Ecological Indicators 110 (2020) 105937

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Original Articles

Sampling effort and information quality provided by rare and common species in estimating assemblage structure

Luciano F. Sgarbi[a], Luis M. Bini[b], Jani Heino[c], Jenny Jyrkänkallio-Mikkola[d], Victor L. Landeiro[e], Edineusa P. Santos[f], Fabiana Schneck[g], Tadeu Siqueira[f], Janne Soininen[d], Kimmo T. Tolonen[c,h], Adriano S. Melo[b,*]

[a] Programa de Pós-Graduação em Ecologia e Evolução, Universidade Federal de Goiás, Goiânia, GO, Brazil
[b] Departamento de Ecologia, Universidade Federal de Goiás, Goiânia, GO, Brazil
[c] Finnish Environment Institute, Freshwater Centre, Oulu, Finland
[d] Department of Geosciences and Geography, University of Helsinki, PO Box 64, Helsinki FIN-00014, Finland
[e] Departamento de Botânica e Ecologia, Universidade Federal do Mato Grosso, Cuiabá, MT, Brazil
[f] Instituto de Biociências, Universidade Estadual Paulista, UNESP, Rio Claro, São Paulo, Brazil
[g] Instituto de Ciências Biológicas, Universidade Federal do Rio Grande, Rio Grande, RS, Brazil
[h] Department of Biological and Environmental Science, University of Jyväskylä, P.O. Box 35, FI-40014 Jyväskylä, Finland

ARTICLE INFO

Keywords:
Community ecology
Biological diversity
Stream insects
Procrustes
Minimal sampling effort

ABSTRACT

Reliable biological assessments are essential to answer ecological and management questions but require well-designed studies and representative sample sizes. However, large sampling effort is rarely possible, because it demands large financial resources and time, restricting the number of sites sampled, the duration of the study and the sampling effort at each site. In this context, we need methods and protocols allowing cost-effective surveys that would, consequently, increase the knowledge about how biodiversity is distributed in space and time. Here, we assessed the minimal sampling effort required to correctly estimate the assemblage structure of stream insects sampled in near-pristine boreal and subtropical regions. We used five methods grouped into two different approaches. The first approach consisted of the removal of individuals 1) randomly or 2) based on a count threshold. The second approach consisted of simplification in terms of 1) sequential removal from rare to common species; 2) sequential removal from common to rare species; and 3) random species removal. The reliability of the methods was assessed using Procrustes analysis, which indicated the correlation between a reduced matrix (after removal of individuals or species) and the complete matrix. In many cases, we found a strong relationship between ordination patterns derived from presence/absence data (the extreme count threshold of a single individual) and those patterns derived from abundance data. Also, major multivariate patterns derived from the complete data matrices were retained even after the random removal of more than half of the individuals. Procrustes correlation was generally high (> 0.8), even with the removal of 50% of the species. Removal of common species produced lower correlation than removal of rare species, indicating higher importance of the former to estimate resemblance between assemblages. Thus, we conclude that sampling designs can be optimized by reducing the sampling effort at a site. We recommend that such efforts saved should be redirected to increase the number of sites studied and the duration of the studies, which is essential to encompass larger spatial, temporal and environmental extents, and increase our knowledge of biodiversity.

1. Introduction

Reliable assessment of biological assemblages demands robust and standardized sampling effort (Bonar et al., 2011), particularly when many species are rare and represented by one or two individuals, or present in one or two sampling units (Gotelli and Colwell, 2009; Kanno

* Corresponding author at: Departamento de Ecologia, ICB, Universidade Federal de Goiás, Goiânia, GO CEP 74690-900, Brazil.
E-mail addresses: luciano.f.sgarbi@gmail.com (L.F. Sgarbi), lmbini@gmail.com (L.M. Bini), jani.heino.eco@gmail.com (J. Heino), jenny.jyrkankallio-mikkola@wwf.fi (J. Jyrkänkallio-Mikkola), vllandeiro@gmail.com (V.L. Landeiro), edineusa.eco@gmail.com (E.P. Santos), fabiana.schneck@gmail.com (F. Schneck), tadeu.siqueira@unesp.br (T. Siqueira), janne.soininen@helsinki.fi (J. Soininen), kimmo.t.tolonen@jyu.fi (K.T. Tolonen), asm.adrimelo@gmail.com (A.S. Melo).

https://doi.org/10.1016/j.ecolind.2019.105937
Received 16 April 2019; Received in revised form 8 November 2019; Accepted 12 November 2019
Available online 25 November 2019
1470-160X/

et al., 2009). These rare species may be common in other habitats (Sgarbi and Melo, 2018), but sampling all habitats and their species in streams and rivers is not an easy task (Hughes et al., 2012; Li et al., 2014). Extensive sampling is rarely feasible in large-scale field surveys, particularly within a short-time frame (Magurran, 2017) or when financial resources are scarce (Smith et al., 2003; Buss et al., 2015; Angelo, 2017; Magalhães, 2017; Siqueira et al., 2017). Thus, we need methods and protocols that increase our knowledge about how biodiversity is distributed in space and time to advance ecological understanding. Yet, caution should be taken when advising such guidelines, as non-ideal sampling may lead to flawed conclusions (Stout and Vandermeer, 1975; Cao and Hawkins, 2005; Chao et al., 2009). In this sense, a key issue in basic and applied ecology is to decide an adequate, minimal sample size to optimize time and financial resources in biodiversity research (Hughes and Peck, 2008).

An adequate sampling effort is intrinsically related to the objective of surveys and the distinctiveness of the assemblages under comparison. For instance, studies aiming to estimate species richness tend to require a higher sampling effort than those aiming to estimate beta diversity (Schneck and Melo, 2010) because the former is very dependent on rare species which are collected at very low rates in the accumulated samplings (Melo, 2004; Kanno et al., 2009). Studies on functional diversity should also require a high sampling effort if rare species contribute disproportionally in terms of uncommon traits (Leitão et al., 2016). Regarding the distinctiveness of the assemblages under comparisons, low sampling effort is enough to detect differences among assemblages if this difference is large. For example, a reduced sampling effort is enough to detect severe human disturbance on ecosystems. Yet, a high sampling effort is needed if disturbance is weak or if the survey is intended to detect early signals of disturbance, for example, when only sensitive or rare species are expected to disappear (Firmiano et al., 2017). Comparisons among assemblages across near-pristine environmental gradients should also require a high sampling effort, because no loss of species is expected and differences should mostly be due to species turnover as a result of local species sorting.

One potential way to attain time- and cost-effective sampling is the use of minimal, but adequate, number of sampled individuals. Determining the adequate sample size may be difficult for metrics based on species richness, due to its strong dependency on accumulated sampling of individuals (Gotelli and Colwell, 2001; Melo, 2004; Cao and Hawkins, 2005). However, this may not be a major problem when estimating assemblages resemblance (Cao et al., 2002; Cao and Hawkins, 2005; Schneck and Melo, 2010). This is because individuals are sampled haphazardly and, on average, the estimated relative abundances of species are close to those observed in nature (Marchant, 2002). Moreover, although some rare species are not detected, these species may have a minor weight in the analyses of assemblages (Yu et al., 2017), particularly for dissimilarity indices based on abundance data (Marchant, 2002; Draper et al., 2019). Accordingly, reduced sample sizes may be enough to estimate common species and, thus, properly reflect compositional resemblance between assemblages (Cao and Hawkins, 2005; Legendre and Legendre, 2012).

A second strategy to obtain reliable and cost- and time-effective data is to use thresholds of individuals counted for a given species. The threshold strategy consists of counting a maximum number of individuals per species within a sample (Blanchet et al., 2016). The rationale is that species presenting more than a given (threshold) number of individuals are regarded as common and that further counting should not bring much more information. The use of thresholds may be time- and cost-effective mainly for sorting small organisms in the laboratory, because it would not be necessary to separate all individuals belonging to the most common species. Yet, this strategy needs the proper identification of individuals, a task that may take a long time to accomplish. Thus, it should be mostly feasible for coarse taxonomic levels (e.g. family) or during the identification phase, after sorting of individuals. However, one advantage of the species-threshold approach, as compared to reducing the total count of individuals per sample, is that all sampled individuals are assessed (but not counted) and thus best estimates of species richness are obtained.

The estimation of resemblance between assemblages using either rare or common species has been a controversial issue among ecologists (Cao et al., 1998; Marchant, 2002; Poos and Jackson, 2012; Yu et al., 2017), although the use of only one of them may have some methodological advantages. For example, low sampling effort is sufficient to detect common species and correctly estimate their relative abundances. In fact, researchers often remove rare species from assemblage datasets because they may add noise to multivariate analyses (Marchant, 2002; Queheillalt et al., 2002). Also, the taxonomy and biology of common species is generally better known than that of rare species because common species are widely distributed and usually present in several biological collections. On the other hand, rare species may sometimes provide better information than common species to estimate resemblance between assemblages and, therefore, beta diversity patterns (Cao et al., 1998; Poos and Jackson, 2012). For instance, common species may be present nearly everywhere and be poor indicators of assemblage dissimilarity, whereas rare species may be more sensitive to environmental differences and best discriminate assemblages (Cao et al., 1998; but see Marchant, 2002). The debate about the use of rare or common species is unresolved, and it is thus necessary that researchers know which group, rare or common, retains more information about variation in assemblage structure (Heino and Soininen, 2010; Siqueira et al., 2012; Alahuhta et al., 2014). One approach to evaluate the effectiveness of the sole use of rare or common species consists of removing species, from the rarest to the commonest species (or vice versa), followed by the comparison of results against random removal of species (Leitão et al., 2016; Roque et al., 2016; Graça et al., 2017).

Many studies have evaluated the minimal sampling effort required for biodiversity analyses (Gotelli and Colwell, 2001; Cao et al., 2002; Cao and Hawkins, 2005; Chao et al., 2009; Saito et al., 2015; Blanchet et al., 2016; Roque et al., 2016). However, particularities of the different regions across the world (e.g. variation in species richness, density and proportion of rare species; Lang et al., 2019) may lead to erroneous conclusions. This is the case of subtropical-tropical *vs.* temperate-boreal systems. Stout and Vandermeer (1975), for example, sampled insects in tropical and temperate streams, and found the former to be more species rich. Yet, this was evident only after a large sampling effort in tropical streams, which showed low density. Also, Heino et al. (2018) found a difference in genus richness between boreal and subtropical streams, with higher richness in the latter, particularly at coarse spatial scales. The abundances also differed between regions, but with subtropical streams harboring five-fold lower densities than boreal streams (Heino et al., 2018).

Despite many studies evaluating minimal sampling effort, few studies have evaluated different sampling effort approaches, for instance, the removal of individuals, samples or species. Also, many of the previous evaluations were done using strong disturbance gradients where small sampling effort may easily detect differences between assemblages. We used stream insects from near-pristine subtropical and boreal regions as model organisms and evaluated the minimal effort needed to properly estimate the similarity between assemblages using five methods grouped in two approaches. The first approach consisted of 1) random removal of individuals within a sample, and 2) removal of individuals based on counting thresholds. The second approach evaluated which subset of species, rare or common, best match the structure observed in the complete set of species using three methods: 1) species removal from the rarest to the commonest species; 2) removal from the commonest to the rarest species and; 3) random species removal. The efficiency of the two methods based on removal of individuals and the three methods using different subsets of species was evaluated using Procrustes analysis, which assesses the correlation between the complete and the reduced datasets.

2. Materials and methods

2.1. Study area

We used two independent datasets that were collected following similar sampling protocols. The first dataset (BR dataset) was obtained in the Carmo River catchment, Parque Estadual Intervales (24°S, 48°W), state of São Paulo, Brazil. The vegetation of the area is tropical ombrophilous submontane-montane forest (Melo and Froehlich, 2001). The whole basin has high vegetation cover with well-preserved streams. The streams were characterized by neutral to slightly alkaline, well-oxygenated and oligotrophic waters (Valente-Neto et al., 2017). The bedrock is composed of several types of rocks, but predominantly by those of sedimentary origin (Melo and Froehlich, 2001).

The second dataset (FIN dataset) was obtained in the Oulankajoki River basin, Oulanka National Park (66°N, 29°E), north-eastern Finland. The study area has considerable altitudinal differences, and the vegetation varies from pristine coniferous forests to mixed-deciduous riparian woodlands (Heino et al., 2013). The headwater streams are generally near-pristine, with alkaline waters, with low to high concentrations of humic substances and low to moderate nutrient concentrations. The bedrock is composed predominantly of calcareous rocks (Heino et al., 2013).

2.2. Sampling design

For both datasets, 10 riffles were sampled in each of nine streams (9 streams × 10 riffles = 90 riffles) in the first half of September 2009 (FIN) and April 2015 (BR). The distance between streams ranged from ~220 m to ~12 km in BR and from ~1100 m to ~19 km in FIN. In each stream, riffles were sampled from near its confluence with the Carmo River (BR) and the Oulankajoki River (FIN) to upstream. The distance between consecutive sampled riffles varied from 25 to 50 m in BR (Valente-Neto et al., 2017) and from 50 to 200 m in FIN (Heino et al., 2013).

For both datasets, aquatic insects were sampled in each riffle site using a kick net (mesh size = 0.33 mm) during 2 min in an accumulated effort of four 30-seconds sample units (Heino et al., 2013; Valente-Neto et al., 2017). The sampling was designed to include all main habitats of each riffle (Heino et al., 2013; Valente-Neto et al., 2017). Insects sampled in each riffle were preserved in ethanol and transported to the laboratory, where all individuals were counted and identified to genus level. Both datasets included the insect orders Ephemeroptera, Plecoptera, Trichoptera, Coleoptera and Odonata.

2.3. Statistical methods

We considered the set of 10 riffles in each stream as being a metacommunity (Heino and Grönroos, 2013). As these two regions differ in a number of ways, we performed the analyses for each region separately. Our interest was in the variation among riffles (local assemblages) within each stream (metacommunities). Thus, the metacommunity was the level for which a pattern was evaluated. Accordingly, despite the high number of sampled riffles, averages used in the comparisons were obtained for the metacommunity level (i.e. nine streams).

2.3.1. Individual-based reduction in sampling effort methods

For each stream data, we generated reduced datasets by randomly removing part of the individuals from each riffle. The proportion of individuals removed varied sequentially from 0 to 0.95 with intervals of 0.05. In other words, we created nested datasets using random sequential subsamples of different proportions of individuals. As the removal was at random, some genera with one or a few individuals in the full riffle data may have been missing in the reduced dataset. However, all riffles remained in the reduced datasets as removal was performed within each riffle. We generated 100 reduced datasets for each proportion value (20) of each stream, totaling 36,000 matrices (2 regions × 9 streams × 20 proportion values × 100 reduced datasets).

We also created reduced assemblage data by defining a threshold of maximum number of individuals counted per genus within a riffle (Blanchet et al., 2016). The threshold values started from one individual and were increased sequentially up to the largest number of individuals found for a genus within a riffle within a given stream. Threshold values were restricted to the observed abundance values in the riffle sample. Note that, in one extreme, the threshold of one individual equals the use of presence-absence data, whereas in the other extreme the threshold of the highest number of individuals in a sample equals the use of the complete abundance data (Blanchet et al., 2016).

2.3.2. Taxon-based reduction in sampling effort methods

We also created reduced data by removing genera sequentially either from the rarest to the commonest (RtoC) or from the commonest to the rarest ones (CtoR). Genus abundance was defined as the total number of individuals per genus within each riffle of a stream. We sequentially removed genera from the rarest (RtoC) or from the commonest (CtoR) until we reached 50% of the genera present in each stream data. We also generated reduced datasets by removing genera randomly. The removal of genera was sequential, producing nested datasets until 50% of the total number of genera was reached in each stream. We generated 100 reduced datasets for each number of random genera excluded and for each stream site.

2.3.3. Concordance between reduced and complete assemblage data

We evaluated the concordance between the ordination produced with the complete species dataset and the ordination produced with reduced datasets using a Procrustes analysis (Peres-Neto and Jackson, 2001). In a typical application using two sets of scores (with the sampling sites as matrix rows) from an ordination method (see below), Procrustes analysis involves, first, scaling the sets of scores so that the sizes of their distributions (in p-dimensions) are similar. Second, ordination axes are rotated until the distances between their sites and the corresponding sites of the other set (which is kept as a reference) are minimized. These procedures produce a badness-of-fit statistic (called m^2), which can be transformed to a goodness-of-fit statistic (r), where $r = \sqrt{1 - m^2}$. In our study, a high value of r indicates that the ordination pattern of the riffles generated by a reduced dataset is concordant with the pattern generated by the complete dataset.

We calculated biological dissimilarities for the complete and corresponding reduced datasets (10 riffles in a stream) using the Bray-Curtis index on log (x + 1) transformed data and the Sørensen index on presence-absence data. Due to the characteristics of the threshold approach to reduce sampling effort (Blanchet et al., 2016), only abundance data were used. Next, we used Principal Coordinates Analysis (PCoA) to ordinate riffles. We performed Procrustes analysis based on the first five PCoA axes as described above.

2.3.4. Efficiency in recovering ordination patterns

We fitted a local polynomial regression (locally estimated scatterplot smoothing – LOESS) to summarize the relationship between our measure of efficiency in the recovery of the patterns of complete datasets (Procrustes correlation r) and the amount of reduction in the sampling effort. We used this class of regression because of the non-linear nature of these relationships. The amount of reduction in the sampling effort was standardized considering the proportion of individuals and genera removed in each assemblage (for analyses based on the removal of individuals and genera, respectively). LOESS was fitted with the α parameter equal to 0.5. The α parameter controls the proportion of points in a neighborhood in relation to x-axis, which influence and are used to fit (using weighted least squares) each y-value. Finally, we computed a 95% confidence interval based on the Student t

distribution for each fitted y-value.

We used the R language to perform all analyses (version 3.4.2; R Core Team, 2017). The multivariate analyses were performed using functions available in the vegan package (Oksanen et al., 2017). The LOESS was computed using the function 'loess' of the stats package (R Core Team, 2017).

3. Results

The BR dataset included a total of 35,382 individuals belonging to 81 genera. The genus richness and the number of individuals per riffle ranged from 18 to 50 (mean = 31.86) and from 142 to 837 (mean = 393.13), respectively. The site-occupancy of genera, pooling all 10 riffles at all nine stream sites, ranged from 1 to 90 riffles (mean = 35.4), and the abundances of genera varied from 1 to 6101 individuals (mean = 436.81).

The FIN dataset included 66,710 individuals belonging to 51 genera. The genus richness and the number of individuals per riffle ranged from 6 to 26 (mean = 15.08) and from 41 to 3490 (mean = 741.22), respectively. The site-occupancy of the genera ranged from 1 to 87 riffles (mean = 26.61), and the abundances of genera varied from 1 to 34,966 individuals (mean = 1308.04).

3.1. Individual-based reduction in sampling effort

We observed moderate to high values of Procrustes correlation (always above 0.62) even with the removal of 95% of the individuals (Fig. 1; see Supporting information S1 for curves of individual streams), with both random and threshold removal approaches, and for both BR and FIN datasets. For the random removal method, the transformation into presence-absence data (Fig. 1A and C) resulted in lower Procrustes correlations than those using abundance data (Fig. 1B and D). Moreover, we observed that the use of threshold counting always produced higher Procrustes correlation values than those produced by the random removal of individuals (Fig. 1B and D).

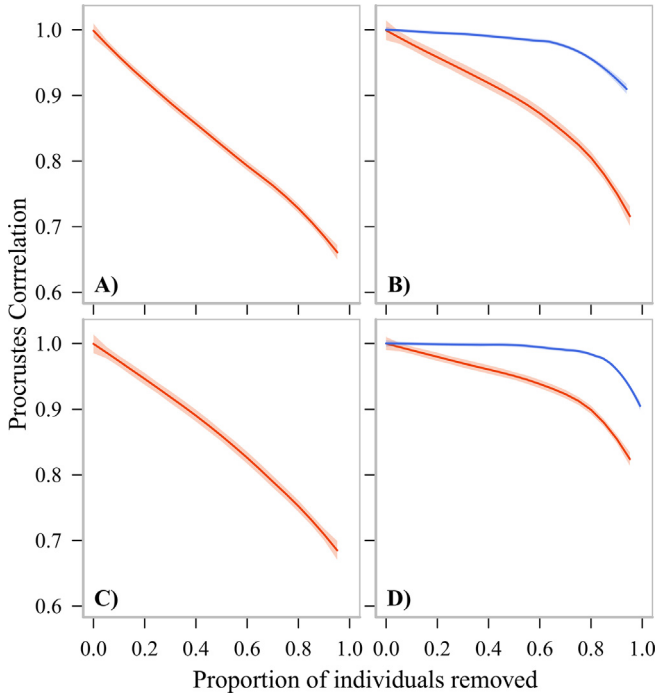

Fig. 1. Smooth curves fitted by LOESS and their respective 95% confidence interval (shaded area) for the effects of the proportion of individuals randomly removed (orange) or threshold counting (blue) on multivariate Procrustes correlation between reduced and complete assemblage data. Shown are the results for Brazil (A and B) and Finland (C and D). Analyses were performed using presence-absence (A and C) and log abundance data (B and D). Threshold counting could be performed only for abundance data. The range of the y-axes follows the variation among individual streams (available in the Supporting information S1).

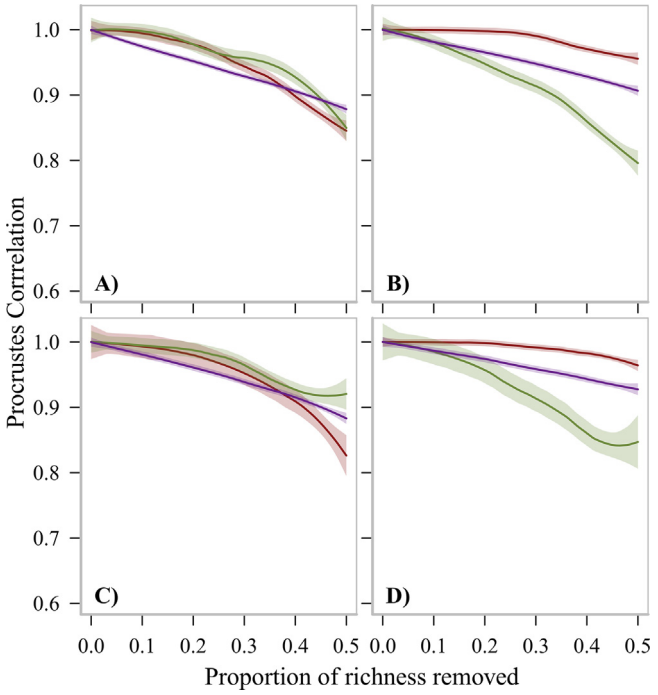

Fig. 2. Smooth curves fitted by LOESS and their respective 95% confidence interval (shaded area) for the effects of the proportion of species richness removed from the rarest to the commonest (RtoC, red), commonest to rarest (CtoR, green) or randomly removed (purple) on multivariate Procrustes correlation between reduced and complete assemblage data. Shown are the results for Brazil (A and B) and Finland (C, D). Analyses were performed using presence-absence (A and C) and log-transformed abundance data (B and D). The range of the y-axes follows the variation among individual streams (available in the Supporting information S1.2).

3.2. Taxon-based reduction in sampling effort

Procrustes correlations were always high (> 0.8), even with the removal of 50% of the genera present in the matrices (Fig. 2; see Fig. S2 for curves of individual streams). Results were consistent between the BR and FIN datasets. For presence-absence data, we did not observe differences among RtoC, CtoR or random removal of species for both datasets (Fig. 2A and C). On the other hand, analyses based on abundance data were dependent on the way in which genera were removed (Fig. 2B and D). Specifically, the highest Procrustes correlation values were obtained in RtoC, while the CtoR approach was the poorest in recovering the structure observed in the complete assemblage data, indicating the importance of common species. Random removal of genera produced intermediate results (Fig. 2B and D).

4. Discussion

The results supported our hypothesis that reduced sets of sampled individuals and species are sufficient to reproduce ordination patterns obtained with the complete set of individuals and species. We also found that the use of abundance data produced higher correlation values than the use of presence-absence data. Despite the high correlation values for both the random removal of individuals and the counting threshold approaches, the latter was able to retain much more information, even when using a notably reduced proportion of the total number of individuals. The three forms of taxon removal on presence-

absence data produced similar correlations to the complete datasets. In contrast, we observed a strong negative effect of the removal of common genera for Procrustes correlation based on abundance data. A striking result was that even removing 80% of the sampled individuals within a stream, or 50% of genera, we still found congruent datasets that indicated the original ordination patterns.

Total genus richness was higher in the subtropical (81 genera) than in the boreal (51 genera) region, whereas abundance was much higher in the boreal than in the subtropical region (see also Heino et al., 2018). Thus, it is interesting that the results were similar despite the striking differences in geographical and ecological settings between the regions. This is in agreement with a study on butterflies and moths in farmlands of Romania, Spain and Sweden, where species richness varied among countries, but adequate sample sizes were similar (Lang et al., 2019). In general, this convergence of patterns suggests that our results are applicable to a wide range of climates and stream benthic macroinvertebrate assemblages provided the spatial and environmental extents are relatively short as the ones studied here.

Presence-absence data are more accessible than abundance data. For instance, while sorting stream invertebrates from sediments researchers may start picking large and easily identified species, and then focus on non-seen species and discard those species already recorded. Abundance data are more informative (Blanchet et al., 2016), but require more time and resources to be obtained, particularly for high-density assemblages such as those of stream macroinvertebrates. We found that as individuals were removed from samples, the presence-absence data lost information about assemblage structure at a much higher rate compared with abundance data. This result has been observed before, where the use of abundance data recovered subtle ecological patterns not detected using presence-absence data (Melo, 2005). This result, however, may depend on the extent of the study. Our study included samples in relatively homogenous environments within short spatial (~1 km) and environmental (e.g. absence of human disturbances) extents. As spatial/environmental extent increases, abundance values may respond more strongly to local processes, such as niche selection and ecological drift, and reduce proper estimates of assemblage resemblance. In these cases, presence-absence may be as good as, or better than, abundance data (Wilson, 2012).

Threshold counting performed very well and produced more reliable results than random removal of individuals. This approach keeps incidence information for all taxa and maintains partial information about their abundances. Therefore, threshold counting recovers much of the information present in the complete dataset, even with low sampling effort (Blanchet et al., 2016). However, its practical use is hampered because individuals must be identified before being discarded from counting. While this may be effective for samples containing abundant and easily-recognized species, it is impractical when individuals need to be carefully examined. In contrast, the random removal of individuals can be easily performed using subsamples (Ligeiro et al., 2013; Saito et al., 2015) and, despite its lower performance in relation to the threshold counting method, it still produced relatively high Procrustes correlation values (i.e. > 0.7) at very low sampling effort.

Although we only partially understand the processes controlling the distribution of rare or common species (Magurran and Henderson, 2003; Alahuhta et al., 2014; Connolly et al., 2014; Sgarbi and Melo, 2018) and the importance of the use of common (Marchant, 2002; Queheillalt et al., 2002) or rare species (Faith and Norris, 1989; Cao et al., 1998) in multivariate analyses, our results provided some practical guidelines to optimize studies of stream benthic macroinvertebrates. The removal of rare or common genera provided similar Procrustes correlations using presence-absence data. On the other hand, common genera were very important to estimate resemblance between assemblages using abundance data. Our results align with those by Draper et al. (2019), who showed, using trees in western Amazonia, that a very small subset of dominant species (99 out of 2031 species) was enough to recover beta diversity and distance-decay patterns. In contrast, the removal of rare species did not affect strongly the estimates of resemblance between assemblages present in the complete abundance-based dataset. There is some evidence that rare species can be as important as the common ones in studies of assemblage-environment associations (Wilson and Meurk, 2011; Siqueira et al., 2012; Leitão et al., 2016). Yet, many rare species may constitute transient species, collected accidentally (Magurran and Henderson, 2003; Sgarbi and Melo, 2018), and it can be argued that their inclusion may bring noise to analyses (Marchant, 2002; Queheillalt et al., 2002). In fact, removal of rare species did not affect substantially the relationship of local assemblages of Amazonian butterflies and plants (Graça et al., 2017), although it affected the resemblance between stream fish assemblages in Canada (Poos and Jackson, 2012), and estimates of functional diversity of fish, birds and trees (Leitão et al., 2016). Accordingly, our finding of the low importance of rare species to recover resemblance between assemblages may be system-specific and likely cannot be widely generalized.

Our results show that random removal of genera performed similarly as the removal of rare or common genera in the recovery of resemblance between assemblages using presence-absence data. This result corroborates a previous study showing that the exclusion of part of the ant genera (where species separation is difficult) did not cause important loss of information (Vasconcelos et al., 2014). It is also in line with studies showing that coarser taxonomic resolution can be used to identify human-impacted and reference sites (Whittier and Van Sickle, 2010) or to separate sites according to ecological factors (Melo, 2005). Accordingly, it can be suggested that, in many cases, selecting only those species that are easy to count (e.g. large individuals) or identify (genera/families including easily recognized species/genera) may still retain good estimates of assemblage compositional variation. Further studies should evaluate whether the use of traits (instead of taxonomic entities) would produce similar results in the assessment of assemblage dissimilarity among pristine streams. One may argue that the trait approach is effective in recovering anthropogenic disturbance gradients (e.g. Castro et al., 2018), but may have low power to detect differences when the study does not include strong gradients that would select specific sets of traits.

We conclude that a reduced sampling effort may be sufficient to recover resemblance between assemblages detected with complete data. Also, the sole use of common species alone or a random subset of species may represent adequately the complete assemblage data. Thus, sampling designs can be optimized in many cases by reducing sampling effort at a site and increasing the number of studied sites and, consequently, enlarging spatial and environmental extents of ecological studies. This is particularly relevant in studies involving small invertebrates, where much time is spent in the sorting and identification of individuals.

CRediT authorship contribution statement

Luciano F. Sgarbi: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Visualization. **Luis M. Bini:** Conceptualization, Methodology, Writing - review & editing. **Jani Heino:** Conceptualization, Methodology, Writing - review & editing, Project administration, Funding acquisition. **Jenny Jyrkänkallio-Mikkola:** Data curation, Writing - review & editing. **Victor L. Landeiro:** Conceptualization, Methodology, Writing - review & editing. **Edineusa Pereira dos Santos:** Investigation, Data curation, Writing - review & editing. **Fabiana Schneck:** Investigation, Writing - review & editing. **Tadeu Siqueira:** Conceptualization, Methodology, Investigation, Data curation, Writing - review & editing, Project administration, Funding acquisition. **Janne Soininen:** Conceptualization, Methodology, Investigation, Data curation, Writing - review & editing, Project administration, Funding acquisition. **Kimmo T. Tolonen:** Data curation, Writing - review & editing. **Adriano S. Melo:**

Conceptualization, Methodology, Software, Formal analysis, Writing - review & editing, Visualization, Supervision.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 (scholarship to LFS). ASM, VLL and LMB received research fellowships from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, process no: 307587/2017-7, 307961/2017-6, and 304314/2014-5, respectively). This work was funded by the FAPESP-AKA Joint Call on Biodiversity and Sustainable Use of Natural Resources: grant 2013/50424-1 from São Paulo Research Foundation (FAPESP) to TS, and grants no. 273557 and no. 273560 from the Academy of Finland (AKA) to JH and JS, respectively.

Declarations of Competing Interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecolind.2019.105937.

References

Alahuhta, J., Johnson, L.B., Olker, J., Heino, J., 2014. Species sorting determines variation in the community composition of common and rare macrophytes at various spatial extents. Ecol. Complex. 20, 61–68. https://doi.org/10.1016/j.ecocom.2014.08.003.

Angelo, C., 2017. Brazilian scientists reeling as federal funds slashed by nearly half. Nature 544, 7648. https://doi.org/10.1038/nature.2017.21766.

Blanchet, F.G., Legendre, P., He, F., 2016. A new cost-effective approach to survey ecological communities. Oikos 125, 975–987. https://doi.org/10.1111/oik.02838.

Bonar, S.A., Fehmi, J.S., Mercado-Silva, N., 2011. An overview of sampling issues in species diversity and abundance surveys. In: Magurran, A.E., McGill, B.J. (Eds.), Biological Diversity: Frontiers in Measurement and Assessment. Oxford University Press, Oxford, pp. 11–24.

Buss, D.F., Carlisle, D.M., Chon, T.S., Culp, J., Harding, J.S., Keizer-Vlek, H.E., Robinson, W.A., Strachan, S., Thirion, C., Hughes, R.M., 2015. Stream biomonitoring using macroinvertebrates around the globe: a comparison of large-scale programs. Environ. Monit. Assess. 187, 4132. https://doi.org/10.1007/s10661-014-4132-8.

Cao, Y., Hawkins, C.P., 2005. Simulating biological impairment to evaluate the accuracy of ecological indicators. J. Appl. Ecol. 42, 954–965. https://doi.org/10.1111/j.1365-2664.2005.01075.x.

Cao, Y., Williams, D.D., Williams, N.E., 1998. How important are rare species in aquatic community ecology and bioassessment? Limnol. Oceanogr. 43, 1403–1409. https://doi.org/10.4319/lo.1998.43.7.1403.

Cao, Y., Larsen, D.P., Hughes, R.M., Angermeier, P.L., Patton, T.M., 2002. Sampling effort affects multivariate comparisons of stream assemblages. J. N. Am. Benthol. Soc. 21, 701–714. https://doi.org/10.2307/1468440.

Castro, D.M.P., Dolédec, S., Callisto, M., 2018. Land cover disturbance homogenizes aquatic insect functional structure in neotropical savanna streams. Ecol. Ind. 84, 573–582. https://doi.org/10.1016/j.ecolind.2017.09.030.

Chao, A., Colwell, R.K., Lin, C.W., Gotelli, N.J., 2009. Sufficient sampling for asymptotic minimum species richness estimators. Ecology 90, 1125–1133. https://doi.org/10.1890/07-2147.1.

Connolly, S.R., MacNeil, M.A., Caley, M.J., Knowlton, N., Cripps, E., Hisano, M., Thibaut, L.M., Bhattacharya, B.D., Benedetti-Cecchi, L., Brainard, R.E., Brandt, A., Bulleri, F., Ellingsen, K.E., Kaiser, S., Kröncke, I., Linse, K., Maggi, E., O'Hara, T.D., Plaisance, L., Poore, G.C.B., Sarkar, S.K., Satpathy, K.K., Schückel, U., Williams, A., Wilson, R.S., 2014. Commonness and rarity in the marine biosphere. Proc. Natl. Acad. Sci. U.S.A. 111, 8524–8529. https://doi.org/10.1073/pnas.1406664111.

Draper, F.C., Asner, G.P., Coronado, E.N.H., Baker, T.R., García-Villacorta, R., Pitman, N.C.A., Fine, P.V.A., Phillips, O.L., Gómez, R.Z., Guerra, C.A.A., Arévalo, M.F., Martínez, R.V., Brienen, R.J.W., Monteagudo-Mendoza, A., Montenegro, L.A.T., Sandoval, E.V., Roucoux, K.H., Arévalo, F.R.R., Acuy, I.M., Pasquel, J.D.A., Casapia, X.T., Llampazo, G.F., Medina, M.C., Huaymacari, J.R., Baraloto, C., 2019. Dominant tree species drive beta diversity patterns in western Amazonia. Ecology 100, e02636. https://doi.org/10.1002/ecy.2636.

Faith, D.P., Norris, R.H., 1989. Correlation of environmental variables with patterns of distribution and abundance of common and rare freshwater macroinvertebrates. Biol. Conserv. 50, 77–98. https://doi.org/10.1016/0006-3207(89)90006-2.

Firmiano, K.R., Ligeiro, R., Macedo, D.R., Juen, L., Hughes, R.M., Callisto, M., 2017. Mayfly bioindicator thresholds for several anthropogenic disturbances in neotropical savanna streams. Ecol. Ind. 74, 276–284. https://doi.org/10.1016/j.ecolind.2016.11.033.

Gotelli, N.J., Colwell, R.K., 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecol. Lett. 4, 379–391. https://doi.org/10.1046/j.1461-0248.2001.00230.x.

Gotelli, N.J., Colwell, R.K., 2009. Estimating species richness. In: Magurran, A.E., McGill, B.J. (Eds.), Biological Diversity: Frontiers in Measurement and Assessment. Oxford University Press, Oxford, pp. 39–54.

Graça, M.B., Souza, J.L., Franklin, E., Morais, J.W., Pequeno, P.A.C.L., 2017. Sampling effort and common species: optimizing surveys of understorey fruit-feeding butterflies in the Central Amazon. Ecol. Ind. 73, 181–188. https://doi.org/10.1016/j.ecolind.2016.09.040.

Heino, J., Grönroos, M., 2013. Does environmental heterogeneity affect species co-occurrence in ecological guilds across stream macroinvertebrate metacommunities? Ecography 36, 926–936. https://doi.org/10.1111/j.1600-0587.2012.00057.x.

Heino, J., Soininen, J., 2010. Are common species sufficient in describing turnover in aquatic metacommunities along environmental and spatial gradients? Limnol. Oceanogr. 55, 2397–2402. https://doi.org/10.4319/lo.2010.55.6.2397.

Heino, J., Grönroos, M., Ilmonen, J., Karhu, T., Niva, M., Paasivirta, L., 2013. Environmental heterogeneity and β diversity of stream macroinvertebrate communities at intermediate spatial scales. Freshw. Sci. 32, 142–154. https://doi.org/10.1899/12-083.1.

Heino, J., Melo, A.S., Jyrkänkallio-Mikkola, J., Petsch, D.K., Saito, V.S., Tolonen, K.T., Bini, L.M., Landeiro, V.L., Silva, T.S.F., Pajunen, V., Soininen, J., Siqueira, T., 2018. Subtropical streams harbour higher genus richness and lower abundance of insects compared to boreal streams, but scale matters. J. Biogeogr. 45, 1983–1993. https://doi.org/10.1111/jbi.13400.

Hughes, R.M., Peck, D.V., 2008. Acquiring data for large aquatic resource surveys: the art of compromise among science, logistics, and reality. J. N. Am. Benthol. Soc. 27, 837–859. https://doi.org/10.1899/08-028.1.

Hughes, R.M., Herlihy, A.T., Gerth, W.J., Pan, Y., 2012. Estimating vertebrate, benthic macroinvertebrate and diatom taxa richness in raftable Pacific Northwest rivers for bioassessment purposes. Environ. Monit. Assess. 184, 3185–3198. https://doi.org/10.1007/s10661-011-2181-9.

Kanno, Y., Vokoun, J.C., Dauwalter, D.C., Hughes, R.M., Herlihy, A.T., Maret, T.R., Patton, T.M., 2009. Influence of rare species on electrofishing distance when estimating species richness of stream and river reaches. Trans. Am. Fish. Soc. 138, 1240–1251. https://doi.org/10.1577/T08-210.1.

Lang, A., Kallhardt, F., Lee, M.S., Loos, J., Molander, M.A., Muntean, I., Pettersson, L.B., Rákosy, L., Stefanescu, C., Messéan, A., 2019. Monitoring environmental effects on farmland Lepidoptera: Does necessary sampling effort vary between different biogeographic regions in Europe? Ecol. Ind. 102, 791–800. https://doi.org/10.1016/j.ecolind.2019.03.035.

Legendre, P., Legendre, L., 2012. Numerical Ecology, third ed. Elsevier, Oxford.

Leitão, R.P., Zuanon, J., Villeger, S., Williams, S.E., Baraloto, C., Fortunel, C., Mendonça, F.P., Mouillot, D., 2016. Rare species contribute disproportionately to the functional structure of species assemblages. Proc. R. Soc. B 283, 20160084. https://doi.org/10.1098/rspb.2016.0084.

Li, L., Liu, L., Hughes, R.M., Cao, Y., Wang, X., 2014. Towards a protocol for stream macroinvertebrate sampling in China. Environ. Monit. Assess. 186, 469–479. https://doi.org/10.1007/s10661-013-3391-0.

Ligeiro, R., Ferreira, W., Hughes, R.M., Callisto, M., 2013. The problem of using fixed-area subsampling methods to estimate macroinvertebrate richness: a case study with Neotropical stream data. Environ. Monit. Assess. 185, 4077–4085. https://doi.org/10.1007/s10661-012-2850-3.

Magalhães, A.L.B., 2017. Brazil: Biodiversity at risk from austerity law. Nature 542, 295. https://doi.org/10.1038/542295e.

Magurran, A.E., 2017. The important challenge of quantifying tropical diversity. BMC Biol. 15, 14. https://doi.org/10.1186/s12915-017-0358-6.

Magurran, A.E., Henderson, P.A., 2003. Explaining the excess of rare species in natural species abundance distributions. Nature 422, 714–716. https://doi.org/10.1038/nature01547.

Marchant, R., 2002. Do rare species have any place in multivariate analysis for bioassessment? J. N. Am. Benthol. Soc. 21, 311–313. https://doi.org/10.2307/1468417.

Melo, A.S., 2004. A critique of the use of jackknife and related non-parametric techniques to estimate species richness. Commun. Ecol. 5, 149–157. https://doi.org/10.1556/ComEc.5.2004.2.1.

Melo, A.S., 2005. Effects of taxonomic and numeric resolution on the ability to detect ecological patterns at a local scale using stream macroinvertebrates. Arch. Hydrobiol. 164, 309–323. https://doi.org/10.1127/0003-9136/2005/0164-0309.

Melo, A.S., Froehlich, C.G., 2001. Macroinvertebrates in neotropical streams: richness patterns along a catchment and assemblage structure between 2 seasons. J. N. Am. Benthol. Soc. 20, 1–16. https://doi.org/10.2307/1468184.

Oksanen J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs E., Wagner H., 2017. Vegan: Community Ecology Package. R package ver. 2.4-2. https://www.R-project.org/.

Peres-Neto, P.R., Jackson, D.A., 2001. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. Oecologia 129, 169–178. https://doi.org/10.1007/s004420100.

Poos, M.S., Jackson, D.A., 2012. Addressing the removal of rare species in multivariate bioassessments: the impact of methodological choices. Ecol. Ind. 18, 82–90. https://doi.org/10.1016/j.ecolind.2011.10.008.

Queheillalt, D.M., Cain III, J.W., Taylor, D.E., Morrison, M.L., Hoover, S.L., Tuatoo-Bartley, N., Rugge, L., Christopherson, K., Hulst, M.D., Harris, M.R., Keough, H.L., 2002. The exclusion of rare species from community-level analyses. Wildlife Soc. Bull. 30, 756–759.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Roque, F.D.O., Zampiva, N.K., Valente-Neto, F., Menezes, J.F., Hamada, N., Pepinelli, M., Siqueira, M., Swan, C., 2016. Deconstructing richness patterns by commonness and rarity reveals bioclimatic and spatial effects in black fly metacommunities. Freshw. Biol. 61, 923–932. https://doi.org/10.1111/fwb.12757 https://doi.org/10.1111/fwb.12757.

Saito, V.S., Fonseca-Gessner, A.A., Siqueira, T., 2015. How should ecologists define sampling effort? The potential of procrustes analysis for studying variation in community composition. Biotropica 47, 399–402. https://doi.org/10.1111/btp.12222.

Schneck, F., Melo, A.S., 2010. Reliable sample sizes for estimating similarity among macroinvertebrate assemblages in tropical streams. Ann. Limnol. Int J. Lim. 46, 93–100. https://doi.org/10.1051/limn/2010013.

Sgarbi, L.F., Melo, A.S., 2018. You don't belong here: explaining the excess of rare species in terms of habitat, space and time. Oikos 127, 497–506. https://doi.org/10.1111/oik.04855.

Siqueira, T., Bini, L.M., Roque, F.O., Couceiro, S.R.M., Trivinho-Strixino, S., Cottenie, K., 2012. Common and rare species respond to similar niche processes in macroinvertebrate metacommunities. Ecography 35, 183–192. https://doi.org/10.1111/j.1600-0587.2011.06875.x.

Siqueira, C.C., Frederico, C., Rocha, D., 2017. Brazil's public universities in crisis. Science 356, 812. https://doi.org/10.1126/science.aan2527.

Smith, R., Muir, R.D., Walpole, M.J., Balmford, A., Leader-Williams, N., 2003. Governance and the loss of biodiversity. Nature 426, 67–70. https://doi.org/10.1038/nature02025.

Stout, J., Vandermeer, J., 1975. Comparison of species richness for stream-inhabiting insects in tropical and mid-latitude streams. Am. Nat. 109, 263–280.

Valente-Neto, F., Durães, L., Siqueira, T., Roque, F.O., 2017. Metacommunity detectives: confronting models based on niche and stochastic assembly scenarios with empirical data from a tropical stream network. Freshw. Biol. 63, 86–99. https://doi.org/10.1111/fwb.13050.

Vasconcelos, H.L., Frizzo, T.L., Pacheco, R., Maravalhas, J.B., Camacho, G.P., Carvalho, K.S., Koch, E.B.A., Pujol-Luz, J.R., 2014. Evaluating sampling sufficiency and the use of surrogates for assessing ant diversity in a Neotropical biodiversity hotspot. Ecol. Ind. 46, 286–292. https://doi.org/10.1016/j.ecolind.2014.06.036.

Whittier, T.R., Van Sickle, J., 2010. Macroinvertebrate tolerance values and an assemblage tolerance index (ATI) for western USA streams and rivers. J. N. Am. Benthol. Soc. 29, 852–866. https://doi.org/10.1899/09-160.1.

Wilson, J.B., 2012. Species presence/absence sometimes represents a plant community as well as species abundances do, or better. J. Veg. Sci. 23, 1013–1023. https://doi.org/10.1111/j.1654-1103.2012.01430.x.

Wilson, J.B., Meurk, C.D., 2011. The control of community composition by distance, environment and history: a regional-scale study of the mountain grasslands of southern New Zealand. J. Biogeogr. 38, 2384–2396. https://doi.org/10.1111/j.1365-2699.2011.02573.x.

Yu, Z., Wang, H., Meng, J., Miao, M., Kong, Q., Wang, R., Liu, J., 2017. Quantifying the responses of biological indices to rare macroinvertebrate taxa exclusion: does excluding more rare taxa cause more error? Ecol. Evol. 7, 1583–1591. https://doi.org/10.1002/ece3.2798.