AN ANALYSIS OF LAKE ERIE’S NEAR SHORE HABITATS WITH THE BENTHIC DIATOM METRIC

KenJon Yeong
John Carroll University, kyeong16@jcu.edu

Follow this and additional works at: http://collected.jcu.edu/mastersessays
Part of the Mathematics Commons

Recommended Citation
http://collected.jcu.edu/mastersessays/14

This Essay is brought to you for free and open access by the Theses, Essays, and Senior Honors Projects at Carroll Collected. It has been accepted for inclusion in Masters Essays by an authorized administrator of Carroll Collected. For more information, please contact connell@jcu.edu.
AN ANALYSIS OF LAKE ERIE’S NEAR SHORE HABITATS
WITH THE BENTHIC DIATOM METRIC

An Essay submitted to the
Office of Graduate Studies
College of Arts & Sciences of
John Carroll University
Partial Fulfillment of the Requirements
For the Degree of
Master of Science.

By
Ken Jon Yeong
2015
This essay of Ken Jon Yeong is hereby accepted:

Thomás H. Short
Advisor – Dr. Thomas H. Short. Faculty

April 6, 2015
Date

I certify that this is the original document

Author - Ken Jon Yeong

04/06/2015
Date
# TABLE OF CONTENTS

1. INTRODUCTION
   ........................................................................................................... 4

2. DATASET DESCRIPTION AND MODIFICATIONS
   ........................................................................................................... 7

3. THE RELATIONSHIPS AMONG BDM, BASIN, AND RIVERS
   ........................................................................................................... 12

4. THE PREDICTIVE ABILITY OF TP, COND AND NO\textsubscript{2}
   ........................................................................................................... 19

5. THE RELATIONSHIP BETWEEN BDM AND THE RIVER MILE
   ........................................................................................................... 24

6. CONCLUSION
   ........................................................................................................... 31

7. REFERENCES
   ........................................................................................................... 34
1. INTRODUCTION
The preservation of Lake Erie is vital to the ecological and economical sustainability of Northeast Ohio. Over the past decade, initiatives such as the Great Lakes Restoration Initiative and the Lake Erie Protection and Restoration Plan have been launched to restore and improve the condition of Lake Erie. These initiatives were launched to ensure the longevity of Lake Erie by restoring its habitats, protecting native species, preventing and controlling invasive species and reducing nutrient run off that contributes to harmful algae blooms [[1], 1].

In order to evaluate these initiatives, there needs to be a metric that accurately measures the condition of Lake Erie’s water and habitats. In the 1980’s the Ohio Environmental Protection Agency (OEPA) began using biota to assess the health of streams after the concept of biologic integrity was introduced [[4], 3]. Unfortunately, the OEPA did not monitor the near shore water of Lake Erie. Furthermore, proven biotic metrics to assess near shore habitats have not been available [[4], 3]. Therefore a new metric had to be developed to assess the near shore habitats of Lake Erie.

The Benthic Diatom Metric (BDM), developed by Dr. Gerald Sgro and colleagues from the University of Minnesota Duluth, was developed as part of the Great Lakes Environmental Indicator project. It was developed to monitor the Lake Erie near shore habitats because the biotic indices currently in used for rivers and streams (The Invertebrate Community Index and the Electro-Fishing Method) were not appropriate for near shore habitats [[4], 4].

The BDM score is based on “weighted average total phosphorus optima for the diatom species from a training set of 155 samples collected in the near shore throughout the Great Lakes” [[4], 4]. The species optima in the training set correlated more strongly with conditions in the watersheds than with chemistry samples because water chemistry fluctuates more rapidly than the turnover of diatom assemblages. Diatom assemblages typically have a two-week turnover [[4], 4]. The training set optima was converted into a
diatom water quality metric by ranking the diatom species by their total phosphorus optima. The diatoms were divided into 10 groups and ranked from one to 10, with 10 representing species with a preference for lowest total phosphorus concentrations (cleaner water).

Thus, the metric of a sample is taken as a measure of the composition of diatoms rather than a measure of the total phosphorus in water sample \([4], 4\). The BDM score ranges from one to 10, with a higher score indicating a higher percentage of pollution sensitive organisms (organisms in a higher ranking group), and thus the lowest level of pollution \([4], 4\).

In his research, Dr. Sgro found that the “best multiple regression model relating water chemistry to BDM score included the log base 10 of conductivity \((\log_{10}(\text{Cond}))\), total phosphorus \((\log_{10}(\text{TP}))\) and nitrogen dioxide \((\log_{10}(\text{NO}_2))\) \([[[4]], 7\). The multiple regression model had an adjusted \(R^2\) of 0.64 that was significant with a p-value of less than 0.001 \(([4]), 7\). Thus 64 percent of the variation in BDM was accounted for by the conductivity, total phosphorus and nitrogen dioxide of the water sample. In simple linear regression models, the BDM score had \(R^2\) values of 0.46, 0.40, and 0.24 with \(\log_{10}(\text{Cond}), \log_{10}(\text{TP})\) and \(\log_{10}(\text{NO}_2)\) respectively.

This essay addresses the distributions of BDM scores between different basins and rivers within and along the shores of Lake Erie. It also addresses the strength of the relationships among BDM, conductivity, total phosphorus, and nitrogen dioxide over the entire region. It is our goal to give the reader an understanding of the water quality of Lake Erie through the use of BDM. More importantly, we want to help determine the factors that influence water quality. We will do that through the scrutiny of the strength of the relationships among BDM, Cond, TP, and NO\(_2\). We hope this paper provides a better understanding of the current challenges faced to accurately assess the state of Lake Erie’s near shore habitats and provides insight to better serve Lake Erie through the continuous efforts of its restoration.
We will begin by describing the dataset provided by Dr. Sgro. We will address the information contained in the dataset and the modifications that were made to make the data suitable for statistical analysis. Next, using this dataset, we will explore the relationships among BDM, basins, and rivers to determine whether the BDM score differs significantly in various basins or rivers. Third, we will analyze the strength of the relationships among BDM and Cond, TP, and NO$_2$ to test whether these three variables are reliable predictors of the BDM score. Lastly, we will study the changes of BDM as we go further upstream, beginning from the harbor, up to the mouth of the river, and finally into the river itself. All analyses were conducted in R Studio using the base package [[3]] and additional packages such as ggplot2 [[5]] and ggmap [[2]].
2. DATASET DESCRIPTION AND MODIFICATIONS
The original data were collected and analyzed by the OEPA. However, the data were not collected specifically for this project. Dr. Sgro used the data provided by the OEPA and put together the dataset used for this project. Dr. Sgro originally provided two datasets, one for the chemical samples of a site and another for the diatom assemblage of the site.

Dr. Sgro mentioned that not all chemical samples were collected at the same time as the diatom samples, but he matched them as closely as he could. He stated that the diatom samples collected represented about a year’s estimate of diatoms. Therefore, hypothetically, the chemical samples did not have to be collected at the same time. We ended up only using the chemical dataset for our analysis. Changes were made to the chemical dataset so that it would be suitable for statistical analysis.

The chemical dataset consisted of 71 observations and 42 variables. Each observation contained a sample ID (“Chem” column), a collection date (for chemicals and diatoms) and information about location (latitude and longitude, site, river mile). Each observation also contained measurements of over 30 different chemicals and properties such as total TP, NO$_2$ and Cond.

An additional four columns (BDM, Site, Basin, LWR) were added to the dataset. These columns were added either by modifying existing information in the dataset or by using information provided by Dr. Sgro. The “BDM” column contains the BDM score of each sample. The BDM scores were obtained through a table provided by Dr. Sgro. [[4], 15-16]. We matched up the “Chem” column between our dataset and the table provided by Dr. Sgro to find each sample’s corresponding BDM score.

The “Site” column provided the name of the river or site of the sample. It was created using existing site names in the “Location” column of the dataset. Although the “Location” column contained the site names, it also specified the location within each site. We wanted to isolate each site’s name so that it could be used as a variable for
statistical analysis. Thus the “Site” column was created by extracting information from the “Location” column. Four out of 71 samples did not come from a river. They were near shore samples in bay areas. Two of these samples came from Cedar Point and the other two came from Immergrum. Immergrum is a site off Bay Shore Road, less than a mile west of Maumee Bay State Park. In total, there were 11 sites, four from the Central basin and seven from the Western basin.

The “Basin” column provided the basin (Central or Western) of the sample. It was created using the longitude of the sample. Samples with a longitude greater than or equal to 82.3 were classified as Western basin samples while samples with a longitude less than 82.3 were classified as Central basin samples. One can think of the Western basin as everything west of Cedar Point, including Cedar Point, and the Central basin as everything east of Cedar Point halfway through Pennsylvania. Our Central basin samples only go as far as the Pennsylvania border. In total, there were 39 samples from the Central basin and 32 samples from the Western basin.

Lastly, the “Location Within River” column (LWR) was created using the “River Mile” column (RM). The river mile measures how far upstream a sample is from the harbor. For example, a sample with a RM value of “2” means that the sample was taken two miles from the mouth of the river. Dr. Sgro told us that the RM used for harbor samples (RM values above “1000”) did not correlate to the actual definition of the river mile. It was simply used by the OEPA to indicate whether a sample belonged to a harbor or not. He suggested changing the RM values of all harbor samples to zero since it would give a more accurate relationship between the BDM score of a sample and its RM value. We followed his suggestion and made the appropriate changes. Samples with a RM value below one were classified as “mouth” samples while all other samples were classified as “river” samples. In total, there were 17 “harbor” samples, 17 “mouth” samples and 37 “river” samples.
One of Dr. Sgro’s main concerns was the lack of independence between samples. Samples in close proximity tend to exhibit similar chemical and diatom compositions. These correlations could introduce bias into estimates produced by statistical models. In the following sections, we will address the lack of independence between samples, how it prevents us from performing some analysis, and the types of analyses we did to account for it.

Table 2.1 provides the reader with a row of what the final dataset looks like while Figure 2.2 and Figure 2.3 provides maps of the locations of the samples by basin.

Table 2.1: Sample row from dataset.

<table>
<thead>
<tr>
<th>Chem</th>
<th>Date</th>
<th>Basin</th>
<th>Site</th>
<th>LWR</th>
<th>Cond</th>
<th>TP</th>
<th>NO2</th>
<th>BDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>204188-6/7</td>
<td>6/7/12</td>
<td>West</td>
<td>Immergrum Harbor</td>
<td>504</td>
<td>0.06</td>
<td>0.04</td>
<td>4.08</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2.2: Map of Central basin samples.
Figure 2.3: Map of Western basin samples.
3. THE RELATIONSHIPS AMONG BDM, BASINS, AND RIVERS

Recall that the BDM score is a measure of water quality through the use of sample sediment diatom assemblages. Thus an analysis of the BDM score is simply an analysis of the water quality of our samples. The BDM score has a minimum of one and a maximum of 10. Dr. Sgro conveyed that samples with a BDM score above five could be thought of as samples from a clean water source.

We wanted to know if the water quality differed significantly based on the location of the water sample. Each sample contains two types of geographic classifications (basin and site) and one geomorphic classification (location within river). In this section, we will analyze whether, on average, samples from the Central basin have significantly different BDM scores than samples from the Western basin. We will also analyze whether, on average, sites within the same basin have different BDM scores. We begin by analyzing descriptive statistics and graphs of our dataset. Figure 3.1 provides a graphical display representing the distribution of BDM scores for all samples.

Figure 3.1: Distribution of BDM scores.
The histogram in Figure 3.1 shows that the majority of our samples have a BDM score between two and five. It also shows that around six and 10 samples have a BDM score below two and above five respectively. Table 3.2 provides a numerical summary for the BDM scores.

Table 3.2: Numerical summary of BDM scores for all samples.

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Mean</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.58</td>
<td>2.64</td>
<td>3.80</td>
<td>3.77</td>
<td>4.75</td>
<td>6.72</td>
</tr>
</tbody>
</table>

Table 3.2 shows that the mean BDM score for all samples is 3.77. Also, the value for the third quartile is 4.75. This is slightly below the BDM score of five that Dr. Sgro deems a clean water sample. Thus more than 75% of our samples are not considered clean water samples. Next, we will look at a comparison of BDM scores between samples from the Central basin and Western basin. Again, we would like to know whether the BDM score differs significantly between both basins. Figure 3.3 and Table 3.4 provide boxplots and a numerical summary to compare the BDM scores from both basins.

Figure 3.3: Distribution of BDM scores by basin.
Table 3.4: Numerical summary of BDM score by basin.

<table>
<thead>
<tr>
<th>Basin</th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Mean</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central</td>
<td>1.97</td>
<td>3.71</td>
<td>4.57</td>
<td>4.37</td>
<td>4.88</td>
<td>6.72</td>
</tr>
<tr>
<td>Western</td>
<td>1.58</td>
<td>2.15</td>
<td>2.97</td>
<td>3.05</td>
<td>3.56</td>
<td>5.91</td>
</tr>
</tbody>
</table>

The boxplots and numerical summary show that on average, samples from the Central basin have a higher BDM score. This suggests that on average, the water from the Central basin is cleaner than the water from the Western basin. Next, we will consider the mean BDM scores for all sites.

Table 3.5: Mean BDM score for all sites.

<table>
<thead>
<tr>
<th>Site</th>
<th>Sample size</th>
<th>Mean BDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cedar Point (not river)</td>
<td>2</td>
<td>5.59</td>
</tr>
<tr>
<td>Conneaut</td>
<td>8</td>
<td>4.88</td>
</tr>
<tr>
<td>Ashtabula</td>
<td>15</td>
<td>4.79</td>
</tr>
<tr>
<td>Black</td>
<td>10</td>
<td>3.79</td>
</tr>
<tr>
<td>Arcola</td>
<td>1</td>
<td>3.74</td>
</tr>
<tr>
<td>Immergrum (not river)</td>
<td>2</td>
<td>3.71</td>
</tr>
<tr>
<td>Grand</td>
<td>5</td>
<td>3.56</td>
</tr>
<tr>
<td>Portage</td>
<td>6</td>
<td>3.30</td>
</tr>
<tr>
<td>Sandusky</td>
<td>11</td>
<td>2.87</td>
</tr>
<tr>
<td>Old Woman</td>
<td>3</td>
<td>2.70</td>
</tr>
<tr>
<td>Maumee</td>
<td>8</td>
<td>2.44</td>
</tr>
</tbody>
</table>

Table 3.5 shows that on average, samples from Cedar Point have the highest BDM scores while samples from Maumee have the lowest BDM score. This suggests that on average, sites within the same basin can have substantially different BDM scores since Cedar Point and Maumee both belong to the Western basin. However, we took this result with a grain of salt given the small sample size from Cedar Point.

So far the data suggest that BDM scores do differ between basins and between sites within the same basin. In the next section, we will describe the statistical model used to analyze whether the BDM scores differs between basins and sites. We will also discuss
the results of our analysis, accounting for potential dependence among observations taken from the same or neighboring locations.
NESTED ANALYSIS OF VARIANCE

A one-way analysis of variance (ANOVA) is a statistical test that compares the mean of a dependent variable for different groups. For example, we would use a one-way ANOVA to compare the average math scores between ninth graders from different schools. A nested ANOVA on the other hand, has more than one level of groupings. It is used to simultaneously compare the mean of a dependent variable for different groups and levels of groupings. For example, if we wanted to compare the average math scores of ninth graders across different states and the average math scores of ninth graders in different schools within the same state, a nested ANOVA would be appropriate because each school can only belong to one state. Thus the school of a student is said to be nested within the state of the student.

In our model, the two levels of groups are the basin and site of a sample. Note that each site belongs to only one basin. Thus the site of a sample is said to be nested within the basin of the sample. The nested ANOVA will test whether, on average, the BDM score differs significantly by basin. It will also test whether, on average, the BDM score differs significantly among sites in the same basin.

In this model, we assumed that samples from different basins and rivers were independent of each other. We understand that the basins are not entirely independent since there is nothing to prevent water from crossing between basins. However, we found that the closest Central and Western basin samples were about seven miles apart. We are assuming that this distance is enough to assume independence between Central and Western basin samples.

We also acknowledge that the water composition of two rivers within the same basin might be correlated. Theoretically, waves could carry the same source of water up both rivers if they are close by. However, we are assuming that samples from different rivers are independent because we do not think that the chemical and diatom composition of a
river is heavily dependent on the chemical and diatom composition of another river. Table 3.6 provides the results of the nested ANOVA model.

Table 3.6: Nested ANOVA, Basins and Locations.

<table>
<thead>
<tr>
<th></th>
<th>Degrees of Freedom</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
<th>F value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basin</td>
<td>1</td>
<td>30.62</td>
<td>30.62</td>
<td>33.52</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Basin: Sites</td>
<td>9</td>
<td>29.64</td>
<td>3.29</td>
<td>3.61</td>
<td>0.001</td>
</tr>
<tr>
<td>Residuals</td>
<td>60</td>
<td>54.80</td>
<td>0.91</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The p-value at the “Basin” level (< 0.001) is statistically significant at the 0.05 alpha level. Thus we can conclude that on average, the BDM score does depend on the choice of basin. As we saw earlier, the Central basin had a mean BDM score of 4.37 and the Western basin had a mean BDM of 3.05. Thus, we can conclude that, on average, samples from the Central basin have a higher BDM score. Therefore our data show that on average, the Central basin is a significantly cleaner source of water than the Western basin.

The p-value at the “Basin: Sites” level (0.001) is also statistically significant at the 0.05 level. Thus we conclude that at least two sites in the same basin have significantly different BDM means from each other. We ran pairwise t-tests comparing BDM means between all sites in the same basin to find out which sites had BDM means that differed significantly from each other. Table 3.7 and 3.8 provide the results of our t-tests.

Table 3.7: T-test p-values for Central basin sites.

<table>
<thead>
<tr>
<th></th>
<th>Ashtabula</th>
<th>Black</th>
<th>Conneaut</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>0.089</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Conneaut</td>
<td>0.709</td>
<td>0.064</td>
<td>-</td>
</tr>
<tr>
<td>Grand</td>
<td>0.046</td>
<td>0.740</td>
<td>0.038</td>
</tr>
</tbody>
</table>
Table 3.8: T-test p-values for Western basin sites.

<table>
<thead>
<tr>
<th></th>
<th>Cedar Point (not river)</th>
<th>Immergrum (not river)</th>
<th>Maumee</th>
<th>Old Woman</th>
<th>Portage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immergrum</td>
<td>0.064</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Maumee</td>
<td>0.011</td>
<td>0.108</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Old Woman</td>
<td>0.029</td>
<td>0.180</td>
<td>0.376</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Portage</td>
<td>0.009</td>
<td>0.511</td>
<td>0.124</td>
<td>0.254</td>
<td>-</td>
</tr>
<tr>
<td>Sandusky</td>
<td>0.006</td>
<td>0.189</td>
<td>0.270</td>
<td>0.635</td>
<td>0.447</td>
</tr>
</tbody>
</table>

Table 3.7 shows that on average, the BDM score does differ significantly between Ashtabula and Grand, and between Conneaut and Grand at the 0.05 alpha level. Recall that Ashtabula, Conneaut and Grand had a mean BDM score of 4.79, 4.88 and 3.56 respectively. Thus, on average, the water quality in Ashtabula and Conneaut is better than the water quality in Grand. Therefore, on average, the water quality of sites does differ significantly in the Central basin.

In Table 3.8, only t-tests involving Cedar Point had a significant p-value. Recall that the BDM mean for Cedar Point was the highest mean (5.59) we found out of all rivers. Therefore, we conclude that on average, Cedar Point’s BDM score is significantly higher than all other sites except Immergrum. Thus, on average, the water quality in Cedar Point is better than the water quality in all other sites except Immergrum. Therefore, we conclude that the water quality of sites does differ significantly in the Western basin. Recall that Cedar Point is not a river. Consequently, none of the rivers in the Western basin have significantly different water quality.

Intuitively, this result makes sense. Cedar Point is a peninsula in Lake Erie. Therefore we would expect its water to be cleaner since it is more removed from pollution sources that affect other rivers. Like Cedar Point, Immergrum is a near shore bay area. Thus, it is more likely to be removed from pollution as well. In conclusion, our analysis shows that water quality does differ significantly between basins and sites.
4. THE PREDICTIVE ABILITY OF TP, COND AND NO$_2$

In his research, Dr. Sgro fit a multiple regression model predicting the BDM score of samples using log$_{10}$(TP), log$_{10}$(Cond), and log$_{10}$(NO$_2$). His model had an adjusted $R^2$ value of 0.64 with a significant p-value. The adjusted $R^2$ tells us how much variability in the dependent variable the independent variables account for. Therefore, based on our data, the total phosphorus, conductivity, and nitrogen oxide accounts for 64% of the variability in the BDM score.

One of the assumptions of a multiple regression model is the independence of observations. In our dataset, we cannot assume we have independence between samples in the same river. The river flows downstream into the mouth of the harbor. Therefore we must assume that the chemical and diatom composition upstream affects the chemical and diatom composition downstream. Consequently, the independence assumption is violated in Dr. Sgro’s multiple regression model. However, Dr. Sgro believes that the adjusted $R^2$ is a measure of the overall quality of the fit of a statistical model. It is not concerned with the contribution of each variable. Therefore, even though some samples have dependence, a strong adjusted $R^2$ could still tell us whether the independent variables as a group are good predictors of the dependent variable.

We wanted to know how reliable the adjusted $R^2$ of 0.64 really was. Can we really claim that in general, log$_{10}$(TP), log$_{10}$(Cond), and log$_{10}$(NO$_2$), account for 64% of the variability in BDM? We wanted to study the sampling distribution of the adjusted $R^2$ to quantify the uncertainty in the estimated adjusted $R^2$, and to provide an interval estimate. Unfortunately, we cannot collect multiple sets of samples to re-analyze. Thus we will quantify the reliability of our model using a resampling technique called bootstrapping.
BOOTSTRAPPING

Bootstrapping is a resampling technique in statistics. It uses the available dataset and creates “new” datasets by sampling, with replacement rows in the current dataset. In our analysis, we create 2000 “new” datasets by bootstrapping our dataset. Then, we calculated the adjusted $R^2$ of our model for each of the 2000 resampled datasets. Finally, we calculated the 95% confidence interval for the adjusted $R^2$ based on the distribution of the recalculated adjusted $R^2$ values. Figure 4.1 and Table 4.2 provide a histogram and a numerical summary for the resampled adjusted $R^2$ values.

Figure 4.1: Distribution of all resampled adjusted $R^2$ values.

Table 4.2: Numerical summary for all resampled adjusted $R^2$ values.

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.41</td>
<td>0.61</td>
<td>0.66</td>
<td>0.70</td>
<td>0.83</td>
</tr>
</tbody>
</table>
Our histogram and numerical summary indicate that the majority of our resampled adjusted \( R^2 \) values fall between 0.6 and 0.7. The closer these adjusted \( R^2 \) are to each other, the more reliable our model is. Table 4.3 provides the 95% confidence interval for the adjusted \( R^2 \).

Table 4.3: 95% confidence interval for the adjusted \( R^2 \).

<table>
<thead>
<tr>
<th>Mean</th>
<th>Std deviation</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.65</td>
<td>0.07</td>
<td>(0.52, 0.77)</td>
</tr>
</tbody>
</table>

The 95% confidence interval from 0.52 to 0.77 tells us that the variability of BDM accounted for by \( \log_{10}(TP) \), \( \log_{10}(Cond) \), and \( \log_{10}(NO_2) \) fluctuates between 52% and 77%. Therefore there is evidence that as a group, TP, Cond and NO\(_2\) accounts for a substantial amount of variability of the BDM score. Dr. Sgro conveyed that any percentage of accounted variability above 40% could be considered a substantial amount of accounted variability.

The next question of interest is whether the predictive ability of the model varies significantly in different basins. We split our samples into two datasets by their respective basins and used the same method to calculate bootstrap 95% confidence intervals for the adjusted \( R^2 \) for each basin. Figure 4.4 provides boxplots of the distribution of all resampled adjusted \( R^2 \) values by basin. Table 4.5 provides a comparison of the 95% confidence interval of adjusted \( R^2 \) by sample group.
Figure 4.4: Adjusted $R^2$ distribution by basin.

Table 4.5: 95% confidence interval for adjusted $R^2$ by sample group.

<table>
<thead>
<tr>
<th>Sample group</th>
<th>Sample Size</th>
<th>Mean</th>
<th>Std deviation</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>All samples</td>
<td>71</td>
<td>0.65</td>
<td>0.07</td>
<td>(0.52, 0.77)</td>
</tr>
<tr>
<td>Central Basin</td>
<td>39</td>
<td>0.62</td>
<td>0.09</td>
<td>(0.43, 0.77)</td>
</tr>
<tr>
<td>Western Basin</td>
<td>32</td>
<td>0.54</td>
<td>0.15</td>
<td>(0.22, 0.80)</td>
</tr>
</tbody>
</table>

The width of the 95% confidence interval for the Central and Western basin samples are both larger than the width of the 95% confidence interval when considering all samples. The width of the Western basin’s 95% confidence interval is the widest among all sample groups. Its lower limit is the smallest among all three confidence intervals (0.22) and its upper limit is the largest (0.80). Therefore we conclude that the distribution of the
strength of the relationship is more variable in the Western basin than in the Central basin. This implies that there are other factors of BDM that might fluctuate in the Western basin but do not fluctuate as much in the Central basin.

Dr. Sgro told us that there were a few diatoms in the Western basin that were dominant but did not have indicator values. He could not put them into one of the ten groups that were ranked in terms of total phosphorus optima. This means some samples’ BDM score misrepresent the water quality of the sample. It would explain why the distribution of the strength of the relationship varies more in the Western basin than in the Central basin.

Another plausible hypothesis deals with the amount of phosphorus in both basins. The Western basin is more polluted than the Central basin. Therefore, on average, the Western basin might have a high level of phosphorus. It may be that some of the diatoms in the Western basin were no longer limited by phosphorus. This means that past a certain level of phosphorus concentration, the level of phosphorus would no longer affect the growth rate of these diatoms. This would affect the adjusted $R^2$ value since the BDM score assumes that the amount of phosphorus drives the diatom assemblage. In conclusion, there is evidence to suggest that as a group, $\log_{10}(TP)$, $\log_{10}(Cond)$ and $\log_{10}(NO_2)$ are not reliable predictors of BDM. However, we have yet to determine a more reliable set of predictors for BDM.
5. THE RELATIONSHIP BETWEEN BDM AND THE RIVER MILE

In Section 3, we analyzed whether on average, the BDM score differed significantly between basins and sites. In this section, we will explore the relationship between the BDM score and the location of a sample within a river. We think that analyzing this relationship will help us identify the factors influencing diatom assemblages.

Recall that the RM column told us how far upstream the sample was from the mouth of the river. We classified our samples into three groups based on their RM value: Harbor, Mouth and River. We wanted to know whether on average, the BDM scores differed significantly between groups. We started by comparing the BDM score between groups. Figure 5.1 and Table 5.2 presents boxplots and numerical summaries of the BDM score for all groups.

Figure 5.1: Boxplots of BDM scores by location within river.
Table 5.2: Numerical summary of BDM scores by location within river.

<table>
<thead>
<tr>
<th>Location</th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Mean</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harbor</td>
<td>2.16</td>
<td>4.60</td>
<td>5.12</td>
<td>4.90</td>
<td>5.62</td>
<td>6.73</td>
</tr>
<tr>
<td>Mouth</td>
<td>1.78</td>
<td>3.03</td>
<td>3.74</td>
<td>3.63</td>
<td>4.57</td>
<td>4.80</td>
</tr>
<tr>
<td>River</td>
<td>1.58</td>
<td>2.34</td>
<td>3.03</td>
<td>3.32</td>
<td>4.24</td>
<td>6.03</td>
</tr>
</tbody>
</table>

Figure 5.1 indicates that on average, harbor samples have a higher BDM score than mouth samples and mouth samples have a higher BDM than river samples. This suggests that the river gets dirtier as we go further upstream.

In Section 3, we ran a nested ANOVA because we assumed that samples from different basins and sites were independent of each other. However, we did not assume that samples within the same river were independent of each other. The lack of independence between samples within the same river prevents us from running a one-way ANOVA to test whether, on average, the BDM score differs significantly between all three locations within the river.

We decided to look directly at the relationship between RM and BDM. Since RM and BDM were both numerical variables, we created a scatterplot with the BDM on the y-axis and the RM on the x-axis. The goal was to determine whether there was a relationship between BDM and RM. Recall that the RM value of harbor samples (initially above 1000) were converted to zeroes to better represent the location of a harbor sample in the river. Figure 5.3 presents our scatterplot of BDM versus RM.
This scatterplot is not very informative. The majority of our samples are clustered to the left of the scatterplot while a handful of samples are spread out to the right. In order to have a better understanding of our scatterplot, we created a second scatterplot with BDM plotted against $\log_{10}(RM)$, the log base 10 of RM. This transformation spreads out the samples on the left of the scatterplot but clusters the samples on the right of the scatterplot.

Initially, we ran into a problem because harbor samples had a RM value of zero and the log of zero is undefined. In the end, we decided to remove the harbor samples from our scatterplot. Since all harbor samples had a RM value of zero, it did not accurately
represent the actual location of our harbor samples. Therefore, it would not take away from what we could learn about the relationship between BDM and RM upward of the river mouth. Figure 5.4 is a scatterplot of the BDM score against $\log_{10}(RM)$ for river and mouth samples only.

Figure 5.4: BDM versus $\log_{10}(RM)$ (River and Mouth samples only).
The scatterplot indicates that there is a piecewise relationship between BDM and \( \log_{10}(RM) \). The BDM value increases as \( \log_{10}(RM) \) increases from negative one to zero and then decreases as \( \log_{10}(RM) \) increases from zero to one. Then it increases again as \( \log_{10}(RM) \) goes beyond one. This suggests that at the mouth of the river, the water gets cleaner as we move upstream. However, after moving about a mile upstream, the water gets dirtier until about the 10-mile mark, after which it starts to get cleaner again.

Dr. Sgro reasoned that a lake seiche, a fluctuation in water levels due to strong wind or storms, could carry harbor and estuary waters upstream thus explaining why water a mile upstream would be cleaner. Dr. Sgro also suggested that samples taken 10 miles upstream could have been taken above major sources of pollution. Consequently, the water would get cleaner as we moved further away from the pollution sources. However, we are unsure why there is a sudden drop in BDM score from \( \log_{10}(RM) = 0 \) to \( \log_{10}(RM) = -0.5 \), a mere distance of 0.4 river miles.

We wanted to see whether this piecewise relationship would hold for all rivers. In order to test that assumption, we recreated our scatterplot but this time with different shapes and colors to distinguish different rivers from each other. Figure 5.5 is a scatterplot of our samples grouped by river.
Figure 5.5: BDM versus log$_{10}$(RM) (River and Mouth samples only).

The scatterplot above does not provide conclusive evidence whether this piecewise relationship holds for all rivers. What we see is a lack of representative samples for different RM values from each river. There are no samples from Sandusky with a log$_{10}$(RM) value below 0.25 and none of Ashtabula’s samples have a log$_{10}$(RM) value above 0.5. There are just not enough samples to accurately capture how the water quality changes for each river as we go further upstream.
We asked ourselves how likely was it for this piecewise pattern to happen by chance alone? We did not know how to quantify this likelihood because of the lack of independence between samples but it was very difficult to believe that it did happen by chance alone. Perhaps there really is a piecewise relationship between BDM and \( \log_{10}(RM) \). We hope that in the future, more samples can be collected across all RM values in order to better understand the relationship between the river mile and water quality.
6. CONCLUSION
In Section 2, we gave the reader an overview of the dataset provided by Dr. Sgro. We discussed the original data provided by Dr. Sgro and the changes that were made to make it suitable for statistical analysis. In Section 3, we used a nested ANOVA to test whether on average, the water quality (measured by BDM) of Lake Erie’s near shore habitats differed by basin and sites. The nested ANOVA model showed us that on average, the BDM score did differ by basin. By comparing the mean BDM for both basins, we found that the mean BDM of the Central basin was significantly higher than the mean BDM of the Western basin. Therefore, we concluded that on average, the water quality of the Central basin was cleaner than the water quality of the Western basin.

The nested ANOVA also showed us that on average, the BDM score did differ by sites within the same basin. Using t-tests, we compared the BDM scores of all sites within the same basin. In the Central basin, we found that on average, the BDM score of the Ashtabula and Conneaut Rivers were significantly higher than the BDM score of the Grand River. Therefore, we concluded that on average, the water quality of the Ashtabula and Conneaut Rivers was cleaner than the water quality of the Grand River. In the Western basin, we found that on average, the BDM score of Cedar Point was significantly higher than the BDM score of all other Western basin sites except Immergrum. Therefore, we concluded that on average, the water quality of Cedar Point was cleaner than all other sites in the Western basin except Immergrum.

In Section 4, we analyzed the multiple regression model provided by Dr. Sgro to quantify the uncertainty in the estimated adjusted $R^2$ of the model. Using bootstrapping, we constructed a 95% confidence interval for the adjusted $R^2$ and found that the variability of BDM accounted for by $\log_{10}(TP)$, $\log_{10}(Cond)$ and $\log_{10}(NO_2)$ fluctuated between 52% and 77%. Then, we separated our samples into their respective basins and constructed a 95% confidence interval for the adjusted $R^2$ for each basin. We found that the variability of BDM accounted for by $\log_{10}(TP)$, $\log_{10}(Cond)$ and $\log_{10}(NO_2)$ fluctuated between 43% and 77% in the Central basin and the variability of BDM accounted for by $\log_{10}(TP)$,
log_{10}(\text{Cond}) and log_{10}(\text{NO}_2) fluctuated between 22\% and 80\% in the Western basin. Therefore, we concluded that by log_{10}(\text{TP}), log_{10}(\text{Cond}) and log_{10}(\text{NO}_2) were not reliable predictors of BDM due to the large fluctuations of variability accounted for by these predictor variables in the Western basin.

As mentioned, Dr. Sgro conveyed that there were a few dominant diatoms in the Western basin that could not be ranked in terms of their total phosphorus optima. That meant that the BDM score of certain Western basin samples might have inaccurately represented the water quality of the sample. We hope that further progress can be made to accurately rank these diatoms. It would be interesting to see how the variability of the adjusted R^2 changes if these diatoms were ranked accurately.

We also hypothesized about the growth rate of certain diatoms in the Western basin leveling off past a certain level of phosphorus. This would also affect the adjusted R^2 since the BDM score assumes that the amount of phosphorus drives the diatom assemblage. We concluded that more work needs to be done to determine a more reliable set of predictors for BDM.

Dr. Sgro also believes that metals such as zinc (Zn), copper (Cu) and cadmium (Cd) might affect diatoms. He wants to know whether there are enough of these metals to provoke a response from diatoms. The metals Zn, Cu and Cd are readily available in our dataset. However, we did not analyze their relationships with BDM due to time constraints. One could begin by examining a simple linear model between all three metals and BDM to see if there is a correlation between these metals and BDM. One could also add these metals to the multiple regression model and check if the 95\% confidence interval of the adjusted R^2 would change from what we saw earlier. Both these methods could be starting points to analyze the relationship between BDM and these metals.

In Section 5, we found that there was a piecewise relationship between BDM and log_{10}(\text{RM}). The BDM score increased as log_{10}(\text{RM}) increased from -0.5 to zero and then
decreased as $\log_{10}(RM)$ increased from zero to one. However, the size of our dataset did not allow us to assess whether this piecewise relationship held for all rivers. This prevented us from making conclusive claims about the relationship between BDM and RM. We hope that the OEPA will make efforts to collect a larger dataset with equal and ample representation from all levels of RM for all rivers. If we had a larger dataset, we could analyze whether the piecewise relationship between BDM and $\log_{10}(RM)$ held for all rivers. This could help determine where significant sources of pollution occur in each river. Consequently, we could identify factors that influence BDM and help improve our multiple regression model.

Lastly, in Section 3, we also showed that 75% of our samples had a BDM below five, the level Dr. Sgro considers to be a clean water source. This suggests that more work is needed to restore the near shore habitats of Lake Erie. It is important for residents of Ohio to see themselves as caretakers of Lake Erie as Lake Erie supplies them with an abundance of ecological and economical opportunities. We hope that the development of the BDM and other measures of near shore habitats’ health will help focus efforts to restore Lake Erie’s near shore habitats with greater efficiency.

We want to thank Dr. Sgro for allowing us to collaborate with him on this project. His guidance, patience and willingness to help made this project more thought provoking, exciting and fulfilling. We hope this paper has provided the reader with an understanding of the distribution of water quality in the Central and Western basins of Lake Erie as well as the challenges to assess the factors that influence the diatom assemblages of Lake Erie’s near shore habitats.
7. REFERENCES

<http://greatlakesrestoration.us/actionplan/index.html>


