Taxonomic harmonization may reveal a stronger association between diatom assemblages and total phosphorus in large datasets

Lee, S.S., Bishop, I.A., Spaulding, S.A., Mitchell, R., and Yuan, L.L. 2019. Ecological Indicators

# ABSTRACT

* Diatom data have been collected in large-scale biological assessments in the United States, such as the U.S. Environmental Protection Agency’s National Rivers and Streams Assessment (NRSA). However, the effectiveness of diatoms as indicators may suffer if inconsistent taxon identifications across different analysts obscure the relationships between assemblage composition and environmental variables. To reduce these inconsistencies, we harmonized the 2008-2009 NRSA data from nine analysts by updating names to current synonyms and by statistically identifying taxa with high analyst signal (taxa with more variation in relative abundance explained by the analyst factor, relative to environmental variables). We then screened a subset of samples with QA/QC data and combined taxa with mismatching identifications by the primary and secondary analysts. When these combined “slash groups” did not reduce analyst signal, we elevated taxa to the genus level or omitted taxa in difficult species complexes. We examined the variation explained by analyst in the original and revised datasets. Further, we examined how revising the datasets to reduce analyst signal can reduce inconsistency, thereby uncovering the variation in assemblage composition explained by total phosphorus (TP), an environmental variable of high priority for water managers. To produce a revised dataset with the greatest taxonomic consistency, we ultimately made 124 slash groups, omitted 7 taxa in the small naviculoid (e.g., *Sellaphora atomoides*) species complex, and elevated *Nitzschia, Diploneis,* and *Tryblionella* taxa to the genus level. Relative to the original dataset, the revised dataset had more overlap among samples grouped by analyst in ordination space, less variation explained by the analyst factor, and more than double the variation in assemblage composition explained by TP. Elevating all taxa to the genus level did not eliminate analyst signal completely, and analyst remained the most important predictor for the genera *Sellaphora, Mayamaea,* and *Psammodictyon*, indicating that these taxa present the greatest obstacle to consistent identification in this dataset. Although our process did not completely remove analyst signal, this work provides a method to minimize analyst signal and improve detection of diatom association with TP in large datasets involving multiple analysts. Examination of variation in assemblage data explained by analyst and taxonomic harmonization may be necessary steps for improving data quality and the utility of diatoms as indicators of environmental variables.
* R ver. 3.4.1

# DATA

**0913\_siteinfo.csv**

Site information and physical habitat data from NRSA 2008-2009 samples included in the study for random forest analysis.

**latlong.csv**

Latitude and longitude of NRSA 2008-2009 sites included in the study for random forest analysis.

**1020\_wq.csv**

Transformed and scaled water chemistry data from NRSA 2008-2009 samples included in the study for random forest analysis.

**waterchem\_rawfull.csv**

Raw water chemistry data from NRSA 2008-2009 samples included in the study for PERMANOVA analysis and creating species response curves.

**NRSA8\_diatom.csv**

Taxon occurrence data from NRSA 2008-2009 samples included in the study.

**0913\_taxonomist info.csv**

Analyst information for the NRSA 2008-2009 samples included in the study.

**20171031.1025.taxon.csv**

Taxon names from USGS Biodata Aquatic - Bioassessment Data for the Nation available on the World Wide Web, accessed [31 October 2017], at [Taxon names from USGS Biodata Aquatic - Bioassessment Data for the Nation](https://my.usgs.gov/confluence/display/biodata/Complete+BioData+Taxonomy+Downloads) .

URL https://my.usgs.gov/confluence/display/biodata/Complete+BioData+Taxonomy+Downloads

Used to run function biodata\_check() in the study.

**C.1 name revisions 20180727.csv**

A file with the species lists for each revised dataset (columns). This file is imported into the R project in B.2. Harmonization Script.R. In each column, rows with the same OTU are combined into a single OTU and the number of valves counted from NRSA8\_diatom are summed.

**C.2 QAQC\_counts.csv**

A file with diatom count data from 143 samples counted by a primary analyst and secondary analyst. Used for finding differences in count data of the same slide from a primary and secondary analyst for OTU identified as having analyst as the factor explaining the most variation in relative abundance. Differences in count data used to inform identification of taxa to make slash groups.

# SCRIPTS

**B.2 Harmonization Script.R**

Method for reformatting taxon names in NRSA8\_diatom to species case, running biodata\_check, and applying revised species lists with slash groups in C.1 name revisions 20180727.csv.

* packages: plyr, stringr, reshape2, qdapTools

**B.3 Random Forest Script.R**

Method for running random forest analysis to determine the 5 most important variables explaining variation in relative abundance of each OTU. For OTUs with analyst as the 1st most important variable (= high analyst signal), C.2 QAQC counts were examined to determine OTUs for inclusion in slash groups. After new slash groups were entered into C.1 name revisions 20180727, scripts B.2 and B.3 were rerun to assess changes to the 5 most important variables explaining variation in OTUs with high analyst signal in the previous run of random forest.

* packages: foreach, doParallel, iterators, parallel, randomForest

**B.4 Comparison Analysis Script.R**

Method for reshaping data frames and analysis with non-metric multidimensional scaling ordination plots, exporting metafiles of ordination plots with samples represented as points and analyst groups represented as ellipsoid hulls, analysis of analyst signal in revised datasets using analysis of similarity and PERMANOVA, analysis of variation explained by total phosphorus using PERMANOVA, and making species response curves for total phosphorus.

* packages: vegan, goeveg, mgcv

# FUNCTIONS

* capitalize: formats taxon names to species case, where the first letter of the genus name is capitalized and the rest of the string is made lowercase.
* truncAuthor: removes all text after the specific epithet or variety or form, etc.
* biodata\_check: checks user taxonomy against the BioData Taxonomic System.
* importSpeciesList: imports a species list given a csv translation file and columns containing different name adjustments. The position of the column with the preferred species list is given as colnumber (the column number).
* nameTranslate: takes a wide dataset (samples as rows and species as columns) and a name translation sheet, outputs new wide dataset with collapsed columns, according to translation sheet.
* rfmod: random forest model for detecting OTUs with high analyst signal.
* ssProcess: Function for processing species data (samples as rows and species as columns). Rare species removed if less than 50 occurrences. Samples removed if less than 400 valves counted.
* plotEhulls: Function for plotting ellipsoid hulls around groups from an ordination object. Prints the goodness of fit information for the analyst groups based on group centroids.
* plotPoints: Function for plotting results of an ordination object with analyst groups (n=9) in different colors.
* specresponse\_sl: Modified function “specresponse” from goeveg package to include more plotting parameters.