7: Water Quality in Lakes

Water Data Analytics | Kateri Salk

Spring 2022

Lesson Objectives

1. Analyze stressor-response relationships between chlorophyll and nutrients in lakes
2. Explore how classification can account for geographic differences in lake water quality
3. Create correlation matrices to explore relationships among variables

Session Set Up

```
getwd()
```

```
## [1] "/Users/katerisalk/Box Sync/Courses/Water Data Analytics/Lessons"
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

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

```
library(lubridate)
```

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

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(LAGOSNE)
#install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
theme_set(theme_classic())
options(scipen = 4)
```

Load the LAGOSNE database

```r
# Load LAGOSNE data into R session
LAGOSdata <- lagosne_load()
```

```
## Warning in (function (version = NULL, fpath = NA) : LAGOSNE version unspecified,
## loading version: 1.087.3
```

```r
names(LAGOSdata)
```

```
##  [1] "county"              "county.chag"          "county.conn"
##  [4] "county.lulc"         "edu"                  "edu.chag"
##  [7] "edu.conn"            "edu.lulc"             "hu4"
## [10] "hu4.chag"            "hu4.conn"             "hu4.lulc"
## [13] "hu8"                 "hu8.chag"             "hu8.conn"
## [16] "hu8.lulc"            "hu12"                 "hu12.chag"
## [19] "hu12.conn"           "hu12.lulc"            "iws"
## [22] "iws.conn"            "iws.lulc"             "state"
## [25] "state.chag"          "state.conn"           "state.lulc"
## [28] "buffer100m"          "buffer100m.lulc"      "buffer500m"
## [31] "buffer500m.conn"     "buffer500m.lulc"      "lakes.geo"
## [34] "epi_nutr"            "lakes_limno"          "lagos_source_program"
## [37] "locus"
```

```r
# If the package installation and data download has not worked, use this code:
# load(file = "./Data/Raw/LAGOSdata.rda")


# Additional help resources for the LAGOS database:
help.search("datasets", package = "LAGOSNE")

# Exploring the data types that are available
LAGOSstate <- LAGOSdata$state
LAGOSnutrient <- LAGOSdata$epi_nutr
LAGOSgeo <- LAGOSdata$lakes.geo
LAGOSlocus <- LAGOSdata$locus
LAGOSlulc <- LAGOSdata$buffer500m.lulc
LAGOSlimno <- LAGOSdata$lakes_limno

# Tell R to treat lakeid as a factor, not a numeric value
LAGOSnutrient$lagoslakeid <- as.factor(LAGOSnutrient$lagoslakeid)
LAGOSgeo$lagoslakeid <- as.factor(LAGOSgeo$lagoslakeid)
LAGOSlocus$lagoslakeid <- as.factor(LAGOSlocus$lagoslakeid)
LAGOSlulc$lagoslakeid <- as.factor(LAGOSlulc$lagoslakeid)
LAGOSlimno$lagoslakeid <- as.factor(LAGOSlimno$lagoslakeid)
```

Stressor-Response Relationships: Chlorophyll and Nutrients

EPA's guidance on stressor-response modeling specifies that regression approaches can be used to connect ecosystem stressors (such as TN and TP) to ecosystem responses that have an impact on designated uses including recreation and aquatic life (such as chlorophyll). This approach can be extended using classification (to partition datasets into similarly behaving groupings), addition of covariates, and inclusion of prediction intervals.

EPA also recently released a suite of stressor-response models built from the National Lakes Assessment that allow a user to build site-specific analyses in an interactive web app. More info here: https://www.epa.gov/nutrient-policy-data/ambient-water-quality-criteria-address-nutrient-pollution-lakes-and-reservoirs. See the chlorophyll vs. TN and TP web app here.

```r
# Join data frames to add state name onto the geo dataset
LAGOSgeo <- left_join(LAGOSgeo, LAGOSstate)
```

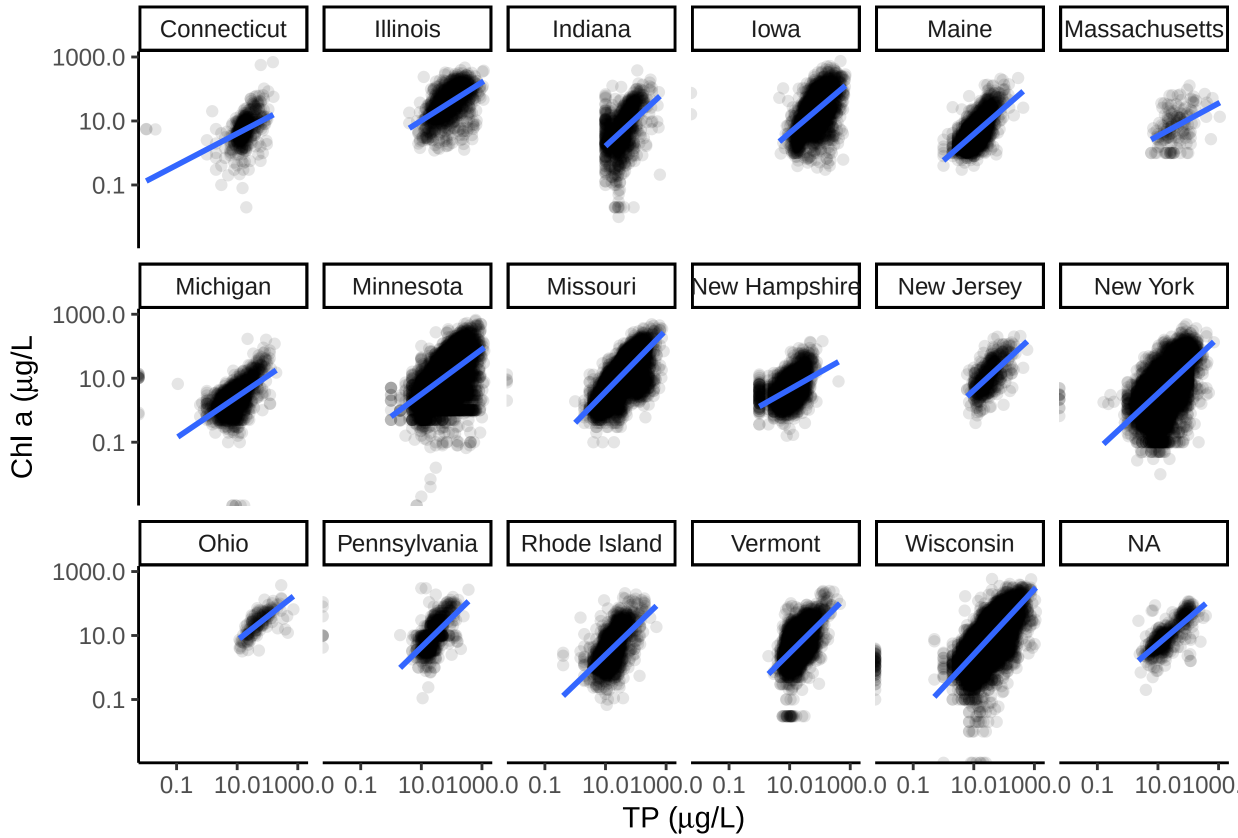
```
## Joining, by = "state_zoneid"
```

```r
LAGOSjoined <- LAGOSnutrient %>%
  left_join(., LAGOSgeo) %>%
  left_join(., LAGOSlocus) %>%
  left_join(., LAGOSlulc) %>%
  left_join(., LAGOSlimno)
```

```
## Joining, by = "lagoslakeid"

## Joining, by = c("lagoslakeid", "iws_zoneid", "hu4_zoneid", "hu6_zoneid", "hu8_zoneid", "hu12_zoneid"

## Joining, by = "lagoslakeid"

## Joining, by = c("lagoslakeid", "nhdid", "nhd_lat", "nhd_long")
```

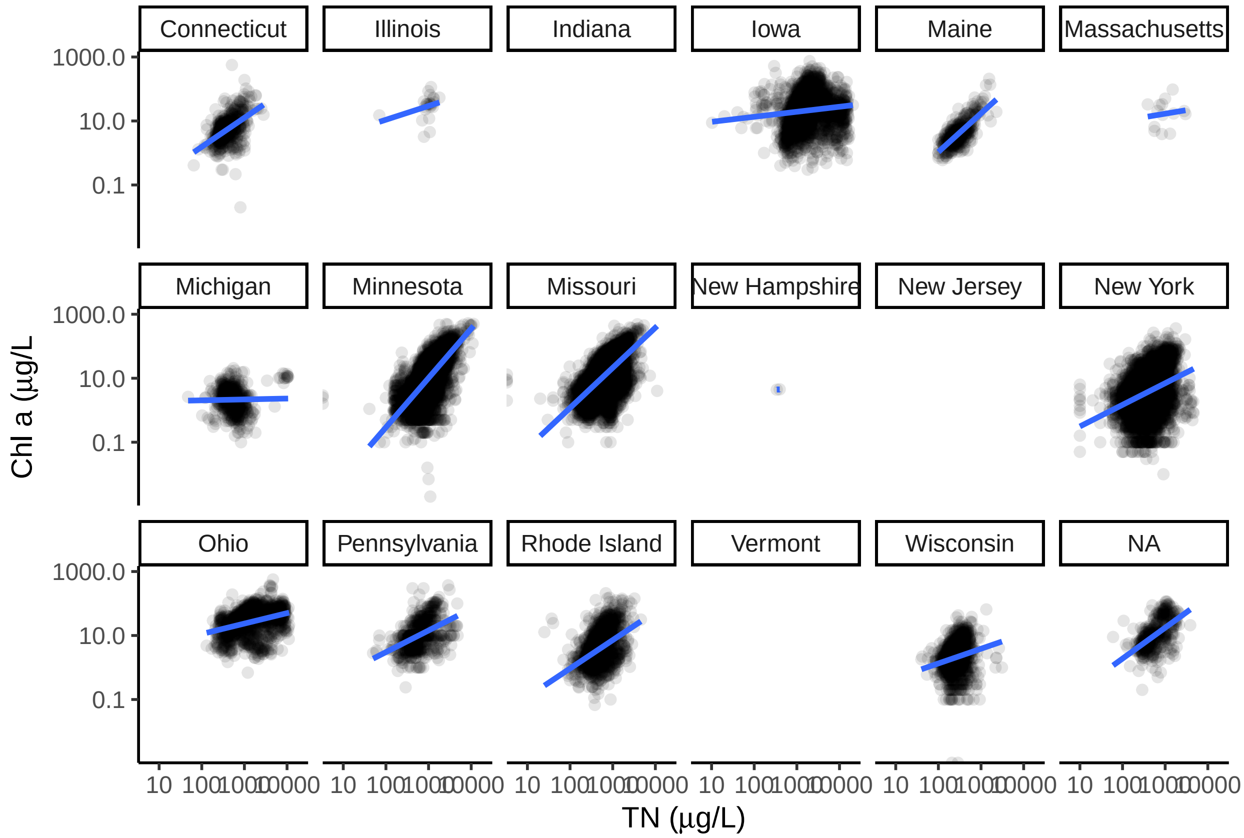
```r
ggplot(LAGOSjoined, aes(x = tp, y = chla)) +
  geom_point(alpha = 0.1) +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(vars(state_name), ncol = 6) +
  scale_y_log10() +
  scale_x_log10() +
  labs(x = expression("TP ("*mu*"g/L)"), y = expression("Chl a ("*mu*"g/L"))
```

```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous x-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous x-axis

## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 728654 rows containing non-finite values (stat_smooth).

## Warning: Removed 728541 rows containing missing values (geom_point).
```




```r
ggplot(LAGOSjoined, aes(x = tn, y = chla)) +
  geom_point(alpha = 0.1) +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(vars(state_name), ncol = 6) +
  scale_y_log10() +
  scale_x_log10() +
  labs(x = expression("TN ("*mu*"g/L)"), y = expression("Chl a ("*mu*"g/L"))
```

```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous x-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous x-axis

## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 779812 rows containing non-finite values (stat_smooth).

## Warning: Removed 779802 rows containing missing values (geom_point).
```



Run a stressor-response relationship for Minnesota lakes

Minnesota has numeric criteria for chlorophyll and total phosphorus in four of its seven level III ecoregions. Lakes have been classified into these separate classes by ecoregion because they have different levels of nutrients and primary production as well as different chlorophyll-nutrient responses.

Classification is used to split lakes into groups with similar states and/or similar stressor-response relationships. Machine learning methods such as CART or MBRP enable analysts to classify lakes from a suite of potential predictor variables, such as ecoregions, land use, hydrologic regime, etc. Soranno et al. 2010 outline these approaches for classification and how they fit into the management and conservation landscape.

LAGOS-NE doesn't have ecoregions readily available, so we will use Hydrologic Unit Codes (HUCs) instead today. Minnesota's HUC 4 boundaries overlap with the level III ecoregions, so these will make a nice proxy for today.

Exercise:

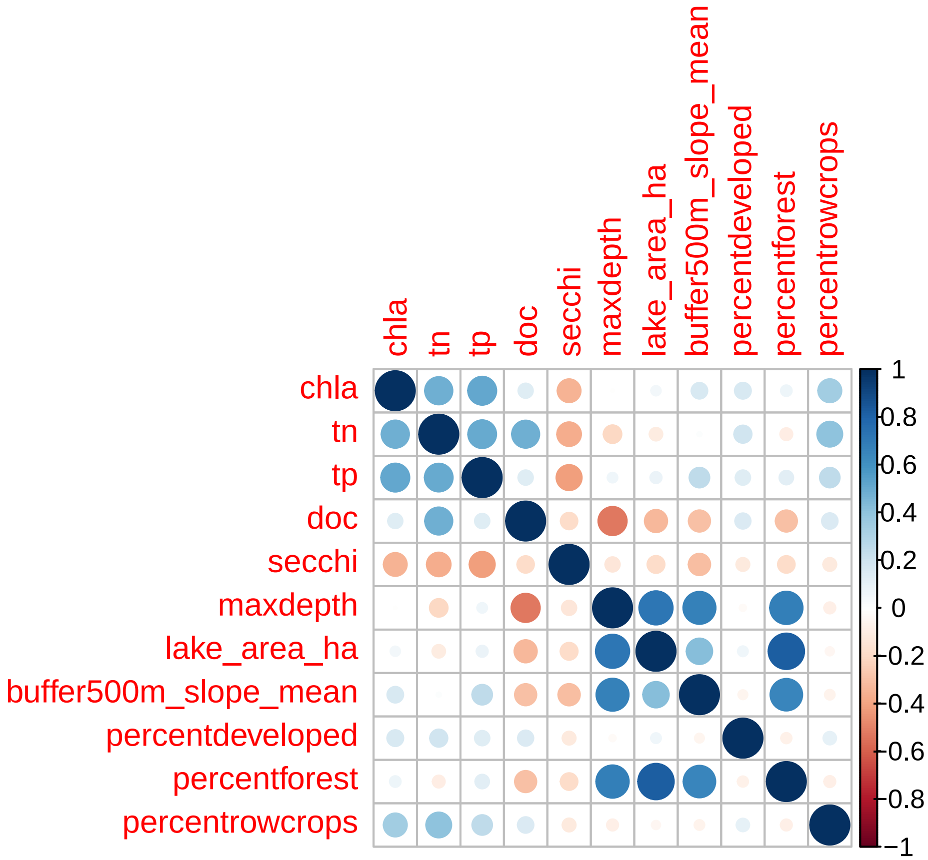
1. Create a new dataframe called LAGOSMN from the LAGOSjoined dataset.
2. Join the "hu4" dataset from LAGOS (hint: use the $ symbol).
3. Select the columns lagoslakeid, sampledate, state_name, hu4, chla, tn, and tp.
4. Add in columns for year and month.
5. Filter your dataset so that only lakes in Minnesota sampled from June through September from 1990 onward are included.
6. Include only the HUC 4 codes 0401, 0701, 0702, 0703, 0704, and 0902.
7. Plot the chlorophyll-TP stressor-response relationship, divided by HUC 4. You could choose to do this as a faceted plot or use color as an aesthetic.
8. Create a linear model to evaluate the chlorophyll-TP relationship, with HUC 4 code as an interaction effect. Is it important to account for the HUC 4 code in the model? How might this classification

impact potential nutrient management?

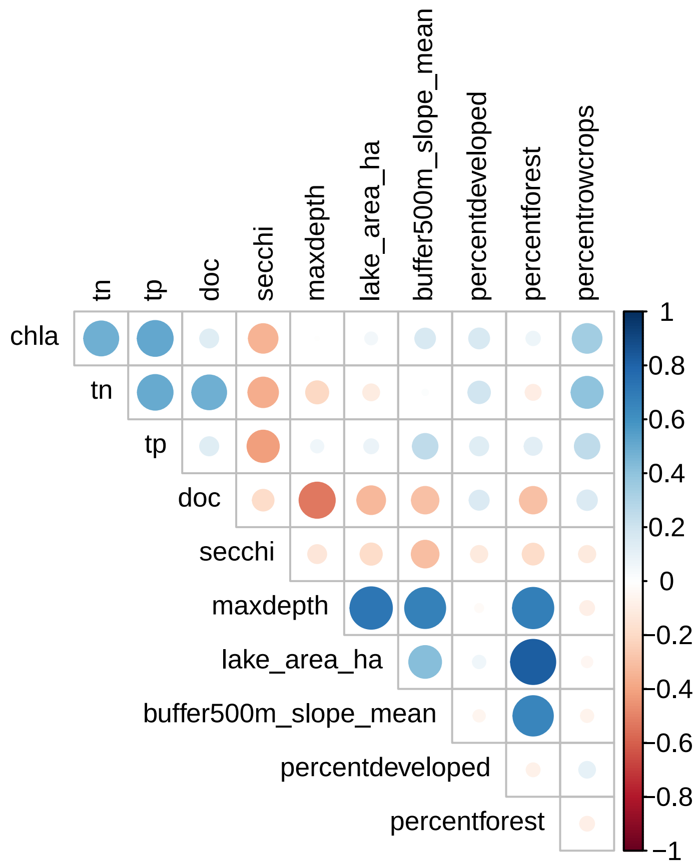
Covariates

If we wanted to explore some of the other variables that might help predict lake water quality, the function `corrplot` will allow us to create a correlation matrix whereby we can assess the correlations among various variables. We could then use this exploration to inform classification and/or covariate analysis with stressor-response modeling.

```
LAGOScovariates <- LAGOSjoined %>%
  mutate(percentdeveloped = buffer500m_nlcd2011_pct_22 + buffer500m_nlcd2011_pct_23 + buffer500m_nlcd20
         percentforest = buffer500m_nlcd2011_pct_41 + buffer500m_nlcd2011_ha_42 + buffer500m_nlcd2011_ha
         percentrowcrops = buffer500m_nlcd2011_pct_82) %>%
  select(chla, tn, tp, doc, secchi,
         maxdepth,lake_area_ha, buffer500m_slope_mean,
         percentdeveloped, percentforest, percentrowcrops)

# create a correlation matrix. Necessitates only complete cases.
LAGOScorr <- cor(LAGOScovariates, use = "complete.obs")

# plot correlation matrix.
corrplot(LAGOScorr)
```




```
# customize.
corrplot(LAGOScorr, type = "upper", diag = FALSE, tl.cex = 0.8, tl.col = "black")
```



tn
tp
doc
secchi
maxdepth
lake_area_ha
buffer500m_slope_mean
percentdeveloped
percentforest
percentrowcrops
chla
tn
tp
doc
secchi
maxdepth
lake_area_ha
buffer500m_slope_mean
percentdeveloped
percentforest
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1