

Metadata from “Data and scripts: Reduced enumeration effort, but not coarse taxonomic resolution, is sufficient to represent beta diversity patterns of stream benthic diatoms”

Authors: Emanuela Castro, Tadeu Siqueira, Adriano S. Melo, Luis M. Bini, Victor L. Landeiro, Fabiana Schneck

1 Brief description of biological variables

Sampling region: Sampling was carried out in southeast Brazil between latitudes 24°18'S, 48°25'W, within the Intervales State Park (part of the largest preserved Atlantic Forest area in the State of São Paulo, Brazil). We sampled 10 riffles of nine streams (totaling 90 riffles) ranging from second- to fourth-order. Streams were isolated from each other and riffles within each stream were approximately 10 to 50 m apart from each other. Samplings occurred in April 2015.

Diatom's sampling and identification: At each riffle site, we sampled 10 stones of similar size and removed 25 cm² of periphytic material from each stone. The material from the 10 stones was pooled and preserved in 4% formalin. We cleaned the samples using acid and then mounted permanent slides using Naphrax®. We counted 500 valves from each site using a light microscope with 1000× magnification. We identified diatoms mostly as species. We recorded 260 species and 44 genera of benthic diatoms (Castro et al. 2019, <https://doi.org/10.1007/s10750-019-04037-8>)

2. Data files accompanying this metadata

script_enumeration_effort.R: this code generates reduced datasets by randomly subsampling individuals within sites. The probability of species to be sampled is weighted by their abundance within each site. The procedure is repeated *n* times for

each of the enumeration efforts of interest. Next, two procedures are conducted using each reduced dataset:

1. a Principal Correspondence Analysis and a Procrustes analysis to compare the reduced dataset with the complete one (500 valves per site identified at the species level);
2. the calculation of the mean distance from sites (riffles) to their group centroid (stream) followed by the Pearson correlation between the mean distances of each reduced dataset and the complete dataset.

The first procedure allows the comparison of multivariate locations (composition of the metacommunities), whereas the second one allows comparing multivariate dispersions (differences in beta diversity among metacommunities).

species.txt: a matrix with the abundance of each species (columns) sampled in each site (rows). The first and second columns represent stream and site identity, respectively.

genera.txt: a matrix with the abundance of each genus (columns) sampled in each site (rows). The first and second columns represent stream and site identity, respectively.

ind_sp_50.txt; ind_sp_100.txt; ind_sp_200.txt; ind_sp_300.txt; ind_sp_400.txt:

each of these five archives contains the results obtained using the code available at the archive `script_enumeration_effort.txt` for enumeration efforts based on 50, 100, 200, 300, and 400 valves identified to the species level. The code was run on a faster computer and the results were saved to construct the plots afterward.