test for single change-point detection
##
```

```
## data:  Peter.N.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
mk.test(Peter.N.surface$tp_ug[57:98]) #pval = 0.129, z = -1.51, insignificant negative trend from 1994-
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.N.surface$tp_ug[57:98]
## z = -1.5172, n = 42, p-value = 0.1292
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## -141.0000000 8514.3333333   -0.1637631
```

```r
#Paul Lake
mk.test(Paul.N.surface$tn_ug) #pval 0.72, z = -0.35, insignificant negative trend
```

```
##
##  Mann-Kendall trend test
##
## data:  Paul.N.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
pettitt.test(Paul.N.surface$tn_ug) #change point at 16, from 1991-08-26
```

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

```r
mk.test(Paul.N.surface$tn_ug[1:15]) #pval = 0.0075, z = -2.67, insignificant negative trend
```

```
##
##  Mann-Kendall trend test
##
## data:  Paul.N.surface$tn_ug[1:15]
## z = -2.6723, n = 15, p-value = 0.007533
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##           S        varS         tau
## -55.0000000 408.3333333  -0.5238095
```

```r
mk.test(Paul.N.surface$tn_ug[16:99]) #pval = 0.0274, z = 2.20, insignificant positive trend
```

```
##
##  Mann-Kendall trend test
##
## data:  Paul.N.surface$tn_ug[16:99]
## z = 2.2058, n = 84, p-value = 0.0274
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 5.720000e+02 6.700867e+04 1.640849e-01
```

```r
pettitt.test(Paul.N.surface$tn_ug[16:99]) #16+36=52, 5-17-1992
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Paul.N.surface$tn_ug[16:99]
## U* = 852, p-value = 0.001403
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              36
```

```r
mk.test(Paul.N.surface$tn_ug[52:99]) #pval = 0.197, z = -1.28, insignificant negative trend
```

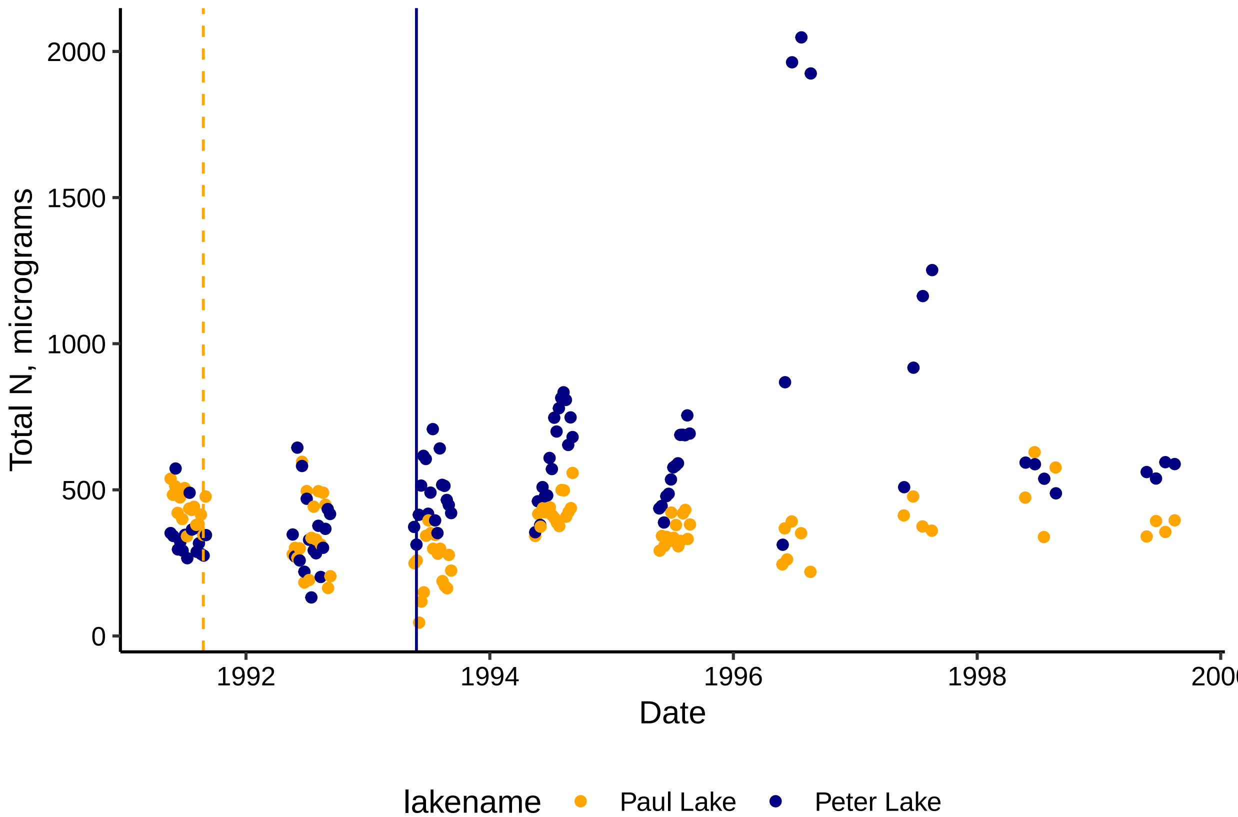
```
##
##  Mann-Kendall trend test
##
## data:  Paul.N.surface$tn_ug[52:99]
## z = -1.2888, n = 48, p-value = 0.1975
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
##  -146.0000000 12658.6666667    -0.1294326
```

What are the results of this test?

> ANSWER: for Peter Lake: z = 7.2927, p-value = 3.039e-13. Since the p-val is so low, we can reject the null, meaning that we see a trend. Since the z-score is not near zero, we can say that there is a positive trend over time, i.e., Total N is getting higher in Peter Lake. However, Paul Lake (pval 0.72, z = -0.35) is not like this: the p-val is high, the z-score is close to zero, so we can't be confident that there's any sort of trend 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).

```r
PeterPaul.N <- ggplot(PeterPaul.N.surface, aes(x = sampledate, y = tn_ug, color = lakename)) +
  geom_point() +
  geom_vline(xintercept = as.Date("1991-08-26"), color = "orange", lty = 2) +
  geom_vline(xintercept = as.Date("1993-05-26"), color = "navy", lty = 1) +
  scale_color_manual(values = c("orange", "navy")) +
  labs(x = "Date", y = "Total N, micrograms")
print(PeterPaul.N)
```



Total N, micrograms
2000
1500
1000
500
0
1992
1994
1996
1998
2000
Date
lakename
Paul Lake
Peter Lake