Assignment 8: Time Series Analysis

Sebastian Bognar

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

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.

```
#determine location of working directory and load necessary packages
getwd()
```

```
## [1] "/Users/Seabass/Documents/Duke/spring_2019/env_872L/lesson_2/ENV_872L/Assignments"
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------- tidyverse 1.2.
```

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

```r
library(trend)
library(ggplot2)
library(lubridate)
```

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

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

```r
library(nlme)
```

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

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

```r
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'.
```

```r
library(multcompView)
library(emmeans)

#upload EPA air quality raw dataset for pm2.5 and processed NTL_LTER dataset for peter and paul

RAW_EPA_PM2.5 <- read.csv("../Data/Raw/EPAair_PM25_NC2018_raw.csv")

Peterpaul_Nutrients_Processed<-read.csv("../Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaul_Processed.

#change name of PM2.5 colname

names(RAW_EPA_PM2.5)[5]<-"PM2.5"

names(RAW_EPA_PM2.5)[3]<-"Site.ID"

names(RAW_EPA_PM2.5)[8]<-"Site.Name"





# change date category from character to date format

#EPA
```

```
RAW_EPA_PM2.5$Date<- as.Date(RAW_EPA_PM2.5$Date, format = "%m/%d/%y")

#peter_paul

Peterpaul_Nutrients_Processed$sampledate<- as.Date(Peterpaul_Nutrients_Processed$sampledate, format = "%

# assignment theme

theme_A8 <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "right",
        axis.ticks = element_line(colour = "black"),
        panel.border = element_rect(fill= NA,color="black", size=0.5,
                                    linetype="solid"),
        panel.grid.major.y =element_line(color = "grey"),
        panel.grid.minor.y = element_blank(),
        panel.grid.major.x  = element_blank(),
        panel.grid.minor.x  = element_blank(),
        axis.text.x = element_text(angle = 40,  hjust = 1))
```

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.

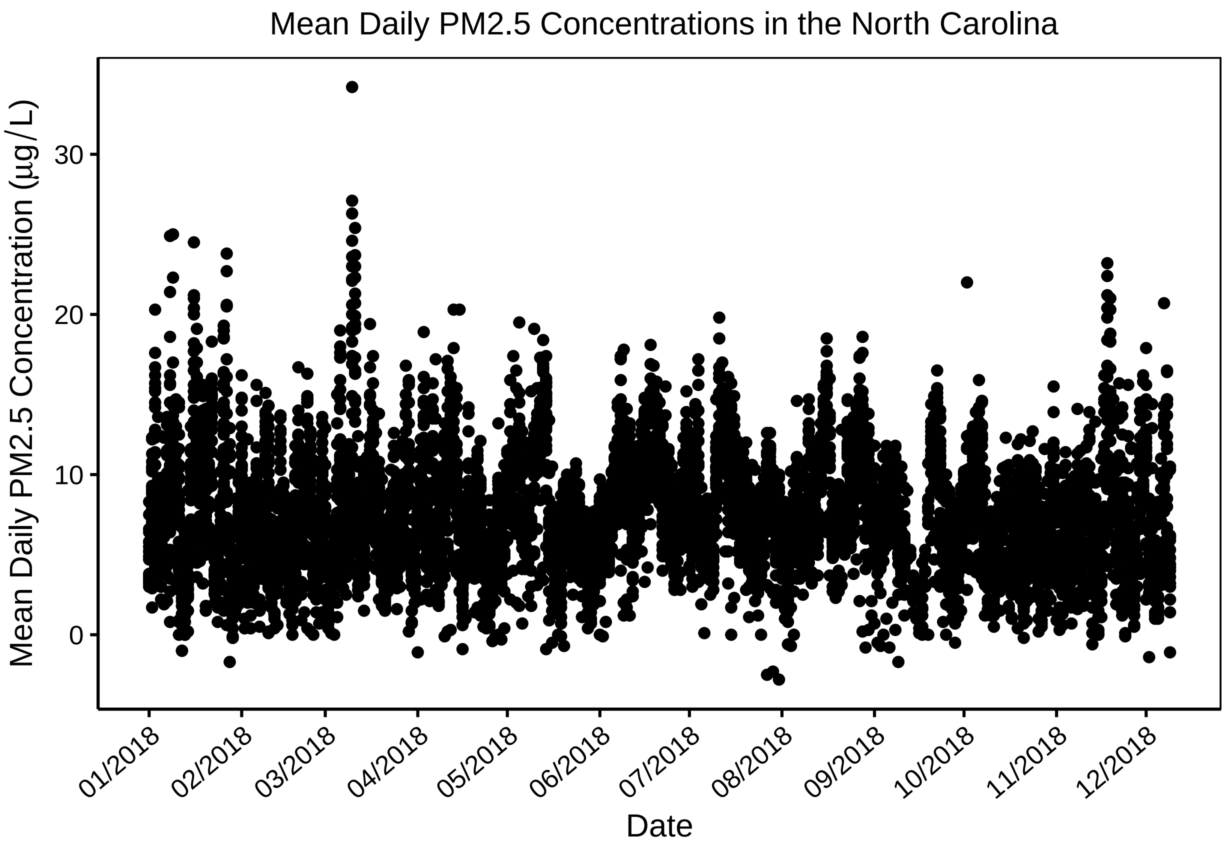
3a. Illustrate PM2.5 concentrations by date. Do not split aesthetics by site.

```
# 3A PM2.5 plot

PM2.5_by_date_plot <- ggplot(RAW_EPA_PM2.5, aes(x = Date, y =PM2.5))+
geom_point()+
theme_A8+
xlab("Date")+
ylab(expression("Mean Daily PM2.5 Concentration"~"("*mu*g/L*")"))+
scale_x_date(date_breaks = "1 month", date_labels = "%m/%Y")+
ggtitle(" Mean Daily PM2.5 Concentrations in the North Carolina")+
theme(plot.title = element_text(size = 12 ))+
theme(plot.title = element_text(hjust = 0.5))



PM2.5_by_date_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.

```
#  3B eliminate duplicate measurements for single dates for each site

RAW_EPA_PM2.5 = RAW_EPA_PM2.5[order(RAW_EPA_PM2.5[,'Date'],-RAW_EPA_PM2.5[,'Site.ID']),]
RAW_EPA_PM2.5 = RAW_EPA_PM2.5[!duplicated(RAW_EPA_PM2.5$Date),]

# 3C determine temporal autocorrelation in the model

AUTO_EPA <- lme(data = RAW_EPA_PM2.5,
                     PM2.5 ~ Date,
                     random = ~1|Site.Name)

AUTO_EPA
```

```
## Linear mixed-effects model fit by REML
##   Data: RAW_EPA_PM2.5
##   Log-restricted-likelihood: -928.6076
##   Fixed: PM2.5 ~ 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(AUTO_EPA)
```

```
##    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 run a mixed effects model

MIXED_EPA <- lme(data = RAW_EPA_PM2.5,
                     PM2.5 ~ Date,
                     random = ~1|Site.Name,
                     correlation = corAR1(form = ~ Date|Site.Name, value = 0.514),
                     method = "REML")

summary(MIXED_EPA)
```

```
## Linear mixed-effects model fit by REML
##  Data: RAW_EPA_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: PM2.5 ~ 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 was no significant decreasing trend in PM2.5 concentrations in 2018 (Mixed Effects Linear model; p = 0.214; DF = 339). The equation for PM2.5 concentration in 2018: [PM2.5] = 83.15 - 0.0043*(Date).

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.

```r
#fixed effects model

Fixed_EPA <- gls(data = RAW_EPA_PM2.5,
                      PM2.5 ~ Date,
                      method = "REML")

summary(Fixed_EPA)
```

```
## Generalized least squares fit by REML
##   Model: PM2.5 ~ Date
##   Data: RAW_EPA_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
```

```r
# use an anova to determine which model has more explanatory power

anova(Fixed_EPA,MIXED_EPA)
```

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

Which model is better?

> ANSWER: The fixed effects model of PM2.5 concentration accounts for more of the variability than the mixed effects model, which indicates that the fixed effects model is better model (ANOVA; $p<0.001$).There was significant decreasing trend in PM2.5 concentrations in 2018 (Fixed Effects Linear model; $p = 0.0091$; DF = 339). The equation for PM2.5 concentration in 2018: [PM2.5] = 98.58 - 0.0051*(Date).

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).

```r
#wrangle the dataset

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

#split the datasets by lakes

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

# Test for change points in Peter Lake (36,57)
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
#test if there is another change point

pettitt.test(Peter.nutrients.surface$tn_ug[37:98])
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Peter.nutrients.surface$tn_ug[37:98]
## U* = 522, p-value = 0.002339
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              20
```

```r
pettitt.test(Peter.nutrients.surface$tn_ug[58:98])
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Peter.nutrients.surface$tn_ug[58:98]
## U* = 120, p-value = 0.5882
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              10
```

```r
#test for change points in 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
```

```r
#mann kendall tests peter lake
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
mk.test(Peter.nutrients.surface$tn_ug[1:36])
```

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

```
##
## data:  Peter.nutrients.surface$tn_ug[1:36]
## z = 0.040863, n = 36, p-value = 0.9674
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 4.000000e+00 5.390000e+03 6.349206e-03
```

```r
mk.test(Peter.nutrients.surface$tn_ug[37:98])
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.nutrients.surface$tn_ug[37:98]
## z = 2.9642, n = 62, p-value = 0.003035
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 4.890000e+02 2.710433e+04 2.585933e-01
```

```r
mk.test(Peter.nutrients.surface$tn_ug[58:98])
```

```
##
##  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
```

```r
#mann-kendall test 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
```

What are the results of this test?

> ANSWER: For Peter Lake, there was a significant positive monotonic trend in total nitrogen concentration (Mann Kendall Test; $p < 0.001$; $S = 2377$). Additionally, there were two change points at 1993-06-02 and 1994-06-29 (Pettitt Test; $p < 0.001$). There were no significant monotonic trends in total nitrogen concentration for Paul lake (Mann-Kendall; $p = 0.73$).

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

```r
TN_PETER_PAUL_PLOT <- ggplot(PeterPaul.nutrients.surface, aes(x = sampledate, y= tn_ug, color = lakenam
geom_point()+
```

```r
theme_A8+
xlab("Date")+
ylab(expression("Total Nitrogen Concentration"~"("*mu*g/L*")"))+
ggtitle(" Total Nitrogen Concentration in Peter and Paul Lake")+
theme(plot.title = element_text(size = 12 ))+
theme(plot.title = element_text(hjust = 0.5))+
scale_color_manual(values = c('#fdae6b','#e6550d'))+
labs(color="Lake Name")+
geom_vline(xintercept=as.Date("1993-06-02"), linetype="dashed", col = 'black', show.legend =TRUE)+
geom_vline(xintercept=as.Date("1994-06-29"), linetype="dashed", col = 'black',show.legend = TRUE)


TN_PETER_PAUL_PLOT
```