13: Time Series Analysis

Environmental Data Analytics | Kateri Salk

Spring 2019

LESSON OBJECTIVES

1. Describe the aspects of hierarchical models, fixed effects, and random effects
2. Choose and justify appropriate statistical models when time is an explanatory variable
3. Apply repeated measures ANOVA to datasets with temporal components

SET UP YOUR DATA ANALYSIS SESSION

```r
getwd()
```

```
## [1] "C:/Users/Felipe/OneDrive - Duke University/1. DUKE/1. Ramos 2 Semestre/EOS-872 Env. Data Analyti
```

```r
library(tidyverse)
#install.packages("lubridate")
library(lubridate)
#install.packages("nlme")
library(nlme)
#install.packages("lsmeans")
library(lsmeans)
#install.packages("multcompView")
library(multcompView)

PeterPaul.chem <- read.csv("./Data/Processed/NTL-LTER_Lake_ChemistryPhysics_PeterPaul_Processed.csv")

# Set date to date format
PeterPaul.chem$sampledate <- as.Date(PeterPaul.chem$sampledate,
                                             format = "%m/%d/%y")

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

HIERARCHICAL MODELS

Hierarchical models, or **mixed-effects models,** are a type of linear model in which explanatory variables are given a model whose parameters are also estimated by the data. The coefficients associated with explanatory variables thus may not be a single value but instead be sampled from a distribution, called the hyper-distribution, which is defined by the modeler. The advantage of the hierarchical model is that it builds capacity to describe multiple layers of stochasticity, which enables accounting of all aspects of uncertainty in a system. Specifically, we can separately model the process of interest and the sampling process.

The coefficients of a hierarchical model are divided into two categories: **fixed effects** and **random effects.** A **fixed effect** is a factor whose levels are experimentally determined or whose interest lies in the effects of each level (e.g., covariates, treatments, interactions). A **random effect** is a factor whose levels are sampled from a larger population, or whose interest lies in the variation among them rather than the specific effect of each level. In choosing whether you are dealing with a fixed or a random effect, consider the following questions:

- Do you have a particular interest in the studied factor level? FRA: all fixe effect
- Have you included all possible levels in the study? FRA: all, fixed effect
- Do you have interest in the variance among levels? Parece que random effect.
- Do you have interest in generalizing to factor levels that you did not study? Random effect

One common variable in hierarchical models is **time.** Time can be a complicated explanatory variable, as it can act as either a fixed or a random effect depending on the study design and research question. Due to **temporal autocorrelation,** conditions measured at a single site will be highly influenced by the conditions preceding the sampling date. Therefore, two samples taken in relatively close temporal proximity may not necessarily be independent of one another. Treating time as a random effect will account for temporal trends in observations (e.g., diel or seasonal patterns) that may not be of interest for your study. FRA: this violates the indepedent assump.

Another common variable in hierarchical models is **space.** In many situations, we may want to infer conditions beyond the sites that we have sampled. By treating space as a random variable, we may be able to extrapolate conditions of the response variable across a spatial gradient. FRA: we have some points but want to generalize –> random effects.

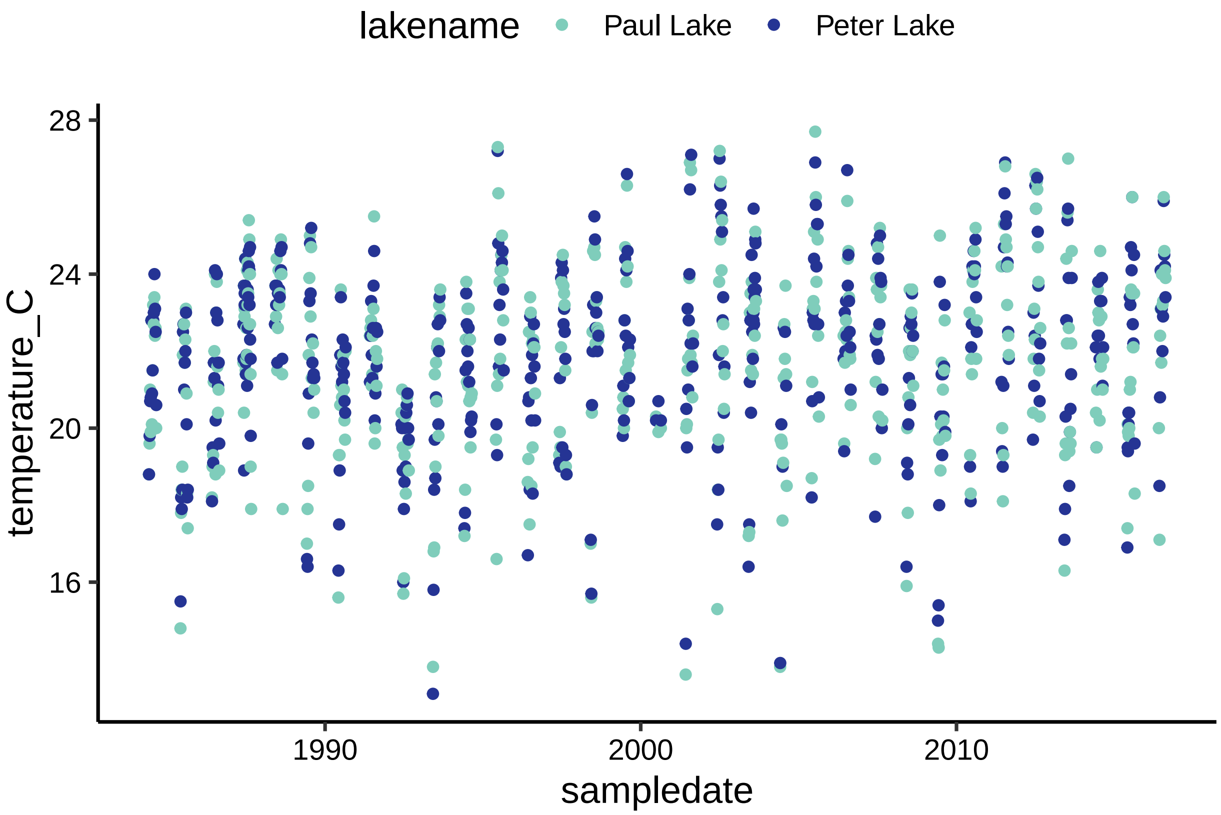
REPEATED MEASUREMENTS AND AUTOCORRELATION

In many situations where monitoring is conducted, samples taken repeatedly at a given site may not be considered truly independent. The conditions present on a given day may be dependent on conditions present earlier in time. This is clearly an issue for the way we might traditionally think of experimental design and statistical independence, but this type of study design is often of interest in the field of environmental science. We can set up models to consider autocorrelation of time within a given place. One example of this type of model is a **repeated measures ANOVA**.

Let's think about the situation of temperature monitoring in the NTL-LTER lake sites, Peter and Paul Lakes. We might be interested to know whether surface temperatures in the summer have increased over time in response to climate change. However, we know that (a) temperature conditions on a given date are dependent on conditions earlier in the season, and (b) there is considerable variability across the summer season within a year (i.e., cooler temperatures occurring in June vs. August). We can set up a hierarchical model to deal with the autocorrelation by date as well as the variability associated with seasonality.

Let's wrangle our data and visualize a preliminary relationship between our variables of interest.

```r
PeterPaul.summertemp <-
  PeterPaul.chem %>%
  select(lakename:temperature_C) %>%
  #filter for Julian days in June-August and surface measurements
  filter(daynum > 151 & daynum < 243 & depth == 0 ) %>% # just a period of time and just surface.
  #add a "week" column to represent seasonality
  mutate(Week = week(sampledate)) %>%
  #code won't work if there are NAs
  na.exclude() # careful that you dont erase useful data.

ggplot(PeterPaul.summertemp, aes(x = sampledate, y = temperature_C, color = lakename)) +
  geom_point() +
  scale_color_manual(values = c("#7fcdbb", "#253494"))
```



Next, we will determine the degree of temporal autocorrelation in our dataset. We will use the package `nlme` for our analyses. Another good package for running hierarchical, or mixed-effects, models is `lme4`. For the basic types of hierarchical models, these packages have about the same functionality.

```
# Determine autocorrelation in residuals
TempTest.auto <- lme(data = PeterPaul.summertemp,
                     temperature_C ~ sampledate * lakename,
                     random = ~1|Week)
TempTest.auto
```

```
## Linear mixed-effects model fit by REML
##   Data: PeterPaul.summertemp
##   Log-restricted-likelihood: -1841.595
##   Fixed: temperature_C ~ sampledate * lakename
##                (Intercept)                       sampledate
##               2.061909e+01                     8.816467e-05
##         lakenamePeter Lake sampledate:lakenamePeter Lake
##               1.093329e-01                     2.204779e-06
##
## Random effects:
##  Formula: ~1 | Week
##         (Intercept) Residual
## StdDev:    1.588171 1.941958
##
## Number of Observations: 863
## Number of Groups: 14
```

```
# what the variability is among the weeks. That is why it gives a StdDev. (the random effect).
# Fixed effects gives coef.

summary(TempTest.auto) # esto lo puse yo
```

```
## Linear mixed-effects model fit by REML
##  Data: PeterPaul.summertemp
##        AIC      BIC    logLik
##   3695.189 3723.724 -1841.595
##
## Random effects:
##  Formula: ~1 | Week
##         (Intercept) Residual
## StdDev:    1.588171 1.941958
##
## Fixed effects: temperature_C ~ sampledate * lakename
##                                 Value Std.Error  DF  t-value p-value
## (Intercept)                 20.619095 0.5218706 846 39.50998  0.0000
## sampledate                   0.000088 0.0000262 846  3.36343  0.0008
## lakenamePeter Lake           0.109333 0.4298095 846  0.25438  0.7993
## sampledate:lakenamePeter Lake  0.000002 0.0000372 846  0.05924  0.9528
##  Correlation:
##                             (Intr) smpldt lknmPL
## sampledate                  -0.553
## lakenamePeter Lake          -0.410  0.671
## sampledate:lakenamePeter Lake  0.390 -0.704 -0.952
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.23374297 -0.61443099  0.05643875  0.61436794  2.96775134
##
## Number of Observations: 863
## Number of Groups: 14
```

```
#
ACF(TempTest.auto)
```

```
##    lag         ACF
## 1    0  1.00000000
## 2    1  0.42324818
## 3    2  0.04801177
## 4    3 -0.04227307
## 5    4 -0.08775368
## 6    5 -0.07702138
## 7    6 -0.13094473
## 8    7 -0.10997063
## 9    8 -0.02905440
## 10   9 -0.02826303
## 11  10 -0.01730103
## 12  11 -0.01476535
## 13  12 -0.04397958
## 14  13 -0.04780381
## 15  14 -0.05377938
## 16  15  0.01680037
## 17  16  0.02257738
## 18  17 -0.03976543
## 19  18 -0.09594045
```

```
# position 2 --> 42.3%
```

This model structure should look familiar, with a typical linear model structure and dataframe defined. The addition here is that we have defined Week as a random variable. Essentially, we are interested not in the specific effects of each week but in the variability among weeks, so we have defined it as a random effect (essentially coming from a larger distribution of seasonal variability). The ~1 statement indicates that each week has its own intercept in the model. From here, we want to take the first order correlation to specify our autocorrelation structure. From the ACF output, we take the 2nd value (the innermost group level) to define the degree of autocorrelation. This number will always fall between 0 and 1. Notice that there is a fairly large degree of autocorrelation in our variables.

We will now create a repeated measures ANOVA model now that we have defined our autocorrelation structure. The way we have set up this model, we are considering temporal autocorrelation within the levels of Week, and we have retained Week as a random effect.

The correlation statement in the model is defined as follows: `correlation = structure(form = ~ time | subjvar)`, where structure is the autocorrelative structure (options in `?corClasses`), time is the temporal variable, and subjvar is the variable for experimental units.

```
TempTest.mixed <- lme(data = PeterPaul.summertemp,
                     temperature_C ~ sampledate * lakename,
                     random = ~1|Week,
                     #specify autocorrelation structure of order 1 (this is the 42.3%)
                     #sampledate is duplicated in some cases, so need to split up by lake.
                     #FRA: This is only beacuse R doesnt like  that the sample were taken the same time
                     correlation = corAR1(form = ~ sampledate/lakename|Week, value = 0.423),
                     # we put manually the 0.423
                     #define method as restricted maximum likelihood
                     method = "REML")
summary(TempTest.mixed)
```

```
## Linear mixed-effects model fit by REML
##  Data: PeterPaul.summertemp
##        AIC      BIC    logLik
##   3678.638 3711.928 -1832.319
##
## Random effects:
##  Formula: ~1 | Week
##         (Intercept) Residual
## StdDev:    1.618912 1.930645
##
## Correlation Structure: ARMA(1,0)
##  Formula: ~sampledate/lakename | Week
##  Parameter estimate(s):
##      Phi1
## 0.5910308
## Fixed effects: temperature_C ~ sampledate * lakename
##                                  Value Std.Error  DF  t-value p-value
## (Intercept)                  20.466650 0.5323160 846 38.44831  0.0000
## sampledate                    0.000099 0.0000266 846  3.73751  0.0002
## lakenamePeter Lake            0.072415 0.4409492 846  0.16423  0.8696
## sampledate:lakenamePeter Lake  0.000005 0.0000379 846  0.12217  0.9028
##  Correlation:
##                              (Intr) smpldt lknmPL
## sampledate                   -0.554
## lakenamePeter Lake           -0.409  0.669
## sampledate:lakenamePeter Lake  0.389 -0.701 -0.953
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.23677606 -0.59679953  0.07085816  0.63629846  3.02528688
##
## Number of Observations: 863
## Number of Groups: 14
```

```
#Results: Random we got a stddev. Fixed effects: lakenamePeter Lake coef --> no diff between lakes.
# The interaction is also not significant.

# Compare the random effects model with the fixed effects model
TempTest.fixed <- gls(data = PeterPaul.summertemp,
                      temperature_C ~ sampledate * lakename,
                      method = "REML")
summary(TempTest.fixed)
```

```
## Generalized least squares fit by REML
##   Model: temperature_C ~ sampledate * lakename
##   Data: PeterPaul.summertemp
##        AIC      BIC    logLik
##   4019.694 4043.473 -2004.847
##
## Coefficients:
##                                    Value Std.Error  t-value p-value
## (Intercept)                    20.800214 0.3770183 55.17031  0.0000
## sampledate                      0.000085 0.0000326  2.59341  0.0097
## lakenamePeter Lake              0.085356 0.5344819  0.15970  0.8732
## sampledate:lakenamePeter Lake  0.000006 0.0000463  0.12712  0.8989
##
##  Correlation:
##                              (Intr) smpldt lknmPL
## sampledate                   -0.951
## lakenamePeter Lake           -0.705  0.671
## sampledate:lakenamePeter Lake  0.670 -0.704 -0.951
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -3.5435857 -0.6593598  0.1113940  0.6964839  2.4025400
##
## Residual standard error: 2.415319
## Degrees of freedom: 863 total; 859 residual
```

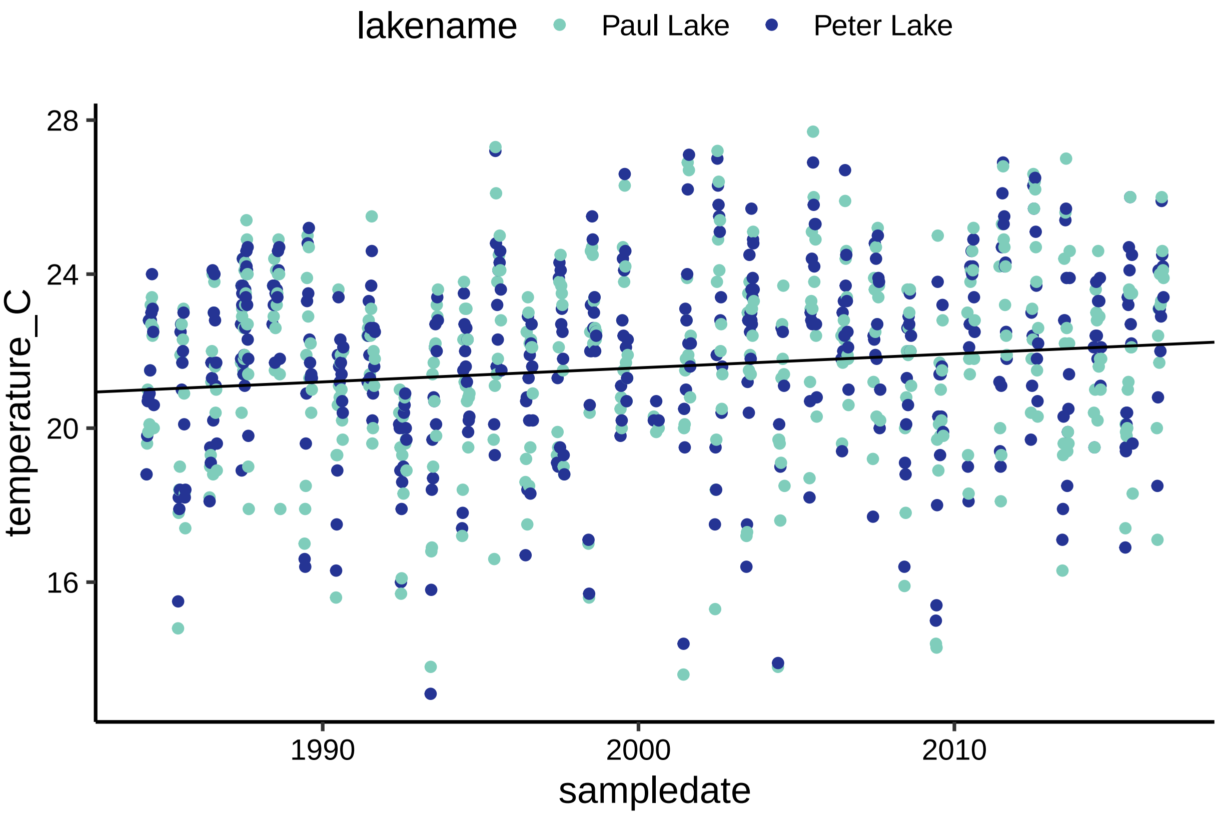
```
# little diff results but same conclusions
# why did we wanto to use random effect? --> We now there is variability. why dont use
# that for our advantage.

anova(TempTest.mixed, TempTest.fixed)
```

```
##                Model df      AIC      BIC    logLik   Test  L.Ratio
## TempTest.mixed     1  7 3678.638 3711.928 -1832.319
## TempTest.fixed     2  5 4019.694 4043.473 -2004.847 1 vs 2 345.0563
##                 p-value
## TempTest.mixed
## TempTest.fixed  <.0001
```

```
# The lower the AIC, the better.
# The p-value tells us whether those models have a significantly different fit

# FRA: the models and signif diff and the mixed is better (AIC)

# Post-hoc test
# This will yield groupings of temperature by lake for the average date value
TempTest.posthoc = lsmeans(TempTest.mixed, ~ sampledate * lakename)
cld(TempTest.posthoc, alpha = 0.05, Letters = letters, adjust = "tukey")
```

```
##  sampledate lakename   lsmean    SE df lower.CL upper.CL .group
##       10989 Paul Lake    21.6 0.443 13     20.4     22.7  a
##       10989 Peter Lake   21.7 0.443 13     20.6     22.8  a
##
## d.f. method: containment
## Confidence level used: 0.95
## Conf-level adjustment: sidak method for 2 estimates
## significance level used: alpha = 0.05
```

```
# FRA: you want this is there is a signif diff between groups. Something about groups.


# display our final relationship
ggplot(PeterPaul.summertemp, aes(x = sampledate, y = temperature_C, color = lakename)) +
  geom_point() +
  scale_color_manual(values = c("#7fcdbb", "#253494")) +
  geom_abline(intercept = 20.47, slope = 0.0001) #rounded coeff.
```



Question: How would you interpret the collective results of your mixed effects model in the context of the study question?

ANSWER: