Analysis of nutrients and factors essential for the maintenance of lake water quality in Wisconsin

https://github.com/mse22/Eadala_ENV872_Project

Monisha Eadala

Contents

List of Tables

List of Figures

1 Rationale and Research Questions

The measurement of eutrophication is not an easy task. The aquatic ecosystem is very complex by constant interactions between physical, chemical and biological components. There are several indicators available to assess the degree of eutrophication. Nutrients such as phosphorus, phosphate, nitrogen and nitrates are the main elements that can be analyzed since high concentrations of these nutrients will mean more algae growth.

The presence of sufficient dissolved oxygen in the water column is very important for all aquatic life. Eutrophic waters (rich in nutrients) have fluctuating amounts of dissolved oxygen. Algae consume oxygen. A great algal biomass can therefore provide very low concentrations of oxygen in the water, sometimes so low that even fish can not survive and die.

Therefore, I hope to analyze variables such as total nitrogen, total phosphorus, dissolved oxygen, depth, temperature (in Celsius), months, days and years to better understand aspects of the eutrophication and lake water quality in general.

Research question 1: Have the lakes been within the limiting nutrient criteria over the years? How does the N:P ratio impact the dissolved oxygen in the lakes?

Research question 2: What are the best set of predictors of dissolved oxygen across the monitoring period at the North Temperate Lakes LTER

2 Dataset Information

The datasets in this repository contain data from studies on several lakes in the North Temperate Lakes District in Wisconsin, USA. Data were collected as part of the Long Term Ecological Research station established by the National Science Foundation.

Data were collected from the North Temperate Lakes Long Term Ecological Research website (https://lter.limnology.wisc.edu/about/overview). Data were collected using the Data tool on the website (https://lter.limnology.wisc.edu/data).

From the Data homepage, the following were typed and searched: 1. *Cascade Project at North Temperate Lakes LTER Core Data Physical and Chemical Limnology 1984 - 2016* 2. *Cascade Project at North Temperate Lakes LTER Core Data Nutrients 1991 - 2016*

Each of relevant files were downloaded through "Download All Data (csv)" button.

The two raw datasets relevant to my project are:

1. 'NTL-LTER_Lake_ChemistryPhysics_Raw.csv' file: This contains data relevant to physical and chemical variables (such as temperature, dissolved oxygen, and irradiance) that are measured at one central station near the deepest point of each lake. In most cases these measurements are made in the morning (8 to 9 am). Vertical profiles are taken at varied depth intervals. Chemical measurements are sometimes made in a pooled mixed layer sample (PML); sometimes in the epilimnion, metalimnion, and hypolimnion; and sometimes in vertical profiles. In the latter case, depths for sampling usually correspond to the surface plus depths of 50%, 25%, 10%, 5% and 1% of surface irradiance. (As noted in https://portal.edirepository.org/nis/metadataviewer?packageid=knb-lter-ntl.352.3)

This dataset contains the below column names and their relevant details:

Table 1: Chemistry-Physics Dataset Content

Column Name	Class	Units	Relevant Dataset Information
lakeid	factor	-	Provides the IDs of the lakes either in the form of capital letters or words; for example, L, R, T, E, Tbog, Roach, Ward, etc.
lakename	factor	-	Provides the names of the lakes; for example, Paul Lake, Peter Lake, Tuesday Lake, East Long Lake, etc.
year4	integer	-	Provides the year in which its respective data was collected in four digits from 1984 to 2016
daynum	integer	-	Provides the number of the day on which its data was collected from from 1 to 366
sampledate	factor	-	Provides the date on which its data was collected in m/d/y format

Column Name	Class	Units	Relevant Dataset Information
depth	numeric	m	Provides the depth at which the data sample was collected in meters
temperature_C	numeric	C	Provides the water temperature in Celsius
dissolvedOxygen	numeric	mg/L	Provides the dissolved oxygen measurement in milligrams per liter
irradianceWater	numeric	uE	Provides the photosynthetically active radiation measured in the water column in micro-Einstein
irradianceDeck	numeric	uE	Provides the photosynthetically active radiation measured on the deck of the sampling boat in micro-Einstein
comments	factor	-	Provides the comments noting departure from standard procedure

2. 'NTL-LTER_Lake_Nutrients_Raw.csv' file: This contains the data relevant to physical and chemical variables (such as total nitrogen, total phosphorus, ammonia and ammonium, nitrite and nitrate, and phosphate concentrations) that are measured at one central station near the deepest point of each lake. In most cases these measurements are made in the morning (8 to 9 am). Vertical profiles are taken at varied depth intervals. Chemical measurements are sometimes made in a pooled mixed layer sample (PML); sometimes in the epilimnion, metalimnion, and hypolimnion; and sometimes in vertical profiles. In the latter case, depths for sampling usually correspond to the surface plus depths of 50%, 25%, 10%, 5% and 1%t of surface irradiance. The 1991-1999 chemistry data was obtained from the Lachat auto-analyzer. Like the process data, there are up to seven samples per sampling date due to Van Dorn collections across a depth interval according to percent irradiance. Voichick and LeBouton (1994) describe the autoanalyzer procedures in detail. Nutrient samples were sent to the Cary Institute of Ecosystem Studies for analysis beginning in 2000. The Kjeldahl method for measuring nitrogen is not used at IES, and so measurements reported from 2000 onwards are Total Nitrogen. (As noted in https://portal.edirepository.org/nis/metadataviewer?packageid=knb-lter-ntl.351.3)

This dataset contains the below column names and their relevant details:

Table 2: Nutrients Dataset Content

Column Name	Class	Units	Relevant Dataset Information
lakeid	factor	-	Provides the IDs of the lakes either in the form of capital letters or words; for example, L, R, T, E, Tbog, Roach, Ward, etc.
lakename	factor	-	Provides the names of the lakes; for example, Paul Lake, Peter Lake, Tuesday Lake, East Long Lake, etc.
year4	integer	-	Provides the year in which its respective data was collected in four digits from 1991 to 2016
daynum	integer	-	Provides the number of the day on which its data was collected from from 1 to 366
sampledate	factor	-	Provides the date on which its data was collected in m/d/y format
depth_id	integer	-	Provides the depth level categorized from -2 to 7; 1 represents 100% light, 2 represents 50% light, 3 represents 25% light; 4 represents 10% light, 5 represents 5% light, 6 represents 1% light, 7 means Hypolimnion, -1 means Epilimnion/PML, and -2 means Metalimnion
depth	numeric	m	Provides the depth at which the data sample was collected in meters
tn_ug	numeric	ug/L	Provides the total nitrogen concentration in micrograms per liter
tp_ug	numeric	ug/L	Provides the total phosphorus concentration in micrograms per liter
nh34	numeric	ug/L	Provides the ammonia and ammonium concentration in micrograms per liter
no23	numeric	ug/L	Provides the nitrite and nitrate concentration in micrograms per liter

Column Name	Class	Units	Relevant Dataset Information
po4	numeric	ug/L	Provides the phosphate concentration in micrograms per liter
comments	factor	-	Provides any additional comments

Wrangling Process from raw to processed data:

The raw data were first explored to understand list of unique columns and rows each contains. Then a call was made to only include columns and data that was available in both the datasets since I intended to combine the two datasets to analyse how elements from the chemistry-physics dataset are impacted by the nutrients in the nutrients dataset. So the chosen years were 1991-20016 since they were available in both the datasets. Similarly, only 6 lakes that were common in both the datasets were selected. Selection was also made based on the quantity of data available for each one of the lakes. Lakes that have very little data available were excluded from the combined processed dataset even though they were common to both the nutrients and the chemistry-physics datasets.

Also, since combining two datasets means more number NAs, I decided to only study selected variables available in both the datasets. For example, I chose not to study the phosphates, nitrates/nitrites and ammonia/ammonium variables from the nutrient dataset to minimize the number of the NAs that might show up in the combined processed dataset. Therefore, I choose to focus on nitrogen (TN) and phosphorus (TP) concentrations along with other physical variables such are days and depth in the nutrients data. Similarly, I choose not to include both the irradiance variables and instead focus on dissolved oxygen and temperature in the chemistry-physics dataset. Data was also wrangled to include a new column, N:P, to understand the nutrient limitations and efficient water management techniques. This was done initially in the data exploring stage and purposes of the project itself.

3 Exploratory Analysis

```r
# Read in nutrients data
Nutrients_raw <- read.csv("./Data/Raw/NTL-LTER_Lake_Nutrients_Raw.csv")
# Read in chemistry-physics data
ChemistryPhysics_raw <- read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")

# Format Dates
Nutrients_raw$sampledate <- as.Date(Nutrients_raw$sampledate,
format = "%m/%d/%y")
ChemistryPhysics_raw$sampledate <- as.Date(ChemistryPhysics_raw$sampledate,
format = "%m/%d/%y")
```

Data exploration of the nutrients and chemistry-physics raw data files associated with NTL-LTER lakes.

```r
dim(Nutrients_raw)
```

```
## [1] 5836   13
```

```r
dim(ChemistryPhysics_raw)
```

```
## [1] 38614    11
```

```r
str(Nutrients_raw)
```

```
## 'data.frame':    5836 obs. of  13 variables:
##  $ lakeid    : Factor w/ 26 levels "B","Berg","Bolg",..: 13 13 13 13 13 13 18 18 18 1
##  $ lakename  : Factor w/ 26 levels "Bergner Lake",..: 16 16 16 16 16 16 17 17 17 17 .
##  $ year4     : int  1991 1991 1991 1991 1991 1991 1991 1991 1991 1991 ...
##  $ daynum    : int  140 140 140 140 140 140 140 140 140 140 ...
##  $ sampledate: Date, format: "1991-05-20" "1991-05-20" ...
##  $ depth_id  : int  1 2 3 4 5 6 1 2 3 4 ...
##  $ depth     : num  0 0.85 1.75 3 4 6 0 1 2.25 3.5 ...
##  $ tn_ug     : num  538 285 399 453 363 583 352 356 364 582 ...
##  $ tp_ug     : num  25 14 14 14 13 37 11 15 28 14 ...
##  $ nh34      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ no23      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ po4       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ comments  : Factor w/ 3 levels "","sample missing",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```r
str(ChemistryPhysics_raw)
```

```
## 'data.frame':    38614 obs. of  11 variables:
##  $ lakeid         : Factor w/ 9 levels "C","E","H","L",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ lakename       : Factor w/ 9 levels "Central Long Lake",..: 5 5 5 5 5 5 5 5 5 5 ..
##  $ year4          : int  1984 1984 1984 1984 1984 1984 1984 1984 1984 1984 ...
##  $ daynum         : int  148 148 148 148 148 148 148 148 148 148 ...
##  $ sampledate     : Date, format: "1984-05-27" "1984-05-27" ...
##  $ depth          : num  0 0.25 0.5 0.75 1 1.5 2 3 4 5 ...
##  $ temperature_C  : num  14.5 NA NA NA 14.5 NA 14.2 11 7 6.1 ...
##  $ dissolvedOxygen: num  9.5 NA NA NA 8.8 NA 8.6 11.5 11.9 2.5 ...
##  $ irradianceWater: num  1750 1550 1150 975 870 610 420 220 100 34 ...
##  $ irradianceDeck : num  1620 1620 1620 1620 1620 1620 1620 1620 1620 1620 ...
##  $ comments       : Factor w/ 2 levels "DO Probe bad - Doesn't go to zero",..: NA NA
```

```r
colnames(Nutrients_raw)
```

```
##  [1] "lakeid"     "lakename"   "year4"      "daynum"     "sampledate"
##  [6] "depth_id"   "depth"      "tn_ug"      "tp_ug"      "nh34"      
## [11] "no23"       "po4"        "comments"  
```

```r
colnames(ChemistryPhysics_raw)
```

```
##  [1] "lakeid"           "lakename"         "year4"            "daynum"         
##  [5] "sampledate"       "depth"            "temperature_C"    "dissolvedOxygen"
##  [9] "irradianceWater"  "irradianceDeck"   "comments"        
```

```r
summary(Nutrients_raw)
```

```
##      lakeid                     lakename        year4          daynum     
##  R      :1387   Peter Lake        :1387   Min.   :1991   Min.   :105.0  
##  L      :1383   Paul Lake         :1383   1st Qu.:1993   1st Qu.:167.0  
##  W      : 954   West Long Lake    : 954   Median :1996   Median :195.0  
##  E      : 926   East Long Lake    : 926   Mean   :1999   Mean   :193.9  
##  T      : 699   Tuesday Lake      : 699   3rd Qu.:2001   3rd Qu.:222.0  
##  C      : 139   Central Long Lake: 139   Max.   :2016   Max.   :254.0  
##  (Other): 348   (Other)           : 348                                 
##    sampledate            depth_id          depth            tn_ug        
##  Min.   :1991-05-20   Min.   :-2.000   Min.   : 0.000   Min.   :   0.0  
##  1st Qu.:1993-09-03   1st Qu.: 1.000   1st Qu.: 0.500   1st Qu.: 351.9  
##  Median :1996-08-01   Median : 3.000   Median : 1.300   Median : 453.9  
##  Mean   :1999-06-04   Mean   : 2.976   Mean   : 2.878   Mean   : 593.3  
##  3rd Qu.:2001-08-16   3rd Qu.: 5.000   3rd Qu.: 3.700   3rd Qu.: 664.9  
##  Max.   :2016-08-17   Max.   : 7.000   Max.   :12.000   Max.   :3497.7  
##                       NA's   :80       NA's   :782      NA's   :2330    
##      tp_ug              nh34               no23              po4          
##  Min.   : -6.349   Min.   : -18.700   Min.   :  -1.856   Min.   : -0.3333  
##  1st Qu.: 11.000   1st Qu.:   9.651   1st Qu.:   1.979   1st Qu.:  1.2437  
##  Median : 18.663   Median :  21.529   Median :   4.087   Median :  2.6055  
##  Mean   : 30.040   Mean   : 155.913   Mean   :  47.218   Mean   :  8.3046  
##  3rd Qu.: 33.999   3rd Qu.: 154.884   3rd Qu.:  40.811   3rd Qu.:  5.0000  
##  Max.   :352.056   Max.   :2713.684   Max.   :1574.824   Max.   :373.8360  
##  NA's   :317       NA's   :3209       NA's   :3149       NA's   :3090      
##                            comments
##                                 :5834  
##  sample missing                 :   1  
##  TP value suspect, far too high:   1  
##                                        
##                                        
##                                        
##                                        
```

```r
summary(ChemistryPhysics_raw)
```

```
##      lakeid                lakename         year4          daynum     
##  R       :11288   Peter Lake    :11288   Min.   :1984   Min.   : 55.0  
##  L       :10325   Paul Lake     :10325   1st Qu.:1991   1st Qu.:166.0  
##  T       : 6107   Tuesday Lake  : 6107   Median :1997   Median :194.0  
##  W       : 4188   West Long Lake: 4188   Mean   :1999   Mean   :194.3  
##  E       : 3905   East Long Lake: 3905   3rd Qu.:2006   3rd Qu.:222.0  
##  M       : 1234   Crampton Lake : 1234   Max.   :2016   Max.   :307.0  
##  (Other): 1567   (Other)       : 1567                                 
##    sampledate             depth       temperature_C   dissolvedOxygen 
##  Min.   :1984-05-27   Min.   : 0.00   Min.   : 0.30   Min.   :  0.00  
##  1st Qu.:1991-08-08   1st Qu.: 1.50   1st Qu.: 5.30   1st Qu.:  0.30  
##  Median :1997-07-28   Median : 4.00   Median : 9.30   Median :  5.60  
##  Mean   :1999-02-05   Mean   : 4.39   Mean   :11.81   Mean   :  4.97  
##  3rd Qu.:2006-06-06   3rd Qu.: 6.50   3rd Qu.:18.70   3rd Qu.:  8.40  
##  Max.   :2016-08-17   Max.   :20.00   Max.   :34.10   Max.   :802.00  
##                                       NA's   :3858    NA's   :4039    
##  irradianceWater     irradianceDeck                                  comments    
##  Min.   :   -0.337   Min.   :   1.5   DO Probe bad - Doesn't go to zero:  206  
##  1st Qu.:   14.000   1st Qu.: 353.0   DO taken with Jones Lab Meter    :  162  
##  Median :   65.000   Median : 747.0   NA's                             :38246  
##  Mean   :  210.242   Mean   : 720.5                                            
##  3rd Qu.:  265.000   3rd Qu.:1042.0                                            
##  Max.   :24108.000   Max.   :8532.0                                            
##  NA's   :14287       NA's   :15419                                             
```

Visual data exploration of the nutrients data raw file in Figure 1, Figure 2, and Figure 3.

Visual data exploration of the chemistry-physics data raw file in Figure 4.

Wrangling of data to include only the most meaningful variables and compatible data for combining the two datasets.

```r
# Find out what years are included in the chemistry-physics dataset
list(unique(ChemistryPhysics_raw$year4))
```

```
## [[1]]
##  [1] 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998
## [16] 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
```

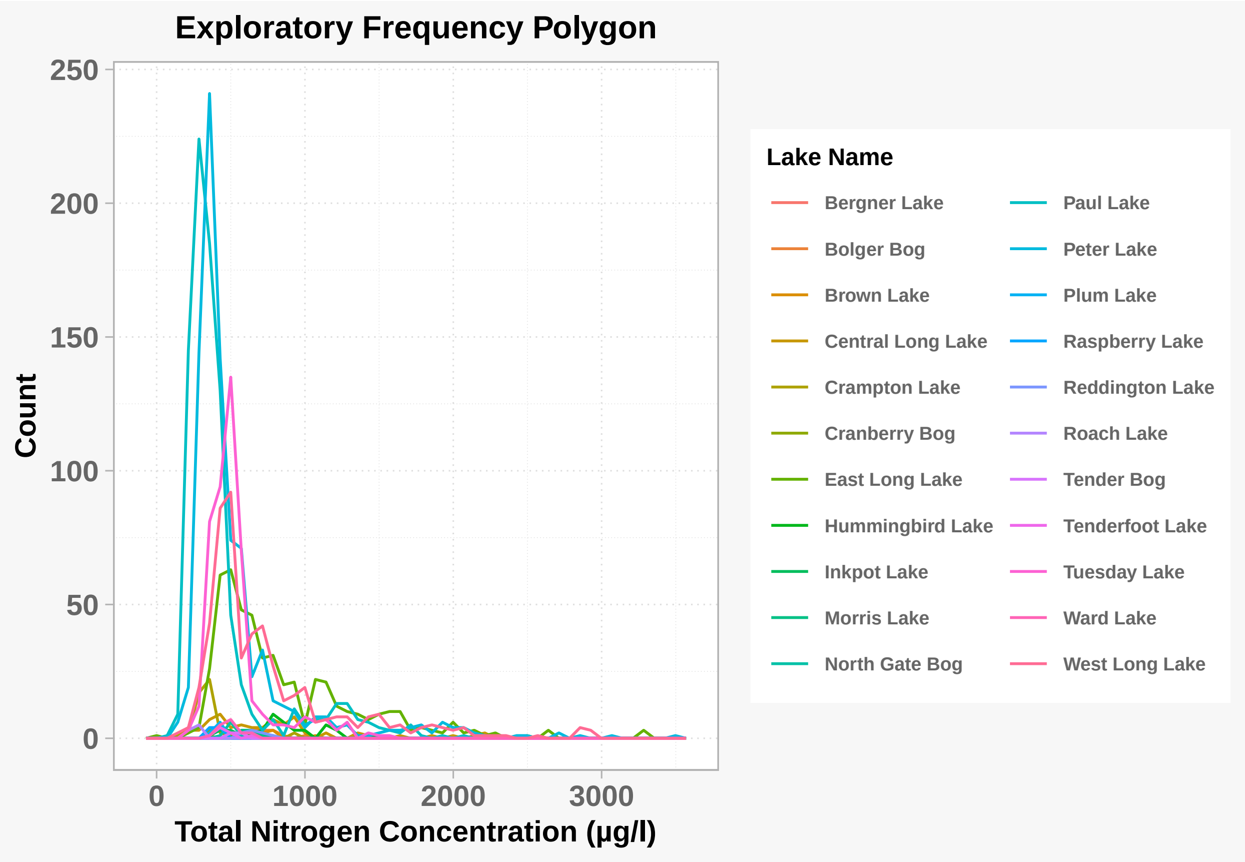


Figure 1: Nitrogen concentration frequency polygon.

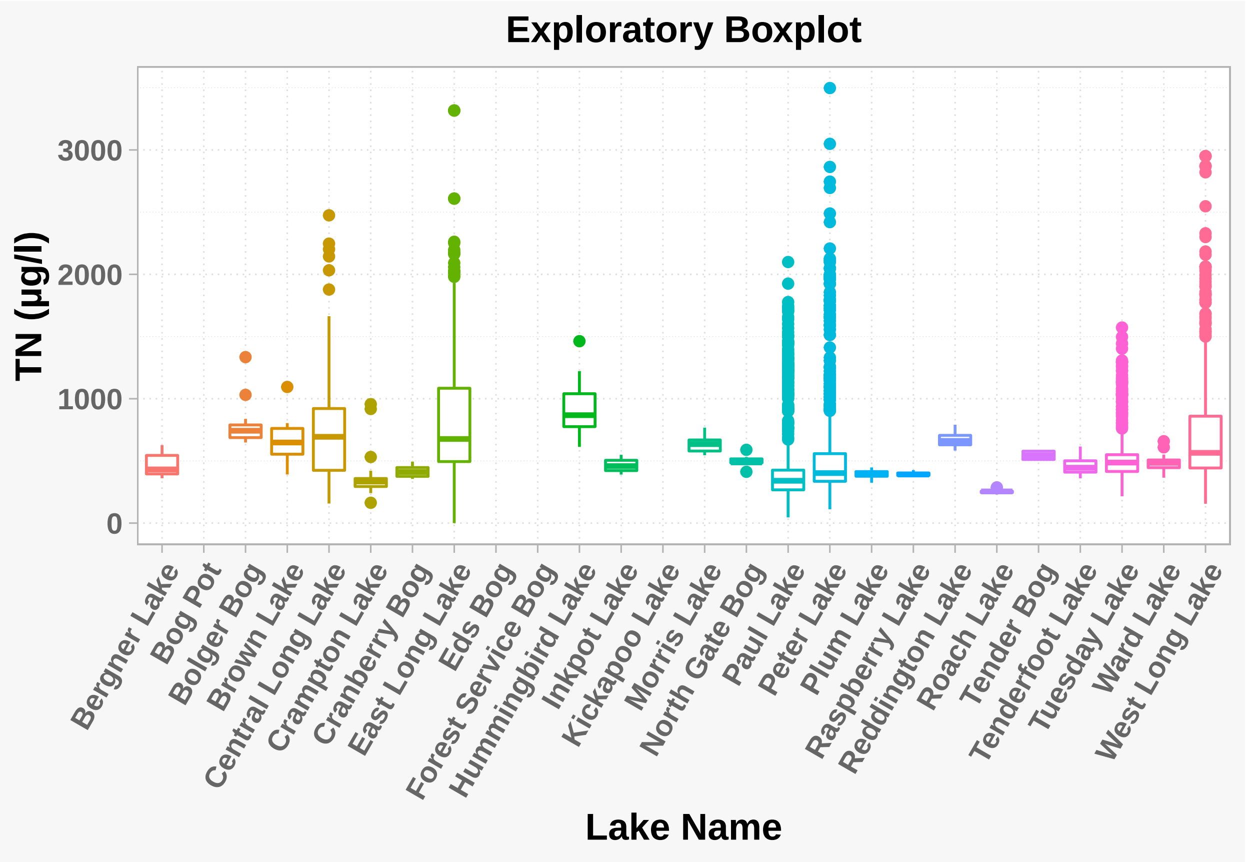


Figure 2: Nitrogen concentration boxplot.

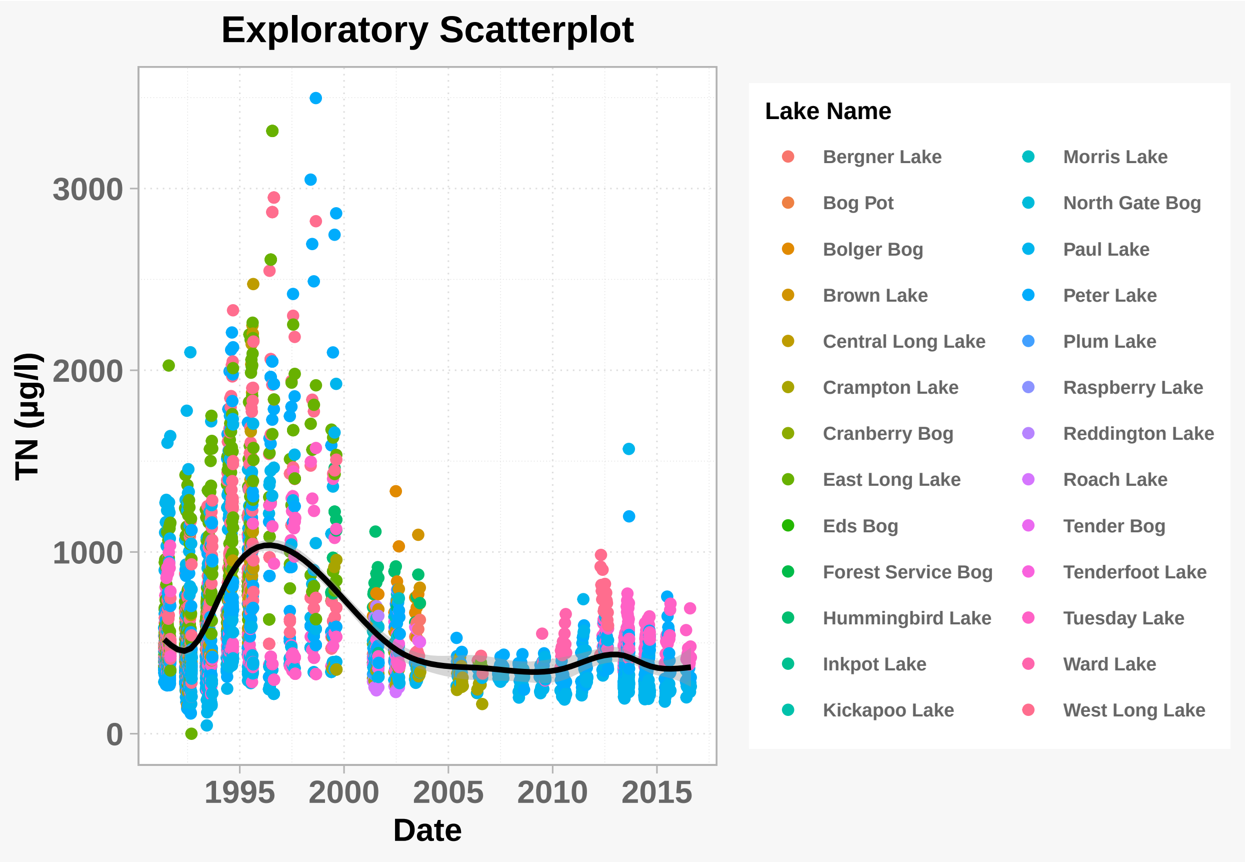


Figure 3: Nitrogen concentration scatterplot.

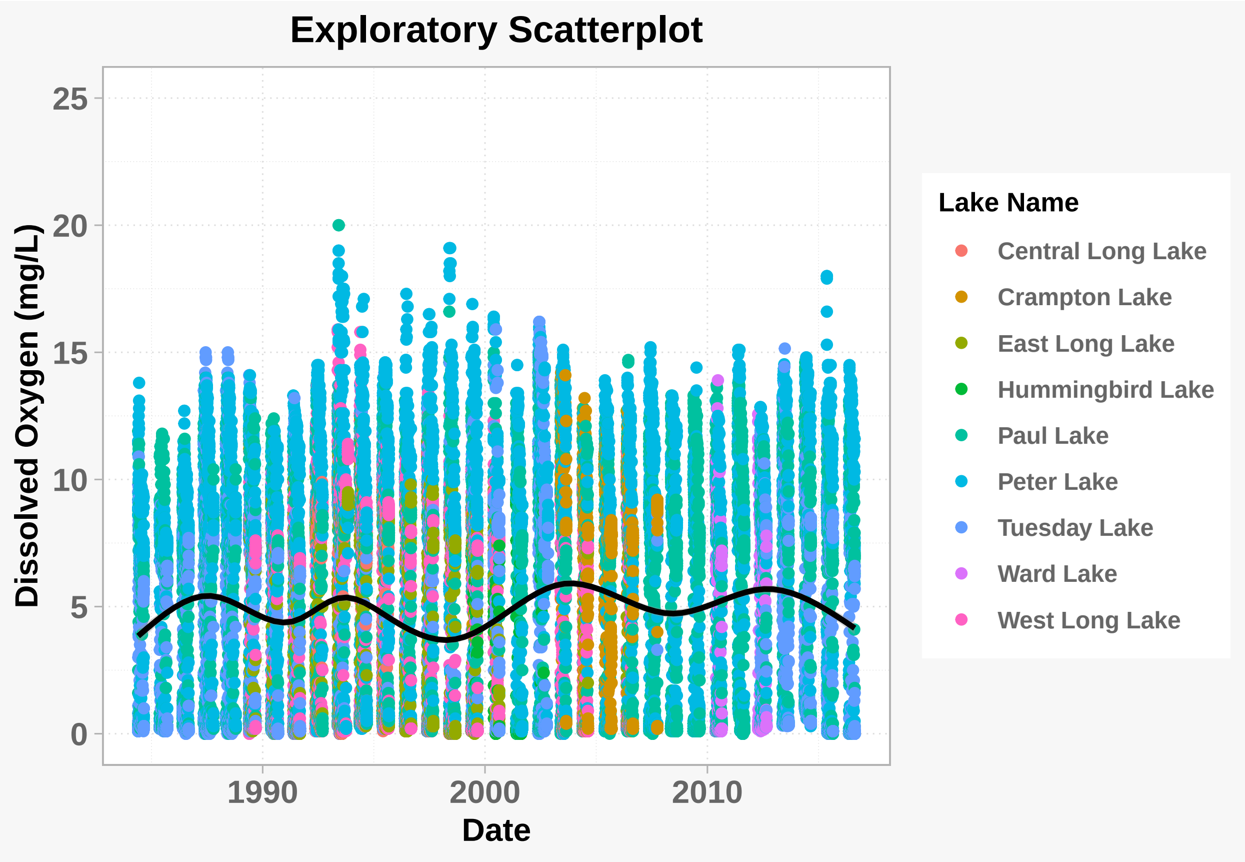


Figure 4: Dissolved oxygen scatterplot.

```
## [31] 2014 2015 2016
```

```
# Find out what years are included in the Nutrients dataset
list(unique(Nutrients_raw$year4))
```

```
## [[1]]
##  [1] 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2005 2006
## [16] 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
```

```
# Keep only useful columns to minimize the number of NAs later and filter only the yea
Nutrients_processed <- Nutrients_raw %>%
  select(-lakeid, -depth_id , -nh34, -no23, -po4, -comments) %>%
  filter(year4 == 1991 | year4 == 1992 | year4 ==  1993 | year4 ==  1994 | year4 == 199

# Keep only useful columns to minimize the number of NAs later and filter only the yea
ChemistryPhysics_processed <- ChemistryPhysics_raw %>%
  select(-lakeid, -irradianceDeck, -irradianceWater, -comments) %>%
  filter(year4 == 1991 | year4 == 1992 | year4 ==  1993 | year4 ==  1994 | year4 == 199

# List the unique lake names in the processed nutrients dataset
list(unique(Nutrients_processed$lakename))
```

```
## [[1]]
##  [1] Paul Lake          Peter Lake         East Long Lake     West Long Lake
##  [5] Tuesday Lake       Central Long Lake  Hummingbird Lake   Crampton Lake
##  [9] Brown Lake         Bergner Lake       Bolger Bog         Bog Pot
## [13] Cranberry Bog      Eds Bog            Forest Service Bog Inkpot Lake
## [17] Kickapoo Lake      Morris Lake        North Gate Bog     Plum Lake
## [21] Raspberry Lake     Reddington Lake    Roach Lake         Tenderfoot Lake
## [25] Ward Lake          Tender Bog
## 26 Levels: Bergner Lake Bog Pot Bolger Bog Brown Lake ... West Long Lake
```

```
# List the unique lake names in the processed chemistry-physics dataset
list(unique(ChemistryPhysics_processed$lakename))
```

```
## [[1]]
## [1] Paul Lake         Peter Lake        East Long Lake    West Long Lake
## [5] Tuesday Lake      Central Long Lake Hummingbird Lake  Crampton Lake
## [9] Ward Lake
## 9 Levels: Central Long Lake Crampton Lake East Long Lake ... West Long Lake
```

```
# Combine the two datasets and filter only the common and interested lakes from the tw
Combined_processed <- full_join(Nutrients_processed, ChemistryPhysics_processed) %>%
  filter(year4 == 1991 | year4 == 1992 | year4 ==  1993 | year4 ==  1994 | year4 == 199
  filter(lakename == "Paul Lake" | lakename == "Peter Lake" | lakename == "East Long Lak

# Save the combined data in the processed folder
```

```
write.csv(Combined_processed, row.names = FALSE, file = "./Data/Processed/Combined_proce
```

A Redfield ratio N:P of 16:1 by moles in general indicates a roughly balanced supply of N and P, and algae assemblages tend to mirror this ratio fairly closely when growing under balanced growth conditions (https://www.sciencedirect.com/topics/earth-and-planetary-sciences/redfield-ratio). Converting the molar ratio to migrogram ratio gives us N:P of 7.24:1, since 1 mole of N = 14.0067g and 1 mole of P = 30.973761g (https://www.convertunits.com/from/grams+Nitrogen/to/moles; https://www.convertunits.com/from/grams+Phosphorus/to/moles)

First question for further data exploring: *Are the lakes phosphorus or nitrogen deficient?*

```
#To exlude some of the valariables from the dataset so that the number of NAs in the d
NitrogenPhosphorus <- Combined_processed %>%
  select(-temperature_C) %>%
  na.exclude()

# Save the new folder
write.csv(NitrogenPhosphorus, row.names = FALSE, file = "./Data/Processed/NitrogenPhosph
```

What is the average total nitrogen/total phosphorus (N:P) ratio for each of the six north temperate lakes?

```
NitrogenPhosphorus$NPRatio <- NitrogenPhosphorus$tn_ug / NitrogenPhosphorus$tp_ug # Give

# To summarize the N:P ratio in each lake
NPRatio.summary <- NitrogenPhosphorus %>%
   group_by(lakename) %>%
  summarise(mean.NPRatio = mean(NPRatio, na.rm = TRUE),
            minimum.NPRatio = min(NPRatio, na.rm = TRUE),
            maximum.NPRatio = max(NPRatio, na.rm = TRUE),
            Standard.dev.NPRatio = sd(NPRatio, na.rm = TRUE)) # Gives us summary stats
```

Figure 5 offers a visual to understand if most of the lakes are above or below the Redfield Ratio. In other words, to understand if most of the lakes are phosphorus or nitrogen limited.

Dashed line indicates the Redfield Ratio of 7.24:1, which indicates optimal conditions for phytoplankton growth. The lakes above the Redfield Ratio would be phosphorus limited while those below it would be nitrogen limited.

Figure 5 Tells us that most of the Lakes are phosphorus deficient, therefore phosphorus should be the limiting nutrient prioritized or addressed during water management.

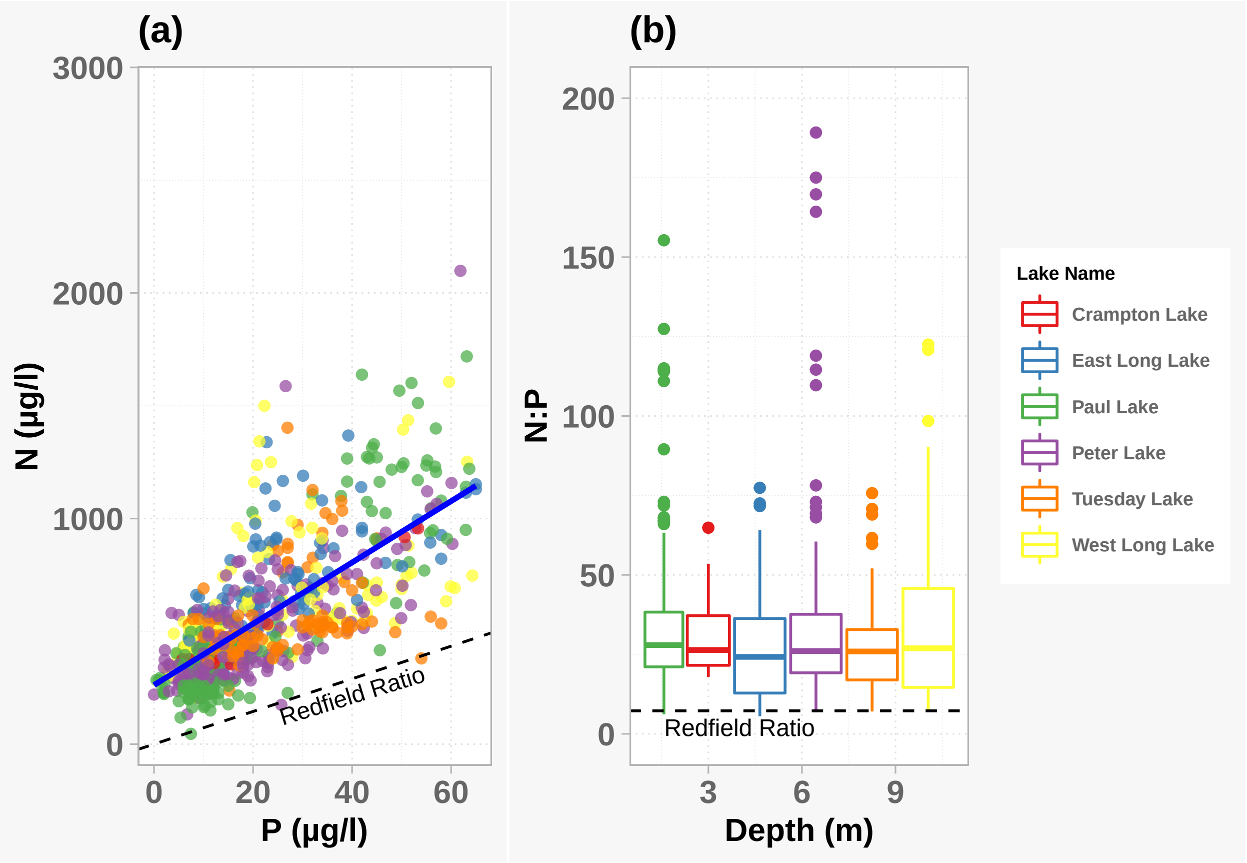


Figure 5: (a) Nitrogen vs phosphorus concentration (b) Distribution of N:P Ratio.

4 Analysis

4.1 Question 1: Have the lakes been within the limiting nutrient criteria over the years? How does the N:P ratio impact the dissolved oxygen in the lakes?

One-sample t-test

The figures under exploratory analysis tell us that all the eight lakes are phosphorus deficient, and therefore phosphorus is the element that needs to be controlled on priority to avoid eutrophication. EPA has compiled state, territorial, and authorized tribal water quality standards that EPA has approved or are otherwise in effect for Clean Water Act purposes. According to this compilation, Wisconsin's stratified lakes' total phosphorus criterion is not more than 30µg/l (https://www.epa.gov/wqs-tech/state-specific-water-quality-standards-effective-under-clean-water-act-cwa#tb2). According to Carlson R.E. and J. Simpson in 1996, phosphorus concentration between 24µg/l and 96µg/l suggests eutrophic conditions, and anything above 96µg/l suggests hypereutrophic conditions (https://en.wikipedia.org/wiki/Trophic_state_index). Therefore, it is important to check if the eight North Temperate Lakes are at least within the 30µg/l criterion or not.

This statistical analysis will test the null hypothesis that the means of the total phosphorus concentrations in the eight North Temperate Lakes are below the regulatory standard of 30µg/l.

First, the assumption of normal distribution is evaluated through the Shapiro-Wilk normality test.

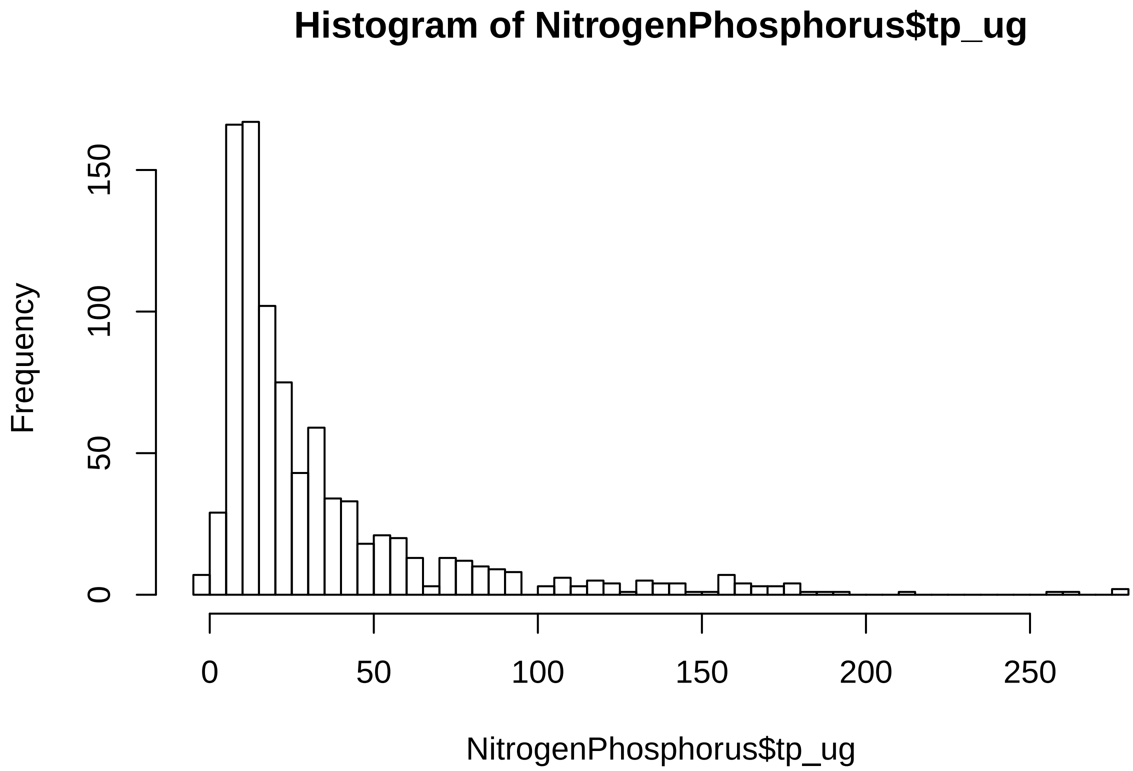
```r
shapiro.test(NitrogenPhosphorus$tp_ug)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  NitrogenPhosphorus$tp_ug
## W = 0.69337, p-value < 2.2e-16
```

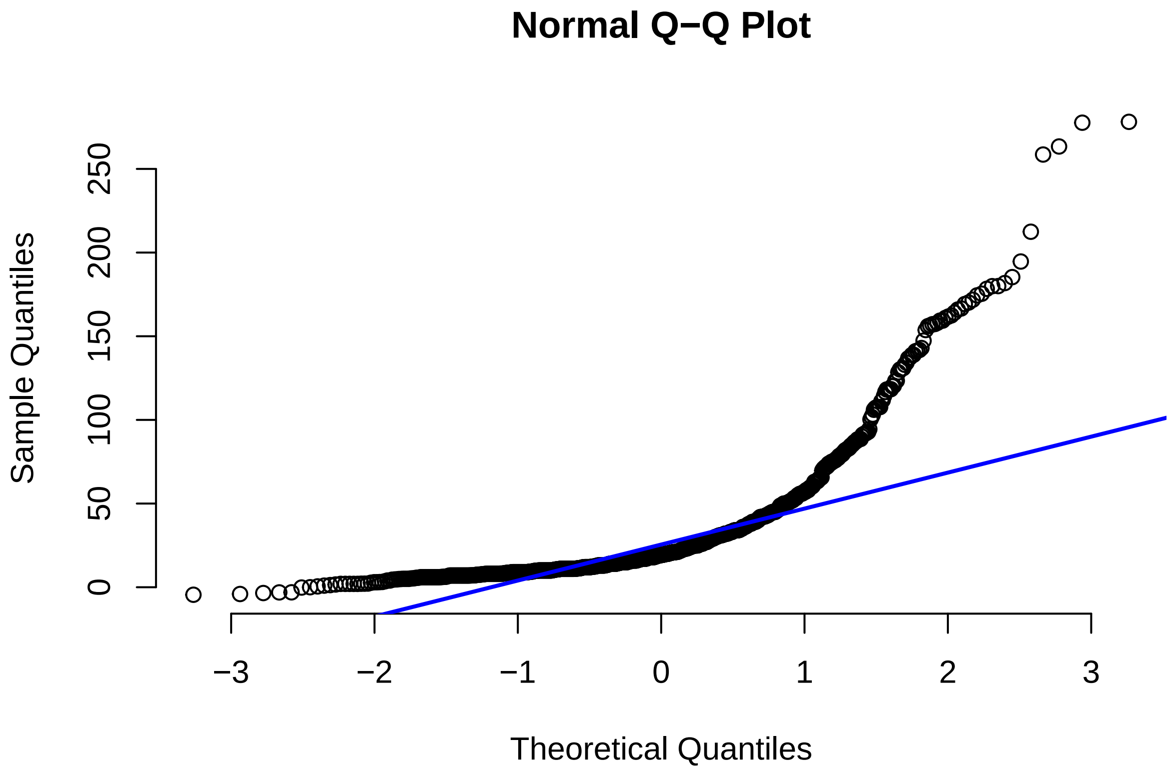
The Shapiro-Wilk normality test says that the total phosphorus concentrations data in the six lakes of Wisconsin are significantly different from a normal distribution (Shapiro-Wilk normality test; W = 0.69337, p-value < 0.0001)

Next, a visual analysis of the data is performed.

```r
hist(NitrogenPhosphorus$tp_ug, breaks = 50) # Phosphorus concentration histogram
```



```r
qqnorm(NitrogenPhosphorus$tp_ug, pch = 1, frame = FALSE)
qqline(NitrogenPhosphorus$tp_ug, col = "blue", lwd = 2)
```



In the above figures it can be seen that the data distribution is right-skewed with a longer tail to the right than a normal distribution; nevertheless, environmental data often violate the assumptions of normality and since the sample size is large, a t-test is performed anyway.

```
t.test(NitrogenPhosphorus$tp_ug, mu = 30, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  NitrogenPhosphorus$tp_ug
## t = 3.2053, df = 907, p-value = 0.9993
## alternative hypothesis: true mean is less than 30
## 95 percent confidence interval:
##      -Inf 36.36745
## sample estimates:
## mean of x
##  34.20656
```

According to the One Sample t-test, the TP concentrations for the North Temperate Lakes of Wisconsin from 1991 to 2016 have not been significantly lower than 30µg/l (one sample t-test; t = 3.2053, df = 907, p-value $>$ 0.0001).

Figure 6 presents a visualization of the TP concentrations in Wisconsin lakes data in comparison with the Wisconsin regulatory standard.

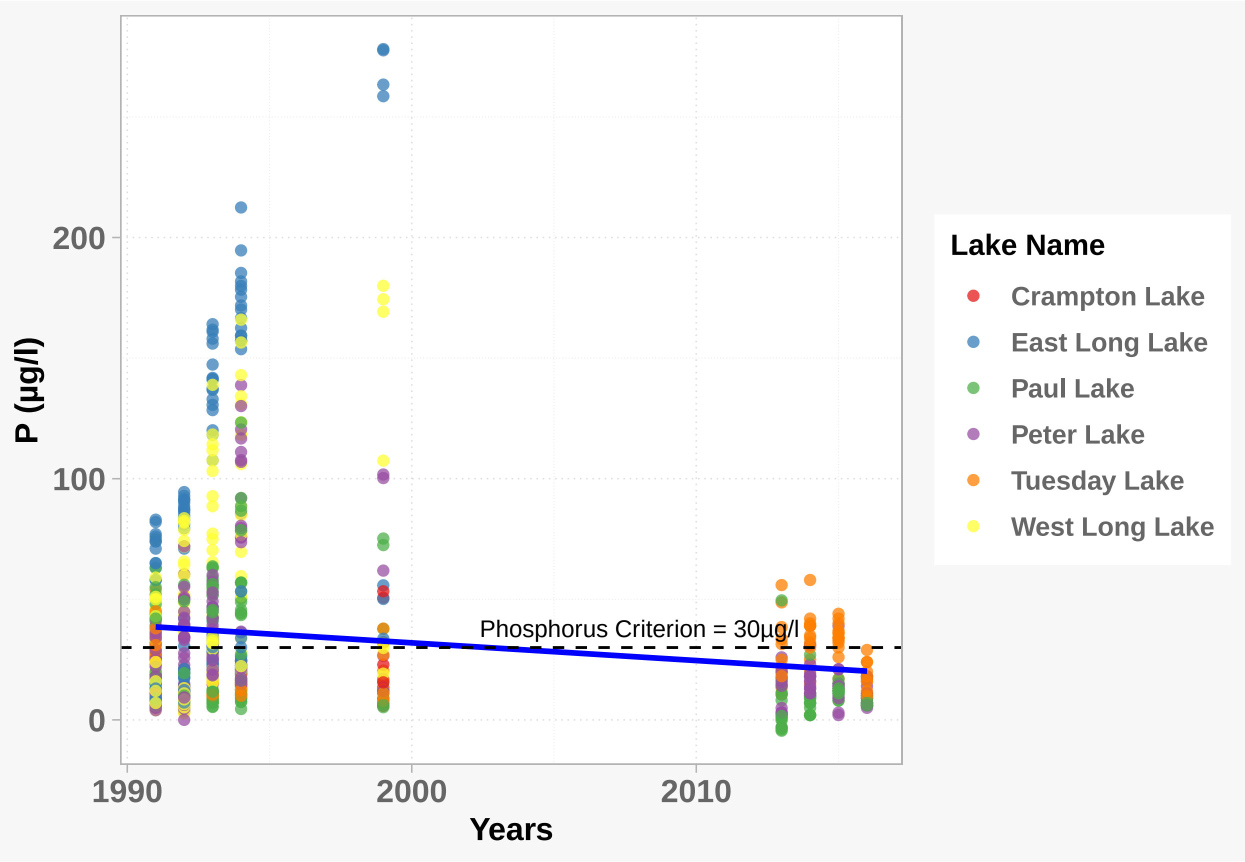


Figure 6: Phosphorus concenration over years.

The blue line in comparison with the black dashed line indicates that most of the Wisconsin lakes started off with phosphorus levels above the criteria but have been able to lower their phosphorus in the most recent recent years.

Coming the the next part of the research question, figure 7 gives us the visual interpretation of how the N:P ratio affects the dissolved oxygen.

Figure 7 tells us that an increase in the N:P ratio increased dissolved oxygen in the lakes. This is consistent with the theory that controlling phosphorus concentration in phosphorus deficient lakes can can preserve their water quality. However, phosphorus might not be the only factor affecting the dissolved oxygen and eventually the water quality of water. Therefore, it is important to understand what other variables affect dissolved oxygen, whose depletion indicates eutrophication in many cases.

4.2 Question 2: What are the best set of predictors of dissolved oxygen across the monitoring period at the North Temperate Lakes LTER?

Linear model

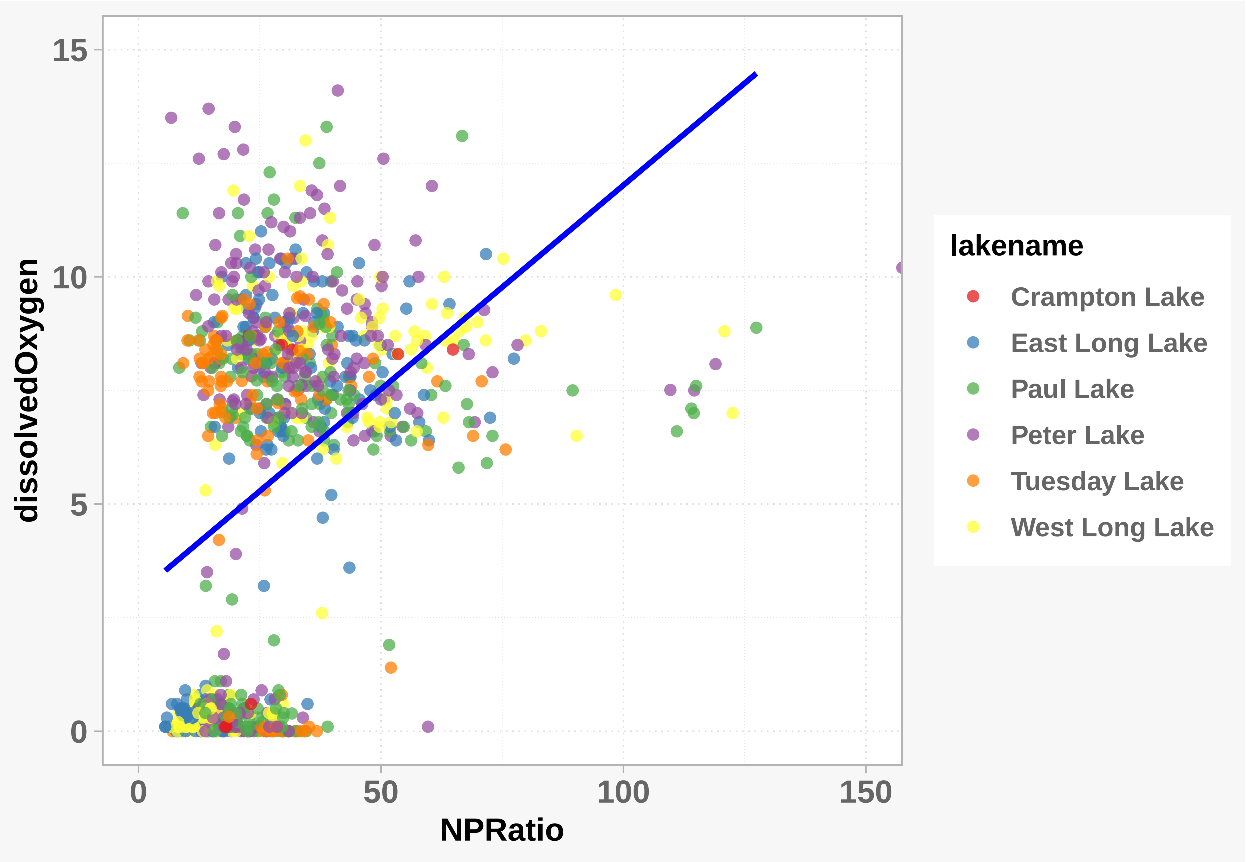


Figure 7: Dissolved oxygen over N:P ratio

To answer this question, a generalized linear model approach was taken.

Since multiple explanatory variables are being considered at a time in the model, it is important to check the model is not over-parameterized. Therefore, the Akaike's Information Criteria (AIC) was used.

```
DO.predictors <- Combined_processed %>%
  na.exclude()

DO.AIC <- lm(data = DO.predictors, dissolvedOxygen ~ year4 + daynum + temperature_C + tn
step(DO.AIC) # The lower the AIC value, the better
```

```
## Start:  AIC=1015.33
## dissolvedOxygen ~ year4 + daynum + temperature_C + tn_ug + tp_ug +
##     depth
##
##                 Df Sum of Sq    RSS    AIC
## <none>                       2735.7 1015.3
## - tp_ug          1     25.85 2761.6 1021.9
## - year4          1     29.13 2764.9 1022.9
## - daynum         1     39.13 2774.9 1026.2
## - tn_ug          1    100.08 2835.8 1045.9
## - temperature_C  1    119.62 2855.3 1052.2
## - depth          1   2046.68 4782.4 1519.9

##
## Call:
## lm(formula = dissolvedOxygen ~ year4 + daynum + temperature_C +
##     tn_ug + tp_ug + depth, data = DO.predictors)
##
## Coefficients:
##   (Intercept)          year4         daynum  temperature_C          tn_ug
##     59.781118      -0.023206      -0.006947      -0.137099      -0.001380
##         tp_ug          depth
##      0.007627      -0.874667
```

The best predictors from the lowest AIC are: year, daynum, temperature_C, tn_ug, tp_ug and depth.

Next, a multiple regression is performed on the recommended set of variables.

```
DO.model <- lm(data = DO.predictors, dissolvedOxygen ~ year4 + daynum + temperature_C +
anova(DO.model)
```

```
## Analysis of Variance Table
##
## Response: dissolvedOxygen
##                Df Sum Sq Mean Sq  F value    Pr(>F)
## year4           1  759.0   759.0  249.705 < 2.2e-16 ***
## daynum          1  275.3   275.3   90.573 < 2.2e-16 ***
## temperature_C   1 8357.3  8357.3 2749.378 < 2.2e-16 ***
## depth           1 2407.5  2407.5  792.007 < 2.2e-16 ***
## tn_ug           1   76.4    76.4   25.141 6.419e-07 ***
## tp_ug           1   25.8    25.8    8.503  0.003634 **
## Residuals     900 2735.7     3.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The full model also reconfirmed the above variables as the best. So model is run with the same variables.

```
DO.model <- lm(data = DO.predictors, dissolvedOxygen ~ year4 + daynum + temperature_C +
summary(DO.model) # Runs a multiple regression on the above recommended set of variabl
```

```
##
## Call:
## lm(formula = dissolvedOxygen ~ year4 + daynum + temperature_C +
##     depth + tn_ug + tp_ug, data = DO.predictors)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1484 -0.7931 -0.1172  0.5387 10.1439
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   59.7811180 14.9185927   4.007 6.65e-05 ***
## year4         -0.0232059  0.0074958  -3.096 0.002023 **
## daynum        -0.0069474  0.0019362  -3.588 0.000351 ***
## temperature_C -0.1370994  0.0218550  -6.273 5.49e-10 ***
## depth         -0.8746669  0.0337079 -25.948  < 2e-16 ***
## tn_ug         -0.0013804  0.0002406  -5.738 1.31e-08 ***
## tp_ug          0.0076268  0.0026155   2.916 0.003634 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.743 on 900 degrees of freedom
## Multiple R-squared:  0.8131, Adjusted R-squared:  0.8118
## F-statistic: 652.6 on 6 and 900 DF,  p-value: < 2.2e-16
```

The final set of variables that best predict the dissolved oxygen across a monitoring period are year, day number, temperature, depth, nitrogen concentration and phosphorus concentration. This set of variables explain 81% of variance (Adjusted R-squared = 0.8118, p-value < 0.001)

The following can be said about Figure 8:

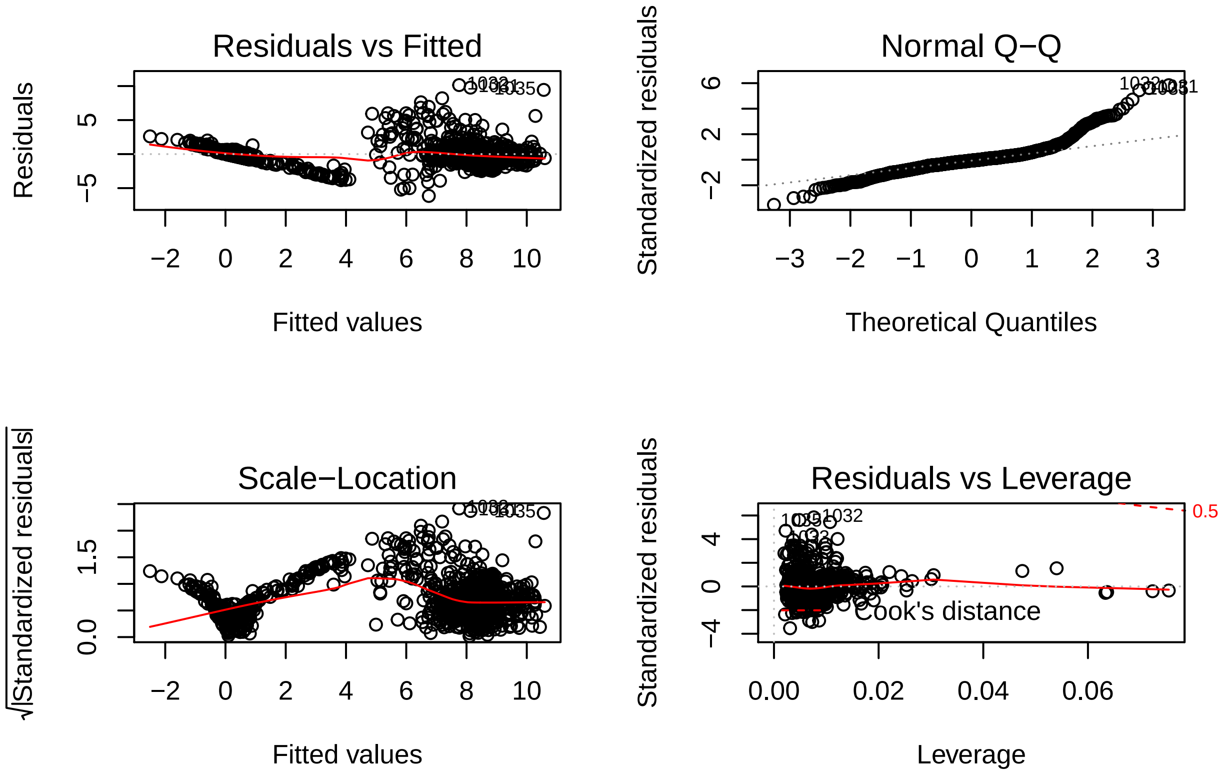


Figure 8: Model diagnostic plots.

1. Residuals vs Fitted: Used to check the linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship; which is the case here.
2. Normal Q-Q: Used to examine whether the residuals are normally distributed. It's good if residuals points follow the straight dashed line. In this case the points do appear to follow an straight line (except for the extremes which is common for environmental data), so the graph is accepted.
3. Scale-Location (or Spread-Location): Used to check the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity. In this case, the plot is difficult to predict so we can rely on the residual vs fitted data to assure homoscedasticity.
4. Residuals vs Leverage: Used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis. In this case the outliers are not so influential that leaving them out might change the structure of the model.

Given that our assumptions have been met, we can move on to the next step. From the model we got that the intercept is 59.7811180, which is the dissolved oxygen in mg/L, when all the other variables are zero. Therefore, we reject the null hypothesis of no effect of the explanatory variables on the response variable.

The year decreases the dissolved oxygen (mg/l) in the lakes significantly. With an increase of 1 year, the dissolved oxygen is decreased by 0.0232059 mg/l ($t = -3.096$, $df = 900$, $p < 0.05$).

The day number decreases the dissolved oxygen (mg/l) in the lakes highly significantly. With an increase of 1 day, the dissolved oxygen is decreased by 0.0069474 mg/l ($t = -3.588$, $df = 900$, $p < 0.001$).

The temperature decreases the dissolved oxygen (mg/l) in the lakes highly significantly. With an increase in temperature by 1 Celsius, the dissolved oxygen is decreased by 0.1370994 mg/l ($t = -6.273$, $df = 900$, $p < 0.001$).

The depth decreases the dissolved oxygen (mg/l) in the lakes highly significantly. With an increase in depth by 1 meter, the dissolved oxygen is decreased by 0.8746669 mg/l ($t = -25.948$, $df = 900$, $p < 0.001$).

The nitrogen concentration decreases the dissolved oxygen (mg/l) in the lakes highly significantly. With an increase in nitrogen concentration by 1 µg/l, the dissolved oxygen is decreased by 0.0013804 mg/l ($t = -5.738$, $df = 900$, $p < 0.001$).

The phosphorus concentration increases the dissolved oxygen (mg/l) in the lakes significantly. With an increase in phosphorus concentration by 1 µg/l, the dissolved oxygen is increased by 0.0076268 mg/l ($t = 2.916$, $df = 900$, $p < 0.05$).

The final linear equation to predict dissolved oxygen (mg/l) from the explanatory variables is:

Dissolved oxygen in mg/l = 59.7811180 - 0.0232059xyear number - 0.0069474xday number - 0.1370994xtemperature in C - 0.8746669xdepth in m - 0.0013804xTN

in µg/l + 0.0076268xTP in µg/l

Figure 9 - Figure 14 offer visual presentations of the dissolve oxygen against each model variable.

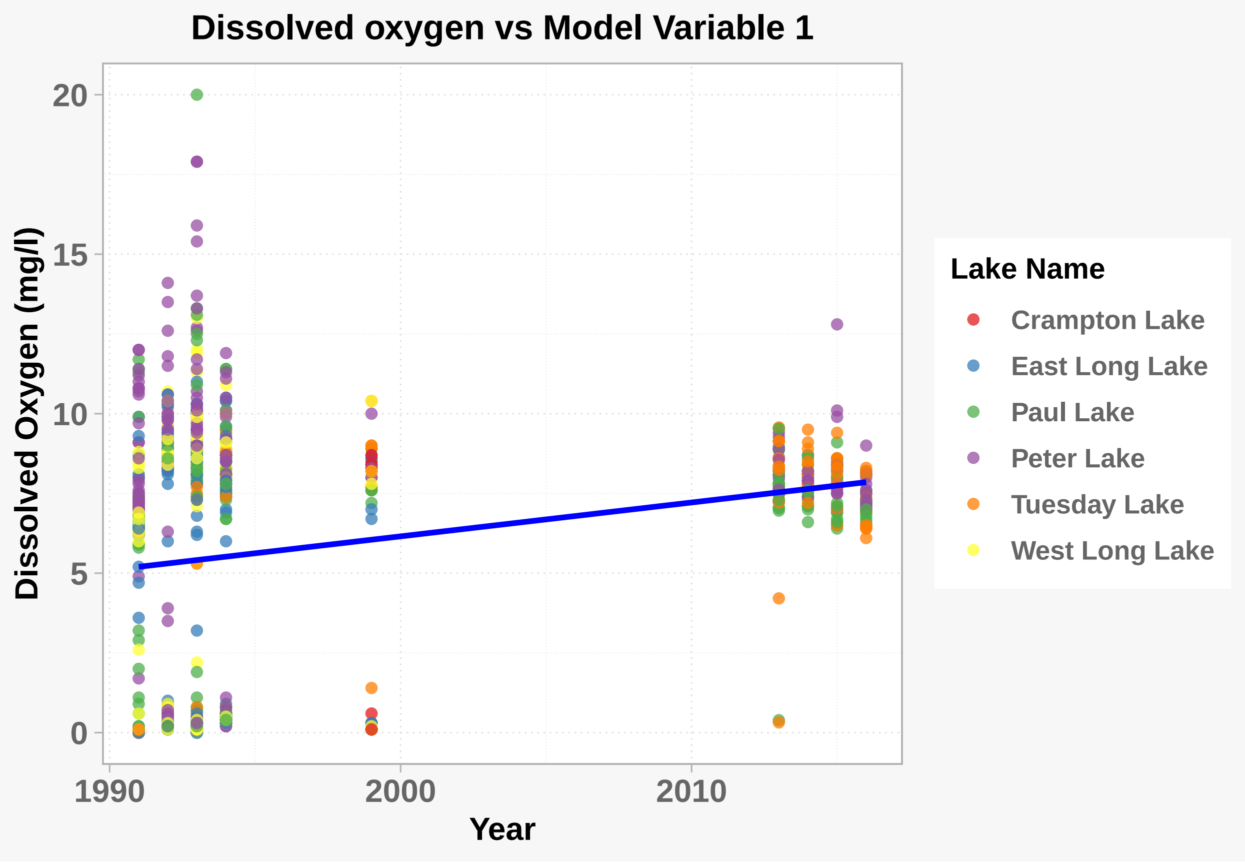


Figure 9: Dissolved oxygen vs Years.

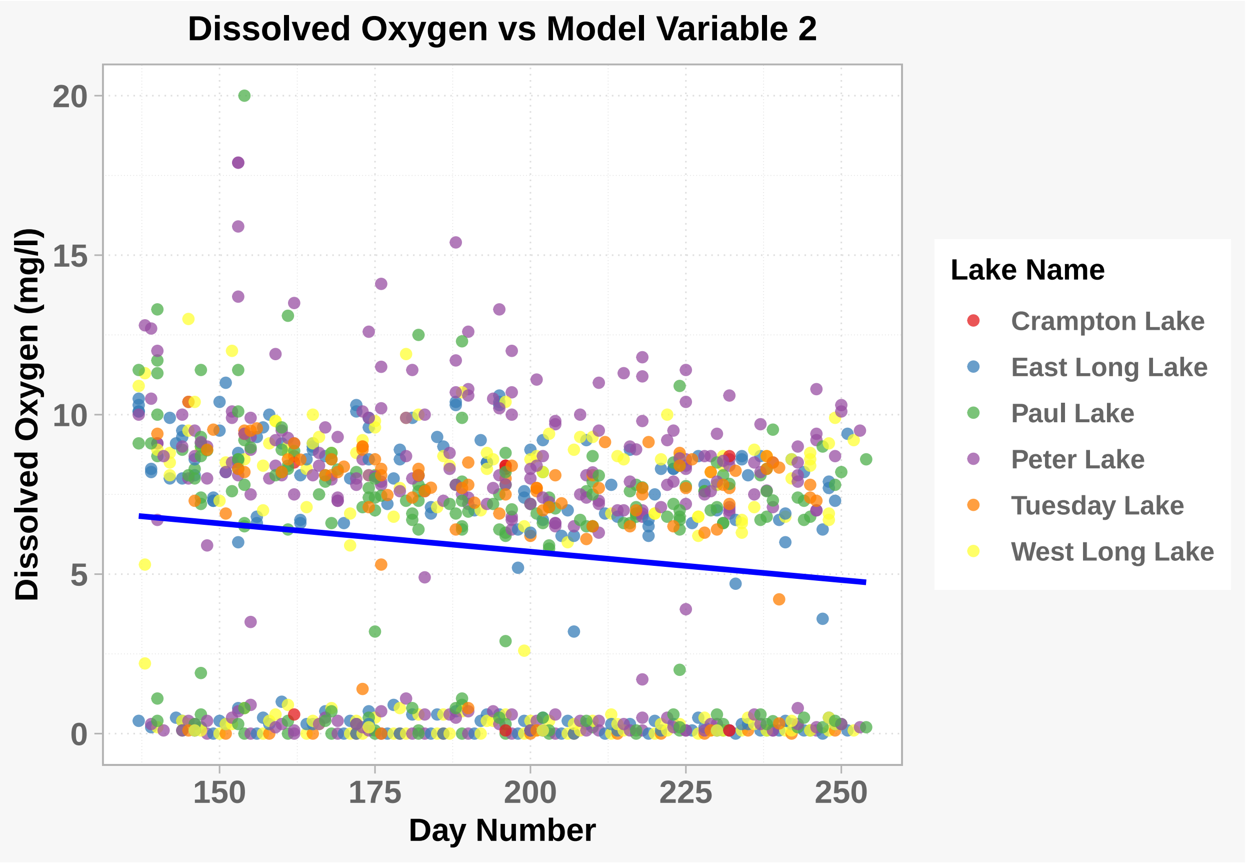


Figure 10: Dissolved oxygen vs day number.

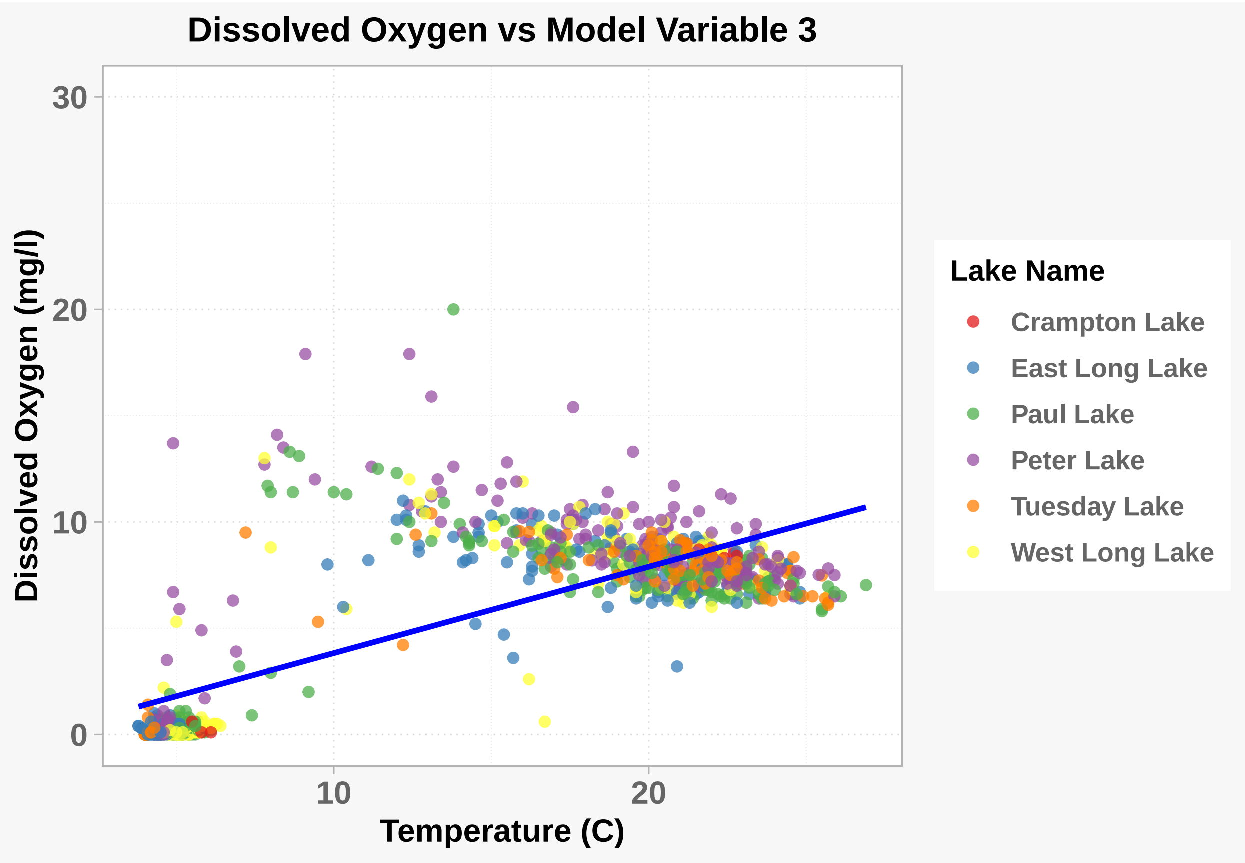


Figure 11: Dissolved oxygen vs temperature.

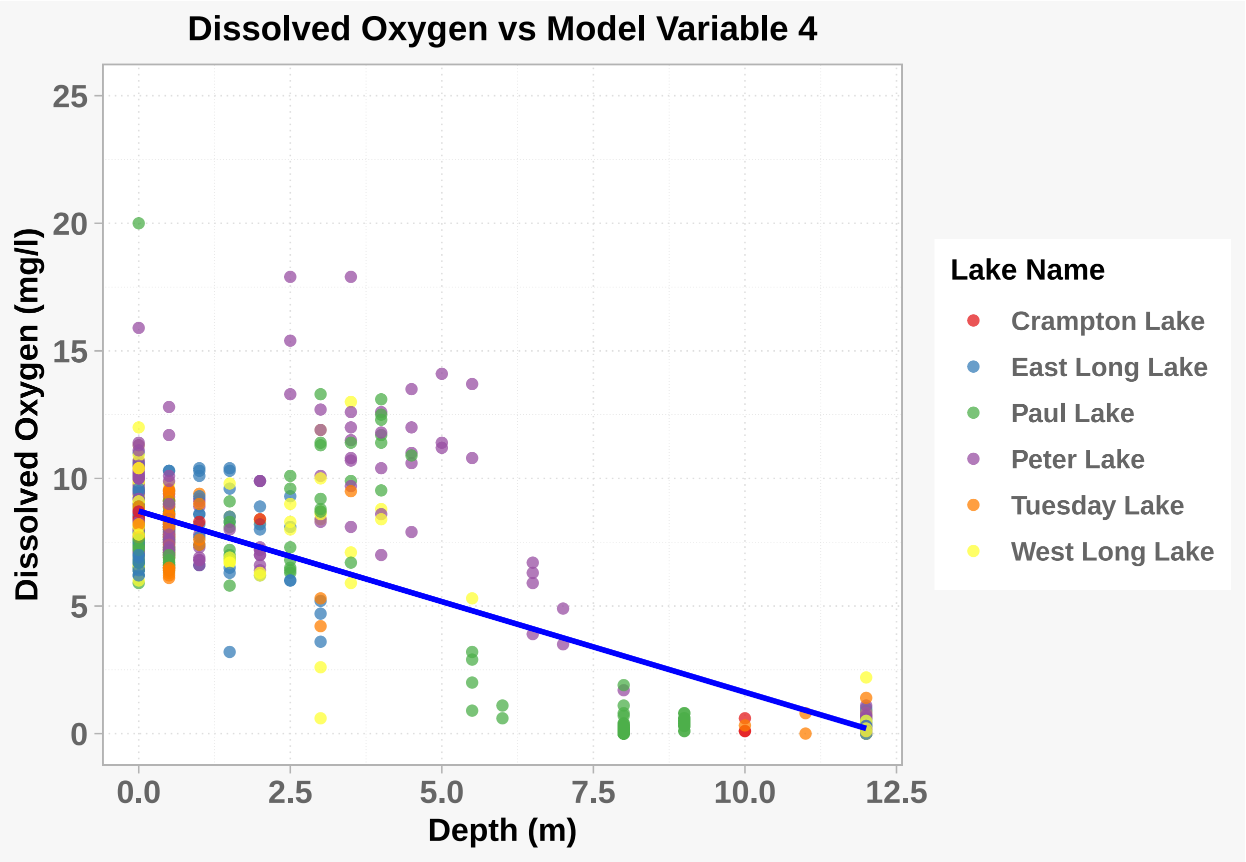


Figure 12: Dissolved oxygen vs depth.

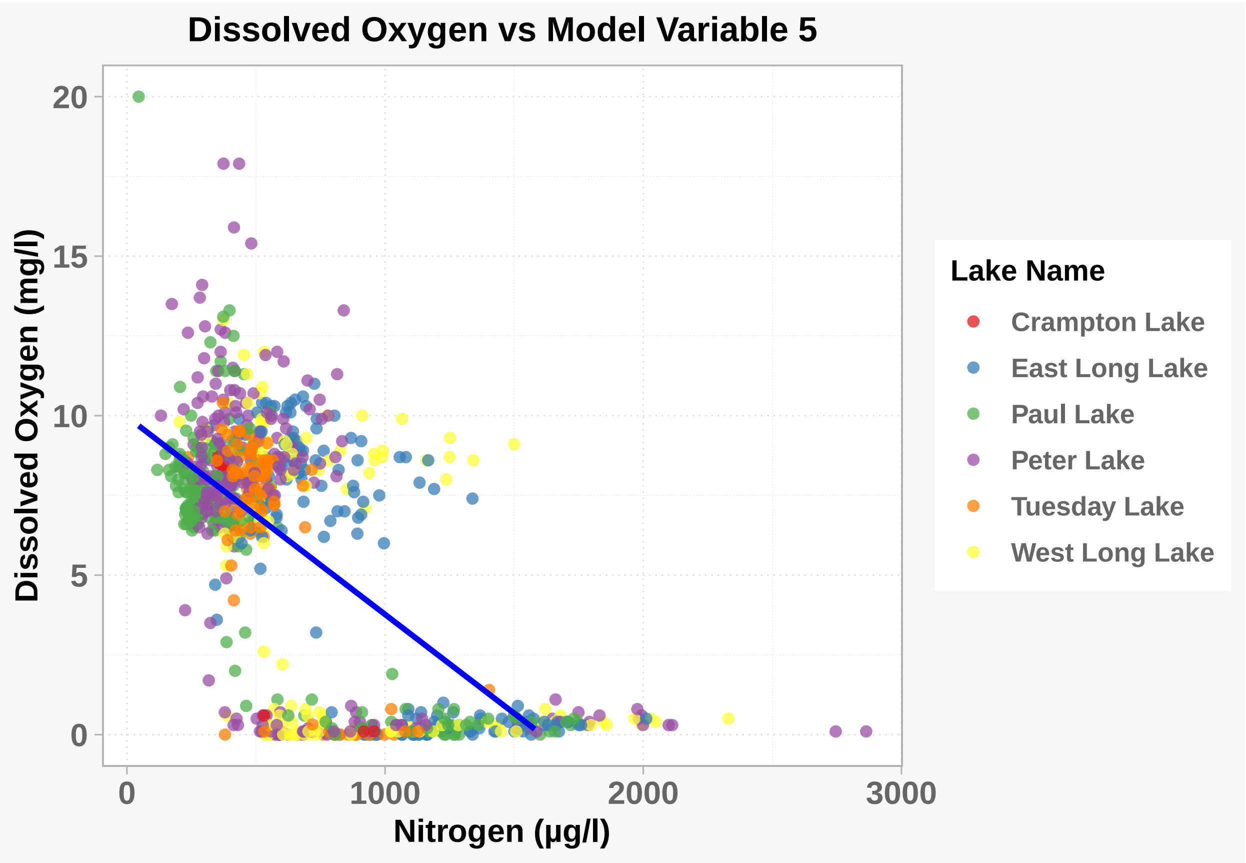


Figure 13: Dissolved oxygen vs nitrogen concentration.

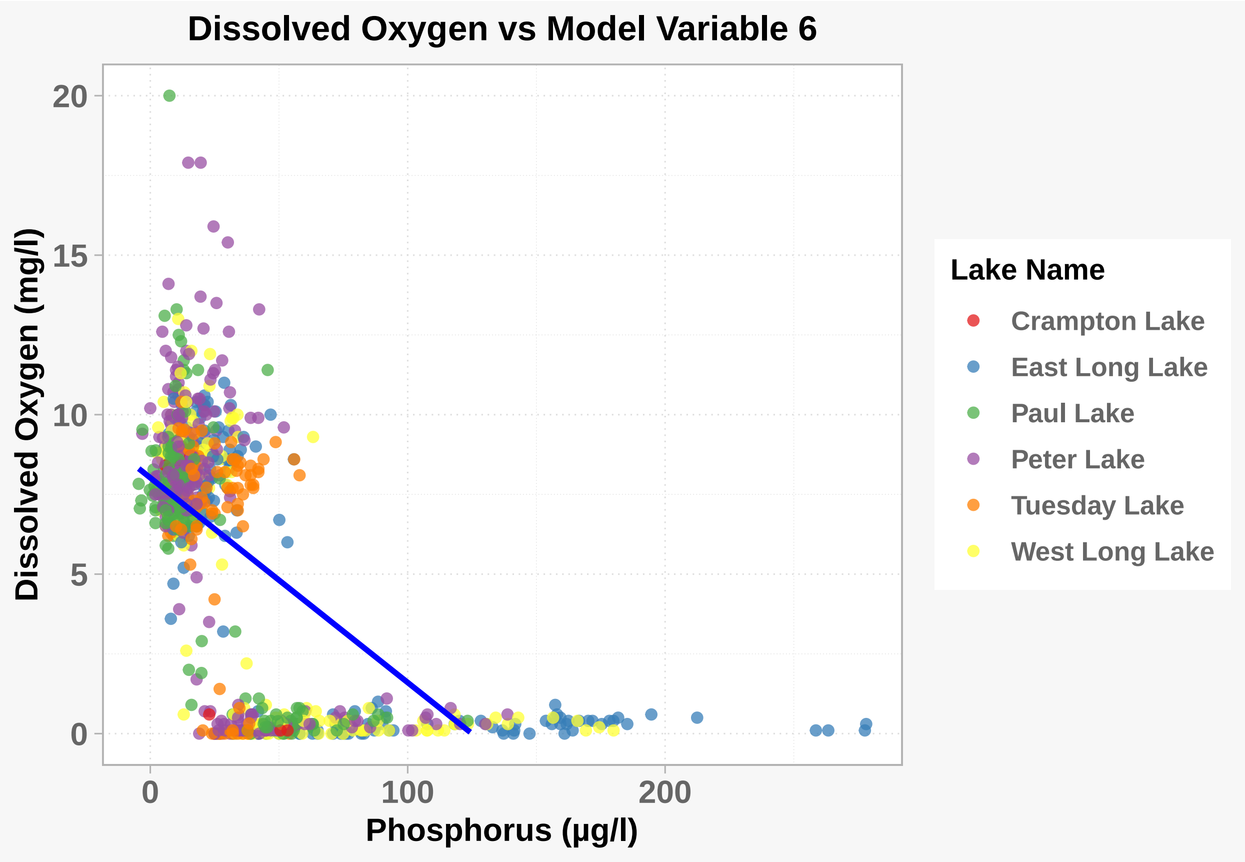


Figure 14: Dissolved oxygen vs phosphorus concentration.

5 Summary and Conclusions

I began this project with a set of exploratory questions that helped arrive at the necessary data analysis methods to eventually answer the the two research questions. Below are the key findings of this data analysis project stemming from key questions asked during this entire process (not just exclusive to the two main research questions or their subsets):

1. Majority of the lakes were found to be above the line of Redfield Ratio, therefore confirming that all the lakes in the dataset used were phosphorus deficient. This means that phosphorus can controlled to avoid eutrophication.

2. Most of the lakes were found to have a mean N:P ratio much above the Redfield Ratio. However, there have been instances of almost all of them having touched the Redfield Ratio.

3. The N:P ratio is found to be somewhat positively correlated to the dissolved oxygen in the lakes. However, the scatterplots show that that there is a huge spread of dissolved oxygen at especially below the ratio 50. This could mean that there are other stronger factors effecting dissolved oxygen

4. The the phosphorus concentration has decreased slowly but consistently over the years, possibly hinting at successful water quality management. Last decade has seen the phosphorus concentration controlled, below the set state criterion.

Other than the above, a major finding is that year number (possibly depicting environmental changes), day number (depicting seasonal changes), temperature, depth, nitrogen concentration and phosphorus concentration together are a good predictors of dissolved oxygen. The multiple regression model highlights several trends: increase in the year number, day number, temperature, depth and nitrogen concentration decreases dissolved oxygen significantly while increase in phosphorus concentration increases dissolved oxygen significantly. However, not all the plots (that include their respective regression line) are consistent with models results. For example, the Lm() line in the dissolved oxygen over temperature graph shows that there is a positive correlation between the tow when it is supposed to the other way around. This is not true in the real world since cold water can hold more dissolved oxygen. However, is it possible that depth has had a more major effect on this trend in this graph. Similarly, even though the model predicts that increase in phosphorus should increase dissolved oxygen, the graph shows the opposite. Once again, it is probably because there are interacting and mixed effects involved. Therefore, there is more scope for further statistical testing and analysis to confirm some of these correlations.