# Template for Thesis Proposal

Feng Gu

July 31, 2025

## Contents

# 1 Preliminary exploration

In this section, I implemented a simplified framework of Jian's analysis of the ZCI-stress score relationship using the quantile regression.

1. **Collect comparable data.**

   I currently have three datasets: zoobenthic community data (311×16), chemical data (104×30), and environmental data (289×7). These datasets were merged by station ID, resulting in a combined dataset with 104 rows and 53 columns, containing all three types of data. Column indices were structured by data type to improve readability and consistency for future processing.

2. **Pre-process taxa data.**

   The zoobenthic dataset has issues such as small sample size and potential outliers. I used the IQR method to detect outliers and then applied octave transformation, as suggested in Jian's analysis, to reduce their impact. The transformed data showed fewer outliers and a more even distribution.

3. **Assess sediment contamination.**

   Instead of standardization, I applied log-transformation to the chemical data to reduce dominance by high-value variables. Then I conducted PCA and selected principal components based on variance explained, pollutant specificity, and balanced loadings. These selected components were normalized and summed (with attention to loading directions) to produce a composite "SumReal" score reflecting sediment contamination. This score was added to the dataset as an indicator of stress level.

4. **Identify reference and degraded sites.**

   Sites were classified by their stress scores. The bottom 20% were considered minimally disturbed (reference sites), and the top 20

5. **Cluster reference sites by community composition.**

   Since taxa composition varies even among reference sites, clustering was applied to identify dominant community patterns under undisturbed conditions. These clusters represent "normal" taxa structures, each likely shaped by distinct environmental conditions.

6. **Build a discriminant model for habitat classification.**

   A discriminant model was trained to predict cluster membership using environmental variables from reference sites. This model was applied to all sites, assigning each to one of the community clusters. This setup allows comparisons between reference and degraded sites within the same habitat type to define "best" and "worst" endpoints.

7. **Construct endpoints and compute ZCI via ordination.**

   Within each cluster, endpoints were constructed using mean taxa abundances from the most reference-like and most degraded sites. These endpoints were used in Bray-Curtis ordination to assign scores to each site, scaled between 0 (worst) and 1 (best) to compute the Zoobenthic Condition Index (ZCI). I do not yet fully understand the ordination method used, and I plan to explore whether a vector-based ZCI might provide a more accurate assessment.

8. **Evaluate the ZCI vs SumRel relationship by quantile regression.**

   I plotted ZCI against SumRel and applied quantile regression to examine their relationship.

## 1.1 Collect comparable data

Currently, there are three available datasets(matrices): zoobenthic community data(311 by 16), chemical data(104 by 30) and environmental data(289 by 7). These metrices can be merged by the station ID, which gives a comparable dataset with 104 rows and 53 columns.

After merging the three data set on station IDs that are common to all three set, there is completed matrix with **104 rows and 53 columns** from the three variable types. [1]

[1] The three types(chemical, environmental, taxa) were denoted and stored as a higher column index in the pandas dataframe, which makes the table more readable. Later, the results from upcoming data operations that apply on across all rows will also be stored in this way.

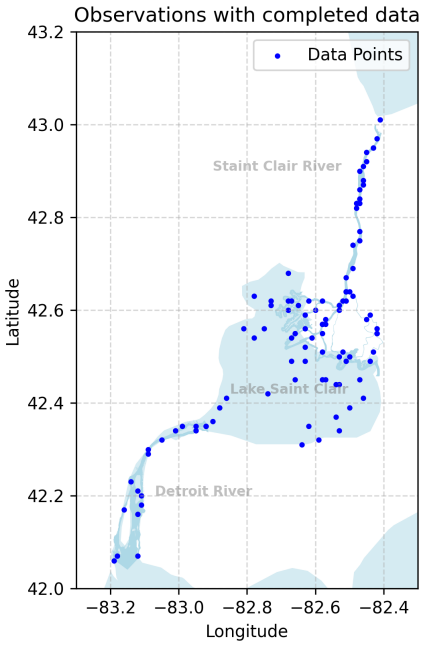


Figure 1: Distribution of the observations with completed three types of data.

## 1.2 Pre-process taxa data

Small sample size and poor data quality are two common and unignorable issues in the zoobenthic dataset, outliers in the data needs to be transformed or removed to improve the data quality, which helps to eliminate the dominance of outliers in later analysis.

In this stage, I first explored the distribution of the taxa data and detected the outliers by using IQR method. After confirming the existence of outliers, octave transformation was applied on the taxa data, as suggested in Jian's analysis, to reduce the impact of outliers and balance the influence of all taxons.

The octave transformation is applied by the following formula, where the proportion of a taxa of a site it the proportion of the taxa in the total density of all taxa in the site.

$$x_{\text{octave transformed}} = \log_2(100 \times (\text{proportion of taxa} + 0.01))$$

After the transformation, I checked the distribution and outliers again, and there were less outliers and flatter distribution of the taxa data.

**—Mathematical coverage of the operation at this stage—**

Let $\mathbf{Y} \in \mathbb{R}^{n \times t}$ denote the raw taxa abundance matrix, where $n$ is the number of sites and $t$ is the number of taxa.

**Outlier detection:** For each taxon $j$, compute the interquartile range (IQR) of the abundances $\{Y_{ij}\}_{i=1}^{n}$ and identify outliers as values outside $[Q_1 - 1.5 \cdot \text{IQR}, Q_3 + 1.5 \cdot \text{IQR}]$, where $Q_1$ and $Q_3$ are the first and third quartiles.

**Octave transformation:** For each site $i$ and taxon $j$, compute the proportion $p_{ij} = \frac{Y_{ij}}{\sum_{k=1}^{t} Y_{ik}}$. The octave-transformed value is:

$$X_{ij} = \log_2 (100 \cdot (p_{ij} + 0.01))$$

where $X_{ij}$ is the transformed abundance for site $i$, taxon $j$.

After transformation, the distribution of taxa abundances is more balanced and outliers are reduced.

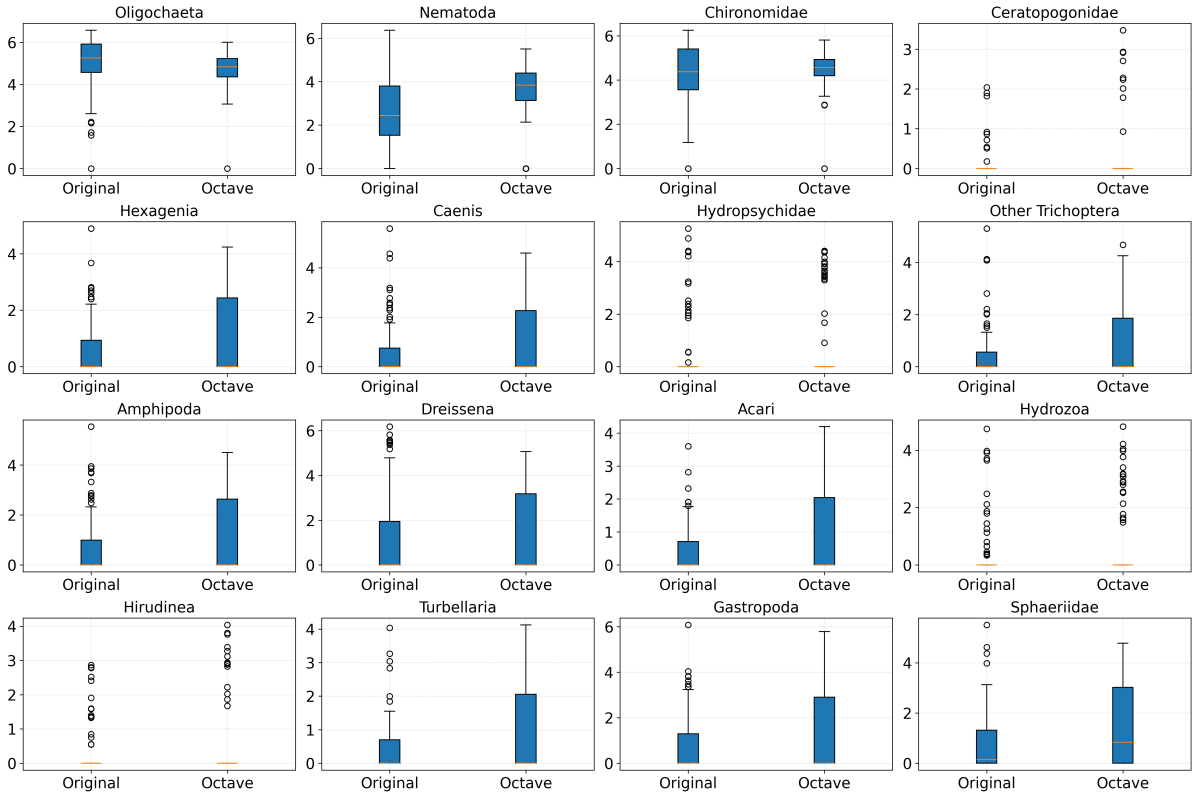


Figure 2: Boxplots of the octave-transformed taxa data, showing reduced outliers and more balanced distribution.

## 1.3 Assess sediment contamination

The sediment contamination is assessed by applying PCA on the chemical data. In this work, instead of standardizing the chemical data, log-transformation was applied on the chemical data to reduce the dominance of some chemical variables that have large values. Then, PCA was applied on the transformed chemical data and the distribution of loadings across the chemical variables was visualized by histogram for each principal component for later selecting step.

Not all PCs are used to compute the stress score, there are three criteria to select the suitable PCs:

- Selected PCs should have a relatively high proportion of variance explained(high eigenvalue).
- Selected PCs should have a high loading on the chemical variables that are pollutants and rarely sourced from nature.
- Selected PCs should avoid the counteracting effect due to uniform distributed positive and negative loadings across the chemical variables.

After applying the above criteria, 'PC1', 'PC2', 'PC3', 'PC5', 'PC6', 'PC7', 'PC9' are selected as the suitable PCs.

Based on the selected PCs, I normalized these PCs to the range [0, 1] to eliminate the effect of differing scales (due to their eigenvalues), reflecting the real-world situation that the toxicity of chemical elements is not directly comparable and not always proportional to their concentrations. Subsequently, I summed the normalized PCs, considering the directionality of positive and negative loadings, to obtain the "SumReal" score[2] for each sample, which serves as a measure of sediment contamination.

After this step, each observation in the dataset has a "stress score"(which is the "SumReal" score as Jian defined in her thesis), the higher the score, the more contaminated the sediment is.

[2]As done by Jian in her sediment contamination assessment.

**—Mathematical coverage of the operation at this stage—**

Mathematically, let $\mathbf{X} \in \mathbb{R}^{n \times p}$ denote the log-transformed chemical data matrix, where $n$ is the number of samples and $p$ is the number of chemical variables. Principal component analysis (PCA) is applied to $\mathbf{X}$ to obtain principal components:

$$\mathbf{Z} = \mathbf{X}\mathbf{W}$$

where $\mathbf{W} \in \mathbb{R}^{p \times k}$ contains the loadings of the selected $k$ principal components via the above criteria. Let $\mathbf{z}_j$ denote the $j$-th selected principal component (column of $\mathbf{Z}$). Each $\mathbf{z}_j$ is normalized to $[0, 1]$:

$$\tilde{\mathbf{z}}_j = \frac{\mathbf{z}_j - \min(\mathbf{z}_j)}{\max(\mathbf{z}_j) - \min(\mathbf{z}_j)}$$

The overall sediment contamination score ("SumReal") for sample $i$ is then computed as:

$$\text{SumReal}_i/(\text{stress score})_i = \sum_{j=1}^{k} \alpha_j \tilde{z}_{ij}$$

where $\alpha_j$ reflects the directionality (sign) and importance of each PC, to each component $z_j$ it should be set to 1 for PCs mainly with positive loadings on pollutants, and -1 for PCs mainly with negative loadings on pollutants.
Thus, each sample receives a "stress score" $\text{SumReal}_i$ that quantifies sediment contamination based on the selected and normalized principal components.

## 1.4 Identify reference and degraded sites

Based on the "SumReal" score, reference and degraded sites were identified by selecting the lowest 20% and highest 20% of the "stress score", respectively. There selected reference sites will be viewed as minimally disturbed sites, and their taxa compositions will be though as determined completely by the environmental factors.

Therefore, these references will support build a purely tidy model that predicts what a pristine taxa composition should be given a specific set of environmental conditions (if the distribution of the environmental factors is uniform across all possible values or fit to the real world distribution).

**—Mathematical coverage of the operation at this stage—**

Mathematically, let $S_i$ denote the "SumReal" stress score for site $i$ ($i = 1, \ldots, n$). Define the $q$-th quantile of the stress scores as $Q_q(S)$. The sets of reference sites ($\mathcal{R}$) and degraded sites ($\mathcal{D}$) are then defined as:

$$\mathcal{R} = \{i : S_i \leq Q_{0.2}(S)\}, \qquad \mathcal{D} = \{i : S_i \geq Q_{0.8}(S)\}$$

where $Q_{0.2}(S)$ and $Q_{0.8}(S)$ are the 20th and 80th percentiles of the stress scores, respectively. Sites in $\mathcal{R}$ are considered minimally disturbed (reference), while sites in $\mathcal{D}$ are considered highly degraded.

## 1.5 Cluster reference sites by community composition

Given the fact that taxa composition still varies across the reference sites due to the various environmental conditions, applying clustering on the reference sites can help us to identify the major patterns of the taxa composition under the minimal disturbance level.

This clustering results will tell us the major groups of taxa compositions, which should correspond to major types of environmental conditions that help shape the taxa compositions into these groups. A taxa composition falling into one of these groups with the same or similar environmental conditions should be viewed as an approximately "normal" taxa composition, which means it may be minimally disturbed by human activities.

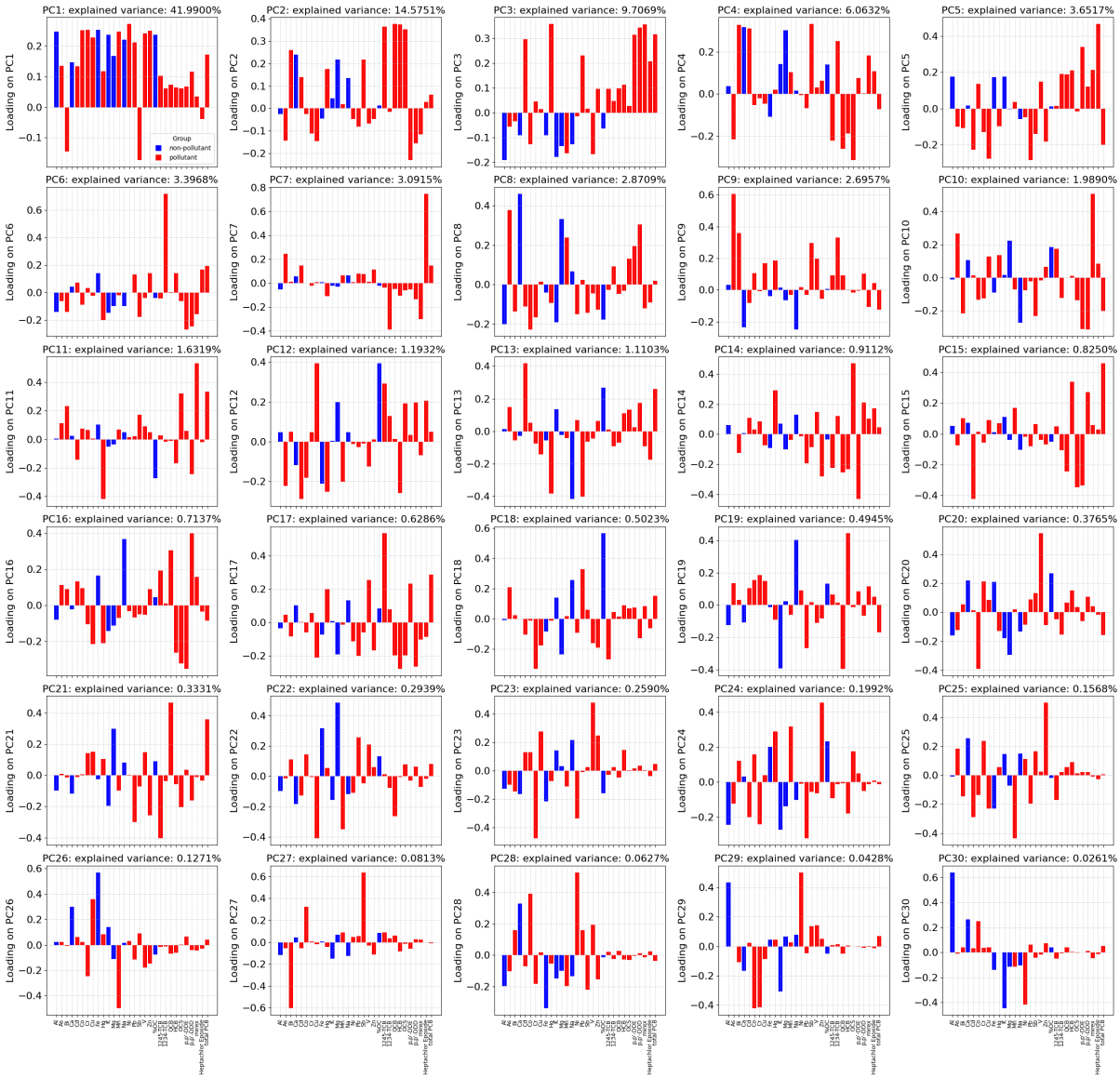


Figure 3: PCA loadings of the chemical data, showing the distribution of loadings across chemical variables.

**—Mathematical coverage of the operation at this stage—**

Let $\mathcal{R}$ denote the set of reference sites identified previously. For each site $i \in \mathcal{R}$, let $\mathbf{y}_i \in \mathbb{R}^t$ represent its taxa composition vector, where $t$ is the number of taxa.
To identify major patterns among reference sites, we apply a clustering algorithm (e.g., $k$-means or hierarchical clustering) to the set $\{\mathbf{y}_i : i \in \mathcal{R}\}$. The goal is to partition the reference sites into $K$ clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$ such that:

$$\mathcal{R} = \bigcup_{k=1}^{K} \mathcal{C}_k, \qquad \mathcal{C}_k \cap \mathcal{C}_l = \emptyset \text{ for } k \neq l$$

and sites within each cluster $\mathcal{C}_k$ have similar taxa compositions.
The resulting clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$ represent the major types of minimally disturbed taxa compositions, which can be further analyzed in relation to environmental variables.

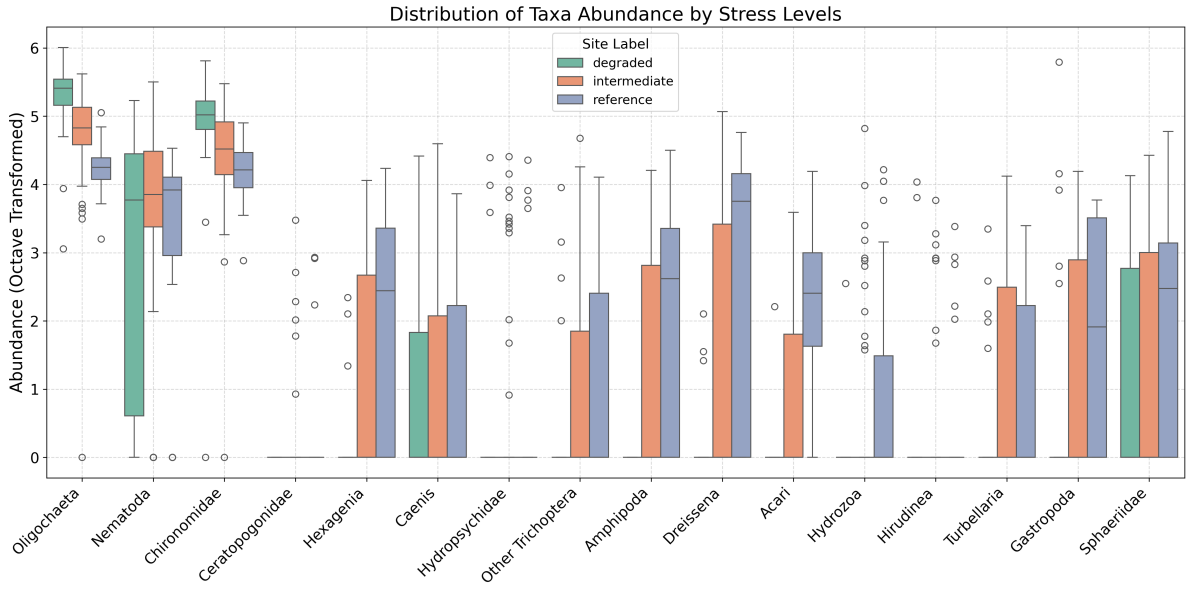


Figure 4: Taxa distribution by site label, showing the potential taxa composition differences across different level of sediment contamination.

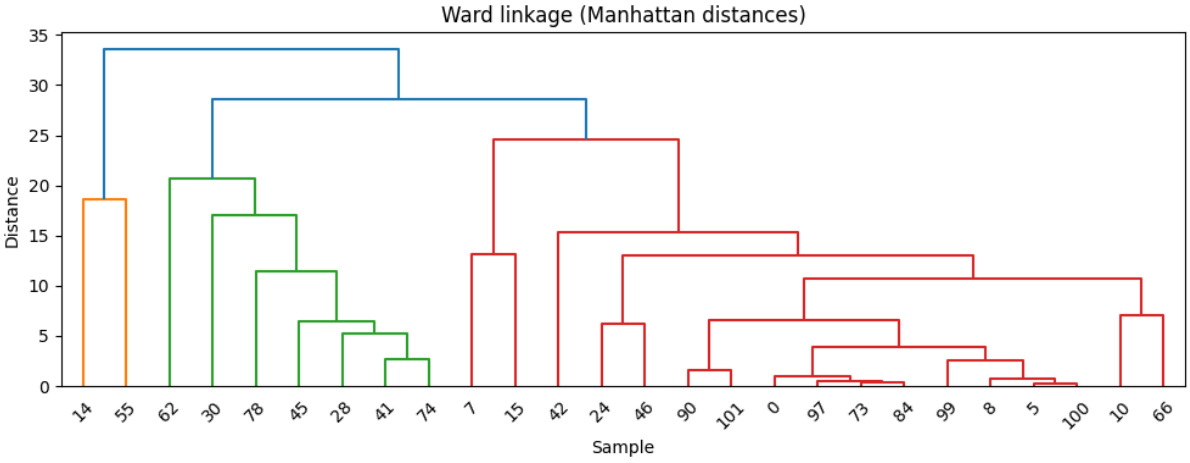


Figure 5: Clustering results of reference sites by taxa composition, showing the major groups of minimally disturbed taxa compositions.

## 1.6 Build a discriminant model for habitat classification.

At this stage, a purely tidy model that predicts the taxa composition with given environmental conditions is built. Discriminant function(classification function) was fitted on these reference sites to predict cluster memberships(labels) given their environmental conditions.

Because the response variable is categorical, applying the fitted discriminant function on all sites given their environmental conditions will group the sites into the selected taxa clusters. There are reference and degraded sites within each cluster, they can be used to construct the theoretical *best* and *worst* individual in aspects of taxa compositions and stress scores, which will be used for ordination method to get ZCI scores later.

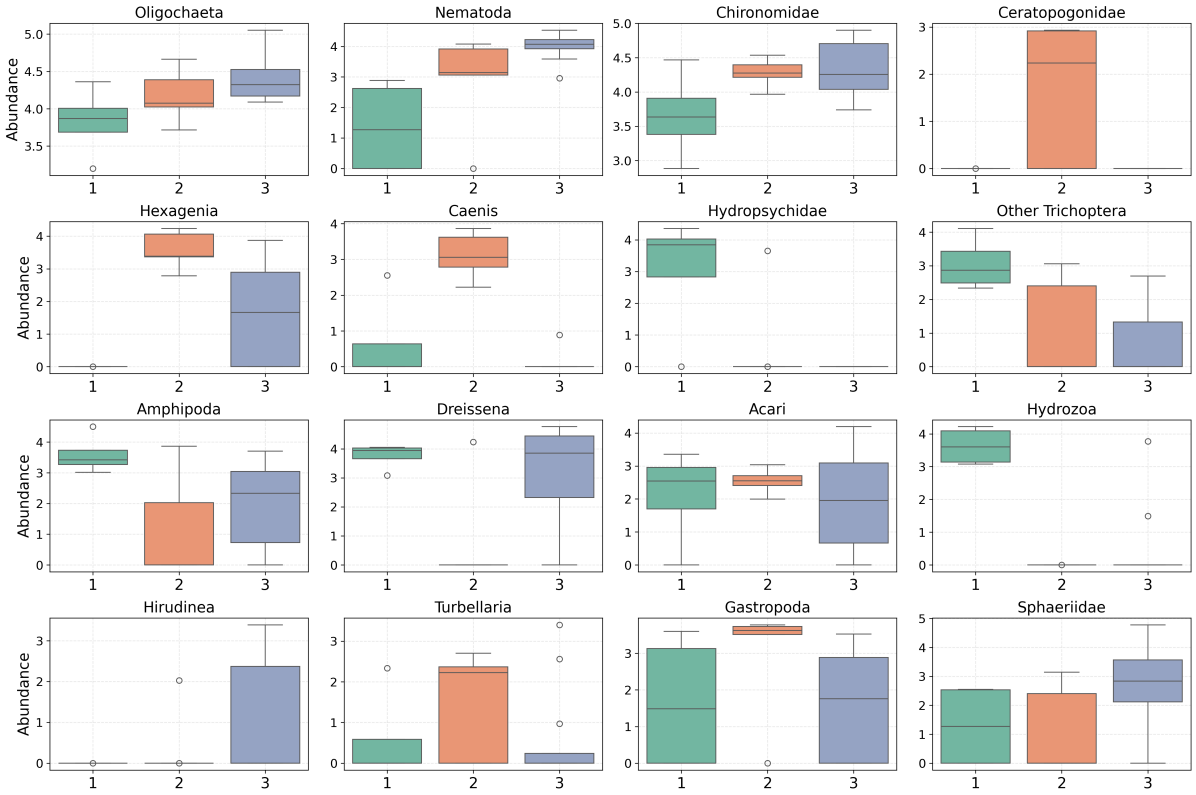


Figure 6: Taxa abundance by cluster, showing the distribution of taxa across different clusters of reference sites.

**—Mathematical coverage of the operation at this stage—**

Let $\mathcal{C}_1, \ldots, \mathcal{C}_K$ denote the clusters of reference sites identified by taxa composition. For each site $i$, let $\mathbf{e}_i \in \mathbb{R}^d$ be its vector of environmental variables, and let $c_i \in \{1, \ldots, K\}$ be its cluster label. A discriminant function $f : \mathbb{R}^d \to \{1, \ldots, K\}$ is trained on the reference sites $\{(\mathbf{e}_i, c_i) : i \in \mathcal{R}\}$ to predict cluster membership from environmental variables:

$$\hat{c}_i = f(\mathbf{e}_i)$$

Applying $f$ to all sites assigns each site to a predicted cluster:

$$\forall i \in \{1, \ldots, n\}, \quad \hat{c}_i = f(\mathbf{e}_i)$$

Within each predicted cluster, sites can be further analyzed to compare reference and degraded sites, enabling the construction of theoretical "best" (reference) and "worst" (degraded) taxa compositions and associated stress scores for each habitat type.

## 1.7 Construct endpoints and compute ZCI via ordination.

After all sites have been assigned to clusters, endpoints need to be constructed within each cluster to construct the "standard" taxa compositions under the lowest(*best individual*) and highest(*worst individual*) stress levels.

The **mean abundance** of taxa in a small subset, like 3 to 5, of the reference sites with the lowest "stress score" can be used as the *best* endpoint, and vice versa, the **mean abundance** of a small subset of the degraded sites with the highest "stress score" can be used as the *worst* endpoint.

The *best* and *worst* endpoints are used to perform Bray-Curtis ordination, which is a method to scale the taxa compositions under the lowest and highest stress levels. The ordination results provide scores

Table 1: Discriminant Coefficients

| | $Class_1$ | $Class_2$ | $Class_3$ |
|---|---|---|---|
| Intercept | -121.154 | -28.769 | 18.883 |
| Total Organic Carbon (LOI %) | 0.394 | -0.429 | 0.156 |
| Water Depth (m) | -0.114 | 0.110 | -0.039 |
| Water Temperature | 4.411 | 0.927 | -0.712 |
| Dissolved Oxygen Concentration (mg/L) | 1.807 | 0.722 | -0.439 |
| Median Particle Size (Phi) | 2.276 | 1.229 | -0.689 |

Table 2: Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.00 | 0.00 | 0.00 | 1 |
| 2 | 0.00 | 0.00 | 0.00 | 1 |
| 3 | 0.60 | 1.00 | 0.75 | 3 |
| accuracy | na | na | 0.60 | 5 |
| macro avg | 0.20 | 0.33 | 0.25 | 5 |
| weighted avg | 0.36 | 0.60 | 0.45 | 5 |

for each site in each cluster, and these scores are converted into range [0, 1] to be used as Zoobenthic Condition Index, where the degraed endpoint as 0 and the reference endpoint as 1.

**—Mathematical coverage of the operation at this stage—**

Mathematically, for each cluster $k$, let $\mathcal{R}_k$ and $\mathcal{D}_k$ denote the sets of reference and degraded sites in cluster $k$, respectively. Define the *best* endpoint $\mathbf{y}_{\text{best}}^{(k)}$ as the mean taxa composition of the $m$ reference sites in $\mathcal{R}_k$ with the lowest stress scores, and the *worst* endpoint $\mathbf{y}_{\text{worst}}^{(k)}$ as the mean of the $m$ degraded sites in $\mathcal{D}_k$ with the highest stress scores:

$$\mathbf{y}_{\text{best}}^{(k)} = \frac{1}{m} \sum_{i \in \mathcal{B}_k} \mathbf{y}_i, \qquad \mathbf{y}_{\text{worst}}^{(k)} = \frac{1}{m} \sum_{i \in \mathcal{W}_k} \mathbf{y}_i$$

where $\mathcal{B}_k \subset \mathcal{R}_k$ and $\mathcal{W}_k \subset \mathcal{D}_k$ are the selected subsets.
For each site $i$ in cluster $k$, compute the Bray-Curtis dissimilarity to both endpoints:

$$d_{\text{best}}(i) = \text{BC}(\mathbf{y}_i, \mathbf{y}_{\text{best}}^{(k)}), \qquad d_{\text{worst}}(i) = \text{BC}(\mathbf{y}_i, \mathbf{y}_{\text{worst}}^{(k)})$$

where $\text{BC}(\cdot, \cdot)$ denotes the Bray-Curtis dissimilarity.
The Zoobenthic Condition Index (ZCI) for site $i$ is then scaled to $[0, 1]$:

$$\text{ZCI}_i = \frac{d_{\text{worst}}(i)}{d_{\text{best}}(i) + d_{\text{worst}}(i)}$$

so that $\text{ZCI}_i = 1$ at the best endpoint and $\text{ZCI}_i = 0$ at the worst endpoint.

**Honestly, I did not what is the ordination method here** and why it can compress the 16 taxa variables into a single value to reflect the distance to the endpoints, I will check this method later. Now, I just asked GPT to provide me code to implement this method and receive the zoobenthic composition scores for each site and convert them into valued-based ZCI scores.

**But intuitively, I think there might be vector-based ZCI scores** that can be computed for each site and compared with a set of reference sites, which should provide a more accurate assessment of the zoobenthic condition than a value-based ZCI.

Table 3: Best and Worst Taxa Compositions by Pattern

| Endpoint | **Best** | | | **Worst** | | |
|---|---|---|---|---|---|---|
| taxa pattern | 1 | 2 | 3 | 1 | 2 | 3 |
| Oligochaeta | 3.645 | 4.376 | 4.448 | 5.513 | 4.884 | 5.169 |
| Nematoda | 1.807 | 2.664 | 4.224 | 4.659 | 4.418 | 4.524 |
| Chironomidae | 3.385 | 4.382 | 4.244 | 4.871 | 4.286 | 4.827 |
| Ceratopogonidae | 0.000 | 1.723 | 0.000 | 0.000 | 0.000 | 0.000 |
| Hexagenia | 0.000 | 3.411 | 0.850 | 0.000 | 0.000 | 0.000 |
| Caenis | 0.000 | 3.423 | 0.000 | 0.000 | 1.472 | 0.000 |
| Hydropsychidae | 4.013 | 1.217 | 0.000 | 0.000 | 0.000 | 0.000 |
| Other Trichoptera | 2.691 | 0.802 | 0.965 | 0.000 | 0.668 | 0.000 |
| Amphipoda | 3.666 | 0.676 | 2.363 | 0.000 | 0.000 | 0.000 |
| Dreissena | 3.976 | 0.000 | 3.034 | 0.000 | 0.473 | 0.000 |
| Acari | 1.697 | 2.716 | 1.000 | 0.000 | 0.000 | 0.000 |
| Hydrozoa | 3.808 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Hirudinea | 0.000 | 0.676 | 0.943 | 0.000 | 1.269 | 0.000 |
| Turbellaria | 0.779 | 1.691 | 1.132 | 0.000 | 0.000 | 0.862 |
| Gastropoda | 2.188 | 2.500 | 1.956 | 0.000 | 1.386 | 1.307 |
| Sphaeriidae | 0.845 | 0.802 | 2.533 | 0.000 | 0.000 | 2.833 |
| stress level | 4.255 | 3.746 | 3.658 | 2.745 | 2.698 | 2.902 |
| ZCI | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 | 0.000 |
| taxa pattern | 1.000 | 2.000 | 3.000 | 1.000 | 2.000 | 3.000 |

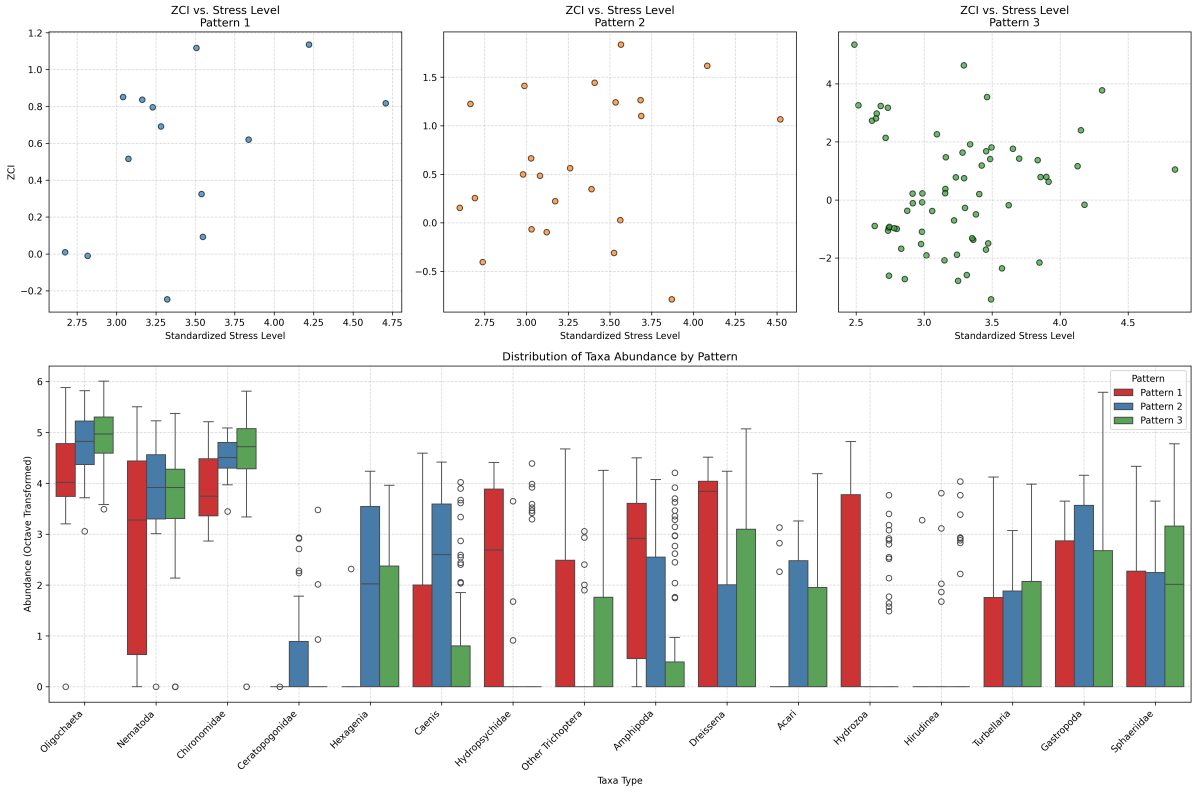


Figure 7: ZCI vs Stress Scores and the Distribution of Taxa Abundance across Taxa Patterns

## 1.8 Evaluate the ZCI vs SumRel relationship by quantile regression

To this step, within each cluster, I have had the ZCI scores and stress scores for each site, and they are value-based scores so that simple linear or quantile regression can be applied to them.

A piecewise quantile regression was applied to the ZCI scores and stress scores within each cluster to

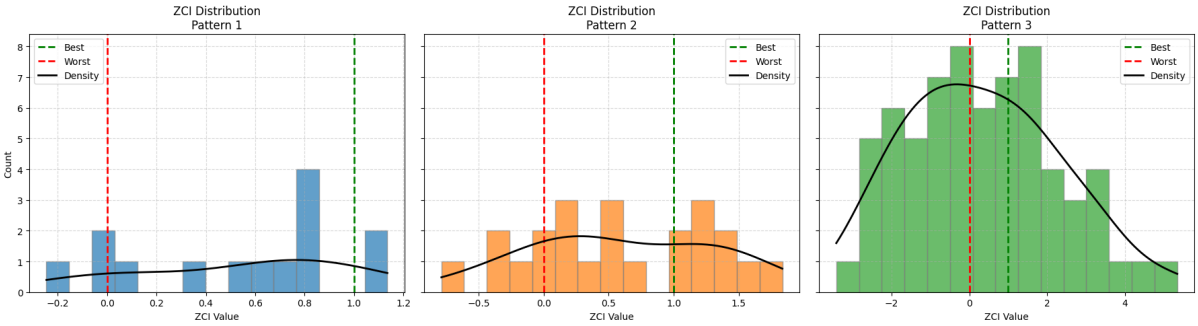


Figure 8: ZCI Distribution across Clusters, showing the variation of ZCI scores within each cluster.

evaluate the relationship between them, a bias-corrected bootstrapping method was applied to estimate the confidence intervals of the piecewise quantile regression coefficients.

**—Mathematical coverage of the operation at this stage—**

Let $\text{ZCI}_i$ denote the Zoobenthic Condition Index for site $i$, and let $\text{SumRel}_i$ denote the corresponding stress score. For each cluster $k$, we fit a quantile regression model to the data $\{(\text{SumRel}_i, \text{ZCI}_i) : i \in \mathcal{C}_k\}$:

$$Q_\tau(\text{ZCI}|\text{SumRel}) = f_\tau(\text{SumRel})$$

where $Q_\tau(\cdot|\cdot)$ denotes the $\tau$-th quantile of ZCI given SumRel, and $f_\tau$ is a piecewise linear function estimated via quantile regression.
The coefficients of $f_\tau$ are estimated using a bias-corrected bootstrap method to obtain confidence intervals for the quantile regression estimates.

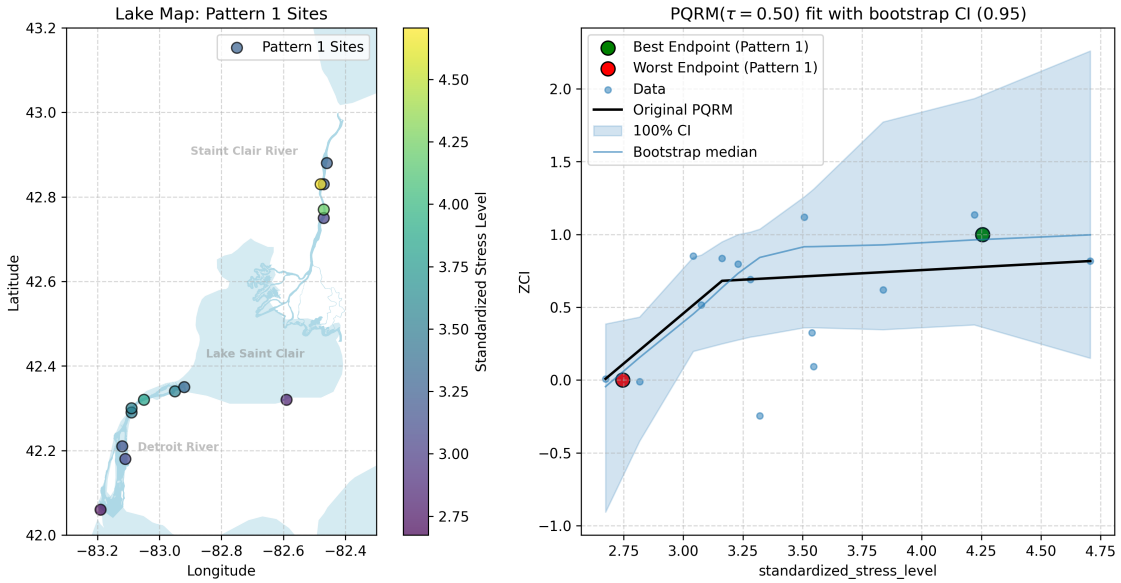


Figure 9: Quantile regression results of ZCI vs Stress Scores for cluster 1

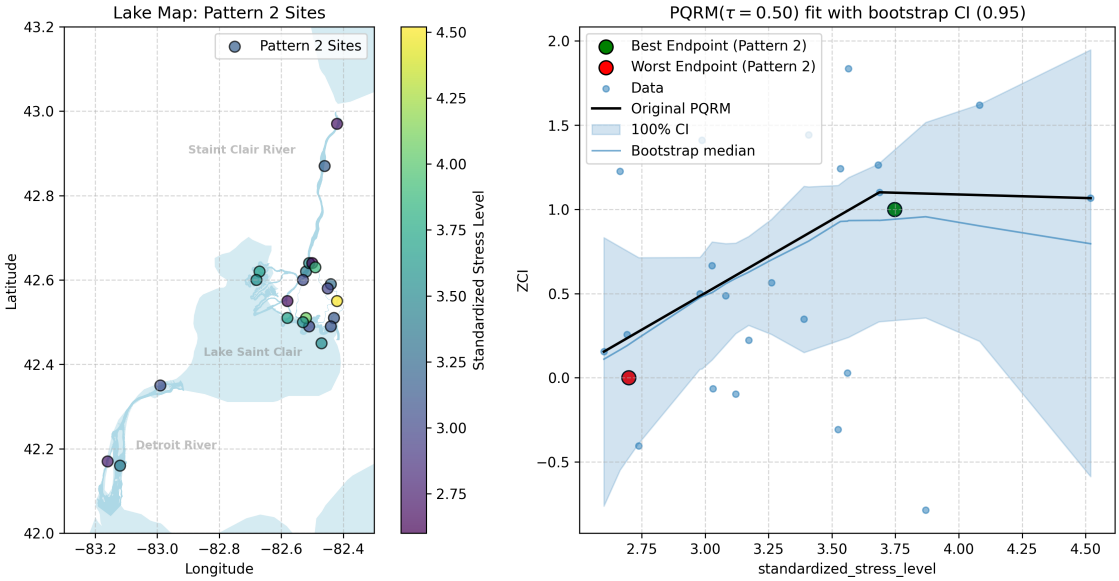


Figure 10: Quantile regression results of ZCI vs Stress Scores for cluster 2

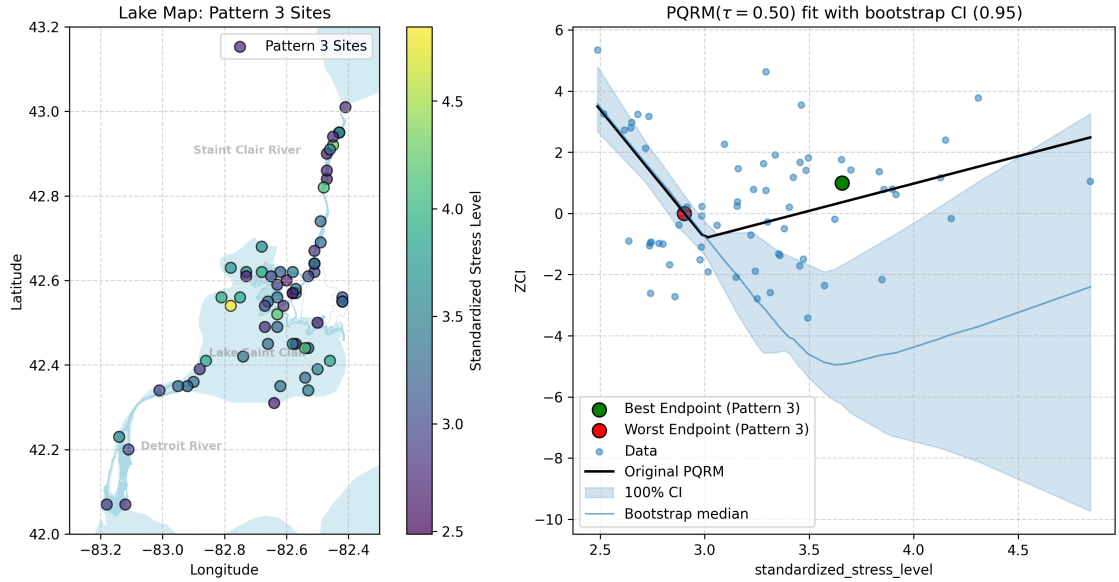


Figure 11: Quantile regression results of ZCI vs Stress Scores for cluster 3

# 2 Appendix

## 2.1 Tables

Table 4: Environmental Variables and Their Explanations

| Variable Name | Explanation |
|---|---|
| Site_ID | Unique identifier for each sampling site |
| Lake_or_River | Indicates whether the site is in a lake or river |
| Latitude | Geographic latitude coordinate |
| Longitude | Geographic longitude coordinate |
| Total_Organic_Carbon_LOI_percent | Total organic carbon content (loss on ignition, as %) |
| Water_Depth_m | Water depth at the sampling location (meters) |
| Water_Temperature_C | Water temperature in degrees Celsius |
| Dissolved_Oxygen_Concentration_mgL | Dissolved oxygen concentration in milligrams per liter |
| Median_Particle_Size_Phi | Median particle size of sediment (Phi scale) |

Table 5: Taxonomic Variables and Their Explanations

| Taxonomic Group | Explanation |
|---|---|
| Oligochaeta | Aquatic segmented worms |
| Nematoda | Roundworms |
| Chironomidae | Non-biting midges (larvae) |
| Ceratopogonidae | Biting midges |
| Hexagenia | Mayfly genus (larvae) |
| Caenis | Mayfly genus (larvae) |
| Hydropsychidae | Net-spinning caddisflies |
| Other Trichoptera | Other caddisfly families |
| Amphipoda | Small crustaceans (e.g., scuds) |
| Dreissena | Zebra/quagga mussels |
| Acari | Aquatic mites |
| Hydrozoa | Small predatory animals (hydroids) |
| Hirudinea | Leeches |
| Turbellaria | Flatworms |
| Gastropoda | Snails and slugs |
| Sphaeriidae | Fingernail clams |

Table 6: Stressors and Their Explanations

| **Chemical/Contaminant** | **Explanation** |
|---|---|
| Al | Aluminum (trace metal) |
| As | Arsenic (toxic element) |
| Bi | Bismuth (trace element) |
| Ca | Calcium (major element, hardness) |
| Cd | Cadmium (toxic metal) |
| Co | Cobalt (trace element) |
| Cr | Chromium (trace metal) |
| Cu | Copper (trace metal, micronutrient) |
| Fe | Iron (major element, micronutrient) |
| Hg | Mercury (highly toxic metal) |
| K | Potassium (major element) |
| Mg | Magnesium (major element, hardness) |
| Mn | Manganese (trace element) |
| Na | Sodium (major element) |
| Ni | Nickel (trace metal) |
| Pb | Lead (toxic metal) |
| Sb | Antimony (trace element) |
| V | Vanadium (trace element) |
| Zn | Zinc (trace metal, micronutrient) |
| %OC | Percent organic carbon |
| 1245-TCB | 1,2,4,5-Tetrachlorobenzene (organic pollutant) |
| 1234-TCB | 1,2,3,4-Tetrachlorobenzene (organic pollutant) |
| QCB | Quintachlorobenzene (organic pollutant) |
| HCB | Hexachlorobenzene (organic pollutant) |
| OCS | Octachlorostyrene (organic pollutant) |
| p,p'-DDE | DDT breakdown product |
| p,p'-DDD | DDT breakdown product |
| mirex | Organochlorine insecticide |
| Heptachlor Epoxide | Organochlorine pesticide breakdown product |
| total PCB | Total polychlorinated biphenyls |

Table 7: Explanation, Habitat, and Survival Rate in Fast/Slow Water for Each Taxon

| **Taxa** | **Explanation** | **Habitat Type** | **Survival Rate (Fast/Slow Water)** |
|---|---|---|---|
| Oligochaeta | Aquatic segmented worms | Both (lentic/lotic) | Moderate/High |
| Nematoda | Roundworms | Both (lentic/lotic) | Moderate/High |
| Chironomidae | Non-biting midge larvae | Both (lentic/lotic) | Moderate/High |
| Ceratopogonidae | Biting midge larvae | Both (lentic/lotic) | Moderate/Moderate |
| Hexagenia | Burrowing mayflies | Lentic (lakes/ponds) | Low/High |
| Caenis | Small mayflies | Both (lentic/lotic) | Low/Moderate |
| Hydropsychidae | Net-spinning caddisflies | Lotic (streams/rivers) | High/Low |
| Other Trichoptera | Other caddisflies | Both (lentic/lotic) | Moderate/Moderate |
| Amphipoda | Small crustaceans | Both (lentic/lotic) | Moderate/High |
| Dreissena | Zebra/quagga mussels | Lentic, large rivers | Low/High |
| Acari | Aquatic mites | Both (lentic/lotic) | Moderate/High |
| Hydrozoa | Predatory invertebrates | Lentic (lakes/ponds) | Low/High |
| Hirudinea | Leeches | Both (lentic/lotic) | Low/High |
| Turbellaria | Flatworms | Both (lentic/lotic) | Low/High |
| Gastropoda | Aquatic snails | Both (lentic/lotic) | Low/High |
| Sphaeriidae | Fingernail clams | Both (lentic/lotic) | Low/High |

Table 8: Chemical Descriptive Statistics by Site Label

| | SumReal | degraded | intermediate | reference |
|---|---|---|---|---|
| Al | mean | 4276.423 | 6380.140 | 4319.381 |
| | std | 2888.769 | 5523.949 | 1767.861 |
| As | mean | 2.186 | 1.777 | 2.232 |
| | std | 1.602 | 1.290 | 1.041 |
| Bi | mean | 17.085 | 17.505 | 17.622 |
| | std | 10.352 | 10.273 | 9.722 |
| Ca | mean | 28180.500 | 33518.930 | 28480.714 |
| | std | 14031.433 | 11400.266 | 11870.107 |
| Cd | mean | 0.535 | 0.351 | 0.271 |
| | std | 0.649 | 0.202 | 0.233 |
| Co | mean | 4.049 | 4.497 | 3.984 |
| | std | 1.733 | 2.209 | 1.118 |
| Cr | mean | 13.254 | 12.830 | 9.007 |
| | std | 16.373 | 11.835 | 2.937 |
| Cu | mean | 16.958 | 18.082 | 12.946 |
| | std | 22.388 | 29.120 | 9.003 |
| Fe | mean | 9495.000 | 11246.789 | 9650.905 |
| | std | 5392.824 | 6804.654 | 3856.739 |
| Hg | mean | 0.474 | 0.324 | 0.196 |
| | std | 1.230 | 0.420 | 0.365 |
| K | mean | 818.927 | 1285.558 | 845.657 |
| | std | 638.053 | 1092.550 | 411.332 |
| Mg | mean | 12849.500 | 15204.175 | 12269.143 |
| | std | 6104.202 | 5764.037 | 5281.794 |
| Mn | mean | 161.228 | 188.900 | 161.905 |
| | std | 76.973 | 86.663 | 57.883 |
| Na | mean | 118.998 | 134.042 | 123.611 |
| | std | 49.081 | 43.693 | 41.021 |
| Ni | mean | 11.225 | 12.399 | 9.136 |
| | std | 8.851 | 8.424 | 3.542 |
| Pb | mean | 12.515 | 8.774 | 8.573 |
| | std | 32.312 | 22.204 | 18.750 |
| Sb | mean | 17.262 | 16.765 | 18.001 |
| | std | 11.879 | 13.115 | 13.743 |
| V | mean | 15.274 | 18.353 | 15.183 |
| | std | 7.012 | 9.560 | 4.408 |
| Zn | mean | 52.732 | 46.181 | 35.677 |
| | std | 48.896 | 44.586 | 17.938 |
| %OC | mean | 2.110 | 2.405 | 1.779 |
| | std | 1.599 | 1.458 | 0.682 |
| 1245-TCB | mean | 0.906 | 1.201 | 0.555 |
| | std | 2.321 | 2.143 | 1.035 |
| 1234-TCB | mean | 0.252 | 0.234 | 0.253 |
| | std | 0.257 | 0.240 | 0.332 |
| QCB | mean | 0.729 | 1.255 | 0.636 |
| | std | 1.015 | 3.055 | 0.871 |
| HCB | mean | 2.759 | 17.713 | 2.904 |
| | std | 4.291 | 83.487 | 6.011 |
| OCS | mean | 1.213 | 1.502 | 0.721 |
| | std | 3.395 | 3.606 | 1.874 |
| p,p'-DDE | mean | 0.679 | 0.485 | 0.324 |
| | std | 1.255 | 0.930 | 0.328 |
| p,p'-DDD | mean | 3.879 | 0.772 | 0.862 |
| | std | 14.634 | 1.039 | 0.923 |
| mirex | mean | 0.253 | 0.212 | 0.134 |
| | std | 0.682 | 0.332 | 0.242 |
| Heptachlor Epoxide | mean | 0.098 | 0.051 | 0.071 |
| | std | 0.235 | 0.250 | 0.211 |
| total PCB | mean | 15.137 | 10.705 | 7.715 |
| | std | 32.189 | 36.285 | 16.795 |