The Quality of Ontario Lakes

How has the water quality of Ontario's inland lakes changed from 2015 to 2022, as measured by total phosphorus and water clarity? What does this suggest about the productivity of Ontario's inland lakes?

Joshua Antonio Crisologo & Jonathan Manuel

Introduction

This report is based on two datasets sourced from the Ontario Lake Partner Program (LPP) via the Ontario Data Catalog. The LPP conducts annual assessments of water quality in inland lakes throughout Ontario, with data collected by volunteers following standardized provincial protocols. The datasets cover total phosphorus (µg / L) and water clarity measured by seechi depth (m) for numerous inland lakes in the Precambrian Shield region. Each dataset includes geospatial information, site descriptions, collection dates, and metrics pertaining to the water quality. The data was last validated on January 17, 2024 and is updated yearly. Both datasets were last updated on December 31, 2022.

Table 1: Dataset Description

Variable	Types	Description
Latitude	integer	The latitude of the lake in DMS
Longitude	character	The longitude of the lake in DMS
Site.ID	character	The site ID of the sampling point
Township	character	The township the lake
Lake.Name	numeric	The name of the lake
Site.Description	character	The description of the sampling point
avg_phos_ug_l	Date	The average total phosphorus (ug / L)
phos_is_outlier	numeric	Whether total phosphorus is an outlier
phos_date	Date	The date that the phosphorus sample was collected
secchi_depth_m	numeric	The depth at which the seechi disk can no longer be distinguished
trans_date	numeric	The date that the seechi disk was inserted into the lake

Purpose

This report aims to address the research question:

How has the water quality of Ontario's inland lakes changed from 2015 to 2022, as measured by total phosphorus and water clarity? What does this suggest about the productivity of Ontario's inland lakes? What about lakes around the world?

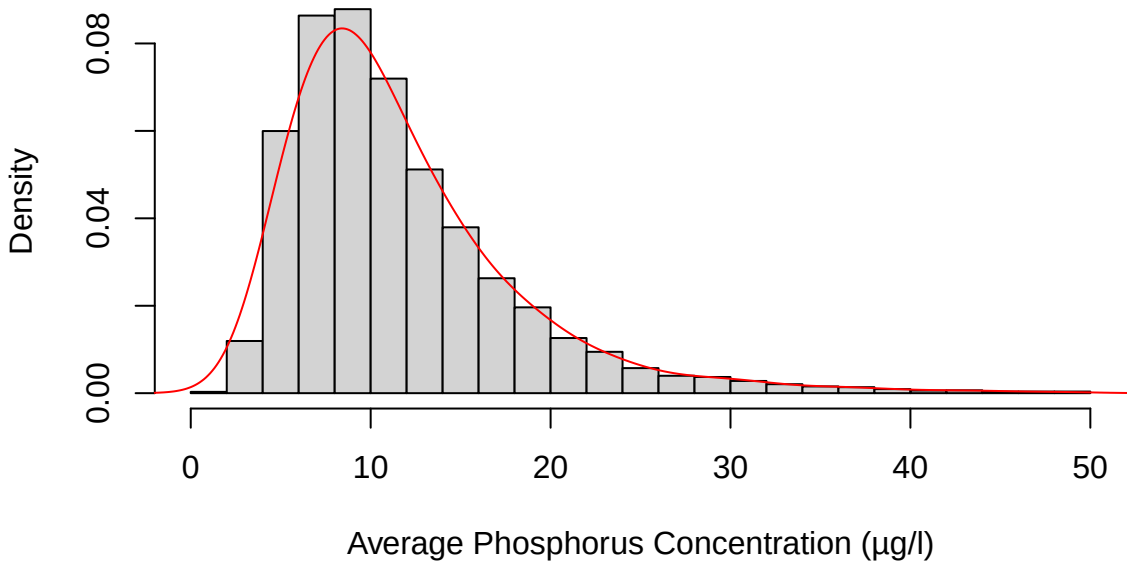
By answering these questions, this report ultimately aims to provide insight on where the government of Ontario can best address its conservation efforts, understand the trend of productivity of Ontario's inland lakes, and use the results of Ontario's lakes to predict the productivity of lakes around the world.

Background

Water quality, as measured by phosphorus levels and secchi depth, has a major influence on the biodiversity of inland lakes and freshwater streams.

In the vast majority of Ontario's inland lakes, phosphorus is the element that controls the growth of algae. As such total phosphorus concentrations (µg/L) are most aptly used to assess lake nutrient status. Limnologists place lakes into three categories based on their total phosphorus concentrations: oligotrophic (less than 10 µg/L), mesotrophic (10-20 µg/L TP), and eutrophic (over 20 µg/L). Oligotrophic lakes are low in nutrients and rarely have algal blooms - the least productive of the three levels. Eutrophic lakes, on the other hand, have high nutrient levels and often suffer from persistent algal blooms - it is most productive of the three trophic levels. In the middle in terms of productivity, Mesotrophic lakes vary in characteristics and may experience moderate blooms. The distribution of average phosphorus concentration in Ontario's inland lakes after filtering out extreme outliers is shown below:

Figure 1: Distribution of Average Phosphorus Concentration



From Figure 1, it can be deduced that the bulk of Ontario's inland lakes possess oligotrophic values suggesting that many provincial lakes lack nutrients and rarely have algal blooms as a result. Thus eluding to the lack of productivity of Ontario's inland lakes.

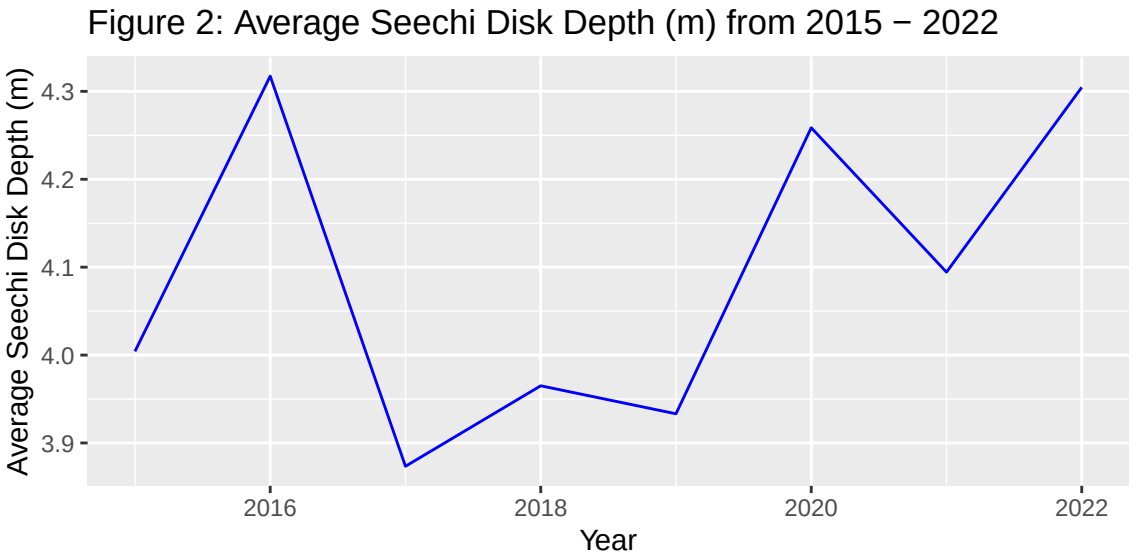
Another metric to classify lakes are secchi disks. These tools are black and white disks used to ascertain water clarity by lowering it into the water and measuring the point at which black and white can no longer be distinguished. Readings of over 5 meters indicate oligotrophic conditions, depths of 3.0 to 4.9 meters suggest mesotrophic conditions and readings less than 2.9 meters signify eutrophic lakes, with higher nutrient levels.

Below Table 2 shows the average point at which the black and white coloring of the Seechi disk could no longer be distinguished for each lake classification.

Table 2: Average Secchi Disk Depth (m) by Lake Classification

Year	Eutrophic	Mesotrophic	Oligotrophic
2015	2.027140	3.803148	6.072994
2016	2.118546	3.833635	6.171241
2017	2.070776	3.813988	6.168022
2018	2.070520	3.817491	6.039375
2019	2.069596	3.815862	6.142789
2020	2.007616	3.917175	6.007801
2021	2.084981	3.865085	6.187396
2022	2.079520	3.848938	6.488841

To gain a better understanding of how the seechi disk depth has changed 2015 to 2022, the average seechi disk depth will be calculated and plotted. Figure 2 is the result of this process.



In an effort to ascertain the proportion of inland lakes whose secchi depth readings indicate eutrophic conditions (less than or equal to 2.9 metres), a one-sample proportion test will be conducted with a confidence interval of 95% and a sample size of 1000. The 25th and 75th percentiles of these proportions will be calculated using bootstrapping to gain a more holistic view of the proportion's variability. Our null hypothesis will be that our mean is at 0.5, while our alternative hypothesis is that our true mean is different from this ideal average (not equal to 0.5). To conduct our test of hypothesis, we use bootstrapping.

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum(sample$secchi_depth_m <= 2.9) out of 1000, null probability 0.5
## X-squared = 241.08, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.2275164 0.2824012
## sample estimates:
##     p
## 0.254
```

From this, it can be determined that the proportion of inland lakes that exhibit eutrophic conditions is estimated to be ~0.25 - 0.28, with a 95% confidence interval ranging from ~0.22- 0.32.

Since the p-value $<$ 2.2e-16, there is strong evidence against the null hypothesis. This suggests that the proportion of lakes that exhibit eutrophic conditions is significantly different than the null proportion of 0.5.

Table 3: Quantile Results of Eutrophic Lakes According to Secchi Depth

Quantile	Value
25%	0.244
75%	0.262

We used bootstrapping to find the quantile values of eutrophic lakes according to secchi depth. The 25th percentile and 75th percentile of the bootstrapped proportions of eutrophic inland lakes are estimated to be

~0.24 to ~0.27, respectively. This suggests 95% of the time, in an area with similar conditions to Ontario, around ~24-27% of lakes are eutrophic. It can then be concluded that roughly a quarter of Ontario's inland lakes have high nutrient levels, often suffer from persistent algal blooms, and are thus extremely productive.

Trophic State Index (TSI)

Another, widely used, way of determining whether a lake is Oligotrophic, Mesotrophic, or Eutrophic is through the Trophic State Index (TSI). Developed by Carson in 1977, the TSI of a lake can be determined with the average of the Trophic State Index in terms of the phosphorus level (µg/L), labelled TSI(TP), and the Trophic State Index in terms of secchi depth, labelled TSI(SD). The formula for TSI(TP) is derived as 14.42ln(TP) + 4.15, where TP represents the total phosphorus level (the average phosphourus level in µg/L). The formula for TSI(SD) is derived as 60 - 14.41ln(SD), where SD is the secchi depth in metres. Thus, the total TSI is (TSI(SD)+TSI(TP)) / 2.

If a lake's TSI is < 30, it is considered Oligotrophic. If it's TSI is from 30-40, it is Mesotrophic, while any TSI level above 40 is considered Eutrophic. The TSI is a clean and an easily identifiable method of classifying a lake's trophic level. However, many lake databases, including Ontario's, do not calculate a lake's TSI level and instead list the lake's respective phosphorus and secchi depth values. In this case, it can be confusing and strenuous to identify a lake's trophic status. It can be very convenient for researchers to know a lake's trophic status at an instant.

Thus, for the purpose of analyzing the productivity level of lakes in Ontario, we will create a model that will analyze a lakes average phosphorus level and secchi depth and instantly determine it's trophic level, without the need for calculating the TSI. The model will consist of a Random Forest model, as well as cross validation. This way, anyone can identify a lake's trophic status, regardless of the dataset. Classifying our lake's trophic level will also help us analyze the productivity of lakes across Ontario when plotted on a map.

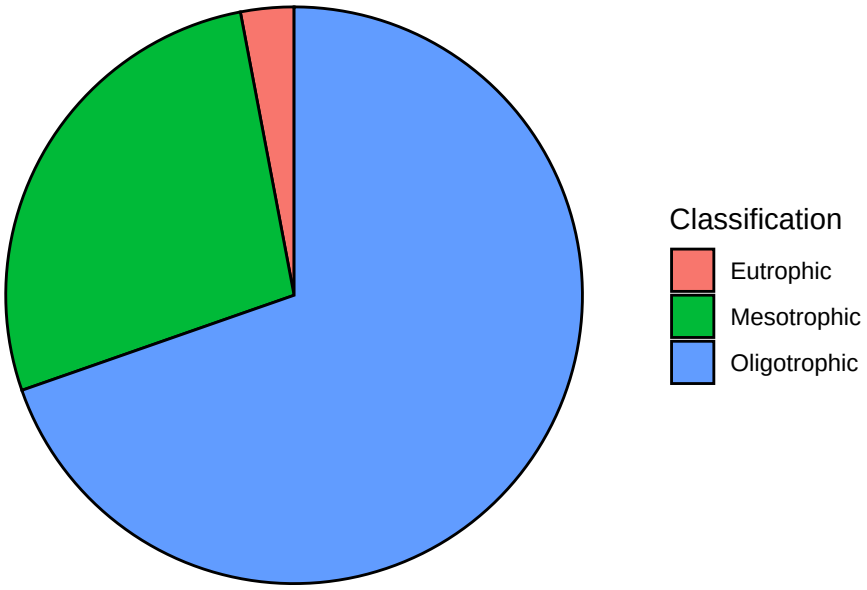
To start, we use our dataset filtered to the most recent date of testing. We do this by filtering by the most recent date of phosphorus testing, followed by the most recent date of secchi depth testing. Then, we classify each lake's trophic status by calculating the actual TSI of the lake using the formulas specified above - (TSI(SD)+TSI(TP)) / 2. The proportion of each lake's trophic status as classified by TSI (as of the most recent date of testing) are as follows:

Table 4: TSI Classification for Filtered Data (as of most recent test date)

Year	Eutrophic	Mesotrophic	Oligotrophic
2015	7	19	28
2016	2	10	34
2017	2	29	51
2018	21	90	69
2019	25	140	139
2020	1	9	26
2021	4	15	47
2022	12	110	280

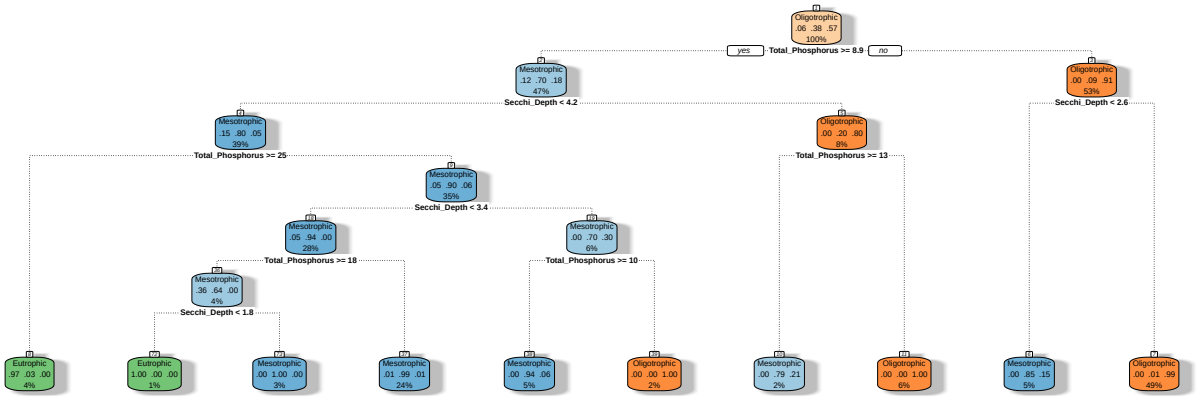
The most recent dates of testing are of the year 2022. Figure 3 shows the proportion of each lake's trophic level as of the most recent date of testing, 2022. Classifying by TSI, considers both phosphorus level and secchi depth (two variables), which creates different results from classification with only one variable. It can be interpreted that in 2022, the vast majority of Ontario's lakes are oligotrophic, with little being eutrophic.

Figure 3: TSI Classification for 2022



After formally classifying each lake using the TSI, we can then train a model to correctly classify each lake without the use of TSI. Creating a predictive model will assist in easily detecting a lake's trophic level and productivity - no matter the data set. Our first step is cross validation - we split our data set as 60% training data, and 40% testing data. To train our model, we will use a Random Forest approach; using bagging (bootstrapped samples of our training data) to create decision trees and creating a prediction based on each tree's features. Each tree analyzes each lake's trophic level, related to their phosphorus level and secchi depth. Such a decision tree is presented below.

Figure 4: Decision Tree for Lake Trophic Level according to phosphorus level and secchi depth



Rattle 2024−Apr−05 15:02:36 rstudio

Decision trees are important in assisting our model in deciding what determining factors and interactions needed to obtain a lake's trophic status. The node at the top of the tree is the head node, which asks a question. Each following branching node contains the % of data that logically answers this question (True

or False). The leaf nodes (nodes at the very bottom) contain the final predictions of data based on the interactions displayed in the tree. Analyzing the decision tree, we can see that a lake's trophic level largely depends on the range of its total phosphorus and secchi depth values.

A Random Forest model generates a large amount of these decision trees, learns and adapts to the interactions and features behind our training set, and uses them in predicting the results of our test set. We generate our Random Forest model based on how our lake's trophic status relates to its average phosphorus level and secchi depth. After creating our Random Forest, we can determine the importance of each variable with respect to predicting our values through a variable importance plot.

Figure 5: Random Forest Model Based on Lake Trophic Status

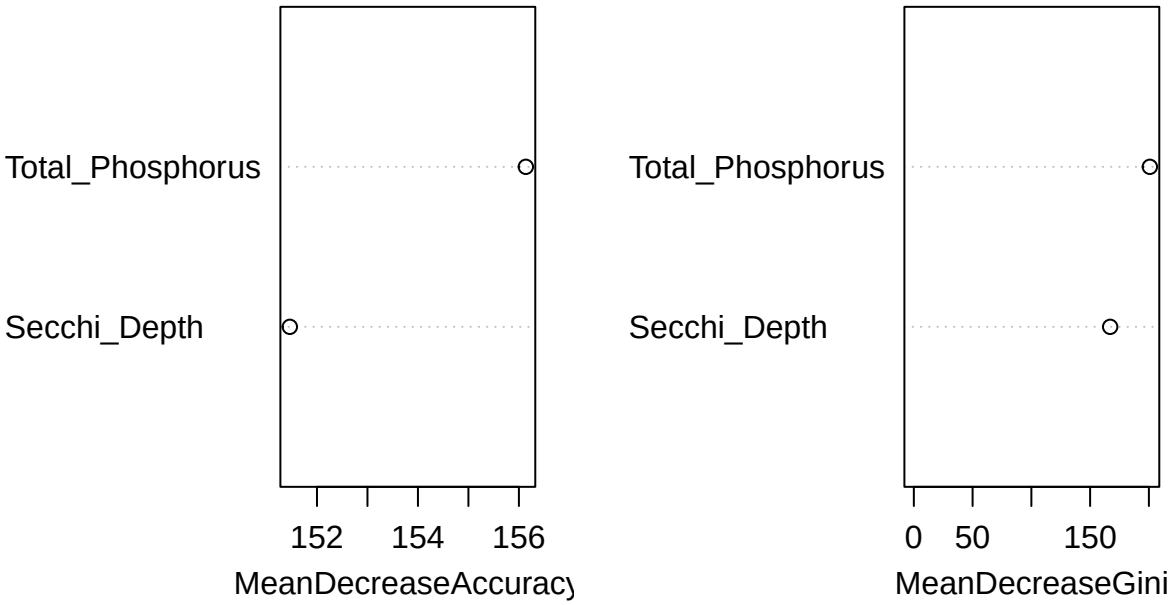


Figure 5 indicates our average phosphorus level (Total Phosphorus) is the most important variable in our prediction, as our accuracy would decrease significantly if we removed this variable from the equation. Thus, using our Random Forest model, we can predict the trophic status of each lake in our test data set simply with the average phosphorus level and secchi depth alone - no need for the complicated calculations of the TSI. We can display the results of our prediction by creating a confusion matrix and matching our model's prediction to the status classified by the TSI.

Table 5: Accuracy of Random Forest Model (as represented by x)

x
0.9743041

Table 5 displays that our model accuracy fluctuates between ~95-98% accuracy in a dataset containing over 1000 lakes, which is extremely accurate. Through the use of cross-validation and Random Forest, we have successfully created a model that can predict a lake's trophic status using it's average phosphorus level and secchi depth. We know that this model works for any dataset with similar parameters due to our use of cross-validation and a train/test model.

Final Summary

Our report has utilized a variety of statistical analysis tools, such as bootstrapping, test of hypothesis, confidence intervals, cross validation, and decision tree/Random Forest models to analyze how the water quality and productivity of Ontario's inland lakes has changed from 2015-2022, as well as the current state of each lake as of their most recent date of testing. Additionally, we have visualized these trends and proportions through histograms, line graphs, pie charts, and tables.

From classfying each lake's trophic status using solely their phosphorus level or their secchi depth, or by using their TSI value, we can conclude that the majority of lake's in Toronto are Oligotrophic. Additionally, we know that overtime, the average secchi depth has increased since 2015 - hinting that Ontario's lakes are actually getting clearer as of recent. From our bootstrapped proportion test and quantile/confidence interval calculation, we can infer that, when measured solely on secchi depth, eutrophic lakes make up around 25% of Ontario's lakes each year. This rejected our hypothesis that they would make up around 50% of the lakes in Ontario. This means that the majority of lakes in Ontario are not optimally productive, by the definition of oligotrophic.

On the other hand, we then classified each lake using their Trophic State Index value and found that the vast majority of lakes are still oligotrohpic, however, the amount of eutrophic lakes decrease compared to classifying based on secchi depth/phosphorus alone. We used cross validation and decision trees to create a Random Forest model, with the purpose of predicting the TSI of a lake based on its phosphorus level and secchi depth alone. With this model, we can easily predict a lake's TSI and its resulting trophic status with any dataset with ~95-98% accuracy, given the phosphorus level and secchi depth.

Since Ontario's dataset for the water quality and phosphorus level of lakes is constantly updating yearly, our model can identify each lake's trophic level at an instant. It is imperative to consistently analyze the quality and productivity of lakes in an area as large as Ontario, as they are a clear indicator as to how the environment around is faring today.

References

DeSellas, Anna. Metadata For: Lake Partner Program Title Ontario Lake Partner Program Alternative Title Status Cited Responsible Parties. 1 Dec. 2023.

Guide to Interpreting Total Phosphorus and Secchi Depth Data from the Lake Partner Program. Lake Partner Ontario, 2013.

"North American Lake Management Society (NALMS)." North American Lake Management Society (NALMS), www.nalms.org/secchidipin/monitoring-methods/trophic-state-equations/.

"Ontario Lake Partner - Ontario Data Catalogue." Data.ontario.ca, 17 Jan. 2024, data.ontario.ca/dataset/ontario-lake-partner.

Prasad, A. Carlson's Trophic State Index for the Assessment of Trophic Status of Two Lakes in Mandya District. 2012.

"Secchi Readings – Measuring Our Water Clarity – Halls & Hawk Lakes Property Owners Association (HHLPOA)." Halls and Hawk Lakes, 19 Mar. 2011, hallshawklakes.ca/featured/secchi-readings/#:~:text=Secchi%20Reading%20and%20Lake%20Nutrient. Accessed 4 Apr. 2024.

Appendix

Inital Data Preprocessing

```
# loading the data

trans <- read.csv("../data/transparency.csv")
phosphorus <- read.csv("../data/phosphorus.csv")

# The common variables between data sets
sims = c("Latitude", "Longitude", "Site.ID",
         "Township","Lake.Name", "Site.Description")

# Selected date-range is 2015-2022
# Dropped Sample 1 & Sample 2 in favor of "Average.Total.Phosphorus..µg.L."

phosphorus <- phosphorus %>%
  mutate(date = as.Date(Date..DD.MMM.YY., format = "%d-%b-%y")) %>%
  select(-Date..DD.MMM.YY., -X_id, -Total.Phosphorus.sample.1..µg.L.,
         -Total.Phosphorus.sample.2..µg.L., -Data.Collector, -STN) %>%
  filter(2015 <= year(date) & year(date) <= 2022) %>%
  rename(Latitude = Latitude..DMS., Longitude = Long..DMS.,
         avg_phos_ug_l = Average.Total.Phosphorus..µg.L.,
         phos_date = date,
         phos_is_outlier = Possible.outlier) %>%
  arrange(Latitude)

trans <- trans %>%
  mutate(date = as.Date(Date..DD.MMM.YY., format = "%d-%b-%y")) %>%
  select(-Date..DD.MMM.YY., -X_id, -STN) %>%
  filter(2015 <= year(date) & year(date) <= 2022) %>%
  rename(Latitude = Latitude..DMS., Longitude = Longitude..DMS.,
         secchi_depth_m = Secchi.Depth..m.,
         trans_date=date,
         Township=TOWNSHIP) %>%
  arrange(Latitude)

# Merge the two dataframes together
df <- merge(phosphorus, trans, by=sims)

# Update latitude and longitude values
data <- df %>% group_by(Lake.Name) %>%
        mutate(char_lat = as.character(Latitude)) %>%
        mutate(lat.deci = as.numeric(substr(char_lat, start = 1, stop = 2)) +
                          as.numeric(substr(char_lat, start = 3, stop=nchar(char_lat)))
                          * (0.1 ^ (nchar(char_lat) - 2))) %>%
        mutate(char_long = as.character(Longitude)) %>%
        mutate(long.deci = as.numeric(substr(char_long, start = 1, stop = 2)) +
                          as.numeric(substr(char_long, start = 3, stop=nchar(char_long)))
                          * (0.1 ^ (nchar(char_long) - 2))) %>%
        select(-char_lat, -Latitude, -char_long, -Longitude) %>%
        rename(Latitude = lat.deci, Longitude = long.deci)
```

Table 1: Dataset Descripton

```
variables <- c("Latitude", "Longitude", "Site.ID", "Township", "Lake.Name",
               "Site.Description", "avg_phos_ug_l", "phos_is_outlier",
               "phos_date", "secchi_depth_m", "trans_date")

types <- sapply(data, class)
descriptions <- c("The latitude of the lake in DMS",
                  "The longitude of the lake in DMS",
                  "The site ID of the sampling point",
                  "The township the lake",
                  "The name of the lake",
                  "The description of the sampling point",
                  "The average total phosphorus (ug / L)",
                  "Whether total phosphorus is an outlier",
                  "The date that the phosphorus sample was collected",
                  "The depth at which the seechi disk can no longer be distinguished",
                  "The date that the seechi disk was inserted into the lake")
summary_df <- data.frame(Variable = variables, Types = types, Description = descriptions)
kable(summary_df, row.names = FALSE, caption="Dataset Description")
```

Figure 1: Distribution of Average Phosphorus Concentration

```
no_phos_outliers <- data %>% filter(avg_phos_ug_l <= 50)
hist(no_phos_outliers$avg_phos_ug_l, freq = FALSE,
     main = "Figure 2: Distribution of Average Phosphorus Concentration",
     xlab = "Average Phosphorus Concentration (µg/l)",
     ylab = "Density")

# Add density line
lines(density(no_phos_outliers$avg_phos_ug_l, adjust = 5), col = "red")
```

Table 2: Average Secchi Disk Depth (m) by Lake Classification

```
lake_classification <- data %>%
  mutate(class = case_when(secchi_depth_m >= 5 ~ "Oligotrophic",
                           secchi_depth_m >= 3 ~ "Mesotrophic",
                           TRUE ~ "Eutrophic")) %>%
  mutate(year = year(trans_date)) %>%
  group_by(year, class) %>%
  rename(Year = year, Classification = class) %>%
  summarise(mean_secchi = mean(secchi_depth_m)) %>%
  pivot_wider(id_cols = Year,
              names_from = Classification,
              values_from = mean_secchi) %>%
  arrange(Year)

kable(lake_classification, caption="Average Secchi Disk Depth (m) by Lake Classification")
```

Figure 2: Average Seechi Disk Depth (m) from 2015 - 2022

```
seechi_by_year <- trans %>% mutate(year = year(trans_date)) %>%
  group_by(year) %>%
```

```
  mutate(avg_seechi = mean(secchi_depth_m)) %>%
  select(year, avg_seechi) %>%
  arrange(year)

seechi_by_year <- unique(seechi_by_year)

fig3 <- ggplot(seechi_by_year, aes(x=year, y=avg_seechi)) +
        geom_line(col="blue") +
        labs(y="Average Seechi Disk Depth (m)", x="Year",
             title= "Figure 3: Average Seechi Disk Depth (m) from 2015 - 2022")
```

Hypothesis Testing and Confidence Interval (1-sample proportion test)

```
sample <- trans %>% select(trans_date, secchi_depth_m) %>%
                    sample_n(1000, replace = FALSE) %>%
                    drop_na()

prop_test_result <- prop.test(
  x = sum(sample$secchi_depth_m <= 2.9),  # Number of successes
  n = 1000,                               # Total number of trials
  conf.level = 0.95                       # Confidence level
)

prop_test_result
```

Table 3: Quantile Results of Eutrophic Lakes According to Secchi Depth

Bootstrapping and Confidence Interval

```
boot_function = function(){
  boot_data <-  sample %>% sample_n(nrow(sample), replace = T)
  boot_prop <-  mean(boot_data$secchi_depth_m <= 2.9)

  return(boot_prop)

}

quan <- quantile(replicate(1000, boot_function()), c(0.25, 0.75))
quantiles_df <- data.frame(Quantile = c("25%", "75%"),
                           Value = round(quan, digits = 5))

# To get rid of irremovable index column
quantiles_df <- data.frame(Quantile = quantiles_df$Quantile,
                           Value = quantiles_df$Value)
kable(quantiles_df, caption="Quantile Results of Eutrophic Lakes According to Secchi Depth")
```

Cross Validation

Filtering data for training and testing in preparation for Random Forest model

```
# filtering data and lakes to most recent date of test
 df = data %>%
      group_by(Lake.Name) %>%
```

```
      filter(phos_date==max(phos_date)) %>%
      filter(trans_date==max(trans_date))

# using formulas in regards to TSI to calculate the TSI of each lake
# We classify each lake's trophic status via TSI Index value
df = df %>% mutate(TSI_Depth = (60 - 14.41*log(secchi_depth_m))) %>%
  mutate(TSI_Phos = (14.42*log(avg_phos_ug_l) + 4.15)) %>%
  mutate(TSI = ((TSI_Depth + TSI_Phos) / 2)) %>%
  mutate(classification = case_when(TSI <= 40 ~ "Oligotrophic",
                                    TSI <= 50 ~ "Mesotrophic",
                                    TRUE ~ "Eutrophic")) %>%
  rename(Total_Phosphorus = avg_phos_ug_l, Secchi_Depth = secchi_depth_m)

# cross validation: splitting df into train and test of 60/40% for model testing
final = df
final = final %>% mutate(group_ind = sample(c("train", "test"),
                                            size=1,
                                            prob = c(0.6, 0.4),
                                            replace = T))

final_train = final %>% filter(group_ind == "train")
final_test = final %>% filter(group_ind == "test")
```

Table 4: TSI Classification for Filtered Data

```
TSI_classification = df %>% mutate(year = year(trans_date)) %>%
  group_by(year, classification) %>%
  rename(Year = year, Classification = classification) %>%
  summarise(mean_class = n()) %>%
  pivot_wider(id_cols = Year,
              names_from = Classification,
              values_from = mean_class) %>%
  arrange(Year)

kable(TSI_classification, caption="TSI Classification for Filtered Data
      (as of most recent test date)")
```

Figure 3: TSI Classification for 2022

```
# Filtering to most recent year
TSI_2022 <- filter(TSI_classification, Year == 2022)

# Preparing our df for the pie chart
TSI_2022_long <- TSI_2022 %>%
  pivot_longer(cols = -Year, names_to = "Classification", values_to = "mean_class")

# Creating our pie chart (ggplot does not provide a geom for this)
pie_chart <- ggplot(data = TSI_2022_long,
                    aes(x = "", y = mean_class, fill = Classification)) +
  geom_bar(stat = "identity", width = 1, color="black") +
  coord_polar("y", start = 0) +
  labs(title = "Figure 4: TSI Classification for 2022") +
  theme_void()
```

```
pie_chart
```

Figure 4: Decision Tree for Lake Trophic Level according to phosphorus level and secchi depth

```
# implementing decision tree analysis
# relating classification level to avg phos lvl and secchi depth
tree.m = rpart(classification ~ Total_Phosphorus + Secchi_Depth, data = final_train,
               method = "class")
fancyRpartPlot(tree.m, main="Figure 5: Decision Tree for Lake Trophic Level
               according to phosphorus level and secchi depth")
```

Figure 5: Random Forest Model Based on Lake Trophic Status

```
# another test: using randomForest with the same parameters to predict

rforest.m = randomForest(as.factor(classification) ~
                           Total_Phosphorus + Secchi_Depth,
                         data=final_train,
                         ntree=500, importance=TRUE)
varImpPlot(rforest.m, main="Figure 6: Random Forest Model
           Based on Lake Trophic Status")
```

Table 5: Accuracy of our Random Forest Model (as represented by x)

```
# prediction using random forest

final_test = final_test %>% ungroup(.) %>%
  mutate(rforest_predict = predict(rforest.m, newdata = final_test))

# create confusion matrix using random forest model and test accuracy
conmat = table(final_test$classification, final_test$rforest_predict)
kable(sum(diag(conmat))/sum(conmat),
      caption = "Accuracy of Random Forest Model (as represented by x)")
```