get real Diatom paper

Jonathan Jupke

April 27, 2021

1 Abstract

2 Introduction

3 Methods

3.1 Preparation of diatom data

We compiled a large data set of diatom samples from rivers in different European countries. Our data set comprises 27509 samples that were taken between 2000 and 2017. Each of the samples was assent to the closest stream segment in the stream network provided by Globevnik (2019). If the closest stream segment was further than 500m removed from the location of the sample we removed the sample as we could no unambiguously assign it to a segment. This step reduced the number of samples to 15275 (56% of all samples). Next, we removed samples with were strongly impacted by humans. To identify non-reference sites, we used the three different diatom indices.

Selection by anthropogenic stressors can cover selection by natural environmental selection processes (selection here is meant in the sense of Vellend (cite)) (**Verdonschot2006?**). In addition, as we are interested in evaluating the capacity of broad river types to delineate spatially stable reference diatom communities, non-reference sites are not essential to our question. Removing non-reference sites further decreased the number of samples to 1318 (5% of all samples).

The data originated from different sources and required adjustments to ensure taxonomic consistency. First we used the Global Biodiversity Information Facility (GBIF) to check whether any of our species were considered synonyms and if so changed their names to

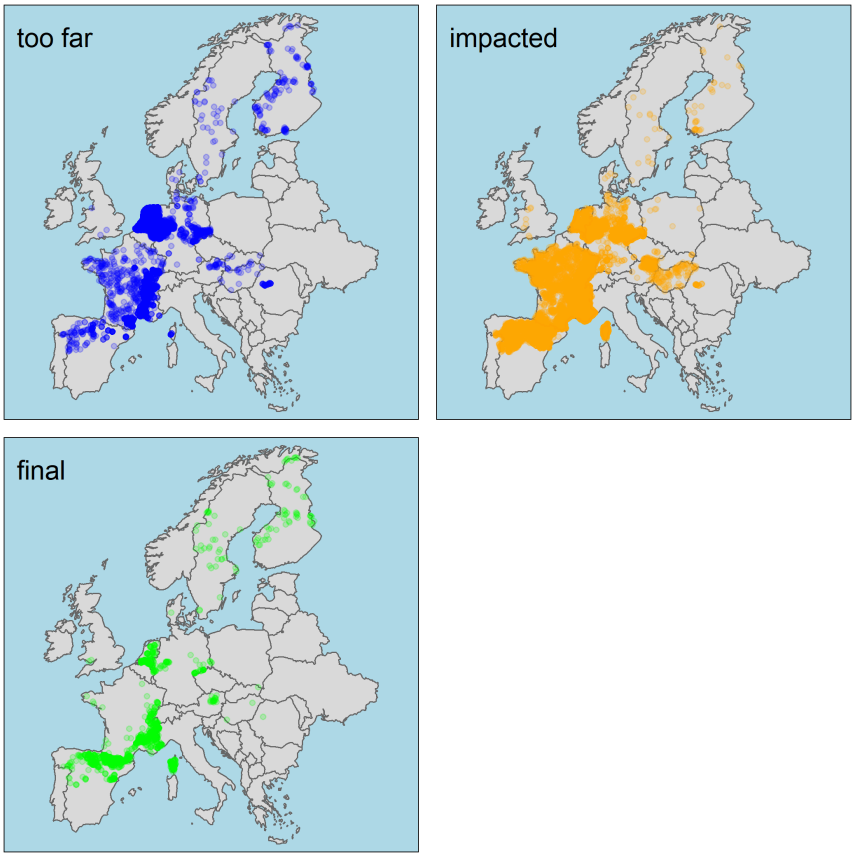

Figure 1: Map of sampling sites. The first map (too far) shows sites that were more than 500 meters removed from the closesed broad river types reach. The second map (impacted) shows sites that were categorized as not in referene condition. The third map (before 2000) shows sites that were sampled before 2000. The fourth map (final) shows that samples that were used for the analyses.

accepted names suggested in GBIF. To integrate the different types of data we transformed all data to presence-absence. Next, we removed all observations that did not belong to one of the following classes: Clitellata, Insecta, Malacostraca, Bivalvia or Gastropoda.

After removing data that did not meet our criteria, we did not have sufficient data to adequately represent all river types. We considered the data for a river type insufficient if it contained to few sites (< 20) or the sites were spatially clustered in a way that was not representative of the river type. Maps of the sampling locations for the individual river types are provided in the Supplementary materials. We were not able to represent: BRT6, 7, 12, 13, 17, 19, 20. Samples from these river types were removed.

3.2 Comparison of typoglogies

To evaluate the two typologies proposed by Lyche Solheim *et al.* (2019) we compared it with the three other typologies: i) the k-means typology of the global river classification framework (GloRiC, Ouellet Dallaire *et al.* 2019); ii) the freshwater ecoregions proposed by Illies (1978); and iii) the biogeographical regions (BGR) proposed by the Europoean Environmental Agency (EEA 2016). These typologies represent two different general kinds of typologies: reach-based typologies (BRT and GloRiC) and regional typologies (Illies and BGR). The former assess single stream reaches. The instances of individual types are not spatially contiguous, can be far apart and are commonly very close to instances of different types. The latter assess large contiguous regions and there is only one instance of each region. Different types are only close at the regional margins. In addition to these five established typologies, we also evaluated BRT_red, a reduced version of BRT20 in which types that were found to have very similar typical assemblages were combined (see Indicator and typical assemblages for more details). As mentioned before, only genus-level data (*data genus*) were used for this analysis. We also ensured that every type in every typology was represented by at least twenty sites. Observations from types that failed to meet this criterion were removed until only adhered types remained. See table **??** for the remaining types per typology and the respective number of samples.

In order to be able to judge the results, we created two additional typologies. As an upper bound of what to expect, we created a classification of biological data using flexible beta clustering (Lance & Williams 1967) with the β parameter set to 0.625. The optimal number of groups was investigated using Average Silhouette Width and determined to be nine. Since this typology is not constrained by environment or space but only represents patterns in the biological data, we expect it to delineate more sharply between biological assemblages than any of the other typologies.

As a lower bound, we created 100 random partitions of the data. For each partition we first drew the number of classes as a random variable from the interval between the lowest number of types in any of the typologies tested (6 in BGR) and the highest number (14 in BRT20 and GloRiC). Then we assigned each observation randomly to one of the groups.

We calculated four cluster quality metrics for each typology: the average silhouette width, the Calinski-Harabasz index, an indicator value score and the classification strength.

The average silhouette width (ASW, Kaufmann & Rousseeuw 1990) is computed as

$$ASW = \frac{1}{n}\Sigma_{i=1}^{n}\frac{a_i - b_i}{max(a_i, b_i)}$$

where a_i is the average dissimilarity of point i to points from its type, b_i is the average dissimilarity of point i to points from the closest other type and n is the number of observations. Positive values indicate, that on average points are more similar to observations from their own type than to those of the most similar one. Therefore high scores imply better typologies. Lengyel & Botta-Dukát (2019) recently proposed a generalized version of the ASW. By using the arithmetic average to compute a_i and b_i, spherical clusters are assumed to be optimal. Using a generalized mean instead, we can flexibly adjust our validity metric to put a stronger emphasis on compactness (a_i) or separation (b_i). The generalized mean of degree p (M^p) is computed as:

$$M^p(\mathbf{x}) = \left(\frac{1}{n}\Sigma_{i=1}^{n}x_i^p\right)^{1/p}$$

This can take the value of common summary statistics such as the minimum ($p = -\infty$), maximum ($p = \infty$) or harmonic mean ($p = -1$). For example, for $p = -\infty$ the silhouette width is the difference between the minimum distance of observation i to any other observation from the same type and the minimum distance from that observation to any observation from the next closest type. This perspective excludes outliers and values separation over compactness. This weighting shifts towards compactness as we increase p. We evaluated the silhouette width for $p \in \{-\infty, -2, -1, 1, 2, \infty\}$. If not further specified, ASW refers to the common average silhouette width (i.e. $p = 1$) in the remainder of the text.

The Calinksi-Harabasz Index (CH, Caliński & Harabasz 1974) is computed as

$$CH = \frac{BGSS}{WGSS} \times \frac{n - k}{k - 1}$$

where $BGSS$ is the squared sum of distances between group centroids and the overall centroid (between group sum-of-squares), $WGSS$ is sum of squared of distances between observations

of one group (within group sum-of-squares), k is the number of clusters and n the number of observations. High values indicate, that variation within types is smaller than between types As the second term controls for the degrees of freedom, it can be understood as an analog to the F-Statistic. The algorithm assumes Euclidean data, but good performance with a similar metrics was shown for binary data in the context of fMRI-scans (Dimitriadou *et al.* 2004).

The indicator value score (IVS) is based on the indicator value (IndVal) proposed by Dufrêne & Legendre (1997). The IndVal itself will be explained in detail below, here we only note that we used 999 permutations to compute p-values and in contrast to the latter application did not control the family-wise error rate. IVS is the fraction of taxa that are statistically significant indicators (at a significance level of 0.01) for some type of a typology. Higher scores indicate a better classification.

Lastly, we computed the classification strength (CS, Van Sickle 1997). Classification strength is the difference between mean within cluster similarity ($\overline{W}$) and mean between cluster ($\overline{B}$) similarity. As such it ranges between 0 ($\overline{W} = \overline{B}$) and 1 ($\overline{B} = 0$), where higher values indicate a stronger classification. A similar and recently applied metric is the partition analysis (Roberts 2019) which is the ratio of $\overline{W}$ and $\overline{B}$.

Based on these four cluster criteria, each typology was assigned a score. We used these scores to evaluate the overall performances of typologies. The typology that performed best in some metric received 6 points, the second 5, the third 4 and so on. Differences smaller than 5% of the range between biological and random partitions were regarded as ties. If two classes were tied, they both received the point for the position reduced by 0.5. For example, if two typologies are tied for the first place, both receive a score of 5.5. A three way tie, was settled by assigning all three classes the middle score. So if three classes are tied and lie at positions 2, 3 and 4, each is assigned 4 points.

3.3 Indicator and typical assemblages

Both indicator and typical assemblages were derived for BRT20. We used the IndVal approach of Dufrêne & Legendre (1997) to identify indicator taxa. For this analysis we used *data genus* which consists of genus level presence-absence data. The IndVal can be understood as the product of the two quantities A and B. For our purposes, A is the relative number of observations of taxon i that are within type j. It was originally described as specificity (Dufrêne & Legendre 1997) but is better understood as concentration (Podani & Csányi 2010) because it is independent of the total number of types. B is the relative frequency with which species i occurs in type j. The maximum score is assigned to a species

which only occurs in one type ($A = 1$) and occurs in all samples of that type ($B = 1$). Here, we used the group-equalized version of the IndVal which accounts for the varying number of samples between types. The statistical significance of the IndVal statistic was assessed with a permutation test that computes IndVal values for random permutations of sites and types and compares the observed IndVal against this empirical distribution. We used $2 * 10^5$ permutations. This procedure ranks tests by their p-values in ascending order. The first p-value is divided by the number of tests (here the number of taxa, M), the second by M-1, the third by M-2 and so on until a p-value exceeds the significance level after the division. We used 0.05 as significance level.

These indicator species provide valuable insight into the communities but miss the ubiquitous generalist species that occur in many types (tramp species *sensu* McGeoch *et al.* (2002)). Even if these taxa are common within a type (high B) they will typically have low concentrations in most types (low A) and hence low and statistically non-significant indicator values. Hence the indicator assemblages do not represent a typical sample, in these sense that these taxa can reasonably expected to occur in samples of the type. We derived such typical assemblages by setting explicit thresholds for B. We used *data all* to derive typical assemblages. These data have different taxonomic levels and we set different thresholds for different taxonomic levels. All species that occurred in more than 25% of samples of a river type (i.e. $B > 0.25$) were considered typical.

After deriving typical assemblages, we evaluated their similarity using the Jaccard similarity. A similarity of 0.5 indicates, that half of the taxa in the combined taxa pool occur in both typical assemblages. If the similarity between two typical assemblages exceeded 0.8, we deemed the river types redundant and combined them. For example, the broad river types BRT2 and BRT3 (medium to large and very small to small siliceous lowland rivers) might be found to be redundant and combined into BRT2_3 (very small to large siliceous lowland rivers). All sites belonging to either of these river types would also be reclassified and the typical assemblages would be derived again. This is repeated until no similarity exceeded 0.8. We did not do the same with indicator assemblages as they are explicitly being optimized for being different from one another (through the equal weighting of A and B). This way we can evaluate whether the ad-hoc combinations of river types used to derive the BRT12 typology are justified by biological homogeneity.

3.4 Software

All computations were conducted in the R Statistical Environment v. 4.0.3 (R Core Team 2020). Data were prepared using data.table 1.14.0 (Dowle & Srinivasan 2021), tidyverse

packages (Wickham *et al.* 2019) and taxize 0.9.98 (**ScottChamberlain2013?**; **Chamberlain2020?**). Geospatial analyses were conducted using sf (Pebesma 2018). Clusters were created and evaluated with fpc (Hennig 2020), indicspecies (Caceres & Legendre 2009), labdsv (Roberts 2019), optpart (**Roberts2020?**). Generalized silhouette widths were computed with the R functions provided in the supplementary materials of Lengyel & Botta-Dukát (2019). Figures and maps were created with ggplot2 (Wickham 2016) and tmap (Tennekes 2018).

4 Results

5 Discussion

References

Caceres, M.D. & Legendre, P. (2009). Associations between species and groups of sites: inindices and statistical inference. *Ecology*, 90, 3566–3574.

Caliński, T. & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3, 1–27.

Dimitriadou, E., Barth, M., Windischberger, C., Hornik, K. & Moser, E. (2004). A quantitative comparison of functional MRI cluster analysis. *Artificial Intelligence in Medicine*, 31, 57–71.

Dowle, M. & Srinivasan, A. (2021). *Data.table: Extension of 'data.frame'*.

Dufrêne, M. & Legendre, P. (1997). Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological monographs*, 67, 345–366.

EEA. (2016). Biogeographical Regions.

Globevnik, L. (2019). Broad typology for rivers and lakes in Europe for large scale analysis.

Hennig, C. (2020). *Fpc: Flexible procedures for clustering*.

Illies, J. (1978). *Limnofauna europaea*. Fischer Stuttgart.

Kaufmann, L. & Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley&Sons.

Lance, G.N. & Williams, W.T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. *The computer journal*, 9, 373–380.

Lengyel, A. & Botta-Dukát, Z. (2019). Silhouette width using generalized mean—A flexible method for assessing clustering efficiency. *Ecology and Evolution*, 9, 13231–13243.

Lyche Solheim, A., Austnes, K., Globevnik, L., Kristensen, P., Moe, J., Persson, J., *et al.* (2019). A new broad typology for rivers and lakes in Europe: Development and application for large-scale environmental assessments. *Science of the Total Environment*, 697, 134043.

McGeoch, M.A., Van Rensburg, B.J. & Botes, A. (2002). The verification and application of bioindicators: A case study of dung beetles in a savanna ecosystem. *Journal of Applied Ecology*, 39, 661–672.

Ouellet Dallaire, C., Lehner, B., Sayre, R. & Thieme, M. (2019). A multidisciplinary framework to derive global river reach classifications at high spatial resolution. *Environmental Research Letters*, 14, 024003.

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10, 439–446.

Podani, J. & Csányi, B. (2010). Detecting indicator species: Some extensions of the IndVal measure. *Ecological Indicators*, 10, 1119–1124.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roberts, D.W. (2019). *Labdsv: Ordination and multivariate analysis for ecology*.

Tennekes, M. (2018). tmap: Thematic maps in R. *Journal of Statistical Software*, 84, 1–39.

Van Sickle, J. (1997). Using Mean Similarity Dendrograms to Evaluate Classifications. *Journal of Agricultural, Biological and Environmental Statistics*, 2, 370–388.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., *et al.* (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4, 1686.