Data Structure Documentation

Your Name

October 14, 2025

1 Overview/Background

There are 3 seperate data files(table formats) store the three types of data respectively:

- **Environmental Data**
- **Taxa Data**
- **Stressor Data**

These original data files carry more columns that this project needs, some information are structured in an unpleasant way and some columns are repeated across the three. **To keep the original integrity and track of changes, I keep the original files and produce new files the wanted data structure from them.**

The motivation and producing process across the three original files are talked as follows.

2 Environmental data

2.1 Raw data structure

The originally complete environmental data set has many columns, they can be categorized into the 3 types of information for each sampled site:

1. **Sample identifier/Sampling information**
   (a) **Integrated Code**: A tidy unique identifier for each sampled site with the naming of "waterbody + number". But it does not exist in other two files.
   (b) **Site Name(all sources)**: The initial identifier from the three surveys, there are no standard naming rules across the surverys, and there are some missing values.
   (c) **Station NO ID(Zhang)**: The exact same identifier as the "Site Name(all sources)" in all non-NA entries, compiled by Zhang. There are also some missing values in it but it complements all the missing values in "Site Name(all sources)". [1]
   (d) **Waterbody**: The waterbody each site belongs to, there are 3 in total: Detroit River(DR), St. Clair River(SCR), and Lake St. Clair(LSC).
   (e) **Year sampled**: The year each sample was collected. There are more than 3 years of the surverys, minority of them were from 2005.
   (f) *** correspondence**: These '* correspondence' like columns are binary indicators to indicate if there are "*" relevant values missing in the row. They are not needed in this analysis and the missing in any type of variables will be detected and handled in later coding steps.

[1] It is reasonable to assume the missing values in it are deliberate for ease of analysis.

(g) *** vs ***: These "* vs *" like columns are binary indicators, they indicate if the associated taxa or stressor data of the same site(Site Name or Station NO ID) are completely available. These columns are also not needed here and this issue will be solved by the inner-join operation and checking the dropped rows in the merged dataframe.

2. **Environmental Variables:**

   (a) **Latitude**: Latitude of the sampled site.

   (b) **Longitude**: Longitude of the sampled site.

   (c) **LOI/Total Organic Carbon (%)**: Loss in Ignition, a measure of organic matter content in the sediment. It sometimes is named as Total Organic Carbon.

   (d) **Measured Depth (m)**: The depth of the water column at the sampled site.

   (e) **Temperature (C)**: The water temperature at the sampled site in Celsius.

   (f) **Water Do Bottom (mg/L)**: The dissolved oxygen at the bottom of the water column at the sampled site, in mg/L.

   (g) **MPS (Phi)**: Mean Particle Size of the sediment at the sampled site, in Phi.

   (h) **Velocity at bottom (m/s)**[2]: The water velocity at the bottom of the water column at the sampled site, in m/s.

   (i) **L ***: The variables with names starting with "L" are the logorithmically transformed variables of the original environmental variables *. The base of the logorithm is 10, except the *LMPS* taking base 2. To avoid issues with log(0), 1 is added to * before the transformation.

$$L* = log_{10}(* + 1) \quad LMPS = log_2(MPS + 1)$$

3. **Clustering labels of reference sites**

   (a) **Clusters:** The cluster labels(integers) of the selected reference sites in the Detroit River case study. The specific values correspond to the clusters that the reference sites belong to. The missing values indicate the non-reference sites.

   (b) **RelMax RefSite:** A binary indicator to indicate if the site is a reference site or not, at a higher level than the cluster labels.

2.2 Produced data structure from the raw environmental data

As talked above, not all columns in the raw data are needed in this analysis. The wanted information should include:

1. A commonly existing unique identifier across the three data files.

2. The original environmental variables and the log-transformation rules(the log-transformed values are not necessary).

3. The previous clustering labels to be used as a comparison basis.

[2]Only samples from Detroit River have this estimated measurement, but not all of DR samples have this measurement.

Table 1: Produced and kept columns from the raw environmental data

Variable Name	Description
StationID	Unique identifier from the union of the two columns: Site Names (all sources) and Station NO ID (Zhang)
Year	Original Year Sampled column
Waterbody	The waterbody column but simplified into three capital letters: DR, LSC, SCR
Latitude	Latitude of the sampled site
Longitude	Longitude of the sampled site
LOI	The original LOI variable in the raw data.
Measured Depth	The original Measured Depth variable in the raw data.
Temperature	The original Temperature variable in the raw data.
Water DO Bottom	The original Water DO Bottom variable in the raw data.
MPS	The original MPS variable in the raw data.
Velocity at bottom	The original Velocity at bottom variable in the raw data.
clusters	The original cluster labels in the raw data for Detroit River study, indicating the cluster labels for reference sites.
RelMax RefSite	The original RelMax RefSite variable in the raw data, indicating if a site is a reference site or not.

StationID	Year	Waterbody	Latitude	Longitude	LOI (%)	Measured Depth (m)	Temperature (°C)	Water DO Bottom (mg/L)	MPS (Phi)	Velocity at bottom (m/sec)	clusters	RelMax RefSite
001A		DR	42.34645	-82.91807	0.95	1.5	16.00	10.00	3.000	0.16		0.00
003ABC	1999	DR	42.35368	-82.94434	1.14	5.9	15.00	9.80	0.556	0.37		0.00
004ABC	1999	DR	42.35031	-82.93425	0.78	3.2	15.50	10.60	0.567	0.44		0.00
005ABC	1999	DR	42.34243	-82.94559	0.53	1.2	17.00	9.00	0.643	0.08		0.00
006B		DR	42.34356	-82.94187	0.75	1.7	17.50	9.40	0.700	0.12		0.00
007ABC	1999	DR	42.34479	-82.93614	4.17	8.1	16.00	9.30	2.238	0.15		0.00
008A	1999	DR	42.35094	-82.92267	0.82	2.1	16.00	12.50	0.647	0.18		0.00
009B	1999	DR	42.36142	-82.92022	0.96	3.2	16.00	10.00	0.905	0.22		0.00
010B	1999	DR	42.35021	-82.98493	1.43	3.4	22.02	8.68	0.950	0.44		0.00
011A	1999	DR	42.34374	-82.99216	1.57	6.2	21.96	8.55	1.580	0.34	1.00	1.00
108B	1999	DR	42.11543	-83.12142	6.91	2.0	20.82	8.77	2.020	0.03		0.00
109C	1999	DR	42.14403	-83.11672	1.10	3.0	19.50	8.12	1.050	0.24		0.00
10FB	1991	DR	42.33556	-83.01556	2.10	5.5	20.10	7.63	0.395	0.26		0.00
110C		DR	42.12248	-83.13727	8.71	1.5	20.15	8.18	1.120	0.34		0.00
111C	1999	DR	42.13442	-83.13522	1.91	1.4	20.03	7.22	1.691	0.29	1.00	1.00
113B	1999	DR	42.14203	-83.17423	1.32	5.0	19.83	6.67	0.583	0.35		0.00
114B	1999	DR	42.13573	-83.17348	2.27	5.5	19.71	6.88	1.022	0.41	1.00	1.00
115ABC	1999	DR	42.10042	-83.14629	3.20	0.8	20.40	9.77	1.660	0.17	1.00	1.00
116B	1999	DR	42.11421	-83.18026	1.13	3.0	19.79	7.49	0.455	0.30	1.00	1.00
117ABC	1999	DR	42.10747	-83.17921	2.14	4.9	19.56	7.64	0.412	0.38	1.00	1.00
012A	1999	DR	42.34182	-82.99074	1.13	1.5	22.52	10.08	1.604	0.28		0.00
118A	1999	DR	42.11104	-83.17762	3.76	5.8	19.58	7.55	0.490	0.44		0.00

Figure 1: Produced environmental data structure

3 Taxa data

3.1 Raw data structure

The original taxa data has a clearer structure than the environmental data. It has the same columns: integrated Code, Site Names(all sources), Station NO ID(Zhang), Waterbody, Latitude, Longitude, Year Sampled. These columns are the same as the ones in the environmental data. The other columns are the 16 taxa variables and two extra columns: Validation Code and Site.

These repeated columns are not shown in this section, the other columns are:

1. **Sampling information*** (×7): the same first 7 columns as the environmental data, showing the original and compiled identifiers, waterbody, location and year information.

2. **Taxa***(×16): The 16 taxa variables, they are the abundances of the 16 benthic invertebrate taxa groups.

3. **Validation Code**: A binary indicator to show if the associated data exists in other two data files for this site (identified by Site Name or Station NO ID).

4. **Site**: The union of the two columns: Site Names (all sources) and Station NO ID (Zhang), as a unique identifier.

On top of the original taxa data, there are 5 sub-tables showing the clustering results for two level studies: Corridor-wide and Detroit River. The first 2 sub-tables are the clustering results(2 clusters) for the Corridor-wide study, containing all 311 sites in the taxa data. The rest 3 sub-tables are the clustering results(3 cluster) for the Detroit River study, containing 213 sites in the taxa data.

Therefore, the first two sub-tables are concated vertically to produce a clustering result for all 311 sites for corridor-wide study. The last three sub-tables are concated vertically to produce a clustering result for all 213 sites for Detroit River study.

1. **Corridor-wide cluster**: The clustering results for all 311 sites across the corridor, 2 clusters in total.

2. **Detroit River cluster**: The clustering results for all 213 sites in Detroit River, 3 clusters in total.

3.2 Produced data structure from the raw taxa data

The sampling information columns contain the same information as the same columns in environmental data, but the unique identifier - StationID - is needed for the merging operation later. Therefore, the wanted columns should include:

Table 2: Kept columns from the raw taxa data

Variable Name	Description
StationID	Unique identifier from the union of the two columns: Site Names (all sources) and Station NO ID (Zhang)
Year	Original Year Sampled column from taxa data. To a site existing in both files, its value equal to the the year value from another data file.
Waterbody	The waterbody column but simplified into three capital letters: DR, LSC, SCR. Same value for the site that exists in both files.
Latitude	Latitude of the sampled site. Same value for the site that exists in both files.
Longitude	Longitude of the sampled site. Same value for the site that exists in both files.
taxa variables(×16)	The original taxa variables in the raw data, indicating the abundances of the 16 benthic invertebrate taxa groups.
Corridor clusters	The clustering results for all 311 sites across the corridor, 2 clusters in total.
Detroit clusters	The clustering results for all 213 sites in Detroit River, 3 clusters in total.

Station	Waterbody	Latitude	Longitude	Year	Oligochaeta	Nematoda	Chironomidae	Ceratopogonidae	Hexagenia	Caenis
003ABC	DR	42.353685	-82.944344	1999	2.269984059	0.563545972	1.655598043	3.20343E-16	0.30911637	3.20343E-16
004ABC	DR	42.350308	-82.934247	1999	3.808503717	3.512090568	6.242625171	3.20343E-16	3.2034E-16	3.20343E-16
005ABC	DR	42.342434	-82.945587	1999	4.820162781	5.519382132	4.733464322	3.20343E-16	0.81594939	3.20343E-16
007ABC	DR	42.344791	-82.936142	1999	1.449186678	0.181595696	2.271083442	3.20343E-16	0.99957288	0.033941359
008A	DR	42.350944	-82.922672	1999	1.082238261	0.246371284	2.499652915	3.20343E-16	0.12843842	3.20343E-16
009B	DR	42.361423	-82.920223	1999	4.412094224	3.20343E-16	6.254789008	3.20343E-16	3.2034E-16	1.29235498
010B	DR	42.350205	-82.984932	1999	0.538526928	3.20343E-16	2.874765816	3.20343E-16	3.2034E-16	3.20343E-16
011A	DR	42.343740	-82.992158	1999	3.713829095	4.96823042	5.71466182	2.010888316	3.2034E-16	3.20343E-16
012A	DR	42.341819	-82.990738	1999	4.573314991	2.172734535	6.117301336	1.461730735	3.2034E-16	2.172734535
013A	DR	42.346341	-82.987057	1999	4.45805522	3.231372253	6.014274131	3.20343E-16	3.2034E-16	3.20343E-16
014B	DR	42.334030	-83.015411	1999	4.530781463	3.648288436	4.268227075	3.20343E-16	3.2034E-16	3.20343E-16
015C	DR	42.337440	-83.010956	1999	5.266786541	3.025535092	4.093511886	3.20343E-16	3.2034E-16	3.20343E-16
016C	DR	42.333507	-82.957874	1999	5.330576553	5.083698118	4.785669733	3.20343E-16	3.2034E-16	3.20343E-16
017B	DR	42.341360	-83.000059	1999	2.884203001	4.544758959	6.110302818	3.20343E-16	1.64508349	3.20343E-16
018A	DR	42.328771	-83.006687	1999	3.20343E-16	0.670582926	1.751405406	3.20343E-16	1.98498296	3.20343E-16
019B	DR	42.332702	-83.000258	1999	4.466034353	3.883764286	5.922031956	3.20343E-16	3.2034E-16	3.20343E-16
021B	DR	42.333326	-82.985069	1999	2.526545814	3.20343E-16	3.395585137	3.20343E-16	3.2034E-16	3.20343E-16
022B	DR	42.332880	-82.980176	1999	3.811828399	3.20343E-16	3.277337944	3.20343E-16	3.2034E-16	3.20343E-16
023C	DR	42.335968	-82.953532	1999	5.442943496	1.807354922	5.599912842	3.20343E-16	1.80735492	1.807354922
024C	DR	42.353578	-82.969012	1999	0.703536296	0.188609565	0.640305535	3.20343E-16	3.2034E-16	3.20343E-16
025B	DR	42.337972	-82.970588	1999	5.039660087	3.270326898	5.917896175	3.20343E-16	3.2034E-16	3.20343E-16
026C	DR	42.337020	-82.966771	1999	3.025535092	1.478047297	2.192645078	3.20343E-16	3.2034E-16	3.20343E-16
027B	DR	42.328035	-83.030038	1999	2.299118212	3.20343E-16	3.20343E-16	3.20343E-16	3.2034E-16	3.20343E-16
029C	DR	42.309683	-83.081584	1999	4.142957954	3.20343E-16	3.20343E-16	3.20343E-16	3.2034E-16	3.20343E-16

Figure 2: Produced taxa data structure from the raw data. (There are resting columns not shown, 16 taxa variables/columns in total)

After the cleaning operation, the produced taxa data structure is shown in Figure 2.

The clustering results for the two studies: corridor-wide and Detroit River, are stored in two seperate tables with StationID as index for easy merging later. These tables of clustering results are shown in Figure 3 and 4.

StationID	corridor_cluster
003ABC	1
004ABC	1
005ABC	1
007ABC	1
008A	1
009B	1
010B	1
011A	1
012A	1
013A	1
014B	1
015C	1
016C	1
017B	1
019B	1
022B	1
023C	1
024C	1
025B	1
027B	1
029C	1
030ABC	1
036C	1
037B	1
042C	1
043ABC	1
044A	1

Figure 3: Clustering results for all 311 sites across the corridor(snapshot)

StationID	DR_cluster
005ABC	1.00
007ABC	1.00
008A	1.00
009B	1.00
011A	1.00
016C	1.00
017B	1.00
019B	1.00
023C	1.00
025B	1.00
027B	1.00
042C	1.00
044A	1.00
045B	1.00
054B	1.00
059ABC	1.00
067B	1.00
070B	1.00
079C	1.00
080C	1.00
081B	1.00
082A	1.00
084A	1.00

Figure 4: Clustering results for all 213 sites in Detroit River(snapshot)

4 Merged environmental and taxa data

At this stage, the produced environmental data and taxa data share the same unique identifier - StationID. An inner-join operation is performed on the StationID column to combine the two datasets, producing a complete dataset of sites with both types of data available. This join operation also reveals the sites that do not have associated data in either of the two datasets, which are dropped in the final merged dataset.

As the data operation logic talked in the proposal, the conceptually merging process and the merged data structure are like:

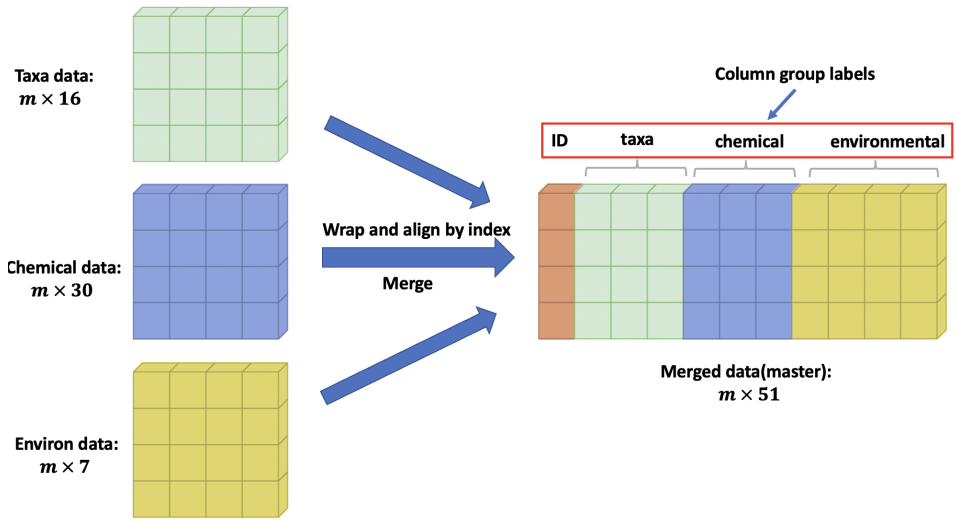

Figure 5: Idealization: how the cleaned data sets are merged together and the resulting data structure.

block	sample_info				environmental						taxa									
subblock	raw				raw						raw									
var	Latitude	Longitude	Waterbody	Year	LOI (%)	MPS (Phi)	easured Depth (m	Temperature (oC)	city at bottom (m)	O Bottom	Acari	Amphipoda	Caenis	ratopogonid	hironomida	Dreissena	Gastropoda	Hexagenia	Hirudinea	d
StationID																				
003ABC	42.3537	-82.9443	DR	1999	1.14	0.55555556	5.9436	15	0.37186	9.8	3.2E-16	4.5299926	3.2E-16	3.2E-16	1.6556	6.08292	3.2E-16	0.30912	3.2E-16	
004ABC	42.3503	-82.9342	DR	1999	0.78	0.56666667	3.2004	15.5	0.44179	10.6	0.80198	3.203E-16	3.2E-16	3.2E-16	6.24263	0.80198	0.45601	3.2E-16	3.2E-16	
005ABC	42.3424	-82.9456	DR	1999	0.53	0.64285714	1.2192	17	0.08216	9	1.01002	3.203E-16	3.2E-16	3.2E-16	4.73346	0.17216	0.46491	0.81595	3.2E-16	
007ABC	42.3448	-82.9361	DR	1999	4.17333	2.23809524	8.0772	16	0.14675	9.3	3.2E-16	1.8402416	0.03394	3.2E-16	2.27108	6.47571	3.2E-16	0.99957	0.23027	
008A	42.3509	-82.9227	DR	1999	0.82	0.64705882	2.1336	16	0.18312	12.5	3.2E-16	4.2242622	3.2E-16	3.2E-16	2.49965	6.26637	0.24637	0.12844	3.2E-16	
009B	42.3614	-82.9202	DR	1999	0.96	0.9047619	3.2004	16	0.22287	10	1.29235	3.203E-16	1.29235	3.2E-16	6.25479	3.2E-16	3.2E-16	3.2E-16	1.29235	
010B	42.3502	-82.9849	DR	1999	1.43	0.95	3.3528	22.02	0.44311	8.68	0.53853	5.5171617	3.2E-16	3.2E-16	2.87477	5.57308	3.2E-16	3.2E-16	3.2E-16	
011A	42.3437	-82.9922	DR	1999	1.57	1.58024691	6.2484	21.96	0.34082	8.55	2.01089	3.203E-16	3.2E-16	2.010888	5.71466	3.2E-16	3.2E-16	3.2E-16	3.2E-16	
012A	42.3418	-82.9907	DR	1999	1.13	1.60377358	1.524	22.52	0.28202	10.08	3.2E-16	3.203E-16	2.17273	1.461731	6.1173	3.2E-16	3.2E-16	3.2E-16	3.2E-16	
013A	42.3463	-82.9871	DR	1999	1.99	1.62264151	1.22	22.36	0.31394	9.3	3.2E-16	0.7649412	3.2E-16	3.2E-16	6.01427	2.55952	3.2E-16	3.2E-16	1.26219	
014B	42.334	-83.0154	DR	1999	1.47	1.54285714	8.2296	21.87	0.34471	8.64	3.2E-16	4.2682271	3.2E-16	3.2E-16	4.26823	4.94522	3.2E-16	3.2E-16	3.2E-16	
015C	42.3374	-83.011	DR	1999	1.27	1	6.5532	22	0.31741	8.24	3.2E-16	3.0255351	3.2E-16	3.2E-16	4.09351	4.59779	3.2E-16	3.2E-16	3.2E-16	
016C	42.3335	-82.9579	DR	1999	0.87	0.87234043	1.524	19.66	0.09968	7.54	3.2E-16	3.203E-16	3.2E-16	3.2E-16	4.78567	3.2E-16	3.2E-16	3.2E-16	3.2E-16	
017B	42.3414	-83.0001	DR	1999	1.69	1.64444444	3.3528	21.96	0.3389	8.77	1.04532	3.203E-16	3.2E-16	3.2E-16	6.1103	3.2E-16	3.2E-16	1.64508	3.2E-16	
018A	42.3288	-83.0067	DR	1999	2.545	1.11764706	12.6492	19.37	0.55189	7.12	3.2E-16	4.0854528	3.2E-16	3.2E-16	1.75141	6.29489	3.2E-16	1.98498	3.2E-16	
019B	42.3327	-83.0003	DR	1999	2.19	1.65909091	3.6576	20.34	0.31106	8.11	3.2E-16	1.5032787	3.2E-16	3.2E-16	5.92203	3.2E-16	3.2E-16	3.2E-16	3.2E-16	
021B	42.3333	-82.9851	DR	1999	0.97	0.25	12.4968	19.68	0.45633	7.16	3.2E-16	3.3955851	3.2E-16	3.2E-16	3.39559	6.27035	3.2E-16	3.2E-16	3.2E-16	
022B	42.3329	-82.9802	DR	1999	1.665	1.74358974	11.8872	19.26	0.53266	7.09	3.2E-16	3.8118284	3.2E-16	3.2E-16	3.27734	6.04914	3.2E-16	3.2E-16	3.2E-16	
023C	42.336	-82.9535	DR	1999	0.97	1.34285714	1.524	19.47	0.04348	7.13	3.2E-16	3.203E-16	1.80735	3.2E-16	5.59991	1.80735	3.2E-16	1.80735	3.2E-16	
024C	42.3536	-82.969	DR	1999	4.89	2	7.9248	21.98	0.47968	8.65	3.2E-16	5.4622681	3.2E-16	3.2E-16	0.64031	5.58488	0.64031	3.2E-16	3.2E-16	
025B	42.338	-82.9706	DR	1999	2.76	2.25423729	1.3716	20.15	0.13463	6.66	3.2E-16	3.203E-16	3.2E-16	3.2E-16	5.9179	3.2E-16	3.2E-16	3.2E-16	3.2E-16	
026C	42.337	-82.9668	DR	1999	1.795	0.3125	8.9916	19.68	0.55893	6.88	3.2E-16	3.5501971	3.2E-16	3.2E-16	2.19265	5.98868	3.2E-16	3.2E-16	3.2E-16	
027B	42.328	-83.03	DR	1999	1.7	2.41666667	11.5824	12.15	0.44522	9.2	1.56598	4.06059	3.2E-16	3.2E-16	3.2E-16	6.27558	3.2E-16	3.2E-16	3.2E-16	
029C	42.3097	-83.0816	DR	1999	1.39	1.22222222	9.144	14.4	0.45044	10.5	3.2E-16	3.203E-16	3.2E-16	3.2E-16	3.2E-16	6.39803	3.2E-16	3.2E-16	3.2E-16	
030ABC	42.322	-83.0554	DR	1999	1.81	0.91891892	11.8872	16.66	0.41049	9.18	3.2E-16	4.0452334	3.2E-16	3.2E-16	1.26499	6.309	3.2E-16	3.2E-16	3.2E-16	
031A	42.318	-83.0519	DR	1999	6.67	-0.16	12.4968	12.04	0.47371	9.36	3.2E-16	3.9341121	3.2E-16	3.2E-16	2.06164	6.17849	3.2E-16	3.2E-16	3.2E-16	
033ABC	42.2868	-83.0941	DR	1999	1.45	1.1	9.144	19.13	0.31292	6.55	3.2E-16	1.5556469	3.2E-16	3.2E-16	5.63537	3.2E-16	3.2E-16	2.76955	0.71945	
034C	42.2973	-83.0933	DR	1999	19.15	-1	11.5824	19.07	0.54223	6.85	3.2E-16	5.8819341	3.2E-16	3.2E-16	3.2E-16	5.27356	3.2E-16	3.2E-16	3.2E-16	
035C	42.2932	-83.0902	DR	1999	3.99	-1	12.4968	18.99	0.51596	7.39	3.2E-16	3.9341121	3.2E-16	3.2E-16	2.52655	6.13026	3.2E-16	3.2E-16	3.2E-16	

Figure 6: Practically: produced merged environmental and taxa data structure. (The multi-level column indices are highlighted in colors; It is a snapshot of first columns, the clustering result columns remained at the right end of the table.)

5 The purpose of maintaining datasets in a wider table with multi-level column indices

There are three main reasons for maintaining the three original datasets in a wider table with multi-level column indices, they are listed as the order of importance as follows:

1. **Avoiding to create and spread many data objects in both coding and storing stages.**

   The framework of this project is designed to be computationally demanding, many intermediate computing results are produced with numbers of testing and tuning tasks, and the results will be used for guiding the next steps. If these intermediate results are stored seperately and instantiated as seperate data objects in the coding space, it will be hard to tract and manage them, even the naming work will be a burden. Additionaly, many test and tuning tasks need to bundle the data and intermediate results, it requires the aligning and matching operations across these objects if they were stored separately, very error-prone and inefficient.

   Therefore, maintaining only the core data and the conclusive intermediate results in a single data object is good for both coding work and data management.

2. **A faster way to do column-wise operation across all types of data.**

   Keeping all important data in one data object makes the column-wise operations broadcasting to all columns across all types of data easy. For example, assuming the pollution scores have been stored for all sites in the merged data structure, the pollution-relevant grouping operations will automatically group the sites with both environmental and taxa data. It saves the time for the process: *grouping on pollution scores* $\rightarrow$ *group indexing on environmental and taxa data* $\rightarrow$ *computing on each group.*

3. **Easier to track, read and inspire ideas from the data.**

   Keeping columns from all types of data together produces a larger(or complete) matrix of data, a more comprehensive view of the data. It is easier to read and understand the data and structure by leveraging the visually clear table format. This sub-block and complete view of the data also inspires ideas for analysis and modeling.

block	sample_info				environmental						taxa								
subblock	raw				raw						raw								
var	Latitude	Longitude	Waterbody	Year	LOI (%)	MPS (Phi)	easured Depth (m	Temperature (oC)	city at bottom (m)	O Bottom	Acari	Amphipoda	Caenis	ratopogonid	hironomid	Dreissena	Gastropoda	Hexagenia	Hirudinea
StationID																			
003ABC	42.3537	-82.9443	DR	1999	1.14	0.55555556	5.9436	15	0.37186	9.8	3.2E-16	4.5299926	3.2E-16	3.2E-16	1.6556	6.08292	3.2E-16	0.30912	3.2E-16
004ABC	42.3503	-82.9342	DR	1999	0.78	0.56666667	3.2004	15.5	0.44179	10.6	0.80198	3.203E-16	3.2E-16	3.2E-16	6.24263	0.80198	0.45601	3.2E-16	3.2E-16
005ABC	42.3424	-82.9456	DR	1999	0.53	0.64285714	1.2192	17	0.08216	9	1.01002	3.203E-16	3.2E-16	3.2E-16	4.73346	0.17216	0.46491	0.81595	3.2E-16
007ABC	42.3448	-82.9361	DR	1999	4.17333	2.23809524	8.0772	16	0.14675	9.3	3.2E-16	1.8402416	0.03394	3.2E-16	2.27108	6.47571	3.2E-16	0.99957	0.23027
008A	42.3509	-82.9227	DR	1999	0.82	0.64705882	2.1336	16	0.18312	12.5	3.2E-16	4.2242622	3.2E-16	3.2E-16	2.49965	6.26637	0.24637	0.12844	3.2E-16
009B	42.3614	-82.9202	DR	1999	0.96	0.9047619	3.2004	16	0.22287	10	1.29235	3.203E-16	1.29235	3.2E-16	6.25479	3.2E-16	3.2E-16	3.2E-16	1.29235
010B	42.3502	-82.9849	DR	1999	1.43	0.95	3.3528	22.02	0.44311	8.68	0.53853	5.5171617	3.2E-16	3.2E-16	2.87477	5.57308	3.2E-16	3.2E-16	3.2E-16
011A	42.3437	-82.9922	DR	1999	1.57	1.58024691	6.2484	21.96	0.34082	8.55	2.01089	3.203E-16	3.2E-16	2.010888	5.71466	3.2E-16	3.2E-16	3.2E-16	3.2E-16
012A	42.3418	-82.9907	DR	1999	1.13	1.60377358	1.524	22.52	0.28202	10.08	3.2E-16	3.203E-16	2.17273	1.461731	6.1173	3.2E-16	3.2E-16	3.2E-16	3.2E-16
013A	42.3463	-82.9871	DR	1999	1.99	1.62264151	1.22	22.36	0.31394	9.3	3.2E-16	0.7649412	3.2E-16	3.2E-16	6.01427	2.55952	3.2E-16	3.2E-16	1.26219
014B	42.334	-83.0154	DR	1999	1.47	1.54285714	8.2296	21.87	0.34471	8.64	3.2E-16	4.2682271	3.2E-16	3.2E-16	4.26823	4.94522	3.2E-16	3.2E-16	3.2E-16
015C	42.3374	-83.011	DR	1999	1.27	1	6.5532	22	0.31741	8.24	3.2E-16	3.0255351	3.2E-16	3.2E-16	4.09351	4.59779	3.2E-16	3.2E-16	3.2E-16
016C	42.3335	-82.9579	DR	1999	0.87	0.87234043	1.524	19.66	0.09968	7.54	3.2E-16	3.203E-16	3.2E-16	3.2E-16	4.78567	3.2E-16	3.2E-16	3.2E-16	3.2E-16
017B	42.3414	-83.0001	DR	1999	1.69	1.64444444	3.3528	21.96	0.3389	8.77	1.04532	3.203E-16	3.2E-16	3.2E-16	6.1103	3.2E-16	3.2E-16	1.64508	3.2E-16
018A	42.3288	-83.0067	DR	1999	2.545	1.11764706	12.6492	19.37	0.55189	7.12	3.2E-16	4.0854528	3.2E-16	3.2E-16	1.75141	6.29489	3.2E-16	1.98498	3.2E-16
019B	42.3327	-83.0003	DR	1999	2.19	1.65909091	3.6576	20.34	0.31106	8.11	3.2E-16	1.5032787	3.2E-16	3.2E-16	5.92203	3.2E-16	3.2E-16	3.2E-16	3.2E-16
021B	42.3333	-82.9851	DR	1999	0.97	0.25	12.4968	19.68	0.45633	7.16	3.2E-16	3.3955851	3.2E-16	3.2E-16	3.39559	6.27035	3.2E-16	3.2E-16	3.2E-16
022B	42.3329	-82.9802	DR	1999	1.665	1.74358974	11.8872	19.26	0.53266	7.09	3.2E-16	3.8118284	3.2E-16	3.2E-16	3.27734	6.04914	3.2E-16	3.2E-16	3.2E-16
023C	42.336	-82.9535	DR	1999	0.97	1.34285714	1.524	19.47	0.04348	7.13	3.2E-16	3.203E-16	1.80735	3.2E-16	5.59991	1.80735	3.2E-16	1.80735	3.2E-16
024C	42.3536	-82.969	DR	1999	4.89	2	7.9248	21.98	0.47968	8.65	3.2E-16	5.4622681	3.2E-16	3.2E-16	0.64031	5.58488	0.64031	3.2E-16	3.2E-16
025B	42.338	-82.9706	DR	1999	2.76	2.25423729	1.3716	20.15	0.13463	6.66	3.2E-16	3.203E-16	3.2E-16	3.2E-16	5.9179	3.2E-16	3.2E-16	3.2E-16	3.2E-16
026C	42.337	-82.9668	DR	1999	1.795	0.3125	8.9916	19.68	0.55893	6.88	3.2E-16	3.5501971	3.2E-16	3.2E-16	2.19265	5.98868	3.2E-16	3.2E-16	3.2E-16
027B	42.328	-83.03	DR	1999	1.7	2.41666667	11.5824	12.15	0.44522	9.2	1.56598	4.06059	3.2E-16	3.2E-16	3.2E-16	6.27558	3.2E-16	3.2E-16	3.2E-16
029C	42.3097	-83.0816	DR	1999	1.39	1.22222222	9.144	14.4	0.45044	10.5	3.2E-16	3.203E-16	3.2E-16	3.2E-16	3.2E-16	6.39803	3.2E-16	3.2E-16	3.2E-16
030ABC	42.322	-83.0554	DR	1999	1.81	0.91891892	11.8872	16.66	0.41049	9.18	3.2E-16	4.0452334	3.2E-16	3.2E-16	1.26499	6.309	3.2E-16	3.2E-16	3.2E-16
031A	42.318	-83.0519	DR	1999	6.67	-0.16	12.4968	12.04	0.47371	9.36	3.2E-16	3.9341121	3.2E-16	3.2E-16	2.06164	6.17849	3.2E-16	3.2E-16	3.2E-16
033ABC	42.2868	-83.0941	DR	1999	1.45	1.1	9.144	19.13	0.31292	6.55	3.2E-16	1.5556469	3.2E-16	3.2E-16	5.63537	3.2E-16	3.2E-16	2.76955	0.71945
034C	42.2973	-83.0933	DR	1999	19.15	-1	11.5824	19.07	0.54223	6.85	3.2E-16	5.8819341	3.2E-16	3.2E-16	3.2E-16	5.27356	3.2E-16	3.2E-16	3.2E-16
035C	42.2932	-83.0902	DR	1999	3.99	-1	12.4968	18.99	0.51596	7.39	3.2E-16	3.9341121	3.2E-16	3.2E-16	2.52655	6.13026	3.2E-16	3.2E-16	3.2E-16

Figure 7: An instance of the merged data structure with multi-level column indices that meet the above purposes.

6 More to come later: Stressor data and the final complete data set.