# Derivation of typical diatom assemblages

21.08.20

Jonathan Jupke

## How were the diatom data sets harmonized?

International diatom occurrence data sets require extensive harmonization because of the taxonomic resolution differing between data sets, different working groups using different nomenclatures, identification errors, and ongoing changes to the accepted nomenclature (Kahlert et al., 2020). Harmonization can reduce overall taxonomic resolution but can improve the detection of large-scale spatio-temporal patterns (Lee et al., 2019). We compared all our data sets against a series of databases that contain accepted names, synonyms with links to the respective accepted names and suggestions for grouping contentious taxa in larger complexes. If a taxon name was found in one of the databases the name was accepted, changed into the accepted name in case it was a synonym, or grouped into the respective complex. Once a taxon was found in a database, it would not be included in queries of subsequent databases. However, if the accepted name differed from the original one, the accepted name would be queried through all previous databases again. The results were also controlled visually for consistency. The following databases were used in the same order:

## Table S2 from (Kahlert et al., 2020)

1. The taxon list associated with the OMNIDA software (Lecointe, 2017)
2. (Mauch et al., 2017)
3. (Rimet et al., 2019)
4. algaebase.org
5. global biodiversity information platform (gbif.org)

## What is the optimal taxonomical level?

To identify the optimal level taxonomical level, we employed the same approach as for the invertebrates. We used a cutoff of 75% instead of 85%, because *Gomphonema*, which is the fourth most common genus in our data set, had 81.43% observations at the species level. In the diatom data, the taxonomical resolution was higher than in the macroinvertebrate data. The lowest resolution in the final data set is the genus level.

## Can we represent the stream types with those samples?

Whether a stream type is sampled extensively enough in our data to derive meaningful Typical assemblage (TA) was determined visually. Maps with all streams that belong to one stream type and all associated sampling sites are uploaded to the [Get Real Drive](https://drive.google.com/drive/folders/1CDXy6Zkwhi_7esAAFYU9rYCBTqtaPr3Q). The representativeness was graded in a three-tier system: good, medium, bad. Further analyses were conducted for all stream types with medium representation, as none were graded as good. More information on the river types is available in the [accompanying publication](https://www.sciencedirect.com/science/article/pii/S0048969719340203). In general, we have fewer sampling sites for the diatoms than for the macroinvertebrates which entails that the representation of stream types is generally worse. Additionally, sites from France often outnumber all others due to the large amount of data we have from France.

Good [n = 0]:

Medium [n = 14]: 1, 2, 3, 4, 5, 6, 8, 9, 12, 14, 16, 17, 18, 19

Bad [n = 6]: 7, 10, 11, 13, 15, 20

# What is a typical assemblage?

As before TAs were derived based on a rule that considered

1. The probability of site *x* belonging to stream type *z* given species *y* is present (a measure of specificity, henceforth **A**)
2. The probability of species *y* being present given that site *x* belongs to stream type *z* (a measure of commonness, henceforth **B**)
3. The Species Indicator Value

The Species Indicator Value (Cáceres & Legendre, 2009; Dufrêne & Legendre, 1997) is the weighted product of **A** and **B** (see Eqn. below)

Here Np is the number of sites that belong to stream type p and np the number of occurrences of the focal species in sites of type p. K is the number of stream types. **A** is weighted by the total number of occurrences to account for unequal sample sizes. The statistical significance of the Indicator Value can be assessed with permutation-based pseudo-*p*-values, which we did with 999 permutations.

 Here, we are not interested in indicator species for each community, but TAs. Hence, simply continuing with those species that have a pseudo-*p-value* below some significance level would not serve our purpose. A species that occurs at each site, across all stream types, highlights the difference: while it would not be indicative of any stream type (low specificity) it should be part of each TA. Hence, we need additional criteria to derive the TAs which can be based on **A**, **B**, and the pseudo*-p-value* of the indicator value. We used the following rules:  Species where considered typical if **B** > 0.4 or **B** > 0.3 and p < 0.05 or **A** > 0.7. Genera where considered typical if **B** > 0.8 or **B** > 0.6 and p < 0.05 or **A** > 0.95.

Note that there was no systematic optimization used to define these thresholds. Such procedures would require optimization criteria, but we are not aware of a criterion that would work in this context. We acknowledged that typical assemblages could be (i) very similar in composition or (ii) harbor strongly differing numbers of taxa so these indicators are not useful to us. It would be possible to try a cross-validation-type approach where each taxon is scored based on the number of random-site-subsets it is included in but such an approach would also entail making essentially arbitrary numerical assumptions. We think the use of subjectively defined thresholds is justified, as long as they are clearly and openly communicated, to be what we define as “typical assemblages”.

However, we conducted a sensitivity analysis to see how much varying the parameters of the rules would alter the results. We altered A and B threshold values. The B threshold for taxa with p 0.05 was 25% below that used for taxa regardless of the *p*-value. Henceforth B refers to the latter B threshold from which the former can be deduced. We varied B in ten steps between 0.10 and 0.75 and A in ten steps between 0.5 and 1.0. Species richness and uniqueness scores of each TA were computed for all 100 combinations of these parameters. Please note that results are only shown and discussed for the non-redundant TAs (see the next section). Species richness decreased with increasing A and B threshold (Figure 1 and Figure 2), while the uniqueness scores increased with B thresholds but decreased with A thresholds (Figure 3 and Figure 4). The rate of change in species richness along gradients in A and B threshold differed markedly between TAs but seemed to be correlated with overall species richness, i.e. more species-rich TA lost species more quickly than less species-rich ones. The TAs of RT03 and RT16 serve as examples at both extremes of our data set. Uniqueness scores decreased noticeably with very high A thresholds (> 0.9), indicating that taxa that are specific to certain river types are an important driver of TA differentiation.

## Redundancy in typical assemblages

We assessed to which degree the different TAs overlap (Table 1). The degree of overlap is the percentage of taxa in a TA that is also present in the most similar (largest overlap) TA.

Table 1: Redundancy between TAs

|  |  |  |
| --- | --- | --- |
| River type | Most similar river type(s) | Degree of overlap |
| RT01: Very Large Rivers | RT02 | 59.5% |
| RT02: Lowland, siliceous, medium-large | RT01 | 46.8% |
| RT03: Lowland, siliceous, very small-small | RT02 | 43.6% |
| RT04: Lowland, calcareous or mixed, medium-large | RT02 | 70.4% |
| RT05: Lowland, calcareous or mixed, very small-small | RT01 + RT02 | 62.1% |
| RT06: Lowland, organic and siliceous, very small-large | RT12 | 33.3% |
| RT08: Mid-altitude, siliceous, medium-large | RT02 | 76.0% |
| RT09: Mid-altitude, siliceous, very small-small | RT05 + RT08 | 48.3% |
| RT12: Mid-altitude, organic and siliceous, very small-large | RT06 | 36.4% |
| RT16: Glacial rivers (all Europe) | RT 01+08+17+18 | 60.0% |
| RT17: Mediterranean, lowland, medium-Large, perennial | RT01 | 73.9% |
| RT18: Mediterranean, mid-altitude, medium-large | RT01 | 78.9% |
| RT19: Mediterranean, very small-small, perennial | RT17 | 66.7% |

Again, choosing a threshold above which we consider two assemblages to be redundant is somewhat arbitrary. We proceeded with 75% but are open to other suggestions. This threshold leads to two redundant assemblages RT08 and RT18. We combined observations from RT02 and RT08 in RT02+08 and observations from RT01 and RT18 in RT01+18. RT02+08 represents siliceous rivers of medium to large size and low to medium altitude while RT01+18 is representing Mediterranean and Western European medium to very large rivers.

The new TAs resulted in overall lower degrees of overlap, none of which exceed the 75% threshold. The highest overlap was between RT17 and RT02\_08 with 62.5%.

## What do the typical assemblages look like?

In all TAs, species is the prevalent taxonomic level (see Figure 5). Some assemblages consist entirely of species (i.e. RT04, 06, 09, 16, 17, and 19). The mean number of species per TA is 30 and the mean number of genera 0.5. RT03 has the most taxa rich TA with 49 taxa and RT19 the most taxa poor with 18.

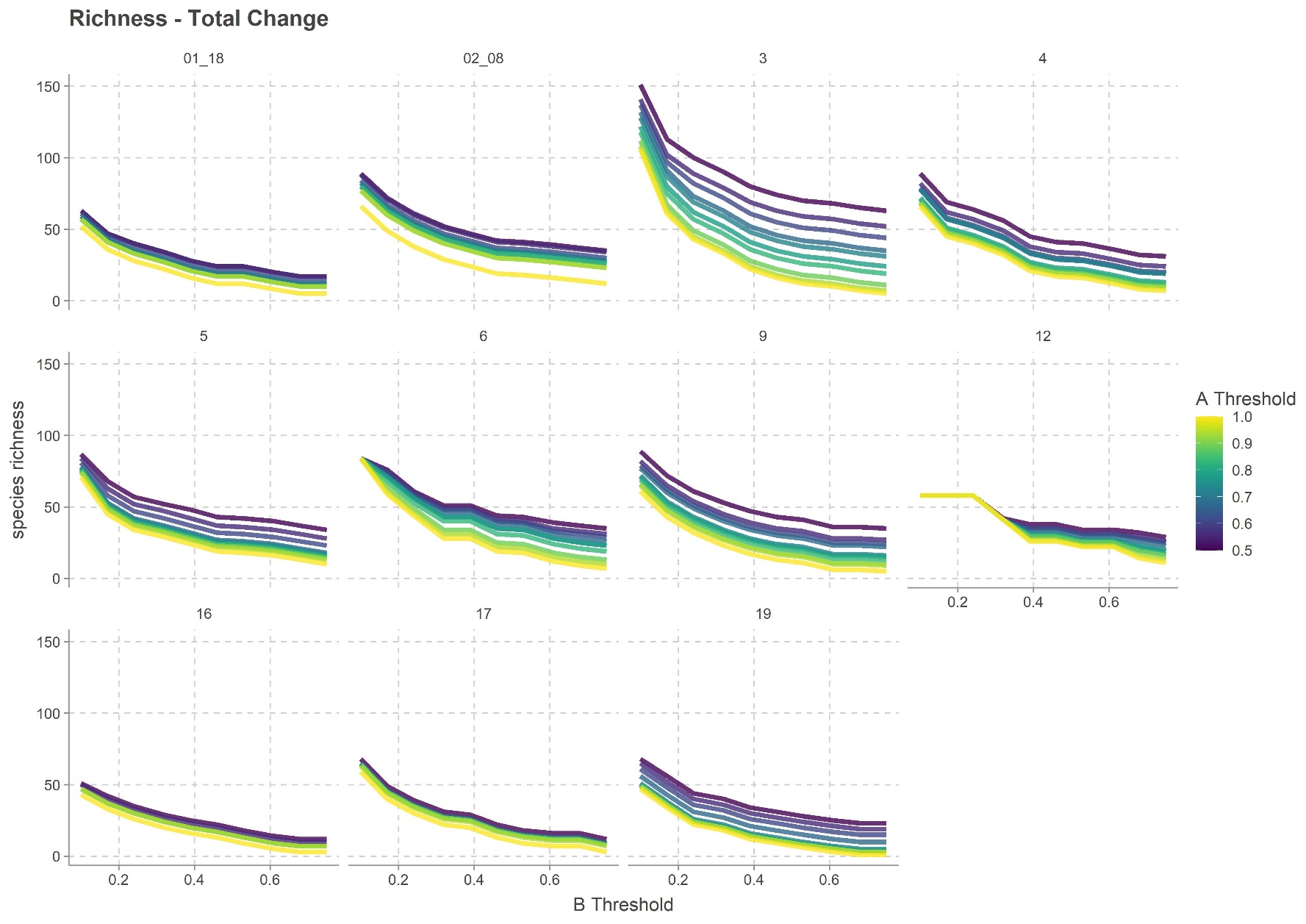


Figure 1: Changes in species richness along a changing B threshold. Line color indicates the employed A threshold.

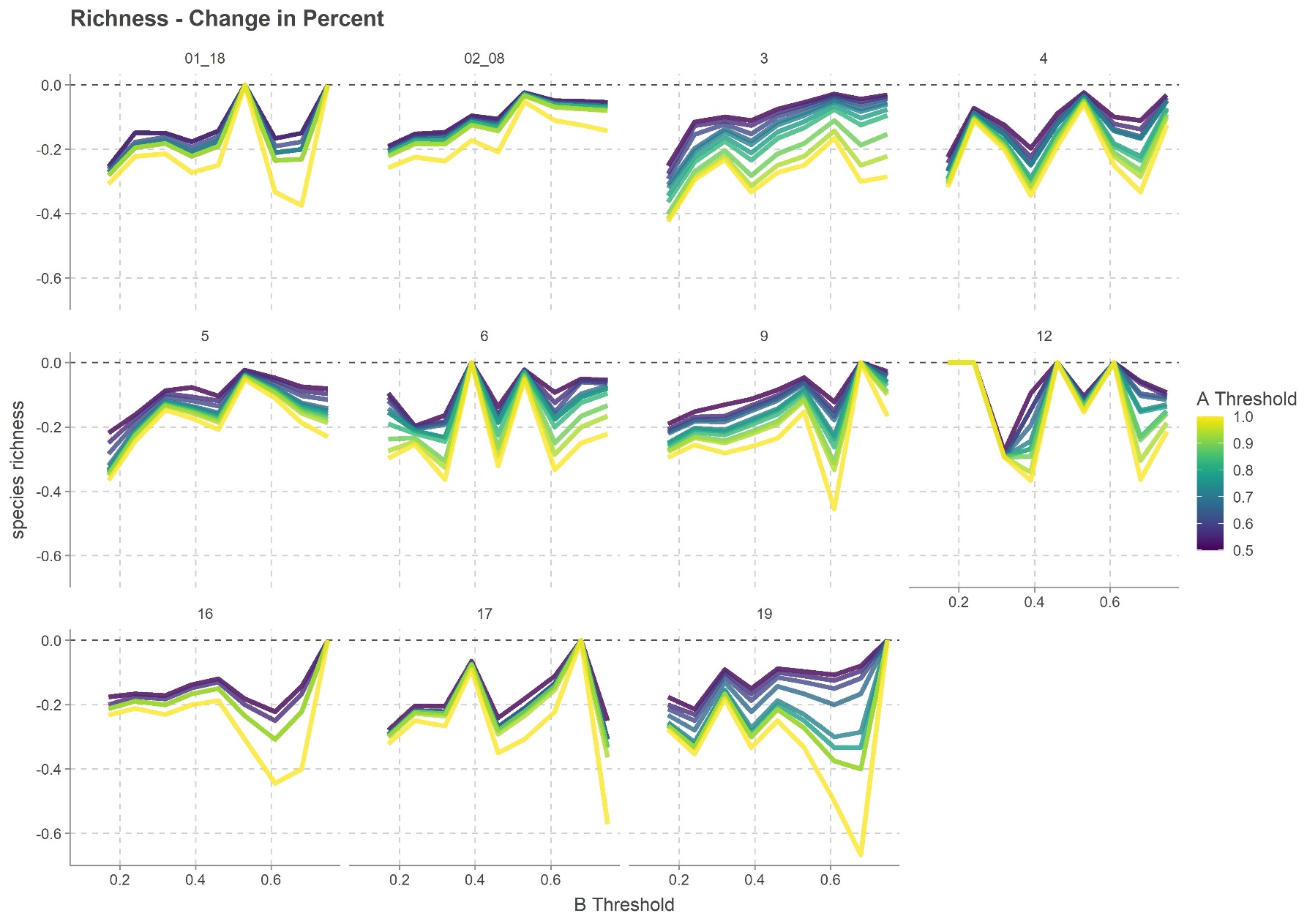


Figure 2: Changes in species richness along a changing B threshold expressed in percent. Line color indicates the employed A threshold. The dashed line marks the change rate of 0%.

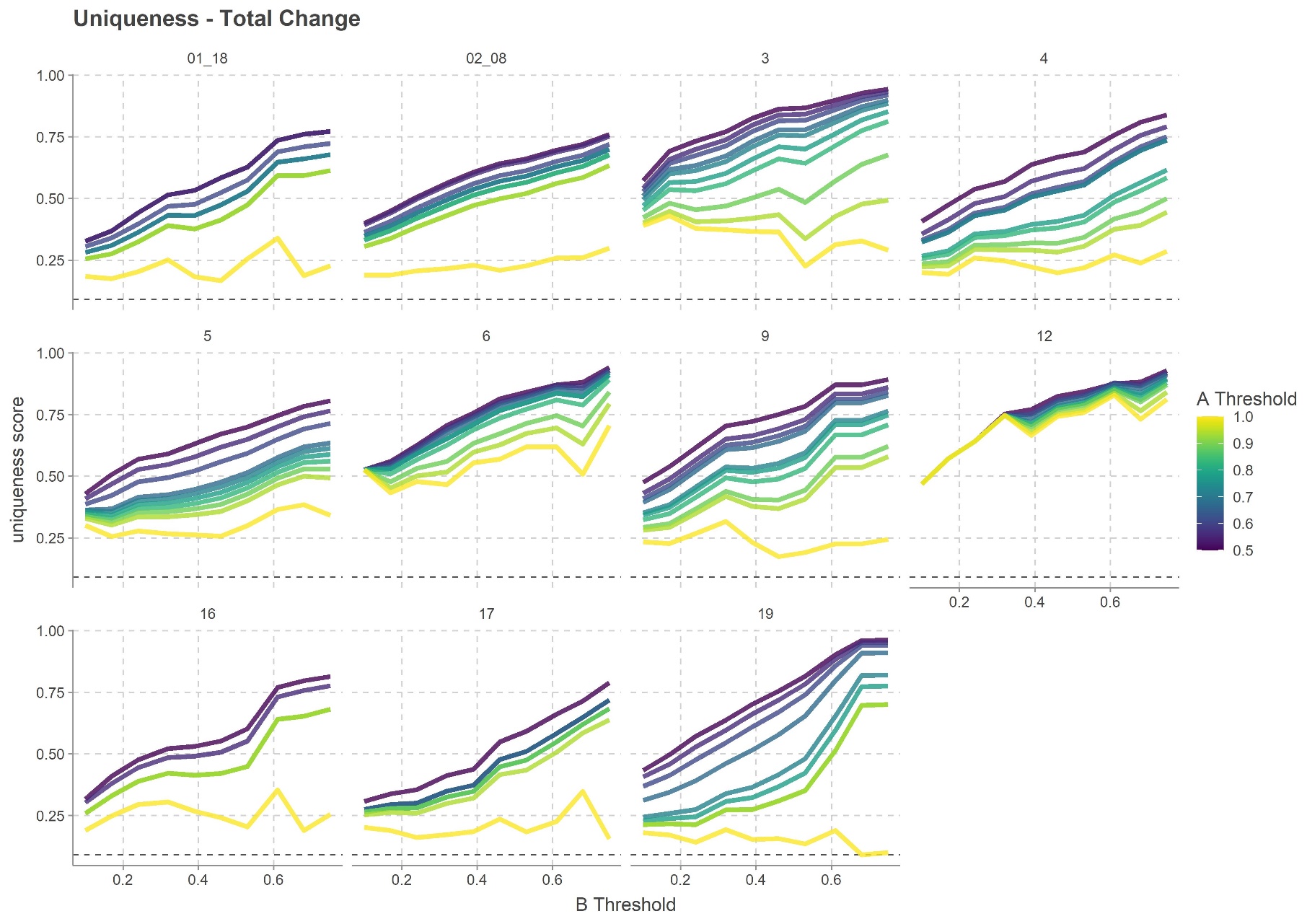


Figure 3: Changes in uniqueness scores along a changing B threshold. Line color indicates the employed A threshold. The dashed line shows the minimum possible uniqueness score of 1/11 = 0.91.

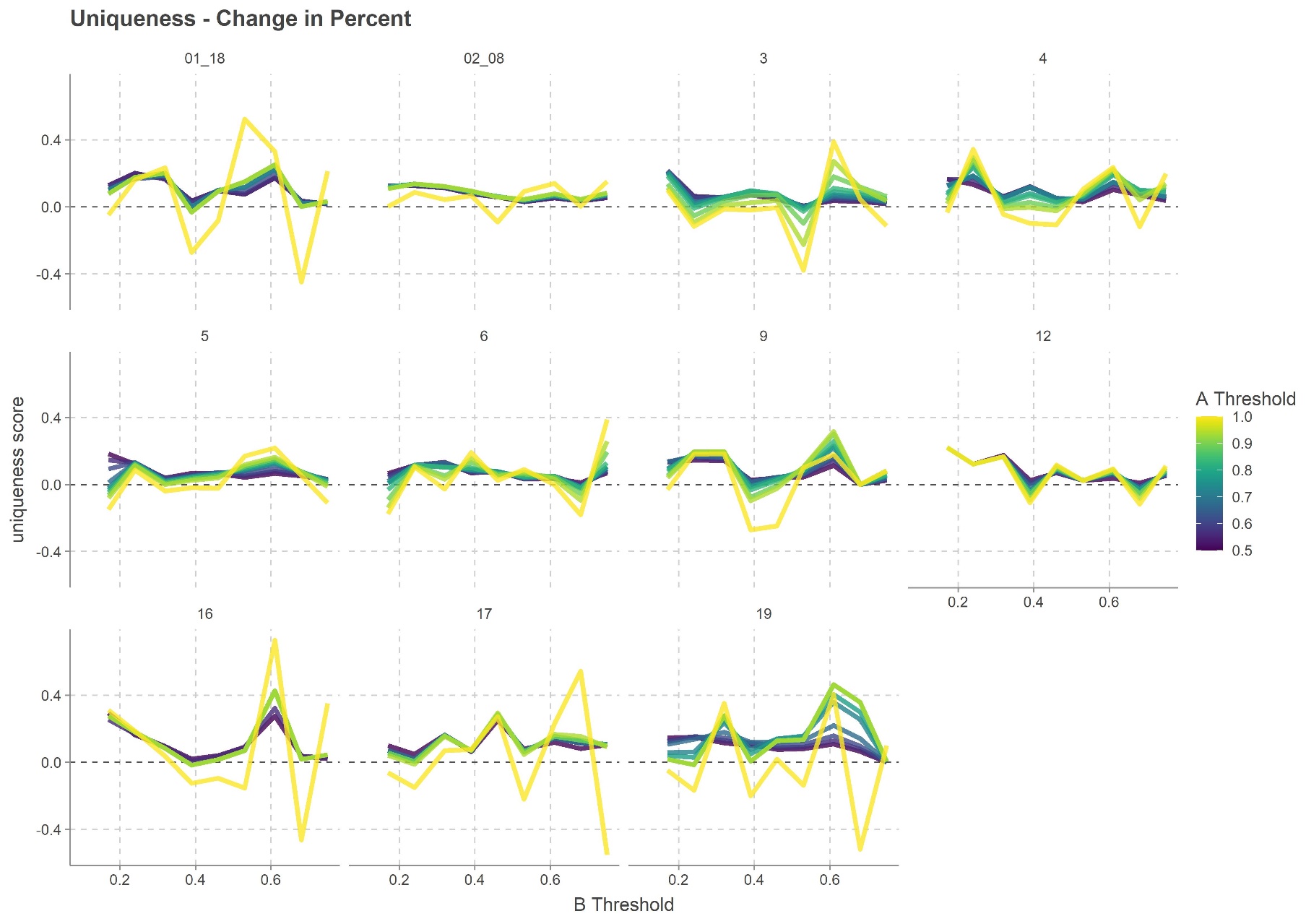


Figure 4: Changes in uniqueness scores along a changing B threshold expressed in percent. Line color indicates the employed A threshold. The dashed line marks the change rate of 0%.

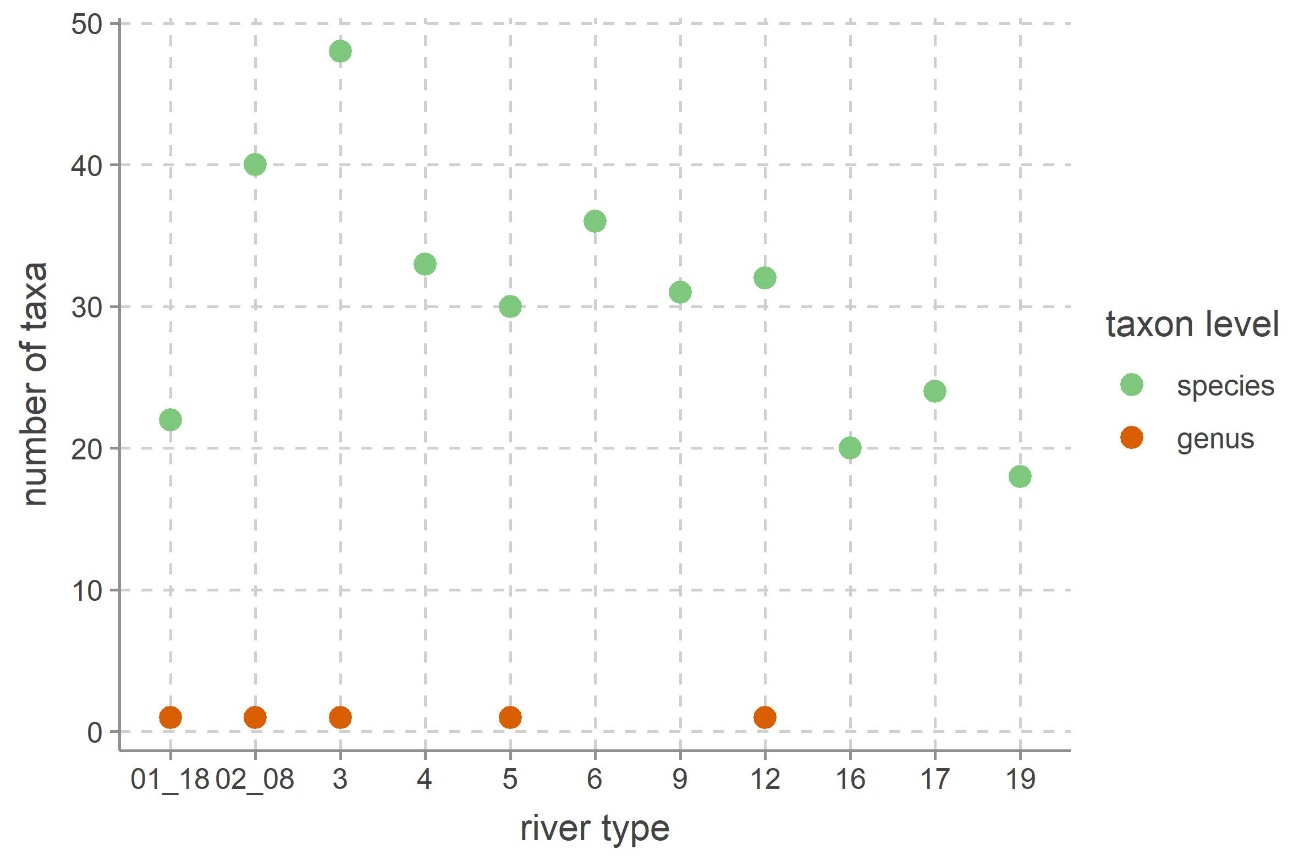


Figure 5: Numbers of taxa on each taxonomical level for all typical assemblages. Dots were removed if the number of taxa was zero.

We can express the uniqueness of a TA with the following score: Each taxon receives a taxon uniqueness score which is one divided by the number of TAs it occurs in. For each river type, we sum the taxon scores of all taxa up and divide it by the number of taxa in the river type’s TA. If all taxa in the TA are unique to that TA the score is one. If all species occur in one other TA the score is 0.5. The minimal score depends on the number of TAs, as it is 1 divided by that number and it signals that all species in that TA occur in all other TAs. These scores are shown in Figure 6. The dashed horizontal line indicates the minimum score.

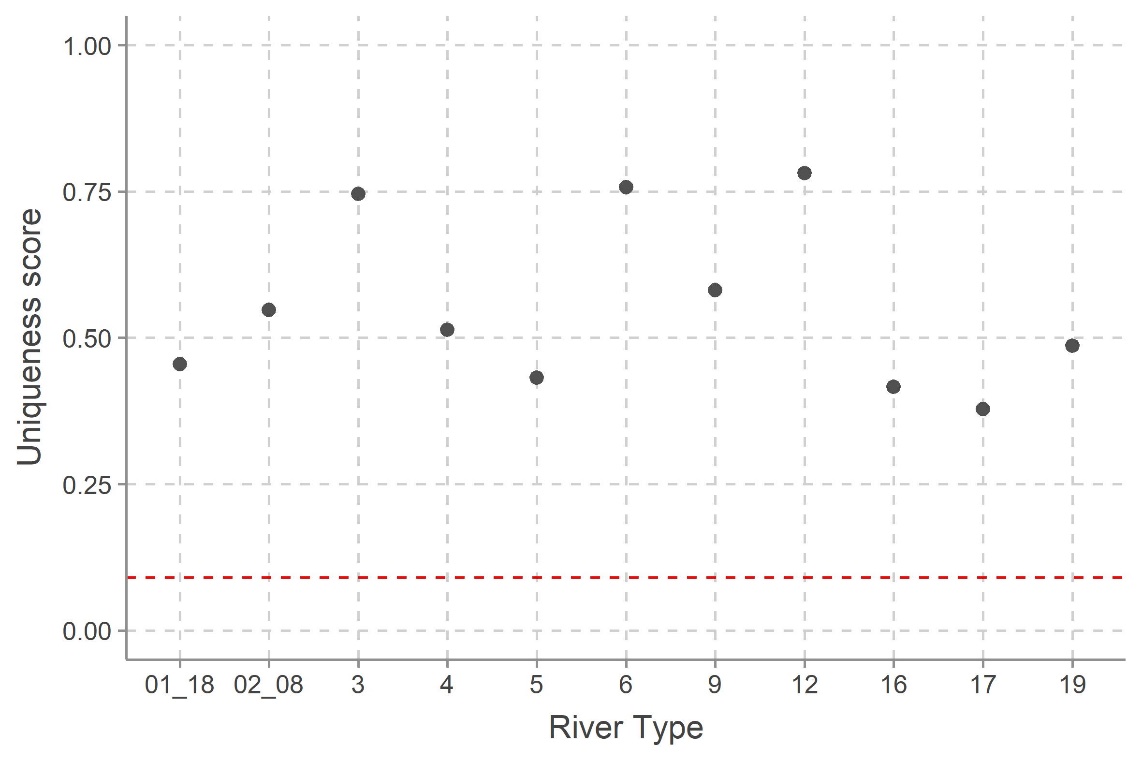


Figure 6: Uniqueness scores for all typical assemblages. The red dashed line indicates the minimum possible score of 0.08.

To visualize how similar or dissimilar the TAs are, we employed three different visualization techniques: t-distributed Stochastic Neighbor Embedding (t-SNE, Figure 7), Principal Coordinates Analysis (PCoA, Figure 8) and Nonmetric multidimensional scaling (NMDS, Figure 8). The latter two are based on a Jaccard distance matrix.

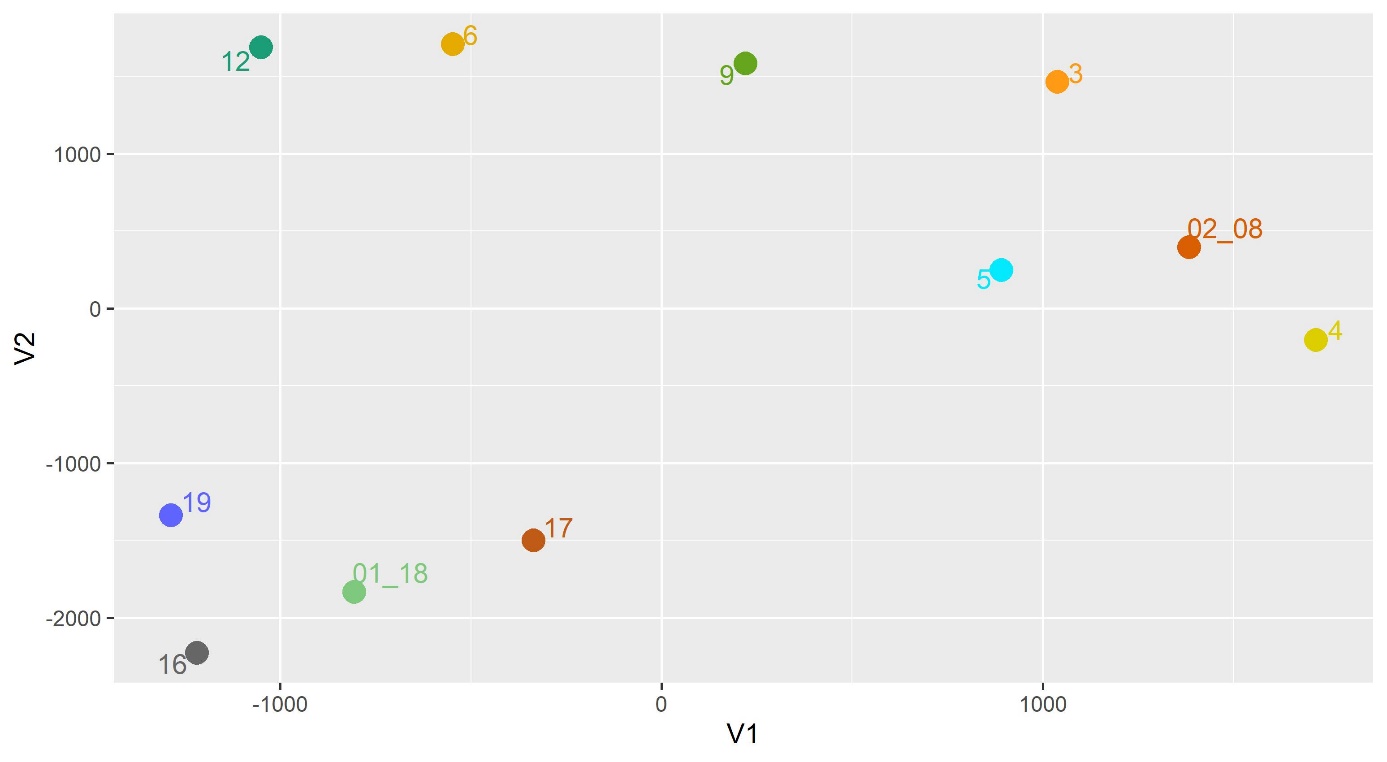


Figure 7: t-SNE of typical assemblages.

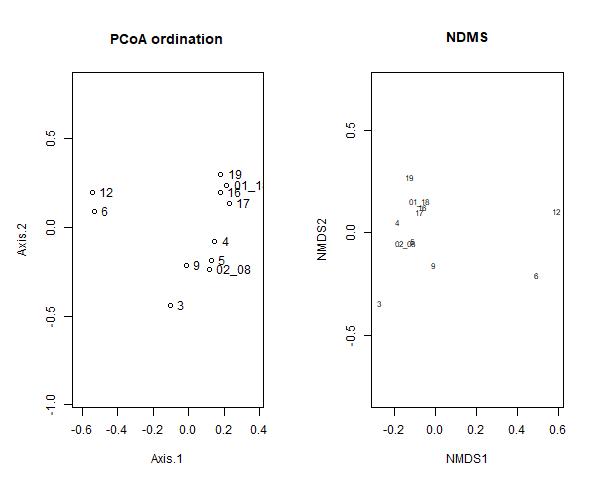


Figure 8: PCoA and NMDS of typical assemblages

In all three representations, the four river types RT01\_18, RT16, RT17, and RT19 are one group. RT17 to 19 are Mediterranean streams and RT16 are glacial streams that also fall mainly within the region (the Pyrenees, the Western Alps, and the Central Alps. We do not have sampling sites in the later). The large similarity between RT1 (Very Large Streams) and RT18 which lead to their combination, might be an artifact of the spatially biased nature of our samples for RT1. For the other TAs, the groupings differ more strongly between visualization approaches. RT02+08 and RT05 are close in all three, with RT04 being a little further away. The TA for RT09 and RT03 are sometimes associated stronger (PCoA for RT09, t-SNE for RT03) and weaker (t-SNE for RT09, PCoA and NMDS for RT03) with this cluster.

The two TAs RT06 and RT12 are separated from the rest (less strongly so in t-SNE but consider that distances beyond the scale of the employed t-distribution are meaningless). Their similarity to each other differs from high (PCoA) to medium (NMDS). RT06 and RT12 are the two stream types highly influenced by organic matter which is known to impact diatom communities (Hering et al., 2006).

## Typical assemblages

Below we list the different taxa that belong to the TAs.

### RT01+18 - Mediterranean and Western European medium to very large rivers.

Achnanthidium atomus, Achnanthidium minutissimum, Amphora pediculus inariensis Complex, Anomoeoneis sphaerophora, Cocconeis pediculus, Cocconeis placentula, Cosmioneis pusilla, Cyclotella wuethrichiana, Cymbella excisa, Encyonema silesicacum minutum lange bertalotii, Eunotia serra Complex, Gomphonema olivaceum olivaceoides, Gomphonema parvulum Complex, Gomphonema pumilum complex, Halamphora thumensis, Navicula cryptotenella cryptotenelloides, Navicula margalithii tripunctata, Navicula reichardtiana caterva, Navicula sorriensis, Nitzschia dissipata recta Complex, Nitzschia fonticola Complex, Reimeria sinuata, Gomphonema

### RT02+08 - siliceous rivers of medium to large size and low to medium altitude

Achnanthidium minutissimum, Amphora pediculus inariensis Complex, Brachysira styriaca, Cocconeis euglypta, Cocconeis placentula, Conticribra guillardii, Delphineis surirella, Encyonema silesicacum minutum lange bertalotii, Entomoneis paludosa, Eolimna minima seminulum atomoides, Eunotia subtilissima, Fragilaria tessella, Gomphoneis eriense, Gomphonema olivaceum olivaceoides, Gomphonema parvulum Complex, Gomphonema saprophilum, Mayamaea alcimonica, Mayamaea permitis, Melosira varians Complex, Navicula concentrica, Navicula cryptotenella cryptotenelloides, Navicula gregaria, Navicula lanceolata complex, Navicula margalithii tripunctata, Navicula menisculus antonii, Navicula reichardtiana caterva, Nitzschia dissipata recta Complex, Nitzschia fonticola Complex, Nitzschia inconspicua Complex, Nitzschia palea paleacea, Nitzschia sociabilis, Nitzschia subcapitellata, Pinnularia divergens complex, Plagiogramma laeve, Planothidium lanceolatum, Reimeria sinuata, Rhaphoneis amphiceros, Sellaphora submuralis Complex, Skeletonema subsalsum, Thalassiosira proschkinae, Rhoicosphenia

### RT03 - Lowland siliceous, very small-small

Achnanthidium minutissimum, Amphora pediculus inariensis Complex, Berkeleya rutilans, Brachysira garrensis, Campylosira cymbelliformis, Catenula adhaerens, Chamaepinnularia soehrensis Complex, Cocconeis placentula, Cyclostephanos Complex, Cyclostephanos tholiformis, Cyclotella hakanssoniae, Cymatosira belgica, Delphineis minutissima, Dimeregramma minor, Discostella complex, Encyonema triangulum, Eolimna minima seminulum atomoides, Fallacia crassicostata, Gomphonema parvulum Complex, Halamphora coffeaeformis, Hippodonta hungarica, Luticola kotschyi, Mayamaea fossalis, Mayamaea ingenua, Melosira varians Complex, Navicula cryptotenella cryptotenelloides, Navicula gottlandica, Navicula gregaria, Navicula lanceolata complex, Nitzschia adamata, Nitzschia amphibia Complex, Nitzschia aurariae, Nitzschia dissipata recta Complex, Nitzschia inconspicua Complex, Nitzschia lacuum alpina bryophila , Nitzschia palea paleacea, Nitzschia perspicua, Nitzschia pura linearis Complex, Nitzschia sociabilis, Opephora guenter, Paralia sulcata, Planothidium lanceolatum, Platessa bahlsii, Staurosira oldenburgiana, Staurosira punctiformis, Stephanodiscus hantzschii, Thalassionema nitzschioides, Thalassiosira decipiens, Rhoicosphenia

### RT04 - Lowland calcareous, medium-large

Achnanthidium minutissimum, Amphora ocellata, Amphora pediculus inariensis Complex, Cocconeis euglypta, Cocconeis pediculus, Cocconeis placentula, Cyclotella striata, Encyonema silesicacum minutum lange bertalotii, Eolimna minima seminulum atomoides, Fragilaria acidoclinata, Gomphonema bavaricum, Gomphonema hebridense, Gomphonema olivaceum olivaceoides, Gomphonema parvulum Complex, Halamphora montana, Navicula capitatoradiata, Navicula cryptotenella cryptotenelloides, Navicula margalithii tripunctata, Navicula menisculus antonii, Navicula reichardtiana caterva, Navicula rotunda, Navicula trophicatrix, Nitzschia amphibia Complex, Nitzschia dissipata recta Complex, Nitzschia fonticola Complex, Nitzschia palea paleacea, Nitzschia sociabilis, Pinnularia alpina lata borealis complex, Planothidium engelbrechtii, Planothidium lanceolatum, Stephanodiscus niagarae complex, Thalassiosira gessneri, Thalassiosira lacustris

### RT05 - Lowland calcareous, very small-small

Achnanthes coarctata, Achnanthidium minutissimum, Amphora pediculus inariensis Complex, Caloneis bacillum Complex, Cocconeis euglypta, Cocconeis placentula, Encyonema silesicacum minutum lange bertalotii, Eolimna minima seminulum atomoides, Eunotia monodon, Fallacia subhamulata helensis, Gomphonema dichotomum, Gomphonema parvulum Complex, Gomphonema pumilum complex, Halamphora montana, Hantzschia abundans, Melosira varians Complex, Navicula cryptocephala, Navicula cryptotenella cryptotenelloides, Navicula gregaria, Navicula margalithii tripunctata, Navicula menisculus antonii, Navicula reichardtiana caterva, Neidiomorpha binodeformis, Neidium iridis Complex, Nitzschia dissipata recta Complex, Nitzschia palea paleacea, Nitzschia sociabilis, Planothidium lanceolatum, Reimeria sinuata, Tabellaria fenestrata Complex, Rhoicosphenia

### RT06 - Lowland, organic and siliceous, very small-large

Achnanthes rupestoides kryophila broenlundense, Achnanthidium minutissimum, Aneumastus stroesei, Brachysira vitrea Complex, Diatoma anceps, Encyonema silesicacum minutum lange bertalotii, Eunotia exigua elegans Complex, Eunotia arcus mucophila bilunaris Complex, Eunotia Complex, Eunotia implicata Complex, Eunotia incisa Complex, Eunotia parallela Complex, Eunotia septentrionalis, Fragilaria capucina complex, Frustulia rhomboides Complex, Gomphonema parvulum Complex, Humidophila schmassmannii, Meridion circulare Complex, Navicula cryptocephala, Navicula rhynchocephala, Nitzschia flexa, Nitzschia palea paleacea, Odontidium mesodon, Pinnularia appenticulata perirrorata silvatica, Pinnularia interrupta Complex, Pinnularia mesolepta Complex, Pinnularia sinistra, Pinnularia viridis complex, Platessa oblongella, Psammothidium curtissimum, Psammothidium subatomoides, Psammothidium ventrale Complex, Rossithidium pusillum anastasiae, Stauroneis phoenicenteron Complex, Tabellaria flocculosa Complex, Ulnaria ulna complex

### RT09 - Mid-altitude siliceous, very small-small

Achnanthidium minutissimum, Achnanthidium subatomus, Amphora micra, Amphora pediculus inariensis Complex, Chamaepinnularia submuscicola, Cocconeis euglypta, Cocconeis placentula, Encyonema silesicacum minutum lange bertalotii, Eolimna minima seminulum atomoides, Eolimna neocaledonica, Eunotia subarcuatoides, Fallacia muraloides, Fragilaria heidenii, Fragilaria neointermedia, Gomphonema parvulum Complex, Navicula cryptocephala, Navicula cryptotenella cryptotenelloides, Navicula gregaria, Navicula lanceolata complex, Naviculadicta cosmopolitana, Nitzschia agnewii, Nitzschia dissipata recta Complex, Nitzschia frequens, Nitzschia inducta, Nitzschia palea paleacea, Nitzschia valdecostata, Nupela lapidosa, Planothidium lanceolatum, Psammothidium subatomoides, Reimeria sinuata, Tabellaria quadriseptata Complex

### RT12 – Mid-altitude, organic and siliceous

Achnanthes acares ricula carissima, Achnanthidium linearioide, Achnanthidium minutissimum, Adlafia Complex, Adlafia minuscula Grunow Lange Bertalot, Brachysira neoexilis, Brachysira zellensis, Caloneis tenuis Complex, Encyonema gaeumannii perpusillum, Encyonema gracile luna, Encyonema silesicacum minutum lange bertalotii, Eunotia exigua elegans Complex, Eunotia arcus mucophila bilunaris Complex, Eunotia implicata Complex, Eunotia naegelii, Eunotia pectinalis Complex, Eunotia praerupta Complex, Fragilaria arcus, Fragilaria capucina complex, Fragilaria tenera complex, Frustulia rhomboides Complex, Gomphonema parvulum Complex, Navicula angusta, Navicula notha leptostriata, Nitzschia bavarica Complex, Nitzschia perminuta Complex, Pinnularia subcapitata Complex, Psammothidium rossii altaica, Rossithidium nodosum, Rossithidium pusillum anastasiae, Tabellaria flocculosa Complex, Ulnaria ulna complex, Stenopterobia

### RT16 – Glacial Rivers

Achnanthidium atomoides, Achnanthidium minutissimum, Amphora pediculus inariensis Complex, Cocconeis euglypta, Cocconeis placentula, Cymbella tridentina, Diatoma moniliformis tenuis, Distrionella incognita, Encyonema silesicacum minutum lange bertalotii, Fragilaria arcus, Fragilaria incognita, Gomphonema olivaceum olivaceoides, Gomphonema parvulum Complex, Gomphonema pumilum complex, Navicula cryptotenella cryptotenelloides, Navicula margalithii tripunctata, Nitzschia dissipata recta Complex, Nitzschia fonticola Complex, Odontidium hyemale, Reimeria sinuata

### RT17 - Mediterranean, lowland, medium-Large, perennial

Achnanthidium minutissimum, Amphora pediculus inariensis Complex, Aneumastus tusculus, Cocconeis euglypta, Cocconeis pediculus, Cocconeis placentula, Cymbella excisa, Cymbella kappii, Diatoma vulgaris, Encyonema silesicacum minutum lange bertalotii, Encyonema simile, Gomphonema parvulum Complex, Gomphonema pumilum complex, Navicula cryptotenella cryptotenelloides, Navicula margalithii tripunctata, Navicula menisculus antonii, Navicula reichardtiana caterva, Nitzschia dissipata recta Complex, Nitzschia fonticola Complex, Nitzschia inconspicua Complex, Nitzschia palea paleacea, Planothidium holstii, Pseudostaurosira neoelliptica, Reimeria sinuata

### RT19 - Mediterranean, very small-small, perennial

Achnanthidium minutissimum, Amphora pediculus inariensis Complex, Cocconeis euglypta, Cocconeis pediculus, Cocconeis placentula, Cymbella excisa, Cymbella tumidula, Encyonema geisslerae, Encyonema silesicacum minutum lange bertalotii, Epithemia goeppertiana, Epithemia turgida, Gomphonema corsicum, Gomphonema pumilum complex, Mastogloia Complex, Navicula cryptotenella cryptotenelloides, Navicula margalithii tripunctata, Nitzschia fonticola Complex, Stauroneis separanda

# References

Cáceres, M. de, & Legendre, P. (2009). Associations between species and groups of sites: inindices and statistical inference. *Ecology*, *90*(12), 3566–3574. https://doi.org/10.1890/08-1823.1

Dufrêne, M., & Legendre, P. (1997). SPECIES ASSEMBLAGES AND INDICATOR SPECIES: THE NEED FOR A FLEXIBLE ASYMMETRICAL APPROACH. *Ecological Monographs*, *67*(3), 345–366.

Hering, D., Johnson, R. K., Kramm, S., Schmutz, S., Szoszkiewicz, K., & Verdonschot, P. F. M. (2006). Assessment of European streams with diatoms, macrophytes, macroinvertebrates and fish: A comparative metric-based analysis of organism response to stress. *Freshwater Biology*, *51*(9), 1757–1785. https://doi.org/10.1111/j.1365-2427.2006.01610.x

Kahlert, M., Rühland, K. M., Lavoie, I., Keck, F., Saulnier-Talbot, E., Bogan, D., et al. (2020). Biodiversity patterns of Arctic diatom assemblages in lakes and streams: Current reference conditions and historical context for biomonitoring. *Freshwater Biology*, (April 2019), 1–25. https://doi.org/10.1111/fwb.13490

Lecointe, M. (2017). OMNIDA. Retrieved from http://www.omnidia.fr

Lee, S. S., Bishop, I. W., Spaulding, S. A., Mitchell, R. M., & Yuan, L. L. (2019). Taxonomic harmonization may reveal a stronger association between diatom assemblages and total phosphorus in large datasets. *Ecological Indicators*, *102*(October 2018), 166–174. https://doi.org/10.1016/j.ecolind.2019.01.061

Mauch, E., Schmedtje, U., Maetze, A., & Fischer, F. (2017). Taxaliste der Gewässerorganismen Deutschlands. *Informationsberichte Des Bayerischen Landesamtes Für Wasserwirtschaft*, *1*(3).

Rimet, F., Gusev, E., Kahlert, M., Kelly, M. G., Kulikovskiy, M., Maltsev, Y., et al. (2019). Diat.barcode, an open-access curated barcode library for diatoms. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-51500-6

# 