

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?

2024-03-30

Introduction

Sources to add while citing: <https://data.ontario.ca/dataset/ontario-lake-partner> & https://files.ontario.ca/moe_mapping/downloads/metadata/opendata/Lake_Partner_Program_metadata_EN2.pdf

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 (ug / L) and water clarity measured by secchi 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.

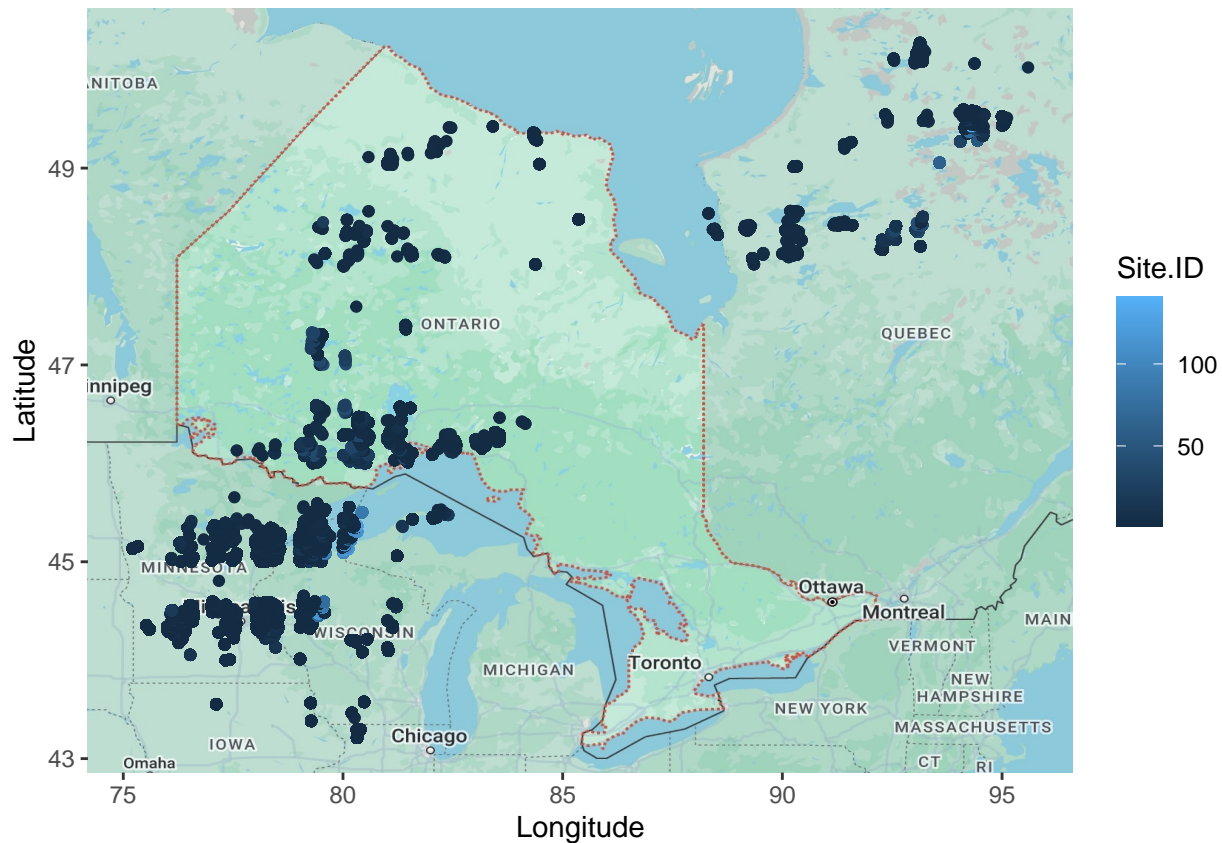
Dataset Description

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

```
library(ggpubr)
library(jpeg)
img=readJPEG("ontario.jpg")

ggplot(data, aes(y=Latitude, x=Longitude, col=Site.ID)) + background_image(img) +
  geom_point()
```

```
## Warning: Removed 208 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



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?

By answering this question, this report ultimately aims to provide insight on where the government of Ontario can best address its conservation efforts and understand the trend of productivity of Ontario's inland lakes.

Background

Source (<https://foca.on.ca/wp-content/uploads/2012/05/Guide-to-Interpreting-TP-and-Secchi-Data-Complete.pdf>)

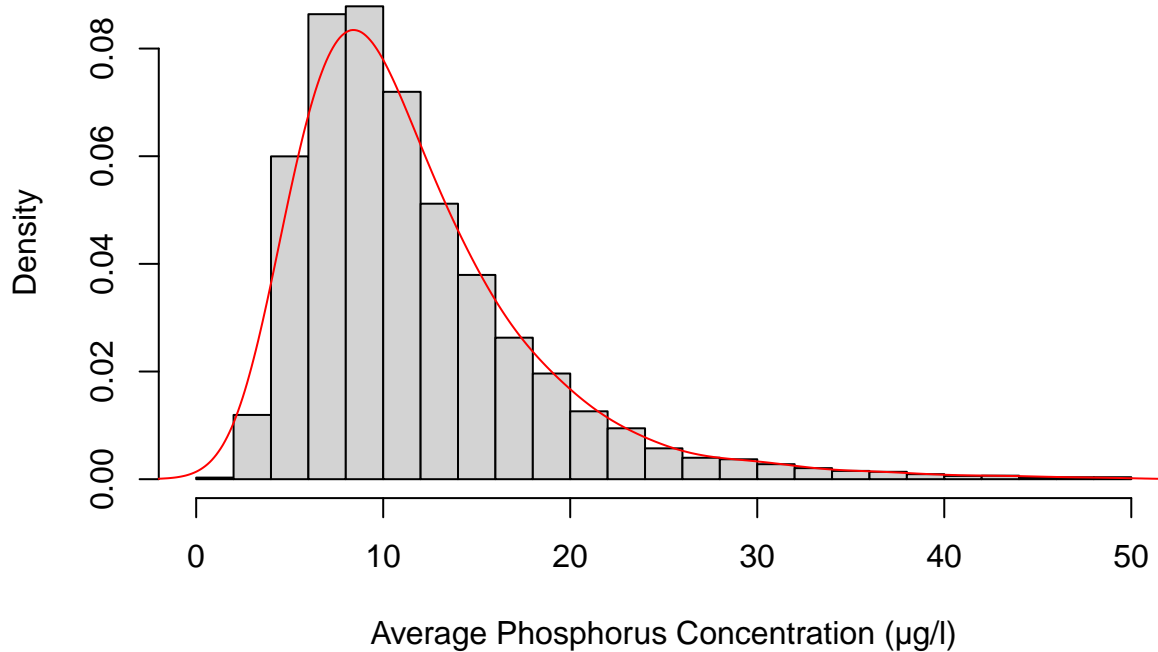
Water quality, as measured by phosphorus levels and secchi depth, has a major influence on the biodiversity of inland lakes and freshwater streams.

In the vast majority of Ontario's inland lakes, phosphorus is the element that controls the growth of algae. As such total phosphorus concentrations ($\mu\text{g/L}$) are most aptly used to assess lake nutrient status. Limnologists place lakes into three categories based on their total phosphorus concentrations: oligotrophic (less than $10 \mu\text{g/L}$), mesotrophic ($10\text{-}20 \mu\text{g/L}$ TP), and eutrophic (over $20 \mu\text{g/L}$). Oligotrophic lakes are low in nutrients and rarely have algal blooms. Mesotrophic lakes vary in characteristics and may experience moderate blooms. Eutrophic lakes have high nutrient levels and often suffer from persistent algal blooms.

A basic graph to give the reader some sense into how our data is structured

Option 1A

Distribution of Average Phosphorus Concentration



Source: <https://hallshawklakes.ca/featured/secchi-readings/#:~:text=Secchi%20Reading%20and%20Lake%20Nutrient,enric>

Secchi disks 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.

Table 2

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```

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

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.

From this, it can be determined that the proportion of inland lakes that exhibit eutrophic conditions is estimated to be 0.254, with a 95% confidence interval ranging from 0.225164 and 0.2824012. Since the p-value $< 2.2\text{e-}16$, 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.

Quantile	Value
25%	0.281
75%	0.300

The 25th percentile and 75th percentile of the bootstrapped proportions of eutrophic inland lakes are estimated to be 0.245 and 0.264, respectively. This suggests **HELP**, so it can be concluded that roughly a quarter of Ontario's inland lakes have high nutrient levels and often suffer from persistent algal blooms.

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 ($\mu\text{g/L}$), labelled TSI(TP), and the Trophic State Index in terms of secchi depth, labelled TSI(SD). The formula for TSI(TP) is derived as $14.42\ln(\text{TP}) + 4.15$, where TP represents the total phosphorus level (the average phosphorus level in $\mu\text{g/L}$). The formula for TSI(SD) is derived as $60 - 14.41\ln(\text{SD})$, where SD is the secchi depth in metres. Thus, the total TSI is $(\text{TSI}(\text{SD}) + \text{TSI}(\text{TP})) / 2$.

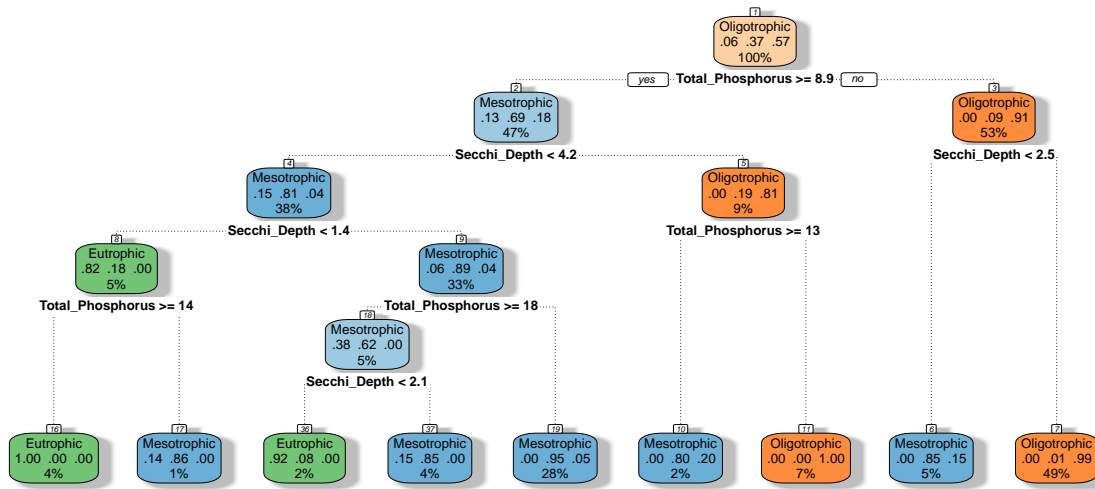
If a lake's TSI is < 30 , it is considered Oligotrophic. If it's TSI is from 30-40, it is Mesotrophic, while any TSI level above 40 is considered Eutrophic. The TSI is a clean and an easily identifiable method of classifying a lake's trophic level. However, many lake databases, including Ontario's do 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 lake's average phosphorus level and secchi depth and instantly determine its 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 - $(\text{TSI}(\text{SD}) + \text{TSI}(\text{TP})) / 2$. The proportion of each lake's trophic status as classified by TSI are as follows:

Year	Eutrophic	Mesotrophic	Oligotrophic
2015	7	19	28
2016	2	10	34
2017	2	28	51
2018	21	89	68
2019	23	139	139
2020	1	9	26
2021	4	15	47
2022	12	109	279

After formally classifying each lake using the TSI, we can then train a model to correctly classify each lake without the use of TSI. 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.

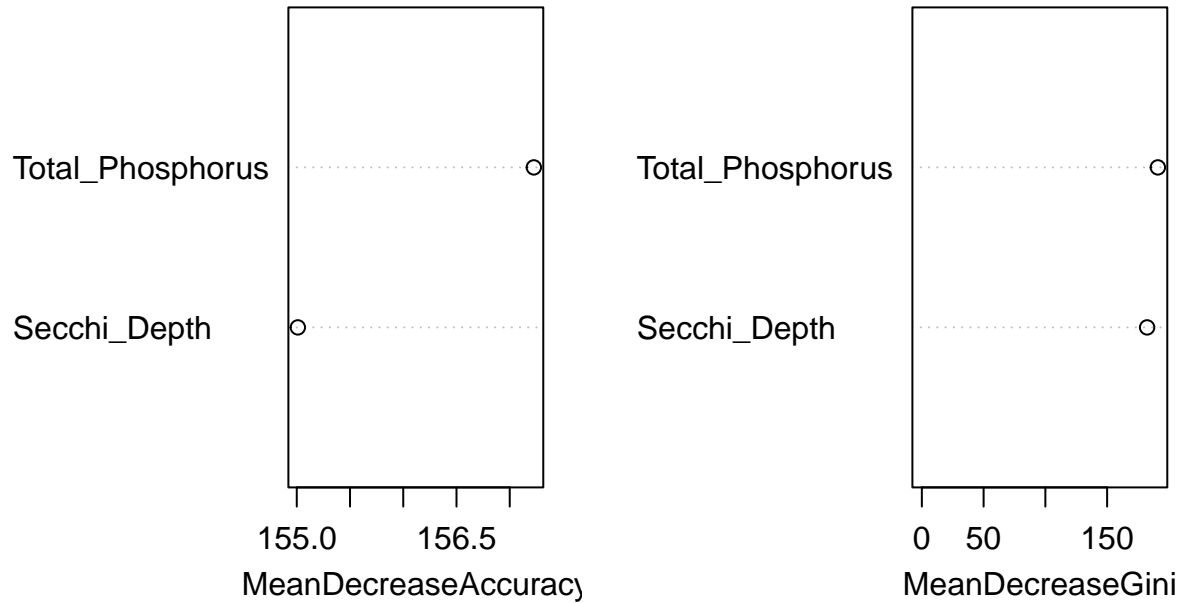


Rattle 2024-Apr-03 03:54:44 rstudio

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.

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 model, we are met with the following graph:

rforest.m



This graph 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 our Random Forest Model (as represented by x)

x
0.9822222

Thus, 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.

Applying our Model

Knowing that our model is extremely accurate, we can use the trophic data predicted by our model to plot each lake to a map of Ontario according to their trophic status.

Plot Map Here

After analyzing our map, we can see that **blank**.

Final Summary

Final summary **here**.