Assignment 8: Time Series Analysis

Caroline Reents

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.

```
knitr::opts_chunk$set(error = TRUE)
getwd()
```

```
## [1] "/Users/carolinereents/Desktop/Data Analytics/EnvironmentalDataAnalytics/Assignments"
```

```
setwd("/Users/carolinereents/Desktop/Data Analytics/EnvironmentalDataAnalytics")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------------------- tidyve

## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
```

```
## v readr   1.1.1     v forcats 0.3.0

## -- Conflicts --------------------------------------------------------------------------- tidyverse_c
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
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

## Warning: package 'emmeans' was built under R version 3.5.2

## 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(trend)




EPA_PM25_2018 <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv")

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

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

EPA_PM25_2018$Date <- as.Date(EPA_PM25_2018$Date, format= "%m/%d/%y")

class(EPA_PM25_2018$Date)
```

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

```r
class(Nutrients_PeterPaul$sampledate)
```

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

```r
mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
```

```
        legend.position = "top")
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.

```
plotconc_date <- ggplot(EPA_PM25_2018, aes(x = Date, y = Daily.Mean.PM2.5.Concentration)) +
  geom_point()

PM2.5<-EPA_PM25_2018

PM2.5 = PM2.5[order(PM2.5[,'Date'],-PM2.5[,'Site.ID']),]

PM2.5 = PM2.5[!duplicated(PM2.5$Date),]
```

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.

```
#3c

ConcTest.auto <- lme(data = PM2.5,
                     Daily.Mean.PM2.5.Concentration ~ Date,
                     random = ~1|Site.ID)

ConcTest.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.ID
##         (Intercept) Residual
## StdDev:    1.650184 3.559209
##
## Number of Observations: 343
## Number of Groups: 3
```

```
ACF(ConcTest.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
#0.514

#3d
ConcTest.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(ConcTest.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: No significant trend

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

summary(ConcTest.fixed)
```

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

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

```r
# mixed has lower AIC
```

```r
ConcTest.posthoc = lsmeans(ConcTest.mixed, ~ Date)
cld(ConcTest.posthoc, alpha = 0.05, Letters = letters, adjust = "tukey")
```

```
##   Date lsmean    SE df lower.CL upper.CL .group
##  17703   7.71 0.341  2     6.25     9.18  a
##
## d.f. method: containment
## Confidence level used: 0.95
## significance level used: alpha = 0.05
```

Which model is better?

> ANSWER: Mixed effects

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
Nutrients_PeterPaul_surface <-
  Nutrients_PeterPaul %>%
  select(-lakeid, -depth_id, -comments) %>%
  filter(depth == 0) %>%
  filter(!is.na(tn_ug))

Nutrients_Peter_surface <- filter(Nutrients_PeterPaul_surface, lakename == "Peter Lake")
Nutrients_Paul_surface <- filter(Nutrients_PeterPaul_surface, lakename == "Paul Lake")

mk.test(Nutrients_Peter_surface$tn_ug)
```

```
##
##  Mann-Kendall trend test
##
## data:  Nutrients_Peter_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
#trend over time, positive
mk.test(Nutrients_Paul_surface$tn_ug)
```

```
##
##  Mann-Kendall trend test
##
## data:  Nutrients_Paul_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
#no trend over time





pettitt.test(Nutrients_Peter_surface$tn_ug)
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Nutrients_Peter_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(Nutrients_Peter_surface$tn_ug[1:36]) #no signifcant trend
```

```
##
##  Mann-Kendall trend test
##
## data:  Nutrients_Peter_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(Nutrients_Peter_surface$tn_ug[37:63]) #no significant trend
```

```
##
##  Mann-Kendall trend test
##
## data:  Nutrients_Peter_surface$tn_ug[37:63]
## z = 1.9179, n = 27, p-value = 0.05512
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
##   93.0000000 2301.0000000    0.2649573
```

What are the results of this test?

> ANSWER: A chaange point occurs at the 36th data point (1992-05-27). Before and after this chaange point there are no significant trends in the data

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

```r
ggplot(Nutrients_PeterPaul_surface, aes(x = sampledate, y = tn_ug, color = lakename)) +
  geom_point() +
  scale_color_manual(values = c("#7fcdbb", "#253494"))+
  geom_vline(xintercept = as.Date("1992-05-27"), color="#253494", lty=2)
```

```
## Error in seq.int(0, to0 - from, by): 'to' must be a finite number
```