Framework about the inference of sediment contamination level based on zoobenthic community structure

Feng Gu

August 7, 2025

Introduction - Motivation

Considering the benefits in research, practical and economical aspects, there is a motivation to develop **a sediment contamination level** inference method based on the **zoobenthic community structure**.

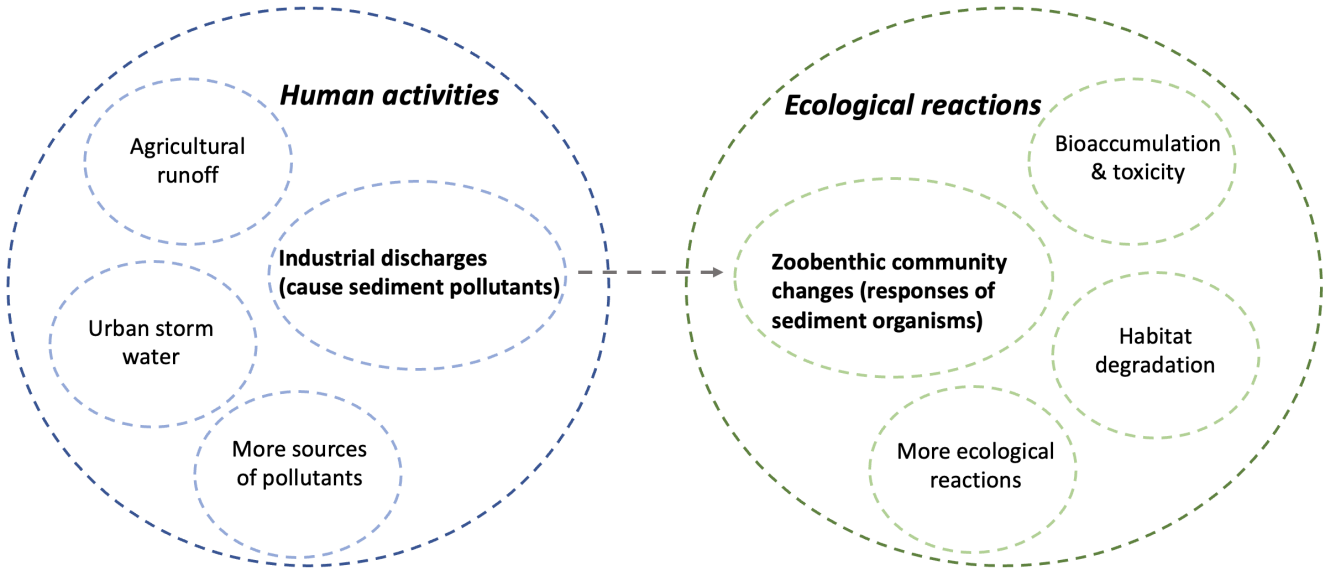
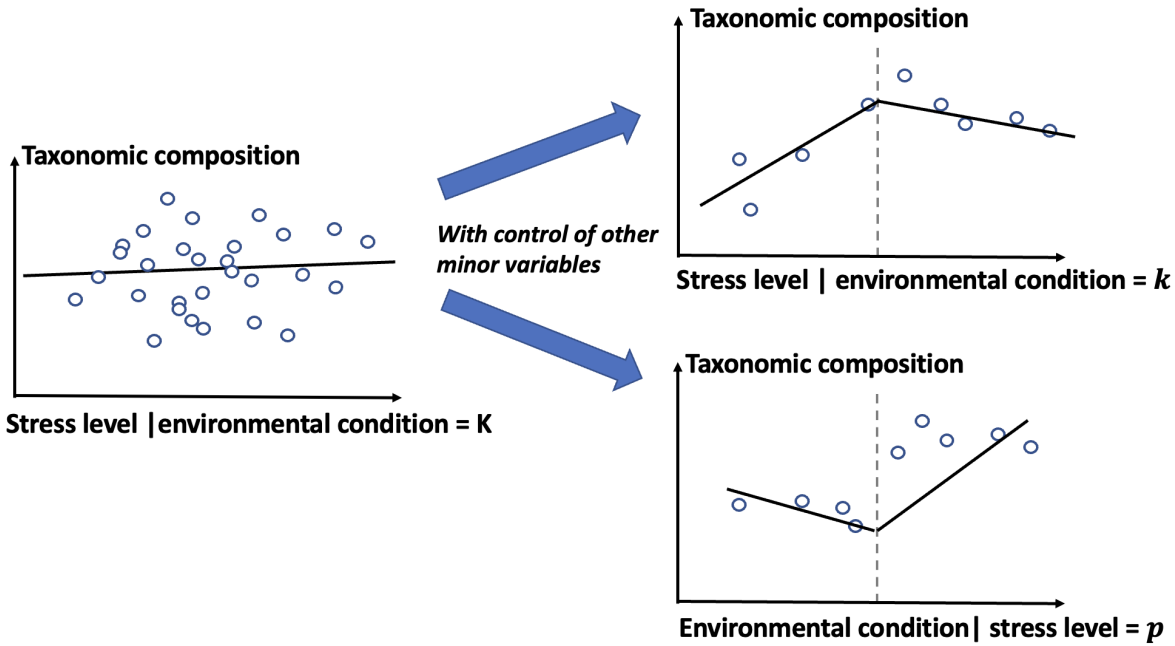


Figure: There are relationships between Human activities and ecological reactions

Introduction - Problem Statement

From natural observations, we need to design **numerical measurements for sediment contamination level and zoobenthic community structure** to describe them in numerical terms and build a model to infer the relationship between them. An unignorable fact is that **the zoobenthic community structure is influenced by environmental factors**, which must be controlled for its different levels of influence to improve the inference accuracy.



Data Collection

The data was prepared by Prof. Jan and his team, includes three types of data collected from the Lake Huron-Erie Corridor:

- Chemical data ($m \times 30$), including: metal concentrations, PCBs and PAHs from factors or mining.
- Environmental data ($m \times 5$ or $m \times 6$), including: temperature, pH, dissolved oxygen, etc.
- Zoobenthic macroinvertebrates data ($m \times 16$), including: selected organisms living in the sampled sediment, where chemical data was collected.

Three separated matrices with shared identical row indices, where each row represents a site sampled at a specific time and location.

Overview to data operation rules along the analysis

The three data matrices are merged into a single matrix by row indices, later processes that bring new information to each site will be merged into this single matrix by row indices as well.

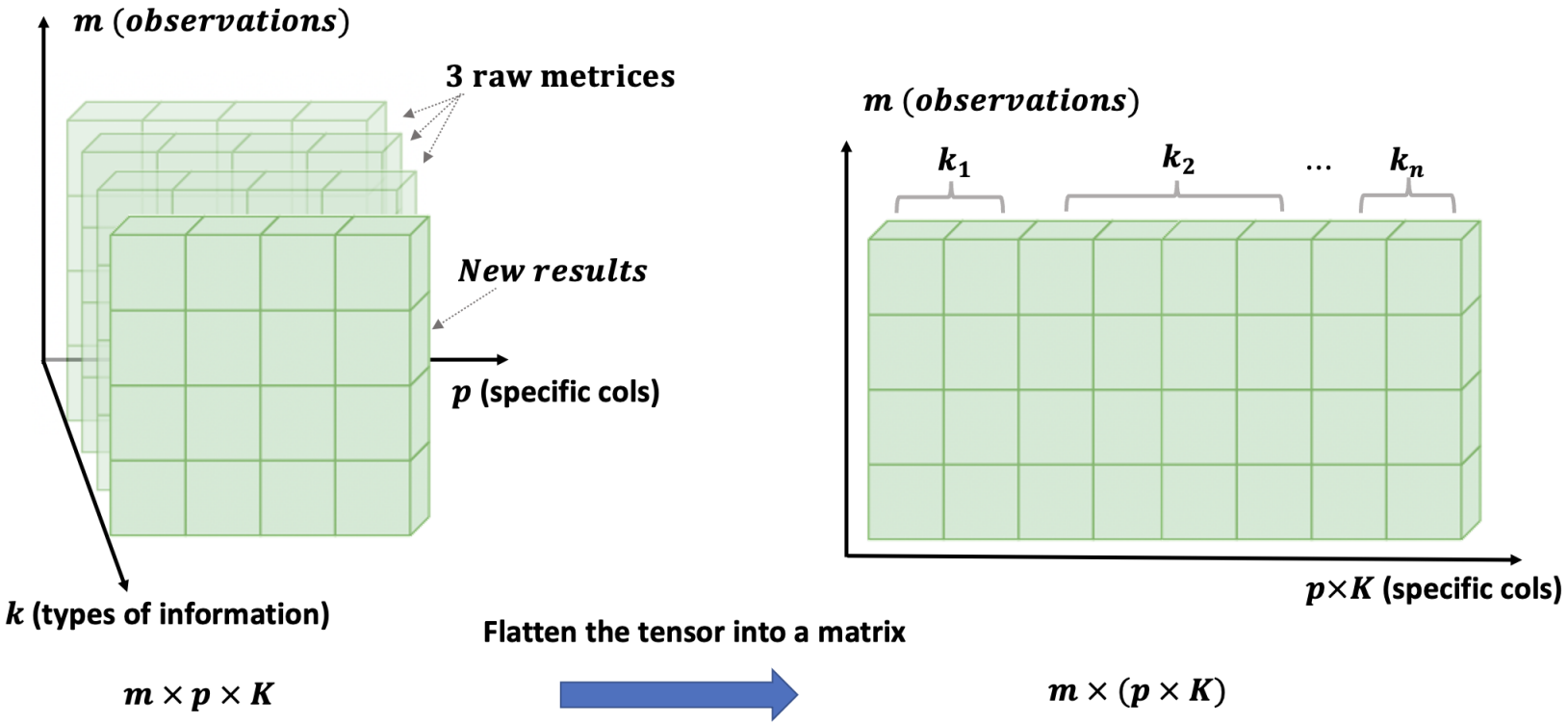
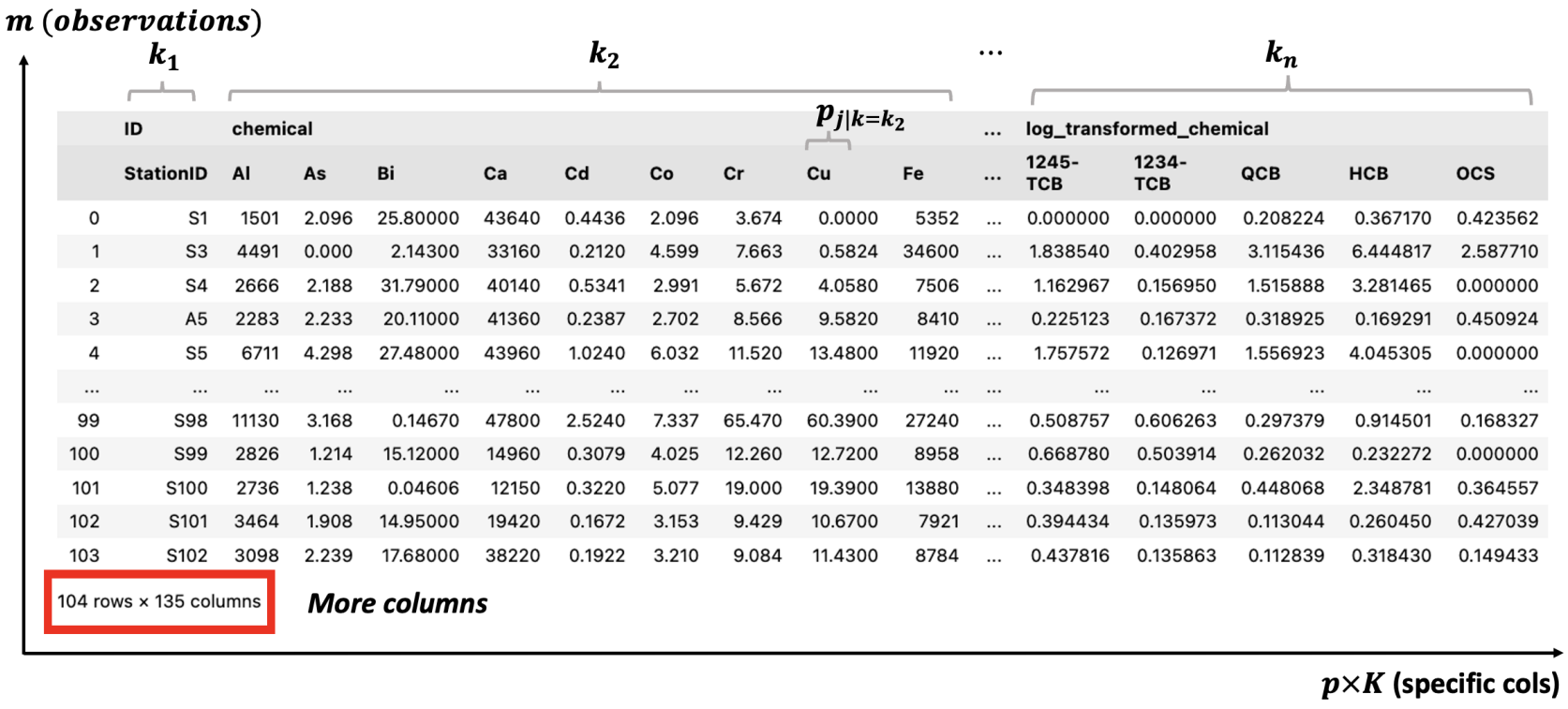


Figure: Data operation rules along the analysis

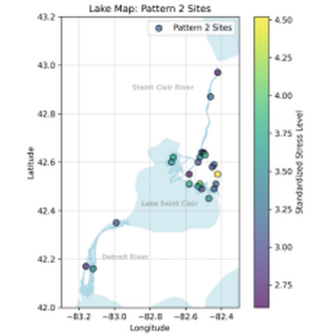
Instance of the data operation rules

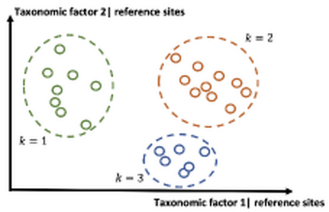


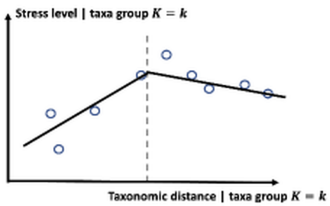
	ID	chemical									...	log_transformed_chemical				
	StationID	Al	As	Bi	Ca	Cd	Co	Cr	Cu	Fe	...	1245-TCB	1234-TCB	QCB	HCB	OCS
0	S1	1501	2.096	25.80000	43640	0.4436	2.096	3.674	0.0000	5352	...	0.000000	0.000000	0.208224	0.367170	0.423562
1	S3	4491	0.000	2.14300	33160	0.2120	4.599	7.663	0.5824	34600	...	1.838540	0.402958	3.115436	6.444817	2.587710
2	S4	2666	2.188	31.79000	40140	0.5341	2.991	5.672	4.0580	7506	...	1.162967	0.156950	1.515888	3.281465	0.000000
3	A5	2283	2.233	20.11000	41360	0.2387	2.702	8.566	9.5820	8410	...	0.225123	0.167372	0.318925	0.169291	0.450924
4	S5	6711	4.298	27.48000	43960	1.0240	6.032	11.520	13.4800	11920	...	1.757572	0.126971	1.556923	4.045305	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
99	S98	11130	3.168	0.14670	47800	2.5240	7.337	65.470	60.3900	27240	...	0.508757	0.606263	0.297379	0.914501	0.168327
100	S99	2826	1.214	15.12000	14960	0.3079	4.025	12.260	12.7200	8958	...	0.668780	0.503914	0.262032	0.232272	0.000000
101	S100	2736	1.238	0.04606	12150	0.3220	5.077	19.000	19.3900	13880	...	0.348398	0.148064	0.448068	2.348781	0.364557
102	S101	3464	1.908	14.95000	19420	0.1672	3.153	9.429	10.6700	7921	...	0.394434	0.135973	0.113044	0.260450	0.427039
103	S102	3098	2.239	17.68000	38220	0.1922	3.210	9.084	11.4300	8784	...	0.437816	0.135863	0.112839	0.318430	0.149433

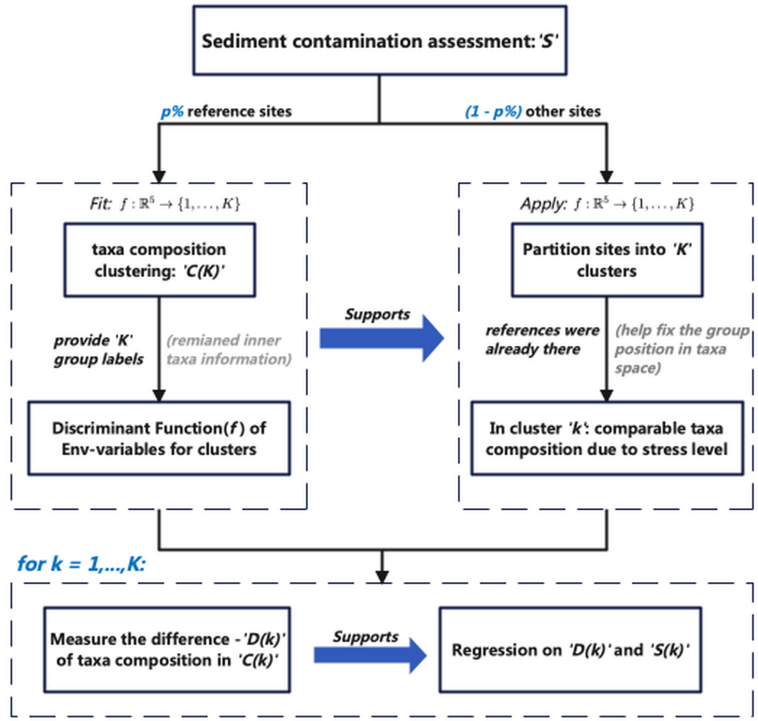
It does not directly relates to the inference model, but it makes data operation easier and smoother to accelerate the analytical process.

Framework of the analysis



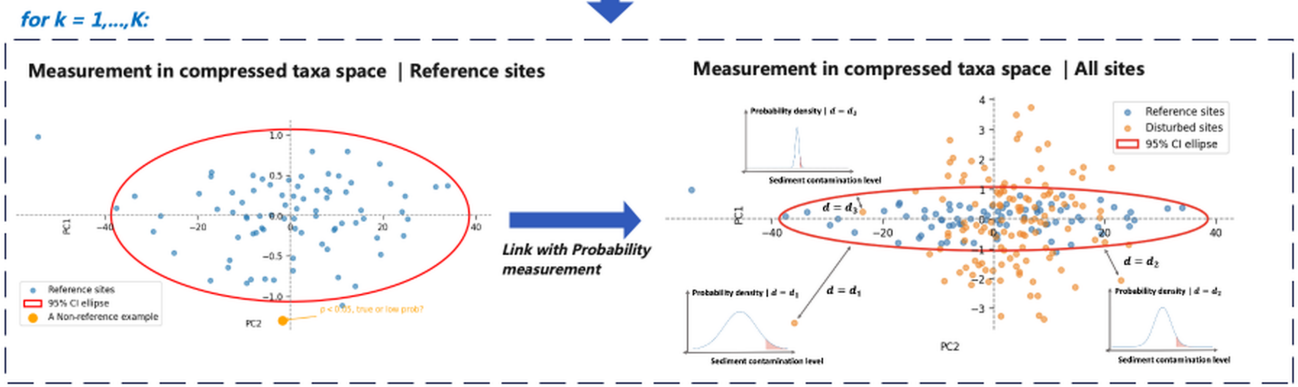






From scale-based to vector-based measurement

(retain more information, improve model performance)



Sediment Contamination Level Assessment

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the log-transformed chemical data. Apply PCA:

$$\mathbf{Z} = \mathbf{XW}$$

where $\mathbf{W}$ contains loadings of k selected PCs. **Normalize each PC**:

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

The contamination score for sample i:

$$S_i/\text{SumReal}_i = \sum_{j=1}^{k} \alpha_j \tilde{z}_{ij}$$

where α_j reflects the direction and importance of each PC.

Pick up references for control of sediment contamination level

Assuming the least stressed sites are references and they were **not or minimally influenced** by human activities, vice versa.

$$\mathcal{R} = \{i : S_i \leq Q_{p\%}(S)\}, \quad \mathcal{D} = \{i : S_i \geq Q_{(100-p)\%}(S)\}$$

where S_i is the stress score for site i, and $Q_{p\%}(S)$, $Q_{(100-p)\%}(S)$ are the p-th and $(100 - p)$-th percentiles.
These references will support building a purely tidy model that predicts what a pristine taxa composition should be given a specific set of environmental conditions [1].

[1] Notice the references-environmental distribution, a uniform or proportional to the real world distribution is better for capturing real-world relationships

Pick up references for control of sediment contamination level

k_{n+1}

...	octave_transformed_taxa		environmental					SumRel		
...	Gastropoda	Sphaeriidae	Total Organic Carbon (LOI %)	Water Depth (m)	Water Temperature (¬∞C)	Dissolved Oxygen Concentration (mg/L)	Median Particle Size (Phi)	stress_level	standardized_stress_level	site_label
...	3.203427e-15	3.203427e-15	0.41	4.6	20.77	10.03	1.43	-4.127157	2.738680	degraded
...	2.804993e+00	3.523769e-15	1.19	3.0	20.33	10.26	1.20	-1.440226	3.093926	intermediate
...	1.922056e-15	2.550633e+00	1.46	4.6	20.03	9.97	1.10	7.029481	4.308701	reference
...	2.242399e-15	2.242399e-15	1.10	5.0	19.78	9.03	-1.30	0.627823	3.281353	intermediate
...	3.269402e+00	3.269402e+00	2.76	1.5	19.86	8.50	1.68	-0.339393	3.171043	intermediate

Figure: New stress scores information added to the data matrix

Determine how environment shapes the taxa composition

This reference sites have taxa composition that were only shaped by the environment, good for exploring how environment determines taxa composition with the control of human impacts. However, **5 environmental factors are not enough to explain the taxa composition measured by 16 taxa**, only a poor model can be built on them. But explaining limited information of the taxa composition may be good, like their clustering patterns C_K.

$$\mathcal{F} : R^{m\times 5} \rightarrow R^{m\times 16}, \quad \text{which is not good}$$

$$\mathcal{F} : R^{m\times 5} \rightarrow C_K^{m\times 1}, \quad \text{which may be good}$$

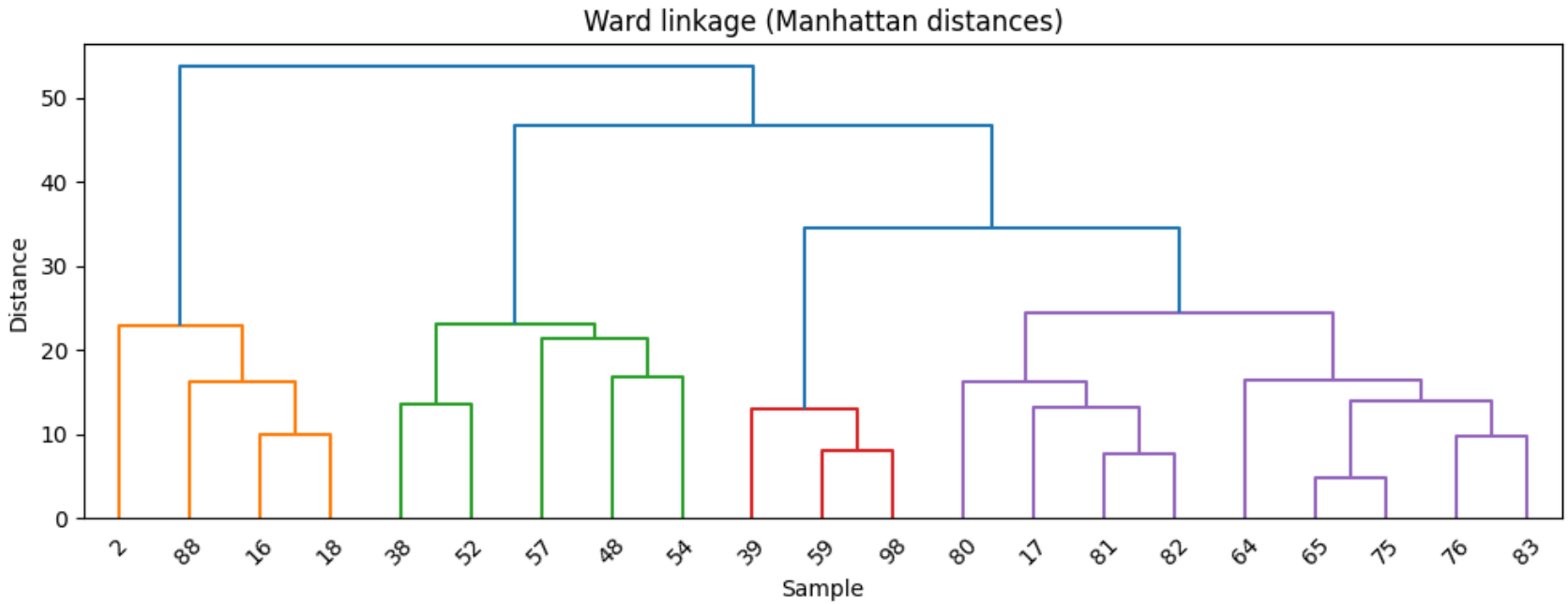
The $C_K^{m\times 1}$ is the clustering label of each site, confined information of its taxa composition.

Clustering on the taxa composition of references

Let $\mathcal{R}$ be the set of reference sites, each with taxa composition $\mathbf{y}_i \in \mathbb{R}^t$. Cluster $\{\mathbf{y}_i : i \in \mathcal{R}\}$ into K groups $\mathcal{C}_1, \ldots, \mathcal{C}_K$:

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

Each $\mathcal{C}_k$ contains sites with similar taxa compositions.



Fit Discriminant Function of environmental factors for clustering labels

Let $\mathcal{C}_1, \ldots, \mathcal{C}_K$ be resulted taxa cluster labels. For each site i, let $\mathbf{e}_i$ be its environmental variables. Fit a discriminant function of environmental variables for cluster labels on the $p\%$ reference sites:

$$\mathcal{F} : e_i^{(1\times 5)} \mapsto \hat{\mathcal{C}}_{i,k}, \quad i \in \mathcal{R}, k \in \{1, \ldots, K\}$$

Apply $\mathcal{F}$ to rest $(1 - p\%)$ sites to group them into the K clusters, where reference sites were already assigned after the fitting stage.

Apply Discriminant Function of environmental factors for clustering labels

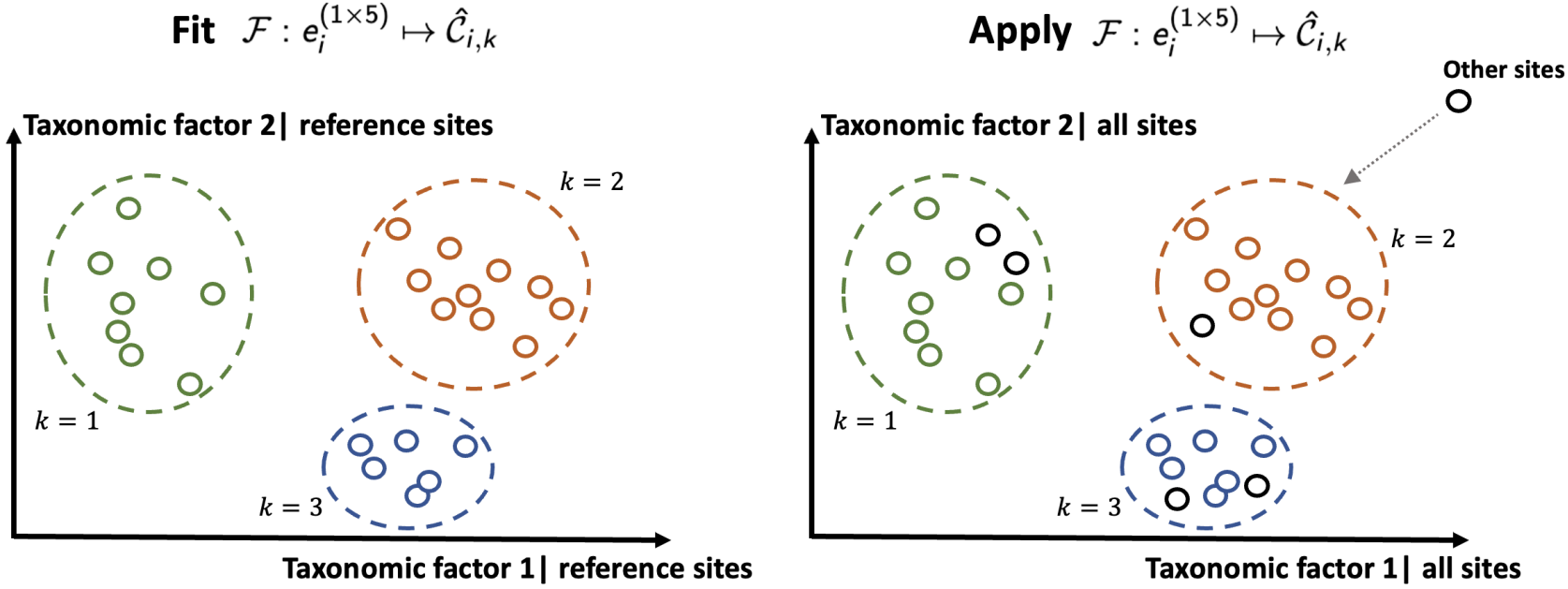
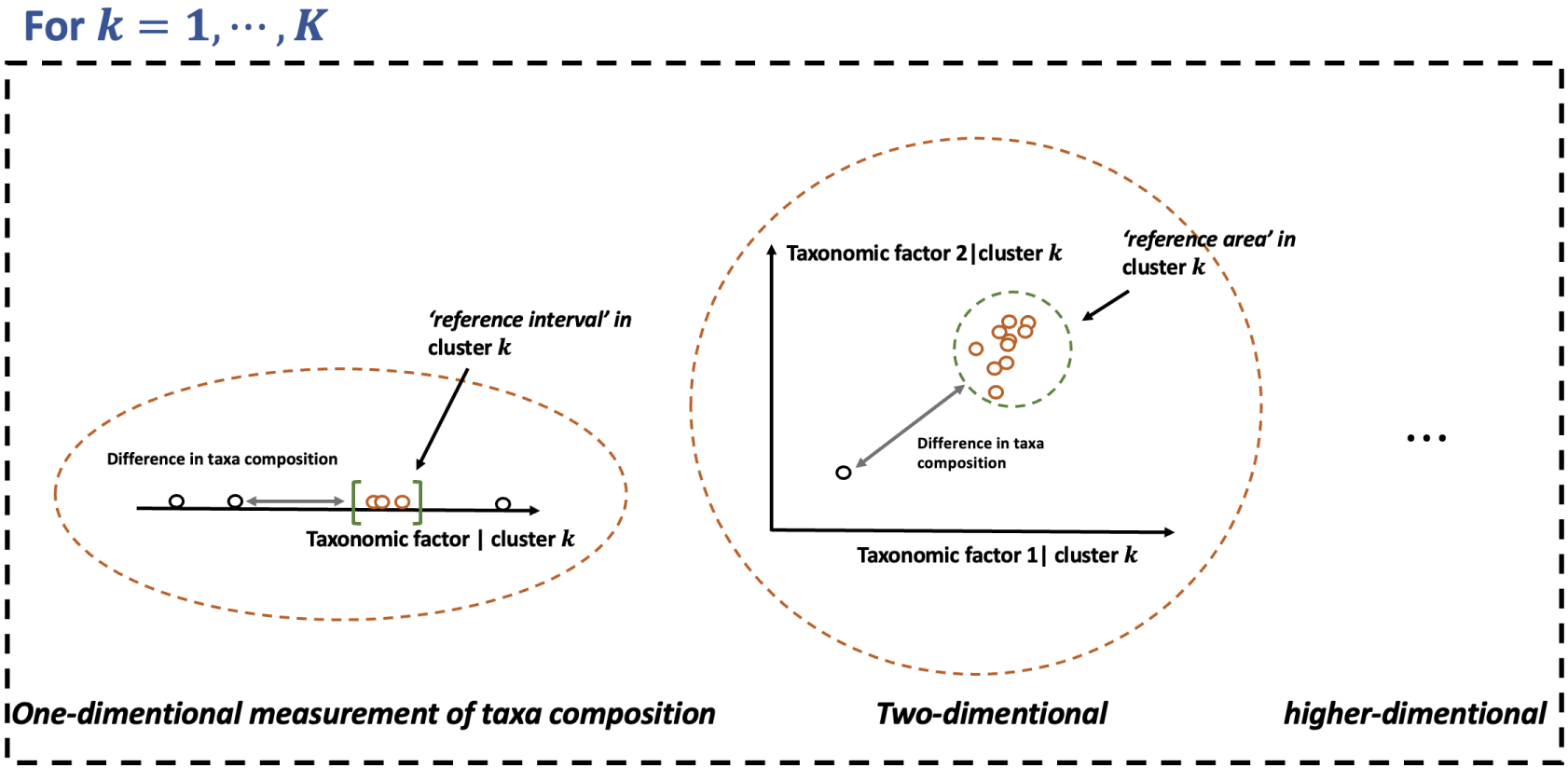


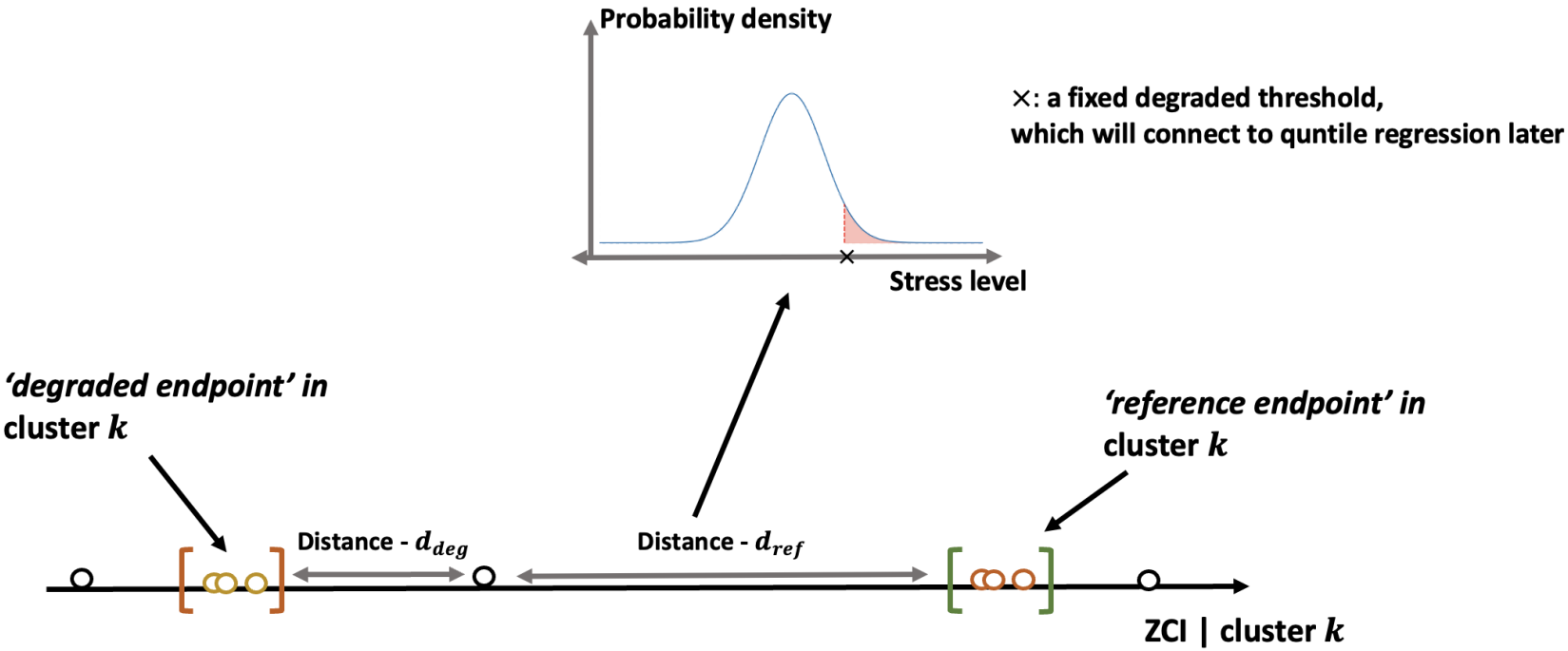
Figure: Discriminant function results on the references and rest sites

Measure the distance between each site to the references in each cluster

In each cluster, the references are of the ideal taxa composition, and the rest sites are not, this difference in taxa composition between them should be caused by human activities - stress level.

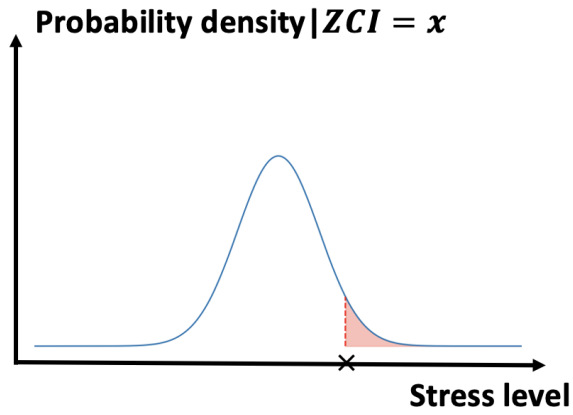


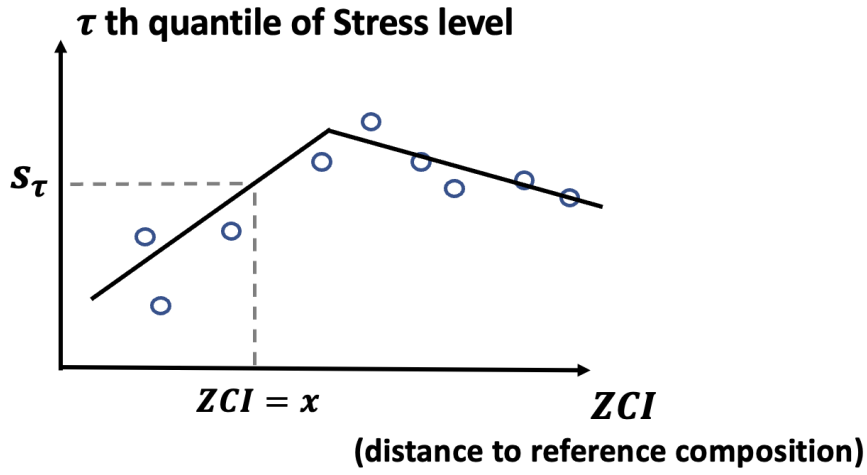
Details in the the one-dimensional taxa difference measurements



Details in one-dimentional measurement of taxa composition

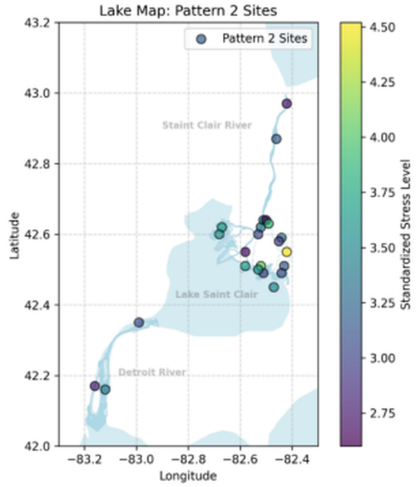
Details in the the one-dimensional taxa difference measurements

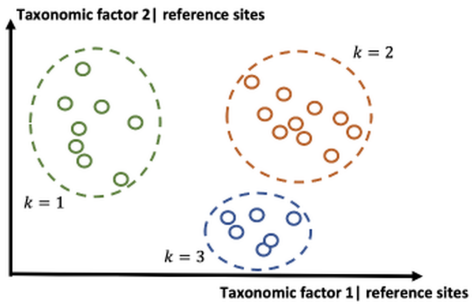


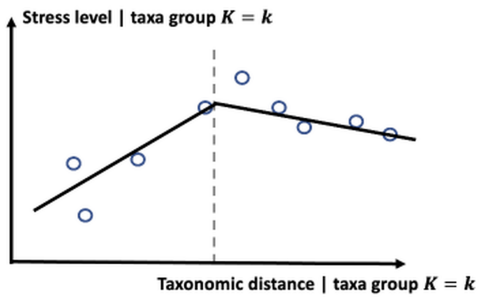


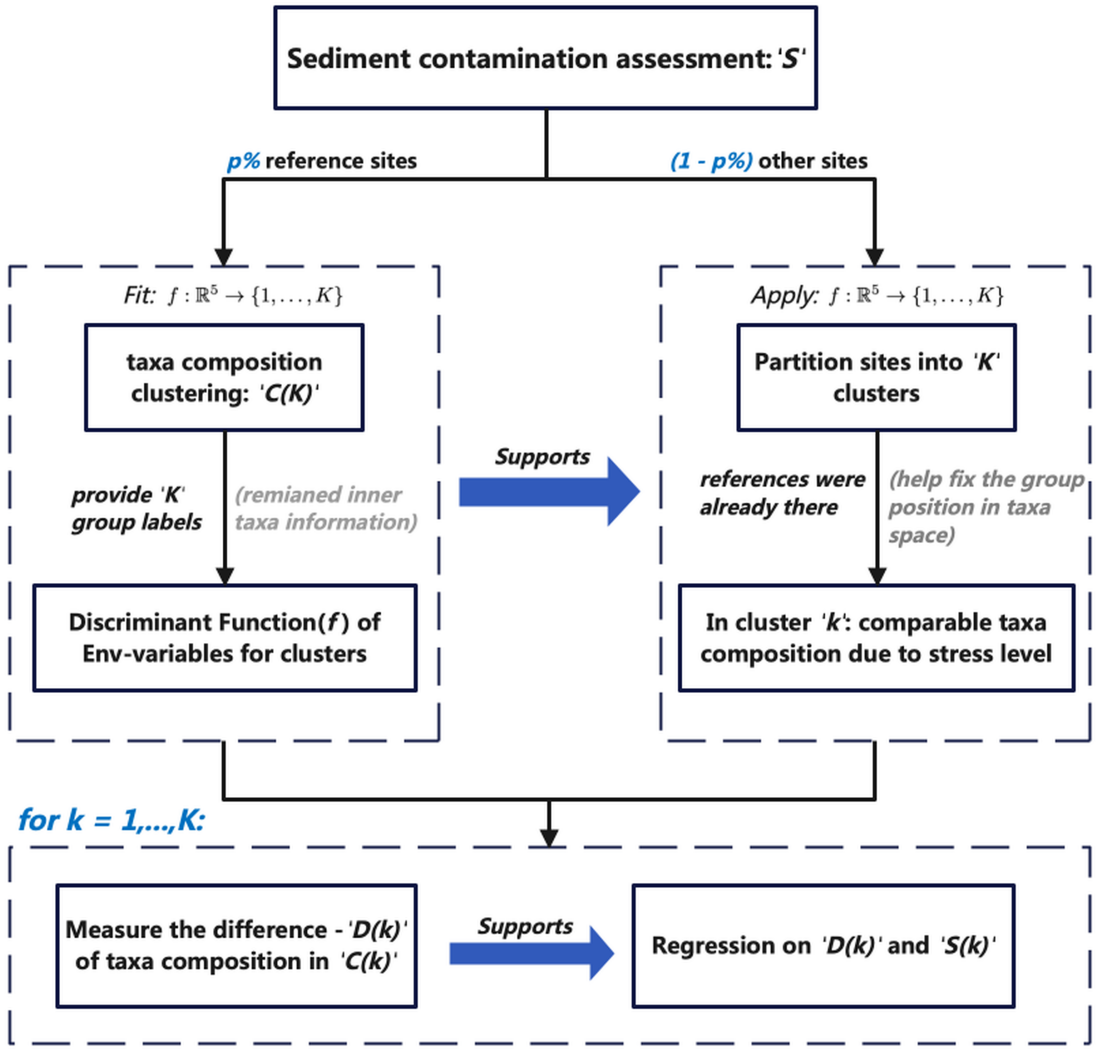
s_τ: the right predicted τ quantile of the stress level that is equal to the threshold value.
The higher the τ parameter required to reach the threshold value, the safer it is to reject 'not degraded' decision.
$\tau = 1$: **all samples** are less than the threshold and support to say 'not degraded'; $\tau \approx 0$: **only one or no sample** is less than the threshold and supports to say 'not degraded'

Recap of the value-based measurements on the stress level and taxa composition







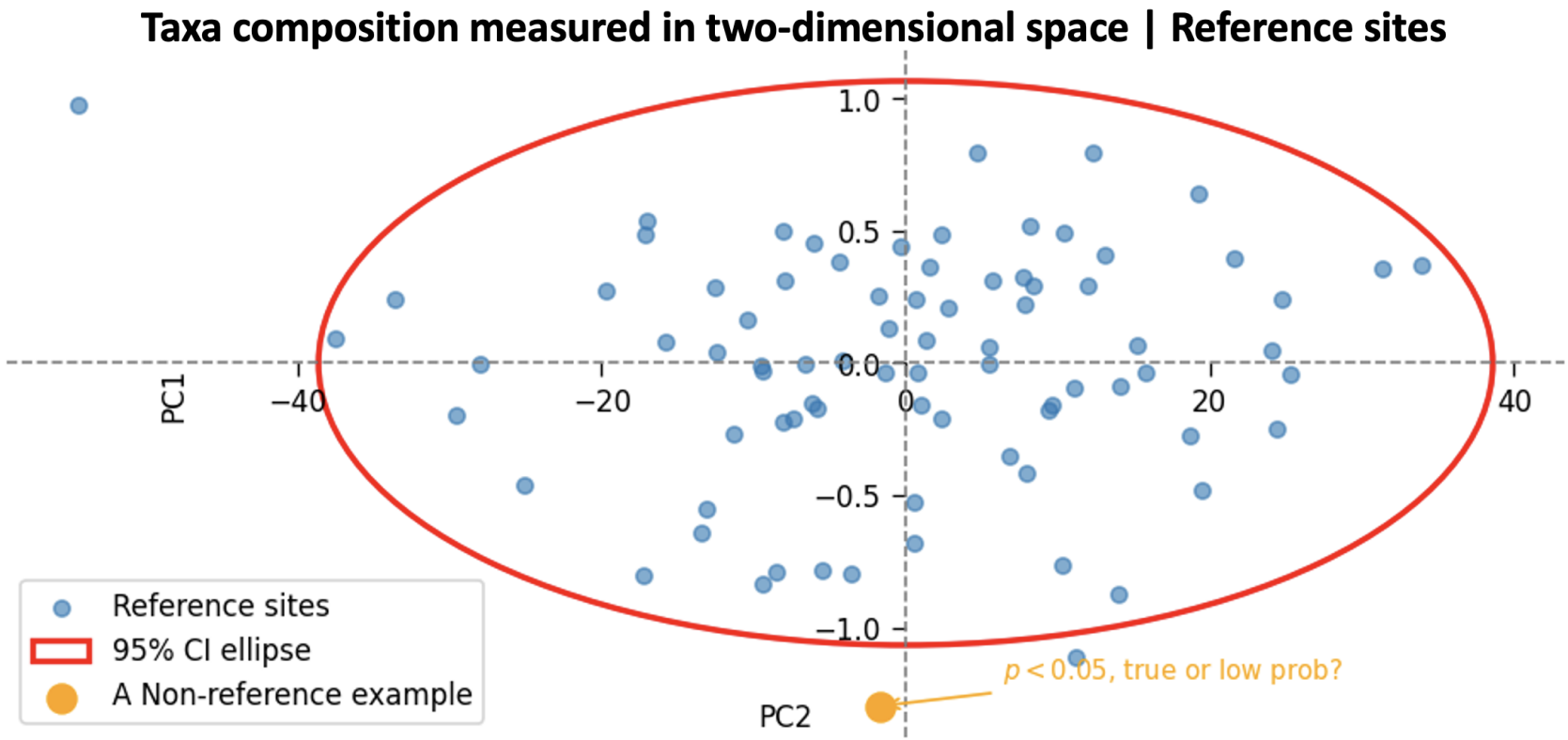


From value-based measurements to vector-based measurements

A value-based measurement, $\mathcal{F} : R^p \rightarrow R$, sacrifices much information in the raw data, even though it carries more information than any single raw data column.
Sacrificing some simplicity, using a vector-based measurement, $\mathcal{F} : R^p \rightarrow R^t (t < p)$, can preserve more information to improve the inference accuracy.
In regression view, it transits from a simple regression to a multivariate regression.

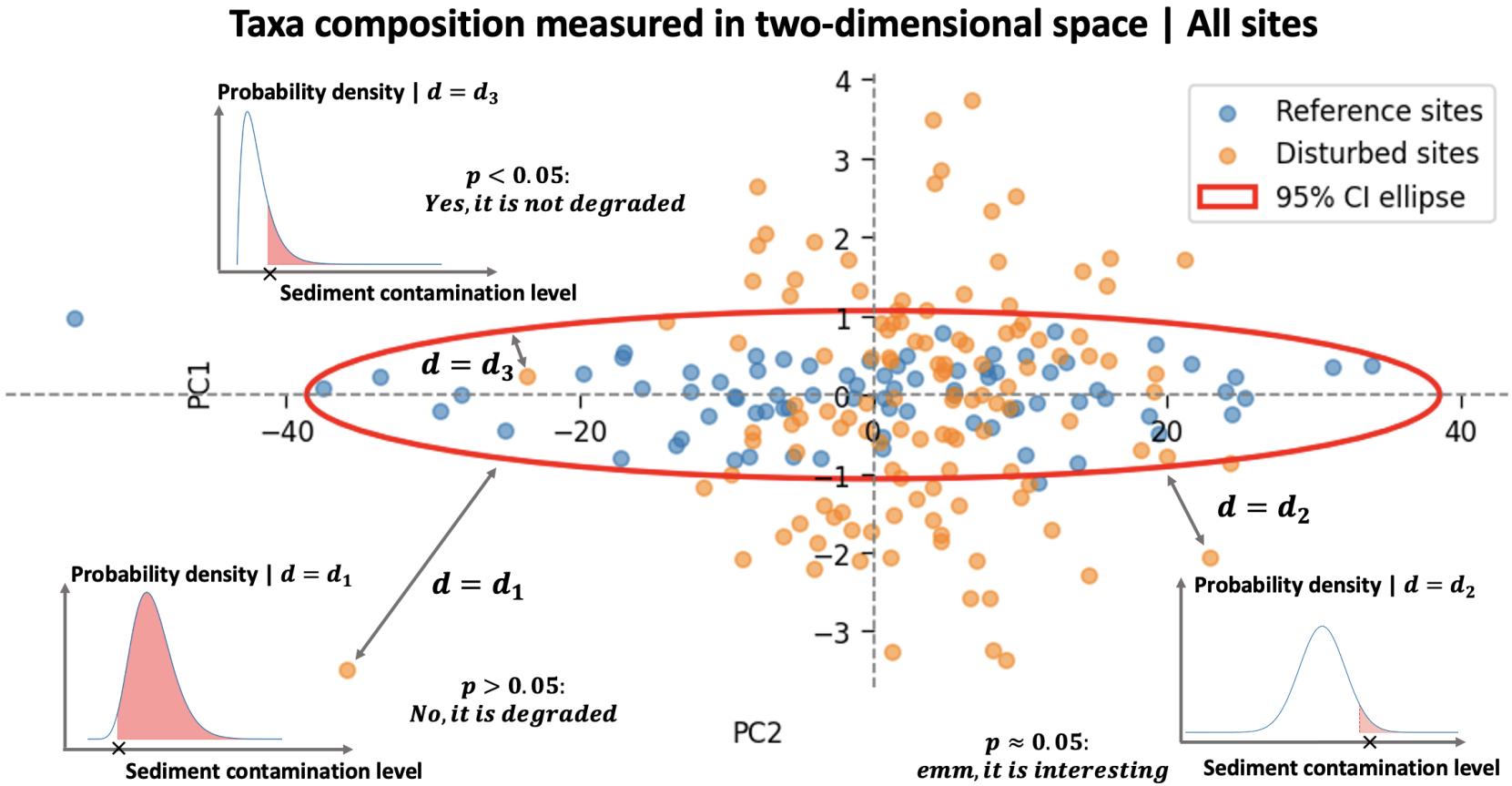
Reference sites in vector-based taxa composition measurements

For $k = 1, ..., K$:



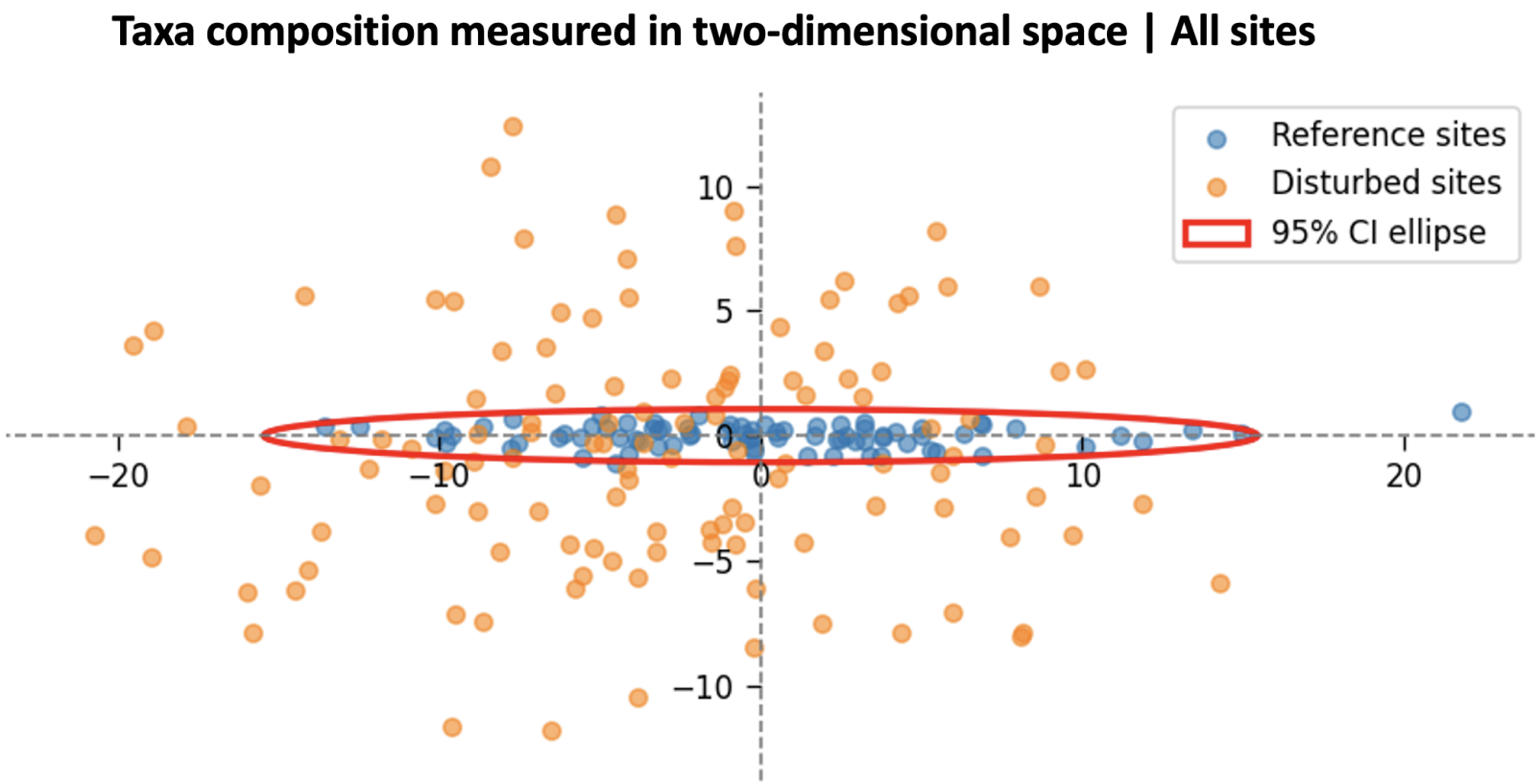
All sites in vector-based taxa composition measurements

For $k = 1, ..., K$:



A hypothetically improved 2-dimensional taxa composition measurements

For $k = 1, ..., K$:



Additional techniques to enhance the process

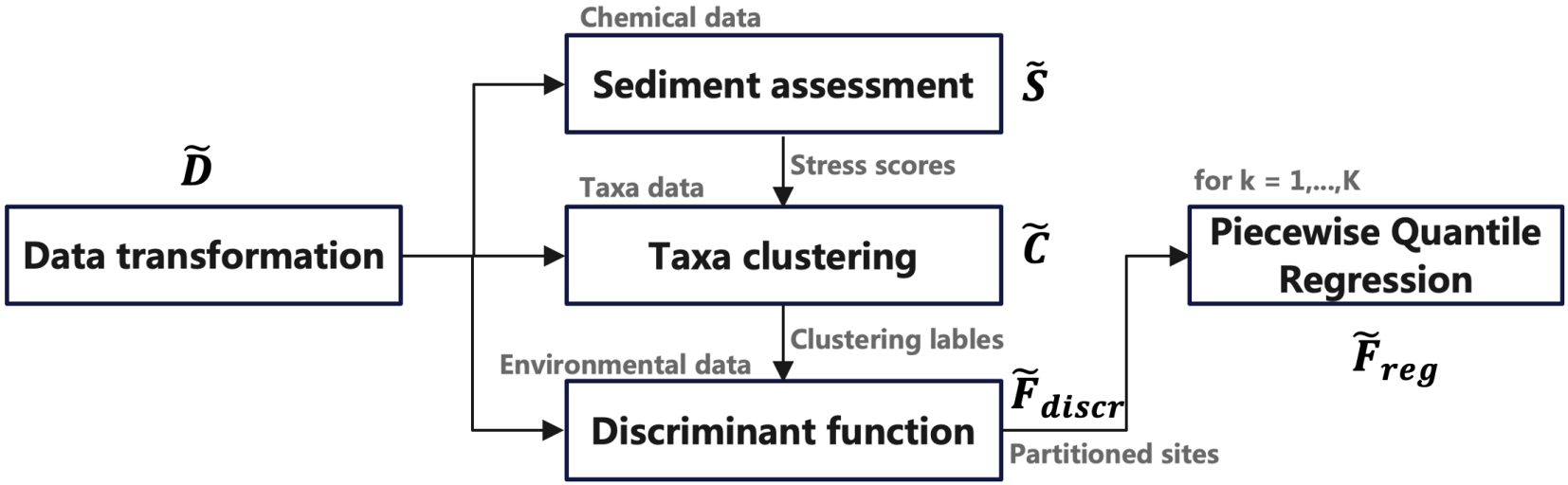
- Supervised learning relevant methods to improve the discriminant function and quantile regression.
- Unsupervised learning relevant methods to improve the clustering of taxa composition.
- Synthetic data generation to partially counter the data scarcity problem.

Potentially core tasks and challenges

- How to define and conduct the measurements of stress level and taxa composition?
- How to balance the correlation and respective interpretability of the measurements? Is it possible to give up the all interpretability of taxa composition and only measure it in the perspective of stress level? (maybe not good)
- Stress level assessment is the foundation influencing all following tasks, how to ensure its accuracy and reliability?

Programming framework and pipeline design

Assuming there are t acceptable methods in each stage, they do not influence our interpretation to the whole process, but they do influence the model performance (the distance measurements in the taxa composition space). One way to find the best combination can be:



$$\left[\tilde{D}, \tilde{S}, \tilde{C}, \tilde{F}_{discr}, \tilde{F}_{reg}\right] \quad Z \times 5$$

Grid search

$$\begin{bmatrix} \tilde{d}_{11}, \tilde{s}_{12}, \tilde{c}_{13}, \tilde{f}_{dis,14}, \tilde{f}_{reg,15} \end{bmatrix} \quad 1 \times 5$$

$$\vdots$$

$$\begin{bmatrix} \tilde{d}_{Z1}, \tilde{s}_{Z2}, \tilde{c}_{Z3}, \tilde{f}_{dis,Z4}, \tilde{f}_{reg,Z5} \end{bmatrix} \quad 1 \times 5$$

Thank you!

Questions?