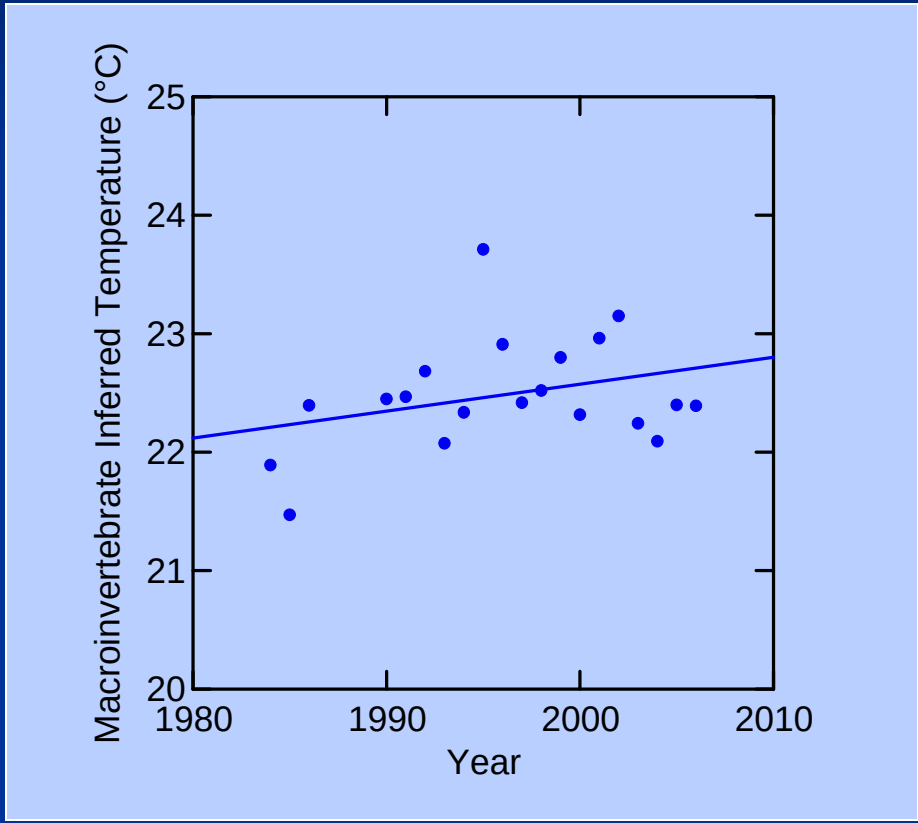
Screening State Biomonitoring Data for Long-Term Trend Analyses



New England Association of Environmental Biologists
March 19, 2009
Jen Stamp, Tetra Tech

OVERVIEW

Why do long-term trend analyses?

How to screen and prepare the data

- Where to start
- Tips
  - Selection of appropriate operational taxonomic units (OTUs)
  - Use of Non-metric Multidimensional Scaling (NMDS)
  - Selection of sites and samples

Reasons why you should consider doing trend analyses with your long-term datasets

WHAT ARE TREND ANALYSES?

Evaluations of how things have changed over time (short-, intermediate- and long-term).

An example of one that is being performed right now is the Global Change Research Program's national initiative examining the effects of climate change on aquatic ecosystems and biological indicators (EPA NCEA/ORD)

Four state biomonitoring databases from different regions are being analyzed for long-term trends related to climate change (Maine, Ohio, Utah and North Carolina)

Abiotic: water and air temperature, annual precipitation, hydrologic parameters, etc.

Biotic: taxonomic composition, traits

MAJOR CONCERN = DETECTION OF FALSE TRENDS

WHERE TO START…

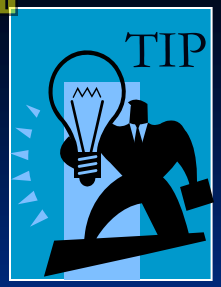
Compile your data into a workable format

We use Ecological Data Application System (EDAS) databases, which are custom database applications that are used with MS Access

Make sure the data has been properly QC'd

A complete and correct master taxa list is required

Note: this may sound very straightforward but it can be a very time-consuming task



Use the online compare function in ITIS to screen for inconsistencies and mistakes, such as misspellings, synonyms and inconsistencies (i.e. group or grp., etc.)

http://www.itis.gov/taxmatch_ftp.html

ITIS Tools

What's New
About ITIS
Data Access
Submit Data
Tools
TRED
Links
Comments
Home

Compare Taxonomy/Nomenclature

Please Note: A prototype ITIS Taxonomic Metadata Tool has been added as an option to enable generation of a component of the FGDC Biological Profile with SGML output. It is currently based on an input file of scientific names only. The prototype will be enhanced in the coming months. Please review the Taxonomic Tool Use Guidelines document for more information.

Step 1 - Upload and Identify File

- In order to perform a taxonomy/nomenclature comparison using the ITIS Online database, the data must first be put into the compare taxonomy/nomenclature import format.
- The next step is to upload your file to the ITIS server:
  Upload File Name: Browse...
  Upload File
- Confirm file name or type in the file name to be matched against ITIS (this entry is case-sensitive, please enter the name exactly as you transferred it) :

Select the character delimiter used to separate the fields within the file: Pipe

Please limit your comparisons to 100 names at a time

to avoid timeout problems.
This function will be replaced with a more efficient version when that version is ready.

Press the Step 2 button to continue: Step 2 .
To reset the file name and delimiter, press the reset button: Reset .

Last Updated: Friday, 19-Dec-2008 10:23:42 MST
http://www.itis.gov/taxmatch_ftp.html

Develop Appropriate Operational Taxonomic Units (OTUs)

OTU = Operational Taxonomic Unit = the taxonomic name or 'unit' being used in the analysis

This may or may not be the identification given by the taxonomist

The intent of OTUs is to exclude ambiguous taxa from analyses and include only distinct/unique taxa

Why the selection of appropriate OTUs is CRITICAL…

In these types of analyses we generally operate under the assumption that differences among sites are due to assemblage differences and not differences in taxonomic resolution

An example of a taxonomic resolution issue: if there are taxa in a sample that are identified to different levels, are the ones ID'd at higher-levels (i.e. genus) different taxa from those within the same genus that are ID'd to the species level, or are they individuals of the same taxon that were unidentifiable (i.e. damaged, immature, etc.)?

SELECTION OF OTUs IS A BALANCING ACT

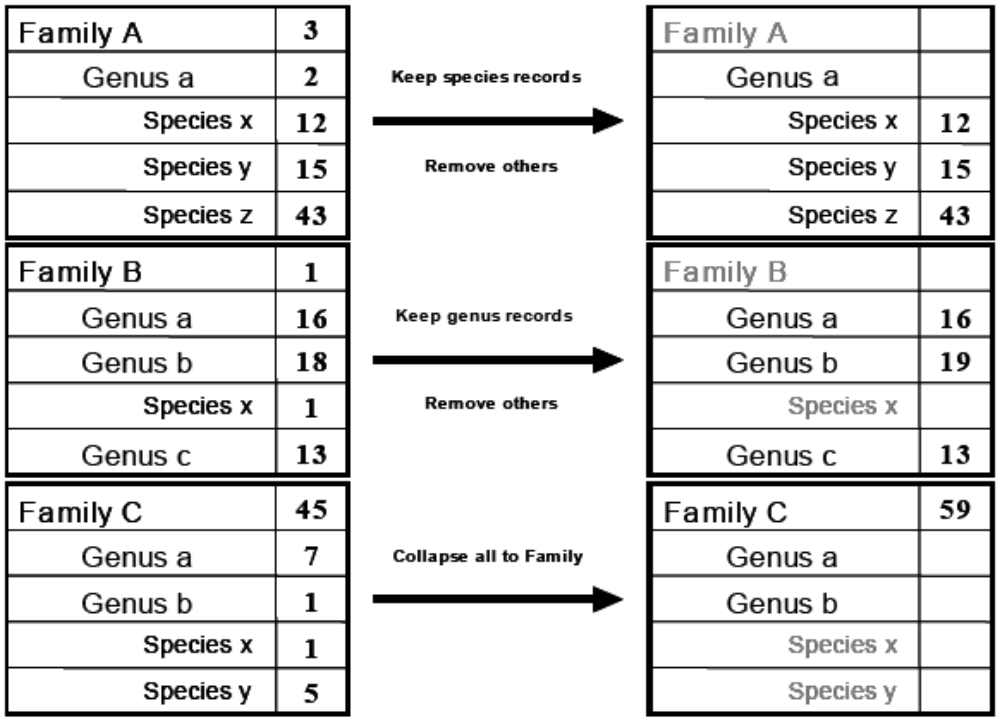


Figure 2.1. This figure describes taxonomic resolution decisions that have to be made for construction of operational taxonomic units. In the first case, most of the information is included in species, but there is no way to know if the genus or family data are the same species or different ones. This could introduce ambiguity into the analysis, so the higher level records are excluded. The second case is similar. Here most of the information is contained in genus records, so the species record is collapsed upwards and the family record excluded. In the last case, most of the information is at the family level. So the records are collapsed upwards. This retains the most records, but loses much of the unique information.

Goal=keep the lowest possible taxa resolution that has minimal effect on # of individuals lost from a sample.

2007. Biological Indicators for Mississippi's Large Rivers: the Pascagoula, Big Black and Tombigbee Rivers. Prepared by Tetra Tech for MS DEQ

OUR APPROACH IN THE CLIMATE CHANGE ANALYSIS

We initially established 3 levels of OTUs:

- Lowest Taxonomic Unit (generally species)
- Genus
- Family

We used the RPMC (remove parent or merge child with parent) methodology outlined in the Cuffney et al. 2007 paper.

We used sample presence/absence data as the basis for deciding whether to merge children into parents or keep children and drop parents.

Note: to automate this and improve efficiency we wrote our own procedure in Access with a mixture of queries and VBA code (Erik Leppo, Tetra Tech)

Next we manually reviewed the list of OTU designations and made final corrections as necessary.

Then we determined which OTU was most appropriate based on the dataset and the type of analysis we were doing. The genus-level turned out to be most appropriate for our purposes.

YOUR STATUS…

Now your data is ready to go.

It has been compiled into a database and QC'd.

You have developed an appropriate OTU

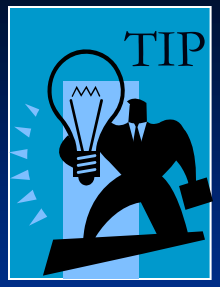
Now what?

SCREEN THE DATA

Screen the data for changes in field and lab protocols

Look for differences or changes in things like:

- collection methods
- sample collection dates
- sample processing/subsampling methods
- how abundance data were recorded
- taxonomists
- taxonomic keys
- selection of sampling locations (i.e. rotational sampling)



AN EFFECTIVE & EFFICIENT SCREENING TECHNIQUE

Non-metric Multidimensional Scaling (NMDS)

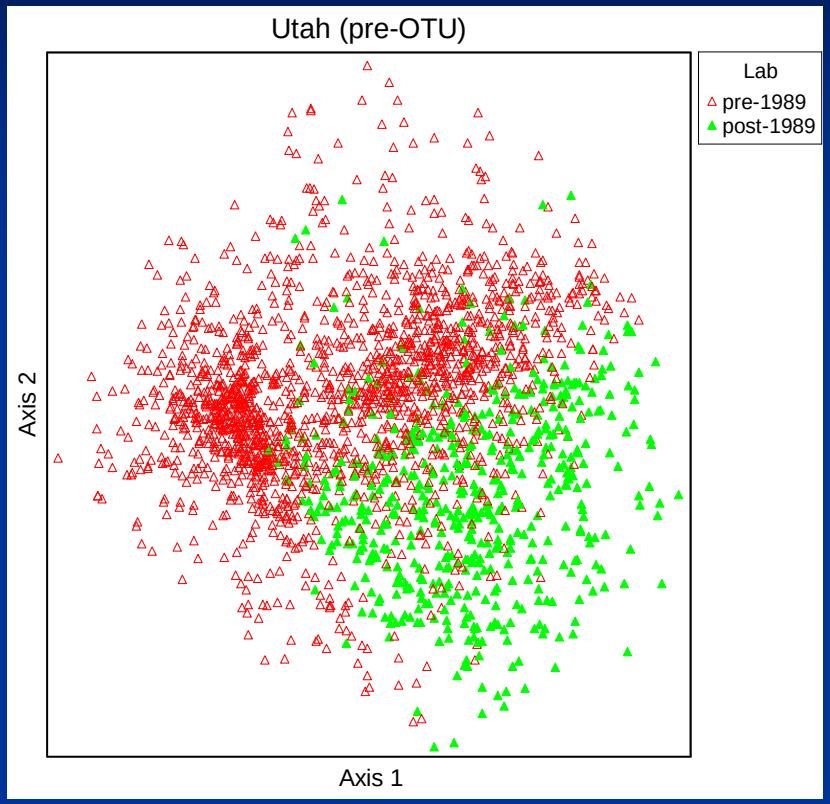
NMDS is an ordination that takes the taxa in the samples and shows in ordination space how closely related the samples and stations are based on their species composition.

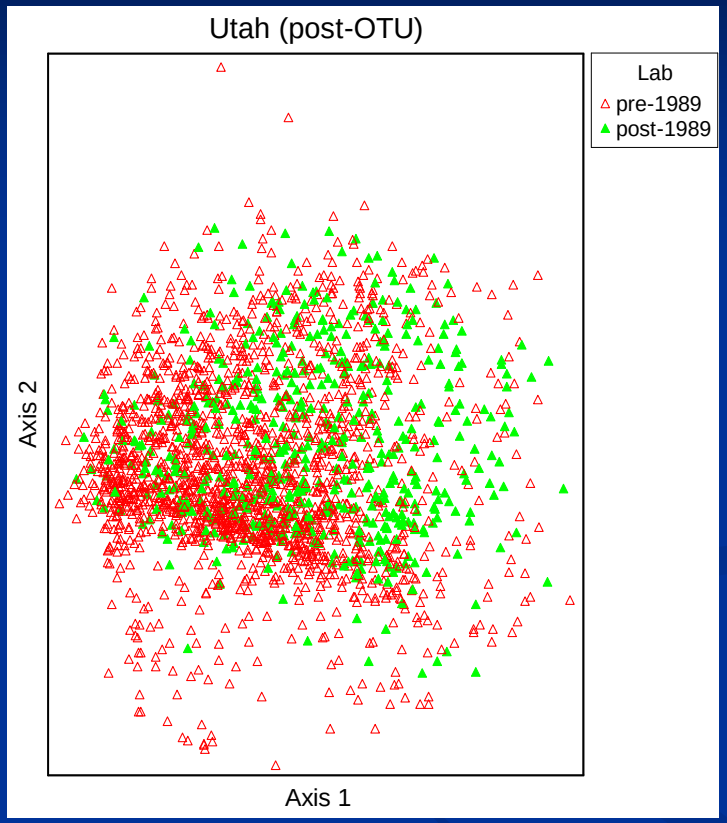
We created pre- and post-OTU NMDS plots to evaluate how effective our selected OTUs were in 'fixing' issues that arose due to changes in field and lab protocols.

In these plots we overlaid different grouping variables (i.e. year, month, collection method, taxonomy lab, ecoregion, watershed, etc.) to look for trends.

The software we used was PCOrd (McCune, B. and M. J. Mefford. 1999. PC-ORD. Multivariate Analysis of Ecological Data. Version 4.41 MjM Software, Gleneden Beach, Oregon, U.S.A.)

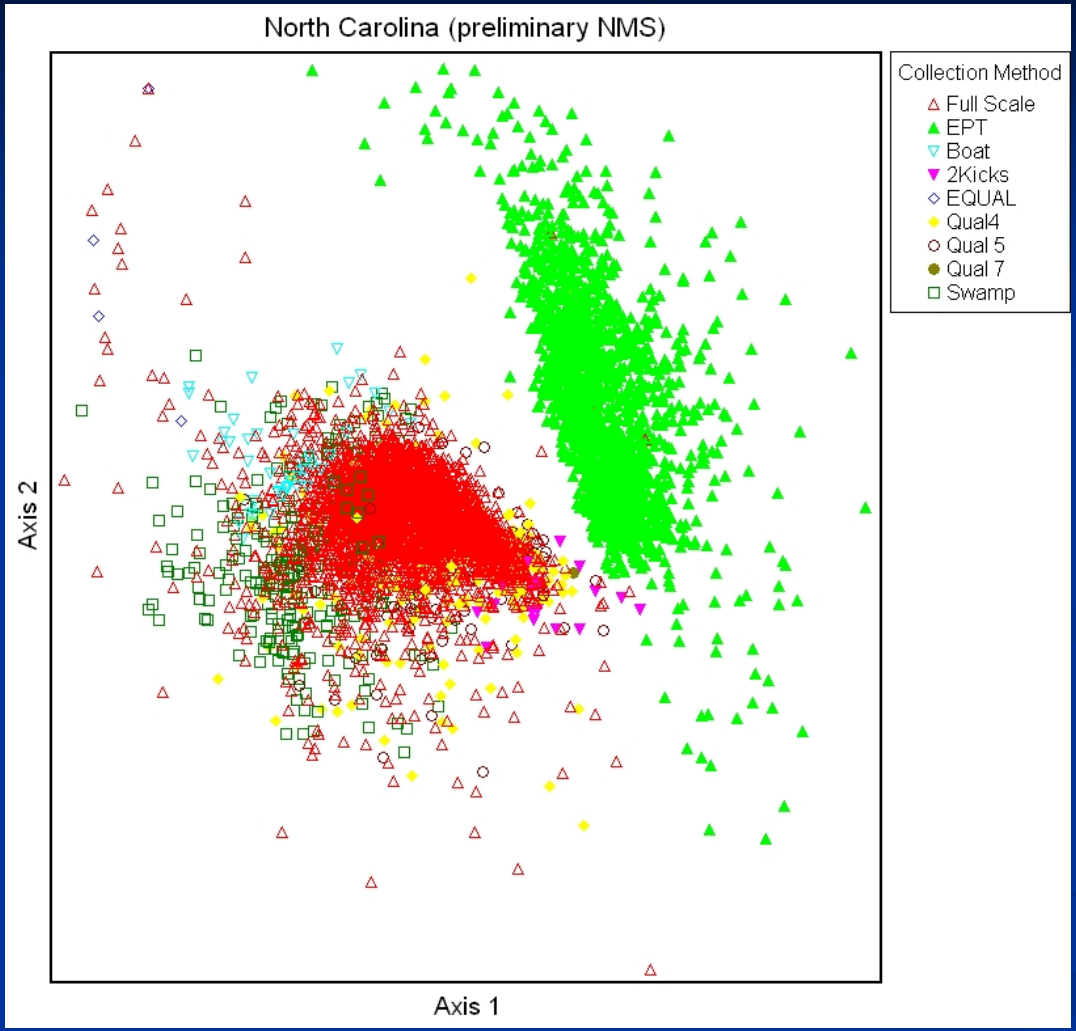
SOME PRE- AND POST-OTU NMDS PLOTS SHOWED DISTINCT PATTERNS





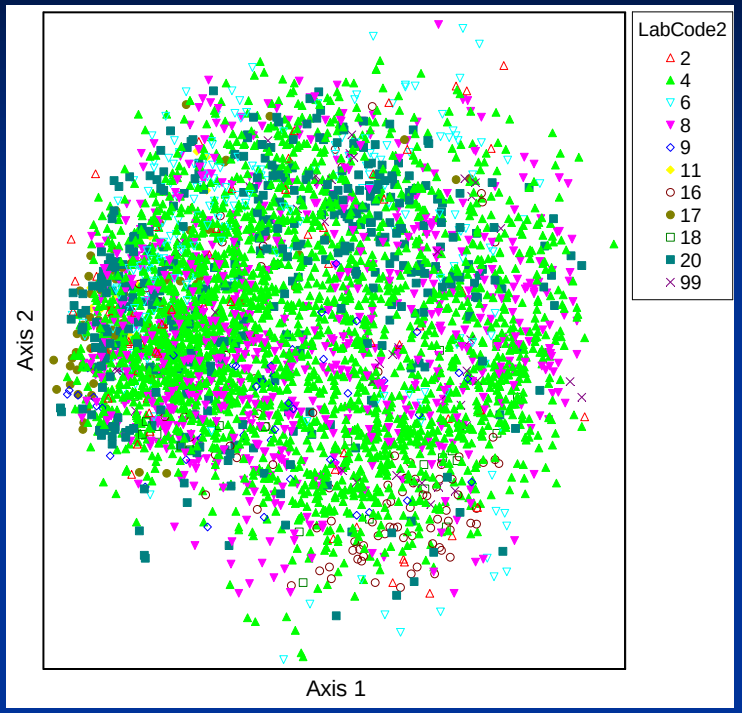
WHAT HAPPENED IN 1989? Some detective work revealed that in Utah they changed taxonomy labs in 1989. We had to make some manual changes or 'fixes' to our OTU to account for changes that resulted from this (i.e. family-level for Chironomidae, adjustments to Ephemerella and Drunella, etc.)

ANOTHER DISTINCT PATTERN…

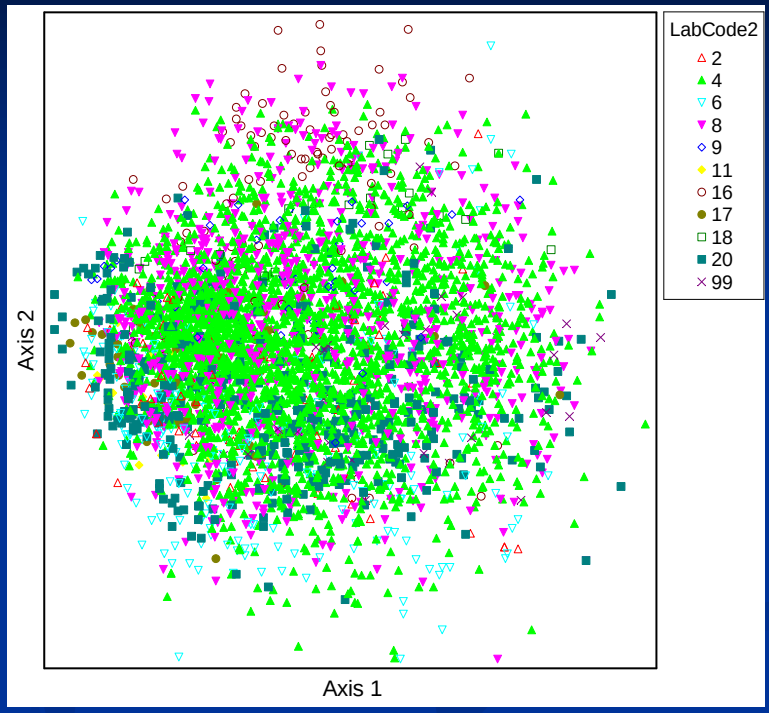


COLLECTION METHOD ISSUE. To fix this we had to limit our analyses in NC to full-scale collection method only.

EXAMPLES OF NON-DISTINCT PATTERNS



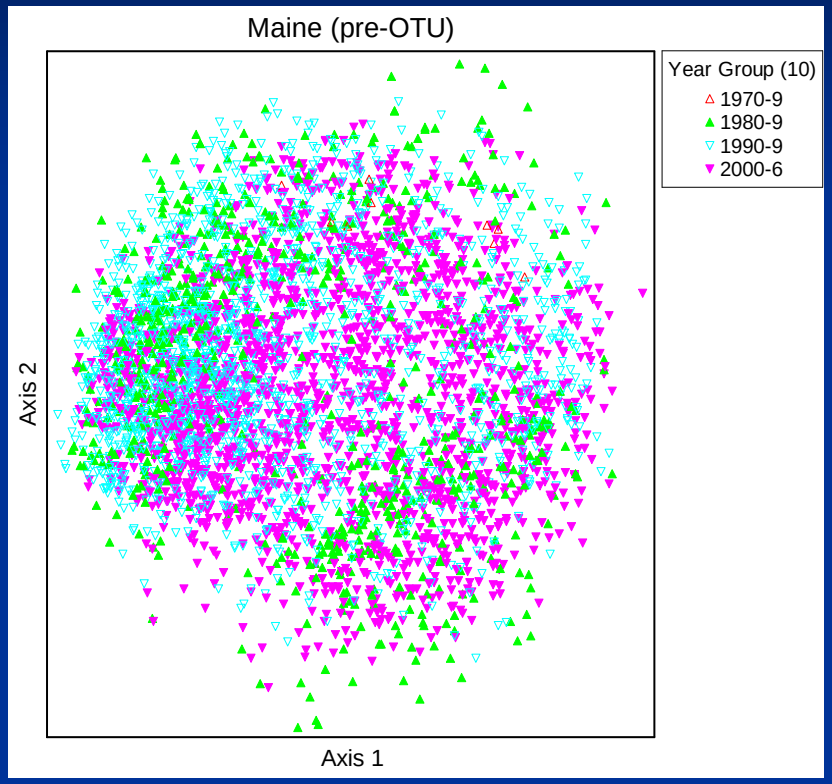
Pre-OTU NMDS plot for Maine

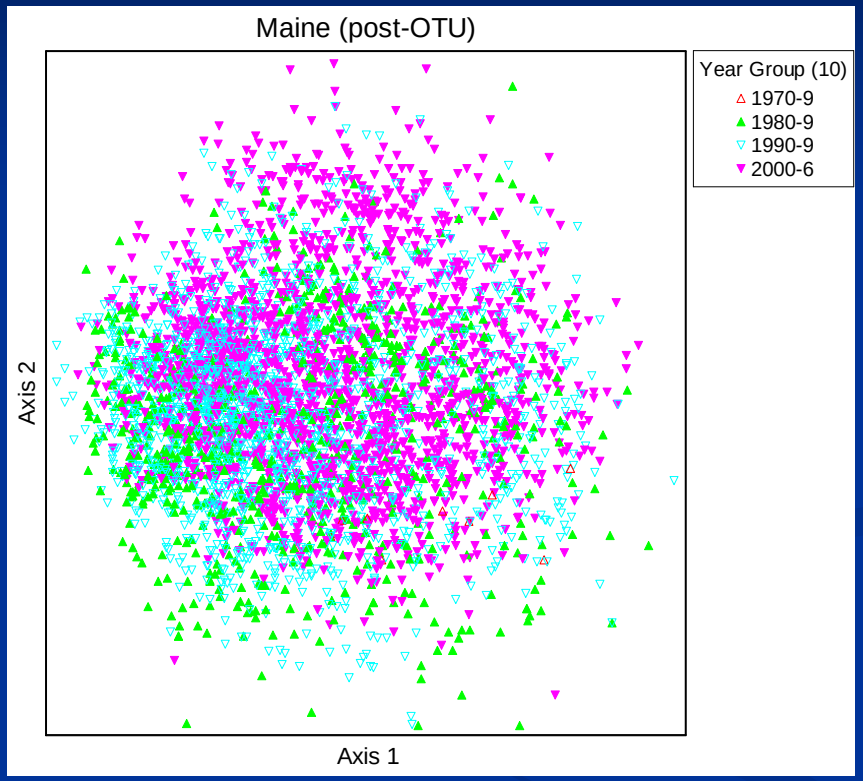


Post-OTU NMDS plot for Maine

MAINE. We limited samples: rock basket and rock cone collection method only; June through November samples only. We had concerns about changes in taxonomy labs (in the 26 years of data that were analyzed, about 16 different taxonomists or labs did varying numbers of IDs). However there were no distinct trends in the plots and any trends that we noticed were inconsistent.

ANOTHER NON-DISTINCT PATTERN…



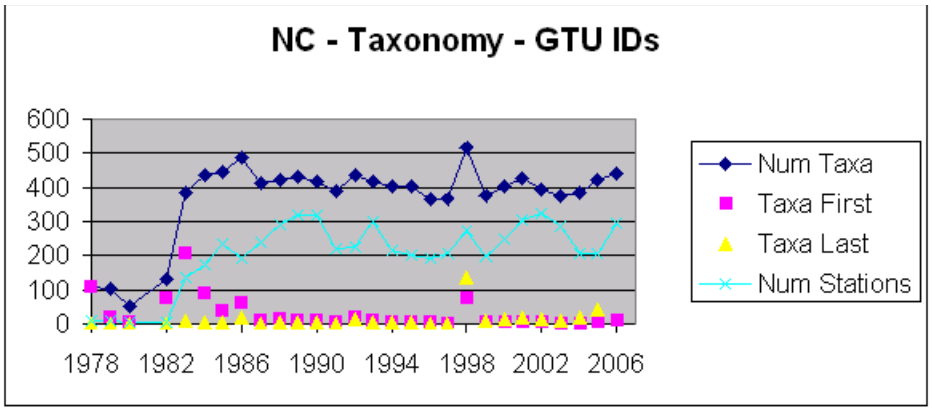


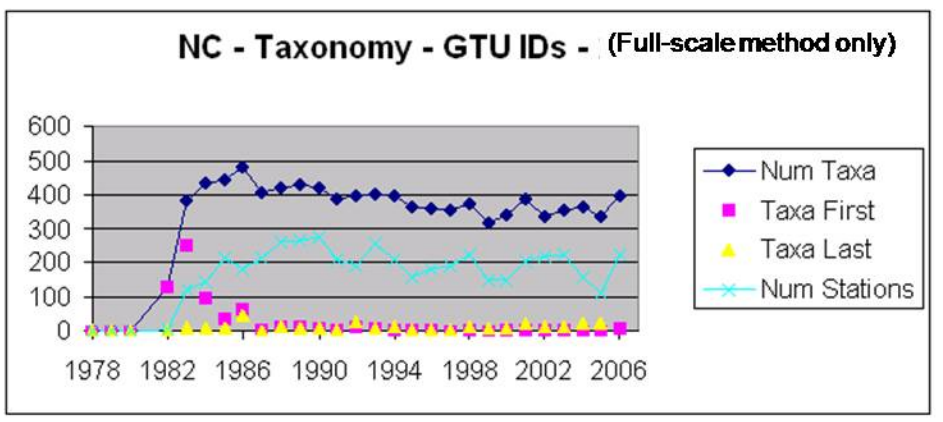
Samples were grouped into 10-year groups. There were no distinct trends in the plots. Any trends that we noticed were inconsistent.

WE ALSO USED SOME OTHER SCREENING TECHNIQUES…

"Num Taxa" = total number of taxa recorded in a particular year
"Taxa First" = number of taxa that appear in the database for the first time in a particular year
"Taxa Last" = number of taxa that appear in the database for the last time in a particular year
"Num Stations" = number of stations sampled in a particular year.

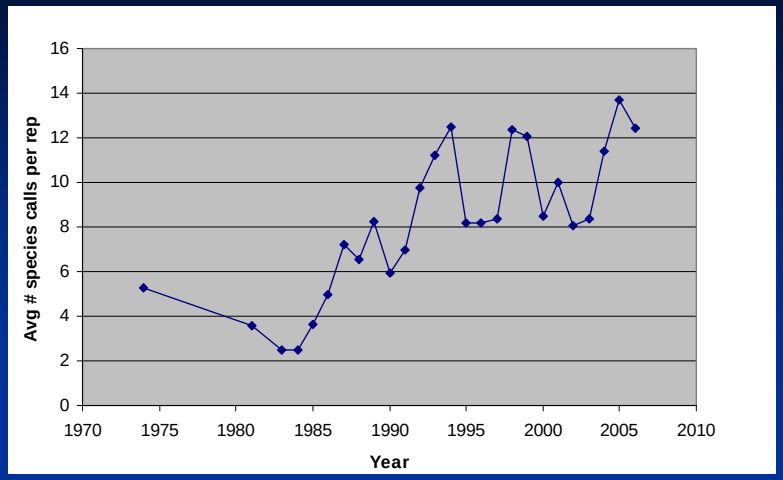
(ALL COLLECTION METHODS)

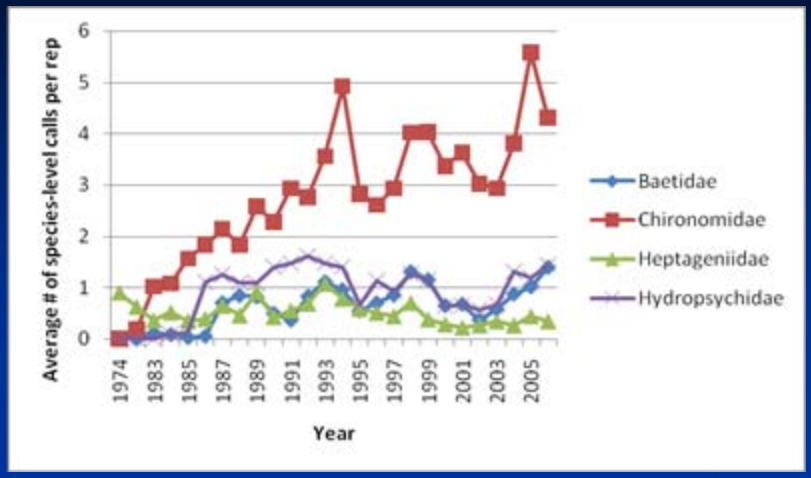




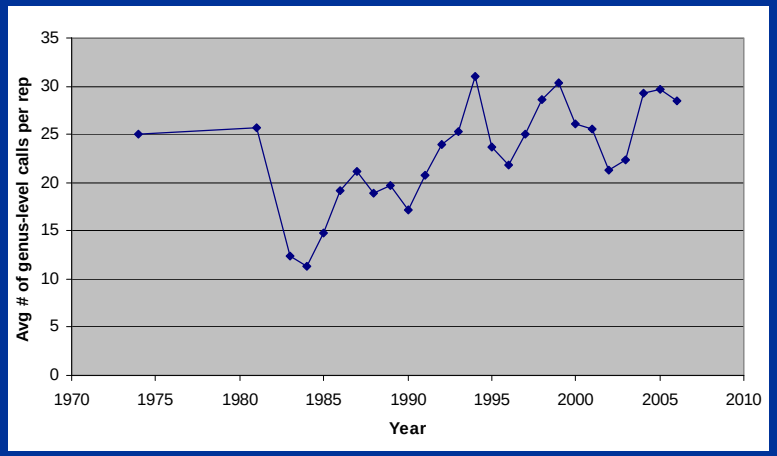
Why the spike in total # of taxa in 1998? Were more stations sampled that year? The number of stations sampled in 1998 was only slightly higher than in previous years, so that probably isn't it. Many of these taxa only occur in the database during 1998. Hmmm…Upon further investigation, we found out that a number of estuarine sites (which have unique taxa) were sampled in 1998. Many of these sites were not sampled prior to 1998 and have not been sampled since. FIX=limit to full-scale method only. We lost 4 years of data, but it was worth the tradeoff.

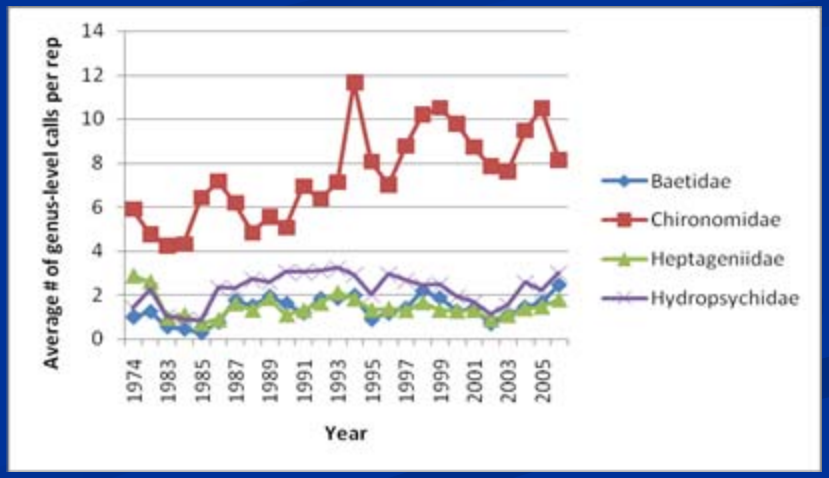
EXAMPLES OF SOME OTHER TYPES OF PLOTS THAT WE LOOKED AT…





Average number of species-level identifications per replicate sample per year in the Maine database (using original data (not adjusted for OTUs)).





Average number of genus-level identifications per replicate sample per year in the Maine database (using original data (not adjusted for OTUs)).

SELECTION OF SITES AND SAMPLES TO USE IN THE ANALYSIS

Some considerations…

SPATIAL – ecoregion, etc., sampling rotations

TEMPORAL – index period, seasonal variation, etc.

DEFINITION OF REFERENCE – different people define it in different ways. Some use biology. Some use things like land use land cover. The bottom line (for our study) was to do our best to limit potential confounding factors.

WE USED SITE-YEAR MATRICES TO FIND REFERENCE SITES WITH LONG-TERM BIOLOGICAL DATA

STATIONID	Yrs	1974	1981	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
56817	23	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
57011	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
56763	11	0	0	1	1	0	0	1	1	1	1	0	1	1	1	0	0	1
56825	11	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1
57008	11	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
57006	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
57065	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
56886	9	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1
56887	9	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1
56819	9	0	0	0	1	1	0	0	1	1	1	0	1	0	0	0	1	0
56821	9	0	0	0	1	1	0	1	1	1	0	0	1	0	0	0	1	0
56987	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
56990	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

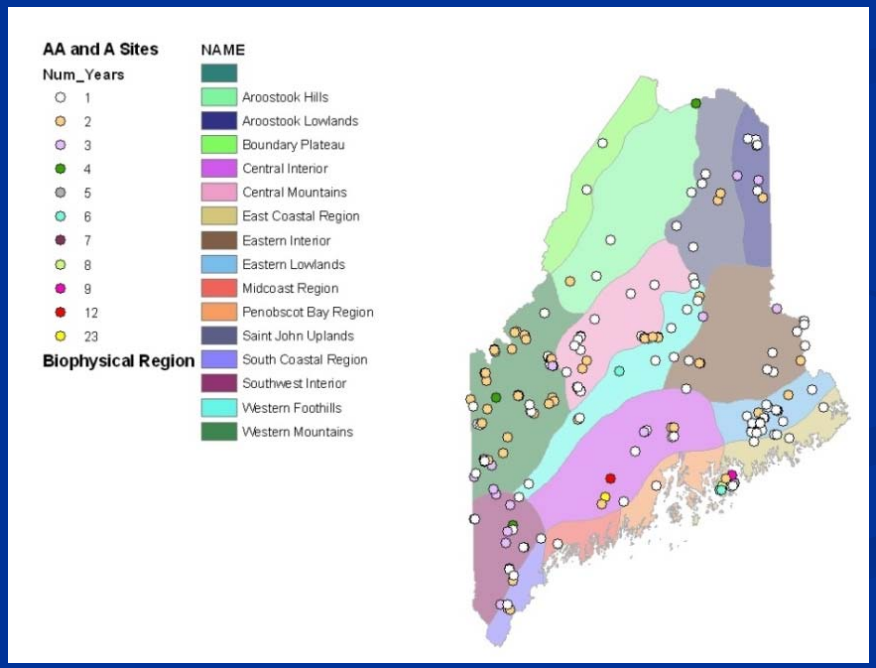
LIMITATIONS

It is tough to find sites that have long-term data

It is tough to find sites with long-term data that haven't been noticeably influenced by other confounding factors

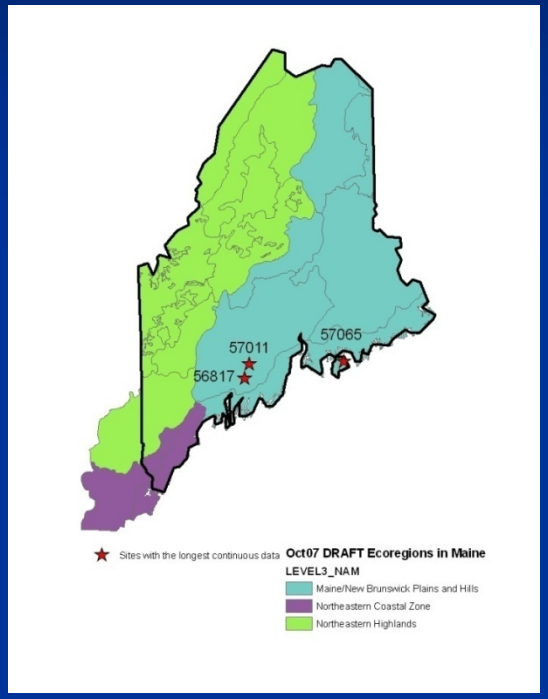
It is tough to draw broad conclusions about long-term trends from data that was collected at a few sites (i.e. how representative are those sites?)

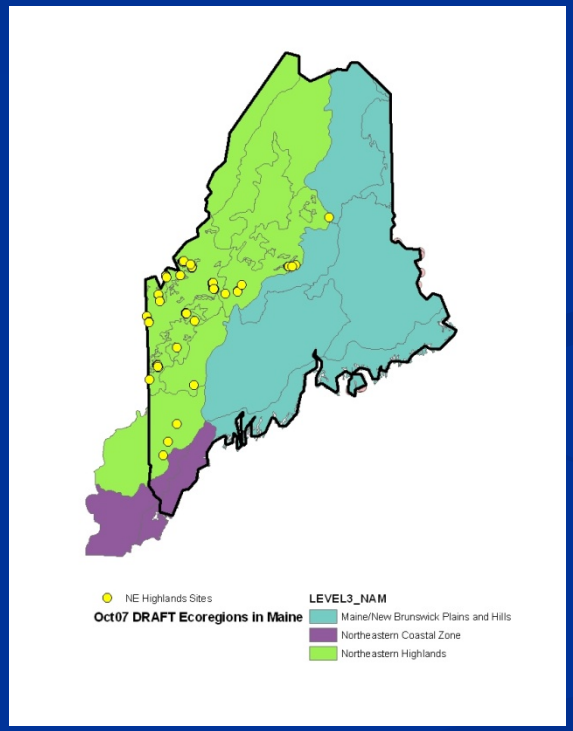
Our knowledge of natural inter-annual variation is limited

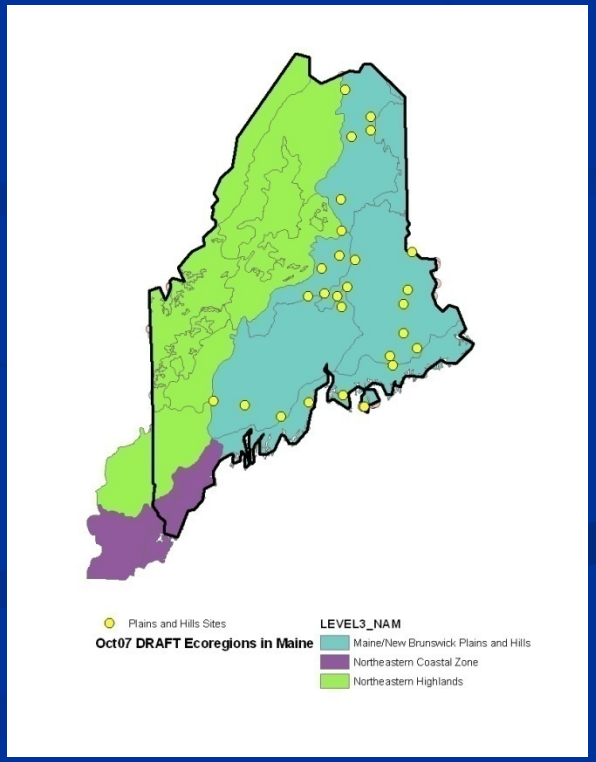


ANOTHER OPTION: GROUP SITES

In an attempt to create longer-term data sets, we combined data from different (carefully selected) sites.

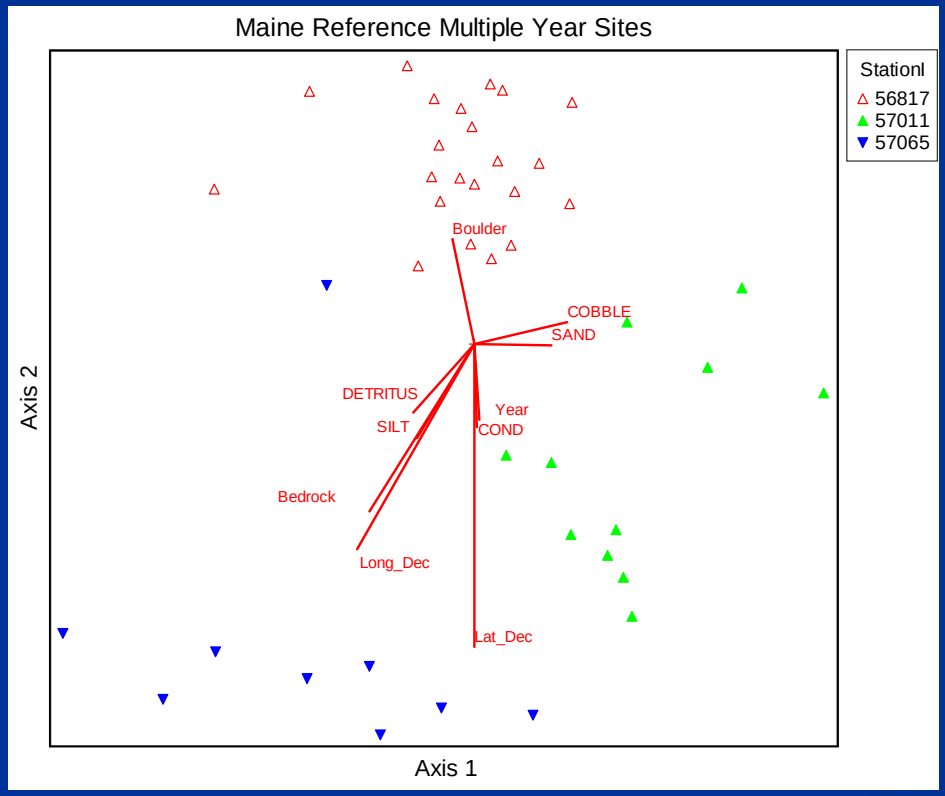


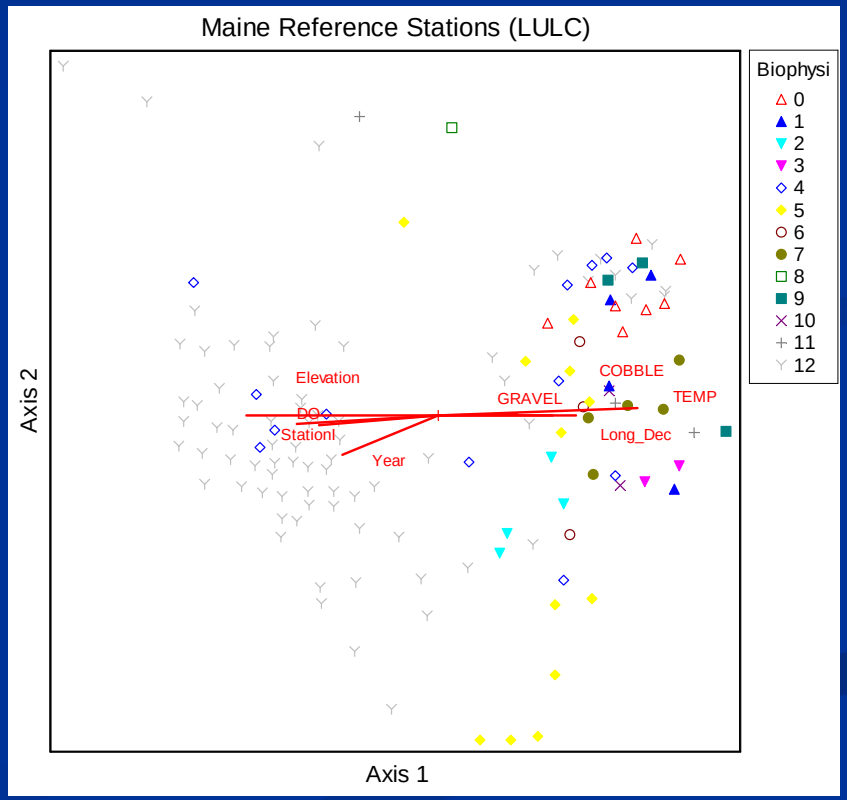




LIMITATIONS

In the site group analyses we have done to date, the sites form distinct groups in our ordinations (based on things like basin, ecoregion, etc.) = not good (it shows that our grouping techniques were not as effective as we had hoped)





RECAP: WHY THIS IS IMPORTANT

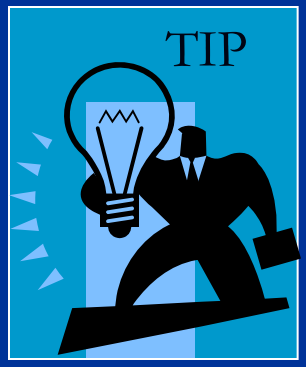
Data preparation and screening are key components of analyses. If you can utilize techniques that improve effectiveness and efficiency, the better off you will be.

Regarding analyses of long-term data, there has been a lot of information collected over the years by a lot of different people for a lot of different reasons. Some of it is scattered in different places and is in different formats. Because of this, some of it tends to fall through the cracks, which is unfortunate because some of it isn't utilized even though it is of high quality and is very informative. Compiling this information into a format that can be readily analyzed is tedious but the types of analyses and the types of information you can pull from it can be incredibly valuable.

WE ENCOURAGE YOU TO TRY LOOKING FOR LONG-TERM TRENDS IN YOUR STATE BIOMONITORING DATA

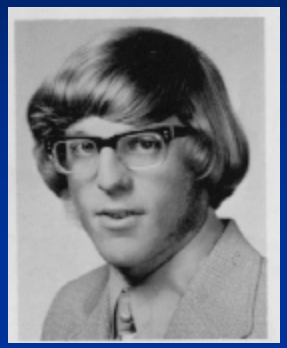
Yes, there will be some limitations (i.e. finding appropriate sites, etc.) but there will be a lot that you will learn despite these limitations. And if there is one thing I have learned in this project, there is lots that we still need to learn!

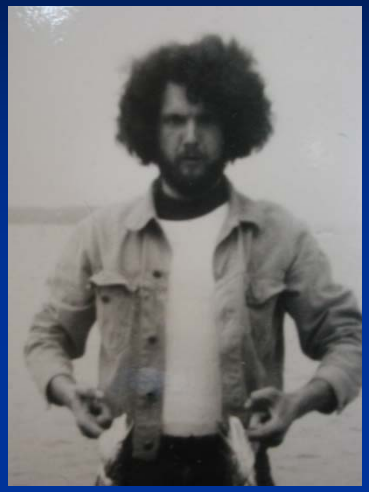
In some states, the same people have been collecting data for many years…



If you are thinking of doing analyses that involve looking at historical data, it would be wise to do them while these employees are still around (and still have good memories). Their long-term local knowledge will be invaluable.

YOU NEVER KNOW…YOU MAY FIND SOME VERY INTERESTING TRENDS…





Some things have changed…

Some things haven't…

ACKNOWLEDGEMENTS

Erik Leppo, Tetra Tech

Anna Hamilton, Tetra Tech

Lei Zheng, Tetra Tech

Mike Paul, Tetra Tech

Britta Bierwagen & Susan Julius
Global Change Research Program
NCEA/ORD

Anonymous photo donors

QUESTIONS? COMMENTS?

Jen Stamp
Tetra Tech
Jen.Stamp@tetratech.com
(802) 229-1059