Ecological Informatics 44: 1-6. (2018)

Through the jungle of methods quantifying multiple-site resemblance

Dénes Schmera[1,2] & János Podani[3,4]

[1]MTA Centre for Ecological Research, Balaton Limnological Institute, Klebelsberg K. u. 3, H-Tihany, Hungary, E-mail: schmera.denes@okologia.mta.hu

[2]MTA Centre for Ecological Research, GINOP Sustainable Ecosystem Group, Klebelsberg K. u. 3, H-Tihany, Hungary

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

[4]MTA-ELTE-MTM Ecology Research Group, Budapest, Hungary

Abstract

Methods that quantify multiple-site resemblance are basic toolkits of ecology for studying community variation in space and time. Although both pairwise and multiple-site coefficients have received increasing attention in the past decade, the high variety of methodologies combined with the absence of a systematic review prevents full understanding and comprehension. To illuminate the situation, we compare and classify methods that use incidence data and propose a unified terminology. The methods can be grouped according to families, approaches and forms. The examination of algebraic expressions and analyses of artificial and actual data sets suggest that inference drawn about communities strongly depends on the methodology applied. We found that the impact of mimicking the original pairwise indices (i.e. the impact of families) was stronger than the impact of components used in formulating the coefficients (i.e. the impact of approach). Our findings suggest that the measures examined quantify drastically different

facets of multiple-site resemblance and therefore they have to be selected with care in community studies.

Keywords

community resemblance, community variation, dissimilarity, multiple-site resemblance coefficients, pairwise resemblance coefficients, similarity

1. Introduction

Understanding spatial variation in species composition is one of the most fundamental challenges of community ecology. This is promoted by testing hypotheses about the processes that generate and maintain biodiversity in ecosystems (Legendre & De Cáceres, 2013). Invasion ecologists, for instance, examine the impact of alien species on native communities, while conservation biologists rely on the measurement of compositional variation in prioritizing areas. The spatial variation of communities can be viewed as either compositional differentiation or similarity (Jost et al., 2011). Beta diversity (Whittaker, 1960, 1972), for instance, expresses compositional differentiation, while community overlap (Arita, 2017, Schmera, 2017) relates to compositional similarity – which are two sides of the same coin.

Community variation has been traditionally studied by examining several pairs of sites from the same locality (but see Legendre & De Cáceres, 2013, for alternative solutions) and quantified by the average value of pairwise resemblance (i.e., similarity or dissimilarity) coefficients (Koleff et al., 2003). Such averages may be used to express both compositional similarity and differentiation. Recently, however, it has been suggested that inference drawn from mean values may be misleading, because pairwise resemblance coefficients cannot account properly for co-occurrence patterns of species in many sites and therefore special indices are required (Diserud & Ødegaard, 2007; Baselga, 2013).

Although multiple-site resemblance coefficients have received increasing attention in contemporary ecology, our knowledge on their relative merits and potential disadvantages is still limited. A recent review on beta diversity deliberately omitted their discussion (Legendre & De Cáceres, 2013) while an even more recent study deepened our understanding of multiple-site overlap measures by providing novel measures and a unified terminology (Arita, 2017). Unfortunately, however, the increasing number of methods, the application of different and often overcomplicated mathematical equations, the ambiguous terminology, as well as the parallel development of similarity and dissimilarity forms impede proper measurement of multiple-site resemblance. Therefore, for the benefit of practicing ecologists, we review the methods quantifying multiple-site resemblance that are based on incidence (presence-absence) data. First, we discuss some basic terms, then we overview pairwise and multiple-site resemblance coefficients. Specifically, we identify and match similarity and dissimilarity forms and simplify some equations. Finally, by using artificial and actual data sets we compare the performance of multiple-site resemblance measures.

2. Basic terms

Originally, pairwise and multiple-site resemblance coefficients have been suggested to measure the (dis)similarity of two or multiple sites based on the presence-absence of species. Consequently, sites are the *objects* of such studies and species are the *descriptors* which characterize the objects. Observed data are commonly arranged in matrix $\mathbf{X} \equiv \{x_{ij}\}$, in which rows represent sites while columns correspond to species (e.g. Legendre & DeCáceres, 2013), a convention followed here as well. *Occurrence* (of species *j* in site *i*) means that species *j* is present in site *i*, coded as $x_{ij} = 1$. In case of species absence, $x_{ij} = 0$. The *species richness* of site *i* (t_i) is the number of occurrences in the given row (row total, $t_i = \sum_{j=1}^{T} x_{ij}$, where *T* is the number of species). The *occurrence frequency* of species *j* (n_j) is the number of sites in which the species is present (called also as range size and calculated as the column total, $n_j = \sum_{i=1}^{N} x_{ij}$, where *N* is the number of sites). Whereas *co-occurrence* is traditionally

understood as the presence of a pair of species in a given site (Mackenzie et al. 2004, Bell 2005, Pollock et al. 2014 and references therein), Arita and co-workers (Trejo-Barocio & Arita 2013, Arita 2017) termed *co-diversity,* with a reference to Bell (2005), as the occurrence of a species in two sites. It follows that the *number of co-occurrences* in a site is the number of species pairs present there, while the *number of co-diversities* is the number of unique site-pair occupancies of a given species. In a more formal way, the number of co-occurrences in site *i* can be expressed as

$$\binom{t_i}{2}, \qquad \text{Eq. 1}$$

while the number of co-diversities of species *j* as:

$$\binom{n_j}{2}. \qquad \text{Eq. 2.}$$

Furthermore, following Schmera (2017) we consider *community overlap* as a phenomenon that represents the intersection in the composition of sites, *overlapping species* as species with at least two occurrences in a set of sites, *overlap size* as a quantitative property of overlapping species that is quantified as the occurrence frequency of the given species minus one:

$$n_j - 1, \qquad \text{Eq. 3.}$$

and *total overlap size* as a quantitative property of community overlap

$$\sum_{j=1}^{T} n_j - T. \qquad \text{Eq. 4.}$$

3. Pairwise resemblance coefficients: a short overview

The literature of numerical ecology abounds in resemblance coefficients (sensu Orlóci 1972) for comparing pairs of sites based on their species composition. We are concerned here with

similarity (*s*) and dissimilarity (*d*) forms which are bounded between 0 and 1, and are therefore complements (*d + s* = 1). Presence-absence versions are commonly expressed in terms of a 2 x 2 contingency table in which *a* refers to the number of species present in both sites being compared (shared species, or the number of overlaps in species composition), *b* to the number of species present only in the first and *c* to the number of species in the second. That is, with respect to a given pair of sites there are *b* and *c* species unique to the first and to the second site, respectively, so that the total number of species in the two sites equals to *a* + *b* + *c*. We shall focus on three well-known resemblance coefficients, namely the Jaccard, the Simpson and the Sørensen indices (Table 1, see Koleff et al. 2003 for further indices).

4. A proposal for a unified terminology to classify methods quantifying multiple-site resemblance

Here we suggest a unified terminology that allows the classification of methods quantifying multiple-site resemblance. The multiple-site indices (see next paragraph for details) mimicking some properties of the original pairwise Jaccard, Simpson and Sørensen indices (Table 2) are termed as different groups (Legendre 2014), types (Arita 2017) or families (Baselga, 2012, Baselga & Leprieur, 2015, Podani & Schmera, 2016). This confused nomenclature, however, does not support the development of the field. We therefore suggest, following the terminology of the first classifier (Baselga 2012), that classes of methods mimicking some properties of the original pairwise coefficients should be termed as *families*. Accordingly, the methods in question can be classified into Jaccard, Simpson and the Sørensen families.

Families, however, do not provide the only way for classifying multiple-site resemblance measures. The next feature on which further grouping is made depends on the type of *components* (mathematical terms) incorporated into the coefficient. Some of the measures

rely only upon pairwise components, some use only general components of the studied presence-absence matrix such as the total overlap size, others use co-diversity and, finally, further ones combine general and pairwise components (see next paragraph for details). We suggest that classes of methods formed according to the components used should be termed as *approaches,* and we can distinguish among mean pairwise, general, co-diversity and mixed components approaches (see below). Finally, as said above, each coefficient can be expressed as similarity or dissimilarity. We will refer to this property of coefficients as *forms*.

Consequently, we suggest a classification of methods quantifying multiple-site resemblance according to families, approaches and forms. The terminology becomes even more complex if we consider that dissimilarity forms (also used as measures of beta diversity) may be partitioned into additive components to separate the effect of various background factors influencing dissimilarity. There are two different *frameworks* for such a partitioning, intensively discussed and debated in the relevant literature (Baselga, 2010, Carvalho et al., 2013, Cardoso et al., 2014, Ensing & Pither, 2015, Chen, 2016, Podani & Schmera, 2016).

5. Multiple-site resemblance: a new classification

Here we suggest a classification of methods assessing multiple-site resemblance by considering families, approaches and forms. In this, we do not suggest any hierarchy among these categories. The classification includes both pairwise and multiple-site coefficients. Pairwise coefficients are used for quantifying multiple-site resemblance by calculating the mean of pairwise coefficients, referred here as "mean pairwise approach".

Multiple-site coefficients express resemblance of more than two sites simultaneously (Table 2). Although the first multiple-site index dates back to the 1950's (Koch 1957, see also Eq.

Tab2/1), further elaboration of such coefficients has started only recently. Some of the new indices follow the logic of pairwise indices and therefore we categorize them into the Jaccard, Simpson and Sørensen families (Table 2). However, the coefficients in either family use different components in quantifying similarity or dissimilarity. In studying the overlap of multiple sites, for instance, Arita (2017) suggested general overlap indices, which use only some general components of the incidence matrix, as well as co-diversity indices, which use the occurrence of two species at a particular site. In other studies, Baselga and co-workers (Baselga et al. 2007, Baselga 2010, 2012) used both general and pairwise components in expressing multiple site resemblance or, in other words, they used mixed components. As said above, we refer to this property of coefficients as *approach* and distinguish among mean pairwise (Table 1), general, co-diversity and mixed components approaches (Table 2). Note that we use the term *general* instead of *general overlap* (sensu Arita 2017), because "general" can reflect both similarity and dissimilarity, whereas "general overlap" intuitively relates to similarity only. Thus, we can distinguish three families (Jaccard, Simpson and Sørensen), four approaches (mean pairwise, general, co-diversity and mixed components) and two forms (similarity and dissimilarity) of methods quantifying multiple site resemblance (Tables 1 & 2).

General similarity indices belonging to different families may be formalized in different ways (Table 2). The observed total overlap size (Eq. 4) may be divided by the maximum number of total overlap size with *N* sites and *T* species (Jaccard family, Eq. Tab2/1), or by the maximum number of total overlap sizes possible if the sites show a nested design (Simpson family, Eq. Tab2/3). Thirdly, average overlap size of species may be divided by the average species richness of sites (Sørensen family, Eq. Tab2/4).

Baselga and co-workers (Baselga et al. 2007, Baselga 2010, 2012), following Diserud & Ødegaard (2007), used $\sum_i t_i - T$ as the "number of shared species" in the multiple-site situation. Since $\sum_{i=1}^{N} t_i = \sum_{j=1}^{T} n_j$ = G, the grand total of **X** (see also Arita et al., 2008, 2012; Arita

2017), we can call $\sum_i t_i - T$ as total overlap size (Eq. 4). Multiple-site "unique species", however, were quantified as the sum of unique species for pairs of sites. It follows that it is a mixed components approach having both pairwise and general constituents. A possible theoretical problem with this is that total overlap size (from the general approach) and the number of site pairs in which the same species occur (pairwise component, called also as co-diversity [Arita 2017]) in the data matrix are not the same (Arita 2017), and therefore the ecological interpretation of these indices is less straightforward.

Moreover, in addition to general indices, Arita (2017) developed a new approach of multiple-site similarity measures he called the co-diversity indices. These indices, in fact, count the sum of the two-site occurrences (co-diversity) of species which is divided either by the sum of the co-diversities when site compositions show a nested design (Simpson family, Eq. Tab2/10), by the possible number of co-diversities when *N* sites are occupied with *T* species (Jaccard family, Eq. Tab2/9) or, finally, by the sum of average species richness for each pair of sites (Sørensen family, Eq. Tab2/11).

6. Simplification of some equations

A couple of mixed-component resemblance coefficients have originally been published with extensive mathematical equations. To make their use easier, we suggest the simplification of two functions. The mixed component Jaccard dissimilarity (Eq. Tab2/6) suggested by Baselga (2012) can be simplified to

$$\frac{\sum_{k<l}(b_{kl}+b_{lk})}{(\sum_i t_i - T)+\sum_{k<l}(b_{kl}+b_{lk})}, \qquad \text{Eq. 5.}$$

while the mixed component Sørensen dissimilarity (Eq. Tab2/8) suggested by Baselga (2010) reduces to

$$\frac{\sum_{k<l}(b_{kl}+b_{lk})}{2(\sum_{i}t_i-T)+\sum_{k<l}(b_{kl}+b_{lk})} \quad \text{Eq. 6.}$$

7. Comparison of methods quantifying multiple-site similarity

7.1 Methods to compare

Here we compare methods that allow quantification multiple-site resemblance. Although we use similarity forms, our conclusions are not restricted to similarity because it is complementary to dissimilarity. Although pairwise coefficients are designed for examining pairs of sites, the mean values of these coefficients are frequently used for assessing multiple site similarity. We will refer to this as mean pairwise approach. We examined also general, mixed components and co-diversity approaches, as well as the Jaccard, Simpson and Sørensen families. In sum, we define any particular method as the combination of an approach and a family, and thus compared 12 methods.

7.2 Artificial data 1

To compare the performance of methods, we examined all possible communities that can be produced by the co-occurrence of 4 species in 4 sites. In order to calculate the number of possibilities, we have to first determine how many ways a single species can be distributed in *N* sites. Since there are two outcomes for each site (the species is present or absent), the possible number of occurrence patterns equals 2^N. However, this includes the situation when the species is absent from all sites. Therefore, the number of occurrence patterns reduces to 2^N-1 (for $N = 4$ we have 15 different patterns). When we have *T* species, then the possible number of co-occurrence patterns increases dramatically $(2^N\text{-}1)^T$ (for $N = T = 4$ we get 50,625). However, these co-occurrence patterns include empty sites (those without species) as well. After removing degenerate matrices, the number of meaningful co-occurrence patterns reduces to 41,503 in the example.

We calculated multiple-site similarities by the different methods (i.e. the combinations of families and approaches) for each of the 41,503 occurrence patterns. When no similarity form was given (the Jaccard and Sørensen families of mixed components), we used the complement of dissimilarity. The resulting scores served as a data set to calculate the Pearson correlation between different methods, in order to express agreement in trends among the measures. We transformed the correlations to distances (distance = 1 – correlation) and analyzed the distance matrix by UPGMA clustering to obtain a dendrogram. The same distance matrix was analyzed by principal coordinates analysis (PCoA). Thus, in these multivariate studies, each object represents a given measure. We used the *gtools* (Warnes et al. 2014), the *betapart* (Baselga et al. 2013) packages in *R* (R Core Team, 2015) and the SYN-TAX 2000 package (Podani 2001) for computations.

The dendrogram (Fig. 1) shows that methods belonging to the Simpson family constitute one group, separated from the methods of the Jaccard and Sørensen families grouped in the other. Within the latter, general coefficients are well-separated and grouping is more strongly influenced by the choice of approach than by the family (Fig. 1). The PCoA ordination of the methodologies (Fig. 2) supports these conclusions. The first axis separates the Simpson family from the Jaccard and Sørensen families, while the second separates the general approach from the others. Since these axes account for 44% and 29% of the total variance, respectively, we can conclude that choice between families had stronger impact on the results than another decision between the general overlap approach and the others.

7.3 Artificial data 2

Artificial data set 1 allowed examining all theoretical possibilities in a matrix with very few sites and species. To obtain a more realistic picture on the relationships among measures, we generated a second artificial data set that is closer to actual community data. We produced 150 sets of 10 sites by 10 species incidence matrices, in which the probability of

281 the occurrence of a species in a particular site was 0.5. We removed degenerate matrices
282 (i.e. those with zero row or column totals) and used the first 100 matrices. We followed the
283 multivariate exploration procedure applied to artificial data set 1. The dendrogram (Fig. 3)
284 shows that methods belonging to the Simpson family form one group, and methods
285 belonging to the Jaccard and Sørensen families appear in another. Within the second group,
286 general coefficients are well-separated. The difference between Figs. 1 and 3 are that Fig. 3
287 shows larger distances among some groups of methods (the maximum distance is larger
288 than 0.5) and at the same time smaller distances among similar methods (the behavior of
289 MC.JAC and MCSOR is similar). The PCoA ordination of the measures (Fig. 4) resulted in
290 much the same conclusions. The separation of the G.SIM from the other methods is clear.
291 On the first axis the Simpson family is distinguished from the Jaccard and Sørensen families,
292 while the second axis separates the general approach from the others. Since these axes
293 account for 68% and 16% of the total variance, respectively, we can conclude that choice
294 between families had stronger impact on the results than the decision between the general
295 overlap approach and the others. We may thus derive the final conclusion from clustering
296 and ordination that the general indices, especially those belonging to the Simpson family,
297 present a rather unique way of calculating multiple-site similarities.

298

299 *7.4 Actual data set*

300

301 Rey (1981) examined the recolonization of islets by arthropods after defaunization by
302 insecticides. The fauna was recorded every week for more than a year; we took the data
303 from the 10th, 13th, 20th and 53rd weeks after treatment. These four data matrices,
304 published in Atmar & Patterson (1995) contain 6 sites and 25, 27, 33 and 33 species,
305 respectively. An analysis equivalent to the mean pairwise Jaccard method indicated a
306 monotonic increase of similarity over the study period (Podani & Schmera 2011).

307

308 We found that all methods compared here indicate a monotonically increasing similarity
309 over time (Fig. 5). Nonetheless, the methods show considerable differences regarding the
310 multiple site similarity in the four assemblages. For instance, in week 53, the Jaccard family

co-diversity index (CD.JAC) yields a similarity value of 0.131, while the Simpson family general index (G.SIM) produces 0.698. This suggests that selection of the methodology (i.e. the choice of the family together with the approach) has significant impact on our inference about community pattern (here similarity). It is important to note that the traditionally used mean of pairwise indices (here abbreviated as mean pairwise method) and the other "true" multiple-site indices produced very different results, suggesting that the pairwise and multiple-site measures are complementary.

8. Conclusions

We emphasized that understanding and interpreting the multiple-site community patterns pose relevant methodological issues of contemporary ecology and biogeography. Our review demonstrated that a wide variety of methods have been available for quantifying multiple-site resemblance patterns. To help the ecologist navigating among them, we suggested a classification of methodology. Accordingly, a method is a combination of an index family and an approach. Analyses of simulated and actual data sets revealed that inference drawn on community pattern strongly depends on the applied method: multiple-site incidence coefficients quantify different facets of multiple-site community patterns. In particular, we found that the impact of choosing from original pairwise index families was stronger on quantifying multiple-site resemblance patterns than the impact of selecting different approaches. Thus, any methodology used for studying multiple-site community patterns should be carefully evaluated before use.

Acknowledgements

We thank Hector Arita and Carlo Ricotta for their constructive comments on an earlier draft of the manuscript. This research was supported by GINOP-2.3.2-15-2016-00019 project.

References

Arita, H.T. (2017) Multisite and multispecies measures of overlap, co-occurrence, and co-diversity. Ecography, 40: 709-718.

Arita, H.T, Christen, J.A., Rodriguez, P. & Soberon, J. (2008) Species diversity and distribution in presence-absence matrices: mathematical relationships and biological implications. The American Naturalist, 172, 519-532.

Arita, H.T, Christen, J.A., Rodriguez, P. & Soberon, J. (2012) The presence-absence matrix reloaded: the use and interpretation of range-diversity plots. Global Ecology and Biogeography, 21, 282-292.

Atmar, W & Patterson, B.D. (1995) The nestedness temperature calculator: a visual basic program, including 294 presence-absence matrices. AICS Research Inc., Univ. Park, NM and The Field Museum, Chicago, IL

Baselga, A. (2010) Partitioning the turnover and nestedness components of beta diversity. Global Ecology and Biogeography, 19, 134-143.

Baselga, A. (2012) The relationship between species replacement, dissimilarity derived from nestedness and nestedness. Global Ecology and Biogeography, 21, 1223-1232.

Baselga, A. (2013) Multiple site dissimilarity quantifies compositional heterogeneity among several sites, while average pairwise dissimilarity might be misleading. Ecography, 36, 124-128.

Baselga, A., Jiménez-Valverde, A. & Niccolini, G. (2007) A multiple-site similarity measures independent of richness. Biology Letters, 3, 642-645.

Baselga, A. & Leprieur, F. (2015) Comparing methods to separate components of beta diversity. Methods in Ecology and Evolution, 6, 1069-1079.

366 Baselga, A., Orme, D., Villeger, S., de Bortoli, J & Leprieur, F. (2013) betapart: Partitioning
367 beta diversity into turnover and nestedness components. R package version
368 1.3.http://CRAN.R-project.org/package=betapart

369 Bell, G. (2005) The co-distribution of species in relation to the neutral theory of community
370 ecology. Ecology, 86, 1757-1770.

371 Cardoso, P., Rigal, F., Carvalho, J.C., Fortelius, M., Borges, P.A.V., Podani, J. & Schmera, D.
372 (2014) Partitioning taxon, phylogenetic and functional beta diversity into replacement
373 and richness difference components. Journal of Biogeography, 41, 749-761.

374 Carvalho J.C., Cardoso, P., Borges, P.A.V., Schmera, D. & Podani, J (2013) Measuring fractions
375 of beta diversity and their relationships to nestedness: a theoretical and empirical
376 comparison of novel approaches. Oikos, 122, 825-834.

377 Chao, A., Chiu, C.C. & Hsieh, T.C. (2012) Proposing resolution to debates on diversity
378 partitioning. Ecology, 93, 2037-2051.

379 Chen, Y. (2016) Partitioning multiple-site tree-like beta diversity into turnover and
380 nestedness components without pairwise comparisons. Ecological Indicators, 61, 413-
381 417.

382 Diserud, O.H. & Ødegaard, F. (2007) A multiple-site similarity measure. Biology Letters, 3, 20-
383 22.

384 Gotelli, N.J. & Chao, A. (2013) Measuring and estimating species richness, species diversity,
385 and biotic similarity from sampling data. Encyclopedia of Biodiversity, vol. 5, pp. 195-
386 211.

387 Jaccard, P. (1912) The distribution of the flora in the alpine zone. New Phytologist, 11, 37-50.

388 Jost, L., Chao, A. & Chazdon R.R. (2011) Compositional similarity and β (beta) diversity. In
389 Magurran A.E. & McGill B.J. (eds) Biological diversity. Frontiers in measurement and
390 assessment. Oxford University Press, pp. 66-84.

391 Koch, L.F. (1957) Index of biotic dispersity. Ecology, 38, 145-148.

392 Koleff, P., Gaston, K.J. & Lennon, J.L. (2003) Measuring beta diversity for presence-absence
393 data. Journal of Animal Ecology, 72, 367-382.

394 Legendre, P. (2014) Interpreting the replacement and richness difference components of
395 beta diversity. Global Ecology and Biogeography, 23, 1324-1334.

396 Legendre, P. & De Cáceres M. (2013) Beta diversity as the variance of community data:
397 dissimilarity coefficients and partitioning. Ecology Letters, 16, 951-963.

398 MacKenzie D.I., Bailey L.L & Nichols J.D. (2004) Investigating species co-occurrence patterns
399 when species are detected imperfectly. Journal of Animal Ecology, 73, 546-555.

400 Orlóci, L. 1972. On objective functions of phytosociological resemblance. American Midland
401 Naturalist, 88, 28-55.

402 Podani, J. (2001) SYN-TAX 2000. Computer programs for data analysis in ecology and
403 systematics. Scientia, Budapest.

404 Podani, J. & Schmera, D. (2011) A new conceptual and methodological framework for
405 exploring and explaining pattern in presence-absence data. Oikos, 120, 1625-1638.

406 Podani, J & Schmera, D. (2016) Once again on the components of pairwise beta diversity.
407 Ecological Informatics, 32, 63-68.

408 Pollock L.J., Tingley R.,Morris W.K., Golding N., O'Hara R.B., Parris K.M., Vesk P.A. &
409 McCarthy M.A. (2014) Understanding co-occurrence by modelling species
410 simultaneously with a Joint Species Distribution Model (JSDM). Methods in Ecology
411 and Evolution, 5, 397–406

412 R Core Team (2015). R: A language and environment for statistical computing. R Foundation
413 for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

414 Rey, J.R. (1981) Ecological biogeography of arthropods on Spartina islands in northwestern
415 Florida. Ecological Monographs, 51, 237-265.

416 Ricotta, C. & Pavoine, S. (2015) A multiple-site dissimilarity measure for species
417 presence/absence data and its relationship with nestedness and turnover. Ecological
418 Indicators, 54, 203-206.

419 Schmera D (2017) On the operative use of community overlap in analyzing incidence data.
420 Community Ecology, 18, 117-119.

421 Simpson, G.G. (1943) Mammals and the nature of continents. American Journal of Science,
422 241, 1-31.

423 Sørensen, T.A. (1948) A method of establishing groups of equal amplitude in plant sociology
424 based on similarity of species content, and its application to analyses of the vegetation
425 on Danish commons. Kongelige Danske Viden- skabernes Selskabs Biologiske Skrifter,
426 5, 1–34.

427 Trejo-Barocio, P. & Arita, H.T. (2013) The co-occurrence of species and the co-diversity of
428 sites in neutral models of biodiversity. Plos One, 8, e79918.

429 Warnes GR, Bolker B, Lumley T (2014) gtools: Various R programming tools. R package
430 version 3.4.1. http://CRAN.R-project.org/package=gtools

431 Whittaker, R.H. (1960) Vegetation of the Siskiyou Mountains, Oregon and California.
432 Ecological Monographs, 30, 279-338.

433 Whittaker, R.H. (1972) Evolution and measurement of species diversity. Taxon, 21, 213-251.

434

435 TABLES

436 Table 1. The most important properties of three well known pairwise resemblance coefficients

Family	Form	Equation	Eq. n.	Interpretation	Reference
Jaccard	similarity	$\frac{a}{a+b+c}$	Tab1/1	the ratio of the number of shared species to the total number of species	Jaccard (1912)
	dissimilarity	$\frac{b+c}{a+b+c}$	Tab1/2	the ratio of the number of unique species to the total number of species	
Simpson	similarity	$\frac{a}{a+\min(b,c)}$	Tab1/3	the ratio of the number of shared species to the number of species at the poorest site	Simpson (1943)
	dissimilarity	$\frac{\min(b,c)}{a+\min(b,c)}$	Tab1/4	the ratio of the number of species unique to the poorest site and the number of species in the poorest site	
Sørensen	similarity	$\frac{2a}{2a+b+c}=\frac{a}{\frac{1}{2}((a+b)+(a+c))}$	Tab1/5	the ratio of the number of shared species to the mean number of species in a single site	Sørensen (1948)
	dissimilarity	$\frac{b+c}{2a+b+c}=\frac{\frac{1}{2}(b+c)}{\frac{1}{2}((a+b)+(a+c))}$	Tab1/6	the ratio of the mean unique species to the mean number of species in a single site	

437

438

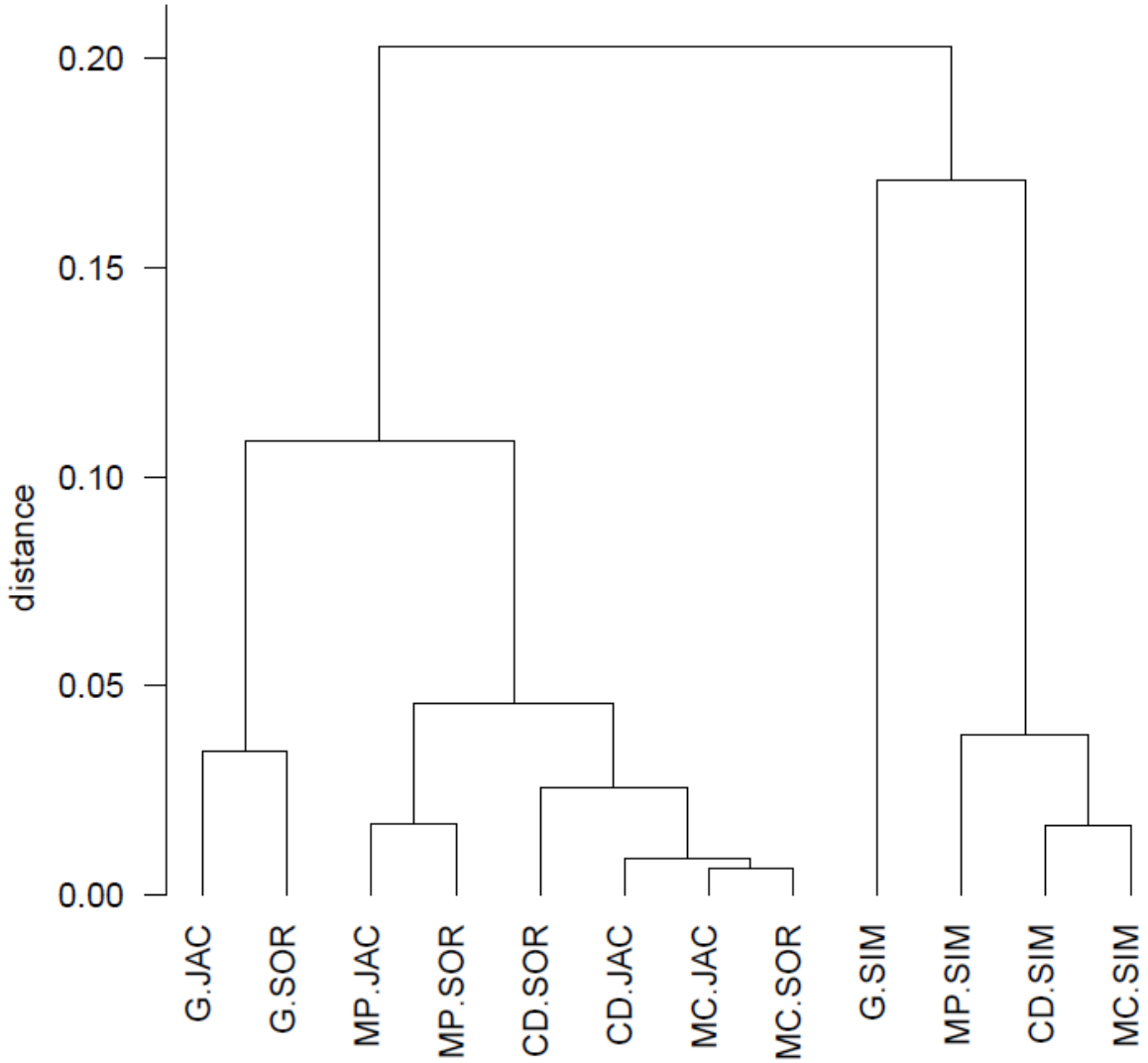
Table 2: Overview of multiple site resemblance coefficients (N: number of sites, T: total number of species, t_i: number of species at site i, n_j: number of sites where species j occurs, o: rank of a species richness value in the order from the smallest to the largest values, g_o: the frequency of sites with species richness of rank o, b_{kl}: number of species unique to site k in pairwise comparison with site l, b_{lk}: number of species unique to site l in pairwise comparison with site k).

Approach	Family	Form	Equation	Eq. n.	Reference
General	Jaccard	similarity	$\frac{\sum_{i=1}^{N} t_i - T}{T(N-1)}$	Tab2/1	Koch (1957), Chao et al. (2012) Gotelli & Chao (2013), Arita (2017)
		dissimilarity	$\sum_{j=1} \frac{N-n_j}{T \times N} \times \frac{N}{N-1}$	Tab2/2	Ricotta & Pavoine (2015, in their Appendix S2)
	Simpson	similarity	$\frac{\sum_{j=1} n_j - T}{\sum_{j=1} p_j - \max(t_i)}$	Tab2/3	Arita (2017)
	Sørensen	similarity	$\frac{N}{N-1}(1-\frac{T}{\sum_{i=1} t_i})$	Tab2/4	Diserud & Ødegaard (2007), Chao et al. (2012), Gotelli & Chao (2013) and Arita (2017)
		dissimilarity	$\frac{\sum_i (T-t_i)}{(N-1)\sum_i t_i}$	Tab2/5	Ricotta & Pavoine (2015, in their Appendix S2)
Mixed components	Jaccard	dissimilarity	$\frac{[\sum_{k<l} \min(b_{kl}, b_{lk})] + [\sum_{k<l} \max(b_{kl}, b_{lk})]}{[\sum_i t_i - T] + [\sum_{k<l} \min(b_{kl}, b_{lk})] + [\sum_{k<l} \max(b_{kl}, b_{lk})]}$	Tab2/6	Baselga (2012)
	Simpson	similarity	$\frac{\sum_i t_i - T}{(\sum_i t_i - T) + \sum_{k<l} \min(b_{kl}, b_{lk})}$	Tab2/7	Baselga et al. (2007)
	Sørensen	dissimilarity	$\frac{[\sum_{k<l} \min(b_{kl}, b_{lk})] + [\sum_{k<l} \max(b_{kl}, b_{lk})]}{2[\sum_i t_i - T] + [\sum_{k<l} \min(b_{kl}, b_{lk})] + [\sum_{k<l} \max(b_{kl}, b_{lk})]}$	Tab2/8	Baselga (2010)
Co-diversity	Jaccard	similarity	$\frac{\sum_{j=1} n^2 - \sum_{j=1} n}{TN(N-1)}$	Tab2/9	Arita (2017)
	Simpson	similarity	$\frac{\sum_{j=1} n^2 - \sum_{j=1} n}{2\sum_{o=1}^{N} (N-o) g_o}$	Tab2/10	Arita (2017)
	Sørensen	similarity	$\frac{\sum_{j=1} n^2 - \sum_{j=1} n}{(N-1)\sum_{i=1} t_i}$	Tab2/11	Arita (2017)

448

449 FIGURES

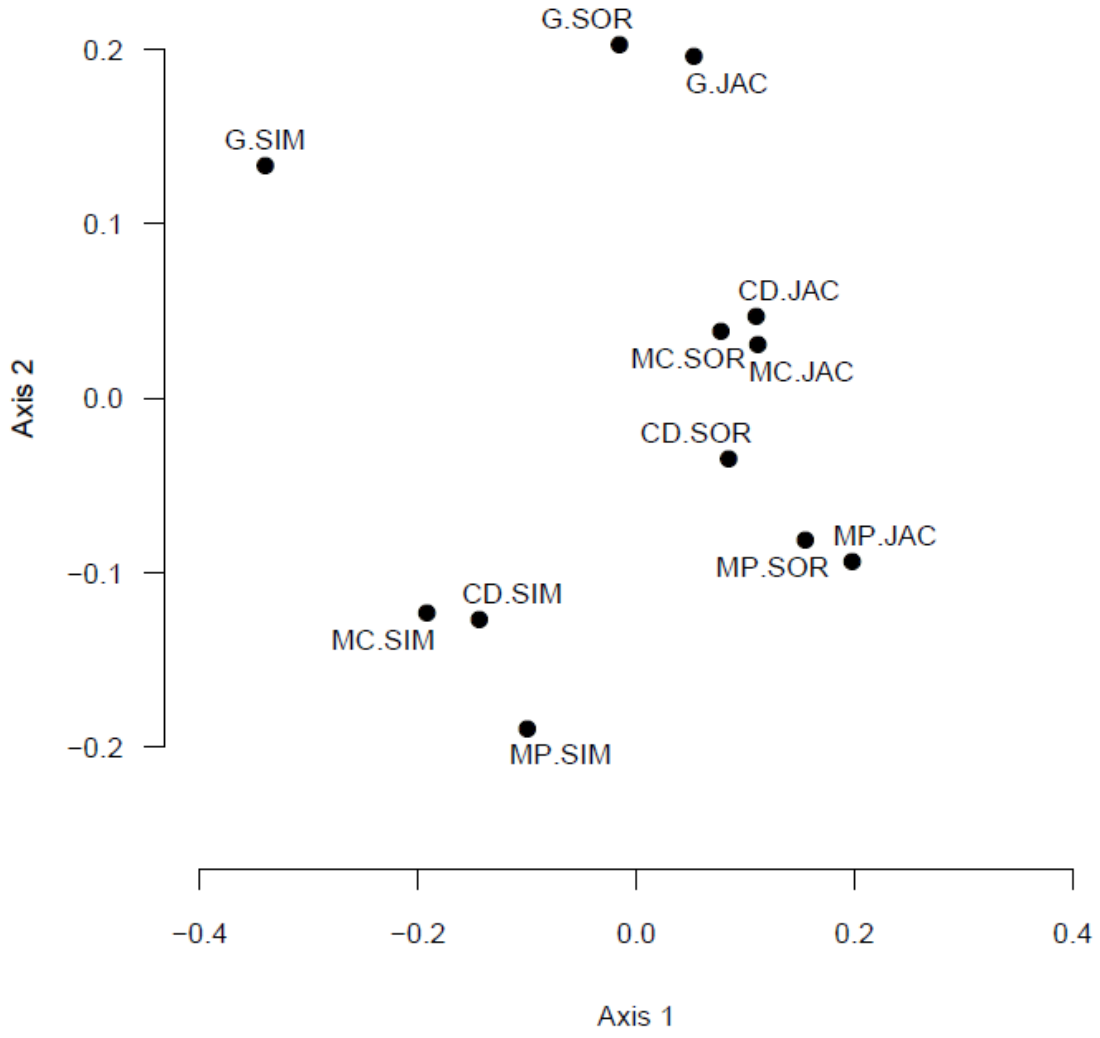
450



451

452 Fig. 1: UPGMA clustering of methods quantifying multiple-site similarities using 1 –
453 correlation as distance for 41,503 different data sets with 4 species and 4 sites.
454 Abbreviations include the combination of an approach and a family, where one or two
455 letters denote an approach (MP: mean pairwise, G: general, MC: mixed component and CO:
456 co-diversity) and after a dot three letters denote a family (JAC: Jaccard, SIM: Simpson and
457 SOR: Sørensen).

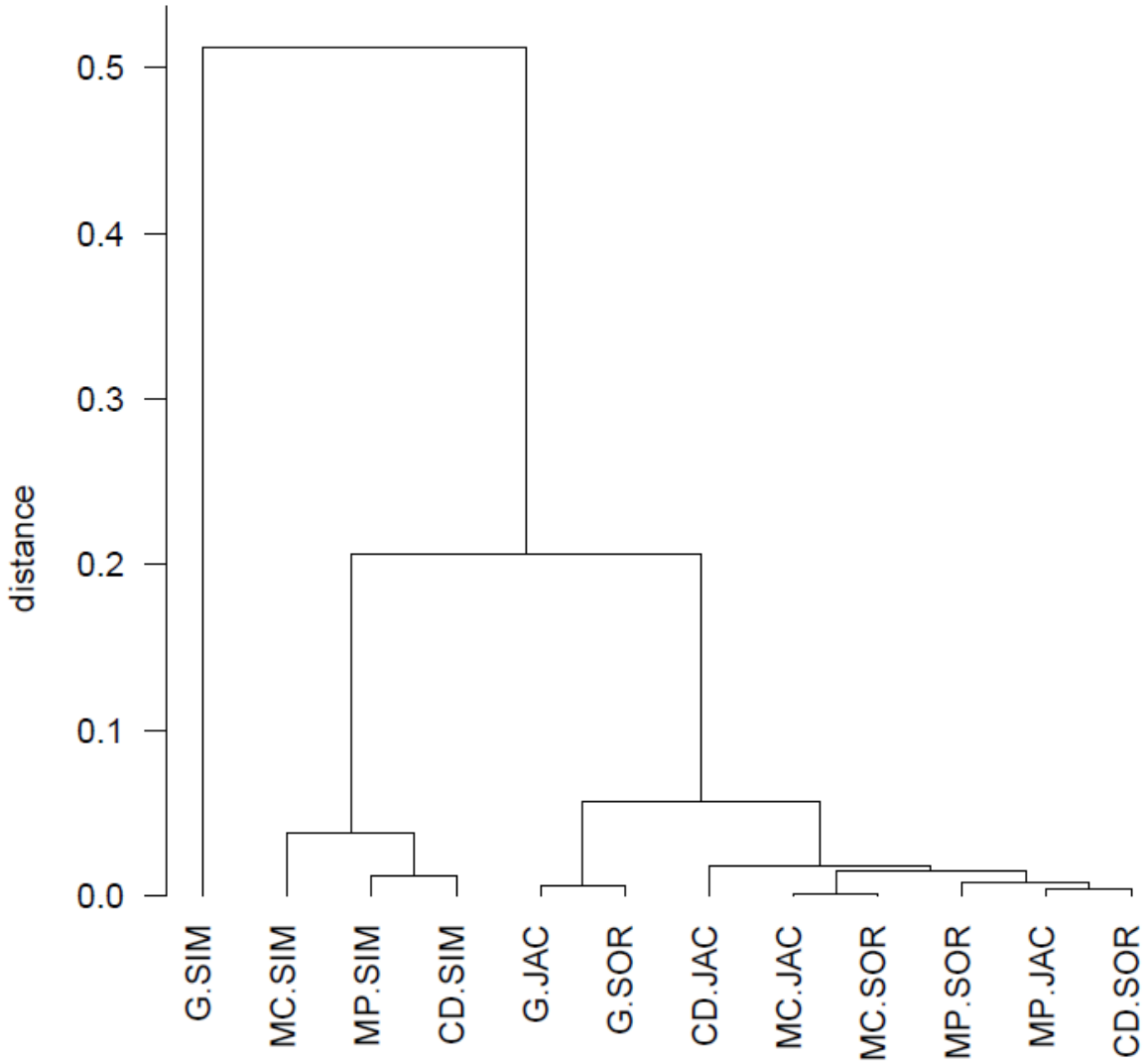
458



459

460 Fig 2: Principal coordinates analysis of methods quantifying multiple-site similarities using 1
461 – correlation as distance for 41,503 different data sets with 4 species and 4 sites. For
462 abbreviations, see caption to Fig. 1.

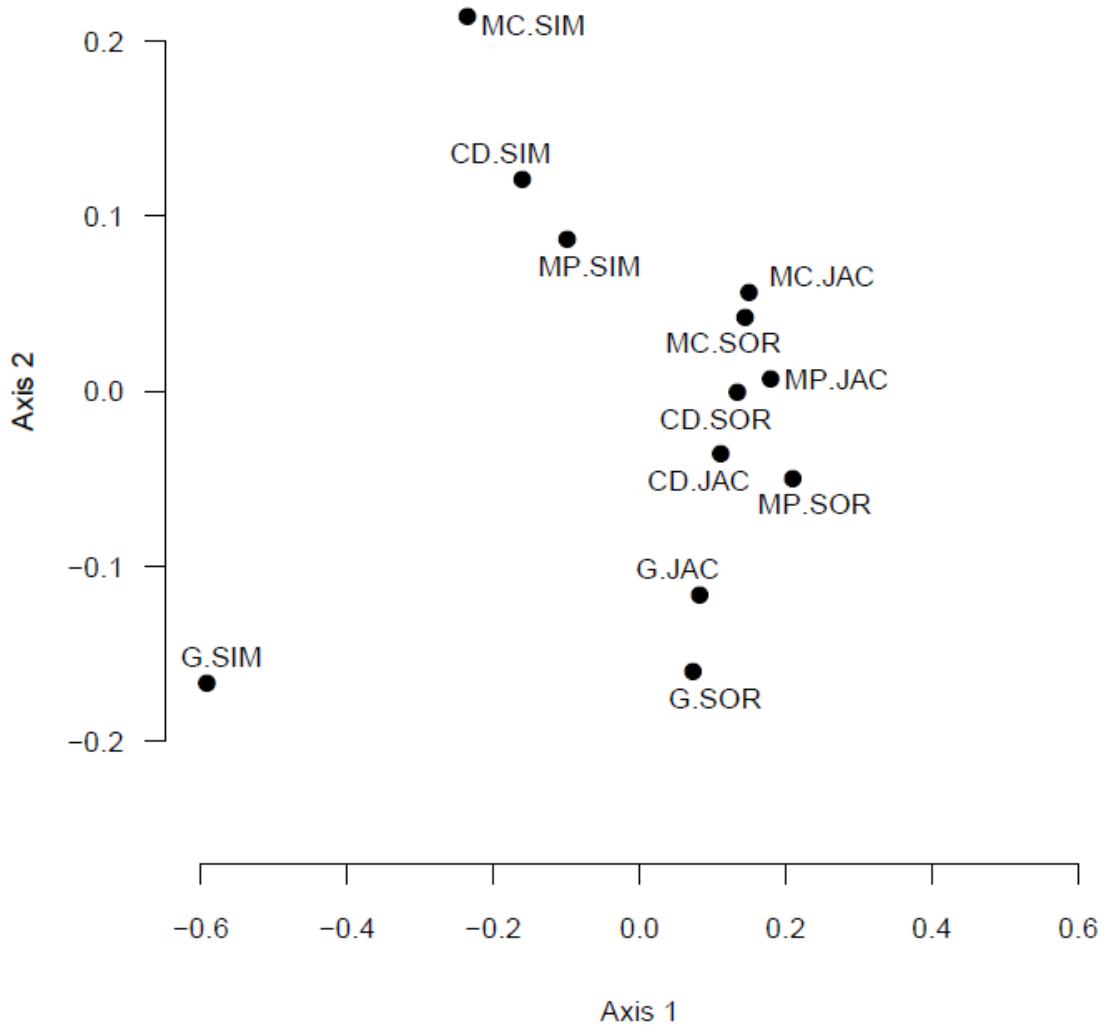
463



464

465 Fig. 3: UPGMA clustering of methods quantifying multiple-site similarities using 1 –
466 correlation as distance for 100 different data sets with 10 species and 10 sites. For
467 abbreviations, see caption to Fig. 1.

468

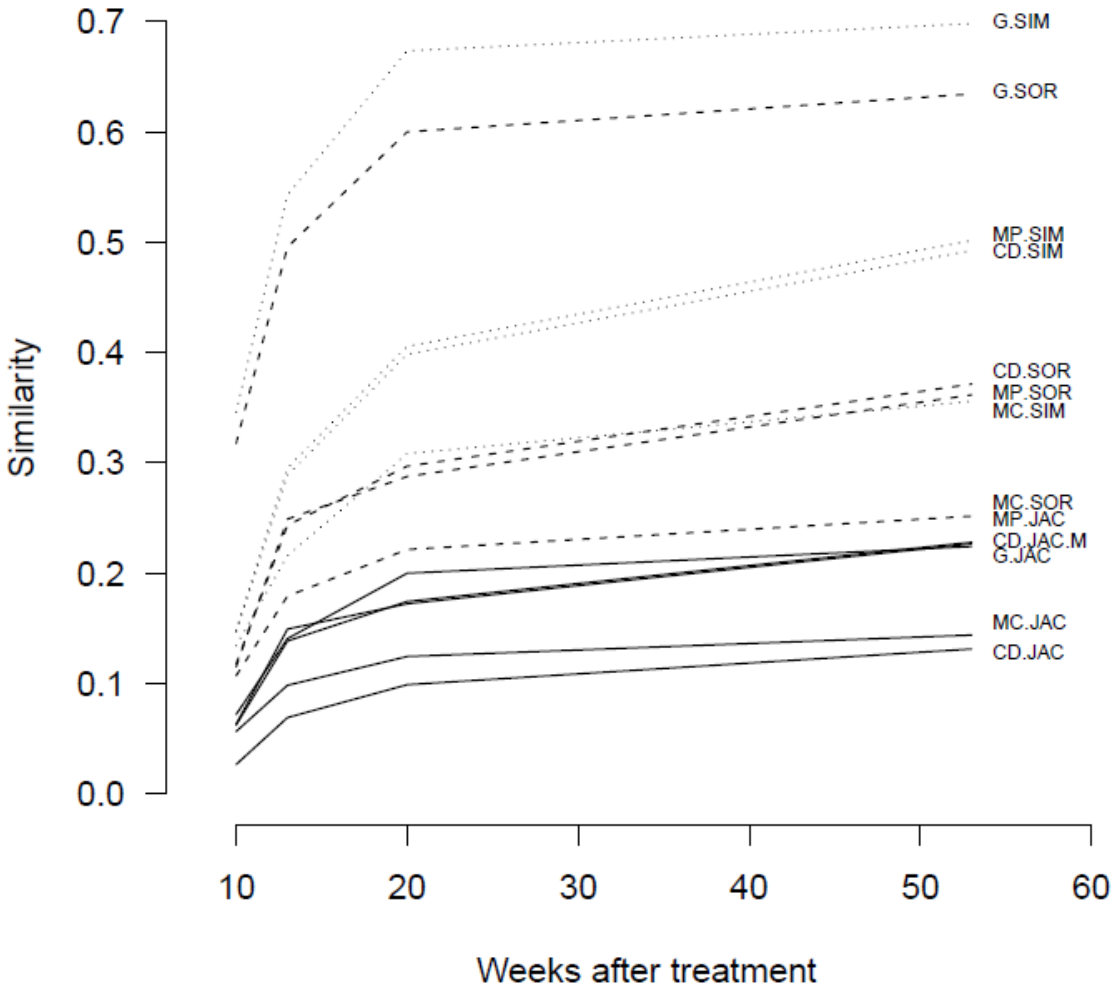


469

470 Fig 4: Principal coordinates analysis of methods quantifying multiple-site similarities using 1-
471 correlation as distance for 100 different data sets with 10 species and 10 sites. For
472 abbreviations, see caption to Fig. 1.

473

474



475

476 Fig. 5: Change of community similarity over time (in weeks) depicted by 13 multiple-site
477 similarity indices. For clarity, data points are connected. For abbreviations, see caption to
478 Fig. 1.

479