9: Water Quality in Lakes

Hydrologic Data Analysis | Kateri Salk

Fall 2019

Lesson Objectives

1. Navigate and explore the LAGOSNE database and R package
2. Evaluate lake water quality using the trophic state index
3. Analyze spatial and temporal patterns of water quality across the northeast U.S.

Opening Discussion

What are the major water quality impairments experienced in lakes?

Session Set Up

```
getwd()
```

```
## [1] "/Users/ks501/Box Sync/Courses/Hydrologic Data Analysis/Lessons"
```

```
library(tidyverse)
```

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

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

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

```
library(lubridate)
```

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

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

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

theme_set(theme_classic())
options(scipen = 100)

#lagosne_get(dest_folder = LAGOSNE:::lagos_path(), overwrite = TRUE)
```

Getting to know the LAGOSNE database

Navigate to https://lagoslakes.org/. We will explore this website to learn about the LAGOS-NE dataset, research, and data management and use initiatives undergone by the resaerch team.

Useful introductions to the LAGOSNE R Package can be found here:

https://github.com/cont-limno/LAGOSNE https://cont-limno.github.io/LAGOSNE/articles/lagosne_structure.html

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

```
## Warning in `_f`(version = version, fpath = fpath): LAGOSNE version
## unspecified, loading version: 1.087.3
```

```
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"
```

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

# Exploring the data types that are available
LAGOSlocus <- LAGOSdata$locus
LAGOSstate <- LAGOSdata$state
LAGOSnutrient <- LAGOSdata$epi_nutr

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

Wrangling data frames in LAGOSNE

LAGOSNE is stored in several pieces, comprising metadata about given lakes (one observation per lake), metadata about each state (one observation per state), and data collected from lakes (one to many observations per lake over time). To connect observations from one data frame to the next, we need to find a common variable between the data frames.

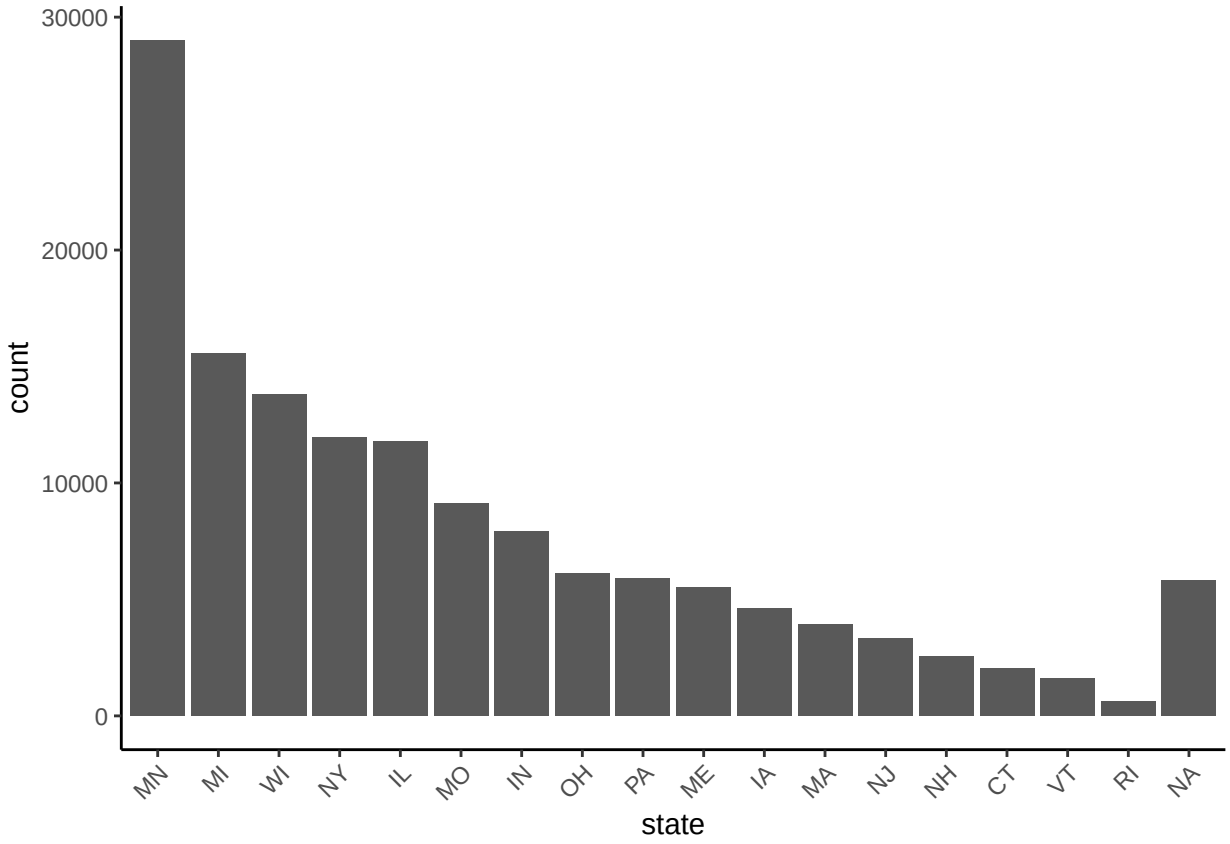
For example, let's find out how many lakes are in each state. Note that LAGOSlocus only includes the state_zoneid, whereas LAGOSstate connects state_zoneid for each state.

Add notes about each line of code as we go along. What does each function do?

```r
# Join data frames
LAGOSlocations <- left_join(LAGOSlocus, LAGOSstate, by = "state_zoneid")

# Order by number of lakes
LAGOSlocations <-
  within(LAGOSlocations,
         state <- factor(state, levels = names(sort(table(state), decreasing=TRUE))))

LakeCounts <- ggplot(LAGOSlocations, aes(x = state)) +
  geom_bar(stat = "count") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1))
print(LakeCounts)
```



Trophic state as a metric for water quality

Robert Carlson's 1977 paper, "A trophic state index for lakes," established the first numeric categorization system for a lake's **trophic state**, the amount of biomass a given system can sustain. Trophic state is a useful water quality metric, as it can give insight into the propensity of a system to develop algal blooms, the degree of nutrient loading in the system, and a range of other potential water quality concerns experienced by other lakes with a similar trophic state (e.g., hypoxia).

To calculate the **Trophic State Index**, three variables can be used. Note these should not be used to define trophic state but as indicators of the broader condition. Comparing these values in a given lake can give insight into the broader mechanisms at play.

- *chlorophyll a concentration*, a proxy for algal (phytoplankton) biomass. Pros: direct measure of primary productivity
- *Secchi disk transparency*, a measure of water clarity. Pros: simple and cheap. Cons: may yield a high TSI in highly colored lakes and in lakes where particulate matter is comprised of non-algal material
- *Total phosphorus (TP)*, a nutrient essential for growth of primary producers. Assumptions: phosphorus is the limiting nutrient for phytoplankton growth (this assumption often holds only for summer months)

$TSI(Chl) = 10(6 - (2.04 - 0.68lnChl/ln2))$

$TSI(SD) = 10(6 - (lnSD/ln2))$

$TSI(TP) = 10(6 - (ln(48/TP)/ln2))$

TSI values correspond to the following trophic states: **0-40**: Oligotrophic **40-50**: Mesotrophic **50-70**: Eutrophic **70-100**: Hypereutrophic

Exploring the LAGOS nutrient data frame

```
dim(LAGOSnutrient)
```

```
## [1] 816095     92
```

```
class(LAGOSnutrient$sampledate)
```

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

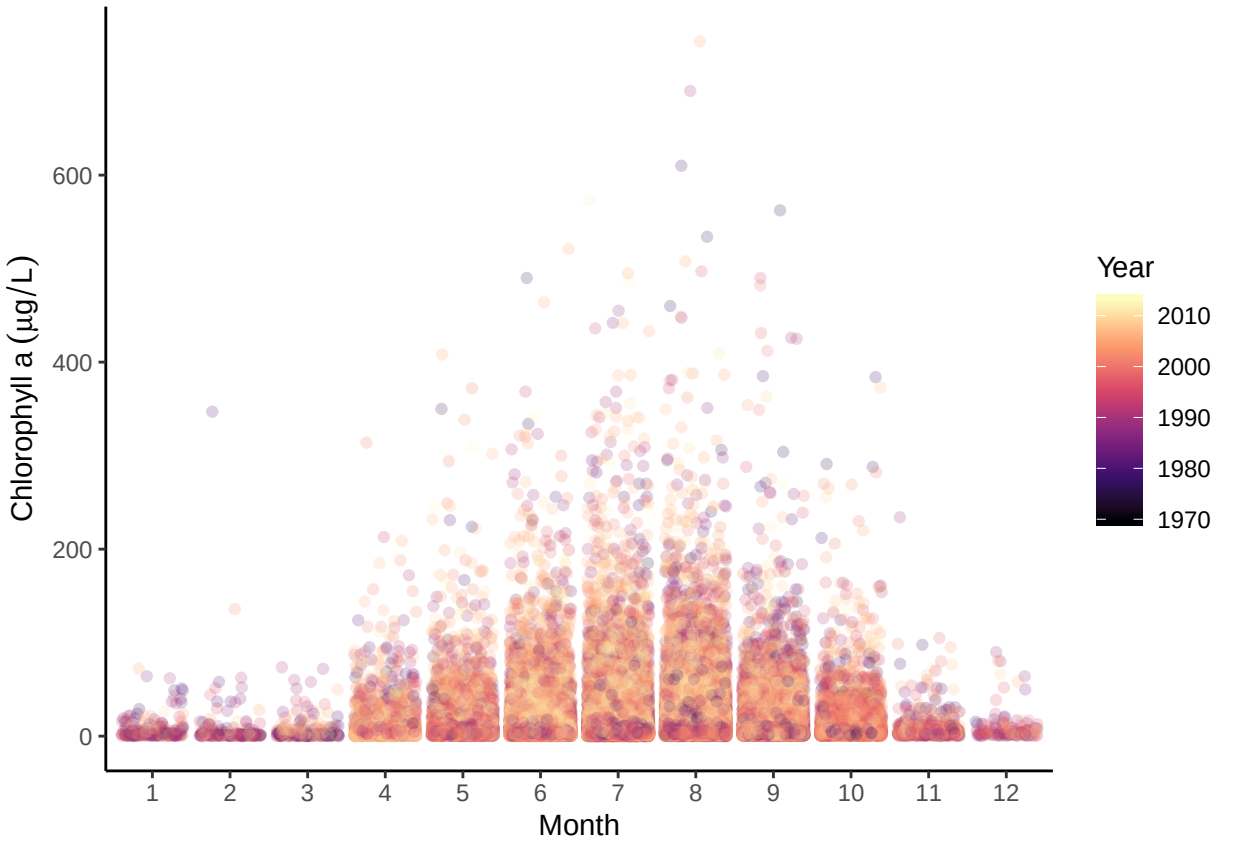
```
LAGOStrophic <-
  left_join(LAGOSnutrient, LAGOSlocations, by = "lagoslakeid") %>%
  select(lagoslakeid, sampledate, chla, tp, secchi,
         gnis_name, lake_area_ha, state, state_name) %>%
  mutate(sampleyear = year(sampledate),
         samplemonth = month(sampledate),
         season = as.factor(quarter(sampledate, fiscal_start = 12))) %>%
  drop_na(chla:secchi)
```

```
## Warning: Column `lagoslakeid` joining factors with different levels,
## coercing to character vector
```

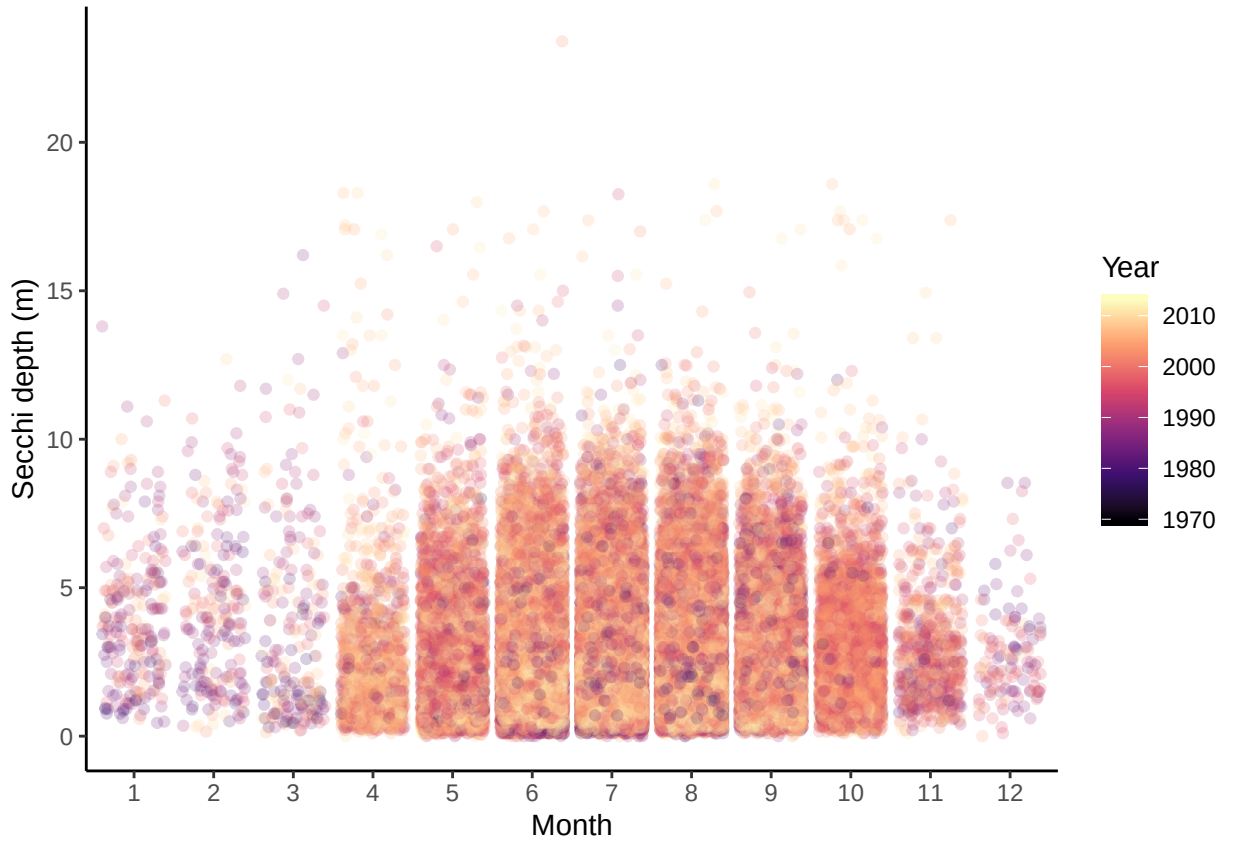
```
levels(LAGOStrophic$season) <- c("Winter", "Spring", "Summer", "Fall")
```

Let's look at observations of chl, secchi depth, and TP seasonally and over the period of study.

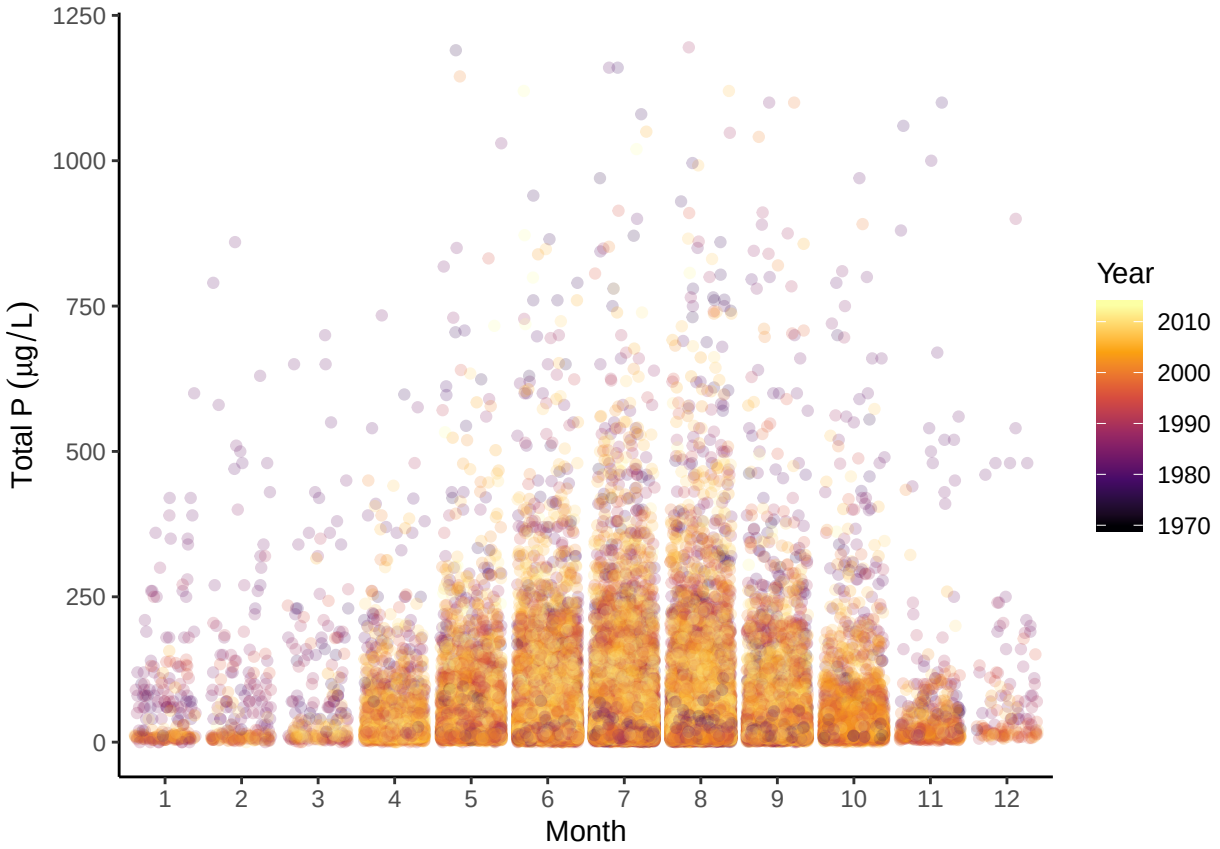
```
chlbymonth <-
ggplot(LAGOStrophic,
       aes(x = as.factor(samplemonth), y = chla, color = sampleyear)) +
  geom_jitter(alpha = 0.2) +
  labs(x = "Month", y = expression(Chlorophyll ~ a ~ (mu*g / L)), color = "Year") +
  scale_color_viridis_c(option = "magma")
print(chlbymonth)
```




```
secchibymonth <-
ggplot(LAGOStrophic,
       aes(x = as.factor(samplemonth), y = secchi, color = sampleyear)) +
  geom_jitter(alpha = 0.2) +
  labs(x = "Month", y = "Secchi depth (m)", color = "Year") +
  scale_color_viridis_c(option = "magma")
print(secchibymonth)
```




```
tpbymonth <-
ggplot(LAGOStrophic,
       aes(x = as.factor(samplemonth), y = tp, color = sampleyear)) +
  geom_jitter(alpha = 0.2) +
  labs(x = "Month", y = expression(Total ~ P ~ (mu*g / L)), color = "Year") +
  scale_color_viridis_c(option = "inferno")
print(tpbymonth)
```



What do you notice about the seasonality of these variables? If we were to characterize a lake based on the value of a given variable, how might seasonality affect our interpretations?

Calculating trophic state index

Let's add a TSI value calculated from each of the three variables to the data frame. Let's also add a column that designates the lake as oligotrophic, mesotrophic, eutrophic, or hypereutrophic based on the TSI.chl value. Make notes about the code as we go along.

```
LAGOStrophic <-
  mutate(LAGOStrophic,
         TSI.chl = round(10*(6 - (2.04 - 0.68*log(chla)/log(2)))),
         TSI.secchi = round(10*(6 - (log(secchi)/log(2)))),
         TSI.tp = round(10*(6 - (log(48/tp)/log(2)))),
         trophic.class =
            ifelse(TSI.chl < 40, "Oligotrophic",
                   ifelse(TSI.chl < 50, "Mesotrophic",
                          ifelse(TSI.chl < 70, "Eutrophic", "Hypereutrophic"))))

LAGOStrophic$trophic.class <-
  factor(LAGOStrophic$trophic.class,
         levels = c("Oligotrophic", "Mesotrophic", "Eutrophic", "Hypereutrophic"))

# LAGOStrophic$season <-
```

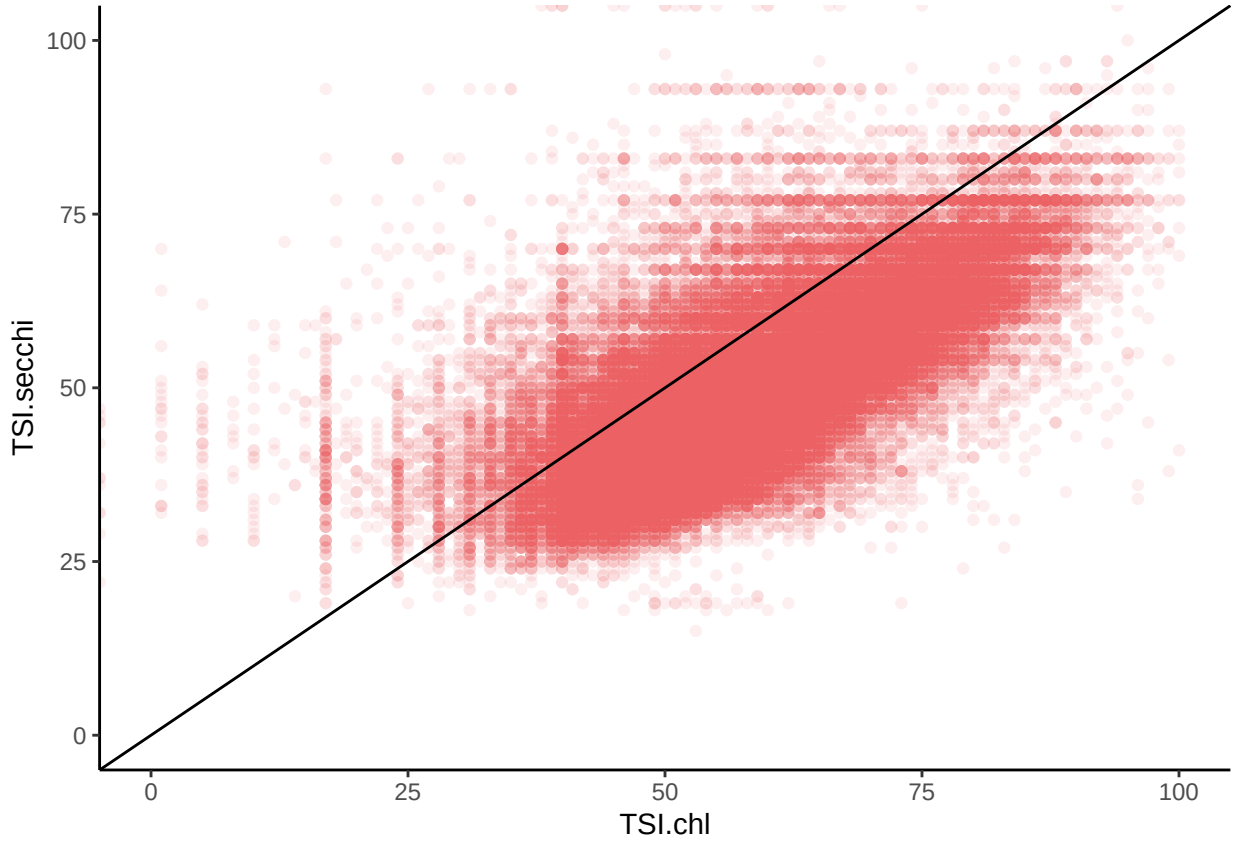
```
#   factor(LAGOStrophic$season,
#          levels = c("Spring", "Summer", "Fall", "Winter"))

#scales::show_col(colormap(colormap = colormaps$magma, nshades=16))
```

Now let's compare TSI values calculated from chl, secchi depth, and TP. If these were all perfectly equivalent metrics, all points should line up on the 1:1 line.

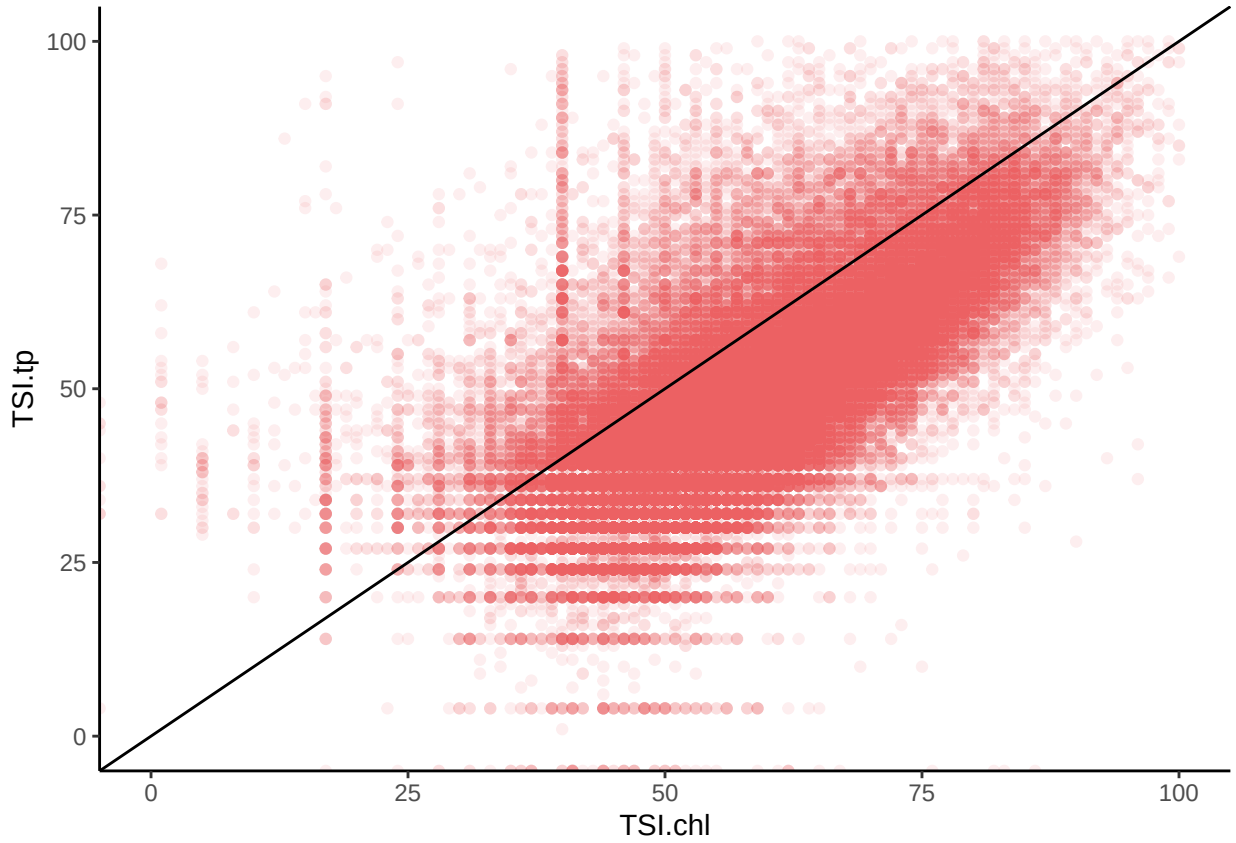
```
chlvssecchi <- ggplot(LAGOStrophic, aes(x = TSI.chl, y = TSI.secchi)) +
  geom_point(alpha = 0.1, color = "#ec6163ff") +
  scale_y_continuous(limits = c(0, 100)) +
  scale_x_continuous(limits = c(0, 100)) +
  geom_abline(slope = 1, intercept = 0)
print(chlvssecchi)
```

```
## Warning: Removed 36 rows containing missing values (geom_point).
```




```
chlvstp <- ggplot(LAGOStrophic, aes(x = TSI.chl, y = TSI.tp)) +
  geom_point(alpha = 0.1, color = "#ec6163ff") +
  scale_y_continuous(limits = c(0, 100)) +
  scale_x_continuous(limits = c(0, 100)) +
  geom_abline(slope = 1, intercept = 0)
print(chlvstp)
```

```
## Warning: Removed 88 rows containing missing values (geom_point).
```



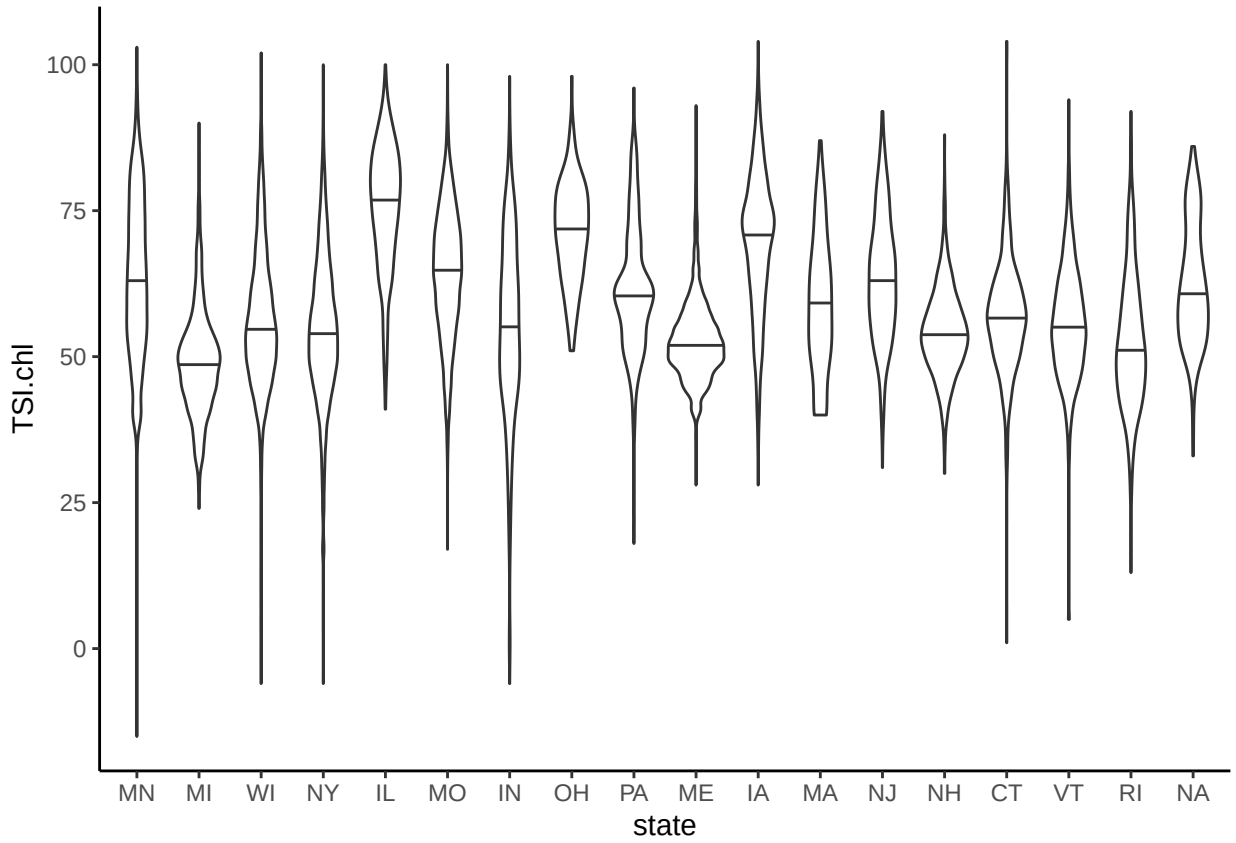
What is the observed relationship, and how does it depart from the 1:1 relationship? What do values above or below the 1:1 line tell us about the conditions present in a lake?

Comparing TSI across states

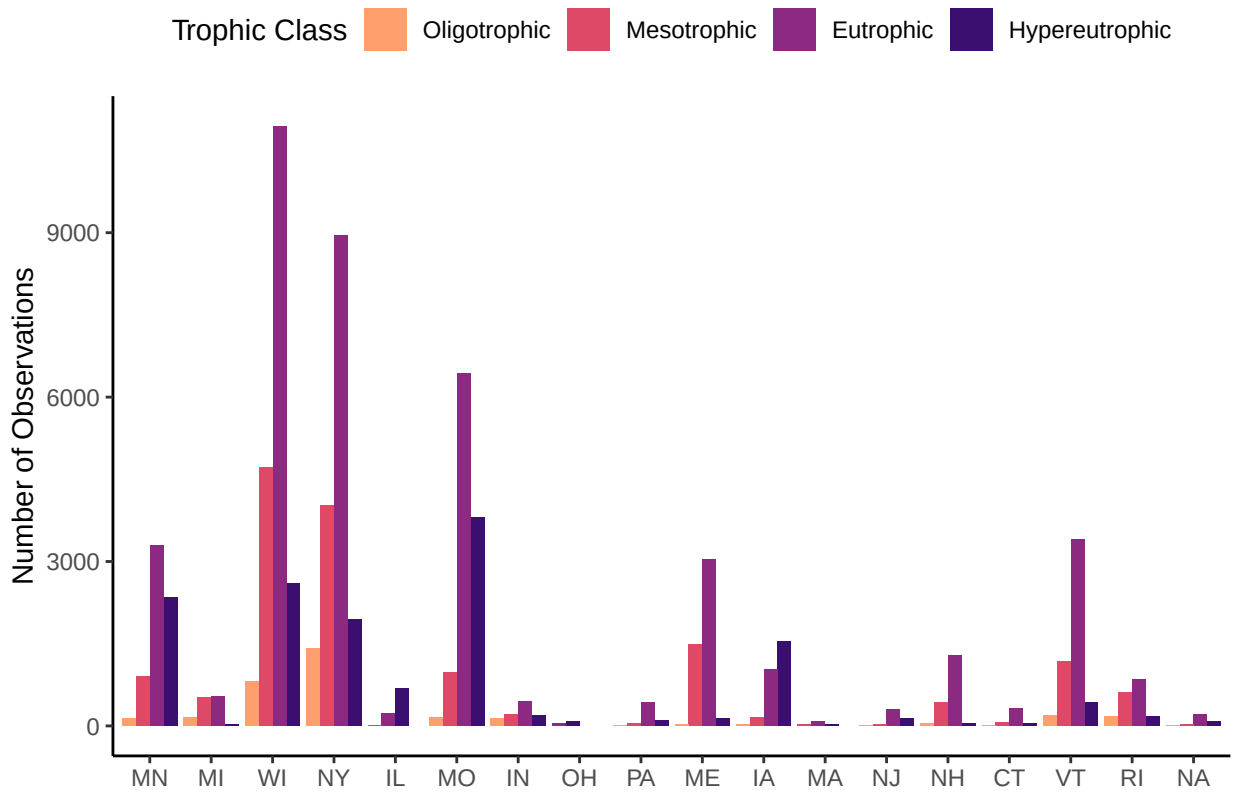
Here are three different ways to envision TSI across the states in the LAGOS-NE database.

```
stateTSIviolin <- ggplot(LAGOStrophic, aes(x = state, y = TSI.chl)) +
  geom_violin(draw_quantiles = 0.50)
print(stateTSIviolin)
```

```
## Warning: Removed 12 rows containing non-finite values (stat_ydensity).
```

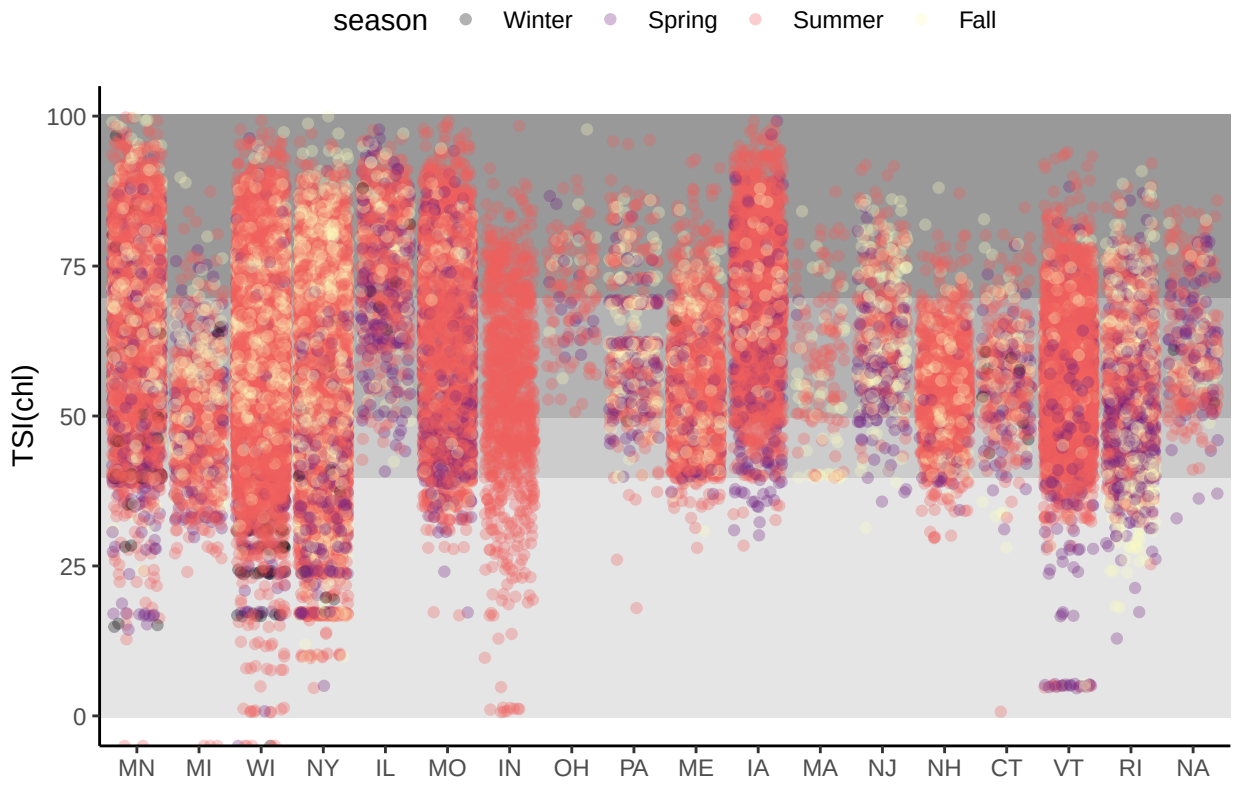



```
stateTSIbar <- ggplot(LAGOStrophic, aes(x = state, fill = trophic.class)) +
  geom_bar(stat = "count", position = position_dodge(preserve = "single")) +
  theme(legend.position = "top") +
  labs(x = "", y = "Number of Observations", fill = "Trophic Class") +
  scale_fill_viridis_d(option = "magma", begin = 0.2, end = 0.8, direction = -1)
print(stateTSIbar)
```




```
stateTSIjitter <- ggplot(LAGOStrophic, aes(x = state, y = TSI.chl, color = season)) +
  geom_rect(xmin = -1, xmax = 19, ymin = 0, ymax = 40,
            fill = "gray90", color = "gray90") +
  geom_rect(xmin = -1, xmax = 19, ymin = 40, ymax = 50,
            fill = "gray80", color = "gray80") +
  geom_rect(xmin = -1, xmax = 19, ymin = 50, ymax = 70,
            fill = "gray70", color = "gray70") +
  geom_rect(xmin = -1, xmax = 19, ymin = 70, ymax = 100,
            fill = "gray60", color = "gray60") +
  geom_jitter(alpha = 0.3) +
  # geom_hline(yintercept = 40, lty = 2) +
  # geom_hline(yintercept = 50, lty = 2) +
  # geom_hline(yintercept = 70, lty = 2) +
  labs(x = "", y = "TSI(chl)") +
  scale_y_continuous(limits = c(0, 100)) +
  theme(legend.position = "top") +
  scale_color_viridis_d(option = "magma")
print(stateTSIjitter)
```

```
## Warning: Removed 23 rows containing missing values (geom_point).
```



What insights do we gain from the different visualizations?

Violin:

Bar:

Jitter:

Closing Discussion

What factors might you expect to influence TSI scores in a given area? What are some variables in the LAGOSNE database that could help you test your hypothesis?