# Derivation of Typical Assemblages

Jonathan Jupke

November 27, 2020

## Contents

# 1 Introduction

In GetReal, we investigate whether selected freshwater and terrestrial assemblages vary spatially and seasonally in their sensitivity towards chemicals and draw conclusions for risk assessment. After we agreed on using the Pan-European river typology proposed by Lyche Solheim *et al.* (2019) to delineate the receiving freshwater systems, the next step was to derive their typical assemblages (TA). Here we describe the methods we used to derive TAs of macroinvertebrates and diatoms for selected river types across Europe. We also describe and briefly discuss the results.

# 2 Harmonizing taxa names

International diatom occurrence datasets require extensive harmonization because of the taxonomic resolution differing between datasets, different working groups using different nomenclatures, identification errors, and ongoing changes to the accepted nomenclature (Kahlert *et al.* 2020). Though harmonizing occurrence datasets can reduce their taxonomic resolution, it also facilitates the detection of large-scale spatio-temporal patterns (Lee *et al.* 2019). We compared all our datasets against six databases that contain accepted names, synonyms with links to the respective accepted names, and suggestions for grouping contentious taxa in larger complexes. If we found a taxon name in one of the databases we either accepted it, changed it into the accepted name in case of a synonym, or included it in a complex. We did not query taxa we found in earlier databases against later ones, but in case the name changed from the original one, we queried the new one against earlier databases. Lastly, we controlled the results visually for consistency. We used the following databases in the same order:

1. Table S2 from Kahlert *et al.* (2020)
2. The taxon list associated with the OMNIDA software (Lecointe *et al.* 1993)
3. The German list of freshwater organisms (Mauch *et al.* 2017)
4. The diat.barcode database (Rimet *et al.* 2019)
5. The website algaebase.org (Guiry 2020)
6. The global biodiversity information platform (gbif) (GBIF.org 2020)

We harmonized the macroinvertebrate data with gbif through the taxize R package (Chamberlain & Szöcs 2013).

# 3 What is the optimal taxonomic level?

One result of the second progress review for GetReal (29.04.2020) was that we should determine an optimal taxonomic level for each taxon separately instead of using one common level for all data. *Oligochaetes*, for example, are usually only determined to subclass level. In a setting with one common taxonomic level (e.g. genus) they would have to be omitted if this level would be higher than subclass. By using taxon-specific levels that take low-resolved taxa into account, we can thus use more of the data. However this poses the challenge of finding an optimal level for each taxon given a dataset.
The following refers exclusively to the macroinvertebrate data, as the taxonomic resolution was not an issue with diatom datasets. We describe the procedure for diatom data at the end of the section. The optimal level was established with a hierarchical approach. First, we removed all observations from Phyla and Classes that were not present in all datasets. We assumed that these represented differences in sampling rather than in communities. The classes Clitellata (Annelida), Insecta, Malacostraca (Arthropoda), Bivalvia and Gastropoda (Mollusca) remained. In the following, a higher taxonomic level refers to levels with higher resolution, i.e. species is the highest taxonomic level and kingdom the lowest. For each taxon, we calculated the percentage of observations represented at each higher level. For example, 4.12% of observations from the order *Lepidoptera* are at the species level, 74.77% at the genus level, 7.75% at the family level, and 13.35% at the order level. Now given a threshold X, we hold a taxon to be optimally represented at a certain taxonomic level if less than X% are represented by higher levels. For example, *Lepidoptera* would be represented on order level if X > 4.12% + 74.77% + 7.75% = 86.64%. As there are no theoretical grounds on which to base such a threshold value we searched for noticeable patterns in the data (Figure 1). The most salient change occurs between 85% and 86%. It occurs because for X > 86% *Chironomidae* are represented at the family level. We used 85% as threshold. Observations that were missed by this procedure, e.g. observations of *Chironomidae* at the family level, were included at their respective level.

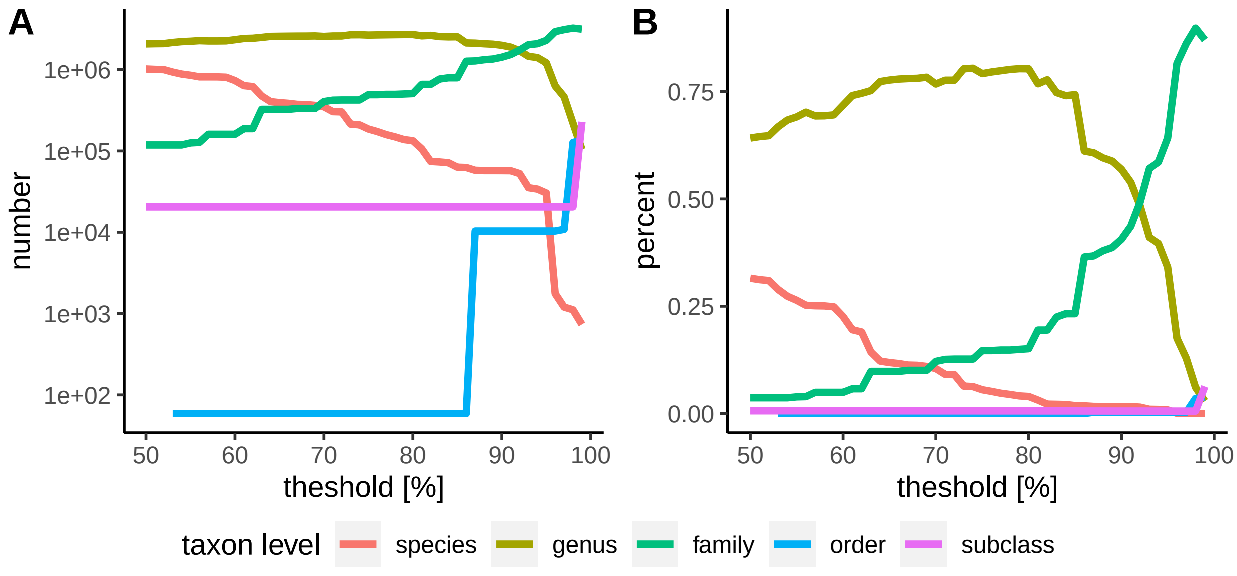

Figure 1: Effect of the treshold value on (A) total observations of each taxonomic level and (B) percentage of each taxonomic level.

For the diatoms, we employed 75% as a threshold, because for *Gomphonema*, which is the fourth most common genus in our dataset, 81.43% of the observations were at the species level. The taxonomic resolution was higher than in the macroinvertebrate data and the lowest resolution is the genus level. The equivalent of Figure 1 for diatoms can be found here.

# 4 Can we represent the stream types with our data?

We determined visually whether our dataset contained enough sampling sites in a given river type to derive meaningful TAs. The degree of representation for river types was graded in a three-tier system: high, medium, and low. A high degree of representation indicates, that we have many sampling locations, which are distributed across the instances of a river type that fall within the countries considered in GetReal. A low degree indicates the opposite, i.e. few and spatially clustered sites. A medium rating implies that we either have many sampling sites, but these only extend over parts of the countries or few sites that extend over most of the countries. The ratings for all river types for macroinvertebrates and diatoms are shown in table 1.

For each river type we provide maps with the associated sampling sites for macroinvertebrates and for diatoms.

Further analyses were conducted for all stream types with a high or medium degree of representation. More information on the river types is available in Lyche Solheim *et al.*

Table 1: The ratings for all river types for macroinvertebrates and diatoms

| Rating | Taxon | River.Types |
|---|---|---|
| high | macroinvertebrates | 4, 5, 9, 10, 11, 12, 13, 16 |
| high | diatoms | |
| medium | macroinvertebrates | 1, 2, 3, 8, 14, 15, 18 |
| medium | diatoms | 1, 2, 3, 4, 5, 6, 8, 9, 12, 14, 16, 17, 18, 19 |
| low | macroinvertebrates | 6, 7, 17, 19, 20 |
| low | diatoms | 7, 10, 11, 13, 15, 20 |

(2019). We have fewer sampling sites for diatoms than for macroinvertebrates which entails that the representation of stream types is mostly lower.

# 5 What is a typical assemblage?

We derived typical assemblages using the commonness of each taxon in a respective river type. Here, the commonness of a taxon refers to the fraction of sampling sites in a river type where that taxon is found. This is equal to the $B$ parameter used in the Species Indicator Value (Dufrêne & Legendre 1997; Cáceres & Legendre 2009). We also considered the second parameter of that statistic, the specificity $A$ which refers to the fraction of occurrences of a taxon that fall within one river type. For example, $A = 1$ indicates that all observations of taxon $x$ occurred in river type $y$ and hence that this taxon is highly specific to the given river type. A commonness of 1 shows that a taxon is present in every sample taken within a river type and therefore that it is very typical for that river type. A taxon belongs to the TA of a river type if it is more common that some threshold, which depends on the taxonomic level of the taxon, or if it is more specific than another threshold, which also depends on the taxonomic level. To prevent rare species and singletons from exerting an undue influence on the TAs we first removed taxa that occurred in less than 5% of the samples. Second, we only included highly specific taxa if they also surpassed a threshold in commonness. The first step was undertaken for each river type separately so that the effective threshold varied between river types. The B-threshold for species was 0.2, for genera 0.4 and for family or lower resolutions 0.6. The A-threshold for species was 0.9 for a species, 0.85 for genus or 0.8 for family or lower taxonomic levels. To be included, a highly specific (i.e. A > the respective threshold value) taxon also had have commonness above 0.05 for species, 0.1 for genera and 0.15 for families or lower taxonomic levels.

We did not systematically optimize these thresholds. Such procedures would require optimization criteria, but we are not aware of any criterion that would work in this context. As TAs can be very similar in composition or harbor strongly differing numbers of taxa, neither criterion would optimize what we would consider a TA. 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 varying these parameters would alter the results. We derived TAs for 50 values of **A** and **B** ranging from 0.01 to 1 and computed taxon richness and uniqueness scores (see Figure 5 and text preceding the Figure) of each TA. Please note that results are only shown and discussed for the non-redundant TAs (see section 6).

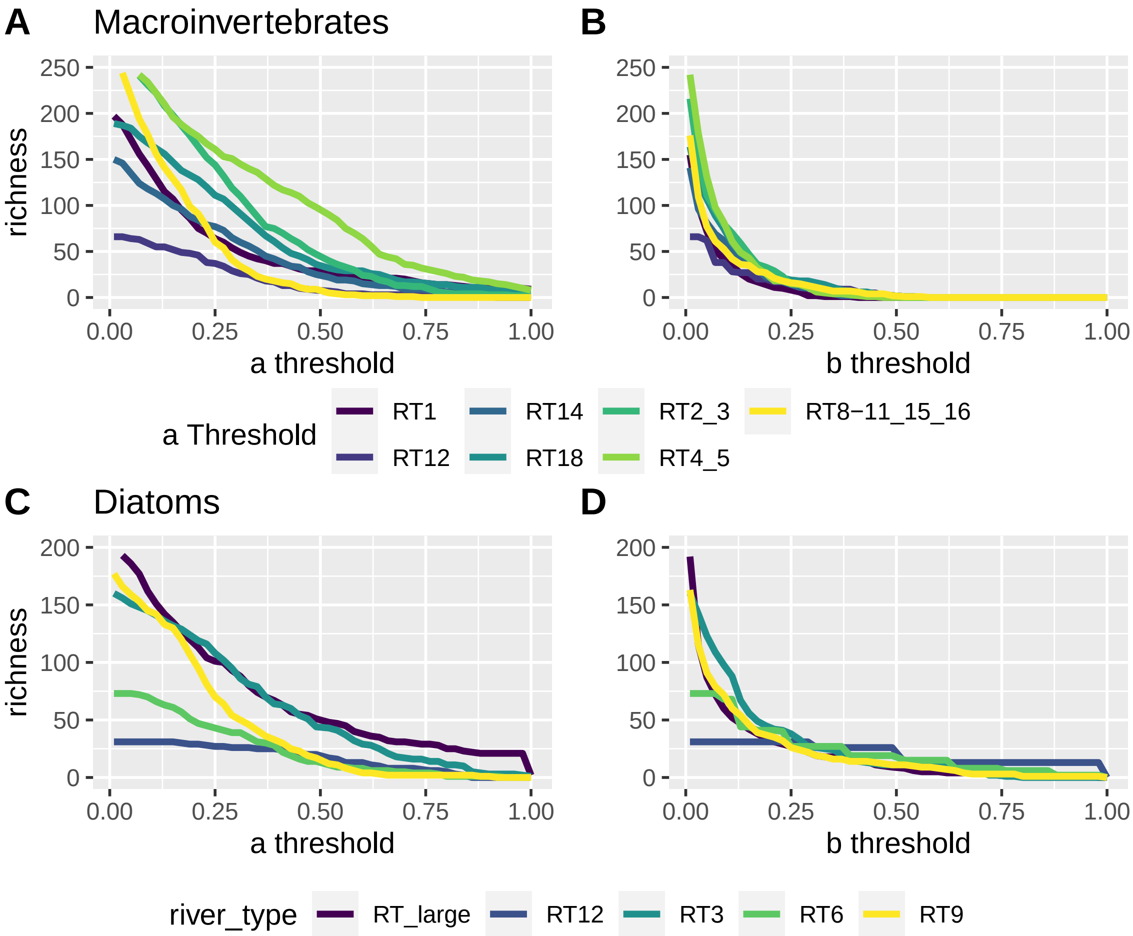

Figure 2: Changes in richness along gradients of A and B thresholds in macroinvertebrates (A, B) and diatoms (C, D). The line color indicates the river type.

Richness decreased with increasing **A** and **B** threshold in macroinvertebrates and diatoms (Figure 2). Uniqueness scores increased with **A** thresholds up to a certain level and then decreased (Figure 3). There is considerable variation between river types as to where this inflection occurs. Along the B-gradient, uniqueness score fluctuate erratically, driven by the relative rate at which other river types loose taxa and the identity of these. At some point no taxa are common enough to surpass the set B-threshold and the TAs are empty.

For macroinvertebrates this takes place for B values between 0.41 (RT1) and 0.58 (RT8-11_15_16). In the diatom assemblages this occurs later (at B = 0.8 in RT3) and not at all in RT9, RT6, RT12. This highlights that there must be several widely distributed diatom taxa, which is likely to some degree also the result of our extensive harmonization efforts.

Plots for each taxon level separately are available for macroinvertebrates and diatoms. However, the general patterns visible in Figure 2 and 3, hold for them as well.

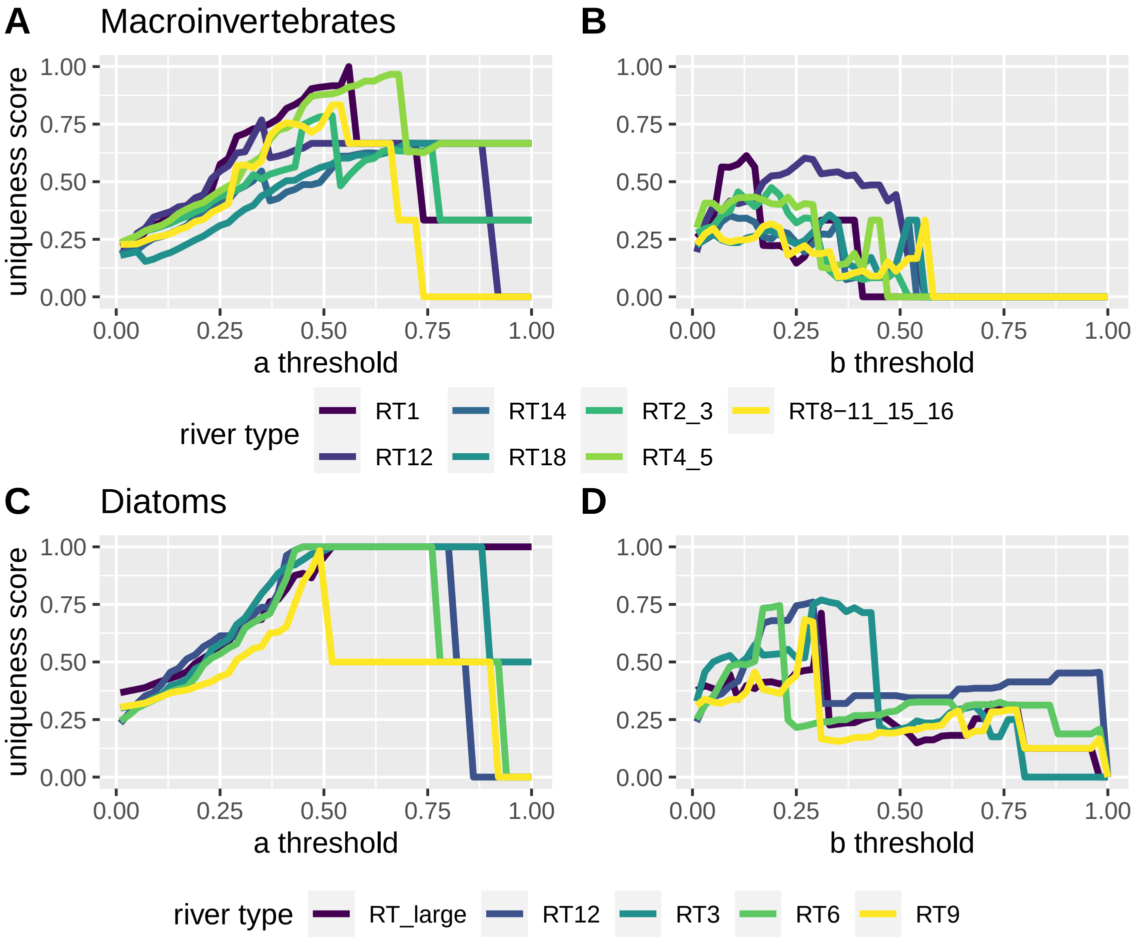

Figure 3: Changes in uniqueness scores along gradients of A and B thresholds in macroinvertebrates (A, B) and diatoms (C, D). The line color indicates the river type.

## 6 Redundancy between typical assemblages

We assessed how similar two TAs are with the Jaccard similarity of their species lists, which is the fraction of the combined taxa pool of both TAs that occurs in both. We joined two river types if their similarity was at or above 0.8. Following a combination we recomputed TAs and similarities until none were above 0.8. In macroinvertebrates, several TAs had similarities above 0.7 but none were above 0.8.

In diatoms, the river types 1,2 and 4 were combined as well as the river types 17 and 18. The first group 1, 2 and 4 constitutes medium sized to very large lowland rivers (< 200 m.a.s.l.). The three combined types are very large rivers (RT1); lowland, siliceous, medium-large (RT2) and Lowland, calcareous or mixed, medium-large (RT4).
The second group, consisting of RT17 and RT18, combines the two larger Mediterranean river types which differ in their altitude. RT17 are lowland river and RT18 mid-altitude (200 - 800 m.a.s.l.). There are two more kinds of Mediterranean river types which are not included in this group: very small to small perennial rivers (RT19) as well as intermittent streams (RT20). However, we did not have any samples for the latter river type.

# 7 Characteristics of typical assemblages

In all macroinvertebrate TAs, except RT14, genus is the prevalent taxonomic level, followed by families or lower taxonomic levels and lastly species (Figure 4A). The mean number of species was 2, mean number of genera 17.08, and the mean number of families or lower 5.85. RT5 had the least taxa with 17 taxa and RT4 was the most taxa rich assemblage with 39 taxa. For diatoms, species is the prevalent taxonomic level in all TAs (Figure 4B). The mean number of species per TA is 38.12 and the mean number of genera 3.5. RT3 has the most taxa rich TA with 67 taxa and RT16 has the least taxa in its TA with 31. Note that diatom TAs encompass more taxa than macroinvertebrate TAs which supports the trends from the sensitivity analysis (Figure 3).

We can express the uniqueness of a TA with the following score: Each taxon receives a taxon uniqueness score that 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 5. The dashed horizontal lines indicate the minimum scores.

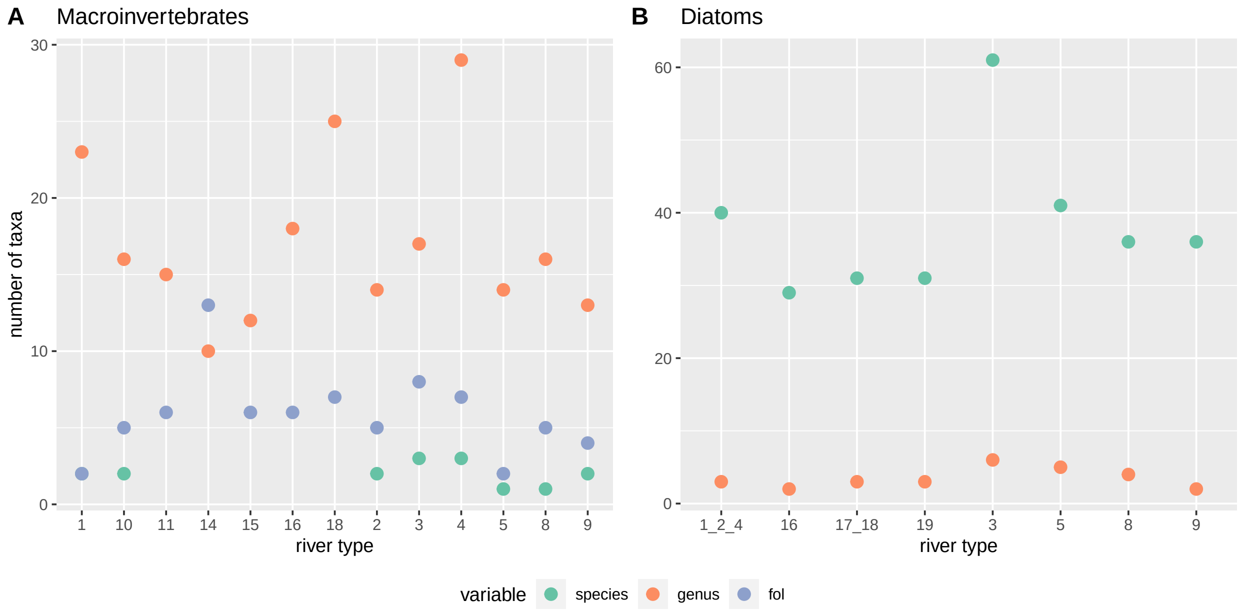

Figure 4: Numbers of taxa on each taxonomical level for all typical assemblages.

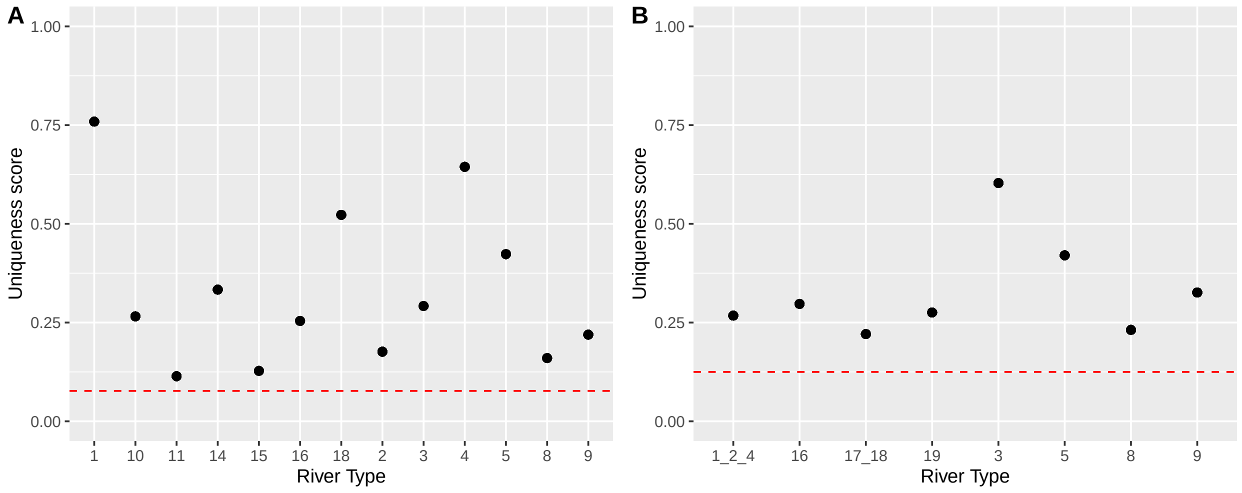

Figure 5: Uniqueness scores for typical assemblages of macroinvertebrates(A) and diatoms (B). The red dashed line indicates the lowest possible score.

We used Non-metric multidimensional Scaling (NMDS, Kruskal (1964)) to visualize the similarity of TAs, based on Jaccard distance matrices (Figure 6).

We can interpret the NMDS plot as showing three groups: one large cluster to the right, RT15 and RT16 in the middle and RT3 to the left. RT3 are lowland, siliceous, very small to small rivers and it is surprising that they are so markedly different of RT4, from which they differ only in size. RT15 and RT16 are high altitude rivers in southern Europe. As indicated

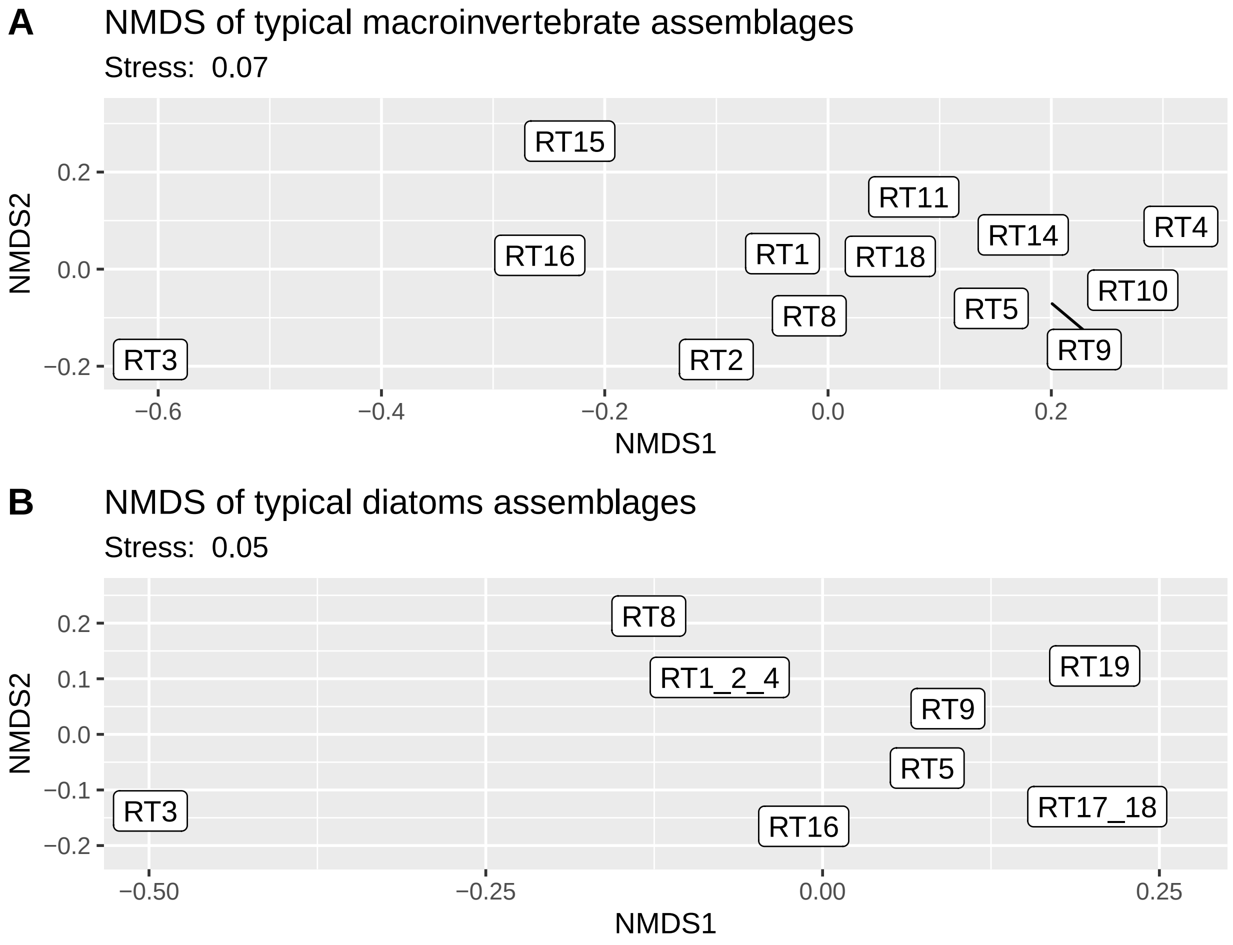

Figure 6: Non-metric multidimensional scaling of typical assemblages based on Jaccard distance matrices. A shows the typical asseblages of macroinvertebrates and B those of diatoms. The postion of RT9 in A was changed to ensure legibility. A black line, starting at the position of RT9 in the plot indicates where RT9 should be located.

before, the other TAs are quite similar to each other and one probably should not give too much meaning to the exact structuring of the larger group.

The NMDS of diatom TAs, also highlights RT3 as most unique. Beyond this obvious pattern we see the three Mediterranean river types (RT17 to 19) aligned on the first axis. Further RT5 and RT9 are close together. Both are very small to small rivers but differ in geology (RT5: calcareous, RT9: siliceous) and altitude (RT5: lowland, RT9: mid-altitude). Lastly, the combined river type RT1,2,4 and RT8 are clustered. As described before, the large cluster includes mid-sized to very large rivers and RT8 are medium to large, mid-altitude, siliceous rivers. The similarities as depicted in this NDMS clearly highlight the role of river size (Lyche Solheim *et al.* (2019) used catchment area as a proxy).

Online, we provide the taxa lists for all macroinvertebrate and diatom TAs.

# 8 Seasonal typical assemblages

In addition to the spatially defined TAs, we derived seasonal TAs (sTA) for a subset of river types. The four seasons were defined as follows: spring is March to May, Summer is June to August, Fall is September to November, and Winter is December to February. To avoid strong spatial signals in the sTAs, we only considered those river types (RT) in which samples were evenly distributed between seasons. In most cases, we had to omit parts of the data (e.g. certain seasons or data sets) to achieve an even spatio-temporal distribution. Online, we provide maps for all river types with all available seasons as well as the respective subsets that we used in the further analyses for macroinvertebrates and diatoms.
As an example, the map of macroinvertebrate samples for the combined RT 4_5 is shown in Figure 7.

To visualize differences between the seasons we used Non-metric multidimensional scaling (NMDS) on Jaccard dissimilarity matrices. The resulting NMDS plots are available for macroinvertebrates and diatoms. Figure 8 shows the NMDS plot for invertebrate samples in RT4_5. For diatoms and macroinvertebrates there are no or little discernible seasonal patterns in most river types. This also shows in high NMDS stress values (typically above 0.2).

Further, we evaluated whether the Jaccard dissimilarity between sites would be better explained by spatial distance or by season. To this end, we employed generalized dissimilarity modeling (GDM, Ferrier *et al.* (2007)). In GDMs, the response variable is the ecological dissimilarity between two sites (expressed in some *a priori* chosen dissimilarity metric, here Jaccard). Smooth functions are fitted to the environmental data and the differences between the values of these functions at the two sites of interest are used as explanatory variables. By using a generalized modeling framework we can account for the bounded nature of dissimilarity metrics (between 0-1) and the smooth functions allow for variation in the rate of

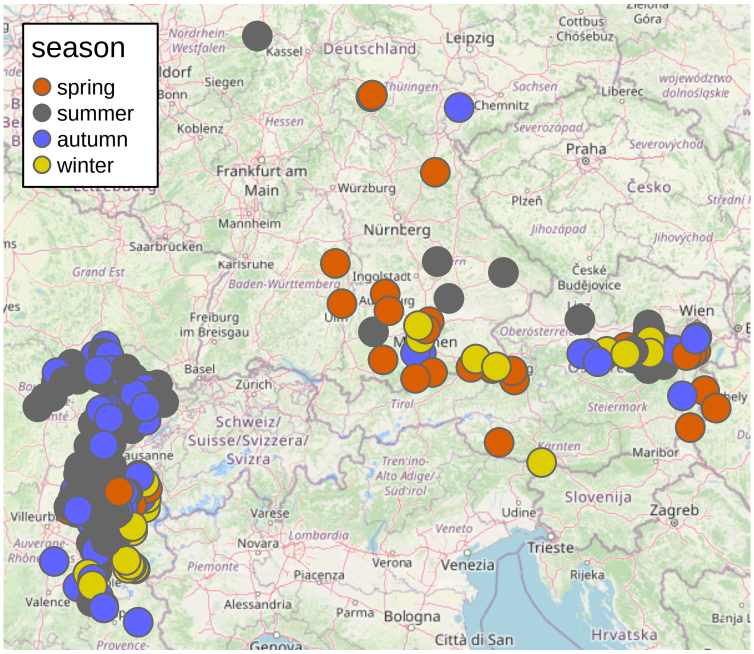

Figure 7: Map of sampling sites for river type 10. The color of the points shows the season of sampling.

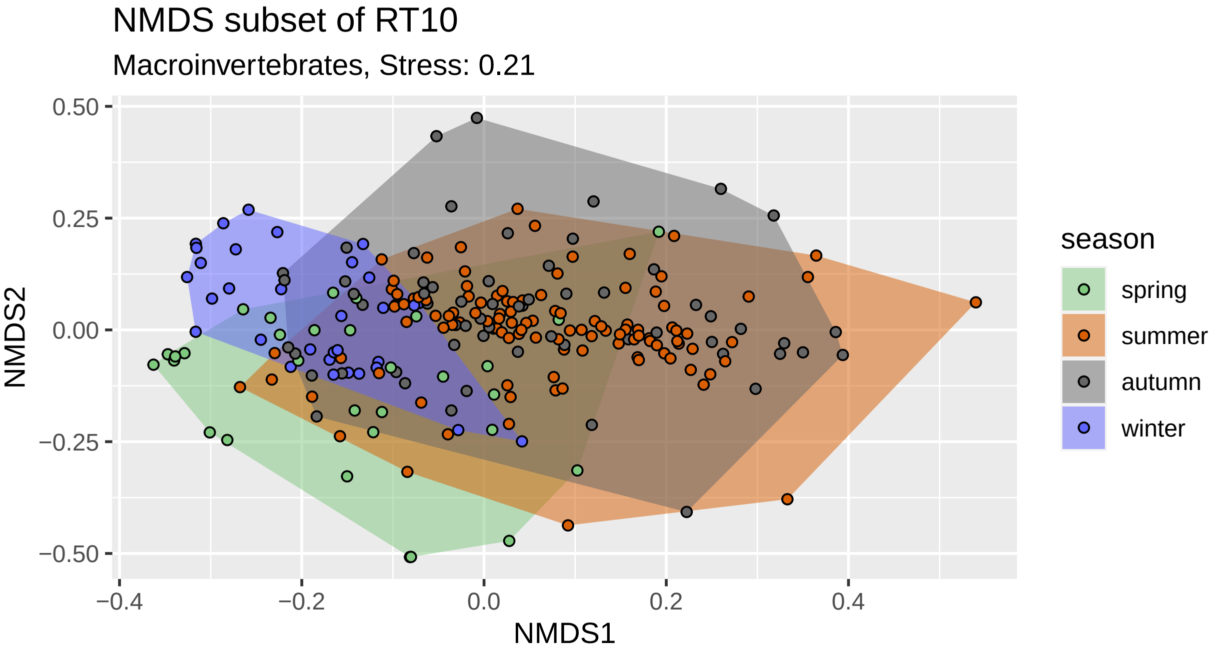

Figure 8: Nonmetric multidimensional scaling plot of Jaccard dissimilarity matrices for macroinvertebrate communities in RT4_5. The color of the points shows the season. Convex hulls surround all sampling points from one season.

compositional turnover along gradients. Plots that show the effect of spatial distance and that of season for all GDMs are available for macroinvertebrates and diatoms. The plot for invertebrates in RT4_5 is shown in Figure 9. The findings from the NMDS are confirmed in the GDMs - spatial distance explains more of the variation in jaccard distances than seasons.

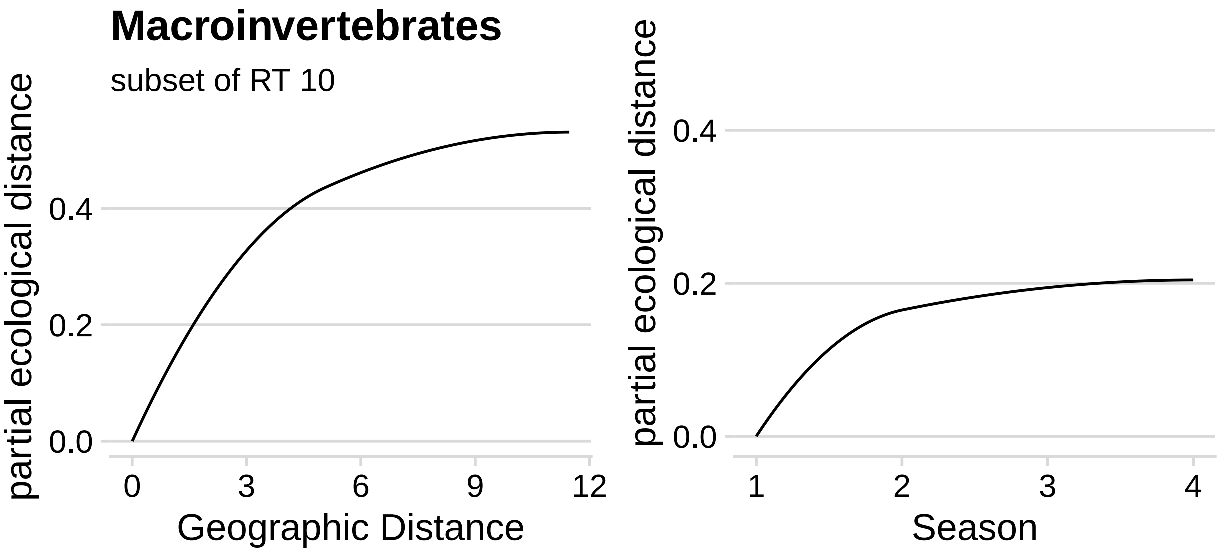

Figure 9: Partial ecological distance between sites with increasing geographic distance or chaning season (1 = spring, 2 = summer, 3 = autumn, and 4 = winter) predicted with Generalized Dissimilarity Models

We selected subsets of RT10 and RT16 (subset 1) for invertebrates as well as a subset of RT16 (subset 1) and all samples from RT17_18 for diatoms because they showed the strongest seasonal variations in GDM and NMDS. For these four river types sTA were derived in the same way as the non-seasonal TAs.

In RT16, the two seasonal TAs shared nine taxa which is about half of the total number of taxa in each sTA (Table 2). The winter sTA had more taxa than the summer sTA.

Table 2: Overlap between seasonal typical assemblages (sTA) of diatoms in river type 16 expressed in percent of taxa in row sTA also present in column sTA. N is the number of taxa in the respective sTA.

| | summer | winter | N |
|---|---|---|---|
| summer | 100.0 | 56.2 | 16 |
| winter | 40.9 | 100.0 | 22 |

In the combined river type 17_18, the number of taxa varied between seasons. Spring had the least taxa in its TA and winter the most. The sTAs from spring, summer and autumn are very similar, with overlaps exceeding 80% with one exception (autumn - spring, 72,7%). The winter sTA differs strongly from the other three and only shares about half of its taxa with them.

\begin{table}[!h]

\caption{Overlap between seasonal typical assemblages (sTA) of diatoms in the combined river type 17_18 expressed in percent of taxa in row sTA also present in column sTA. N is the number of taxa in the respective sTA.}

| | spring | summer | autumn | winter | N |
|---|---|---|---|---|---|
| spring | 100.0 | 100.0 | 94.1 | 64.7 | 17 |
| summer | 81.0 | 100.0 | 90.5 | 57.1 | 21 |
| autumn | 72.7 | 86.4 | 100.0 | 63.6 | 22 |
| winter | 42.3 | 46.2 | 53.8 | 100.0 | 26 |

\end{table}

For the macroinvertebrates, the number of taxa in the sTAs is little lower than for diatoms but the difference is less pronounced than in the TAs. In RT10, the spring sTA shares nine of its eleven taxa with all sTA of that river type and hence only includes two unique taxa (Table 3). The winter sTA also shares the same nine taxa with summer and spring and one additional taxon with the autumn sTA. Summer and autumn sTA include more taxa than the other two (16 and 21 respectively). The summer TA shared 13 of its 16 taxa with the autumn TA and hence has three unique taxa.

Table 3: Overlap between seasonal typical assemblages (sTA) of macroinvertebrates in river type 10 expressed in percent of taxa in row sTA also present in column sTA. N is the number of taxa in the respective sTA.

| | spring | summer | autumn | winter | N |
|---|---|---|---|---|---|
| spring | 100.0 | 81.8 | 81.8 | 81.8 | 11 |
| summer | 56.2 | 100.0 | 81.2 | 56.2 | 16 |
| autumn | 42.9 | 61.9 | 100.0 | 47.6 | 21 |
| winter | 75.0 | 75.0 | 83.3 | 100.0 | 12 |

In RT16, the sTA richness also varies strongly (Table 4). Summer has the highest richness (26 taxa), followed by autumn (17) and winter (11). The winter TA is completely nested within the autumn TA, which shares eleven of its 17 taxa with the summer TA. Hence there is considerable turn-over between summer and autumn TAs but not between autumn and winter.

Online, we provide the complete taxa lists for macroinvertebrate and diatom sTAs.

Table 4: Overlap between seasonal typical assemblages (sTA) of a subset of diatoms from the river type 16. Overlap is expressed in percent of taxa in row sTA also present in column sTA. N is the number of taxa in the respective sTA.

| | summer | autumn | winter | N |
|---|---|---|---|---|
| summer | 100.0 | 42.3 | 38.5 | 26 |
| autumn | 64.7 | 100.0 | 64.7 | 17 |
| winter | 90.9 | 100.0 | 100.0 | 11 |

# References

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

Chamberlain, S. & Szöcs, E. (2013). Taxize - taxonomic search and retrieval in r. *F1000Research*.

Dufrêne, M. & Legendre, P. (1997). Species Assemblages and Indicator Species: The need for a flexible asymmetrical Approach. *Ecological Monographs*, 67, 345–366.

Ferrier, S., Manion, G., Elith, J. & Richardson, K. (2007). Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions*, 13, 252–264.

GBIF.org. (2020). *GBIF home page.*

Guiry, G.M., M. D. & Guiry. (2020). *AlgaeBase. World-wide electronic publication, national university of ireland, galway.*

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*, 1–25.

Kruskal, J.B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, 115–129.

Lecointe, C., Coste, M. & Prygiel, J. (1993). "Omnidia": Software for taxonomy, calculation of diatom indices and inventories management. *Hydrobiologia*, 269, 509–513.

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, 166–174.

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.

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

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, 1–12.