Zoobenthic Community Indicators of Sediment Contamination in a Large River: Applications of Data Science

Feng Gu

October 10, 2025

Contents

1 Introduction

1.1 Sediment contamination in aquatic ecosystems

Rivers and lakes provide essential services for human well-being and biodiversity [61]. With intensive industrial and urban development along the shorelines, anthropogenic activities have released significant amount of inputs into the these aquatic ecosystems [2], elevating ecological risks as these contaminants bioaccumulate in aquatic organisms and eventually enter human food chains through fish consumption, posing health risks [27, 56]. These ecological and human health risks underscore the urgent need for scientific assessment to inform management strategies [44].

Most of these anthropogenic inputs are released into the water column first through various point and non-point sources, carrying two major groups of components: nutrients and hydrophobic contaminants [16]. These nutrients (especially phosphorus) and harmful chemicals quickly attach to floating particles and settle into the muddy bottom sediments through natural binding processes [2]. As these particles settle and accumulate at the bottom where water meets sediment, they trap sediment contaminants through similar binding and absorption processes, turning bottom sediments into a combined storage area of past pollution [27, 56, 18, 11]. These binding-produced particles remain chemically stable for extended periods, effectively preserving pollutant signatures and creating a persistent record of contamination events across multiple temporal scales [18, 11]. This integrative role establishes sediments as a biogeochemical archive that smooths short-term water column fluctuations, making sediment contaminant measurements a robust long-term manifestation of anthropogenic stress [37].

1.2 Benthic communities as bioindicators for sediment contamination

Chemical-only sediment assessment is resource-intensive and overlooks bioavailability, necessitating complementary indicator approaches [17]. In most natural ecosystems, individual organisms exhibit physiological and behavioral adjustments under external pressures; aquatic benthic macroinvertebrates likewise express stressor-mediated responses that scale to detectable community change [33]. Benthic macroinvertebrate assemblages respond to changes in sediment-borne contaminants and habitat conditions [57], rendering them as effective bioindicators for sediment contamination levels [23]. These responses can be observed through changes in benthic community composition and many attempts on small aquatic systems, such as small lakes, rivers, and streams, have been shown successful [15, 3], motivating extension to large lakes and great rivers where fewer successful implementations exist [6, 19].

Effective bioindicator development requires balancing ecological relevance with practical applicability. Ideal taxa should exhibit sensitivity to pollution gradients, maintain wide distribution across environmental gradients within the study area, and be readily sampled and taxonomically identified [41]. In large aquatic ecosystems, environmental complexity and heterogeneity present additional challenges for bioindicator development, necessitating the inclusion of diverse taxa assemblages that influence both indicator construction and interpretation [43, 7]. Through strategic selection of an appropriate taxonomic pool, it becomes feasible to develop robust and sensitive bioindicators for sediment contamination assessment [41].

1.3 Variation in community composition across environmental gradients

Zoobenthic communities respond to both anthropogenic stressors (e.g., sediment contaminants) and natural environmental variability (e.g., substrate, temperature, flow) [46]. Under consistent environmental conditions, community composition prior to anthropogenic influence represents the naturally-shaped baseline for comparison following pollution onset [49].

In pervasively human-influenced regions, most investigated aquatic sites exhibit zoobenthic community composition obscured by the combined influence of anthropogenic stressors and natural environmental variability [55]. This necessitates identifying and isolating the pre-pollution community composition from observed levels, since only the anthropogenically-driven component is relevant for developing bioindicators of sediment contamination [49].

Therefore, partitioning community composition into naturally-driven and pollution-driven components is essential for sediment contamination bioindicator development [49]. The natural component provides the taxonomic baseline and enables quantification of anthropogenic deviations, thereby isolating the pollution-driven signal across temporal scales for indicator construction [40].

1.4 Accounting for spatial structure and autocorrelation

Spatial location of sampling sites represents easily collected yet frequently overlooked information. Many studies assume that zoobenthic communities are shaped identically by natural attributes irrespective of their spatial positions [9], bringing simplicity in quantitative analysis but this may not hold in all real scenarios.

Advances in spatial statistics and computational power now enable incorporation of ecosystem spatial structure and linking environmental complexity to geo-heterogeneity [8, 32]. Accumulating studies demonstrate that natural heterogeneity is spatially structured across ecosystems and that modeling this structure substantially improves explanation for community variation [59, 9, 8]. The primary rationale is that many unmeasured or unmeasurable natural environmental attributes have their own spatial signatures, enabling spatial variables to serve as proxies for these hidden drivers [1] [24, 59]. Therefore, incorporating essential spatial analysis is promising to control more natural variation in community composition, ultimately bringing a clearer picture of zoobenthic community responses to sediment contamination [40].

1.5 Study system: Huron-Erie corridor and Detroit River

These mechanistic principles provide the theoretical foundation for developing a bioindicator of sediment contamination that controls for environmental and spatial variability. To implement this framework, this study will apply quantitative methods within the Great Lakes region.

As one of the world's largest surface freshwater ecosystems, the Great Lakes region contains 21% of the global supply of surface fresh water and supports the livelihoods of millions of people [60]. While much of the Great Lakes region remains minimally disturbed, certain areas suffer from intensive anthropogenic activities [1]. These heavily impacted areas are primarily concentrated along Lake Erie's shorelines and its connecting waterways, particularly the Detroit River[1], making these regions intensively monitored and studied.

Leveraging available data from Zhang [63] and her team, this study will use data collected from the Huron-Erie corridor that comprises the St. Clair Lake, St. Clair River, and Detroit River. This corridor forms a connected water system characterized by fast flow due to channel constriction [2] and by extensive anthropogenic pollution from upstream sources and surrounding shorelines. Among the three segments, the Detroit River will be the primary focus due to its intensive human influence and the availability of extensive benthic community and sediment contamination data ([63]). By developing this bioindicator framework on the surveyed area, this study is expected to advance bioindicator research for large lakes and rivers and provide a methodological foundation for broader application across the Great Lakes region.

[1] This approach parallels the use of instrumental variables and proxy variables in econometrics, where measurable variables are employed to control for unmeasured confounders and enable causal inference.

[2] Channel constriction occurs when water flow is forced through a narrower cross-sectional area, resulting in increased velocity and turbulence according to the continuity equation in fluid dynamics.

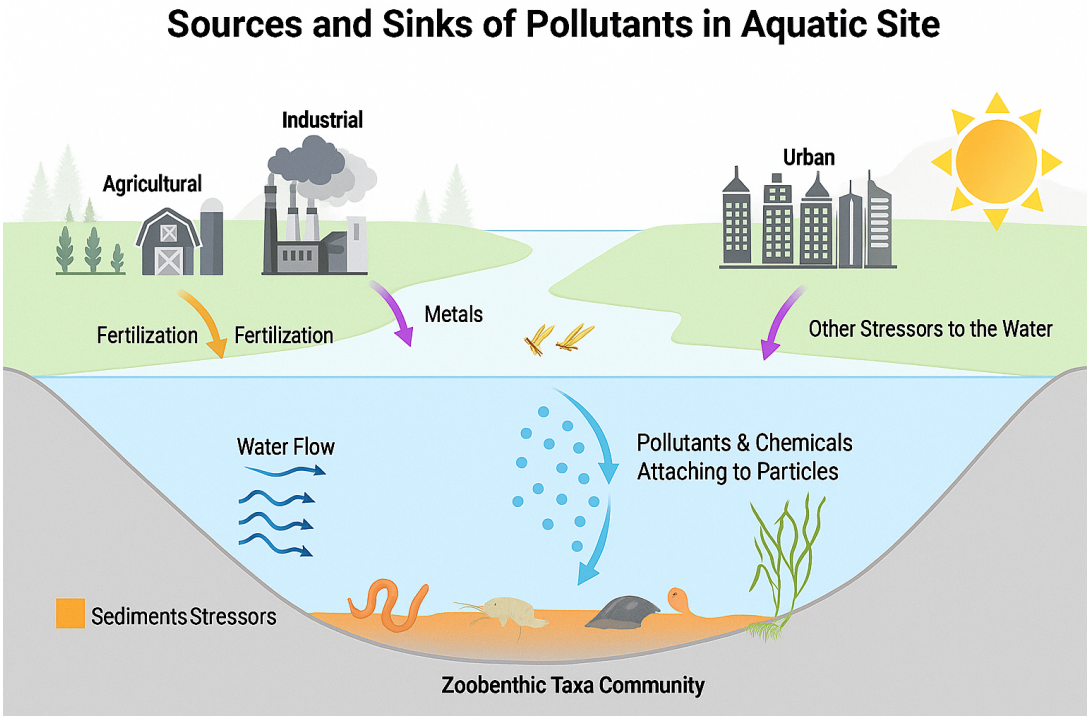

Figure 1: Overview of how zoobenthic community responses to both sediment contamination and natural environmental conditions

2 Research Objectives

The goal of this study is to build a zoobenthic community indicator(ZCI) of sediment contamination in a large aquatic ecosystem. This work can be divided into five specific objectives with respects to aquatic ecology principles and statistical methodologies:

1. Build quantitative measurements for sediment contamination levels and zoobenthic community composition.
2. Control for natural variability in community composition and isolate anthropogenic effects on community composition.
3. Incorporate spatial features to account for potential spatial heterogeneity in community composition.
4. Model the relationship between the sediment contamination levels and zoobenthic community composition with piecewise quantile regression and extend it to a bioindicator(ZCI) of sediment contamination level.
5. Evaluate the indicator power and robustness of the developed ZCI with respect to sample size and other relevant global factors that can influence the estimates.

These listed objectives are made in a tentatively sequential order. Some objectives have been well-supported by existing researches while others are more exploratory. To the objectives 1, 2 and 4, they form a good benchmark of building the zoobenthic community indicator and are supported by a well-established **Multivariate Community Indicator Framework** [3] [10, 38, 34, 19] with a piecewise quantile regression model. To objectives 3 and 5, they are relatively exploratory and will serve the purpose of enhancing and understanding the indicator's power and sensitivities.

The following subsections disccussed the general implementations, highlighting the focuses where possible contributions could be made.

2.1 Multivariate community indicator framework

Multivariate community indicator framework is a comprehensive approach for achieving objectives 1, 2 and 4, we will build it with possible improvements. Three data matrices will be used for it: the taxa matrix, the environmental matrix, and the stressor(chemical) matrix. On top of them, quantitative measurements for sediment contamination and zoobenthic community composition will be designed and applied meanwhile controlling the natural variability. Its basic steps are summarized as follows:

1. **Measure sediment contamination and select reference sites**

   Motivation: establish a defensible contamination gradient and set least disturbed sites as benchmark.

   PCA the stressor matrix; filter and interpret qualified pollutant PCs to form a composite contamination score. Filter sites with the top least disturbance levels as reference sites and they approximate the "naturally shaped" assemblages.

2. **Predict 'naturally shaped' community composition across environmental gradients**

   Motivation: predict the naturally-driven community composition to disentangle the pollution-driven deviation in the community composition.

   Cluster reference-site on taxa information to form groups balancing within-group biotic homogeneity and environmental gradient coverage. Train a discriminant model to predict closest community cluster for any site given its environmental conditions. This sites within the same community cluster are considered environmentally homogeneous.

3. **Measure community composition(ZCI) for environmentally homogeneous sites**

   Motivation: construct a numerical index to compare community composition across sites with similar environmental conditions.

   Within each predicted cluster, apply ordination (e.g., PCA/NMDS) to taxa information; derive the measurement results as ZCI with defined scoring rules. Sites deviating from the reference site centroid are considered disturbed by pollution, they compromise the disturbance-relevant ZCI results.

[3] Its details are provided in following content and in the Methodology section

4. **Quantify the relationship between contamination level and ZCI**

   Motivation: quantify how community composition responds to contamination levels

   Fit (piecewise quantile) regressions of ZCI deviation vs. contamination score to capture distributional responses, test and locate breakpoints (thresholds) across ZCI values to understand the ecological implications.

2.2 Spatial heterogeneity test and incorporation

Objective 3 is to ensure the unmodelled spatial patterns do not confound the regression results. Ecological data often exhibit spatial autocorrelation—nearby sites tend to have similar communities—so failing to include spatial structure can lead to biased estimates and inflated error rates. The principal coordinates of neighbour matrices (PCNM) method will transform spatial distances among sites into orthogonal spatial predictors that can be incorporated into regression or canonical analyses [8, 25, 30].

PCNM eigenvectors from the Euclidean distance matrix will capture spatial structure at different scales [8, 30]. For each environmental cluster, we will test for spatial heterogeneity by regressing ZCI against PCNM vectors to identify significant spatial predictors not explained by environmental variables. Selected PCNM variables will be incorporated as covariates in quantile regression models to control spatial autocorrelation.

2.3 Piecewise quantile regression for breakpoint in quantile relationship

Objective 4 is to model a piecewise quantile regression between sediment contamination levels and ZCI, which enhances the ZCI-contamination regression part in the multivariate community indicator framework. Piecewise quantile regression (PQRM) is chosen for its ability to estimate conditional quantiles of a response without assuming a specific parametric form [14, 35], to model the entire conditional distribution rather than just the mean [14], and to provide results that are robust to outliers or heteroscedastic variance [35]. In the ecological context, quantile regression can reveal changes in zoobenthic community composition associated with increasing contamination levels at different quantiles (e.g., high quantile representing sensitive taxa) and detect breakpoints that indicate ecological thresholds [58, 54, 20].

It is worth to detect threshold(s) in contamination levels across which the slope of the ZCI–contamination relationship changes, indicating points where benthic communities change its way in responding to contamination. Higher quantiles may reveal steeper changes than median quantiles, highlighting vulnerable taxa. Mapping breakpoints across clusters will show whether thresholds vary among environmental gradients.

2.4 Indicator power and robustness with respect to sample size

Objective 5 evaluates how reliably the ZCI detects sediment contamination gradients under varying sampling effort. Robustness analyses are essential because bioindicator performance can deteriorate when sample sizes are small, variance components are poorly estimated, or site selection produces hidden pseudoreplication [36]. Power and precision directly influence management credibility; underpowered indicators risk Type II errors (failing to flag degraded conditions) while unstable estimates inflate Type I error rates in threshold detection [45, 28].

The analytical approach is tentatively to use bootstrap resampling and subsampling to evaluate ZCI robustness. Bootstrap resampling may generate sampling distributions for key parameters (slopes, breakpoints, pseudo-R^2), while subsampling could construct precision curves by repeatedly subsetting sites across sample sizes. Power analysis will potentially simulate datasets under null and alternative hypotheses to compute detection power for contamination effects and threshold shifts. Spatial autocorrelation may need to adjust effective sample sizes, and breakpoint reliability could be assessed through confidence interval width criteria or other suitable metrics.

Results will identify optimal sampling design (sites per cluster) and provide decision-support tables linking confidence levels to required sampling effort. Higher quantiles may require larger samples due to greater dispersion, while also elevating the potential concerns regarding spatial autocorrelation.

3 Data Description

3.1 Data Collection

The data used in this program is provided by Dr. Ciborowski, collected and processed by Zhang [63]. It consists of data from three separate surveys conducted in: 1991, 1999 and 2004, all following the same field protocols [4].

The 2004 data set was majorly collected across the whole zone of Lake Huron–Lake Erie Corridor. The collected information includes location information (longitude and latitude), 16 taxonomic variables, 5 environmental variables, and 30 stressors. The data from two previous studies, which collected data from the Detroit River zone in 1991 and 1999 (Farara and Burt 1993; Wood 2004)—were compiled and incorporated into the 2004 data. This combination enhances the dataset's robustness by providing a more comprehensive perspective on the benthic community dynamics, environmental conditions, and sediment contamination across the entire Corridor over time.

Given the temporal and spatial distribution of sampling across three survey years, **StationID serves as the primary key for data integration and site identification.** Each StationID uniquely identifies a sampling site in both temporal and spatial dimensions, meaning observations with the same StationID represent identical location-time combinations from the three survey years (1991, 1999, 2004).

3.2 Prepare a complete sub-dataset for preliminary analysis

Taxonomic, environmental, and stressor data for sampled sites were originally stored in three separate tabular files. For preliminary analysis, we made a quick check on the data quality and completeness, by the submission of the proposal the latest check was done on July 24, 2025.

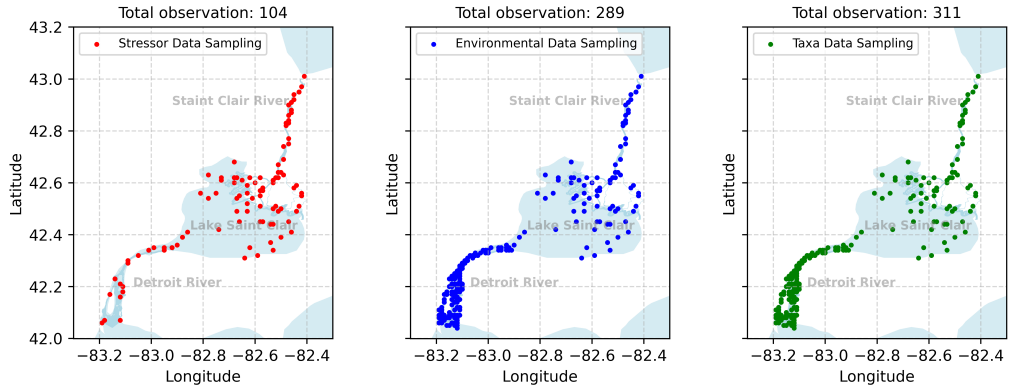

Figure 2: Different data sampling in temporal and spatial dimensions of the three datasets in the survey area.

Figure 2 shows the numbers and locations of observations in each table. Note that some locations may be sampled in different years so only a StationID [5] (rather than a geographical location) can be used to confirm the identity of a sampling site. The three datasets exhibit misalignment in sample sizes. The stressor dataset contains the fewest observations (104), while the taxonomic and environmental datasets contain substantially more observations (289 or more). To prepare a complete dataset for preliminary analysis, the three datasets were merged by StationID using inner-join operations, resulting in a comprehensive dataset with 104 observations containing taxonomic, environmental, and stressor data across the Lake Huron-Erie Corridor.

[4] These sampling locations were determined prior to fieldwork by a stratified random sampling design to ensure representative coverage.

[5] StationID uniquely identifies a sampling site in both temporal and spatial dimensions.

This sample size misalignment will be resolved when complete data becomes available in the near future. Since sample size is the primary difference between current and forthcoming datasets, the preliminary analysis framework is designed for easy scalability when additional data becomes available.

3.3 Large River Case Study: extra water velocity data in Detroit River

Previous work [63] identified limitations in model performance due to insufficient environmental variable coverage. In a Detroit River-focused analysis [63], a new environmental variable—bottom **water velocity**—was added, which was derived by Dr. Reitsma using a three-dimensional water flow model. The focused analysis demonstrated that water velocity is a critical environmental variable for controlling environmental variation. Therefore, the Detroit River will serve as a focused study area with water velocity data included, and it will be conducted once complete stressor data becomes available.

3.4 Environmental attributes and samples

Across the three Lake Huron-Lake Erie Corridor surveys, location information (longitude and latitude recorded via GPS readings) and 5 environmental attributes were measured at each sampling site.

Temperature (°C) and **dissolved oxygen concentration** (mg/L) were measured using a Hydrolab multimeter. **Water depth** (m) was recorded from the Ponar rope. **Loss on ignition (%)** and **median particle size (phi-units)** were determined during sediment processing but are treated as environmental attributes due to their fundamental roles in habitat characterization (details on their analysis are provided in a later subsection). **Water velocity** (m/s) was estimated for the Detroit River area across the three surveys, displayed in the Table 1 along with other cross-corridor attributes.

Table 1: Environmental Attributes

Attribute	Ecological Relevance
Temperature (°C)	Controls metabolic rates and organism distribution patterns.
Dissolved Oxygen Concentration (mg/L)	Determines survival and excludes oxygen-sensitive taxa when low.
Water Depth (m)	Affects light penetration and benthic habitat availability.
Loss on Ignition (%)	Indicates organic matter content and food availability for benthos.
Median Particle Size (phi-units)	Determines substrate stability and habitat suitability for taxa.
Water Velocity * (m/s)	Controls flow regime and determines which taxa can colonize sites.

(i) * Water velocity was estimated from the model by Dr. Reitsma (Detroit River area only, 225 observations) [63].
(ii) Other environmental attributes were measured at all sampling sites across the entire survey area (311 observations).
(iii) All three surveys (1991, 1999, 2004) followed identical collection protocols and were merged for comprehensive temporal analysis.

These attributes are commonly used to describe baseline environmental conditions in aquatic habitats, as they are primarily governed by natural physical processes that influence taxonomic composition [21, 59].

By including these variables as covariates to partially partition the zoobenthic community composition, we can partially control for natural variation contributed by habitat characteristics, thereby isolating the effects of anthropogenic stressors in subsequent analyses of community composition patterns.

3.5 Taxonomic attributes and samples

The zoobenthos were collected with a Ponar grab sampler. After considering the fullness of each grab and the removing of fine materials, the team applied multiple grabs at each site until a total volume of $2L$ sediment was collected. The sediment samples for organic and metals analysis were preserved in corresponding professional containers, all these samples were stored frozen.

One zoobenthic sample replicate from each site was randomly selected and processed, while the other two were archived. Samples were sieved into size fractions (4 mm, 1 mm, 0.5 mm, 0.25 mm), then elutriated to separate lighter detritus and animals from inorganic sediments. Each fraction was sorted

under a microscope and organisms were identified to the lowest possible taxonomic rank using standard keys. Zoobenthos were preserved in 70% ethanol in labeled vials and archived at the University of Windsor[63]. Immediately after the initial sorting of samples, ten samples were randomly selected to assess the sorting efficiency. One sample had a sorting efficiency of 91%, while the remaining samples had efficiencies of 96% or higher.

Specifically, there were 16 taxa recorded from the sediment samples, as shown in the table 2. According to their creature characteristics and preferred habitat, these taxa can be gently divided into three groups according to their preferred habitat.

Table 2: Benthic Taxa and Preferred Habitat Features

Taxa	Explanation	Preferred Habitat
Nematoda	Roundworms	Broad
Chironomidae	Non-biting midges (larvae)	
Ceratopogonidae	Biting midges	
Amphipoda	Small crustaceans (scuds)	
Acari	Aquatic mites	
Hydrozoa	Small predatory animals	
Gastropoda	Snails and slugs	
Oligochaeta	Aquatic segmented worms	Depositional zone
Hexagenia	Mayfly genus (larvae)	
Dreissena	Zebra/quagga mussels	
Hirudinea	Leeches	
Turbellaria	Flatworms	
Sphaeriidae	Fingernail clams	
Caenis	Mayfly genus (larvae)	Erosional zone
Hydropsychidae	Net-spinning caddisflies	
Other Trichoptera	Other caddisfly families	

- **Broad habitat:** Characterized by a wide range of environmental tolerance. Species in this group can inhabit both depositional and erosional areas, adapting to variable flow velocities, sediment types, and oxygen levels.
- **Depositional zone:** Areas with low to moderate water velocity where fine sediments (silt, clay, and organic matter) settle. These habitats often exhibit higher organic content and reduced oxygen penetration, favoring taxa adapted to softer substrates and potentially more enriched nutrient conditions.
- **Erosional zone:** Areas with higher flow velocity, coarser substrates (gravel, cobble, or sand), and well-oxygenated water. These habitats support taxa adapted to cling or attach to stable surfaces and withstand stronger currents.

3.6 Stressors attributes and samples

Sediment samples from each site were thoroughly mixed to ensure homogeneity. The homogenized samples were then split into separate portions for different analyses, including median particle size, total organic carbon (TOC), organic contaminants, and metals.

- **Particle Size:** Median particle size analysis was performed by sieving dried sediment through a series of sieves of decreasing mesh size. Each size fraction was weighed and described using phi units ($\phi = -\log_2 d$), where d is particle size in mm.
- **Total Organic Carbon (TOC):** Sediment TOC(%OC) was determined using loss on ignition (LOI). Pre-weighed, dried sediment samples were combusted at 450°C for 24 hours, and organic carbon was determined gravimetrically by subtracting the remaining mass.
- **Organic Contaminants:** The concentrations of organic contaminants (including 1245-TCB, 1234-TCB, QCB, HCB, OCS, p,p'-DDE, p,p'-DDD, mirex, Heptachlor Epoxide, total PCB) were measured using a gas chromatograph equipped with a 63Ni electron capture detector, following standard operating procedures.

- **Metals:** Metal concentrations (including Al, As, Ca, Cd, Co, Cr, Cu, Fe, Mn, Ni, Pb, and Zn) were analyzed using an Inductively Coupled Plasma Optical Emission Spectrophotometer (ICP-OES). For total mercury (Hg), an atomic absorption spectrophotometer (AAS) was used with a vapor generation accessory for increased sensitivity. Liquid samples were introduced into the instrument for metal analysis.

Quality assurance and chemical analyses were performed in collaboration with the Great Lakes Institute for Environmental Research (GLIER) at the University of Windsor[63]. Among the chemical variables analyzed, major earth elements (Al, Ca, Fe) are generally non-toxic at typical environmental concentrations, reflecting natural sediment composition. However, elevated levels from industrial activities can make them potential stressors, leading to difficulties in assessing their impacts due to naturally high background concentrations.

In contrast, trace metals (As, Bi, Cd, Co, Cr, Cu, Hg, Mn, Ni, Pb, Sb, V, Zn) are anthropogenic pollutants that bioaccumulate in sediments and cause toxicity in benthic organisms.

Persistent organic pollutants (PCBs, QCB, HCB, OCS, p,p'-DDE, p,p'-DDD, mirex, Heptachlor Epoxide) from industrial activities and pesticide use persist in sediments and accumulate in food webs, causing chronic effects on aquatic organisms.

Percent organic carbon (%OC) from decomposed organic matter influences contaminant binding and bioavailability, modulating the ecological impact of other pollutants.

By distinguishing between natural major elements (Al, Ca, Fe), trace elements (Co, Mn, Sb, V), toxic metals (As, Cd, Cr, Cu, Ni, Pb, Zn), the highly toxic mercury (Hg), organic matter content (%OC), and persistent organic pollutants (including pesticides and industrial compounds), these measurements enable a more comprehensive assessment of the sediment stress level and its ecological implications for the benthic community.

4 Literature Review

4.1 Sediment Contamination Assessment

Sediment contamination assessment serves as a fundamental tool for understanding and protecting aquatic ecosystem health, as contaminated sediments pose significant risks to benthic communities and serve as long-term sources of pollutants that can affect entire food webs [42, 12]. Sediments act as major repositories for a wide range of contaminants including heavy metals, persistent organic pollutants, and industrial chemicals, making their assessment critical for evaluating ecological risks and informing environmental management decisions [17, 56].

Some popular evaluation methods include contaminant index-based approaches such as the Enrichment Factor (EF) and the Geoaccumulation Index (Igeo). These index-based approaches provide pollution scores relative to known background or reference values, often focusing on individual or selected chemical elements [5] (Birch, 2022). For example, the Igeo index [4] (Birch, 2013) compares current concentrations to pre-industrial levels, while the EF index compares current concentrations to regional background levels. However, these approaches are limited by their reliance on accurate background concentration estimates and consideration of only a few chemical elements. When applied to datasets with many variables, they become less effective as they treat elements independently and fail to capture multivariate contamination patterns [31].

A more general class of evaluation methods relies on multivariate ordination and data reduction techniques. These data-driven approaches consider all chemical variables simultaneously, revealing interaction patterns through dimensionality reduction without requiring background or reference values, thus avoiding the challenge of defining appropriate benchmarks [19, 49]. However, such methods often yield composite axes or scores that lack intuitive interpretation and clear pollution thresholds, limiting their ability to provide comparable pollution categories like index-based methods [48]. This interpretability challenge has been recognized in ecological assessment contexts, where practitioners need actionable thresholds for management decisions [50].

Turn to large river and lake ecosystems, various pollution sources contribute to a complex array of contaminants, escalating the need for approaches that can handle complex, multivariate contamination patterns. Considering there are 30 stressor variables in my project and many of them share common pollution sources and interact via similar transport mechanisms, a data-driven and multivariate approach is more suitable to capture the contamination levels and patterns. Some ordination methods, such as Principal Component Analysis (PCA), can effectively reduce dimensionality and eliminate redundancy among

Table 3: Sediment Stressors and Classifications

Variable	Description	Type
Metals (mg/kg sediment)		
Al	Aluminum concentration (nontoxic)	Earth Element
Ca	Calcium concentration (nontoxic)	Earth Element
Fe	Iron concentration (nontoxic)	Earth Element
K	Potassium concentration (nontoxic)	Earth Element
Mg	Magnesium concentration (nontoxic)	Earth Element
Na	Sodium concentration (nontoxic)	Earth Element
As	Arsenic concentration (pollutant)	Trace Metal
Bi	Bismuth concentration (pollutant)	Trace Metal
Cd	Cadmium concentration (pollutant)	Trace Metal
Co	Cobalt concentration (pollutant)	Trace Metal
Cr	Chromium concentration (pollutant)	Trace Metal
Cu	Copper concentration (pollutant)	Trace Metal
Hg	Mercury concentration (highly pollutant)	Trace Metal
Mn	Manganese concentration (pollutant)	Trace Metal
Ni	Nickel concentration (pollutant)	Trace Metal
Pb	Lead concentration (pollutant)	Trace Metal
Sb	Antimony concentration (pollutant)	Trace Metal
V	Vanadium concentration (pollutant)	Trace Metal
Zn	Zinc concentration (pollutant)	Trace Metal
Organic Carbon (mg/kg sediment)		
%OC	Organic carbon content	Binding agent
Organic Contaminants (mg/kg sediment)		
1245-TCB	1,2,4,5-Tetrachlorobenzene (hydrocarbon pollutant)	Industrial compound
1234-TCB	1,2,3,4-Tetrachlorobenzene (hydrocarbon pollutant)	Industrial compound
QCB	Pentachlorobenzene (hydrocarbon pollutant)	Industrial compound
HCB	Hexachlorobenzene (hydrocarbon pollutant)	Industrial compound
OCS	Octachlorostyrene (hydrocarbon pollutant)	Industrial compound
p,p'-DDE	Dichlorodiphenyldichloroethylene (pesticide)	Organochlorine
p,p'-DDD	Dichlorodiphenyldichloroethane (pesticide)	Organochlorine
mirex	Mirex (pesticide)	Organochlorine
Heptachlor Epoxide	Heptachlor Epoxide (pesticide)	Organochlorine
total PCB	Total polychlorinated biphenyls	Sum of all PCBs

correlated variables [64]. In addition to reducing dimensionality and eliminating redundancy, advanced PCA methods can also enhance interpretability by incorporating variable weights [22] (Delchambre, 2015) or spatial information [32] (Harris et al., 2011)., making it possible to derive more meaningful contamination scores.

4.2 Control of confounding environmental gradients

In researching relationships between zoobenthic community composition and sediment contamination levels, naive associations can be confounded by environmental conditions that are correlated with community composition or even contamination. This induces omitted-variable bias and can either inflate or mask the true signal we seek for the zoobenthic condition index (ZCI). Accordingly, quantify the unique contribution of sediment contamination to community composition, after conditioning on measured environment, is essential for deriving a reliable and interpretable ZCI.

To overcome this confounding issue, some common fixes include: *(1) regression adjustment / partial regression; (2) variance partitioning; (3) blocking / stratification* [26, 62]. Approaches (1) and (2) are most often used in ecological regression as adjustments to the naïve model, whereas (3) is more common in experimental designs. The third approach represents a design/estimator-based solution that mitigates latent confounding when environmental information is limited or unaddressed [13].

Within the regression framework, adjustments are useful for statistically disentangling environmental influences [40]. However, model performance can be sensitive to the sampling configuration and to the explanatory power of predictors [53, 29].

Simulation studies show that model-revealed relationships depend not only on the true ecological relationships but also on the sampling–environmental distribution (SED), underscoring the importance of evenness and representativeness in sampling design [29]. These limitations produce high uncertainty in estimates when any of the following occur: missing environmental predictors, poor coverage across environmental gradients, or insufficient sample size [26].

Experimental designs can mitigate these issues when regression adjustment is not feasible, and analytical tools such as clustering or stratification can make the design more objective and data-driven [62]. However, these designs are also subject to SED and sample-size constraints.

A notable avenue is to infer *counterfactual outcomes* for the response variable conditional on specified environmental gradients; these outcomes represent the environmentally deterministic component of the response [39, 13]. This approach combines regression adjustment with experimental design: initial data for learning the counterfactual surface are selected via design (e.g., blocking/stratification), and the subsequent inference step uses trained regression models to estimate these deterministic counterfactual outcomes.

4.3 Zoobenthic Community Composition Measurement

Zoobenthic community composition measurement depends fundamentally on how we define and quantify the community, making it nearly impossible to establish a perfect measurement framework in real ecological systems [57]. Common descriptors of zoobenthic community composition include species richness, abundance, and biomass, which can be used individually or in combination to characterize the community [52]. Each metric presents trade-offs between the amount of ecological information captured and the cost of data acquisition [43].

The individual organism represents the minimal unit in a zoobenthic community and serves as the foundational unit for measuring abundance and biomass. While individual responses to external conditions ultimately comprise the community response, their high variability and inherent stochasticity present significant challenges for using single or few individuals to adequately describe community-level responses. In contrast, taxonomically rich assemblages can better represent community responses by averaging out individual-level variability and capturing broader ecological patterns [43].

Beyond the fundamental function of representing community structure, the choice of measurement approach depends critically on the study's primary objectives. In this project, the main purpose is to assess sediment contamination levels using zoobenthic community indicators [19]. Rooted in this purpose, we seek measurements that demonstrate sensitivity to contamination gradients— whether reflecting raw stressor concentrations or evaluated contamination levels—such that pollution-sensitive metrics can outperform alternatives in mathematically predicting contamination levels. However, such forecast-oriented measurements may exhibit limited mechanistic interpretability regarding the zoobenthic community composition itself, necessitating caution when interpreting results from a biological perspective. To

address this limitation, comprehensive mechanistic investigation of the taxa-based measurements and cross-validation with existing ecological knowledge becomes essential [50].

4.4 Quantitative Regression Beyond the Mean and Threshold Detection

Benefiting from the Central Limit Theorem and classical statistical theory, mean-based regression has dominated ecological analysis [14]. Traditional linear regression performs adequately under weak correlations and homogeneous variance, but degrades substantially when these assumptions are violated—a common occurrence in ecology due to natural system complexity and unmeasurable latent factors.

Beyond statistical limitations, mean-based approaches can miss important ecological patterns, elevating the meaningful investigation of relationships beyond the mean. Ecological drivers often act in a heterogeneous way across the distribution of responses, such that effects emerge only under extreme conditions and not in the center of the distribution. For instance, [51](Schmidt et al. 2012) found that the influence of metal concentration was much stronger at higher quantiles of the biological response than on the mean. Such upper-tail–specific associations underscore the limitation of mean regression approaches when the primary ecological signal lies in extremes. They thus motivate the use of "mean-beyond" methods - such as quantile regression - that can capture effects under high-stress or extreme conditions.

Nonlinear and threshold effects are another phenomenon observed in many ecological systems [14] (Cade & Noon, 2003). The concept of ecological thresholds emerged in the 1970s with the recognition that ecosystems often exhibit multiple 'stable' states under different conditions [33] (Holling, 1973, as cited in Groffman et al., 2006). Peterson et al. [47] (Peterson et al., 2006) summarized three main applications of ecological thresholds: (1) **ecosystem state shifts** from small driver changes, (2) **critical loads** that ecosystems can absorb without functional changes, and (3) **extrinsic factor thresholds** where large-scale changes modify driver–response relationships. These ecological recognitions motivate and justify the use of threshold-detecting methods in ecological regression modeling.

The threshold usually functions as a crucial decision point for management, policy, or conservation actions. Therefore, identifying and validating ecological thresholds is vital in both statistical and practical senses. Spake et al. [54] emphasized the importance of *scale*, highlighting four dimensions affecting threshold detection: (1) **grain** (measurement resolution), (2) **extent** (total area or duration covered), (3) **organizational level** (biological hierarchy), and (4) **analytical method** (statistical approach employed). Turning to threshold detection, testing their existence and reliability is crucial. Daily et al. [20] (Daily et al., 2012) showed that small sample sizes increase false detections, while larger samples improve reliability. The sample–environment distribution (SED) strongly influences detection and threshold location, with uneven SEDs producing bias. The rate of ecological change, whether linear or nonlinear, interacts with the chosen method, and user-defined parameters (e.g., quantile τ, bandwidth) play a critical role.

5 Methodology

This methodology section details the overall analytical framework and specific techniques to be employed in this research.

The proposed methodology consists of the following key steps:

1. Dimension reduction on stressors that produce pollution scores and arranges samples in a low-dimensional pollutant space to reveal main synthetic stressor gradients.

2. Find and prepare least polluted sites, construct an estimator that estimates the counterfactual surface (minimal pollution level) of taxa community composition in environmental space.

3. Ordinate the taxa variables, construct comprehensive measurement for the taxa community.

4. Spatial heterogeneity test and geo-information incorporated for latent environmental variables.

5. Piecewise quantile regression models on ZCI and pollution scores, revealing relations at different quantile levels and detecting thresholds.

6. Step out of the framework, change global factors, like sample size and partitioning samples by biased environmental distribution. Study how the series of estimates would change accordingly.

Various symbols are used throughout this section, including function names, vectors, and matrices. The following table summarizes these symbols:

Table 4: Summary of major mathematical symbols and their meanings, organized by subsection.

Subsection	**Symbol**	**Meaning**
Data Description and Sediment Contamination Assessment	m	Number of sampled sites
	$X \in \mathbb{R}^{m\times 30}$	Elemental concentration matrix (30 chemical elements)
	$E \in \mathbb{R}^{m\times 5}$	Environmental variable matrix (5 variables)
	$T \in \mathbb{R}^{m\times 16}$	Taxa abundance matrix (16 taxa)
	$s \in \mathbb{R}^m$	Composite stressor score (from PCA)
	$p\%$	Percentage of least stressed sites chosen as reference
	$I_{\text{ref}} \in \mathbb{R}^m$	Indicator: 1 = reference site, 0 = disturbed site
Reference site clustering and Discriminant Function	$\mathcal{C}_K$	Cluster label (taxa composition group) from reference sites
	$\hat{\mathcal{C}}_K$	Predicted cluster label for disturbed sites
	$\mathcal{F}_{\text{dis}}$	Discriminant function mapping E_{ref} to $\mathcal{C}_K$
	$\delta T_{i,j}$	Taxa community structure relative to the scale of reference site
	$\delta X_{k,j}$	Stress level relative to group-k reference median
Multivariate Gaussian modeling for constructing Zoobenthic Community Index (ZCI)	ϕ_{Hel}	Hellinger transformation
	$\mathcal{R}_k$, $\mathcal{D}_k$	Sets of reference and disturbed sites in group k
	$\boldsymbol{\mu}_k$	Mean taxa composition vector for group k reference sites
	$\boldsymbol{\Sigma}_k$	Covariance matrix of taxa composition in group k
	λ	Ridge regularization term
	I_{16}	16×16 identity matrix
	$\tilde{T}_{k,j}$	Whitened deviation vector for site j in group k
	$\text{ZCI}_{k,j}$	Scalar Mahalanobis distance from reference centroid
	$\text{ZCI}_{k,j}^{(\text{diag})}$	Diagonal approximation ignoring correlations
	$\text{ZCI}_{k,j}^{(1)}, \text{ZCI}_{k,j}^{(2)}$	First two components of multi-dimensional ZCI
	$\text{ZCI}_{k,j}^{\star}$	0–100 scaled ZCI score
	V_k	PCA loading matrix from whitened reference deviations
Spatial basis expansion for spatial predictors	x_i, y_i	Spatial coordinates of site i
	$D \in \mathbb{R}^{m\times m}$	Euclidean distance matrix(symmetrical) between sites
	$A \in \mathbb{R}^{m\times m}$	Double-centered distance matrix from truncated distance matrix
	$S_{\text{sel}} \in \mathbb{R}^{m\times d}$	Selected eigenvectors for explaining spatial variation ($d < m$)
Quantile regression modeling	$\mathcal{F}_k$	Regression function linking ZCI and spatial features to stress level
	$\delta X \mid Z, S_{\text{sel}}$	Relative stress level given ZCI and spatial features
	$Q^{(k)}_{\delta X\mid Z,S_{\text{sel}}}(\tau \mid z, s)$	Conditional τ-quantile of δX given ZCI and spatial features
	$f_{k,\tau}(z, s)$	Quantile regression function for group k at quantile τ
	κ_m	Fixed breakpoint in piecewise regression
	$\gamma^{(k)}_{m,\tau}$	Slope change after breakpoint κ_m
	$\hat{\theta}^{(k)}_{\tau}$	Estimated parameter for quantile regression
Hypothesis testing for degradation	$F^{(k)}_{\delta X\mid Z}(x \mid z)$	Conditional CDF of stress level given ZCI
	x^*_k	Group-k stress threshold for degradation classification
	p	One-sided p-value for degradation test

At the initial stage, the whole information about the sites can be shown in the matrix form:

$$\begin{bmatrix} X & E & T \end{bmatrix} \in \mathbb{R}^{m\times(30+5+16)}$$

where $X \in \mathbb{R}^{m\times 30}$ (elemental concentrations), $E \in \mathbb{R}^{m\times 5}$ (environmental variables), and $T \in \mathbb{R}^{m\times 16}$ (taxa abundances).

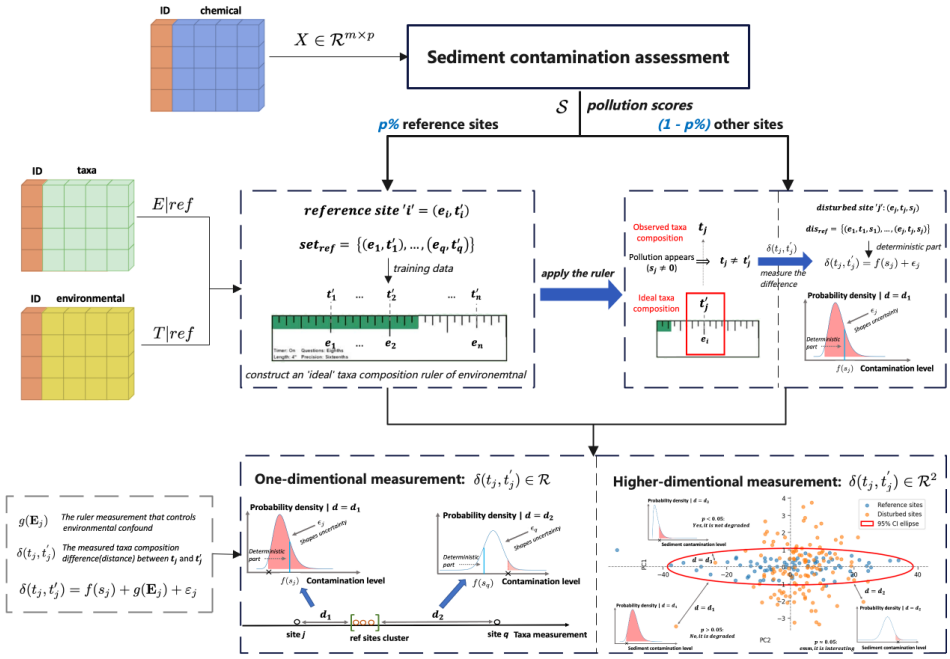

Figure 3: *Overview of workflow for the proposed methodology.*

5.1 Find Reference Sites - Sediment Contamination Assessment

External Reference: Data-driven PCA-based Pollution Assessment

Detailed methodology for sediment contamination assessment has been developed separately. The method report that is under development is available at:

- Click to access: Method report draft (Elsevier format)

To assess the sediment contamination and find the reference sites, we need to compute a composite stressor score s based on the chemical data.

Let m be the number of sampled sites and $X \in \mathbb{R}^{m \times 30}$ denote the matrix of chemical element concentrations (each row represents a site and each column represents an element). Doing a principal component analysis (PCA) on X transforms it into a set of uncorrelated and high variation-loading components Z. On top of the Z, we can select $k(< 30)$ proper components with defined criteria to cover the most variation in pollutant elements and define a composite stressor score $s \in \mathbb{R}^m$ by summing or weighting the selected raw principal components or their normalised variants:

1. **Principal component reduction** – Apply principal component analysis (PCA) to X. PCA transforms X into a set of uncorrelated components $Z = XW$, where $W \in \mathbb{R}^{30 \times k}$ holds loadings of the first k principal components.

2. **Composite stress score** – Let $Z = [\, z_1, \ldots, z_k \,]$ with $z_i \in \mathbb{R}^m$ the vector of scores on the i-th principal component. Define a composite stressor score $s \in \mathbb{R}^m$ by summing or weighting the selected raw principal components:

$$s_j \;=\; \sum_{i=1}^{k} \omega_i \, z_{i,j}, \qquad j \in \{1, \ldots, m\}$$

where $z_{i,j}$ is the i-th PC score at site j and ω_i are predetermined weights (often set to 1 when components contribute equally).

After computing the composite stressor score, we can add this new information to the originally compound matrix:

$$\begin{bmatrix} X & E & T & s \end{bmatrix} \in \mathbb{R}^{m \times (51+1)}$$

This s vector is used to rank the sites with respect to the stress level and filter the pristine reference sites where human impact is minimal or absent. Specifically, we rank sites by s and retain the least-stressed $p\%$ of the sites, create an indicator vector $I_{\text{ref}} \in \mathbb{R}^m$ where $I_{\text{ref},j} = 1$ if site j is a reference site and $I_{\text{ref},j} = 0$ otherwise.

$$\left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} \end{array}\right] \in \mathbb{R}^{m\times(52+1)}$$

To this sites with $I_{\text{ref},j} = 1$, we assume they represent the ideal taxa composition that is shaped by the given environmental conditions, supported by the minimal or absent human disturbance.

$$\left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} \end{array}\right]_{I_{\text{ref}}=1} \in \mathbb{R}^{(p\%\times m)\times(53)}$$

Therefore, in this submatrix, the X matrix only contains the minimal $p\%$ stress levels across all sites, controlling the human disturbance on the taxa composition.

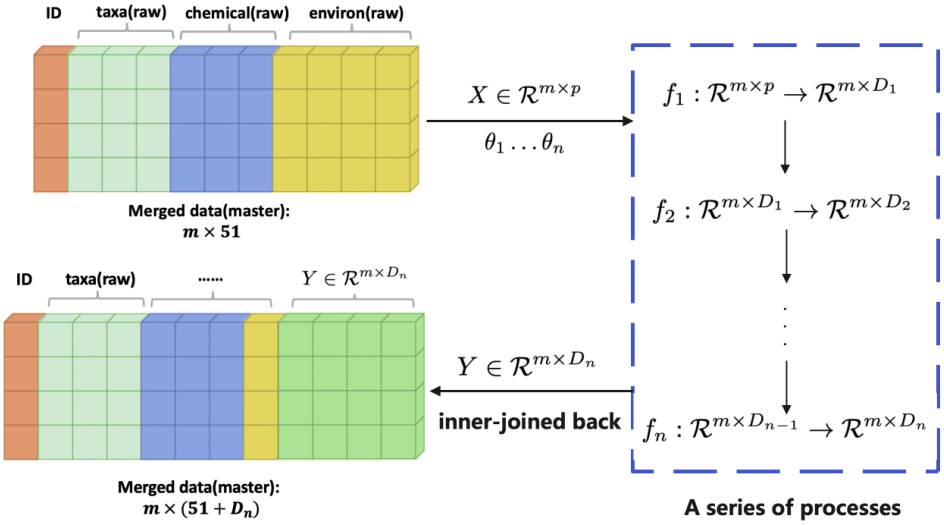

Figure 4: *Visualization of how the new information is generated and integrated into the existing matrix.*

5.2 Prepare metrics of 'ideal' taxa composition - Cluster Analysis on References

In the matrix $\left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} \end{array}\right]_{I_{\text{ref}}=1}$, the set of taxa composition T_{ref} is assumed to be shaped by the environmental variables E_{ref}, a well-fitted regression model between the E_{ref} and T_{ref} matrices can numerically tell us how the taxa composition is multidimensionally shaped by the environmental variables.

However, considering that the $E_{\text{ref}} \in \mathbb{R}^{(p\%\times m)\times 5}$ only provides 5 environmental variables, and there are many other potentially unmeasured and unmeasurable environmental factors, it is nearly impossible to train a fully quantitative inference model that describes the below relationship well:

$$\mathcal{F} : E_{\text{ref}}^{(p\%\times m)\times 5} \to T_{\text{ref}}^{(p\%\times m)\times 16}, \quad \text{poorly fitted model}$$

To solve this issue, we can construct constrained predicted values $T_{\text{ref}}^{q}(q < 16)$ from the T_{ref} matrix, which provides limited yet information about the community structure, so that the model $\mathcal{F} : E_{\text{ref}}^{(p\%\times m)\times 5} \to T_{\text{ref}}^{(p\%\times m)\times q}$ can be trained to avoid overfitting and improve its prediction performance.

$$\mathcal{F} : E_{\text{ref}}^{(p\%\times m)\times 5} \to T_{\text{ref}}^{(p\%\times m)\times q}, \quad \text{improved fitted model}$$

One ideal way to do this information compression is to partition the reference sites into K different groups via clustering methods.

$$T_{\text{ref}}^{(p\%\times m)\times q} = \mathcal{C}_K^{(p\%\times m)\times 1} = clustering(T_{\text{ref}}^{(p\%\times m)\times 16}), \quad \text{where } q = 1$$

By the clustering analysis and merging the resulting information into the reference-base matrix, the reference-base matrix can be updated as:

$$\left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \mathcal{C}_K \end{array}\right]_{I_{\text{ref}}=1} \in \mathbb{R}^{(p\%\times m)\times(53+1)}$$

Even though the C_K is computed from the clustering analysis on taxa composition matrix T_{ref}, the underlying environmental conditions(E_{ref}) are the actual drivers to lead to the clustering results, based on the fundamental assumption that "the reference taxa-composition is shaped by the environmental conditions".

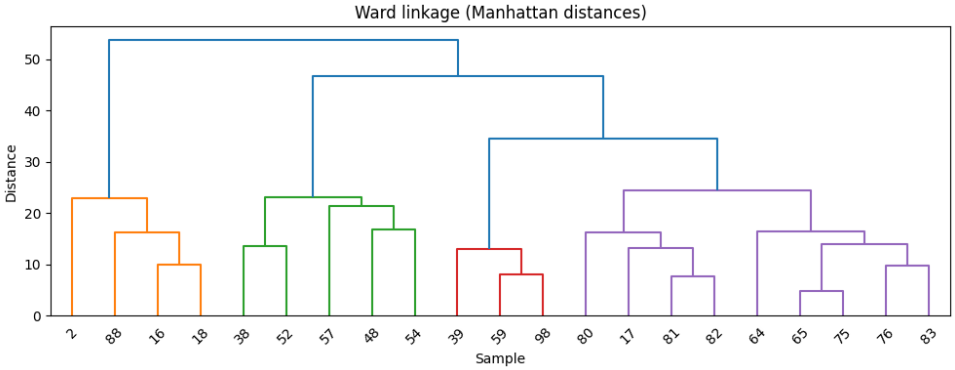

Figure 5: *An example of hierarchical clustering results on the taxa composition matrix of the references with selected clustering number K.*

5.3 Construct 'ideal' taxa composition ruler of environmental factors - Fit a Discriminant Function

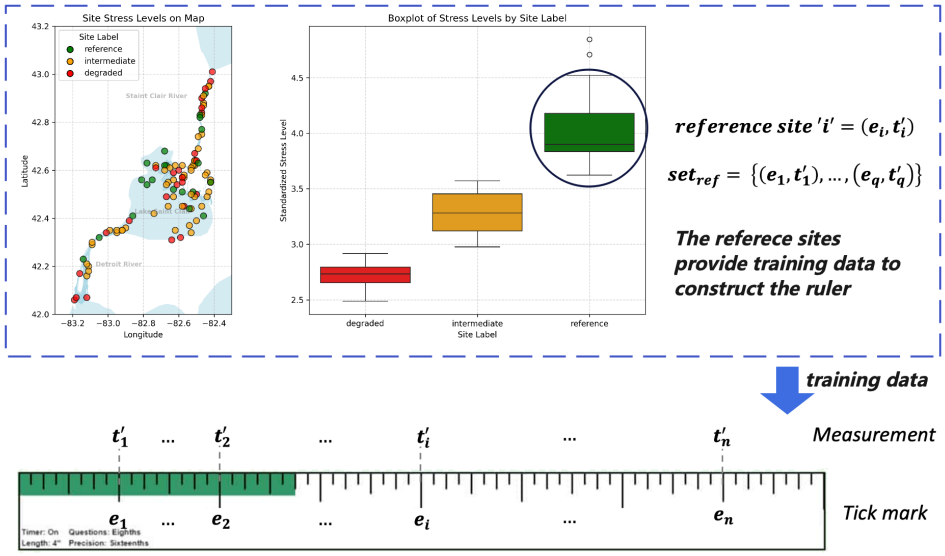

Figure 6: *The reference sites are used as training data to construct this 'ideal' taxa composition ruler. (t'_j represents the raw taxa composition of site j, it has not been transformed into the cluster label $\mathcal{C}_K$ yet.)*

At this stage, there are constrained taxa composition information - cluster labels $\mathcal{C}_K$ that can be used as response variables in training the environmental-taxa composition regression. Specifically, a discriminant function can be fitted here:

$$\mathcal{F}_{\text{dis}} : E_{\text{ref}}^{(p\%\times m)\times 5} \rightarrow \mathcal{C}_K^{(p\%\times m)\times 1}$$

This discriminant function $\mathcal{F}_{\text{dis}}$ fitted on the reference sites tells us how the environmental variables - E roughly shape the taxa composition by assigning each site to one of the taxa composition groups $\mathcal{C}_K$.

During the training stage, the reference sites are partitioned into the K taxa composition groups, helping to fix the group positions in taxa composition space with the pristine taxa composition part in each group. **However, it does not mean there is only pristine taxa composition in each cluster. When human disturbance appears, the pristine taxa composition should be shifted to a new position in the taxa composition space, which is how the disturbed sites look like in the same taxa composition space.**

Therefore, these reference sites are partitioned (by clustering) into different clusters to play the role of 'ideal' metrics on a ruler of taxa composition (by discriminant function), this ruler measures the 'ideal' taxa composition structure that a site should have given its environmental conditions.

An imaginable scenario is that, when we use the fitted $\mathcal{F}_{\text{dis}}$ as a ruler to measure the taxa composition of sites that are affected by human disturbance, the measured 'ideal' taxa composition is not equal to the truly observed taxa composition. And this difference in taxa composition is caused by the human disturbance, which was measured by the sediment contamination assessment in the previous step.

5.4 Mark the 'ideal taxa composition' for disturbed sites - Apply the Discriminant Function

Given the fitted discriminant function $\mathcal{F}_{\text{dis}}$, we can classify the rest $1 - p\%$ of the sites into the taxa composition groups, where reference sites with similar environmental conditions are already assigned into.

Because the clustering analysis was done on the reference sites, the known information on the disturbed sites should look like:

$$\left[\begin{array}{ccccc} X & E & T & s & I_{\text{ref}} \end{array}\right]_{I_{\text{ref}}=0} \in \mathbb{R}^{((1-p\%)\times m)\times(53)}$$

After applying the discriminant function on these disturbed sites, we can know their environmental-deterministic taxa composition groups, $\mathcal{C}_K^{((1-p\%)\times m)\times 1}$. It expands the disturbed-base matrix to:

$$\left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \hat{\mathcal{C}}_K \end{array}\right]_{I_{\text{ref}}=0} \in \mathbb{R}^{((1-p\%)\times m)\times(53+1)}$$

Compare it with the reference-base matrix, we can see that the sites having the same taxa composition cluster $\mathcal{C}_K$ are now comparable with the control of environmental variables E.

To the i th site in the matrix:

$$\left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \mathcal{C}_K \end{array}\right]_{I_{\text{ref}}=1} \in \mathbb{R}^{(p\%\times m)\times(53+1)}$$

If the site has $\mathcal{C}_{K_i} = \hat{\mathcal{C}}_{K_j}$, then the i-th reference site is comparable with the disturbed site j-th site in the taxa composition space with the control of environmental conditions. The difference in their taxa composition, $\delta T_{i,j}$, is caused by the human disturbance, $\delta X_{i,j}$, between the two sites.

$$\mathcal{C}_{K_i} = \hat{\mathcal{C}}_{K_j} \Rightarrow E_{\text{ref},i}^{(1\times 5)} \approx E_{\text{dis},j}^{(1\times 5)} \Rightarrow \delta T_{i,j} = \mathcal{F}_{reg}(\delta X_{i,j})$$

Therefore, the sites within the same taxa composition group will be used to fit the regression model - $\delta T_{i,j} = \mathcal{F}_{reg,k}(\delta X_{i,j})$, and these completed groups can be found through the merging-dismantle process of the two base matrices.

Merging the reference-base matrix and the disturbed-base matrix:

$$\text{stack}\left(\left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \mathcal{C}_K \end{array}\right]_{I_{\text{ref}}=1}, \left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \hat{\mathcal{C}}_K \end{array}\right]_{I_{\text{ref}}=0}\right)$$
$$\Rightarrow \left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \hat{\mathcal{C}}_K \end{array}\right]$$

Split the merged matrix into K submatrices, where each submatrix contains the same cluster label $\mathcal{C}_k$:

$$\left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \hat{\mathcal{C}}_K \end{array}\right] = \begin{cases} \left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \mathcal{C}_1 \end{array}\right] & \text{if } \mathcal{C}_K = 1 \\ \left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \mathcal{C}_2 \end{array}\right] & \text{if } \mathcal{C}_K = 2 \\ \vdots & \vdots \\ \left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \mathcal{C}_K \end{array}\right] & \text{if } \mathcal{C}_K = K \end{cases}$$

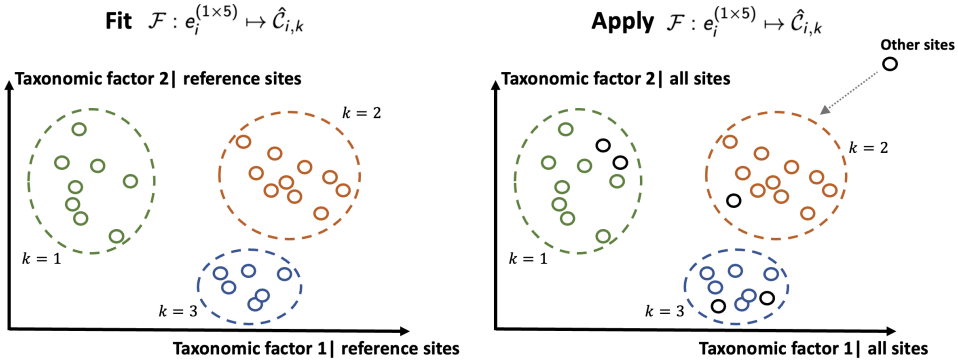

Figure 7: *Visualization of the fitting and application of the discriminant function that assigns disturbed sites to the environmentally determined taxa composition groups.*

Within each submatrix, $\begin{bmatrix} X & E & T & s & I_{\text{ref}} & \mathcal{C}_k \end{bmatrix}$, we will numerically measure the difference in the taxa composition between the degraded sites and the reference sites, δT_k, this distance in taxa composition will be explained by the relative stress level of each site, δX_k.

5.5 Measure the difference from 'pristine' to 'true' taxa composition - Multivariate Gaussian Deviation Index

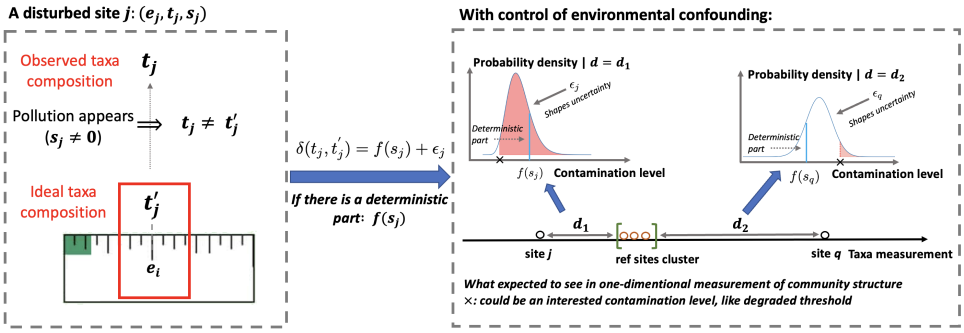

Figure 8: *The difference in taxa composition between the observed(t_j) and ruler measured(t'_j) is connected to the sediment contamination level(s).*

Within each taxa-composition group $\mathcal{C}_k$, let $\mathcal{R}_k$ denote the set of reference sites ($I_{\text{ref}} = 1$) and $\mathcal{D}_k$ the set of disturbed sites ($I_{\text{ref}} = 0$). We construct a site-level deviation metric that quantifies how far a site's observed community is from the pristine expectation of its group while controlling for environmental setting via $\mathcal{C}_k$.

Because taxa compositions are multivariate and often compositional/zero-inflated, we first work on a transformed scale using the Hellinger transformation:

$$\phi_{\text{Hel}} : \mathbb{R}^{16}_{\geq 0} \to \mathbb{R}^{16}, \quad \phi_{\text{Hel}}(\mathbf{t}) = \left(\sqrt{\frac{t^{(1)}}{\sum_{\ell=1}^{16} t^{(\ell)}}}, \ \cdots, \ \sqrt{\frac{t^{(16)}}{\sum_{\ell=1}^{16} t^{(\ell)}}} \right).$$

This transformation converts each taxon abundance to the square root of its relative abundance, reducing the influence of highly dominant taxa while preserving ecological distance relationships. All subsequent quantities are computed on $\phi_{\text{Hel}}(T)$; to simplify notation we overwrite $T \leftarrow \phi_{\text{Hel}}(T)$.

There are other transformations that may be preferred depending on the context (e.g., log-ratio transforms, or raw counts), the Hellinger transformation is tentative and can be replaced as needed.

After the transformation, for group k, compute the reference centroid and covariance

$$\boldsymbol{\mu}_k = \frac{1}{|\mathcal{R}_k|} \sum_{i \in \mathcal{R}_k} T_i, \qquad \boldsymbol{\Sigma}_k = \text{Cov}\{T_i : i \in \mathcal{R}_k\} + \lambda I_{16},$$

where $\lambda > 0$ is a small ridge term to ensure invertibility and numerical stability, and I_{16} is the 16×16 identity matrix. These parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ define a multivariate Gaussian-like distribution in the 16-dimensional taxa space, representing the *pristine community cloud* for group k. Under this view, each reference site is a draw from $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, and the geometric shape of this cloud is an ellipsoid whose orientation and size are determined by $\boldsymbol{\Sigma}_k$.

5.5.1 Value-based Measurement: Z-score Community Index (ZCI)

For any site j in group k (reference or disturbed), define the multivariate standardized deviation from the pristine centroid as the Mahalanobis distance:

$$\mathrm{ZCI}_{k,j} \;=\; \sqrt{\left(T_j - \boldsymbol{\mu}_k\right)^{\top} \boldsymbol{\Sigma}_k^{-1} \left(T_j - \boldsymbol{\mu}_k\right)}.$$

This measures how far T_j lies from the center of the pristine Gaussian cloud, in units that account for both taxon-specific variability and cross-taxon correlations. It effectively reduces the 16-dimensional deviation vector to a *single scalar score* while preserving the anisotropic geometry of the reference distribution.

When a diagonal approximation is preferred, use the "sum of squared z-scores" variant:

$$\mathrm{ZCI}_{k,j}^{(\mathrm{diag})} \;=\; \sqrt{\sum_{\ell=1}^{16} \left(\frac{T_j^{(\ell)} - \mu_k^{(\ell)}}{\sigma_k^{(\ell)}} \right)^2},$$

where $\sigma_k^{(\ell)}$ is the reference standard deviation of taxon ℓ in group k (robust alternatives such as median absolute deviation may also be used). This ignores inter-taxon correlations, treating the pristine cloud as an axis-aligned hypersphere, which can be more stable when the number of reference sites is small relative to the number of taxa.

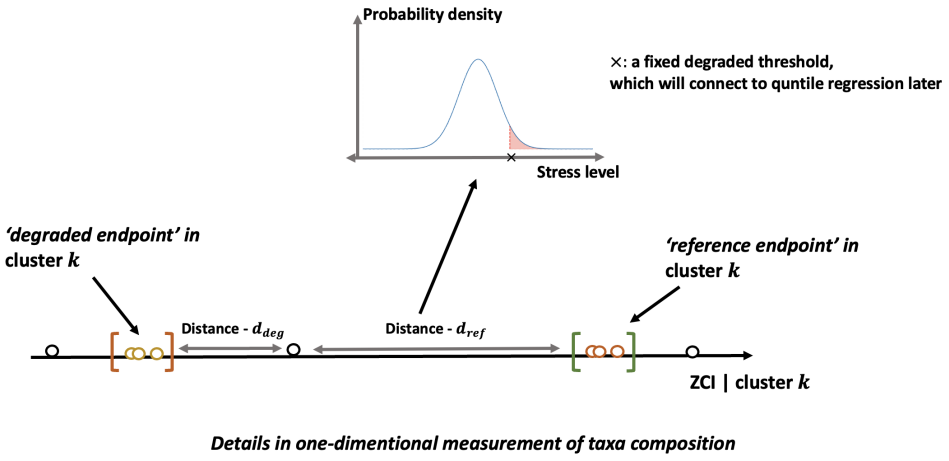

Figure 9: *Visualization of the details of taxa community structure differences measured in one-dimensional ZCI.*

5.5.2 Vector-based Measurement: multi-dimensional ZCI

The scalar $\mathrm{ZCI}_{k,j}$ summarizes deviation magnitude but discards the *direction* of change in community composition. To retain more structure, the same Gaussian framework can be used to construct a multi-dimensional ZCI:

1. **Whitening of deviations** [6] **:** For each site j in group k, compute the whitened deviation vector

$$\tilde{T}_{k,j} = \boldsymbol{\Sigma}_k^{-1/2}(T_j - \boldsymbol{\mu}_k),$$

[6] Whitening means: Centering (subtracting μ_k), rescaling and rotating so that the reference covariance becomes the identity matrix. Knowing that $\Sigma_k = \frac{1}{|T|-1}(T-\mu)^T(T-\mu)$, replacing the T with $\tilde{T} = \Sigma_k^{-1/2}(T-\mu)$, the new covariance matrix $\tilde{\Sigma}_k$ becomes $\frac{1}{|\tilde{T}|-1}(\tilde{T}-\tilde{\mu})^T(\tilde{T}-\tilde{\mu}) = I$. Here, T and μ are both matrices and Σ_k is a non-singular matrix.

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are estimated from the reference sites. Denote the matrix of whitened deviations for *reference* sites as $\tilde{T}_{k,\mathrm{ref}} \in \mathbb{R}^{n_{\mathrm{ref},k} \times 16}$. In this whitened space, the reference cloud is isotropic and centered at the origin.

2. **PCA fitted on whitened reference sites:** Perform principal component analysis (PCA) on $\tilde{T}_{k,\mathrm{ref}}$ to obtain the loading matrix V_k. Retain the first d principal axes $V_{k,(1:d)}$, where $d = 2$ gives a two-dimensional ZCI.

3. **PCA applied to disturbed sites:** For each disturbed site j, compute its whitened deviation $\tilde{T}_{k,j}$ using the *same* $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k^{-1/2}$ from the reference sites, and project it onto the retained principal axes:

$$\left(\mathrm{ZCI}_{k,j}^{(1)}, \mathrm{ZCI}_{k,j}^{(2)}\right) = \tilde{T}_{k,j}\, V_{k,(1:2)}.$$

These coordinates preserve both magnitude and orientation of deviation in the most informative subspace of the pristine community cloud, enabling more nuanced comparisons between sites that have similar scalar ZCI values but differ in the *type* of community shift. The scalar $\mathrm{ZCI}_{k,j}$ can be recovered as the Euclidean norm of these coordinates.

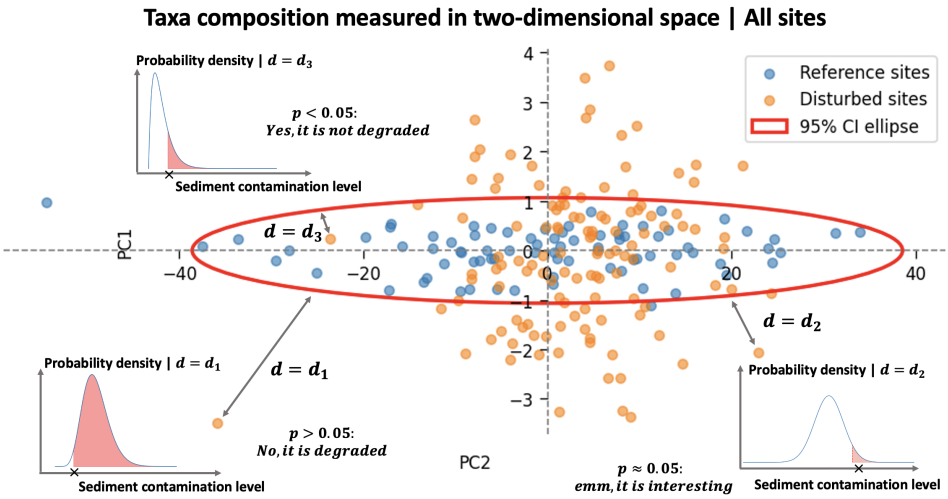

Figure 10: *Visualization of the details of taxa community structure differences measured in two-dimension ZCI.*

5.5.3 Direction, interpretation, and optional 0–100 scaling.

By construction, smaller values indicate communities closer to the pristine expectation for their environment; larger values indicate stronger deviation (putative impact). For reporting, we optionally map ZCI to a condition scale where larger is better:

$$\mathrm{ZCI}_{k,j}^{\star} \;=\; 100\,\big(1 - \widehat{F}_k(\mathrm{ZCI}_{k,j})\big),$$

with $\widehat{F}_k$ the empirical CDF of ZCI computed from *reference* sites in group k. Under this calibration, reference sites cluster near higher scores (closer to 100), while progressively disturbed sites trend toward 0.

5.6 Principal Coordinates of Neighbour Matrices (PCNM) for spatial eigenvectors

To adjust the ZCI–stress relationships for residual spatial structure, we derive spatial eigenvectors (PCNM variables) that capture multi-scale spatial autocorrelation among the same set of m sites in a specific cluster. These act as candidate covariates prior to fitting the piecewise quantile regression model.

Let (x_j, y_j) be the planar (or projected) coordinates for site $j = 1, \ldots, m$. Let $\mathbf{1}$ be an m-vector of ones, $\mathbf{I}_m$ the $m \times m$ identity, and the centering matrix $\mathbf{J} = \mathbf{I}_m - \frac{1}{m}\mathbf{1}\mathbf{1}^T$.

To derive spatial eigenvectors, we first compute Euclidean (or hydrologic, if river network) distances to form the $m \times m$ distance matrix $\mathbf{D}$ with entries:

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

Next, we choose a connectivity threshold d_0 as the maximum edge length of the minimum spanning tree to ensure a connected neighbour graph. Alternative approaches include using the maximum nearest-neighbour distance.

We then construct a truncated distance matrix $\mathbf{T}$ by defining

$$T_{ij} = \begin{cases} D_{ij} & \text{if } D_{ij} \le d_0 \\ 4d_0 & \text{otherwise} \end{cases}$$

where the large constant enforces separation of non-neighbours.

The PCoA transform is applied by forming $\mathbf{A} = -\frac{1}{2}\,\mathbf{J}\,\mathbf{T}^{\circ 2}\,\mathbf{J}$ (where $^{\circ 2}$ denotes elementwise square), followed by eigen-decomposition $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ with eigenvalues $\lambda_1 \ge \cdots \ge \lambda_m$ and eigenvectors $\mathbf{v}_k$.

$$\mathbf{A} = -\tfrac{1}{2}\,\mathbf{J}\,\mathbf{T}^{\circ 2}\,\mathbf{J} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T,$$

We retain eigenvectors with positive eigenvalues $\lambda_k > 0$ and scale them as $\mathbf{s}_k = \mathbf{v}_k\sqrt{\lambda_k}$. These orthogonal PCNM vectors span spatial patterns from broad (large λ_k) to fine scales. For screening, we compute Moran's I for each $\mathbf{s}_k$ using a binary (or inverse-distance) weight matrix based on d_0, retaining only spatially autocorrelated vectors using FDR or adjusted p-values. Forward selection based on AIC or adjusted R^2 can be applied to avoid overfitting.

Finally, we collect the retained spatial vectors in $\mathbf{S}_{\text{sel}} \in \mathbb{R}^{m \times q_s}$ and add these to subsequent ZCI quantile regression models (using the same $\mathbf{S}_{\text{sel}}$ across τ for comparability). We test residual Moran's I and iterate if needed.

5.7 Build ZCI indicator of sediment contamination levels – Piecewise Quantile Regression Model

The ZCI score reflects the degree of deviation in taxa composition from the pristine expectation within each group k, given the same environmental context. Because both the stress level and the community condition are influenced by a range of measured and unmeasured factors, it is reasonable to model the conditional distribution of the stress level *given* the community departure. This regression is used only as a statistical association to infer likely stress levels from observed ZCI values and does **not** imply a causal relationship between stress and community departure.

Within each group k, we relate the relative stress level to the ZCI deviation *and* the selected spatial eigenvector predictors derived earlier. Let

$$z_{k,j} := \text{ZCI}_{k,j}, \qquad \mathbf{s}_{k,j} \in \mathbb{R}^{q_s} \text{ be row } j \text{ of } \mathbf{S}_{\text{sel}} (q_s \text{ retained PCNM vectors}).$$

We model

$$\delta X_{k,j} = \mathcal{F}_k\big(z_{k,j},\, \mathbf{s}_{k,j}\big) + \varepsilon_{k,j},$$

where $\delta X_{k,j}$ is the relative stress (e.g., $s_j - \text{median}\{s_i : i \in \mathcal{R}_k\}$). Using the *same* spatial basis $\mathbf{S}_{\text{sel}}$ across all quantiles τ preserves comparability of slope and breakpoint inference by ensuring spatial adjustment does not vary with τ.

Given potential nonlinearity in z and heteroscedasticity, we use a piecewise linear quantile formulation in z while treating spatial terms additively:

$$Q^{(k)}_{\delta X|Z,S}(\tau \mid z, \mathbf{s}) = f_{k,\tau}(z, \mathbf{s}) = \beta^{(k)}_{0,\tau} + \beta^{(k)}_{1,\tau} z + \sum_{m=1}^{M} \gamma^{(k)}_{m,\tau}(z - \kappa_m)_+ + \sum_{r=1}^{q_s} \alpha^{(k)}_{r,\tau} s^{(r)},$$

with breakpoints $\kappa_1 < \cdots < \kappa_M$ placed on the ZCI axis only (spatial covariates are not segmented). Parameters solve the check-loss minimization

$$\widehat{\boldsymbol{\theta}}^{(k)}_{\tau} \in \arg\min_{\boldsymbol{\theta}} \sum_{j \in \mathcal{C}_k} \rho_\tau\Big(\delta X_{k,j} - f_{k,\tau}(z_{k,j}, \mathbf{s}_{k,j})\Big), \qquad \rho_\tau(u) = u\{\tau - \mathbf{1}(u < 0)\}.$$

The fitted conditional quantile surface is then

$$\widehat{Q}^{(k)}_{\delta X|Z,S}(\tau \mid z, \mathbf{s}) = \widehat{\beta}^{(k)}_{0,\tau} + \widehat{\beta}^{(k)}_{1,\tau} z + \sum_{m=1}^{M} \widehat{\gamma}^{(k)}_{m,\tau}(z - \kappa_m)_+ + \sum_{r=1}^{q_s} \widehat{\alpha}^{(k)}_{r,\tau} s^{(r)}.$$

Including the spatial term ensures that breakpoint and slope interpretations are attributed to contamination-driven community departure rather than residual spatial patterning.

5.7.1 Hypothesis testing for degradation – Quantile-based threshold inference

In many applications, a binary classification of a site as "degraded" or "non-degraded" is more actionable than estimating its exact stress level. This decision problem can be formulated as a one-sided hypothesis test, conditioning on the observed community departure (ZCI):

- **Step 1 – Define degradation threshold.** For each group k, choose a stress threshold x^*_k (e.g., a regulatory limit or an ecologically relevant benchmark) that separates degraded from non-degraded sites.
- **Step 2 – Predict conditional stress distribution.** For a site j with observed $\mathrm{ZCI}_{k,j} = z$, use the fitted quantile regression model $Q_{\delta X|Z,S}(\tau|z, \mathbf{s})$ over a grid of quantile levels $\tau \in (0, 1)$ to approximate the conditional distribution $F^{(k)}_{\delta X|Z,S}(x|z, \mathbf{s})$. This is done by inverting the quantile function across τ.
- **Step 3 – Compute p-value for degradation.** The hypothesis test is:

  $$H_0 : \delta X_{k,j} \le x^*_k \quad \text{vs.} \quad H_a : \delta X_{k,j} > x^*_k.$$

  The conditional p-value is then

  $$p = 1 - F^{(k)}_{\delta X|Z,S}(x^*_k \mid z, \mathbf{s}),$$

  which represents the probability, given the site's ZCI and spatial context, that the stress level exceeds the degradation threshold.

If p is below a chosen significance level α (e.g., 0.05), the site is classified as degraded; otherwise, it is classified as non-degraded. This approach converts the regression output into a probabilistic decision rule while controlling for environmental setting via the group-specific model.

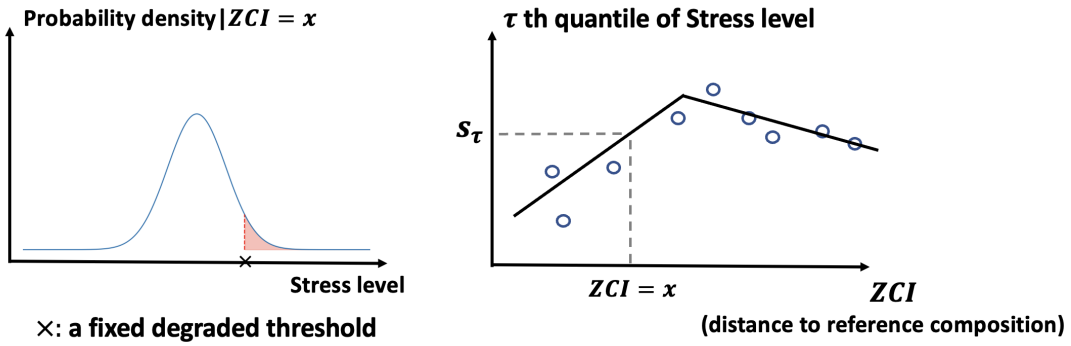

Figure 11: *A pre-fixed degradation threshold on the conditional stress level distribution and the correspondingly predicted quantile value* $\hat{s}_\tau$

5.8 Indicator power and robustness with respect to sample size (tentative)

Power and robustness evaluation of the indicator can be divided into the following specific aims: (i) quantify precision of slopes and breakpoint effects and conditional quantiles as sample size varies and (ii) estimate power for detecting contamination structure and degradation thresholds, all *within each group* k.

Targets: slopes $\beta^{(k)}_{1,\tau}$, changes $\gamma^{(k)}_{m,\tau}$, breakpoint reliability, pseudo-R^2, degradation test (Type I / power), and conditional quantiles $\widehat{Q}_{\delta X|Z,S}$.

Procedure (tentatively designed steps):

1. *Baseline fit:* Fit full piecewise quantile model (fixed κ_m, fixed S_{sel}) for $\tau \in \mathcal{T}$; store parameters and residuals; test residual Moran's I.

2. *Bootstrap (uncertainty):* If no residual spatial autocorrelation: site bootstrap; else block bootstrap. Refit using original S_{sel}; derive percentile CIs and relative widths.

3. *Subsampling (precision curves):* For a grid of reduced sizes $n_k^{(g)}$, draw R subsamples preserving ratio of $\frac{reference}{disturbed}$; refit; summarize bias and RMSE of slopes / predicted quantiles; locate diminishing returns size.

4. *Power simulation:* H_0: $\beta_{1,\tau} = \gamma_{m,\tau} = 0$; H_1: baseline (and reduced effect). Simulate L datasets holding (z, S_{sel}) fixed; estimate power for joint slope/breakpoint test and degradation classification at threshold x_k^*.

5. *Breakpoint reliability:* Accept κ_m if CI width < 0.3 of ZCI span *and* $\Pr(\gamma_{m,\tau} \neq 0 \text{ for some } \tau) \geq 0.8$; otherwise reduce segments or increase sample size.

6. *Outputs:* (a) precision and power curves vs. n_k; (b) recommended minimum n_k meeting power ≥ 0.8 (median quantile) and reliability rule; (c) table of median (IQR) parameter estimates.

Interpretation: Early plateau in precision implies reallocating effort to under-sampled groups or spatial gaps. If slopes are consistent but breakpoints unreliable, report continuous gradients instead of thresholds. Reliable, sharp breakpoints support categorical management triggers.

6 Preliminary exploration

In this section, I implemented a simplified completed workflow based on the methodology described in Section Methodology. It explored the practical application of the proposed method to support the later composite workflow. Specifically, the major preliminary steps are as follows:

1. **Collect comparable data.**

   There are three available datasets, containing the three raw data types: zoobenthic community data (311×16), chemical data (104×30), and environmental data (289×7). The three shared the same identical index - StationID, which supports data merging to prepare completed data for each site. After merging there is a combined dataset with 104 rows and 53 columns, containing all three types of data.

   Key Results: $\begin{bmatrix} X & E & T \end{bmatrix} \in \mathbb{R}^{m\times(30+5+16)}$ was prepared [7].

2. **Assess sediment contamination and Identify reference and degraded sites.**

   A log-transformation was applied to the chemical data to reduce dominance by high-value variables. Then PCA was performed on the transformed chemical data and several principal components were selected to explain the major variation of pollutant elements. By simply standardizing and summing these PCs, comprehensive stress values were computed for each site. To keep consistency with Jian's analysis, I used the name "SumReal" to refer these stress values (levels). Current statistics results shown "the higher are the stress scores, the less are the pollutant elements concentrations", but this will be further explored in the future.

   Based on the computed stress scores, $p\%$ was temporarily set to 20% to identify reference sites and the degraded sites were symmetrically defined. Out of assumption, there were no or minimal human disturbances on the reference sites, their taxa composition was shaped by environmental conditions only.

   Key Results: $\begin{bmatrix} X & E & T & s & I_{ref} \end{bmatrix} \in \mathbb{R}^{m\times(51+2)}$, was prepared

3. **Cluster reference sites by taxa community composition.**

   Turn to the zoobenthic community data of these references, IQR method was applied to detect outliers and then octave transformation was applied to reduce extreme values' impact, considering that taxa in low abundance do not mean they are not important.

[7] The column numbers were not consistent($51 \neq 53$) because StationID and location information were not used in the analysis. Later, the location information can be included to support spatial analysis.

Then clustering was applied to identify major taxa community patterns across different environmental conditions, and K clusters were identified. Here K was set to 3 through the hierarchical clustering method. These sites assigned into each k cluster represent "ideal" taxa structures, each totally shaped by the range of environmental conditions of the corresponding cluster.

Key Results: $\mathcal{C}_K$ and $\left[\begin{array}{cccccc} X & E & T & s & I_{ref} & \mathcal{C}_K \end{array} \right]_{I_{ref}=1} \in \mathbb{R}^{(p\%\times m)\times(53+1)}$, were prepared.

4. **Fit Discriminant Function of environmental factors for taxa clusters.**

   Based on the identified clusters of these reference sites, a discriminant function was fitted to predict the taxa cluster membership of each site based on its environmental variables. [8]

   Key Results: $\mathcal{F}_{dis} : e^{(1,5)} \to \hat{C}_K$ was fitted.

5. **Apply the Discriminant Function to rest disturbed sites to group them**

   To the (1 - p%) disturbed sites, including the degraded sites, $\mathcal{F}_{dis}$ was applied to assign each site to one of the identified taxa clusters.

   Key Results: $\mathcal{F}_{dis} : e^{(1,5)} \to \hat{C}_K$ was applied, $\left[\begin{array}{cccccc} X & E & T & s & I_{ref} & \hat{\mathcal{C}}_K \end{array} \right]_{I_{ref}=0}$ was prepared.

6. **Construct endpoints and compute ZCI in each cluster.**

   Within each cluster, the reference sites and degraded sites were used to construct endpoints, unlike the Multivariate Gaussian cloud of reference sites, the endpoints were simply computed as the means of 3 selected references and 3 degraded sites. The endpoints were used to numerically scale the taxa community structure and compute the Zoobenthic Condition Index via Bray-Curtis ordination method. The ZCI_k reflected the distance of the disturbed sites to the endpoints in cluster k.

   Key Results: $ZCI_k, (k \in 1, ..., K)$ was computed; $ZCI_{k,j}$ reflects the distance of site j to its cluster k endpoints.

7. **Evaluate the ZCI vs SumReal relationship by quantile regression.**

   On top of the ZCI and stress level s, a piecewise quantile regression was fitted to evaluate their relationship across the three taxa clusters. The breakpoints were found by grid search method with a pre-defined searching range, which was set manually based on the preliminary exploration.

   Key Results: $f_{k,\tau}(z), (k \in 1, ..., K)$ was fitted; $\hat{\theta}_\tau^{(k)}$ was solved for each cluster k.

8. **Preliminary exploration of spatial structure with simulation**

   An isolated PCNM (spatial eigenfunction) simulation was run, separate from previous workflow steps, to illustrate extraction of spatial structure across increasing complexity. Three scenarios were simulated (low to higher complexity): (i) equally spaced 1D sites, (ii) equally spaced 2D lattice, and (iii) irregular 2D sites with two clusters. The first two yield regular distance matrices and smoothly ordered (broad- to fine-scale) eigenvectors; the clustered irregular layout produces wave-like spatial patterns in the leading eigenfunctions.

   Key Results: $PCNM : D \to S_{\text{sel}}$ was carried out; S_{sel} comprises the Moran's I–screened spatial eigenvectors derived from the truncated neighbour distance matrix D.

9. **Power and sensitivity analysis on quantile regression with simulation**

   A deterministic piecewise quantile regression model with fixed breakpoint κ and heteroscedastic noise was used to generate synthetic data. Subsamples of sizes $n = 20, 50, 80, \ldots, 500$ were drawn to assess estimation precision of key parameters ($\hat{y}_\tau$, $\beta_{1,\tau}$, δ_τ). Bootstrap 95% CIs and precision curves (RMSE) were computed. Power analysis for detecting non-zero hinge effects ($\delta \neq 0$) was conducted via repeated simulation under alternative and null conditions.

   Key Results: Preliminary patterns show (i) diminishing RMSE gains beyond moderate n; (ii) wider CIs and reduced power at extreme quantiles; (iii) hinge (δ_τ) estimates stabilising more slowly than primary slope estimates. Several implementation and interpretation issues (e.g., behaviour at high τ with small n) were noted for later, more detailed exploration.

[8] It is both acceptable to say "environmental clusters" or "taxa clusters", owing to the assumption that these reference sites have their taxa composition totally shaped by the environmental conditions.

6.1 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 matrices can be merged by the station ID, which gives a comparable dataset with 104 rows and 53 columns.

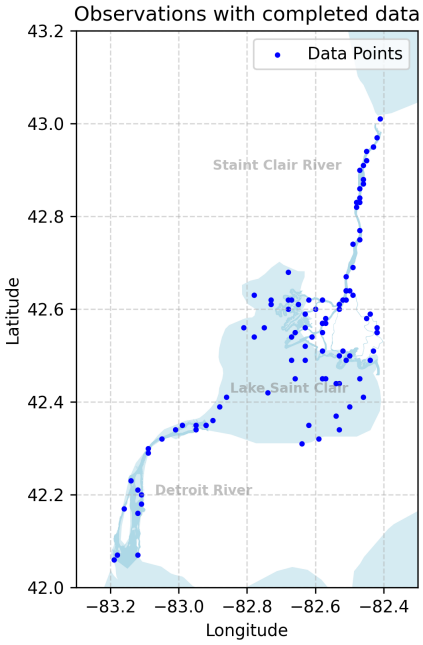

Figure 12: Distribution of the 104 sites with completed three types of data.

Figure 12 shows the distribution of the 104 sites with completed three types of data, these observations are generally evenly distributed across the Huron-Erie Corridor, which backs up the belief in sampling aspect that the top $p\%$ sites in stress level should have no or minimal human disturbances.

6.2 Assess sediment contamination and Identify reference and degraded sites

In this stage, 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 were computed for each principal component for the upcoming selecting PCs 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, these PCs were normalized 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, the normalized PCs were summed, considering the directionality of positive and negative loadings, to obtain the "SumReal" score[9], for each sample, which serves as a measure of sediment contamination.

[9] As done by Jian in her sediment contamination assessment

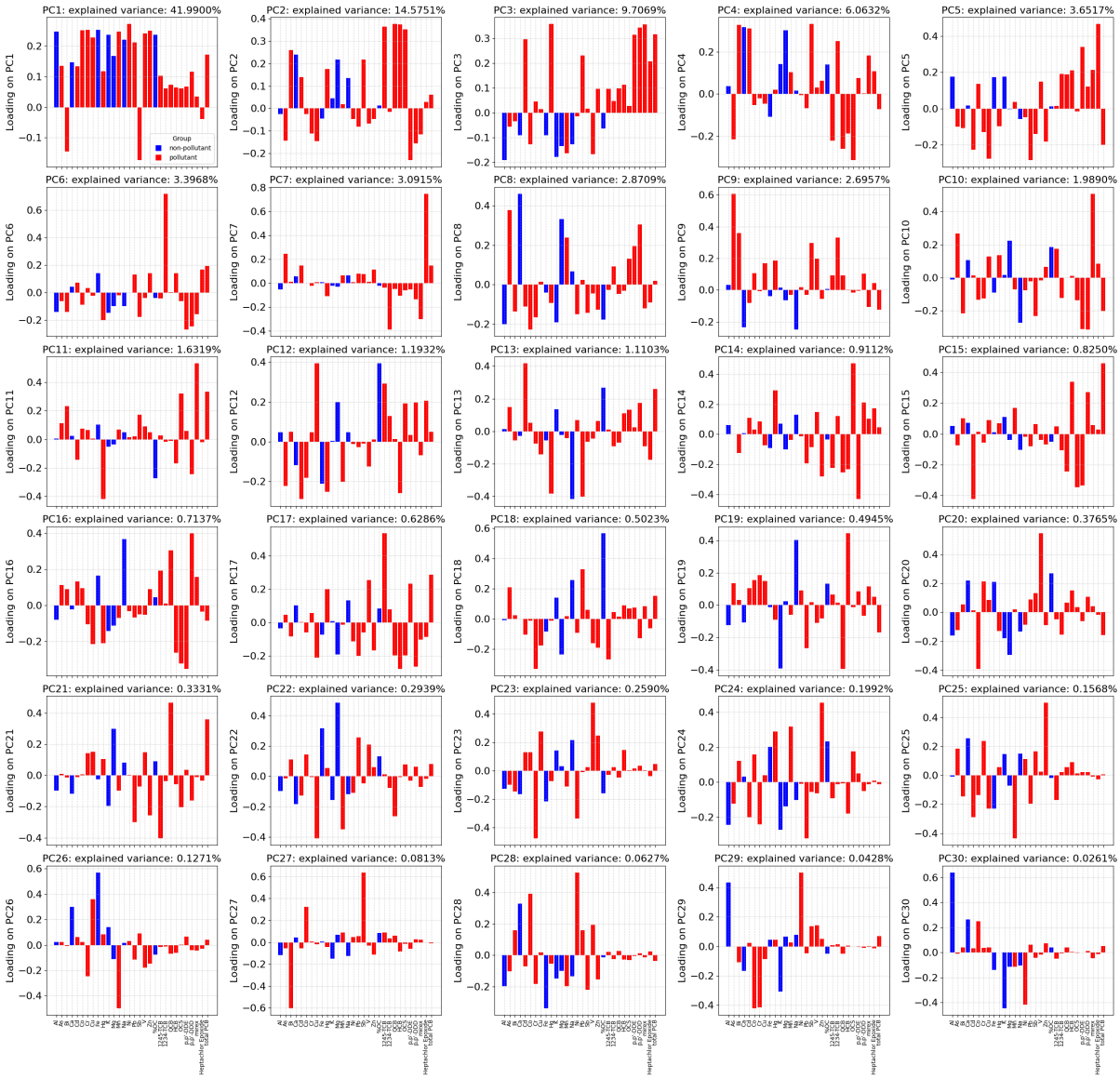

Figure 13: PCA loadings of the chemical data, showing the distribution of loadings across chemical variables

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

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 thought as determined completely by the environmental factors.

These references support building 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).

Figure 15 shows the box plots of taxa abundance across the three blocks. Even though it is not accurate to know the taxa composition across stress levels, but a good result is that the taxa abundance distribution is different across the types: degraded, intermediate and reference. Because such blocking operation controls the human disturbances to some extent, and we hope to see the taxa composition differences across these disturbance levels. Although Figure 15 might not provide a complete and satisfying picture, but it is acceptable considering that we have not controlled the environmental factors yet.

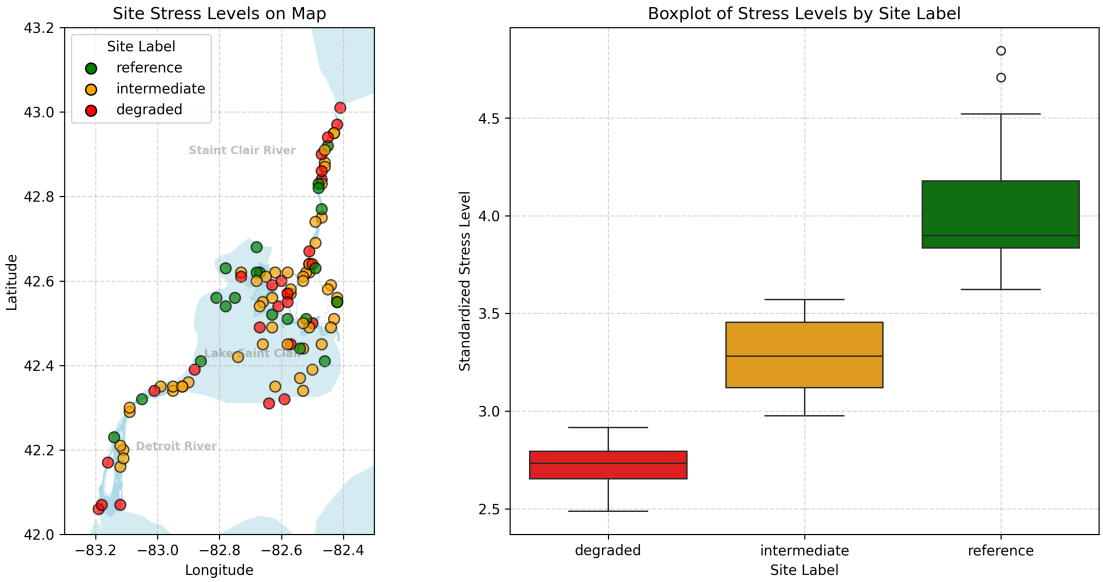

Figure 14: Map and boxplot of stress levels across three blocks of stress levels: reference(top 20% of stress scores), degraded(bottom 20% of stress scores) and others

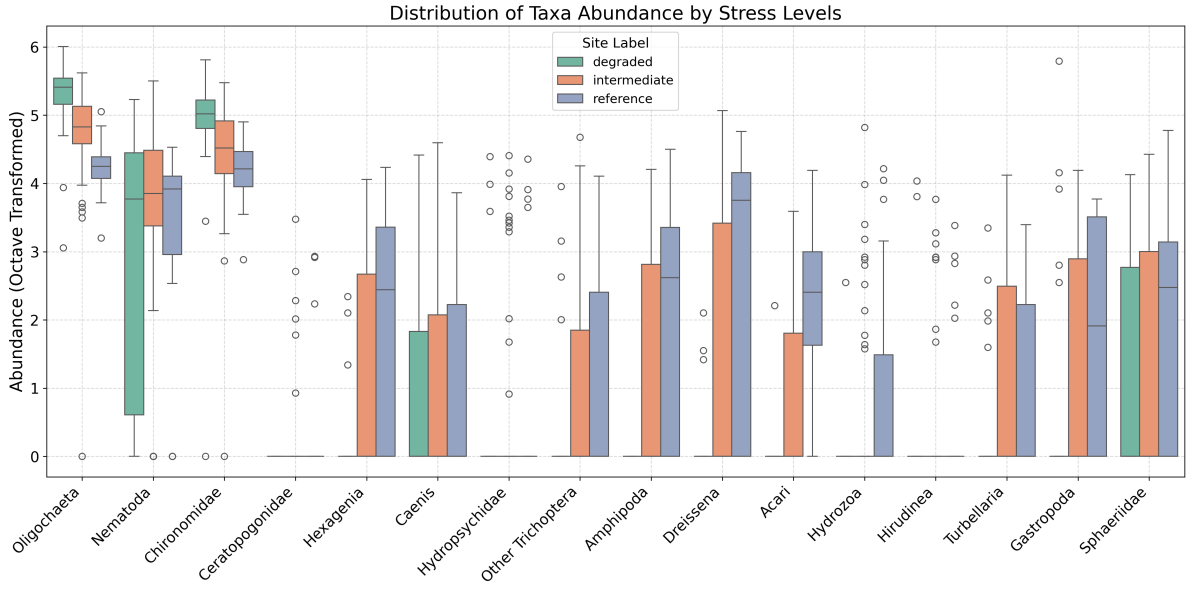

Figure 15: Taxa abundance distribution by site label, showing the potential taxa composition differences across three blocks of sediment contamination

6.3 Cluster reference sites by taxa community composition

In this step, IQR method was applied to detect outliers in the taxa data, and octave transformation was applied as suggested in Jian's analysis, to reduce the impact of outliers and balance the influence of all taxa.

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

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

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

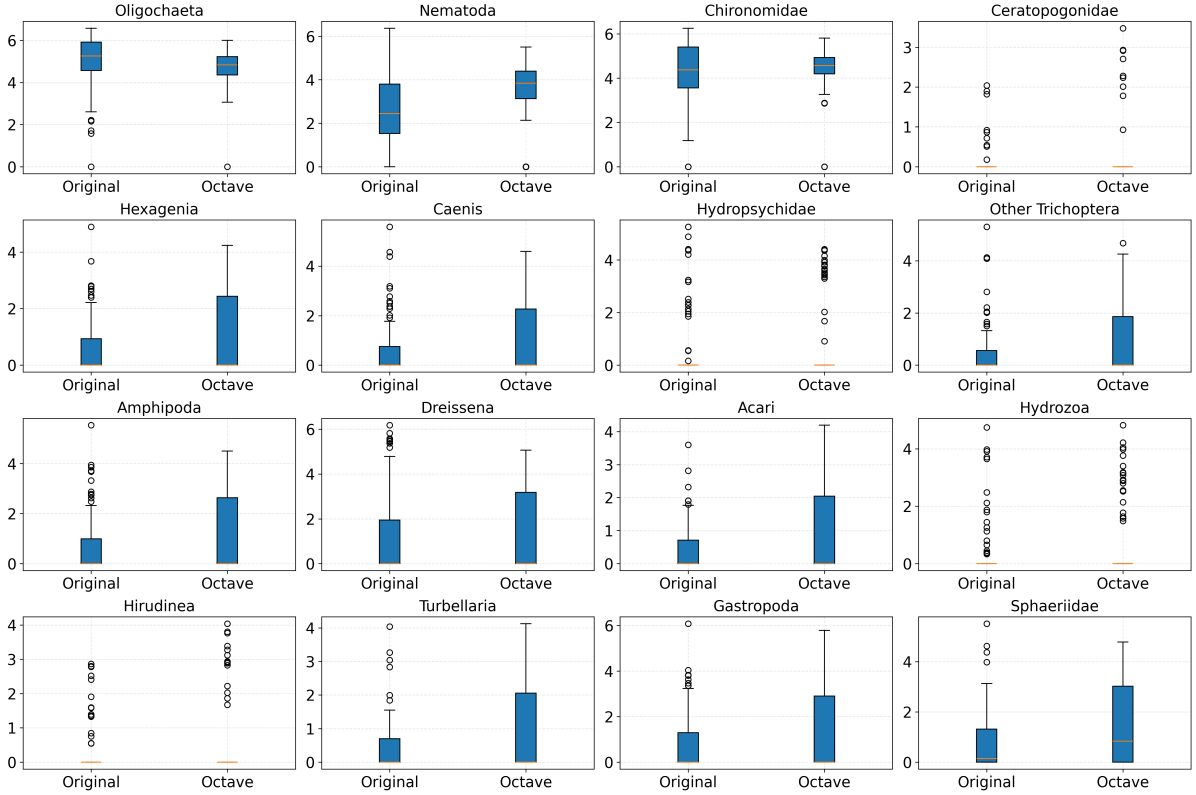

Figure 16: Boxplots of the original vs octave-transformed taxa data, showing reduced outliers and more balanced distribution.

Based on the octave-transformed taxa data, clustering was applied to identify major taxa community patterns across different environmental conditions. The clustering was performed using hierarchical clustering method, and the number of clusters K was set to 3 based on the dendrogram analysis.

Figure 17 is an example of the clustering results on references, and it is not the final clustering result. It was noticeable that one cluster (orange color) has only 2 reference sites in it, which is not enough to represent a stable "ideal" taxa composition and to fit a discriminant function for its cluster.

Considering it is in the preliminary stage, this example is presented here only for showing the clustering process and the potential taxa community patterns across different environmental conditions. A better clustering result needs to include enough reference sites in each cluster for constructing a stable "ideal" taxa composition benchmark and to support the upcoming fitting of a discriminant function, $f_{k,\tau}(z)$. However, the sampled environmental conditions are also important to be considered, a uniform sample-environment distribution or a proportional to the real world distribution is better for the clustering to be meaningful.

Figure 18 shows the taxa density distribution across clusters. The sites in each cluster are both controlled by the environmental conditions(clustering results) and the human disturbances(reference sites, no or minimal human disturbances), so we hope to see their taxa community compositions(density curves) difference across the current clusters.

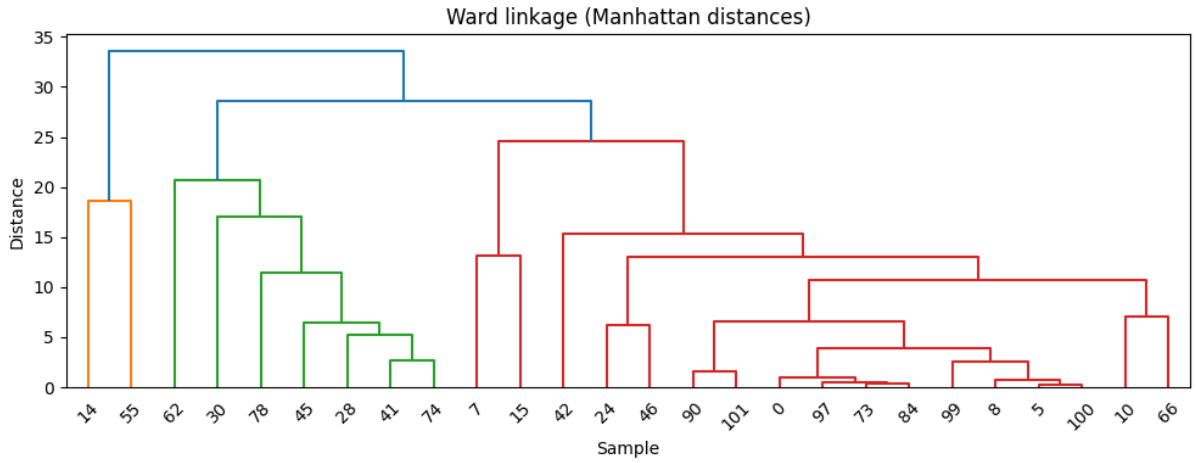

Figure 17: A possible example of clustering results of reference sites by taxa composition, showing the major groups of minimally disturbed taxa compositions.

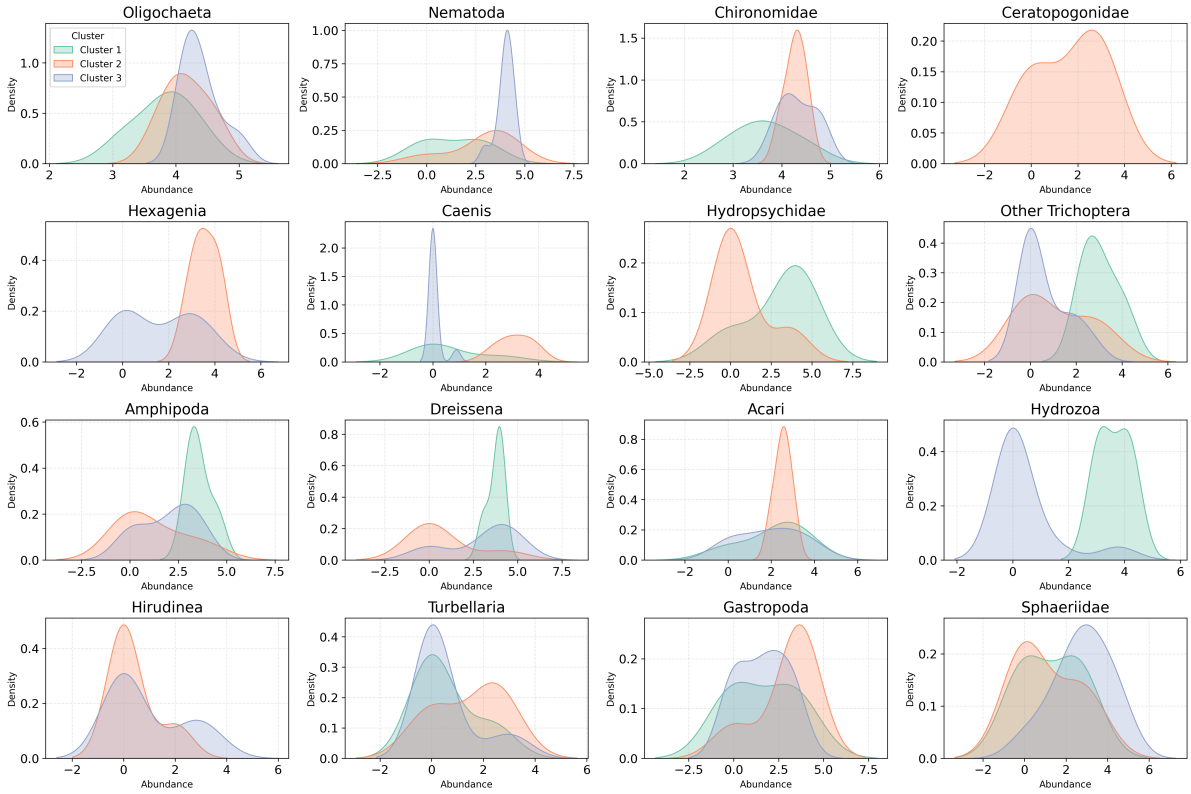

Figure 18: "Ideal" individual taxa density across clusters, showing the distribution of taxa across different clusters of reference sites. (The disappearance of the density curve of some taxa in a cluster means that the taxa is not present in the cluster.)

6.4 Fit Discriminant Function of environmental factors for taxa clusters

At this stage, a discriminant function(classifier) was fitted on these reference sites to predict cluster memberships(labels) with their environmental factors. Table 5 shows the coefficients of the fitted discriminant function, and Table 6 shows the classification report of the fitted discriminant function on the training set of reference sites.

Table 5: Discriminant Coefficients

	$Class_1$	$Class_2$	$Class_3$
Intercept	-121.154	-28.769	18.883
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	2.276	1.229	-0.689

Table 6: 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

6.5 Apply the Discriminant Function to rest disturbed sites

Given the fitted discriminant function, apply it on the rest disturbed sites with their environmental factors to group these sites into the existing taxa clusters.

In this preliminary analysis, in addition to the reference sites, there were degraded sites within each cluster. These degraded sites were used to construct the theoretical *worst* endpoints and the reference sites were used to construct the theoretical *best* endpoints. The two endpoints in each cluster have best and worst taxa compositions and highest and lowest stress scores, and they will be used in ordination method to construct the ZCI scores for other sites in the cluster.

6.6 Construct endpoints and compute Zoobenthic Condition Index (ZCI)

After all sites were assigned to clusters, endpoints were constructed by taking the mean of selected subset of reference and degraded sites in each cluster.

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

Table 7 shows the taxa compositions of the endpoints in each cluster, and the Figure 19 shows the distribution of ZCI scores across clusters.

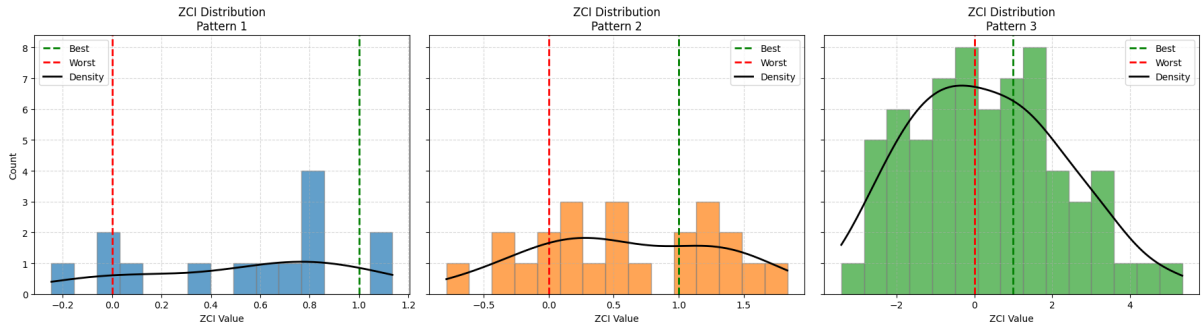

Figure 19: ZCI Distribution across Clusters, showing the variation of ZCI scores within each cluster.

The *best* and *worst* endpoints are used to perform Bray-Curtis ordination, which scales the taxa compositions given the endpoints and compresses the 16 taxa variables into a single value to reflect the distance to the endpoints. In each cluster, the resulting ZCI scores of the two endpoints were converted

Table 7: 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

to 0 and 1, and the ZCI scores of other sites were transformed with the same scale, which is not a [0 ,1] range mapping, but a value-based ZCI score that reflects the distance to the endpoints.

Figure 20 shows the scatter plot of ZCI scores vs stress scores and the box plot of taxa abundance across taxa clusters. One of the encouraging results is that the ZCI scores have correlation with the stress scores across three clusters, and the scatter plots in Figure 20 show very likely correlations between ZCI scores and stress scores.

And to the lower half of the Figure 20, we hope to see the taxa abundance distribution differences across the three clusters due to the different environmental conditions. But this boxplot is not comparable to the boxplot shown in Figure 15, because the Figure 15 is the taxa abundance distribution controlled by the human disturbance levels, while the Figure 20 shows the taxa abundance distribution controlled by environmental conditions. The differences in controlled conditions make it unreasonable to compare these two boxplots, even though they may look similar.

6.7 Evaluate the ZCI vs SumRel relationship by quantile regression

Now, there were complete ZCI scores and stress scores for all sites, and their relationships were evaluated by quantile regression across clusters.

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 quantile regression was applied to the ZCI scores and stress scores within each cluster to evaluate the relationship between them, a bias-corrected bootstrapping method was applied to estimate the confidence intervals of the quantile regression coefficients.

Figure 21 to Figure 23 show the quantile regression results of different τ values for each cluster. Stress score of 4, the horizontal red line, is marked as degraded threshold for all three clusters for convenience, and there are one or two visible quantile regression lines crossing the threshold value. The intersection of a τ^* level quantile regression line and the horizontal red line provides the estimated percent of samples conditioned on the corresponding ZCI score to be less than or equal to the degraded threshold.

The one-side p-value for null hypothesis that the relative stress level conditional on such ZCI score x_k^* is less than the degraded threshold $H_0 : \delta X < x_k^*$ can be computed as $1 - \tau^*$. As a potential example like Figure 21, quantile regression lines of $\tau = 1$ and $\tau = 0.77$ cross the degraded threshold at around $ZCI = 0.25$ and $ZCI = 0.8$ respectively, such results indicate that there are other quantile regression lines intersecting the degraded threshold between the ZCI scores of 0.25 and 0.8, and these quantile levels

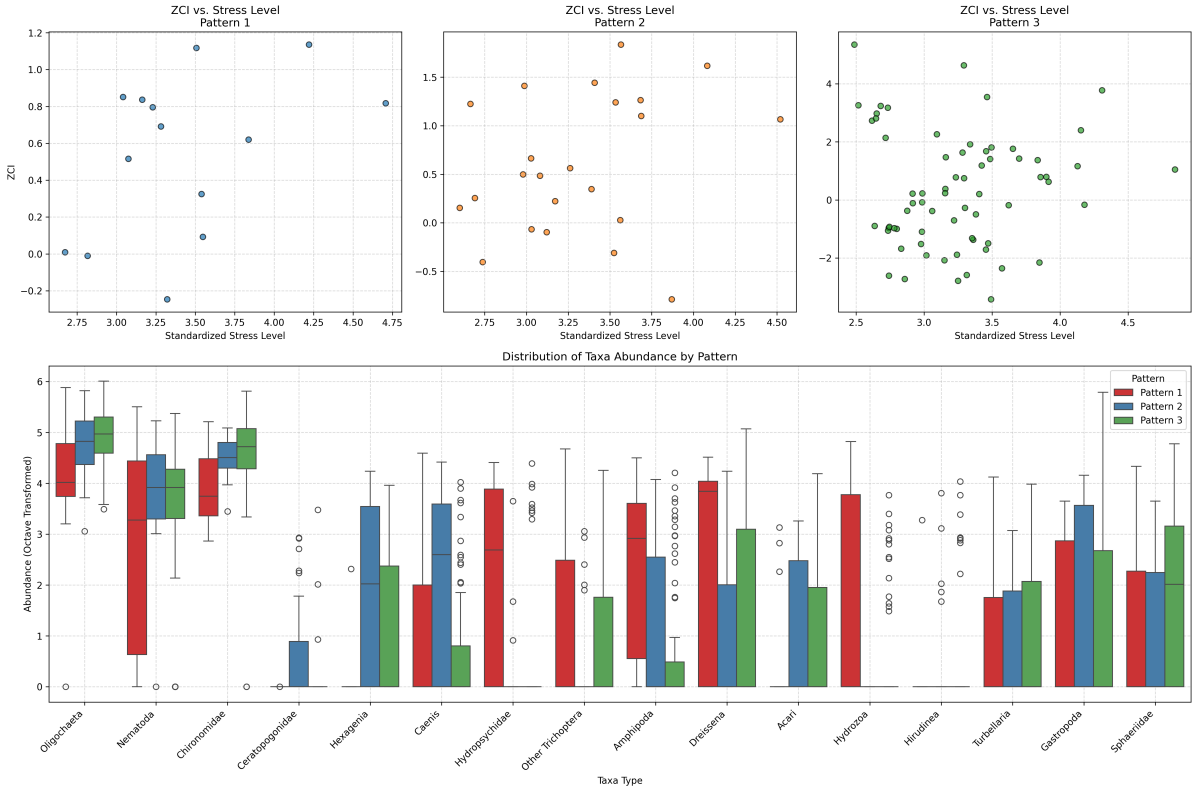

Figure 20: ZCI vs Stress Scores and the Distribution of Taxa Abundance across Taxa Patterns

must fall between 0.77 and 1.

In Appendix, there are piecewise quantile regression results for each cluster and other techniques, like bootstrap or synthetic data generation, can be applied to enhance our understanding toward this inference method on top of the collected original data.

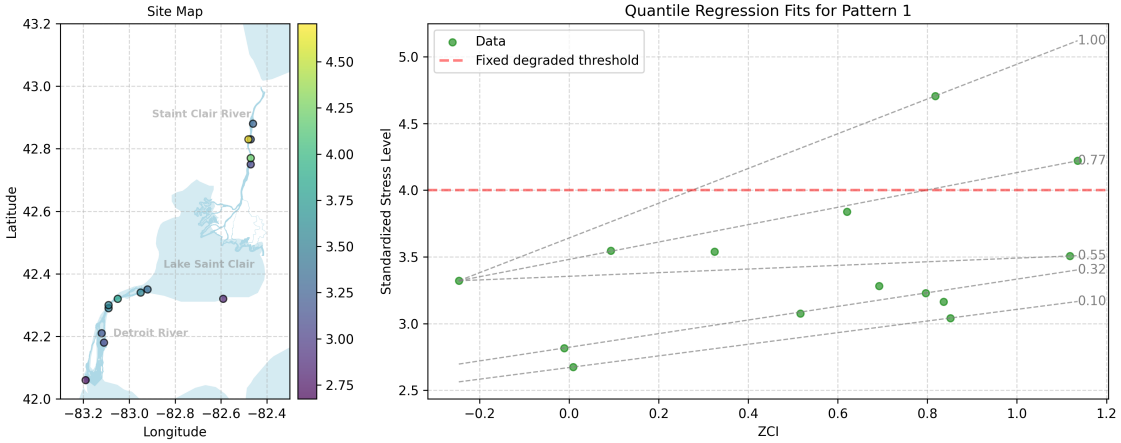

Figure 21: Quantile Regression Results for Cluster 1, showing the relationship between ZCI and stress scores.

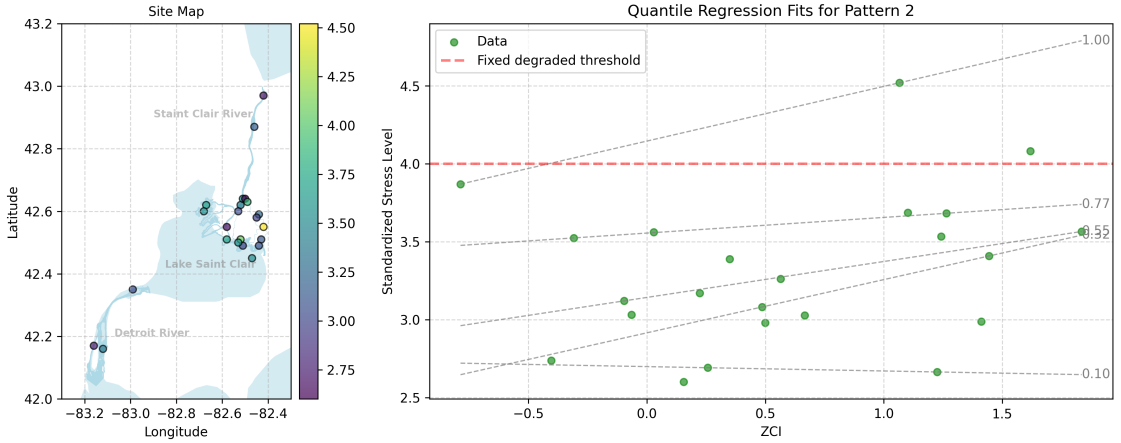

Figure 22: Quantile Regression Results for Cluster 2, showing the relationship between ZCI and stress scores.

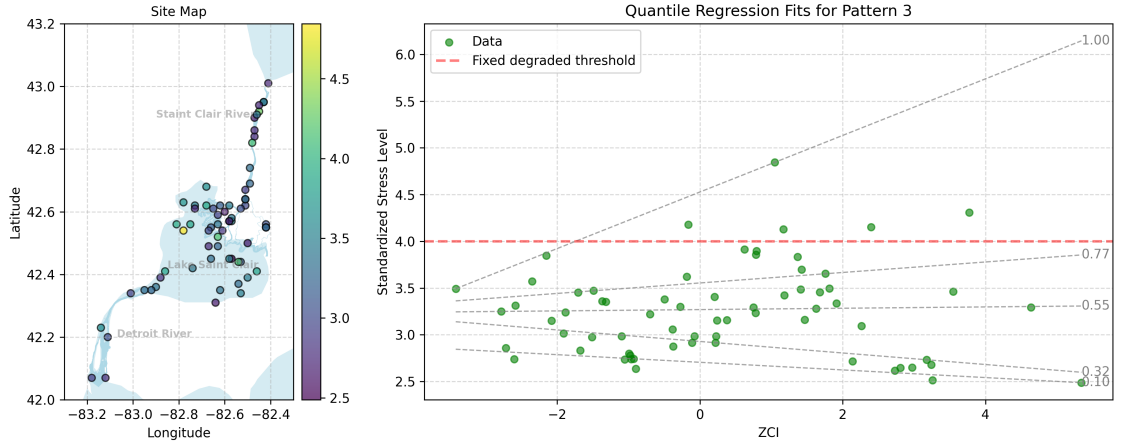

Figure 23: Quantile Regression Results for Cluster 3, showing the relationship between ZCI and stress scores.

6.8 Extract spatial eigenvector by PCNM on simulated data

PCNM (Principal Coordinates of Neighbour Matrices) provides a set of orthogonal spatial eigenvectors that can be used as covariates to model spatial structure (broad to fine scale) in subsequent analyses (e.g., quantile or generalized regression on taxa or stress scores). The workflow applied in the simulated data mirrors what will be done with true geographic information (projected site coordinates of the Huron-Erie Corridor and Detroit River sites):

1. Compute pairwise Euclidean (or great-circle then projected) distances among sites. Real coordinates will first be projected to an equal-area / appropriate planar CRS to avoid distortion.
2. Determine a truncation threshold (here: maximum edge length of the minimum spanning tree / maximum nearest-neighbour distance) to ensure the distance graph is fully connected while suppressing unrealistically long links.
3. Replace distances greater than the threshold with a large value (or leave as is in a truncated matrix) and convert to a truncated distance matrix.
4. Apply double-centering (Gower centering) to obtain a spatially informed matrix whose eigen decomposition yields orthogonal axes ordered from broad (large-scale) to fine (small-scale) spatial structures.
5. Retain eigenvectors with positive spatial autocorrelation (screened by Moran's $I > 0$ and, if needed, significance tests under permutation). These become the candidate spatial predictors (denoted S). A subset S_{sel} is chosen to avoid overfitting (e.g., forward selection with adjusted R^2 or information criteria) when added to ecological models.

The three simulation figures illustrate how sampling design influences the scale and appearance of retained eigenvectors:

Figure 24 (1D regular): With evenly spaced sites along a line, early (broad) eigenvectors resemble low-frequency sine/cosine waves capturing gradual longitudinal gradients. Progressively higher-order (not all shown) eigenvectors would introduce finer oscillations. The regular spacing produces a clean hierarchy of spatial scales.

Figure 25 (2D regular grid): On a lattice, broad-scale eigenvectors express smooth surfaces across both axes. Subsequent eigenvectors increasingly capture checkerboard or banded patterns that partition the grid into larger then smaller spatial patches. The symmetry of the grid yields balanced representation of directions, which helps interpret broad-scale ecological gradients (e.g., upstream–downstream or depth-related trends) when transposed to real data.

Figure 26 (2D irregular): Irregular (clustered / uneven) sampling causes eigenvectors to localize structure around denser site clusters and to stretch patterns across sparser regions. Broad-scale eigenvectors still reflect overall spatial extent, but intermediate components may appear asymmetric or warped because inter-site distances are uneven. This underscores the need, in the real dataset, to evaluate whether spatial coverage is sufficiently even; otherwise, model selection may prioritize localized eigenvectors that compensate for sampling gaps rather than true ecological gradients.

Application to real data: Once applied to actual site coordinates, we will generate candidate PCNM eigenvectors, screen for positive spatial autocorrelation, and incorporate a subset into quantile regression models. This aims to reduce spatial autocorrelation in residuals and improve inference on non-spatial covariates while enabling spatially explicit variance interpretation.

The simulations therefore validate that the procedure recovers interpretable multi-scale patterns under contrasting sampling schemes and highlight considerations (threshold choice, uneven spacing) prior to deploying PCNM on the empirical dataset.

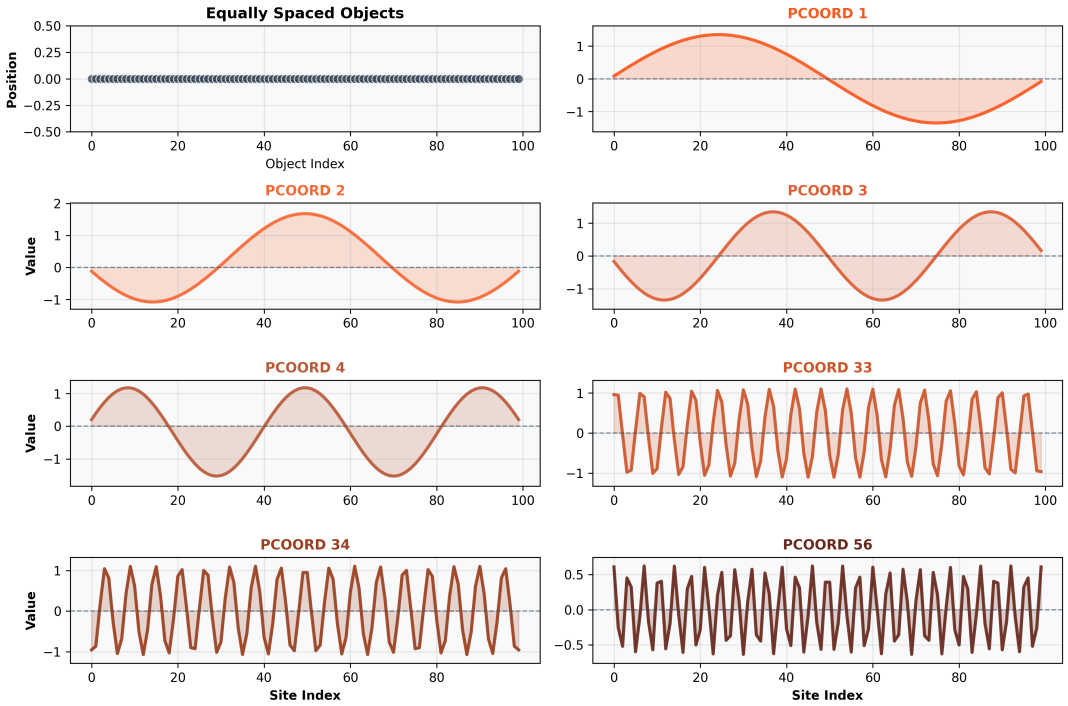

Figure 24: PCNM simulation results for regular sampling in 1D space, showing the original sites and selected spatial eigenvectors.

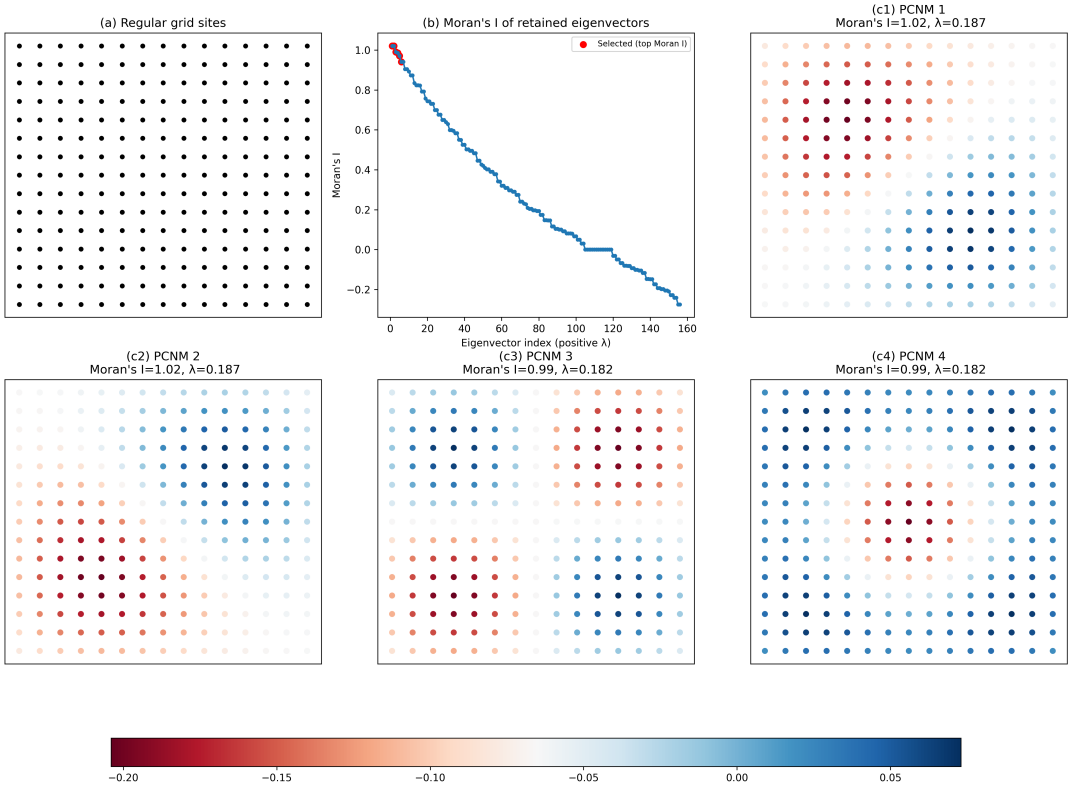

Figure 25: PCNM simulation results for regular sampling in 2D space, showing the original sites and selected spatial eigenvectors.

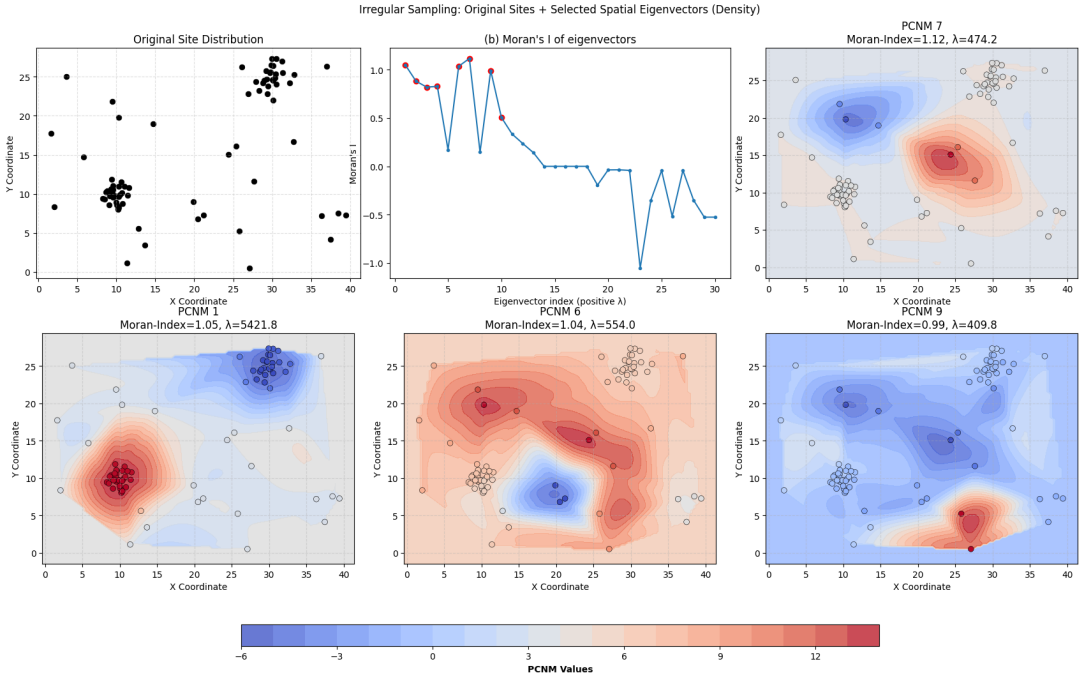

Figure 26: PCNM simulation results for irregular sampling in 2D space, showing the original sites and selected spatial eigenvectors.

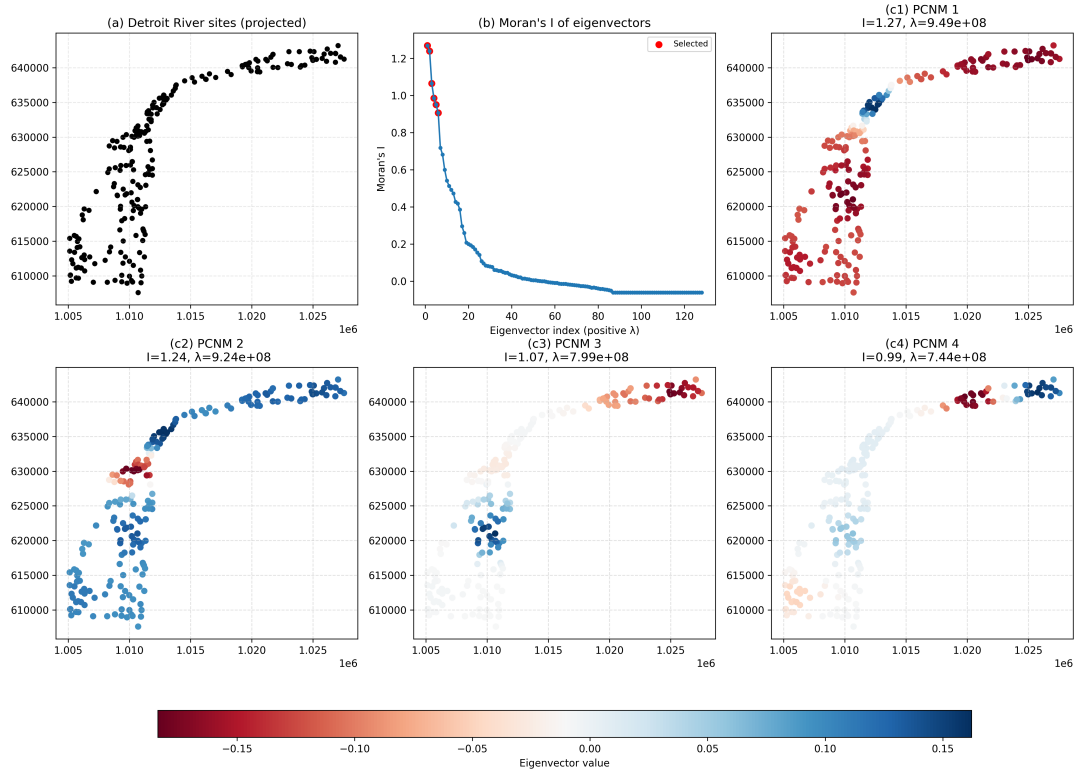

Figure 27: PCNM simulation results for Detroit River site coordinates, showing the original sites and selected spatial eigenvectors.

6.9 Power and sensitivity analysis on quantile regression with simulated data

This simulation evaluates two practical design questions for a piecewise (hinge) quantile regression intended for later ecological inference: (i) how precisely can the slope (β_1) and post-break change (δ) be recovered across quantile levels τ, and (ii) what sample size is required to attain acceptable power (≥ 0.8) to detect a true hinge effect at a high conditional quantile while controlling the Type I error when the hinge effect is absent.

Model and data generation. We generated responses under a fixed breakpoint ($\kappa = 5$) hinge model

$$Y = \beta_0 + \beta_1 X + \delta(X - \kappa)_+ + \varepsilon,$$

with parameters ($\beta_0 = 1.0$, $\beta_1 = 0.5$, $\delta = -0.6$) and heteroscedastic noise whose standard deviation increases linearly with X. A broad quantile grid $\tau \in \{0.05, 0.15, \ldots, 0.95\}$ allows inspection of effect heterogeneity across the conditional distribution.

Estimation and hinge effect uncertainty. For an initial sample ($n = 500$) we fit separate quantile regressions (fixed κ) and constructed bootstrap confidence intervals ($B = 300$) for the hinge coefficient δ_τ using shared resampling indices across all τ within each bootstrap replicate. Sharing indices stabilizes cross-τ comparisons of δ_τ and reduces Monte Carlo variability in the δ_τ profile.

Finite-population subsampling for precision. To evaluate how precision scales with added observations we first generated a large synthetic "population" ($n = 5000$) from the same model. Without replacement subsamples of sizes $n \in \{20, 50, 80, \ldots, 500\}$ were repeatedly drawn ($R = 40$) and refit to obtain RMSE curves for the slope $\beta_{1\tau}$ and hinge δ_τ relative to their generating values. This mimics augmenting an already characterized system rather than refitting to entirely new stochastic realizations, providing a conservative (finite-population) view of diminishing returns.

Power and Type I operating characteristics. For each candidate n, $L = 70$ fresh datasets were simulated under (a) the alternative ($\delta = -0.6$) and (b) the null ($\delta = 0$). For every τ we computed a Wald statistic $|\hat{\delta}_\tau|/\mathrm{SE}(\hat{\delta}_\tau)$ and declared detection if it exceeded 1.96 (two-sided $\alpha \approx 0.05$). Bootstrap resampling of detection indicators ($B = 350$) supplied 95% confidence intervals for the empirical detection probability (power) and false positive rate (Type I). All proportions and interval bounds were clipped to $[0, 1]$ for coherence. The high quantile $\tau = 0.95$ was designated the primary design target because upper conditional responses often reflect limiting processes of ecological interest; the smallest n achieving mean power ≥ 0.8 at $\tau = 0.95$ is reported as the recommended minimum sample size.

Key results (Figure 28 panels). (a) Fitted quantile functions (colored) closely envelop simulated data and track the deterministic mean (black), with clear post-break downward adjustment at higher X. (b) The δ_τ profile shows moderately consistent negative hinge effects across mid-to-upper quantiles; bootstrap intervals exclude 0 for higher τ, indicating stronger evidence of a breakpoint effect in the upper tail. (c) RMSE declines rapidly up to roughly $n \approx 200$–250, after which gains flatten, suggesting diminishing precision returns beyond this range for $\tau = 0.95$. (d) Power increases with n and crosses 0.8 at the recommended n (vertical dashed line), while the empirical Type I error under $\delta = 0$ remains near the nominal 0.05 across sample sizes, confirming adequate test calibration under the simulation settings.

Implications for study design. These results provide a defensible sample size target and demonstrate that (i) synchronised bootstrap resampling enhances interpretability of cross-τ hinge effect patterns, (ii) finite-population style subsampling yields realistic diminishing returns guidance, and (iii) the proposed Wald-based detection maintains nominal error rates. Prior to applying the framework to empirical ZCI–stress analyses, sensitivity analyses around κ mis-specification and noise structure can refine robustness expectations, but the current evidence supports proceeding with the indicated minimum n.

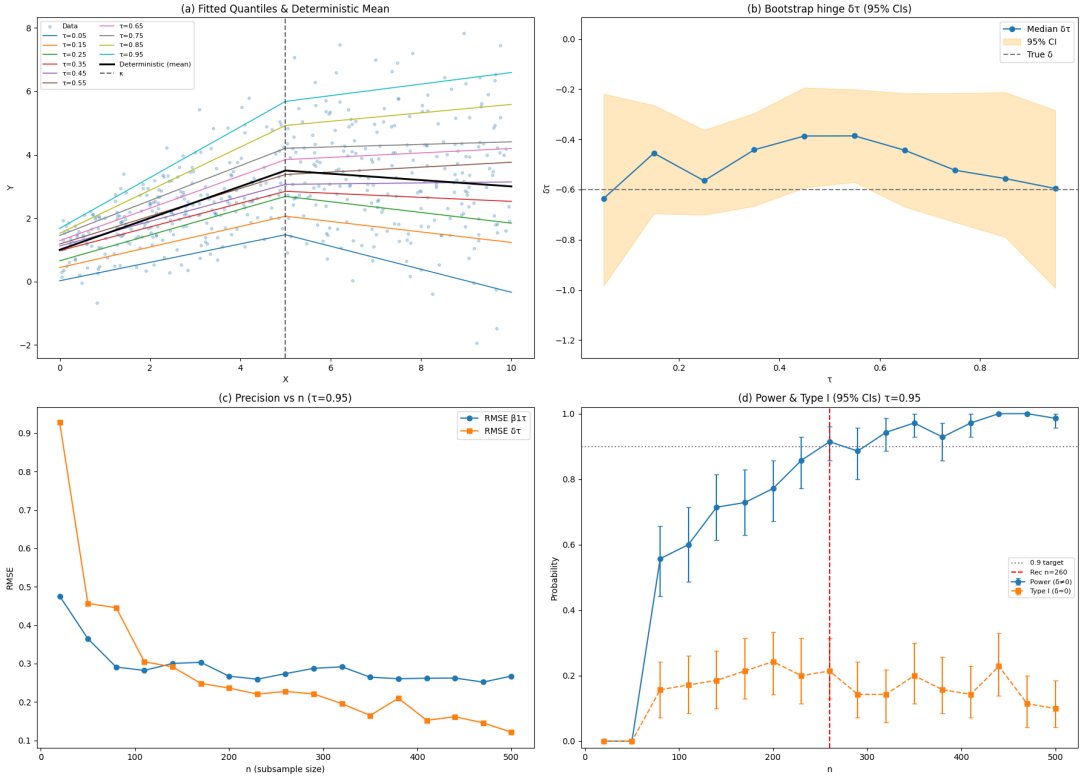

Figure 28: Power and Sensitivity Analysis on Quantile Regression with Simulated Data, showing the impact of sample size and noise level on the detection of significant quantile regression coefficients.

7 Practical Implementation Plan

This section outlines the specific implementation plan for the thesis project, detailing the project structure, work patterns and the benefits of this approach.

The current project structure is designed into 9 major folders, each with a specific storage function as briefly described in Table 8.

Table 8: Project folder structure and description

Folder	Description
`src/`	Core source code package (`ecoindex`) with reusable analysis modules.
`notebooks/`	Interactive Jupyter notebooks for prototyping and exploratory work that majorly achieves by modules from *src*.
`tests/`	Unit tests for packages wrote in `ecoindex`, using `pytest` to ensure correctness and reliability.
`configs/`	Centralized configuration (random seeds, parameters, file paths), collective containers for important parameters.
`data/`	Raw and processed datasets, organized for reproducibility.
`artifacts/`	Outputs not easily reproducible, e.g., trained models or serialized results.
`figures/`	Generated plots, charts, and visualizations for reporting.
`reference/`	External references, guidelines, and supporting documentation.
`documents/`	Thesis drafts, LaTeX files, lying close to results folders to keep synchronous with the results.

The following two subsections will elaborate on the principles and rationales behind this project structure design, and use a specific example to demonstrate the codebase design and working pattern.

7.1 Research Professional Codebases

To make a good research professional codebase, I want to follow the principles of modularity, reusability and maintainability. This means organizing code into reusable components, documenting functionality clearly, doing tests for all components, and ensuring that the codebase can be easily understood and modified.

Out of these principles, the code structure and organization become paramount in achieving the work. It has to be admitted that such structure would cost more time in setting up and analysing the meta data, significant amount of time when compared to directly writing the code in a single file. However, with the progress of the project, the benefits of this structure will become apparent and obviously accelerate the research process, which is visible and achievable to a non-computer science student with the support of these Generative AI tools.

The following example shows a quick starting point for the codebase, where I wrote and wrote some functions to achieve a given data transformation operation, wrapping it in a clear structured way into original data frame and testing the correctness of these operations.

7.1.1 Modularized Function implementation and lightweight application

As it was mentioned in the methodology part, I wanted to continuously add computed results, which were sitewise specific, into the original data frame meanwhile keeping the whole data frame tiny and efficient. Assuming I had the three raw data sets in my *data/raw* folder, and I wanted to do a hellinger transformation on the taxa data set and export the transformed data into the *data/processed* folder. To achieve this complete operation, I would need the following modules and functions within them:

The whole modules are not only designed for Hellinger Transformation, but it is a good example and there will be more functions and module added in this structure to make the project concise and reproducible.

Table 9: Temporary Code Structure for Hellinger Transformation and Integrity of Data Frame

Module	Description
src/config.py	Configuration management for centralized parameter storage and project settings, like file paths and processing options.
src/cleaning.py	Data cleaning utilities for handling missing values, outliers, and data quality issues.
src/ingest.py	Data ingestion functions for loading raw datasets from various sources and formats.
src/dataframe-ops.py	DataFrame manipulation utilities for wrapping new information into layered data frames
src/transform.py	Data transformation functions including ecological transformations like Hellinger.
src/pipeline.py	Pipeline orchestration for chaining multiple data processing steps together.
notebooks/0-interim-data.py	Notebook to invoke the above functions in practical scenarios, lightweight code snippets for quick testing and validation.
tests/test-.py*	Unit tests for validating the functionality of individual modules and functions in above modules(*src/**).

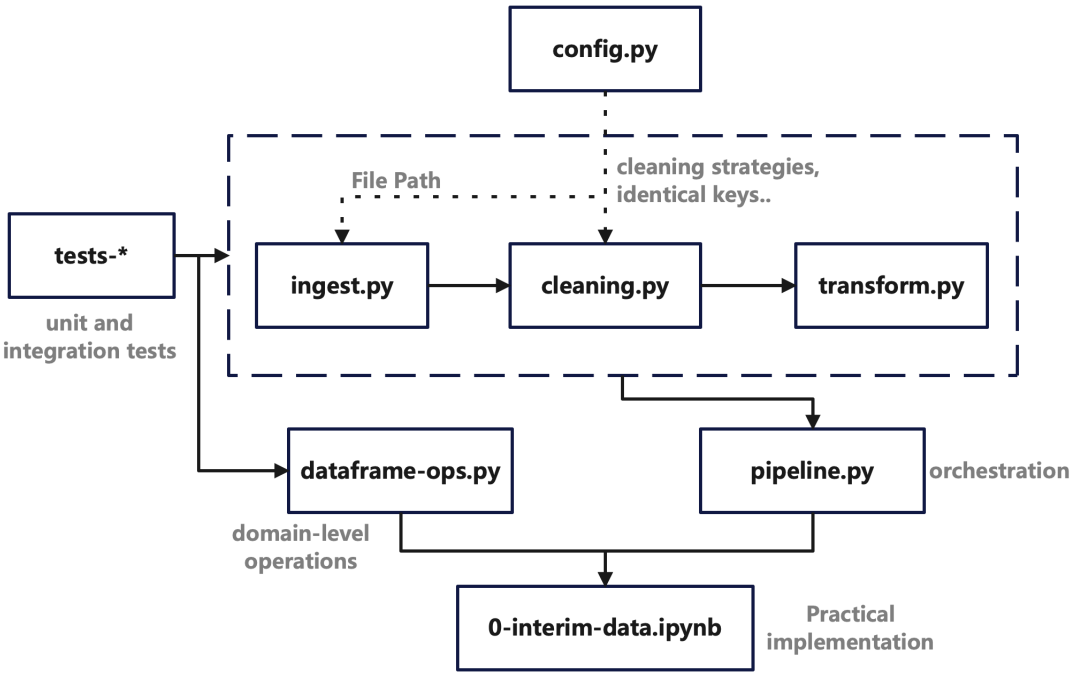

Figure 29: Demonstration of Workflow for Implementing and Integrating Hellinger Transformation into Complete Data Processing Pipeline

With this highly integrated modularization, the previously long and complex code can be enclosed in these separated modules and the application work can be achieved by a few lines of code. Like the following lines of code in Figure 30 achieve the whole transformation and data merging work in the Jupyter notebook, which used to require much more extensive and vulnerable code within the same notebook:

```python
# Apply hellinger transformation on the raw taxa data
taxa_hell = hellinger_transform(get_block(master, "taxa", "raw"))
# Wrap and add the transformed data into the 'taxa' block with 'hellinger' subblock
master    = add_site_block(master, taxa_hell, "taxa", "hellinger")

# Do similar work to logarithmic transformation on the raw chemical data
chem_logz = log1p_standardize(get_block(master, "chemical", "raw"))
master    = add_site_block(master, chem_logz, "chemical", "logz")

# check the added blocks
master
```

✓ 0.0s Python

clear, maintainable, extendable

raw chemical data *Log-trans chemical data*

block	chemical										...	logz			
subblock	raw										...	logz			
var	1234TCB	1245TCB	Al	As	Bi	Ca	Cd	Co	Cr	Cu	...	OCS	Pb	QCB	Sb
StationID															
A10	0.835583	0.775732	3041	1.939	18.45000	28170	0.2950	2.723	8.766	17.64	...	-0.312165	0.280413	-0.737303	0.703623
A23	0.639983	0.697265	4483	2.512	17.03000	42110	0.3986	4.009	10.850	17.28	...	-0.687740	-0.023640	0.328420	-0.117357
A27	0.451838	0.815149	13620	2.759	0.05370	41610	0.2180	6.273	21.080	25.00	...	-0.123819	0.006849	-0.756687	-1.917417
A28	0.224379	0.483363	12750	2.609	0.06617	33280	0.1197	5.824	18.700	24.07	...	-0.483904	-0.088969	-0.570351	-1.917417
A29	0.299715	0.695356	23740	3.735	0.15290	40450	0.1536	9.618	44.370	44.72	...	0.223999	0.703332	-0.343399	-1.831041
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
S99	0.655186	0.951855	2826	1.214	15.12000	14960	0.3079	4.025	12.260	12.72	...	-0.687740	0.137733	-0.452841	-0.096333
UBC1	0.000000	12.135559	6757	1.581	21.01000	43310	0.5146	5.263	12.800	16.00	...	3.159716	0.382723	2.446885	0.191014
UCC1	0.000000	4.319792	5945	2.260	23.62000	50540	0.5658	5.303	11.400	15.81	...	1.655922	0.234947	1.577347	0.282724
UCE1	0.000000	0.552417	7050	4.475	0.00010	38090	0.3449	6.232	11.860	14.62	...	0.945726	0.942971	0.296544	-1.916620
UJC1	0.000000	11.243048	5070	2.713	19.72000	45230	0.6496	5.213	11.220	17.55	...	3.597840	0.379396	3.163397	0.134830

104 rows × 100 columns

Figure 30: The lines of code needed to achieve the above discussed operations in practical scenarios and the resulting integrated data frame from the operation

7.2 General Project Structure and Rationale

Generally speaking, the project follows a modular and well-organized structure to support both research reproducibility and software engineering best practices. The folder layout is as follows:

- **configs**: centralizes configuration files (e.g., paths, random seeds, parameter settings), ensuring experiments are reproducible and adjustable without modifying code.
- **data**: stores raw, interim, and processed datasets in a structured hierarchy, making data provenance transparent and simplifying pipeline automation.
- **documents**: contains draft texts, LaTeX notes, and thesis-related writeups, providing a bridge between research outputs and manuscript preparation.
- **figures**: keeps all generated visualizations, plots, and diagrams organized for easy reuse in the thesis and publications.
- **notebooks**: hosts exploratory Jupyter notebooks, which serve as an interface for practical implementation, quick testing, and visualization.
- **reference**: used for bibliographic resources, papers, and supplementary literature, ensuring traceability of scientific background.
- **src**: holds the main source code in a modular format (e.g., `ingest.py`, `cleaning.py`, `pipeline.py`), which separates concerns between ingestion, cleaning, transformation, and higher-level operations.
- **tests**: includes unit tests and validation scripts to ensure correctness and robustness of each module across different scenarios.
- **artifacts**: stores intermediate results, logs, and model checkpoints, preserving computational outcomes for reproducibility and debugging.
- **pyproject.toml**: defines the project environment, dependencies, and metadata, which standardizes reproducibility across systems.

This design provides several benefits:

1. **Reproducibility**: Configurations, raw data, and processed results are explicitly separated, making workflows transparent and repeatable.
2. **Scalability**: Modular code in `src/` and standardized data storage enable extensions (e.g., adding new datasets or models) with minimal disruption.
3. **Clarity**: Clear distinction between code, data, results, and documentation reduces confusion and facilitates collaboration or future reuse.
4. **Professionalism**: The structure aligns with common practices in industrial data science and academic research, ensuring maintainability and credibility of the project.

8 Supervisory Dissolution

The student agrees that supervision will be dissolved if any of the following happen:

- Two consecutive progress reports are unacceptable
- Three consecutive progress reports are concerning/unacceptable
- An academic integrity violation is suspected by the supervisor and suspected by at least one other faculty member.

9 Timeline(temporary)

The below months are in 2025, each enumerated body in each month is the 1st, 2nd, 3rd and 4th week of that month.

- August
  1. Make Preliminary Analysis on the raw data, explore the achievability of the general framework.
  2. Make adjustments on the previous proposal based on the validated preliminary results.
  3. Finish the proposal, check and submit the proposal.
  4. Prepare and study the relevant theoretical foundations, about the math, statistical and coding tools: for example: Discriminant Function and Ordination method and Project management.
- September
  1. To finish and test the ***data cleaning, operation and transformation modules***, design and prepare the ***PCA sediment contamination module***.
  2. To finish and test the ***PCA sediment contamination module***, design and prepare the ***Cluster Analysis module***.
  3. To finish and test the ***Cluster Analysis module***, design and prepare the ***Discriminant Function Analysis module***.
  4. To finish and test the ***Discriminant Function Analysis module***, design and prepare the ***Ordination-ZCI constructor module***.
- October
  1. To finish and test the ***Ordination-ZCI constructor module***, design and prepare the ***Piecewise Quantile Regression module***.
  2. To finish and test the ***Piecewise Quantile Regression module***, design and prepare the ***Degradation Detection module***.
  3. To finish and test the ***Degradation Detection module***, organize the whole basic framework.
  4. Prepare and practice the Synthetic Data Method for the project, design the ***Synthetic Data module***.

Complete basic framework at the finish of Degradation Detection Module

---

- November
  1. To finish and test the ***Synthetic Data module***, figure out where to integrate it in the overall framework and how to ensure its compatibility with existing modules and test its performance.
  2. Prepare and practice the spatial analysis into the overall framework, study and practice deep learning methods, theoretically integrate them into the framework.
  3. Design, finish and test the ***geographical weighted PCA module***, integrate into the established framework and interpret its results against the classic PCA results.
  4. Create new modules or extend existing modules to support spatial analysis and test them.
- December
  1. Introduce deep learning methods into the existing framework, figure out how to integrate them into specific steps of modules and how to validate their performance and interpret their results.
  2. Integrate and adjust the whole modules, ensuring smooth interaction and data flow between them, like a software test suite.
  3. Finish 80% of the first draft.
  4. Include the Satellite-derived data about wildfire into the project, explore the spatial analysis combined piecewise quantile regression using the established modules.

The below months are in 2026, each enumerated body in each month is the 1st, 2nd, 3rd and 4th week of that month, '.' symbol means still to be determined.

- January
    1. Analyse the Satellite-derived data about wildfire and its relevance to the established methods, identifying potential improvements and adjustments.
    2. Finish the complete first draft.
    3. Make necessary revisions based on feedback and testing results.
    4. Conduct thorough testing and validation of the entire framework.

Complete First draft of the thesis

- February
    1. Finish the second draft of the thesis.
    2. .
    3. .
    4. .
- March
    1. Finish the final draft of the thesis.
    2. .
    3. .
    4. .

Complete Second and final draft of the thesis

- April
    1. Preparing presentation materials and slides.
    2. Rehearsing the presentation and addressing potential questions.
    3. Presentation and defence of the thesis, determined by the coordinator.

Presentation and defence

References

[1] U.S. Environmental Protection Agency and Environment Canada. State of the great lakes 2007: Status and trends of great lakes shoreline hardening. Technical report, U.S. Environmental Protection Agency, Great Lakes National Program Office, 2007. Accessed: 2025-07-15.

[2] J. David Allan, Peter B. McIntyre, Sigrid D. P. Smith, Benjamin S. Halpern, Gregory L. Boyer, Andy Buchsbaum, G. A. Burton, Linda M. Campbell, W. Lindsay Chadderton, Jan J. H. Ciborowski, Patrick J. Doran, Tim Eder, Dana M. Infante, Lucinda B. Johnson, Christine A. Joseph, Adrienne L. Marino, Alexander Prusevich, Jennifer G. Read, Joan B. Rose, Edward S. Rutherford, Scott P. Sowa, and Alan D. Steinman. Joint analysis of stressors and ecosystem services to enhance restoration effectiveness. *Proceedings of the National Academy of Sciences*, 110(1):372–377, 2013.

[3] Virginie Archaimbault, Philippe Usseglio-Polatera, Jeanne Garric, Jean-Gabriel Wasson, and Marc Babut. Assessing pollution of toxic sediment in streams using bio-ecological traits of benthic macroinvertebrates. *Freshwater Biology*, 55(7):1430–1446, 2010.

[4] G. Birch. 14.24 use of sedimentary-metal indicators in assessment of estuarine system health. In John F. Shroder, editor, *Treatise on Geomorphology*, pages 282–291. Academic Press, San Diego, 2013.

[5] Gavin F. Birch. A review and critical assessment of sedimentary metal indices used in determining the magnitude of anthropogenic change in coastal environments. *Science of The Total Environment*, 823:153623, 2022.

[6] Sebastian Birk, Wim Bonne, Angel Borja, Sandra Brucet, Amandine Courrat, Sandra Poikane, Angelo Solimini, Wouter van de Bund, Nikolaos Zampoukas, and Daniel Hering. Three hundred ways to assess europe's surface waters: An almost complete overview of biological methods to implement the water framework directive. *Ecological Indicators*, 18:31–41, 2012.

[7] Nùria Bonada, Narcìs Prat, Vincent H. Resh, and Bernhard Statzner. Developments in aquatic insect biomonitoring: a comparative analysis of recent approaches. *Annual Review of Entomology*, 51:495–523, 2006.

[8] Daniel Borcard and Pierre Legendre. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling*, 153:51–68, 2002. Introduces spatial eigenfunction (PCNM) approach to model spatially structured ecological variation across multiple scales.

[9] Daniel Borcard, Pierre Legendre, and Pierre Drapeau. Partialling out the spatial component of ecological variation. *Ecology*, 73(3):1045–1055, 1992. Foundational paper showing that accounting for (or assuming away) spatial structure simplifies disentangling environmental vs. other sources of community variation.

[10] J. C. Brazner, N. P. Danz, Gerald J. Niemi, R. R. Regal, A. S. Trebitz, R. W. Howe, J. M. Hanowski, Lucinda B. Johnson, J. J.H. Ciborowski, C. A. Johnston, Euan D. Reavie, Valerie J. Brady, and G. V. Sgro. Evaluation of geographic, geomorphic and human influences on great lakes wetland indicators: A multi-assemblage approach. *Ecological Indicators*, 7(3):610–635, 2007.

[11] G. Allen Burton. Sediment quality criteria in use around the world. *Limnology*, 3(2):65–76, 2002.

[12] G. Allen Burton and Eric L. Johnston. Assessing contaminated sediments in the context of multiple stressors. *Environmental Toxicology and Chemistry*, 29(12):2625–2643, 2010.

[13] Jarrett E. K. Byrnes and Laura E. Dee. Causal inference with observational data and unobserved confounding variables. *Ecology Letters*, 28(1):e14384, 2025.

[14] Brian S. Cade and Barry R. Noon. A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1(8):412–420, 2003.

[15] Daniel J. Cain, Samuel N. Luoma, Janet L. Carter, and Sarah V. Fend. Aquatic insects as bioindicators of trace element contamination in cobble-bottom rivers and streams. *Canadian Journal of Fisheries and Aquatic Sciences*, 49(10):2141–2154, 1992.

[16] Stephen R. Carpenter, Nina F. Caraco, David L. Correll, Robert W. Howarth, Andrew N. Sharpley, and Val H. Smith. Nonpoint pollution of surface waters with phosphorus and nitrogen. *Ecological Applications*, 8(3):559–568, 1998.

[17] Peter M. Chapman. The sediment quality triad approach to determining pollution-induced degradation. *Science of the Total Environment*, 97/98:815–825, 1990.

[18] Aurea C. Chiaia-Hernández, Carmen Casado-Martinez, Pablo Lara-Martin, and Thomas D. Bucheli. Sediments: sink, archive, and source of contaminants. *Environmental Science and Pollution Research*, 29:85761–85765, 2022.

[19] Jan Ciborowski, Lucinda Johnson, Joseph Gathman, Valerie Brady, Jeffrey Holland, Tom Hollenhorst, J. Schuldt, G. Host, and Caryll Richards. Zoobenthic indicators of environmental condition at great lakes coastal margins derived from standardized multivariate analyses across anthropogenic stress gradients. *AGU Spring Meeting Abstracts*, 05 2005.

[20] Jonathan P. Daily, Nathaniel P. Hitt, David R. Smith, and Craig D. Snyder. Experimental and environmental factors affect spurious detection of ecological thresholds. *Ecology*, 93(1):17–23, 2012.

[21] Susan P. Davies and Susan K. Jackson. The biological condition gradient: A descriptive model for interpreting change in aquatic ecosystems. *Ecological Applications*, 16(4):1251–1266, August 2006.

[22] L. Delchambre. Weighted principal component analysis: a weighted covariance eigendecomposition approach. *Monthly Notices of the Royal Astronomical Society*, 446(4):3545–3555, 2015.

[23] Mélanie Desrosiers, Bernadette Pinel-Alloul, and Charlotte Spilmont. Selection of macroinvertebrate indices and metrics for assessing sediment quality in the st. lawrence river (qc, canada). *Water*, 12(12):3335, 2020.

[24] Carsten F. Dormann, Jana M. McPherson, Miguel B. Araújo, Roger Bivand, Janine Bolliger, Gudrun Carl, Richard G. Davies, Alexandre Hirzel, Walter Jetz, W. Daniel Kissling, Ingolf Kühn, Ralf Ohlemüller, Pedro R. Peres-Neto, Björn Reineking, Boris Schröder, Frank M. Schurr, and Robert Wilson. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5):609–628, 2007. Reviews that spatial autocorrelation often arises from unmeasured, spatially structured environmental variables; supports using spatial proxies to capture hidden environmental heterogeneity.

[25] Stephane Dray, Pierre Legendre, and Pedro R. Peres-Neto. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (pcnm). *Ecological Modelling*, 196(3-4):483–493, 2006.

[26] L. L. Eberhardt and J. M. Thomas. Designing environmental field studies. *Ecological Monographs*, 61(1):53–73, 1991.

[27] John Eggleton and Karen V. Thomas. A review of factors affecting the release and bioavailability of contaminants during sediment disturbance events. *Environment International*, 30(7):973–980, 2004.

[28] Peter G. Fairweather. Statistical power and design requirements for environmental monitoring. *Marine and Freshwater Research*, 42(5):555–567, 1991.

[29] Matthias M. Fischer. Quantifying the uncertainty of variance partitioning estimates of ecological datasets. *arXiv preprint*, 2018. Demonstrates large uncertainty in variance partitioning, especially when predictor variance is low.

[30] Daniel A. Griffith and Pedro R. Peres-Neto. Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology*, 87(10):2603–2613, 2006.

[31] Katrin Grünfeld. Dealing with outliers and censored values in multi-element geochemical data – a visualization approach using xmdvtool. *Applied Geochemistry*, 20(2):341–352, 2005.

[32] Paul Harris, Chris Brunsdon, and Martin Charlton. Geographically weighted principal components analysis. *International Journal of Geographical Information Science*, 25(10):1717–1736, 2011. Introduces GWPCA with implementation and case study applications.

[33] C. S. Holling. Resilience and stability of ecological systems. *Annual Review of Ecology and Systematics*, 4:1–23, 1973.

[34] George E. Host, Katya E. Kovalenko, Terry N. Brown, Jan J. H. Ciborowski, Lucinda B. Johnson, et al. Risk-based classification and interactive map of watersheds contributing anthropogenic stress to laurentian great lakes coastal ecosystems. *Journal of Great Lakes Research*, 45(3):537–545, 2019.

[35] Qi Huang, Hanze Zhang, Jiaqing Chen, and Mengying He. Quantile regression models and their applications: A review. *Journal of Biometrics & Biostatistics*, 8(3):354, 2017.

[36] Stuart H. Hurlbert. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, 54(2):187–211, 1984.

[37] S. W. Karickhoff. Organic pollutant sorption in aquatic systems. *Journal of Hydraulic Engineering*, 110(6):707–735, 1984.

[38] Katya E. Kovalenko, George E. Host, Terry N. Brown, Jan J. H. Ciborowski, and Lucinda B. Johnson. Congruence of community thresholds in response to anthropogenic stress in great lakes coastal wetlands. *Freshwater Science*, 33(3):958–971, 2014.

[39] Ashley E. Larsen, Kyle Meng, and Bruce E. Kendall. Causal analysis in control–impact ecological studies with observational data. *Methods in Ecology and Evolution*, 10(7):924–934, 2019.

[40] Pierre Legendre and Louis Legendre. Studying beta diversity: ecological variation–partitioning by multiple regression and canonical analysis. *Journal of Plant Ecology*, 1(1):3–11, 2008.

[41] David R. Lenat and Vincent H. Resh. Taxonomy and stream ecology—the benefits of genus- and species-level identifications. *Journal of the North American Benthological Society*, 20(2):287–298, 2001.

[42] Donald D. MacDonald, Christopher G. Ingersoll, and Thomas A. Berger. Development and evaluation of consensus-based sediment quality guidelines for freshwater ecosystems. *Archives of Environmental Contamination and Toxicology*, 39(1):20–31, 2000.

[43] Salomé Menezes, David J. Baird, and Amadeu M. V. M. Soares. Beyond taxonomy: a review of macroinvertebrate trait-based community descriptors as tools for freshwater biomonitoring. *Journal of Applied Ecology*, 47(4):711–719, 2010.

[44] Gerald J. Niemi and Michael E. McDonald. Application of ecological indicators. *Annual Review of Ecology, Evolution, and Systematics*, 35:89–111, 2004.

[45] Craig W. Osenberg, Russell J. Schmitt, Sally J. Holbrook, Kate E. Abu-Saba, and A. Russell Flegal. Detection of environmental impacts: natural variability, effect size, and power analysis. *Ecological Applications*, 4(1):16–30, 1994.

[46] Marianne Pasanen-Mortensen, Markku Pyyk"onen, and Bodil Elmhagen. Where lynx prevail, foxes will fail – limitation of a mesopredator in eurasia. *Global Ecology and Biogeography*, 22(7):868–877, 2013.

[47] Garry D. Peterson, Stephen R. Carpenter, William A. Brock, Jonathan M. Hanson, Joel Carson, Lynne Haskins, Milos Holmgren, Tim Eason, Christine Engels, et al. Ecological thresholds: The key to successful environmental management or an important concept with no practical application? *Ecosystems*, 9(1):1–13, 2006.

[48] Clemens Reimann, Peter Filzmoser, Robert G. Garrett, and Rudolf Dutter. *Statistical Data Analysis Explained: Applied Environmental Statistics with R.* John Wiley & Sons, 2008. Covers multivariate and ordination methods in geochemical and environmental data analysis.

[49] T. B. Reynoldson, R. H. Norris, V. H. Resh, K. E. Day, and D. M. Rosenberg. The reference condition: A comparison of multimetric and multivariate approaches to assess water-quality impairment using benthic macroinvertebrates. *Journal of the North American Benthological Society*, 16(4):833–852, December 1997.

[50] T.B̃. Reynoldson, D.M̃. Rosenberg, K.Ẽ. Day, and R.H̃ Norris. Biological guidelines for freshwater sediment based on benthic assessment of sediment (the beast) using a multivariate approach for predicting biological state. Technical report, Canadian Aquatic Biomonitoring Network (CABIN), Environment Canada, 1999. An early and influential application of clustering reference sites and using discriminant models to define and assess reference conditions.

[51] Travis S. Schmidt, William H. Clements, and Brian S. Cade. Estimating risks to aquatic life using quantile regression. *Freshwater Science*, 31(3):709–723, 2012.

[52] Augustine Sitati, Frank O. Masese, Mourine J. Yegon, Alfred O. Otieno, and Phillip O. Raburu. Abundance- and biomass-based metrics of functional composition of macroinvertebrate communities as surrogates of ecosystem attributes in afrotropical streams. *Aquatic Sciences (preprint)*, 2021. Preprint posted on ResearchGate; abundance-based metrics were found to outperform biomass-based ones in predicting ecosystem attributes.

[53] Tyler W. Smith and Jeremy T. Lundholm. Variation partitioning as a tool to distinguish between niche and neutral processes. *Ecography*, 33(4):648–655, 2010.

[54] Rebecca Spake, Martha Paola Barajas-Barbosa, Shane A. Blowes, Diana E. Bowler, Corey T. Callaghan, Magda Garbowski, Stephanie D. Jurburg, Roel van Klink, Lotte Korell, Emma Ladouceur, Roberto Rozzi, Duarte S. Viana, Wu-Bing Xu, and Jonathan M. Chase. Detecting thresholds of ecological change in the anthropocene. *Annual Review of Environment and Resources*, 47(Volume 47, 2022):797–821, 2022.

[55] John L. Stoddard, David P. Larsen, Charles P. Hawkins, Richard K. Johnson, and Richard H. Norris. Setting expectations for the ecological condition of streams: The concept of reference condition. *Ecological Applications*, 16(4):1267–1276, 2006.

[56] U.S. Geological Survey. Sediment-associated contaminants, February 27 2019. Water Resources Mission Area, U.S. Department of the Interior; accessed 2025-09-09.

[57] Lallébila Tampo, Idrissa Kaboré, Elliot H. Alhassan, Adama Ouéda, Limam M. Bawa, and Gbandi Djaneye-Boundjou. Benthic macroinvertebrates as ecological indicators: Their sensitivity to water quality and human disturbances in a tropical river. *Frontiers in Water*, 3:662765, 2021.

[58] Jabed H. Tomal and Jan J. H. Ciborowski. Ecological models for estimating breakpoints and prediction intervals. *Ecology and Evolution*, 11(5):2053–2065, 2021.

[59] Monica G. Turner. Landscape ecology: The effect of pattern on process. *Annual Review of Ecology and Systematics*, 20:171–197, 1989. Seminal review showing that spatial pattern (multi-scale environmental heterogeneity) governs ecological processes and community variation.

[60] U.S. Environmental Protection Agency. Great lakes facts and figures, 2024. Accessed: 2025-07-15.

[61] Charles J. V"or"osmarty, Peter B. McIntyre, Mark O. Gessner, David Dudgeon, Alexander Prusevich, Pamela Green, Stephen Glidden, Stuart E. Bunn, Caroline A. Sullivan, Carli R. Liermann, and Peter M. Davies. Global threats to human water security and river biodiversity. *Nature*, 467:555–561, 2010.

[62] John A. Wiens and Kenneth R. Parker. Analyzing the effects of accidental environmental impacts: approaches and assumptions. *Ecological Applications*, 5(4):1069–1083, 1995.

[63] Jian Zhang. Zoobenthic community composition and chironomidae (diptera) mouthpart deformities as indicators of sediment contamination in the lake huron-lake erie corridor of the laurentian great lakes. Master's thesis, University of Windsor, Windsor, Ontario, Canada, 2008. A thesis for the degree of Master of Science.

[64] V. Zitko. Principal component analysis in the evaluation of environmental data. *Marine Pollution Bulletin*, 28(12):718–722, 1994.

10 Appendix

10.1 Tables

Table 10: Descriptive Statistics of Major Metals 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

Table 11: Descriptive Statistics of Organic Carbon and Chlorinated Benzenes

	SumReal	**degraded**	**intermediate**	**reference**
%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

Table 12: Descriptive Statistics of Pesticides and PCBs

	SumReal	**degraded**	**intermediate**	**reference**
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

10.2 Figures

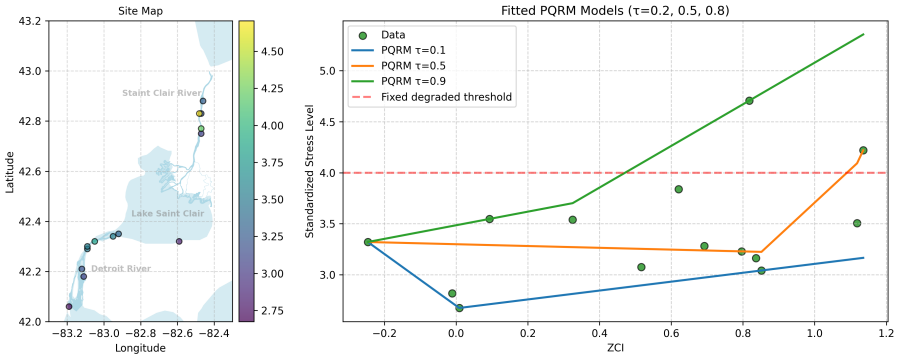

Figure 31: Piecewise Quantile Regression for cluster 1

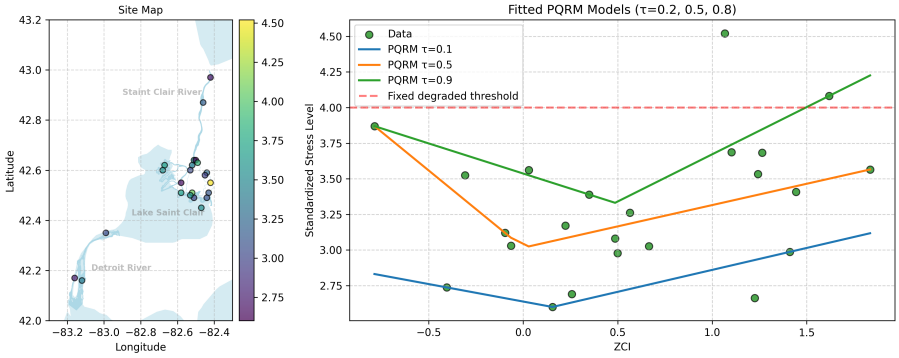

Figure 32: Piecewise Quantile Regression for cluster 2

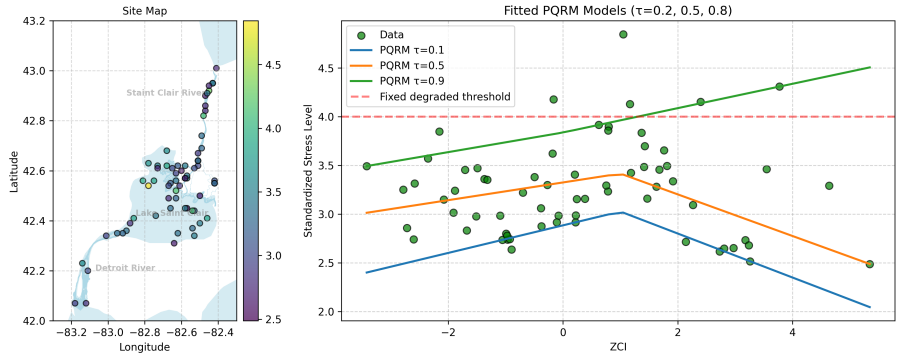

Figure 33: Piecewise Quantile Regression for cluster 3

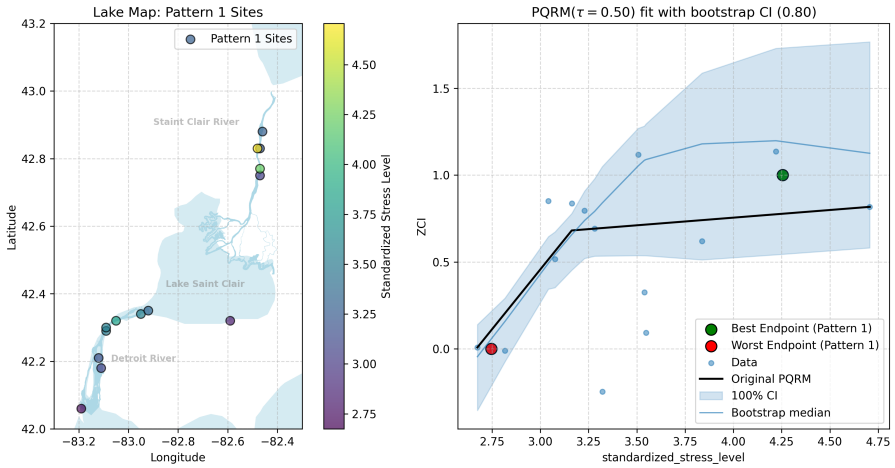

Figure 34: Piecewise Quantile Regression for cluster 1 with inference of confidence intervals

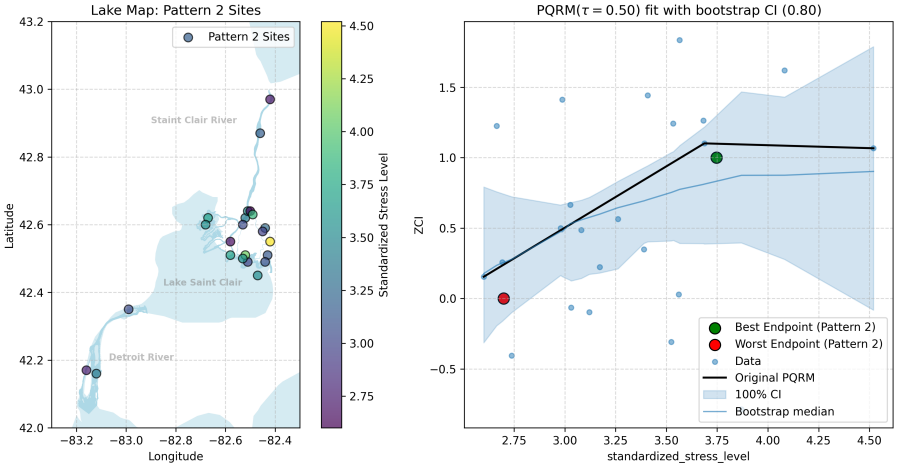

Figure 35: Piecewise Quantile Regression for cluster 2 with inference of confidence intervals

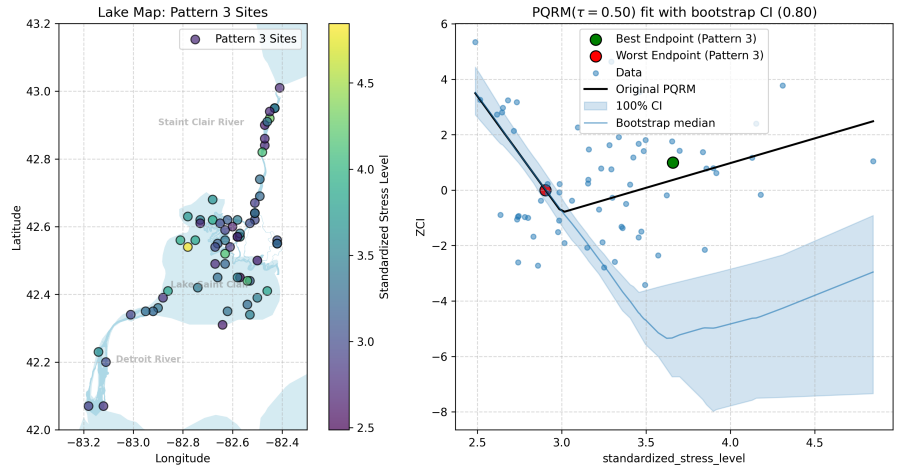

Figure 36: Piecewise Quantile Regression for cluster 3 with inference of confidence intervals

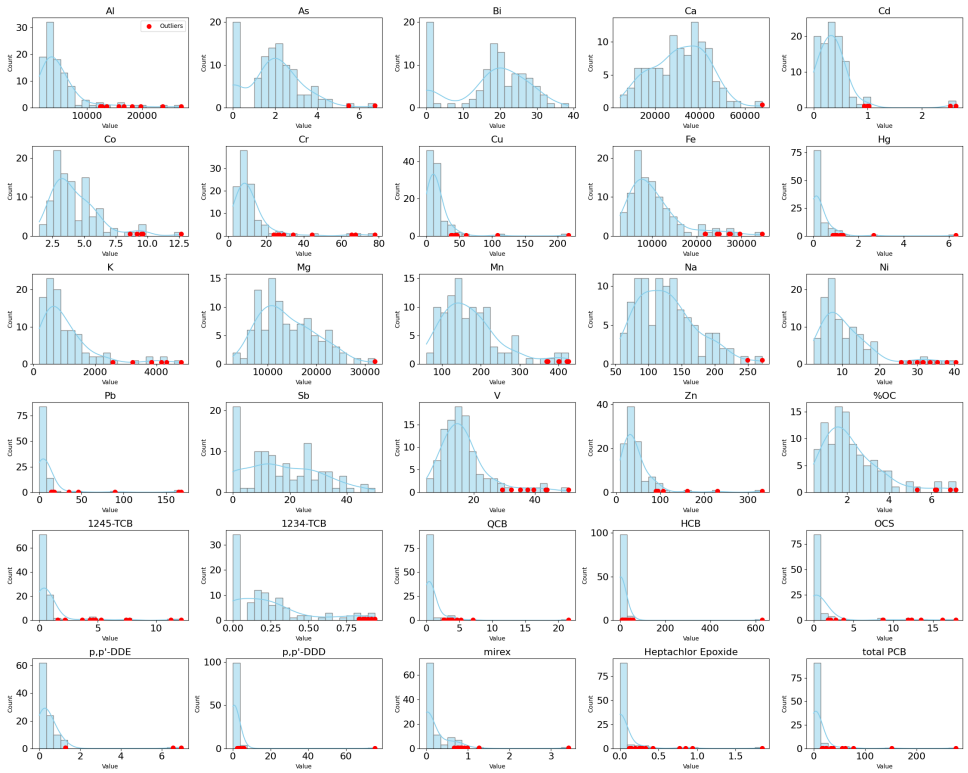

Figure 37: Example of outlier detection: Raw chemical histogram with IQR-detected outliers in red