

Derivation of Typical Assemblages

Jonathan Jupke

Oktober 02, 2020

Contents

1	Introduction	3
2	Harmonizing taxa names	3
3	What is the optimal taxonomic level?	3
4	Can we represent the stream types with our data?	5
5	Notes for Traits	5
	References	5

1 Introduction

Here we describe the methods we used to derive typical assemblages of macro-invertebrates and diatoms for selected river types across Europe. We also describe and briefly discuss the results.

2 Harmonizing taxa names

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 also 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:

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 (Rimmet *et al.* 2019)
5. The website algaebase.org (Guiry 2020)
6. The global biodiversity information platform (gbif) (GBIF.org 2020)

The harmonization of macro-invertebrate data required less effort, and was achieved with gbif (GBIF.org 2020) through the taxize R package (Scott Chamberlain & Eduard Szocs 2013).

3 What is the optimal taxonomic level?

One result of the last progress review for Get Real (held on the 29.04.2020) was that taxa in the TA should be included on their respective optimal taxonomic level instead of using one level (e.g. Genus) for all. *Oligochaetes*, for example, are usually only determined to subclass level, which should not prevent them to be part of a TA in the case that they are common in a given river type. Thus, the question arises: given a data set, what is the optimal taxonomic 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, 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 that are 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, which is to be determined, we would call a taxon 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 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 threshold. Observations that were missed by this procedure, e.g. observations of *Chironomidae* at the family level, were included at their respective level.

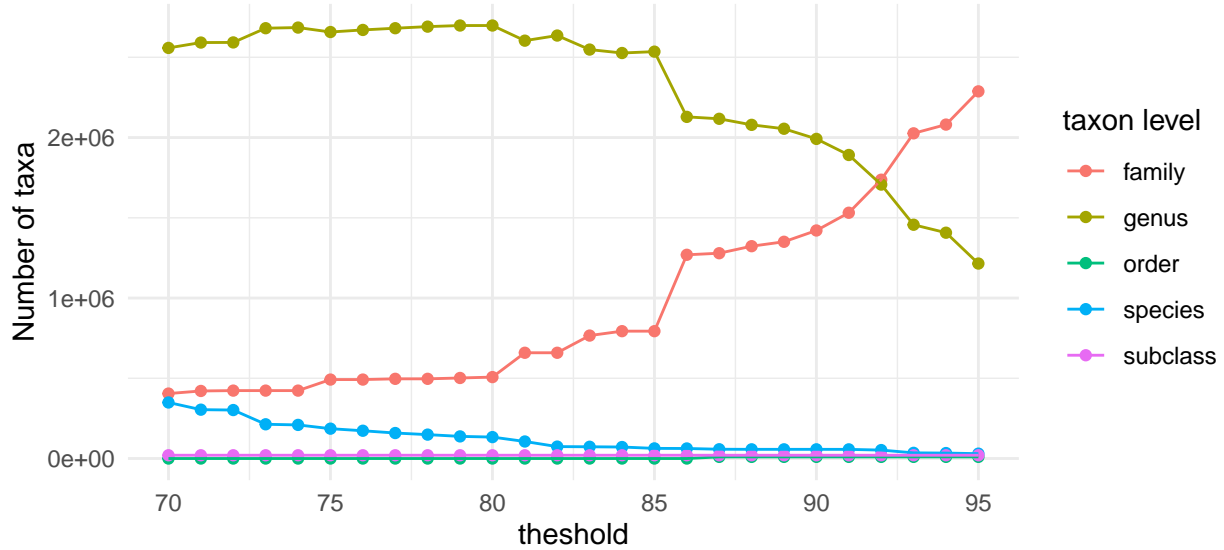


Figure 1: Number of taxa represented at a given taxonomic level as function of the threshold value

For the diatoms, we employed 75% as threshold, because Gomphonema, which is the fourth most common genus in our data set, had 81.43% observations at the species level. The taxonomic resolution was higher than in the macroinvertebrate data and the lowest resolution is the genus level.

Table 1: The ratings for all river types for macro-invertebrates and diatoms

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

4 Can we represent the stream types with our data?

We determined visually whether a our data set contained enough sampling sites in a given river type to derive meaningful TAs. The degree of representation for river type 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 which 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 both macro-invertebrates and diatoms are shown in table 1.

For each river type x taxon combination, you can find maps with the associated sampling sites [for macro-invertebrates](#)

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. 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.

5 Notes for Traits

CWM with B value then RR-VGLM (glm RDA)

References

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.
- 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.
- 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.
- Scott Chamberlain & Eduard Szocs. (2013). Taxize - taxonomic search and retrieval in r. *F1000Research*.