# Derivation of typical invertebrate assemblages

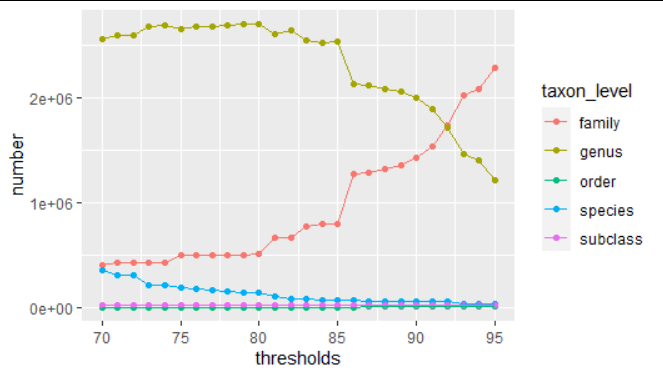
Please note: everything below refers only to the invertebrate data as the diatoms are still in a taxonomical limbo. However, I plan to use the same procedure on them.

## What is the optimal taxonomical level?

One result of the last progress review was that the taxa that make up the typical assemblages (TA) need not be on the same taxonomical level (species, genus, family, …). Instead, we can 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, I used a hierarchical approach. First, I removed all observations from Phyla and Classes that were not present in all data sets. I 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).

Then I looked at each subsequent level (subclass, order, family, genus) and calculated the percentage of observations within that level that is 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 percent at the order level. Now given a threshold X which is to be determined, I would call a taxon optimally represented at a certain taxonomical level if less than X% are represented by this and lower levels. For the example of *Lepidoptera*, I would 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 I searched for noticeable patterns in the data (see 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, I 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 (Dufrêne & Legendre, 1997; Cáceres, & Legendre, 2009) is the weighted product of A and B (see eqn. below)

Here N is the total number of sites, is the number of sites belonging to stream type p, *n* is the total number of occurrences of the focal species, and the number of occurrences of the focal species in sites of type p. The 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 9999 permutations.

Here, we are not interested in indicator species for each community, but in 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, serves to highlight 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. I used the following rules:

Species where considered typical if B > 0.25 or B > 0.2 and p < 0.05 or A > 0.8

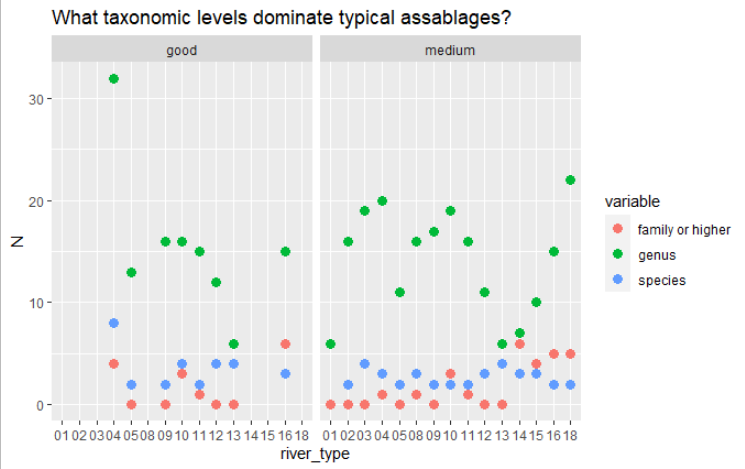
Genera where considered typical if B > 0.5 or B > 0.33 and p < 0.05 or A > 0.95

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 or sensitivity analysis used to define these thresholds. Such procedures would require optimization criteria, but I am 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. I 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”.

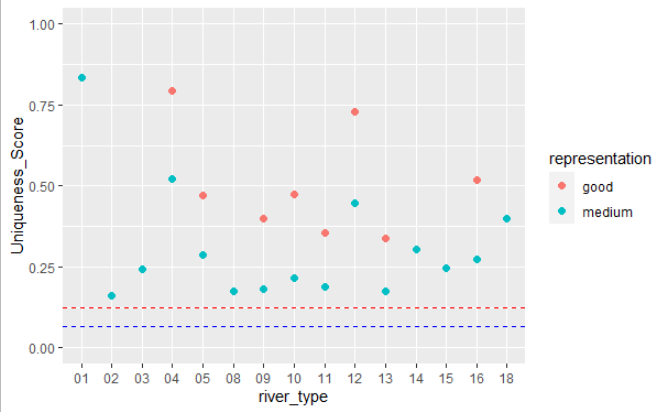
## What do the typical assemblages look like?

In all TA, genus is the prevalent taxonomic level and, in most cases, species follows suit (see Figure 2). However, some assemblages have more families than species (i.e. RT10, RT14, RT 15, RT16, RT18). This is more common in assemblages whose river type’s representation is only medium. The total number of taxa per assemblage changes when medium river types are included, most notably in RT04 which drops from 44 to 24.



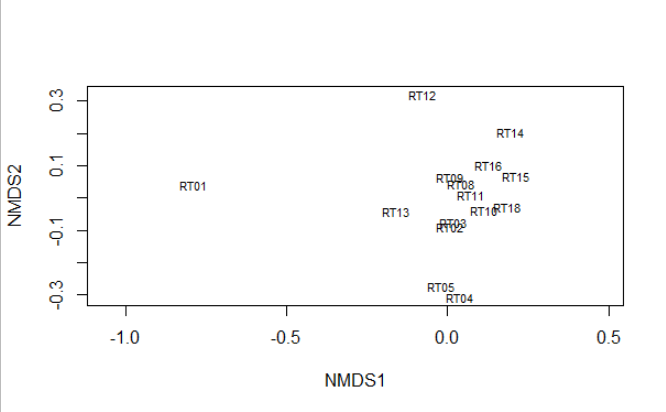
*Figure 2: Numbers of taxa on each taxonomical level for all typical assemblages*

We can express the uniqueness of a TA with the following score: Each taxon receives a taxon uniqueness score which is 1 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 3. The dashed horizontal lines indicate the minimum score for each representation level. In general, these scores are better in TAs of well-represented river types but within each representation level, there are strong variations.



*Figure 3: Uniqueness scores of all typical assemblages.*

An NDMS of the Jaccard Distance between TAs shows several plausible patterns (see Figure 4). Very similar assemblages are found in RT02 (Lowland siliceous, medium-large) and RT03 (Lowland siliceous, very small-small), RT04 (Lowland calcareous, medium-large) and RT05 (Lowland calcareous, very small-small), RT08 (Mid-altitude siliceous, medium-large) and RT09 (Mid-altitude siliceous, very small-small), and somewhat more surprising RT10 (Mid-altitude, calcareous or mixed, medium-large) and RT18 (Mediterranean mid-altitude, medium-large, perennial). RT01 (Very Large rivers) has the most distinct assemblage.



*Figure 4: Non-metric multidimensional scaling of Jaccard distances between typical assemblages*

## Typical assemblages

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

### RT01 - Very Large rivers

*Ampullaceana balthica, Borysthenia naticina, Esperiana esperi, Hippolyte desmaresti, Microcolpia daudebartii, Obesogammarus obesus, Chelicorophium, Corbicula, Dikerogammarus, Heptagenia, Jaera, Potamopyrgus*

### RT02 - Lowland siliceous, medium-large

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

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

*Alainites muticus, Euglesa casertana, Peregriana peregra, Serratella ignita, Amphinemura, Ancylus, Baetis, Brillia, Dicranota, Ecdyonurus, Elmis, Gammarus, Hydropsyche, Leuctra, Limnius, Micropsectra, Oulimnius, Pisidium, Polycentropus, Potamopyrgus, Protonemura, Rhithrogena, Rhyacophila*

### RT04 - Lowland calcareous, medium-large

*Odhneripisidium moitessierianum, Peregriana peregra, Serratella ignita, Ancylus, Asellus, Baetis, Bithynia, Caenis, Cataclysta, Cricotopus, Elmis, Erpobdella, Gammarus, Glossiphonia, Harnischia, Helobdella, Hydropsyche, Micropsectra, Microtendipes, Oulimnius, Pisidium, Potamopyrgus, Sphaerium, Viviparidae*

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

*Peregriana peregra, Serratella ignita, Asellus, Baetis, Calamoceras, Ephemera, Erpobdella, Gammarus, Glossiphonia, Hydropsyche, Oulimnius, Pisidium, Potamopyrgus*

### RT08 - Mid-altitude siliceous, medium-large

*Alainites muticus, Oreodytes sanmarkii, Serratella gnita, Ancylus, Baetis, Dicranota, Ecdyonurus, Elmis, Ephemera, Esolus, Gammarus, Hydraena, Hydropsyche, Isoperla, Leuctra, Limnius, Odontocerum, Polycentropus, Rhyacophila, Simuliidae*

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

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

### RT10 - Mid-altitude, calcareous or mixed, medium-large

*Propappus volki, Serratella ignita, Ancylus, Baetis, Caenis, Ecdyonurus, Elmis, Ephemera, Esolus, Gammarus, Hydraena, Hydropsyche, Hydroptila, Leuctra, Limnius, Oulimnius, Pisidium, Radix, Rhithrogena, Rhyacophila, Riolus, Chironomidae, Limoniidae, Simuliidae*

### RT11 – Mid-altitude, calcareous or mixed, very small-small

*Serratella ignita, Torleya major, Baetis, Ecdyonurus, Elmis, Ephemera, Esolus, Gammarus, Hydraena, Hydropsyche, Leuctra, Limnius, Odontocerum, Pisidium, Protonemura, Rhithrogena, Rhyacophila, Riolus, Simuliidae*

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

*Ameletus inopinatus, Diura nanseni, Nigrobaetis niger, Amphinemura, Baetis, Elmis, Ephemerella, Heptagenia, Isoperla, Lepidostoma, Leptophlebia, Leuctra, Polycentropus, Rhyacophila*

### RT13 – Mid-altitude, organic and calcareous/ mixed

*Alainites muticus, Ampullaceana balthica, Peregriana peregra, Serratella ignita, Baetis, Caenis, Ecdyonurus, Elmis, Limnius, Rhithrogena*

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

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

### RT15 – Highland, calcareous/mixed

*Nilotanypus dubius, Propappus volki, Serratella ignita, Baetis, Ecdyonurus, Gammarus, Hydraena, Hydropsyche, Leuctra, Nemoura, Protonemura, Rhyacophila, Riolus, Chironomidae, Limnephilidae, Limoniidae, Simuliidae*

### RT16 – Glacial Rivers

*Serratella ignita, Stilocladius montanus, Baetis, Ecdyonurus, Elmis, Epeorus, Hydraena, Hydropsyche, Isoperla, Leuctra, Limnius, Nemoura, Odontocerum, Perlodes, Protonemura, Rhithrogena, Rhyacophila, Chironomidae, Limnephilidae, Limoniidae, Oligochaeta, Simuliidae*

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

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

References

Dufrêne, M., and P. Legendre. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecological Monographs 67:345–366.

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