Diatom population structure in 58 lakes in the Hudson Bay Lowlands: An Exploratory Data Analysis

Molly M. Morrissey

Presented to:

Dr. David Schneider

Quantitative Methods in Biology

December 4, 2020

Introduction

The Far-North region of Ontario is a wet terrain abundant with surface waters and peatland. This area, south of Hudson Bay and southwest of James Bay, is known as the Hudson Bay Lowlands (HBL). The vast size of the HBL separates it from other peatlands. Typically, peatlands are confined to geographic depressions and are isolated, surrounded by different biomes. The HBL is a vast "sea" of peatlands and bogs over a flat landscape, interrupted occasionally by soil "islands", and rivers and lakes. The HBL extends 800 miles from the Nottaway and Hurricana Rivers, Quebec, to Churchill, Manitoba (Sjörs, 1959).

Little baseline ecological information exists in the HBL because of its remoteness. Few people live there, and little infrastructure exists. Recently, interest in the HBL has increased because of two seemingly opposing reasons: the hopes of mining a newly discovered large chromite and nickel-copper deposits, and a commitment by the government of Ontario in the 2010 Far North Act to preserve ~50% of the Far North in interconnected tracts of land. Balancing these two disparate endeavors will be a challenge, and regardless of future actions, there is a great need to understand the ecology of the HBL (Jeziorski et al., 2015).

The present study took sediment core samples from 58 lakes in the HBL to check for diatom abundances in the top of the sediment. The lakes were chosen randomly. All lakes were shallow, with the depth ranging from 0.9m—3m, with an average depth of 1.7 m. Environmental variables were measured from each lake, with the lakes showing a broad range in characteristics. The data collected will examine relationships between diatom communities and environmental conditions.

Exploratory Data Analysis (EDA)

Typically, most ecological research is done with hypothesis testing, or with the ultimate goal of creating data models. However, this does not always capture the complexity of ecological relationships. It can be helpful to first perform an exploratory data analysis (EDA) for multidimensional data. EDA is a method of making sense of a large data set to design future analyses (Borcard et al., 2011). With EDA, the researchers need not set a hypothesis for the data they collect, rather they can make use of visual representations of observed phenomena to make inferences about relationships. For the purposes of our study, we felt EDA was appropriate due to how little is known about the ecology of the HBL.

In 1977, John Tukey published the seminal book *Exploratory Data Analysis*. As a statistician, Tukey began to have doubts on the emphasis that had been placed on confirmatory data analyses (hypothesis testing), and it was his belief that we should instead rely on the data to suggest hypotheses before we actually tested them (Tukey, 1962, 1977). EDA uses graphical representations to interpret data rather than statistical tests relying on mathematics. Tukey states: "**The greatest value of a picture** is when it *forces* us to notice **what we never expected to see.**" (1977, bold and italics formatting from source). Thus, it was his belief that EDA could remove the ever-present confirmation bias amongst scientists testing their own hypotheses. With EDA, we exit the world of p-values and goodness of fit tests and enter the world of graphs, using our eyes to judge patterns.

> **It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it.**
>
> --John Tukey, *Exploratory Data Analysis*

Methods

In 2016, we conducted a survey of 58 lakes in the Hudson Bay Lowlands region in north-west Ontario. We collected sediment samples using an OGS designed gravity corer. Sediment samples were collected from the tops of the sediment, so 0.0-5.0 cm depth. For our analyses, we will be looking at the top 1 cm of sediment; the rest will be freeze dried for future analyses. From the top 1 cm, we counted and separated individual diatoms. We identified and tabulated species using a Leica DRMB microscope. We counted at least 90 diatom individuals from each sediment sample. To eliminate bias from non-random distributions of diatoms within the sediment, we counted every individual on the prepared slides. Environmental measurements were taken from each of the 58 lakes (**Table 1**). For water samples tested in the lab, we collected 0.5L from each lake. Methods were adapted from Jeziorski et al. (2015).

Environmental Variable (abbrev.)	**Unit**	**Measurement method/tool**
Lake depth (Depth)	m	Hawkeye DepthTrax Handheld Depth Finder*
Temperature (Temp)	°C	OHAUS ST20 Pen Reader*
Specific conductivity (Spec Cond)	µS/cm	OHAUS ST20 Pen Reader*
Conductivity (Cond)	µS/cm	OHAUS ST20 Pen Reader*
Dissolved organic carbon (DOC)	mg/L	Water sample; high temperature combustion
Lake color (Col)	TCU	Secchi disk*
Nitrogen: $NH_3 + NH_4^-$ (N:NH3+NH4)	µg/L	Water sample, salycylate chemical titration method
Nitrogen: $NO_3^- + NO_2^-$ (N:NO3+NO2)	µg/L	Water sample; spectrophotometric cadmium reduction
Total Kjeldahl Nitrogen (TKN)	µg/L	Water sample; digestion, distillation, ammonia method
Reactive silicate (SiO4)	mg/L	Water sample; gas segmented continuous flow colorimetric analysis
Total phosphorus (P)	µg/L	Water sample; ascorbic acid test
Resistance (Resist)	ohms	Water sample; conductivity reader
Total dissolved solids (TDS)	g/L	Water sample; gravimetric analysis
pH	N/A	OHAUS ST20 Pen Reader*

Table 1. *Environmental variables measured in lakes. *Denotes measurement taken in the field, all other measurements taken in the lab.*

In total, we counted 381 diatom species. To avoid an absurdly large analysis, we discarded any diatom species that did not appear in at least two samples or had <1% relative abundance. We then combined species into larger taxa groups. We ended up with 21 distinct taxa groups (**Table 2**).

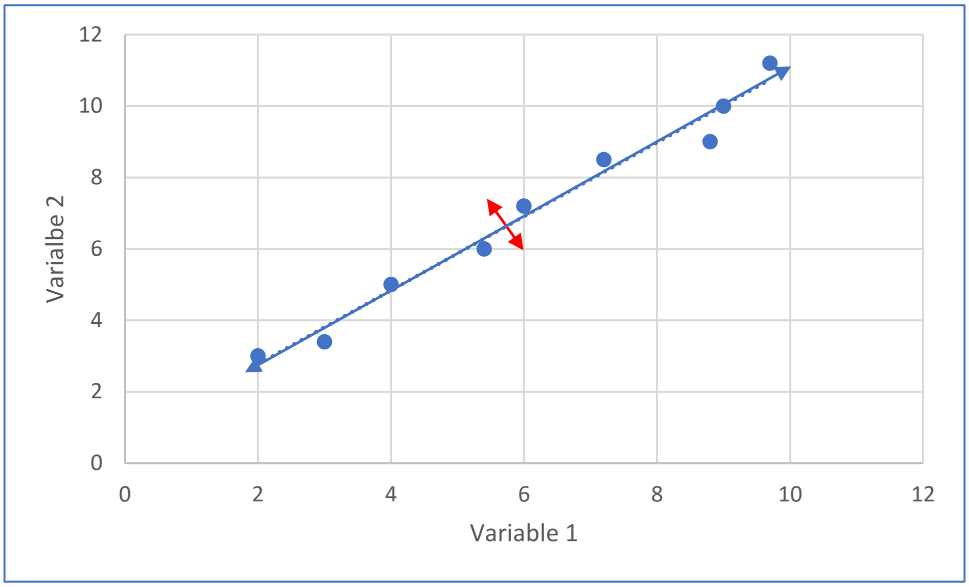
Species	***Abbreviation***
Achnanthes spp	Ach
Achnanthes minutissima	Achm
Aulacoseira spp	Aul
Brachysira spp	Bra
Cymbella spp	Cym
Eunotia spp	Eun
Benthic Fragilaria spp	FraBen
Fragilaria construens	Frac
Fragilaria pinnata	Frap
Fragilaria (virescins) v. exigua	Frav
Fragilaria brevistriata	Frab
Neidium spp	Nei
Nitzschia spp	Nit
Pinnularia spp	Pin
Stauroneis spp	Sta
Planktonic	Plank
Navicula jaagii	Navj
Large Navicula complex	NavLrg
Small Navicula complex	NavSml
Navicula kuelbsii/vitiosa	Navkv
Navicula minima	Navm

Table 2. *Diatom species abbreviations.*

Analysis

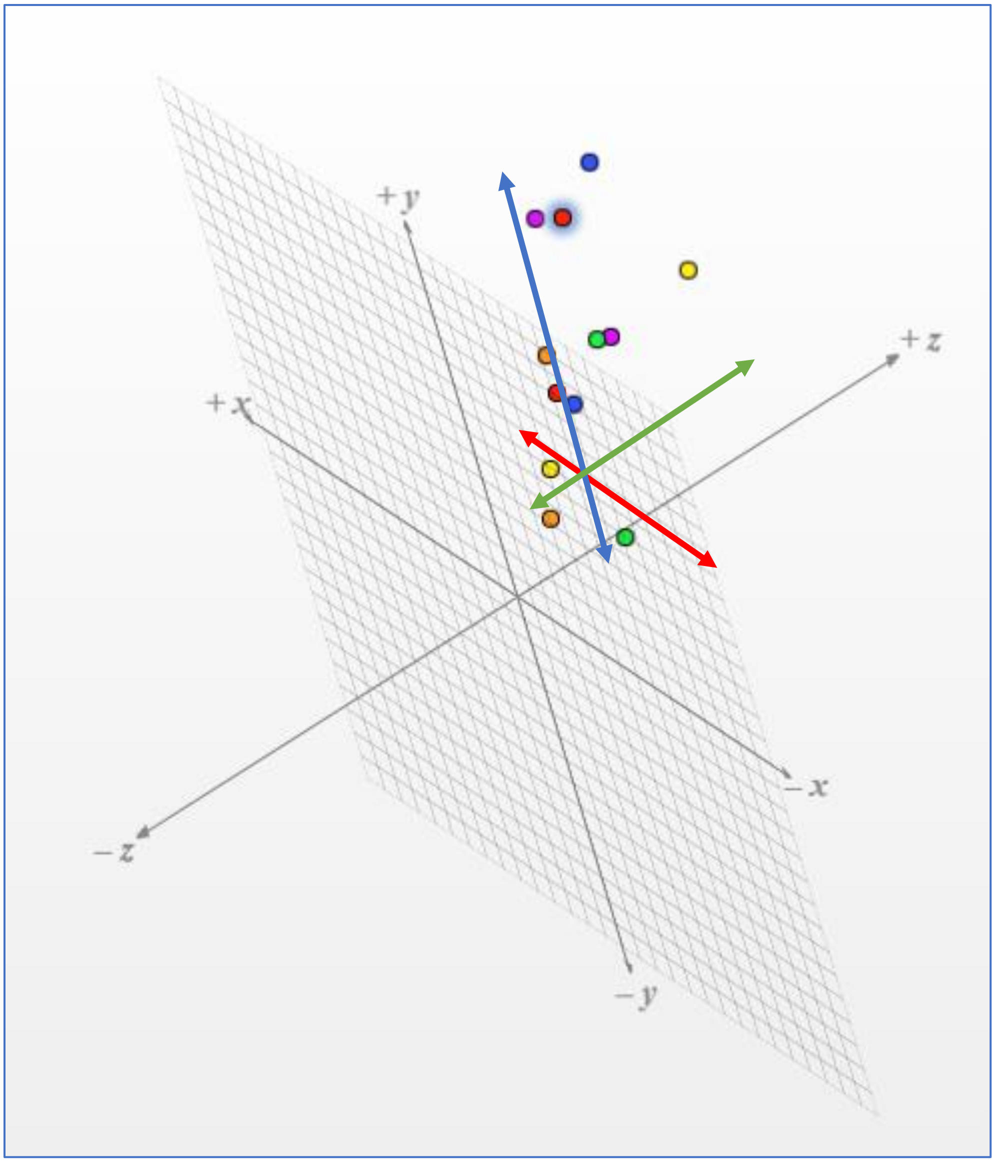
Principle Component Analysis (PCA)

All analyses and graphical models were done in RStudio (R version 4.0.3, 2020-10-10). I used principle component analysis (PCA) methods for my analysis. PCA allows relationships between multiple variables to be displayed graphically, reducing the dimensions needed to do so. For example, if you were to graph out two variables to look at their relationship, you could do so on a simple 2-dimensional graph:



The blue line accounts for all variation from left to right, and the red line accounts for all variation above and below the blue line. We can think of the blue line as principle component (PC) 1, and the red line as PC2.

If we add a third variable, and want to see the relationships of all three variables, we need to create a 3-dimensional graph:



It is somewhat challenging to see the relationships between the three variables without rotating the graph around. In this graph, we have three principle components. The blue line is PC1, the red line PC2, and the green line PC3. We are unable to add a fourth (or more) variable to a graph because of the limitations of the human brain, however we understand that as you add more variables, you will have more principle components. There is a principle component for every variable you have in a dataset. The two strongest principle components—that is, the two explaining the most variance between variables—can be used to plot the subjects from where the variables were collected to look at relationships.

Environmental variables

I first determined the strength of the principle components for the environmental data collected. There were 14 variables collected from the 58 lakes, hence 14 variables. An ***eigenvalue*** is a measure of the importance of an axis (Jeziorski et al., 2015). The eigenvalues for the 14 variables are as follows:

```
Eigenvalues for unconstrained axes:

  PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9  PC10  PC11
5.274 3.027 1.569 1.302 0.906 0.728 0.440 0.303 0.242 0.122 0.060
 PC12  PC13  PC14
0.026 0.000 0.000
```

I further examined the data to see the ***proportion explained*** by each eigenvalue:

```
Eigenvalues, and their contribution to the correlations
Importance of components:
                         PC1    PC2   PC3     PC4     PC5     PC6
Eigenvalue            5.2744 3.0268 1.569 1.30212 0.90620 0.72817
Proportion Explained  0.3767 0.2162 0.112 0.09301 0.06473 0.05201
Cumulative Proportion 0.3767 0.5929 0.705 0.79800 0.86273 0.91474
                          PC7     PC8     PC9     PC10     PC11
Eigenvalue            0.43998 0.30330 0.24156 0.121852 0.060244
Proportion Explained  0.03143 0.02166 0.01725 0.008704 0.004303
Cumulative Proportion 0.94616 0.96783 0.98508 0.993788 0.998091
```

```
                          PC12      PC13      PC14
Eigenvalue            0.026266 3.685e-04 9.521e-05
Proportion Explained  0.001876 2.632e-05 6.801e-06
Cumulative Proportion 0.999967 1.000e+00 1.000e+00
```

Now, we can plot the eigenvalues as a histogram:

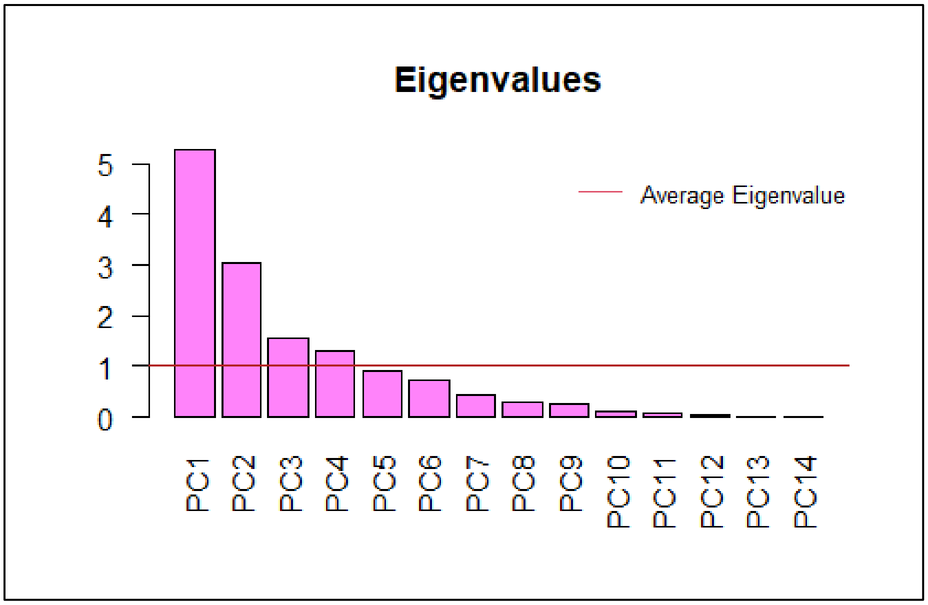


Figure 1. *Principle component eigenvalues, environmental variables. Average eigenvalues shown with red line. PC1 = 37.67%, PC2 = 21.62%, PC3 = 11.20%, PC4 = 9.30%*

PC1 and PC2 account for almost 60% of variation amongst environmental data; thus, I feel confident using these two PCs to interpret these data. I plotted the environmental data on a graph to determine which variables were driving PC1 and PC2 (**Figure 2**). See **Table 1** for abbreviations for each environmental value.

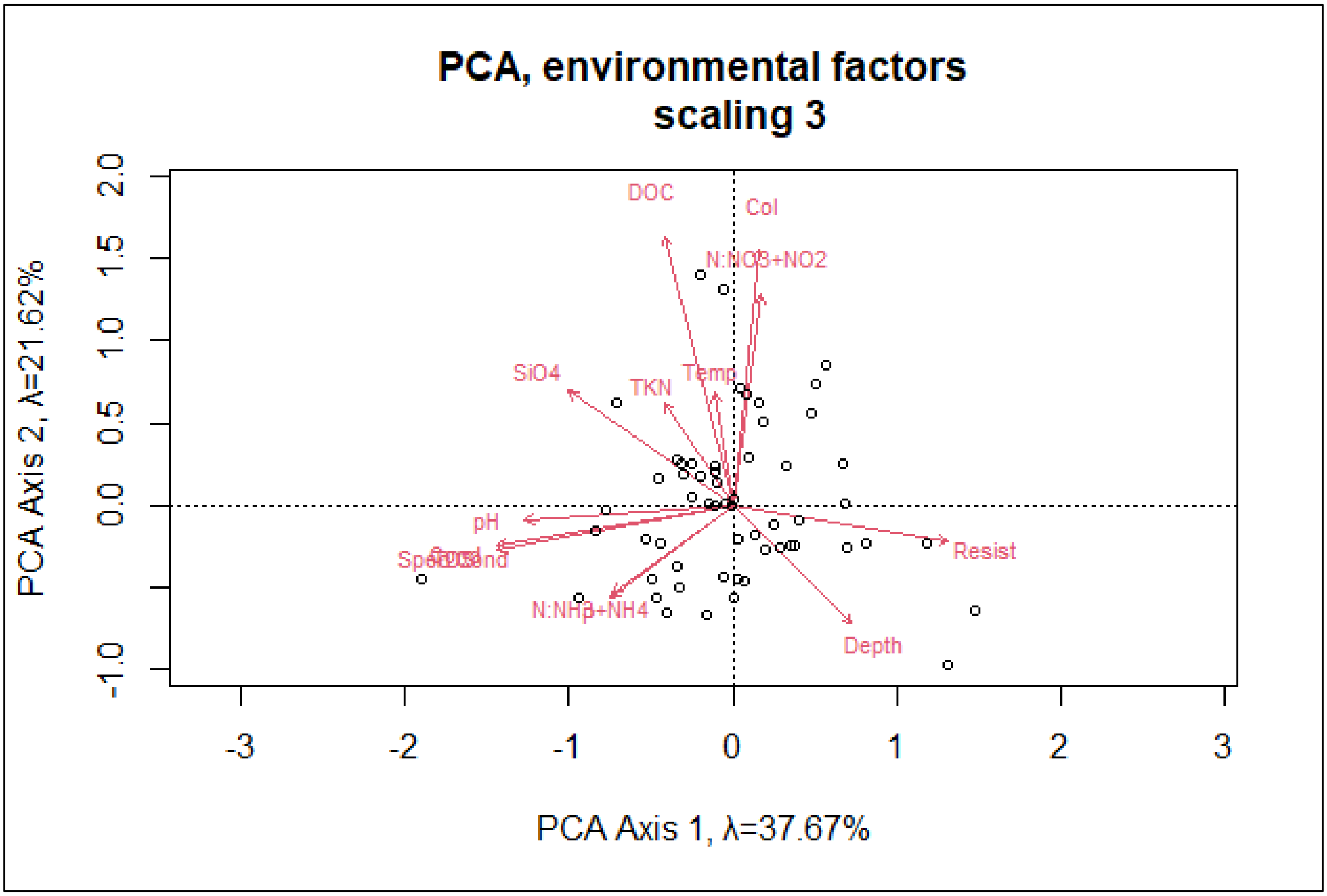


Figure 2. *Environmental factors plotted against PC1 (x-axis) and PC2 (y-axis). Lakes plotted as circles.*

Now I can interpret the graph. On this graph, the environmental variables are called ***loading vectors***. The loading vectors refer to the variables driving the PCs ("loading" the PCs). To determine which loading vectors are the most important drivers of variation, I look for the longest lines. Resistance is positively correlated with PC1, while pH, conductivity, specific conductivity, and total dissolved solids are negatively correlated with PC1. Dissolved organic carbon, lake color, and the NO_3/NO_2^- group are positively correlated with PC2. Total Kjeldahl nitrogen, temperature, phosphorus, depth, reactive silicate, and NH_3/NH_4 group do not appear to be strong drivers of PC1 or PC2. For my interpretation, I am saying a “strong” correlation is any value ≥ |1|. Some sources will give a smaller absolute value when determining strong correlation, however I am choosing |1| to reduce the number of significant vectors.

	PC1	PC2
Depth	0.6710	-0.58134
Temp	-0.1053	0.54847
Spec Cond	**-1.3220**	-0.22040
Cond	**-1.3195**	-0.19584
DOC	-0.3844	**1.31043**
Col	0.1483	**1.25192**
N:NH3+NH4	-0.6757	-0.42846
N:NO3+NO2	0.1644	**1.03168**
TKN	-0.3837	0.50096
SiO4	-0.9235	0.55568
P	-0.6862	-0.45323
Resist	**1.2074**	-0.17987
TDS	**-1.3231**	-0.21634
pH	**-1.1662**	-0.07069

Table 3. *Loading values for environmental factors. Significant loadings (i.e. value ≥ |1| are highlighted.*

Next, I performed a correlation analysis on the lakes (**Figure 3**). This can help explain any differences between the 58 lakes.

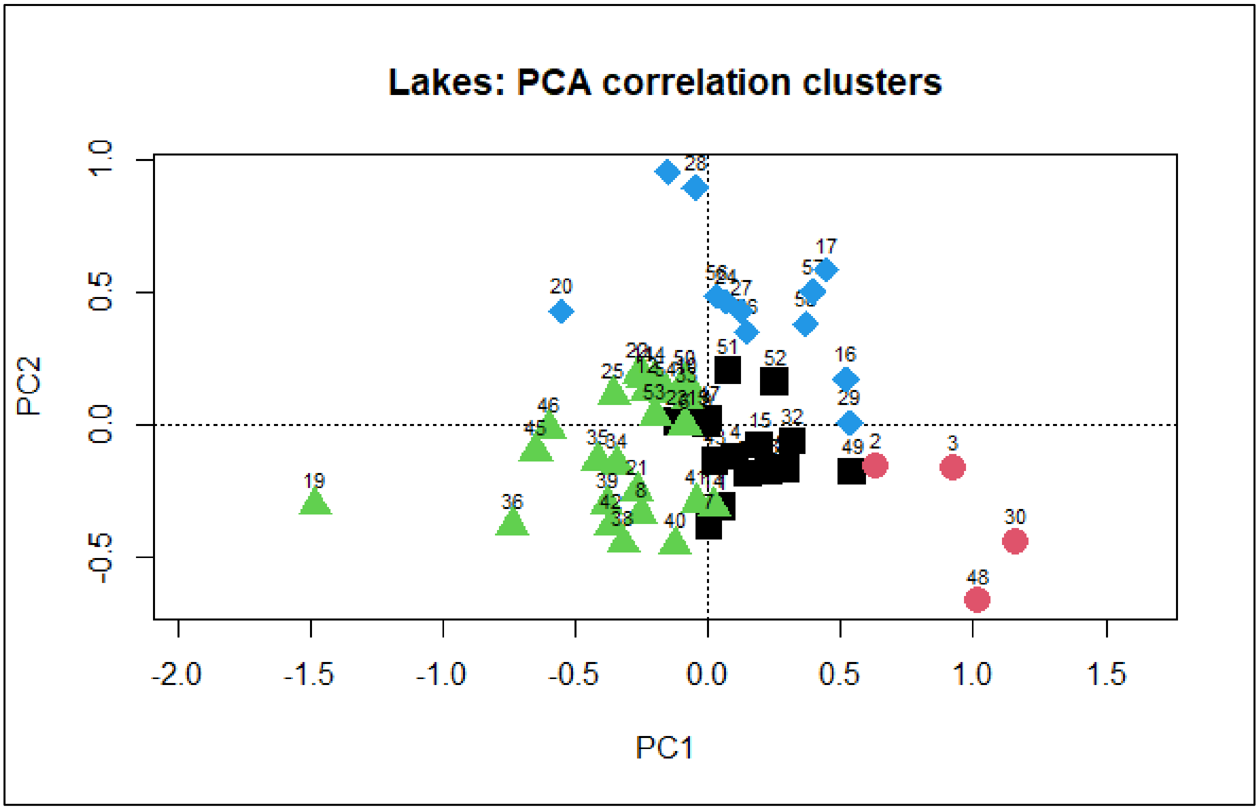


Figure 3. *Lakes plotted against PC1 and PC2.*

Lakes are plotted by their ***site scores***. A site score measures how much an object, in this case the lakes, correlates with the two PCs. The four colors on the graph represent four groups with similar characteristics. Using this figure with **Figure 2** can elucidate environmental patterns within the lakes.

Finally, we want to see where the diatom species land in this PCA (**Figure 4**). See **Table 2** for diatom species abbreviations.

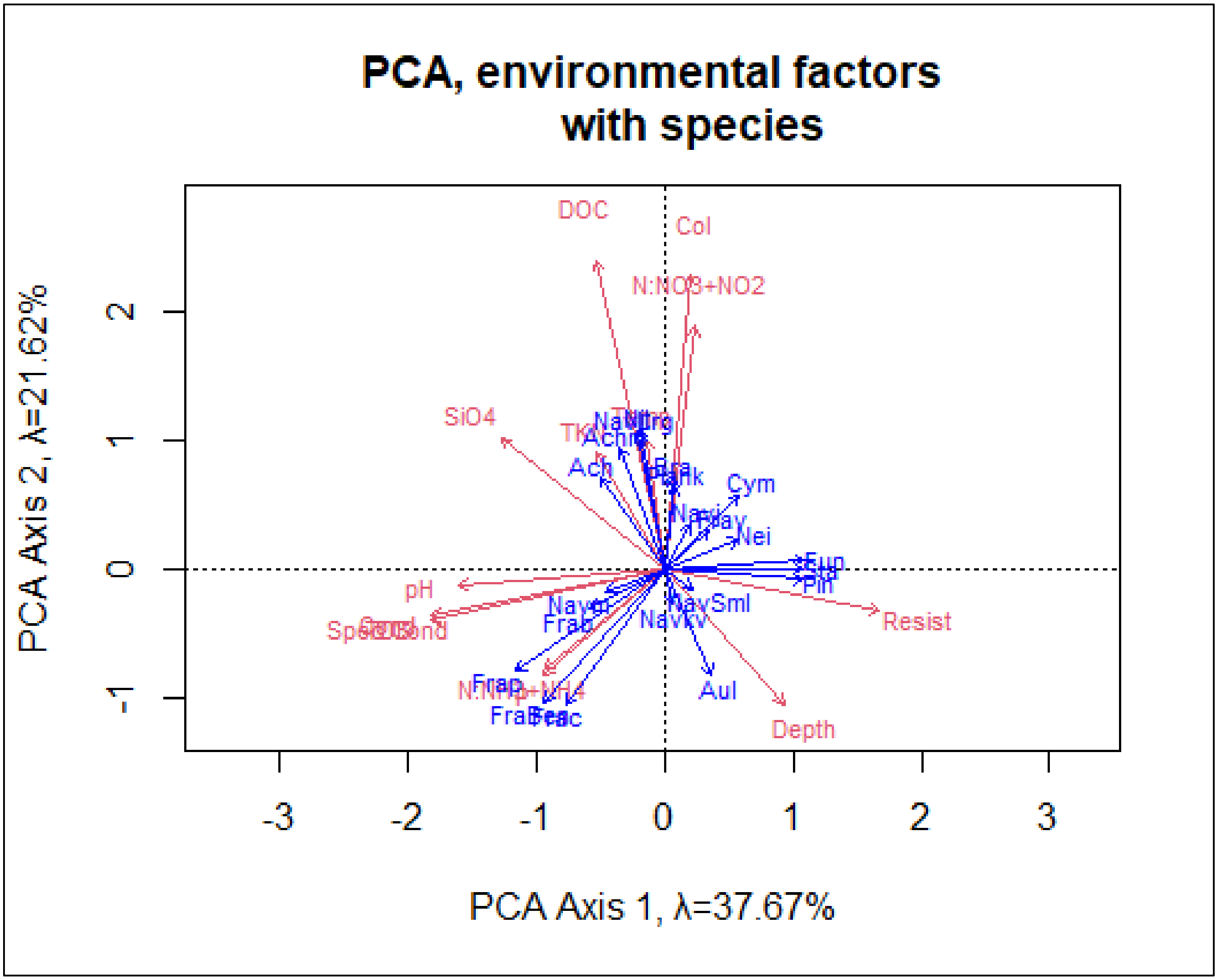


Figure 4. *Environmental factors and species plotted against PC1 and PC2.*

We can see some patterns emerge. *Pinnularia spp*., *Eunotia spp.*, and *Stauroneis spp*. are grouped together, and they are positively aligned to the PC1 axis. *Fragilaria construens, Fragilaria pinnata*, and the benthic *Fragilaria spp*. are grouped together, aligning about equally negatively with PC1 and PC2. *Nitzschia spp.*, *Achnanthes minutissima*, and the large *Navicula* complex are clustered together and aligned positively with PC2.

Species

Next, I wanted to see if it would be worthwhile to perform a PCA using the species data rather than the environmental data. I first calculated eigenvalues for species:

```
Eigenvalues for unconstrained axes:
  PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8
6.291 4.049 2.378 1.690 1.325 0.965 0.919 0.762
(Showing 8 of 21 unconstrained eigenvalues)
```

I further examined the data to see the proportion explained by each eigenvalue:

```
Eigenvalues, and their contribution to the correlations


Importance of components:
                         PC1    PC2    PC3     PC4     PC5
Eigenvalue            6.2913 4.0495 2.3776 1.69004 1.32491
Proportion Explained  0.2996 0.1928 0.1132 0.08048 0.06309
Cumulative Proportion 0.2996 0.4924 0.6056 0.68611 0.74921
                          PC6     PC7     PC8     PC9    PC10
Eigenvalue            0.96467 0.91927 0.76150 0.48435 0.40903
Proportion Explained  0.04594 0.04377 0.03626 0.02306 0.01948
Cumulative Proportion 0.79514 0.83892 0.87518 0.89824 0.91772
                         PC11    PC12    PC13     PC14     PC15
Eigenvalue            0.36237 0.31527 0.27186 0.196747 0.175421
Proportion Explained  0.01726 0.01501 0.01295 0.009369 0.008353
Cumulative Proportion 0.93498 0.94999 0.96293 0.972304 0.980657
                          PC16     PC17     PC18     PC19
Eigenvalue            0.137915 0.125655 0.068053 0.045400
Proportion Explained  0.006567 0.005984 0.003241 0.002162
Cumulative Proportion 0.987224 0.993208 0.996449 0.998611
```

```
                         PC20      PC21
Eigenvalue             0.02877 4.070e-04
Proportion Explained   0.00137 1.938e-05
Cumulative Proportion  0.99998 1.000e+00
```

Finally, I displayed the eigenvalues for species on a histogram:

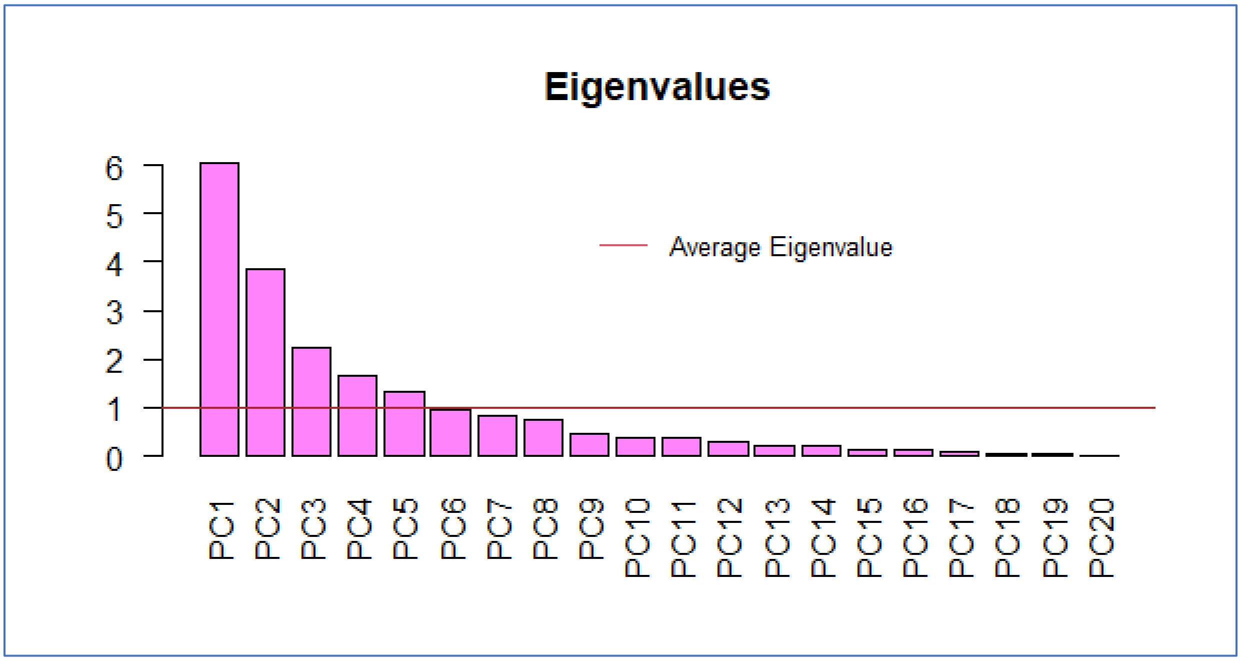


Figure 5. *Eigenvalues for species. PC1=29.96%, PC2=19.28%, PC3=11.32%, PC4=8.05%*

The data are not as compelling as they were for the environmental factors. We must go out to PC3 before we reach 60% variance explained. Including PC4 brings us to almost 70% explained. It is my opinion that running at PCA on the species data will not yield better relationship inferences than with the environmental data. However, I will use the species PCs to explore species relationships with environmental predictors.

I will first look at the variables that were most important to the environmental PC1 (resistance, pH, conductivity, specific conductivity, and TDS) and see how they relate to species PC1:

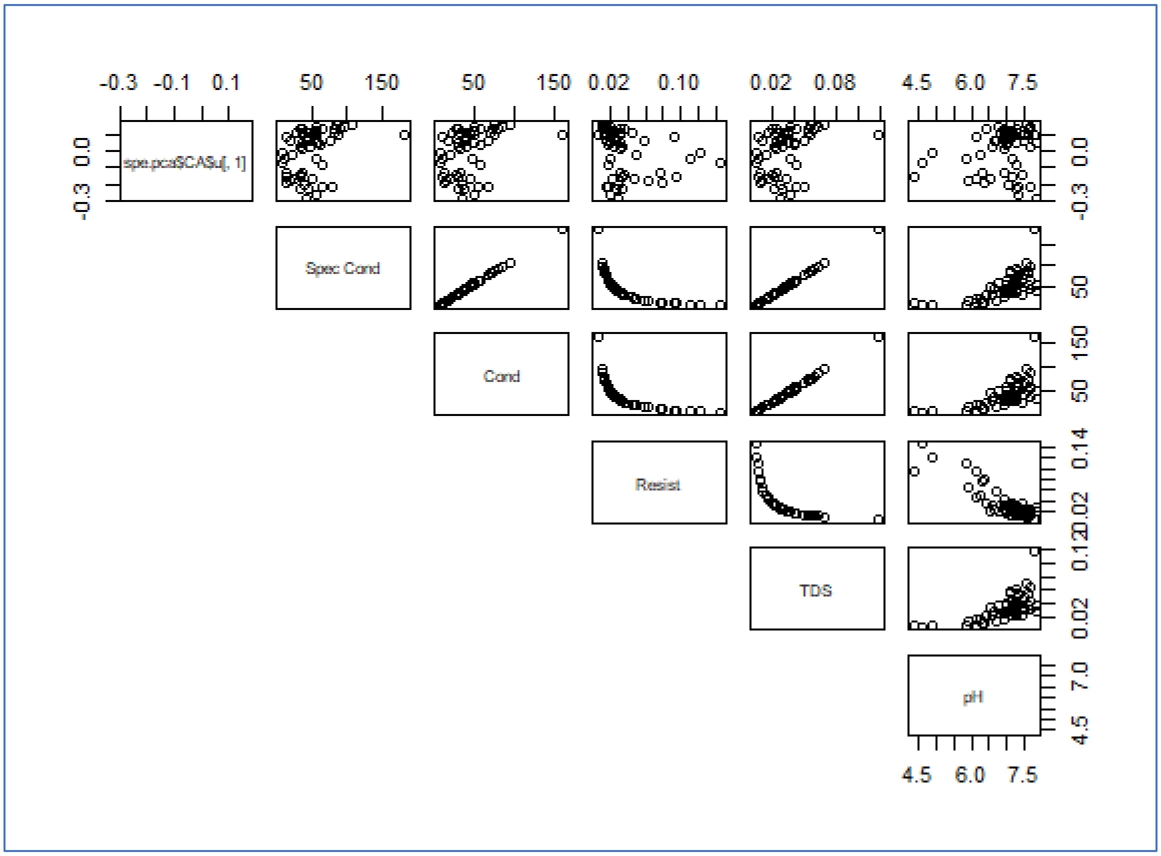


Figure 6. *Pairwise matrix of species PC1 with specific conductivity, conductivity, resistance, TDS, and pH.*

We can see strong relationships between all of these variables, however we do not see much of a relationship between PC1 for the species with any of these variables. Sometimes, when doing PCA on abundance data, the drivers of PC1 are the most abundant diatoms collected, which can shadow any relationships between species. It can be helpful to look at the next most important PC to see patters. Let’s run this analysis again with PC2 for species:

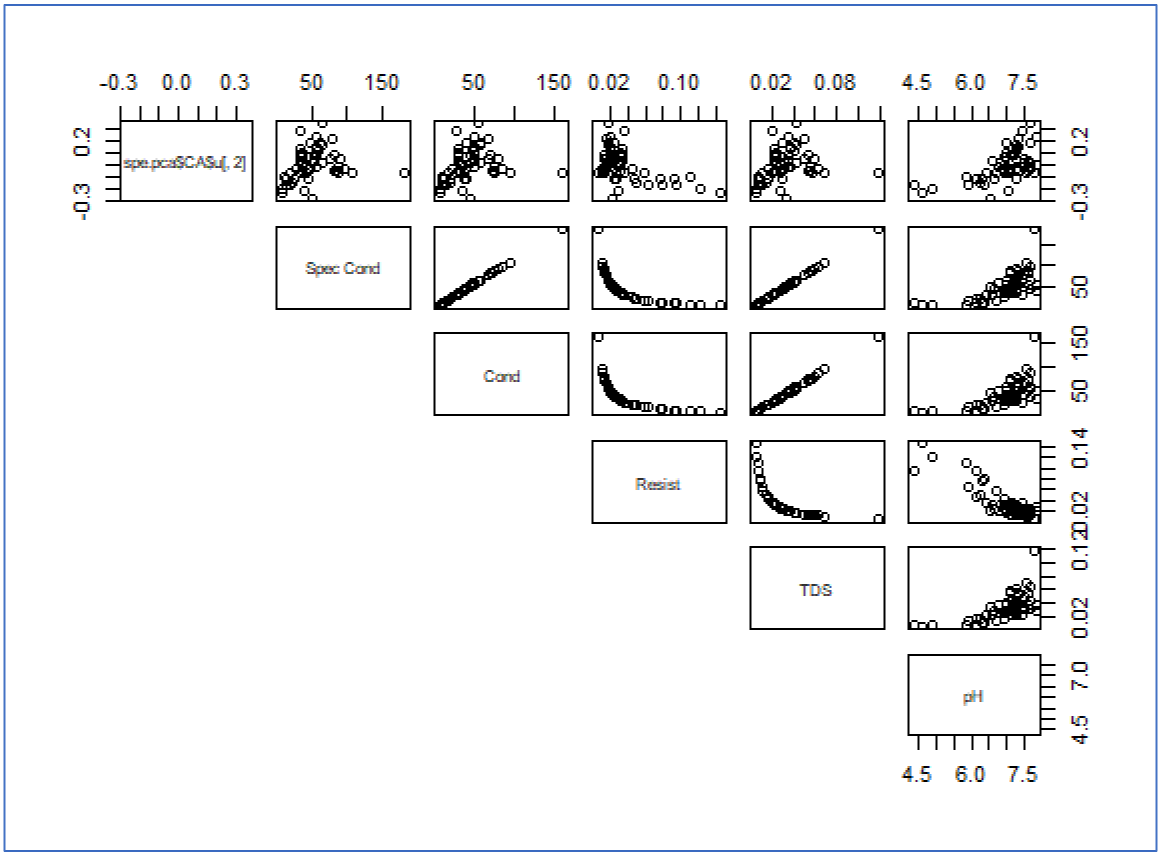


Figure 7. *Pairwise matrix of species PC2 with specific conductivity, conductivity, resistance, TDS, and pH.*

We can see there is a little bit more of a relationship between species and the environmental variables, although it is still not very clear. Since there are 21 diatom taxa involved, we can assume they all have varying environmental preferences, therefore we would not get as clear of a relationship as we would if we were looking at one taxa at a time.

Next, I will look at the variables that were important to environmental PC2: dissolved organic carbon, lake color, and the NO_3/NO_2^- group. We will only look at species PC2 this time:

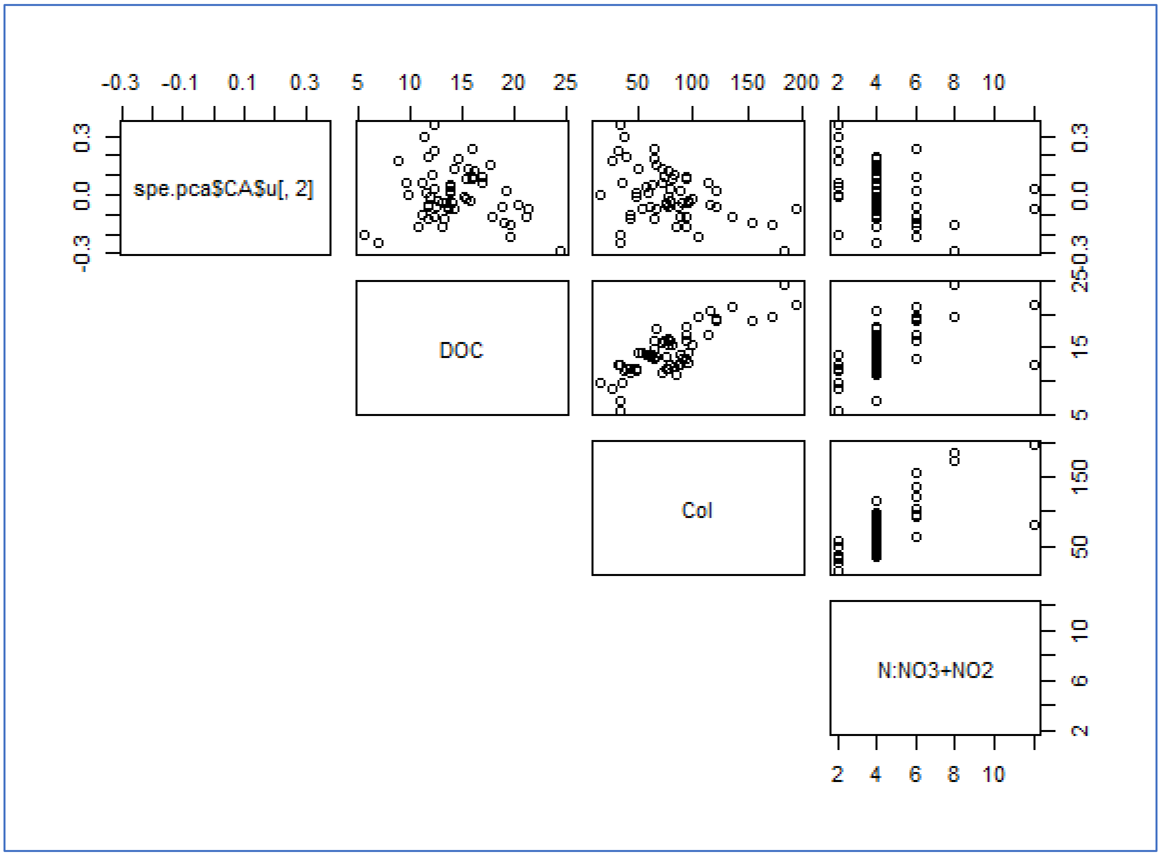


Figure 8. *Pairwise matrix of species PC2 with dissolved organic carbon, lake color, and the NO_3/NH_2^- group.*

With this graph, we do not see any discernable relationship between the species and any of the environmental variables. We do see dissolved organic carbon and color strongly positively related.

Discussion

We can make some inferences about the data based on the figures we created and what we know about freshwater biology. From the loadings on **Figure 2**, we see that pH, conductivity, specific conductivity, and total dissolved solids (TDS) have small angles between each vector, hence they are closely related. Conductivity is a proxy for salt water—salt water is a better conductor of electricity than freshwater. Increases in TDS result in increases in conductivity. pH increases as water gets saltier as well. Therefore, it makes sense that conductivity, specific conductivity, and pH are correlated. *Fragilaria construens, Fragilaria pinnata,* and the benthic *Fragilaria spp*. all seem to be associated with pH, conductivity, specific conductivity and TDS. We can assume that these three taxa of diatoms prefer saltier, higher pH waters.

Resistance measures the ability of a substance to resist an electrical current. Increased resistance can be thought of as a proxy for fresher water. Therefore, it is obvious why we see resistance on the opposite side as pH, conductivity, specific conductivity and TDS. We have three diatom species strongly correlated positively with PC1: *Pinnularia spp*., *Eunotia spp.,* and *Stauroneis spp.* We can assume that these three taxa are more sensitive to increases in salinity, thus are more likely to be found is lower pH and lower salinity water.

Color, dissolved organic carbon (DOC), and the NO_3/NO_2^- group are closely associated together, and with PC2. Color is a measurement of the transparency of a body of water, hence the number of dissolved solids and nutrients. It is curious that both TDS and phosphorus are correlated in opposite directions of color, since increases in both TDS and phosphorus cause higher color readings. This is something that could be explored further in subsequent studies. *Nitzschia spp.*, *Achnanthes minutissima,* and the large *Navicula* complex are positively correlated with PC2, meaning these species thrive in waters rich with DOC and NO_3/NO_2^-.

For future studies, I would recommend focusing on the strongest environmental variables and how they affect diatom abundances. Some environmental variables that may be important but did not make my arbitrary loading value ≥ |1| cutoff are depth and reactive silicate. We see depth in opposition to color, DOC and NO_3/NO_2^-. Nutrients tend to aggregate in the shallow parts of water, with lower depths decreasing nutrient loading.

EDA is an excellent way of taking a large collection of data and letting it tell you what is important, rather than taking selected data and trying to prove importance. With EDA, you may see relationships you could have otherwise missed.

References

Borcard, D., Gillet, F., & Legendre, P. (2011). *Numerical Ecology with R*. Springer-Verlag. https://doi.org/10.1007/978-1-4419-7976-6

Jeziorski, A., Keller, B., Dyer, R., Paterson, A., & Smol, J. (2015). Differences among modern-day and historical cladoceran communities from the "Ring of Fire" lake region of northern Ontario: Identifying responses to climate warming. *Fundamental and Applied Limnology / Archiv Für Hydrobiologie*, *186*, 203–216. https://doi.org/10.1127/fal/2015/0702

Sjörs, H. (1959). Bogs and Fens in the Hudson Bay Lowlands. *ARCTIC*, *12*(1), 2–19. https://doi.org/10.14430/arctic3709

Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics*, *33*(1), 1–67.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Pub. Co.