Freshwater Biology (2014) 59: 1551-1557. -- DOI: 10.1111/fwb.12369

Combining taxon-by-trait and taxon-by-site matrices for analysing trait patterns of macroinvertebrate communities: a rejoinder to Monaghan & Soares (2014)

Dénes Schmera[1,2], János Podani[3,4],Tibor Erős[2] and Jani Heino[5,6,7]

[1]Section of Conservation Biology, Department of Environmental Sciences, University of Basel, Basel, Switzerland

[2]Balaton Limnological Institute, Centre for Ecological Research, Hungarian Academy of Sciences, Tihany, Hungary

[3]Department of Plant Systematics, Ecology and Theoretical Biology, Institute of Biology, L. Eötvös University, Budapest, Hungary

[4]Ecology Research Group of the Hungarian Academy of Sciences, Budapest, Hungary

[5]Natural Environment Centre, Finnish Environment Institute, Oulu, Finland

[6]Deparment of Biology, University of Oulu, Oulu, Finland

[7]Correspondence: Jani Heino, Natural Environment Centre, Finnish Environment Institute, P.O. Box 413, FI-90014 Oulu, Finland. E-mail: jani.heino@environment.fi

Running head: How to combine taxa-by-trait and taxa-by-site matrices?

Keywords: macroinvertebrates, trait-based analyses, weighting

Summary

1. Monaghan & Soares (2014) suggested that combining traits with log-transformed abundance of taxa may cause anomalies in analyses of stream macroinvertebrate communities. While they addressed an important issue in stream ecology, here we present an opposite view. To identify the causes of these contrasting opinions, we carefully examined the examples provided by Monaghan & Soares (2014) and demonstrated how traits can be weighted by the presence, abundance and log-transformed abundance of the taxa in a meaningful way.

2. We found that Monaghan & Soares (2014), following other authors, use the term 'weighting' differently from classical papers of stream ecology. The general practice is to calculate the sum of trait values multiplied by the abundance of each taxon and divide it by the total invertebrate abundance to get a community-level trait value. In contrast, Monaghan & Soares (2014) did not perform the final division and consequently did not get a standardized community-level trait value. It follows that the term "weighting" is used with different meanings in stream ecology, and ecologists should keep these differences in mind.

3. We agree with Monaghan & Soares (2014) that the addition of log-transformed data is equivalent to multiplication on an arithmetic scale. However, we disagree that this provides an inconsistent scaling that confounds quantitative analyses. Using example data sets, we illustrate how trait-based data analysis can be preformed in community ecology in a meaningful way.

Introduction

In a recent paper, Heino, Schmera & Erős (2013) provided an overview of trait patterns of stream communities from a macroecological perspective. In this paper, reference was made to Gayraud *et al*. (2003), who showed that abundance-weighted traits were less powerful than presence-absence weighted traits in discriminating communities under different degrees of human impacts, and to Statzner & Beche (2010), who therefore advocated the use of the second type of weighting for practical biomonitoring due to logistic constrains (e.g. sorting of qualitative samples makes assessment programmes more cost effective). Heino *et al*. (2013) pointed out that these findings contradict the results of taxon-based analyses which suggest that, if taxa are weighted by their abundance, then communities are better separated in relation to environmental variation than when taxa are weighted only by their presence. Based on an overview of the literature, Heino *et al*. (2013) concluded that results of presence- and abundance-based analyses should be evaluated carefully when examining traits of organisms, because differences among studies can reflect both methodological (i.e. handling of data) and real ecological differences (see p. 1549 in Heino *et al*., 2013). More recently, Monaghan & Soares (2014) stated that (1) Heino *et al*. (2013) identified the weak explanatory power of abundance data as a major limitation of macroinvertebrate trait analysis and that (2) the log-transformation of abundance data may cause anomalies in trait-based analyses. We disagree with both conclusions, because (1) Heino *et al*. (2013) did not actually state this (see above) and because (2), in our view, log-transformation of abundance data in trait-based analyses can also be meaningful. To reveal the causes of these differing views, we go through the examples provided by Monaghan & Soares (2014) and examine how traits can be weighted by the presence, abundance and log-transformed abundance of the taxa. To do this, first we define the terminology used here, comment on the approach of Monaghan & Soares (2014) and show how this procedure should be performed.

Terminology

Characterizing taxa by their traits is a challenging task in ecology. A research team (Stazner *et al*., 1994) studying the Upper Rhone River in France made a seminal contribution by determining which traits can be used and how to examine macroinvertebrate communities from a ’functional’ perspective. High heterogeneity of organisms, remarkable variation within taxa, gaps and uncertainties in our knowledge led them to adopt a “*fuzzy coding”* system (Chevenet *et al*., 1994). Fuzzy coding is based on expert opinion and assigns an integer score, ranging from 0 (no affinity) to an arbitrary maximum, often 5 (high affinity), to express the relationship of a taxon to a particular trait. To improve comparability, scores are often standardized by the total number of scores within a taxon (Usseglio-Polatera *et al.*, 2000, Bady *et al.*, 2005, Dolédec *et al.*, 2006).

Information on taxa characterized by different traits can be summarized by a taxon-by-trait matrix (**A**) of size s × t, in which rows represent taxa and columns correspond to traits describing a well-defined character (e.g. the feeding habit) of macroinvertebrates. More formally,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & & a_{1t} \\ a_{21} & a_{22} & & a_{2t} \\ & & a_{ij} & \\ a_{s1} & a_{s2} & & a_{st} \end{bmatrix}$$

Each entry in the matrix (a_{ij}) is determined such that $\sum_{j=1}^{t} a_{ij} = 1$ for every row i (see Bady *et al.*, 2005). In a sense, the a_{ij} values express the relative importance of trait j for taxon i or, in other words, the relative affinity of taxon i to trait j. For example, if a given taxon acts exclusively as a shredder, it cannot be a predator or anything else: only one value can be 1, and all the others are zero (see also Chevenet *et al.*, 1994; Dolédec *et al.*, 2006). Other possibilities, with $0 < a_{ij} < 1$ are shown in the upper part of Table 2 of Monaghan & Soares (2014). Note, however, that zero values are missing from that matrix, which takes a mathematically correct form as given below:

$$\mathbf{A} = \begin{bmatrix} 0.17 & 0 & 0.5 & 0.33 & 0 \\ 0 & 0.75 & 0.25 & 0 & 0 \\ 0 & 0.2 & 0 & 0.2 & 0.6 \\ 0 & 0.5 & 0 & 0 & 0.5 \\ 0 & 0.4 & 0.6 & 0 & 0 \\ 0 & 0 & 0 & 0.75 & 0.25 \\ 0.25 & 0 & 0.75 & 0 & 0 \end{bmatrix}$$

Sampling sites characterized by different taxa can be summarized by a taxon-by-site matrix (**B**) of size s × z, in which rows represent taxa, and the columns correspond to sampling sites. Each entry b_{ik} in **B** is the abundance of taxon i at site k:

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & & b_{1z} \\ b_{21} & b_{22} & & b_{2z} \\ & & b_{jk} & \\ b_{s1} & b_{s2} & & b_{sz} \end{bmatrix}$$

The top left part of Table 2 in Monaghan & Soares (2014) displays such a matrix for seven taxa and a single site, so it is in fact a column vector, abbreviated as **b**:

$$\mathbf{b} = \begin{bmatrix} 80 \\ 50 \\ 100 \\ 20 \\ 40 \\ 30 \\ 20 \end{bmatrix}$$

Comments on the approach of Monaghan & Soares (2014)

Monaghan & Soares (2014) write that, if standardized trait values are used "as a non-linear weighting for trait descriptions", then the "practitioners should be aware of the potential of inconsistent weighting". This causes some uncertainty, because the term ‘weighting’ in Monaghan `& Soares (2014) is different from that used by Charvet *et al.* (2000) and Statzner *et al.* (2001). Charvet *et al.* (2000) and Statzner *et al.* (2001) summed the trait values multiplied by the abundance of each taxon and divided by the total invertebrate abundance to get a community-level trait value. In contrast, Monaghan & Soares (2014), following Archaimbault *et al.* (2005) and Larsen & Ormerod (2010), did not perform the final division, only summed the trait values multiplied by the abundance of each taxon. In other words, some studies weight by proportions, whereas others by abundance. It follows that the term "weighting" has been used with at least by two different meanings in stream ecology. Although the proper measurement of the overall community trait values is still debated (Ricotta & Moretti, 2011), we argue that, from a statistical point of view, *weighting* means that some elements have more impact on the result than other elements in the same set. Stream ecologists should be aware of these two variants when comparing different results.

Monaghan & Soares (2014) exemplify this inconsistent weighing by back-transformation, and emphasise that the "addition of log-transformed data is equivalent to multiplication on an arithmetic scale" (point 2 in their Summary). From a mathematical point of view we agree, but note that accepting this point would exclude all kinds of nonlinear transformations (square-root, logarithmic, etc.) from the toolkit of statistical data analysis. Nonlinear transformations modify the original data structure in a meaningful way and, if properly used, facilitate significance tests by decreasing skewness in the data (Elliot, 1977; Podani, 2000; Legendre & Legendre, 2012). Furthermore, community ecologists very often use log transformation for 'equalization' purposes, that is, to balance the importance of variables (species, or other taxa) before cluster analysis or ordination. The result of this operation is that species with highly unequal abundance will be almost equally influential in multivariate analysis. Non-linear transformation can be conceived as a parameterized series, with raw abundances at one end and presence-absence data at the other (e.g., Clymo transformation, see Podani, 2000). In any case, it is fundamental that all data are transformed by the same method *before* any other calculations are performed. Since the scale of data (i.e. linear vs. nonlinear) is selected completely arbitrarily by the investigator (pH is measured mostly at a nonlinear scale, while individuals are counted mostly at a linear scale), changing from a linear to a nonlinear scale, or *vice versa*, is acceptable. Finally, we do not know of any situation when back-transformation to the original scores would be necessary, so that the warning by Monaghan & Soares (2014) is mainly of academic significance.

Combining traits with the abundance of a single taxon

Monaghan & Soares (2014) state that "the abundance of taxa represented by respective trait classes is obtained by multiplying the total abundance by the proportionate frequencies of the traits". Moreover, their Table 1 exemplifies the situation and shows how this procedure runs with the feeding habits of larvae of the trichopteran *Psychomyia* using raw abundance and standardized trait values (Table 1a of Monaghan & Soares, 2014), using the log-transformation of the product (their Table 1b,c), and using the log-transformed abundance multiplied by the standardized trait values (their Table 1d). However, Monaghan & Soares (2014) give no information about the purpose of these calculations.

Basically, this table is unnecessary because the utility and the effects of log transformation of community data is only apparent when several species appear in the data simultaneously. For a single species, we cannot see why these operations are important. Multiplying a standardized trait value by the raw abundance of that single species: (a) may be correct but not useful, because the standardized trait values of a species provide sufficient information about the (feeding habits of the) taxon anyway. Operations (b-c) are illogical (but are in fact identical) because the raw abundances are multiplied first by the relative frequency of the trait while log transformation is undertaken only subsequently. As mentioned previously, handling the data this way is not correct, because transformation should be done first in order to ensure consistency in subsequent analyses of the data.

Combining traits with the abundance of several taxa in a community

In examining communities, Monaghan & Soares (2014) suggest that "the relative abundance of trait classes at assemblage level is given by the sum of the trait frequencies of respective taxa. This can be calculated by multiplying the respective taxon abundances by their corresponding proportionate frequencies, followed by the summation of the resultant trait abundances" and exemplify this and some log-transformed versions in their Table 2. They write in their Summary that, compared to other versions, the summation of the product of log-transformed abundance and standardized trait values "represents a non-linear abundance weighting of trait frequencies, as opposed to an expression of trait abundance *per se*, because the addition of logarithmic data is equivalent to multiplication on an arithmetic scale, summing this abundance-weighted frequencies provides an inconsistent scaling of trait abundance that might confound quantitative comparison" .

This also risks confusion because Monaghan & Soares (2014) use the term "*relative* abundance of trait classes at assemblage level" (italics ours) to specify the trait value of the community (*community traits* in Heino *et al*., 2013). It is not clear why is this “relative”, which generally means a comparison to some standard (or total). Another source of confusion, compared to the methodology of Charvet *et al*. (2000) and Statzner *et al*. (2000), is that the "relative abundance of trait classes at the assemblage level" is the sum of products ("This can only be calculated by multiplying the respective taxon abundances by their corresponding proportionate frequencies, followed by the summation of the resultant trait abundances") without dividing by the total abundance of invertebrates. In the following, we show how to combine standardized trait values of several taxa in a community to calculate community-level trait values, where weighting includes a division.

Let us start with the taxa-by-traits and taxa-by-sites matrices provided in Table 2 of Monaghan & Soares (2014) and, for the time being, focus only on the presence/absence of the taxa and on the trait "fine sediments". The data suggest that *Psychomyia* has a standardized trait value of 0.17 and *Agapetus* a value of 0.25. We should not forget, however, that the other five taxa (*Sericostoma*, *Polycentropus*, *Odontocerum*, *Drusus* and *Hydropsyche*) have a standardized trait value of 0 (no affinity for feeding on fine sediment). It follows that our community contains five taxa with a standardized trait value of 0, one taxon (*Psychomyia*) with a standardized trait value of 0.17 and one (*Agapetus*) with a value of 0.25. We can illustrate this on a frequency histogram where the horizontal axis shows the standardized trait values while the vertical axis corresponds to their frequencies (top-left subplot in Fig. 1). The mean of these values gives the community trait value (indicated by the arrow). How can we interpret this value? It tells us that if we consider only the presence of the taxa then the community is represented by an average value of 0.06 (= [0.17 + 0.25 + 5 x 0]/7) with respect to the trait ‘feeding on fine sediment’. In other words, members of the community, based on the presence of taxa, have a low affinity for feeding on fine sediment. It follows that weighting by the presence of taxa means that we give more importance to those elements of the set (to those standardized trait values of the observed trait values), which are most frequently represented by the taxa in the community. Examining the issue from a taxonomic point of view, we give equal importance to each taxon. We can do the same procedure for the other traits (top row in Fig. 1). We should note that it is really beneficial that our calculation is based on standardized trait values, because the sum of community trait values equals 1. It follows that we can clearly interpret which function is the most characteristic of the community (in this case, feeding as a scraper).

The logic of using abundances is similar to that used for presence: we count the frequencies of standardized trait values and use these frequencies to calculate community trait values. However, compared to using the *presence of the taxa*, here we use the numerical abundance of *individuals* to weight the standardized trait values. In other words, we have 240 individuals (50 *Sericostoma*, 100 *Polycentropus*, 20 *Odontocerum*, 40 *Drusus* and 30 *Hydropsyche*) showing the standardized trait value 0 for fine sediment trait, we have 80 individuals (80 *Psychomyia*) showing the standardized trait value 0.17 and, finally we have 20 individuals (20 *Agapetus*) with a standardized trait value 0.25 (bottom-left subplot in Fig. 1). The community trait value is the weighted average of these values, and the measure is usually known as community-weighted mean trait value (Ricotta & Moretti, 2011). In the example, (80x0.17 + 20x0.25 + 240x0)/340 = 0.055. For the other four traits we obtain 0.245, 0.269, 0.202 and 0.228. We should note that the sum of these community trait values is 1 and that the community trait value corresponds to the mean standardized trait value of an *individual* in the community. The latter calculations (without the final division) are missing from Table 2 of Monaghan & Soares (2014), although the weighted values are presented in row a (for example, 80 x 0.17 + 20 x 0.25 = 18.6).

Log-transformation of abundance in trait-based community analyses

What happens if we use the log-transformed abundance for weighting? First, we transform the abundance of each taxon. While the original column vector contains abundances of 80, 50, 100, 20, 40, 30 and 20, the log-transformed column vector contains abundances of ln(80) = 4.38, ln(50) = 3.91, ln(100)= 4.61, ln(20) = 2.99, ln(40) = 3.69, ln(30) = 3.40 and ln(20) = 2.99. Note that, in order to make the example comparable with Monaghan & Soares (2014), we use also $\log_e$ transformation (abbreviated to ln), although we suggest that logarithms to the base of 10 are easier to interpret (see below.) We then use these log-transformed abundances, instead of the original abundances. To calculate the frequency of standardized trait value of 0 for fine sediment, we should sum the frequency of taxa with 0 standardized trait value for fine sediment: ln(50) + ln(100) + ln(20) +ln(40) + ln(30) = 18.6. The transformed abundance of standardized trait value of 0.17 for fine sediment is ln(80)= 4.38. Finally, the transformed abundance of the standardized trait value 0.25 equals ln(20) = 2.99. The community trait value can be calculated by the mean of these values [(18.6 x 0 + 4.38 x 0.17 + 2.99 x 0.25) / (18.6 + 4.38 + 2.99) = 0.057. We can calculate frequencies and community trait values for the other traits in similar manner. We should note that the sum of these community trait values equals to 1 (Fig. 1), as required (see also Usseglio-Polatera *et al*., 2000; Dolédec *et al*., 2006).

There are, however, some problems with such analyses. First, this transformation is not able to handle 0 abundances (in comparing multiple communities, some species may be absent from several sites) because ln(0) is mathematically not interpretable. Therefore, we suggest the use of the log(x+1) transformation, as used by Statzner *et al*. (2001), Gayraud *et al*. (2003), Dolédec *et al*. (2006) and also suggested by Monaghan & Soares (2014). Moreover, although the base of logarithms is again an arbitrary decision, we suggest a $\log_{10}(x+1)$ transformation to facilitate interpretation. To demonstrate this, we start with the same taxon-by-trait matrix discussed before, but our community contains only three taxa: 1000 individuals of *Psychomyia*, 100 individuals of *Sericostoma* and 10 individuals of *Polycentropus* (other four taxa are represented by 0 individuals). Here, we do not intend to discuss all details of the analyses, but only emphasize some important differences among different weighting schemes. Regarding the trait ‘scraper’, we have one taxon (*Polycentropus*) showing a standardized trait value of 0, one showing a value of 0.25 (*Sericostoma*), and one a value of 0.5 (*Psychomyia*). Consequently, the community trait value in presence-weighted analyses equals 0.25 (Fig. 2). In abundance-weighted analyses, however, there are 10 individuals with a standardized trait value of 0 (10 individuals of *Polycentropus*), 100 individuals with a standardized trait value of 0.25 (100 individuals of *Sericostoma*) and 1000 individuals with a standardized trait value of 0.5 (1000 individuals of *Psychomyia*). The community trait value equals to 0.47, very close to the value of *Psychomyia*, the dominant species in the assemblage. In log-transformed abundance-weighted analyses [using a $\log_{10}(x+1)$ transformation], the standardized trait values 0, 0.25 and 0.5 will be multiplied by $\log_{10}(11)$=1.04,

$\log_{10}(101) = 2.00$ and $\log_{10}(1001) = 3.00$, respectively (Fig. 2). We should recognize that the $\log_{10}$-transformed abundance expresses the magnitude of the number of individuals in the decimal system. The community trait value will be lower than above, 0.33, reflecting our decision to give less weight to the most common species. It is interesting to see how the community trait value of the scraper trait increases as we change from presence-weighting, through log-transformed abundance weighting to abundance weighting (Fig. 2). All of these suggest that, although Monaghan & Soares (2014) are right in stating that the addition of log-transformed data is equivalent to multiplication on the arithmetic scale, the interpretation of logarithmic transformation is straightforward. Therefore, in our view, papers using log-transformed abundances in trait-based analyses provide not only a useful indicator of environmental conditions (as acknowledged by Monaghan & Soares, 2014), but also a clear and easily-interpretable mathematical procedure.

Acknowledgments

We thank David G. Angeler and an anonymous reviewer for their comments on the manuscript. The research was supported by the Hungarian Scientific Research Fund (OTKA K104279). The work of Tibor Erős was supported by the János Bolyai Research Scolarship of the Hungarian Academy of Sciences. The work of Jani Heino was supported by the Academy of Finland and Emil Aaltonen Foundation.

References

Archaimbault, V., P. Usseglio-Polatera & J.-P. Vanden Bossche (2005) Functional differences among benthic macroinvertebrate communities in reference streams of same order in a given biogeographic area. *Hydrobiologia*, **551**, 171-182.

Bady, P., Dolédec, S., Fesl, C., Gayraud, S., Bacchi, M. & Schöll, F. (2005) Use of invertebrate traits for the biomonitoring of European large rivers: the effects of sampling effort on genus richness and functional diversity. *Freshwater Biology*, **50**, 159-173.

Chevenet, F., Doledec, S. & Chessel, D. (1994) A fuzzy coding approach for the analyses of long-term ecological data. *Freshwater Biology*, **31**, 295-309.

Dolédec, S., Phillips, N., Scarsbook, M., Riley, R. H. & Townsend, C. R. (2006) Comparison of structural and functional approaches to determining landuse effects on grassland stream invertebrate communities. *Journal of the North American Benthological Society*, **25**, 44-60.

Elliot, J. M. (1977) *Some methods for the statistical analysis of samples of benthic invertebrates*. Freshwater Biological Association, Ambleside

Gayraud, S., Statzner, B., Bady, P., Hyabach, A., Scholl, F., Usseglio-Polatera, P. & Bachhi, M. (2003) Invertebrate traits for the biomonitoring of large European rivers: an initial assessment of alternative metrics. *Freshwater Biology*, **48**, 2045-2064.

Heino, J., Schmera, D. & Erős, T. (2013) A macroecological perspective of trait patterns in stream communities. *Freshwater Biology*, **58**, 1539-1555.

Larsen, S. & Ormerod, S. J. (2010) Combined effects of habitat modification on trait composition and species nestedness in river invertebrates. *Biological Conservation*, **143**, 2638-2646.

Legendre, P. & Legendre, L. (2012) Numerical Ecology. 3rd English Edition. Elsevier, Amsterdam, The Netherlands.

Monaghan, K. A. & Soares, A.M.V.M. (2014) The deviation of long-transformed abundance data for the qualitative analysis of macroinvertebrate traits - an addendum to "A macroecological perspective of trait patterns in stream communities" by Heino et al. (2014). *Freshwater Biology*, (in press)

Podani, J. (2000) *Introduction to the Exploration of Multivariate Biological Data*. Backhuys, Leiden.

Ricotta C. & Moretti, M. (2011) CWM and Rao's quadratic diversity: a unified framework for functional ecology. *Oecologia*, **167**, 181-188.

Statzner, B., Bis, B., Doledec, S, & Usseglio-Polatera, P. (2001) Perspectives for biomonitoring at large spatial scales: a unified measure for the functional composition on invertebrate communities in European running waters. *Basic and Applied Ecology*, **2**, 73-85.

Statzner, B. & Beche, L. (2010) Can biological invertebrate traits resolve effects of effects of multiple stressors on running water ecosystems? *Freshwater Biology*, **55**, 80-119.

Usseglio-Polatera, P., Bournaud, M., Richoux, P. & Tachet, H. (2000) Biomonitoring through biological traits of benthic macroinvertebrates: how to use species trait databases? *Hydrobiologia*, **422/423**, 153-162.

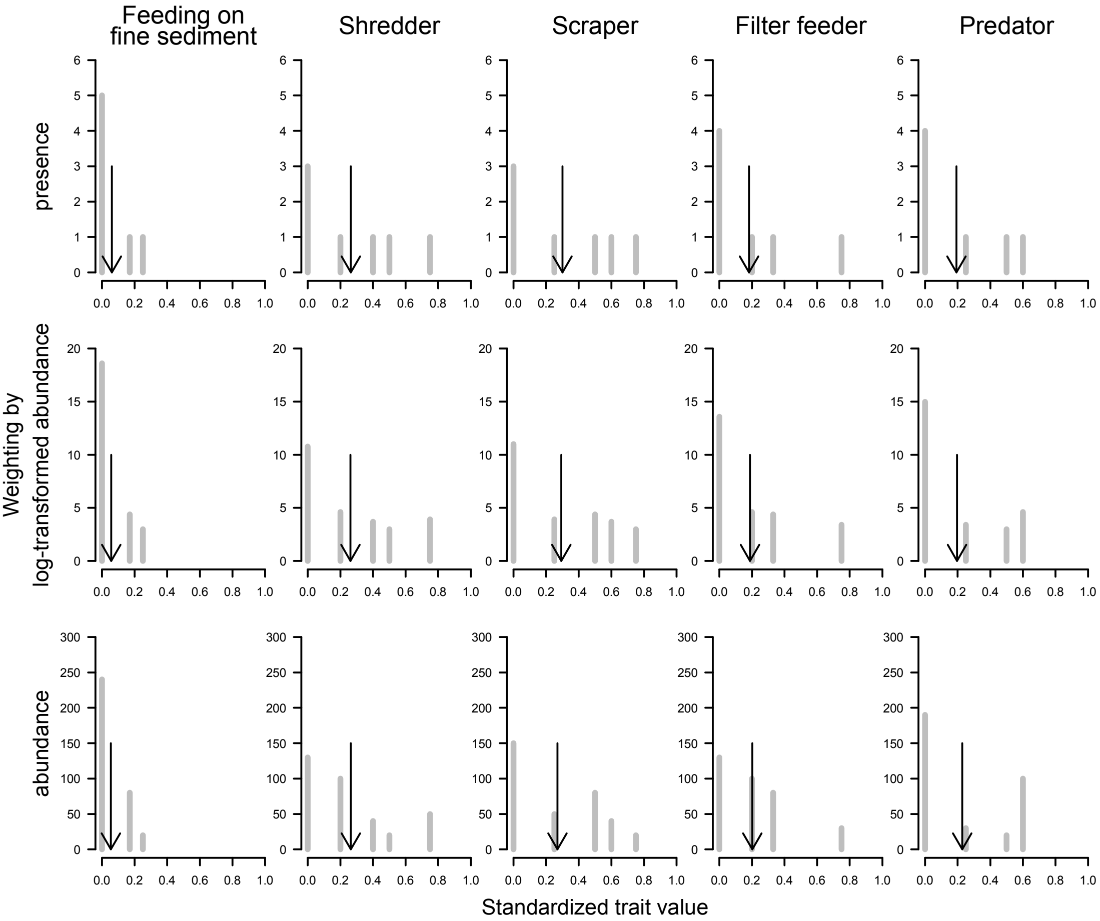

Fig. 1: The frequency distribution of standardized trait values weighted by taxon presence (top row), log-transformed abundance (middle row) and raw abundance (lower row) for five different traits. Grey columns represent frequencies of standardized trait values of the taxa, while arrows show the community-level trait values. The community contains 80 individuals of *Psychomyia*, 50 individuals of *Sericostoma*, 100 individuals of *Polycentropus*, 20 individuals of *Odontocerum*, 40 individuals of *Drusus*, 30 individuals of *Hydropsyche* and 20 individuals of *Agapetus* as displayed in Table 2 of Monaghan & Soares (2014).

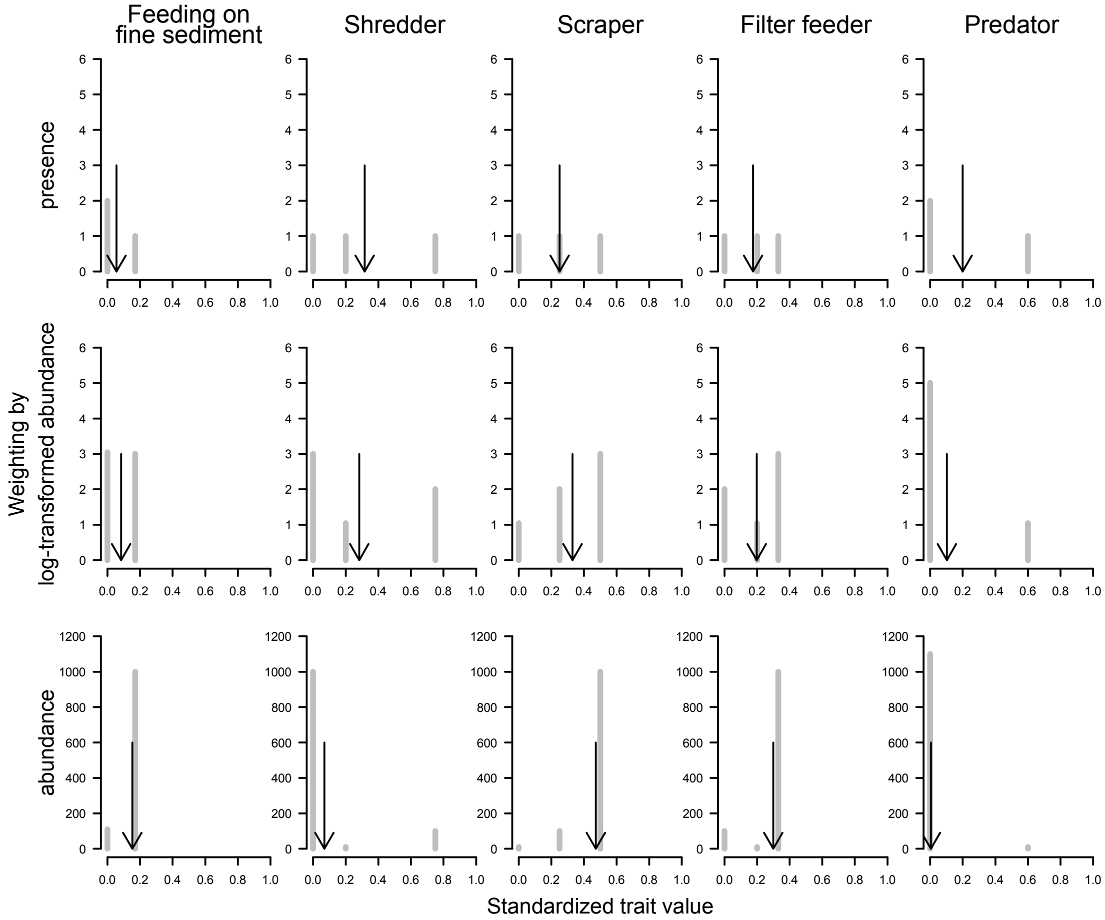

Fig. 2: The frequency distribution of standardized trait values weighted by taxon presence (top row), $\log_{10}$-transformed abundance (middle row) and raw abundance (lower row) for five different traits. Grey columns represent frequencies of standardized trait values of the taxa, while arrows show the community-level trait values. The community contains 1000 individuals of *Psychomyia*, 100 individuals of *Sericostoma* and 10 individuals of *Polycentropus*.