# Derivation of typical invertebrate assemblages

21.08.20

Jonathan Jupke

## What is the optimal taxonomical level?

One result of the last progress review was that the taxa that make up the typical assemblages (TA) could be on different taxonomic levels (species, genus, family, …). We should include taxa that are typically represented at higher taxonomical levels at those levels (e.g. *Oligochaetes* at subclass level) and taxa whose taxonomy is usually highly resolved at higher levels. Thus, the question arises: given a dataset, what is the optimal taxonomical level to represent a specific taxon?

To establish the optimal level, we used a hierarchical approach. First, we removed all observations from Phyla and Classes that were not present in all data sets. We assumed that these represented differences in sampling rather than in communities. That left us with the classes Clitellata (Annelida), Insecta, Malacostraca (Arthropoda), Bivalvia and Gastropoda (Mollusca).

In the following higher taxonomic level refers to levels that are resolved more highly, i.e. species is the highest taxonomic level and kingdom the lowest. We inspected each higher level (subclass, order, family, genus) and calculated the percentage of observations that are represented at each 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 which is to be determined, we would call a taxon optimally represented at a certain taxonomical level if less than X% are represented by higher levels. For example, *Lepidoptera* would be represent 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 noticeable jump occurs between 85 and 86%. It occurs because for X > 86 *Chironomidae* are represented at the family level. Hence, we used 85% as a threshold. Observations that were not captured by this, *e.g.* observations of *Chironomidae* at the family level were included at their respective level.

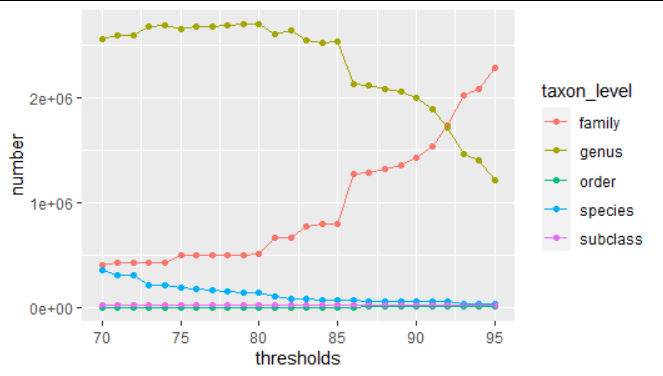


Figure 1: Number of observations at the different taxonomical resolutions as a function of the threshold value

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

Whether a stream type is sampled extensively enough in our data to derive meaningful 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/1xi3m9ZV38TjDYhaced7PEztP3rQoFdmY). The representativeness was graded in a three-tier system: good, medium, bad. Further analyses were conducted for all stream types with good and good or medium representation. More information on the river types is available in the [accompanying publication](https://www.sciencedirect.com/science/article/pii/S0048969719340203).

Good [n = 8]: 4,5,9,10,11,12,13,16

Medium [n = 7]: 1,2,3,8,14,15,18

Bad [n = 5]: 6,7,17,19,20

## 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 and Legendre 2009; Dufrêne and 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 > 0.25 or B > 0.2 and p < 0.05 or A > 0.8

Genera where considered typical if B > 0.5 or B > 0.33 and p < 0.05 or A > 0.95

Families where considered typical if B > 0.95 or B > 0.8 and p < 0.01 or A > 0.99

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. The A thresholds for genera and families were 1.25 and 1.5 times that for species, respectively. The B threshold were two and three times that of the species.

Please note that results are only shown and discussed for the non-redundant TAs (see the next section). Taxa richness decreased with increasing A and B threshold (Figure 2), while the uniqueness scores increased with B thresholds but decreased with A thresholds (Figure 3). 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. Note that graphs are only shown for all taxa levels combined. Plots for each taxon level separately are available in the [GetReal Drive](https://drive.google.com/drive/folders/1Gtm5ah9uouZmmi-Ti7l5dOJHlNvVUEma). However, the general patterns visible in Figure 2 and Figure 3, hold for them as well.

## Redundancy in typical assemblages

We assessed to which degree the different TAs overlap (Table 1). The degree of overlap is the percentageof 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 | RT2+3+4+5+12+13+14+18 | 8.3 % |
| RT02: Lowland, siliceous, medium-large | RT3 | 88.9 % |
| RT03: Lowland, siliceous, very small-small | RT2 | 69.6 % |
| RT04: Lowland, calcareous or mixed, medium-large | RT2+3+5 | 45.8 % |
| RT05: Lowland, calcareous or mixed, very small-small | RT4 | 84.6 % |
| RT08: Mid-altitude, siliceous, medium-large | RT9 | 75.0% |
| RT09: Mid-altitude, siliceous, very small-small | RT8 | 78.9 % |
| RT10: Mid-altitude, calcareous or mixed, medium-large | RT11+18 | 66.7 % |
| RT11: Mid-altitude, calcareous or mixed, very small-small | RT10 | 84.2 % |
| RT12: Mid-altitude, organic and siliceous, very small-large | RT9 | 50.0 % |
| RT13: Mid-altitude, organic and calcareous/mixed | RT2+3 | 80.0 % |
| RT14: Highland (all Europe), siliceous, incl. organic (humic) | RT16+18 | 66.7 % |
| RT15: Highland (all Europe), calcareous/mixed | RT10+RT16 | 76.5 % |
| RT16: Glacial rivers (all Europe) | RT9+10+11+15 | 59.1 % |
| RT18: Mediterranean, mid-altitude, medium-large | RT10 | 57.1 % |

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 lead to five

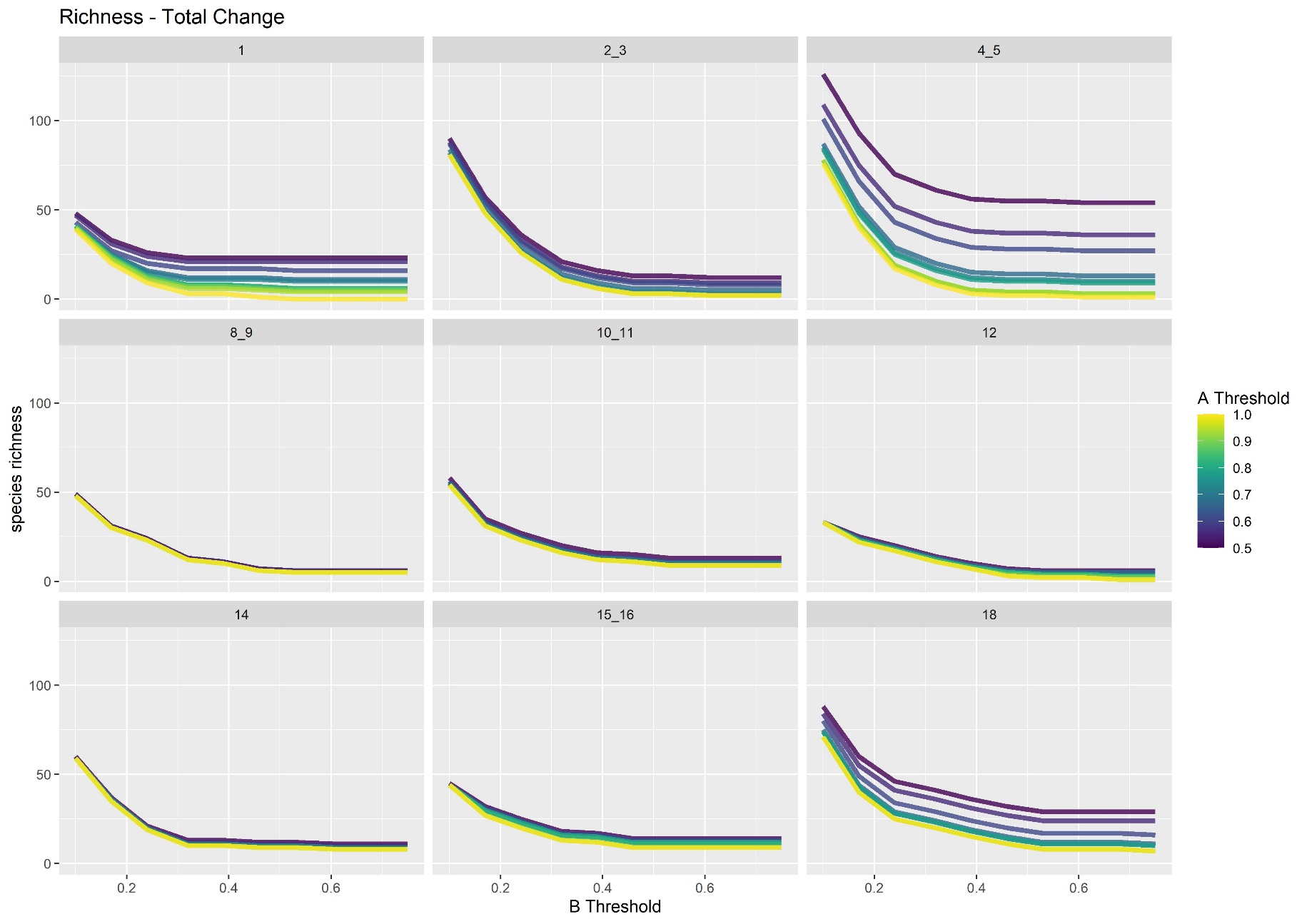


Figure 2: Changes in taxa richness along a changing B threshold. Line color indicates the employed A threshold

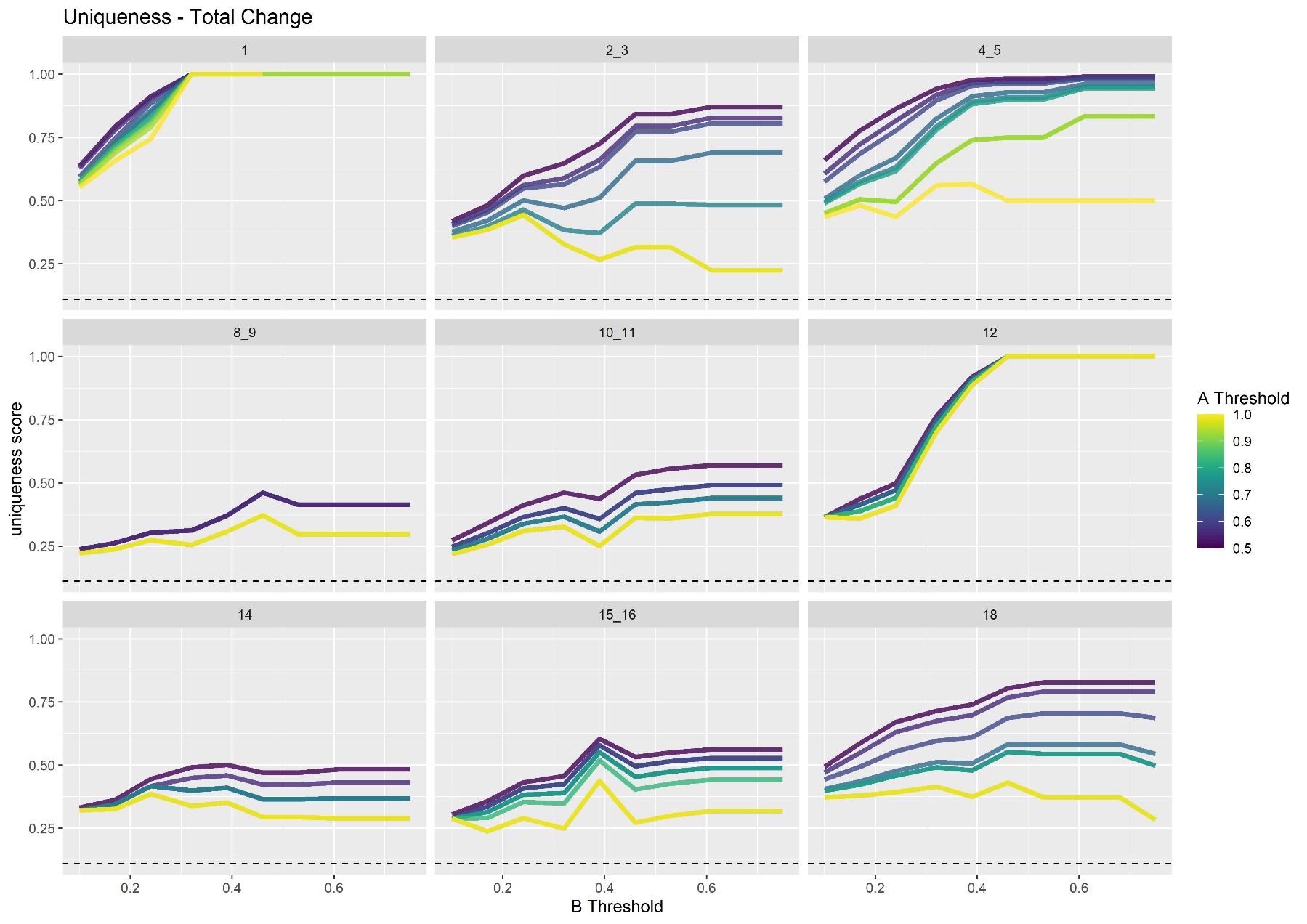


Figure 3: Changes in taxa uniqueness scores along a changing B threshold. Line color indicates the employed A threshold.

redundant assemblages. Several of those are two river types that only differ in river size: RT02 and 03, 04 and 05, 08 and 09, 10 and 11. The only exception is the combination of RT15 and 16. Both are high altitude river types that occur mainly in southern Europe, which differentiates them from the northern high altitude rivers in RT14. RT13 is also redundant with RT02 and 03 however joining it with these two river types led to a drastically reduced number of taxa in the TA, when compared to that of the combined river type RT02\_03. Since RT13 represents an exceedingly rare river type we decided to rather omit it from the analysis and proceed with RT02\_03 instead of RT02\_03\_13.

The new TAs resulted in overall lower degrees of overlap, none of which exceed the 75% threshold. The largest overlaps were between RT10\_11 and RT18+8\_9, RT14 and RT16+18 as well as RT8\_9 and RT10\_11, all of which overlapped in 66.7% of taxa.

## What do the typical assemblages look like?

In all TAs, genus is the prevalent taxonomic level (Figure 4). The numbers of species and families or higher (foh) are similar with both exceeding the other in four assemblages. The mean number of species was 3.2, mean number of genera 14.3, and the mean number of foh 2.4.

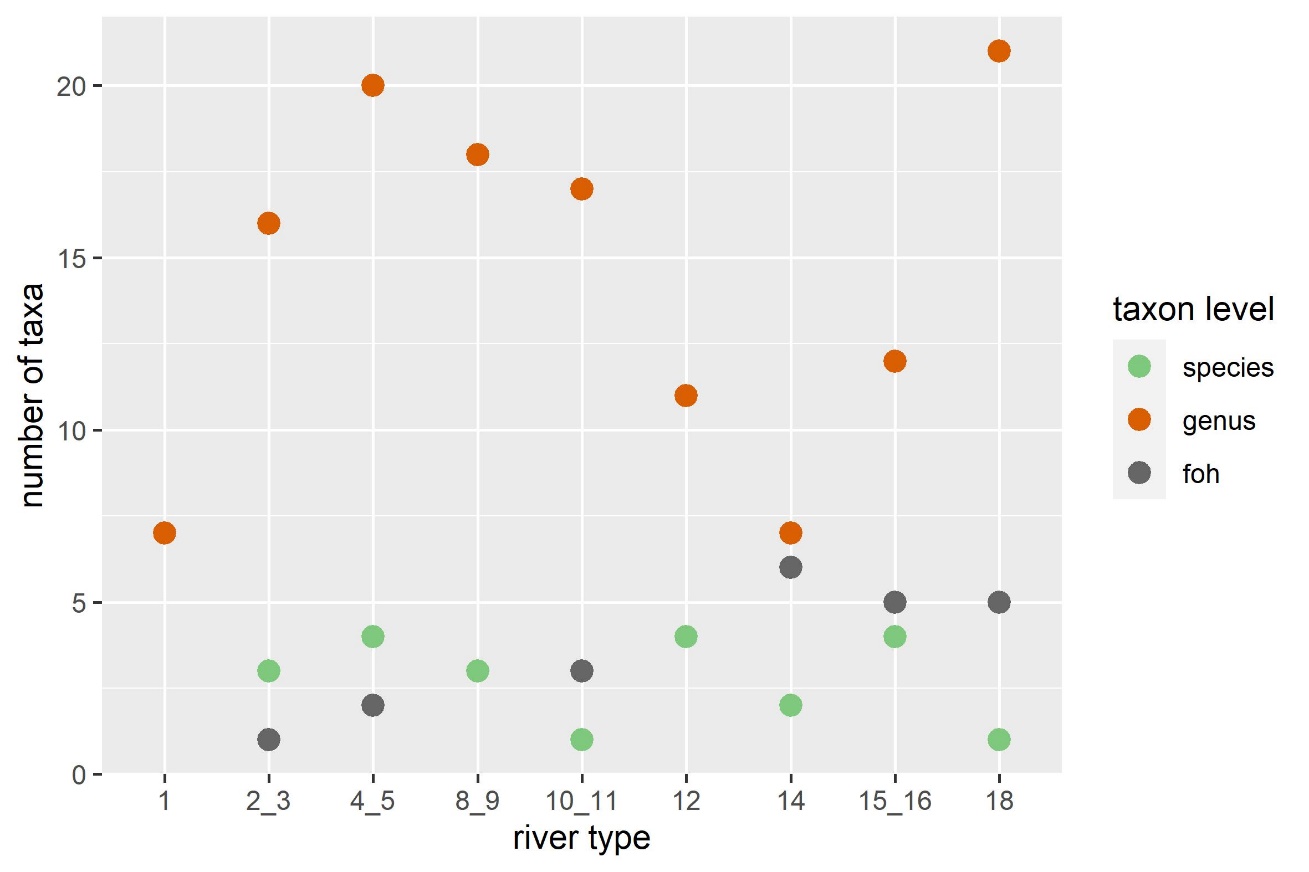


Figure 4: Numbers of taxa on each taxonomical level for all typical assemblages

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 5. The dashed horizontal lines indicate the minimum score for each representation level.

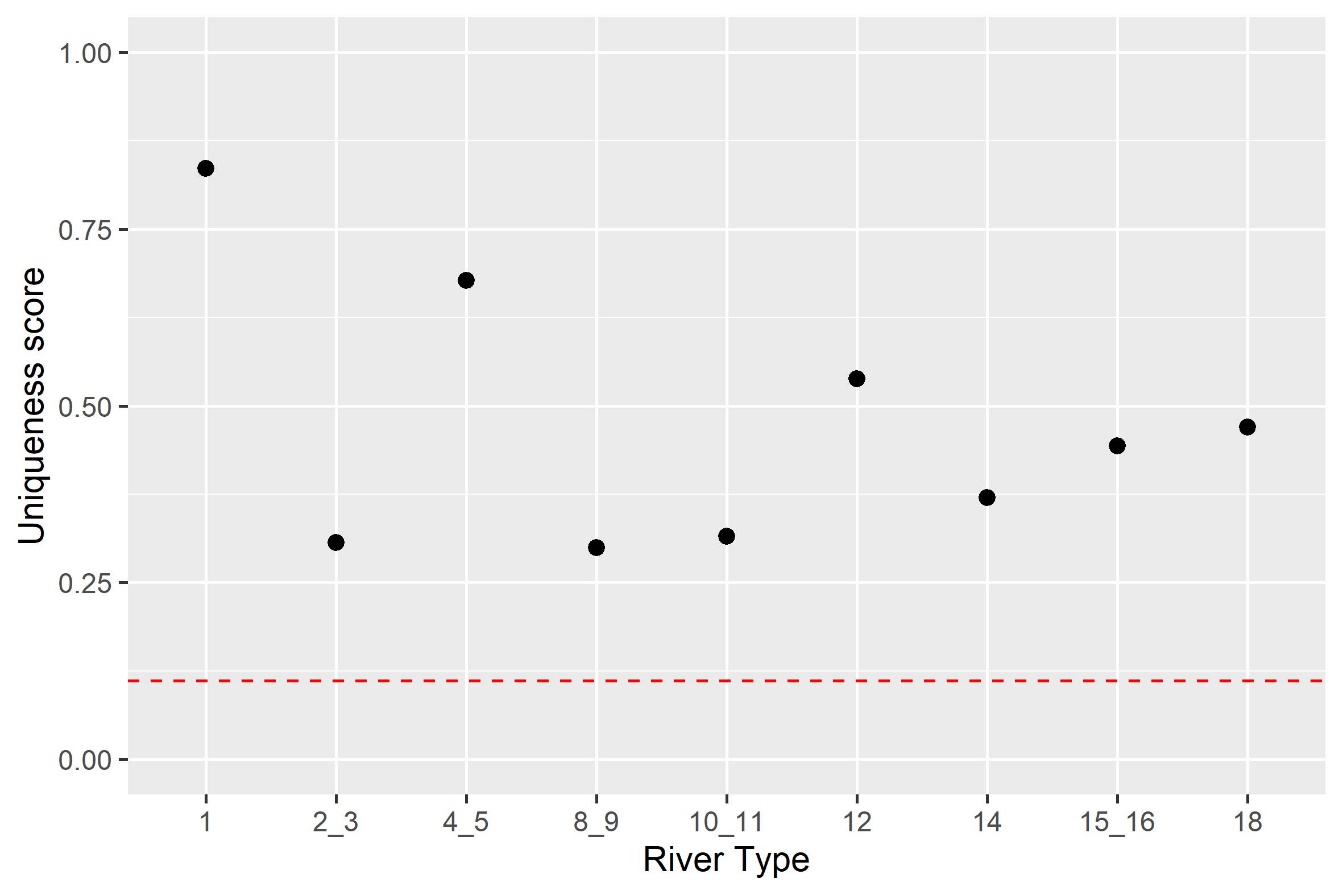


Figure 5: Uniqueness scores of all typical assemblages.

A Nonmetric multidimensional scaling (NMDS) of the data shows several plausible patters (Figure 6). It is based on a custom distance metric which is based on the degree of overlap between assemblages. Most similar are RT10\_11 and RT18, which have many close sites in southern France. The three assemblages RT2\_3 (lowland siliceous), 8\_9 (mid-altitude siliceous), and 15\_16 (high altitude) are approximately as similar to each other, as they are to RT10\_11 and RT18. Four assemblages are more distinct from this central cluster. RT1 (very large streams) has the least overlap with the other assemblages. RT4\_5 are lowland calcareous or mixed rivers takes a middle position between lowland siliceous (2+3) and mid-altitude calcareous or mixed (10\_11) and is very distinct from high altitude rivers and mid-altitude organic and siliceous (RT12) rivers. In general, one can see a gradient from higher to lower altitudes from the upper right to the lower left/middle of the plot. It is surprising to see how similar RT12 and RT15\_16 seem to be, given the large geographic distance between sampling sites (see [maps in GetReal Drive](https://drive.google.com/drive/folders/1xi3m9ZV38TjDYhaced7PEztP3rQoFdmY)).

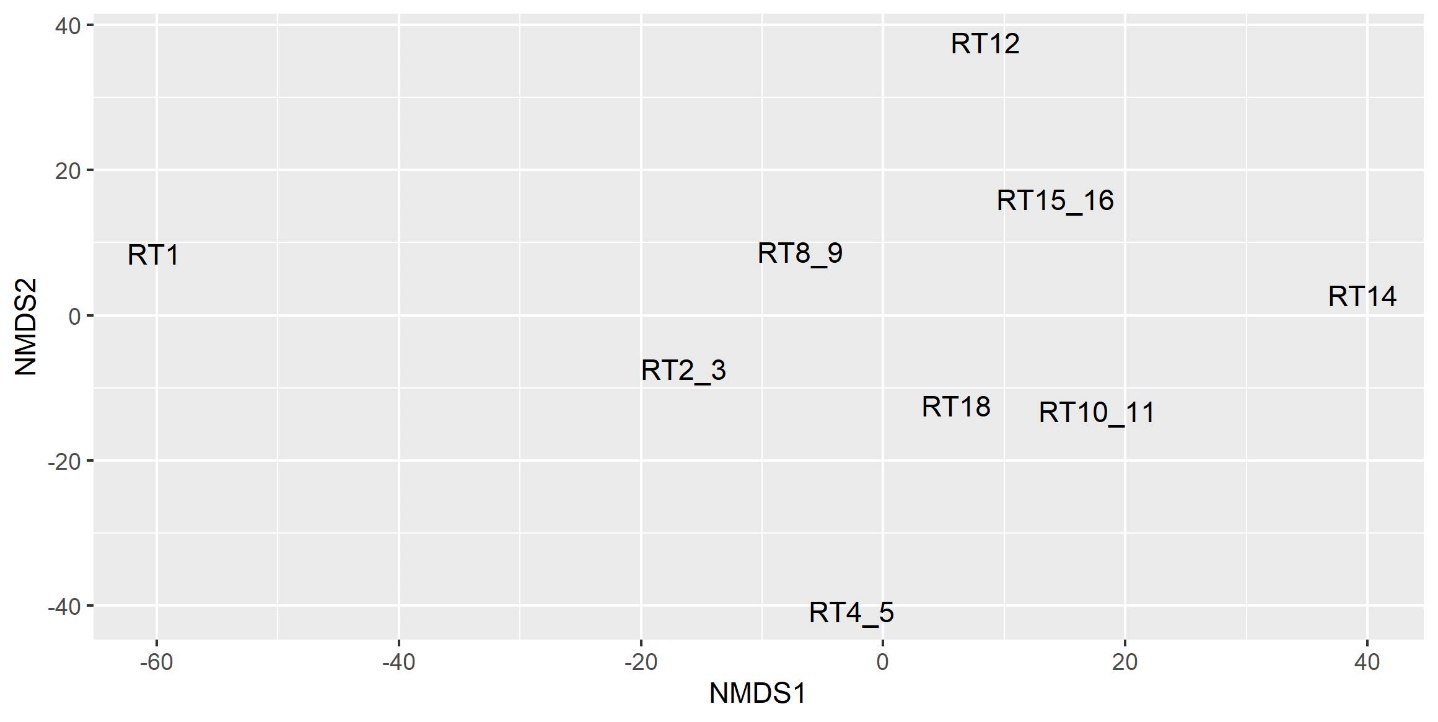


Figure 6: Non-metric multidimensional scaling of Jaccard distances between typical assemblages

## Typical assemblages

Below we list the different taxa that belong to the typical assemblages.

### RT01 - Very Large rivers

*Ampullaceana balthica, Ancylus, Borysthenia naticina, Caenis, Chelicorophium, Corbicula, Dikerogammarus, Esperiana esperi, Hippolyte desmaresti, Isogenus nubecula, Jaera, Microcolpia daudebartii, Obesogammarus obesus, Potamopyrgus*

### RT02\_03 - Lowland siliceous, very small -large

*Alainites muticus, Ameletidae, Ancylus, Baetis, Caenis, Dicranota, Ecdyonurus, Elmis, Gammarus, Hydropsyche, Lepidostoma, Limnius, Oulimnius, Peregriana peregra, Pisidium, Polycentropus, Potamopyrgus, Rhithrogena, Rhyacophila, Serratella ignita*

### RT04\_05 - Lowland calcareous, very small -large

*Ancylus, Asellus, Baetis, Caenis, Calamoceras, Cataclysta, Elmis, Elophila, Erpobdella, Gammarus, Glossiphonia, Harnischia, Hydropsyche, Ilyocoris, Metreletus balcanicus, Micropsectra, Odhneripisidium moitessierianum, Oulimnius, Peregriana peregra, Pisidium, Potamopyrgus, Ranatra, Serratella ignita, Sphaerium, Unionidae, Viviparidae*

### RT08\_09 - Mid-altitude siliceous, very small-large

*Alainites muticus, Amphinemura, Ancylus, Baetis, Dicranota, Ecdyonurus, Elmis, Ephemera, Gammarus, Hydraena, Hydropsyche, Isoperla, Leuctra, Limnius, Odontocerum, Oreodytes sanmarkii, Polycentropus, Protonemura, Rhithrogena, Rhyacophila, Serratella ignita*

### RT10\_11 - Mid-altitude, calcareous or mixed, very small -large

*Ancylus, Baetis, Chironomidae, Ecdyonurus, Elmis, Ephemera, Esolus, Gammarus, Hydraena, Hydropsyche, Hydroptila, Leuctra, Limnius, Micronectidae, Odontocerum, Pisidium, Rhithrogena, Rhyacophila, Riolus, Serratella ignita, Simuliidae*

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

### *Ameletus inopinatus, Amphinemura, Baetis, Dicranota, Diura nanseni, Elmis, Heptagenia, Isoperla, Kageronia fuscogrisea, Lepidostoma, Leptophlebia, Leuctra, Nigrobaetis niger, Polycentropus, Rhyacophila*

### RT14 – Highland, siliceous, incl. organic (humic)

### *Baetis, Chironomidae, Empididae, Ephemerella, Heptagenia, Hydraenida, Hydropsyche, Leuctra, Limnephilidae, Limoniidae, Oligochaeta, Oreodytes sanmarkii, Rhyacophila, Serratella ignita, Simuliidae*

### RT15\_16 – Highland, calcareous/mixed and glacial rivers

### *Baetis, Chironomidae, Ecdyonurus, Elmis, Epeorus, Hydraena, Hydropsyche, Isoperla, Leuctra, Limnephilidae, Limoniidae, Nemoura, Oligochaeta, Protonemura, Pseudodiamesa branickii, Rheosmittia spinicornis, Rhithrogena, Rhyacophila, Serratella ignita, Simuliidae, Stilocladius montanus*

### RT18 - Mediterranean mid-altitude, medium-large, perennial

*Baetis, Caenis, Calopteryx, Chironomidae, Ecdyonurus, Elmis, Empididae, Ephemerella, Esolus, Gammarus, Hydropsyche, Hydroptila, Hydroscapha, Lestidae, Leuctra, Limnius, Oligochaeta, Onychogomphus, Oulimnius, Potamopyrgus, Procambarus, Psychomyia, Rhyacophila, Riolus, Serratella ignita, Simuliidae, Sympecma*