

Ecological resemblance

Theory R functions Examples

Quantifying ecological resemblances between samples, including similarities and dissimilarities (or distances), is the basic approach of handling multivariate ecological data. Two samples, which contain the same species with the same abundances, have the highest similarity (and lowest dissimilarity or distance); the similarity decreases (and dissimilarity/distance increases) with the differences in their species composition. All cluster and ordination methods operate with similarities or distances between samples.

Similarity, dissimilarity and distance

Intuitively, one thinks about **similarity** among objects - the more are two objects similar in terms of their properties, the higher is their similarity. In the case of species composition data, the similarity is calculated using similarity indices, ranging from 0 (the samples do not share any species) to 1 (samples have identical species composition). Ordination techniques are usually based on distances, because they need to localize the samples in a multidimensional space; clustering methods could usually handle both similarities or distances. **Distances** are of two types, either dissimilarity, converted from analogous similarity indices, or specific distance measures, such as Euclidean, which doesn't have a counterpart in any similarity index. While all similarity indices can be converted into distances, not all distances could be converted into similarities (as is true, e.g. for Euclidean distance).

There is a number of measures of similarities or distances ([Legendre & Legendre 2012](#) list around 30 of them). The first decision one has to make is whether the aim is R- or Q-mode analysis (R-mode focuses on differences among species, Q-mode on differences among samples), since some of the measures differ between both modes (e.g. Pearson's r correlation coefficient makes sense for the association between species (R-mode), but not for the association between samples (Q-mode); in contrast, e.g. Sørensen index can be used in both Q- and R-mode analysis, and is called Dice index in R-mode analysis). Further, if focusing on differences between samples (Q-mode), the most relevant measures in ecology are asymmetric indices ignoring double zeros (more about *double-zero problem* below). Then, it also depends whether the data are qualitative (i.e. binary, presence-absence) or quantitative (species abundances). In the case of distance indices, an important criterium is whether they are metric (they can be displayed in Euclidean space) or not, since this influences the choice of the index for some ordination or clustering methods.

[Legendre & Legendre \(2012\)](#) offers a key how to select an appropriate measure for given data and problem (check their Tables 7.4-7.6). Generally, as a rule of thumb, Bray-Curtis and Hellinger distances may be better choices than Euclidean or Chi-square distances.

Double-zero problem

“Double zero” is a situation when certain species is missing in both compared community samples for which similarity/distance is calculated. In [Fig. 1](#) you can see an ecological example of the “double zero” problem. Samples 1 to 3 are sorted according to the wetness of their habitat - sample 1 is the wettest, and sample 3 is the driest. In samples 1 and 3, the species which is labelled as “mesic” does

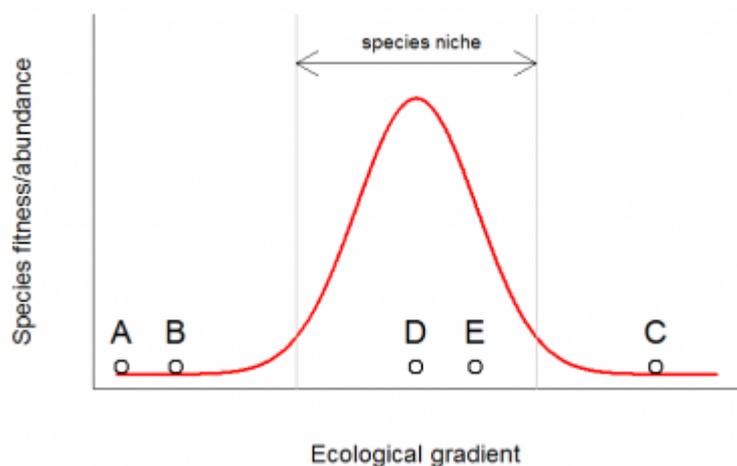
not occur, since sample 1 is too wet and sample 3 too dry - these are the two zeros forming the "double zero" problem. The fact that the mesic species is missing does not say anything about ecological similarity or difference between both samples.

Figure 1: For details see the text.

	preferences of species for moisture				
	← wet-site indicators				dry-site indicators →
	hygrophilic species 1	hygrophilic species 2	mesic species	xeric species 1	xeric species 2
sample 1	1	1	0	0	0
sample 2	0	1	1	1	0
sample 3	0	0	0	1	1

Species missing simultaneously in two samples can mean the following: (1) samples are located outside of the species ecological niche, but one cannot say whether both samples are on the same side of the ecological gradient (i.e. they can be rather ecologically similar, samples A and B on Fig. 2) or they are on the opposite sides (and hence very different, samples A and C). Alternatively, (2) samples are located inside species ecological niche (samples D and E), but the species in given samples does not occur since it didn't get there (dispersal limitation), or it was present, but overlooked and not sampled (sampling bias). In both cases, the double zero represents missing information, which cannot offer an insight into the ecological similarity of compared samples.

Figure 2: Response curve of a single species along environmental gradient; A, B..., E are samples located within or outside the species niche.



Both similarity and distance indices differ in a way how they approach the double-zero problem. **Symmetrical indices** treat double zero (0-0) in the same way (symmetrically) as double presences (1-1), i.e. as a reason to consider samples similar. As explained above, this is not meaningful for species composition data, but can be meaningful for other multivariate data (e.g. environmental dataset containing chemical measurements; for example, the fact that heavy metals are missing in both samples really indicates similarity between both samples). **Asymmetrical indices** treat double zeros and double presences asymmetrically - they ignore double zeros, and focus only on double presences when evaluating the similarity of samples; these indices are usually more meaningful for species composition data. In the context of the ecological example (Fig. 1), the absence of mesic species in sample 1 and sample 3 (0-0, double zero) will increase the similarity of sample 1 and 3 if measured by symmetrical indices of similarity, but will not influence the similarity of these samples if measured by asymmetrical indices (double zeros will be ignored, and only presences, namely 1-1,

1-0, and 0-1, will be considered).

Similarity indices

Categories of similarity indices are summarized in [Tab. 1](#). Symmetrical indices, i.e. those which consider double zeros as relevant, are not further considered here, since they are not useful for analysis of ecological data (although they may be useful for other types of data, e.g. environmental variables). Here we will consider only asymmetrical similarity indices, i.e. those ignoring double zeros. These can be divided into two types, according to the data which they are using: **qualitative** (binary) indices, applied on presence-absence data, and **quantitative** indices, applied on raw (or transformed) species abundances. It is important to mention that presence-absence (qualitative) and abundance (quantitative) species composition data carry different type of information, and their analysis may have a different meaning - see the section [Presence-absence vs quantitative species composition data](#)). Some of the similarity indices also have multi-sample alternatives (i.e. they could be calculated on more than two samples), which could be used to calculate beta diversity.

Similarity indices		How they deal with <i>double zero</i> problem?	
		symmetrical (treat double zeros as important information)	asymmetrical (ignore double zeros)
Which type of data indices use?	qualitative (binary = presence absence data)	not suitable for ecological data	Jaccard similarity, Sørensen similarity, Simpson similarity
	quantitative (species abundances)	not suitable for ecological data	Percentage similarity ¹⁾

Table 1: Similarity indices classified according to their properties.

Qualitative (binary) asymmetrical similarity indices use information about the number of species shared by both samples, and numbers of species which are occurring in the first or the second sample only (see the schema at [Tab. 2](#)). **Jaccard similarity index** divides the number of species shared by both samples (fraction a) by the sum of all species occurring in both samples ($a+b+c$, where b and c are numbers of species occurring only in the first and only in the second sample, respectively). **Sørensen similarity index** considers the number of species shared among both samples as more important, so it counts it twice. **Simpson similarity index** is useful in a case that compared samples largely differ in species richness (i.e. one sample has considerably more species than the other). If Jaccard or Sørensen are used on such data, their values are generally very low, since the fraction of species occurring only in the rich sample will make the denominator too large and the overall value of the

$$\text{Jaccard similarity: } J = \frac{a}{a+b+c}$$

$$\text{Sørensen similarity: } S = \frac{2a}{2a+b+c}$$

$$\text{Simpson similarity: } S_i = \frac{a}{a+\min(b,c)}$$

index too low; Simpson index, which was originally introduced for comparison of fossil data, eliminates this problem by taking only the smaller from the fractions *b* and *c*. (Note that there is yet another Simpson index, namely *Simpson diversity index*; each of the indices was named after different person surnamed Simpson; while the Simpson similarity index is calculating the similarity between pair of compositional samples, the Simpson diversity index is calculating diversity of a single community sample; you may find details in my [blog post](#) on this topic).

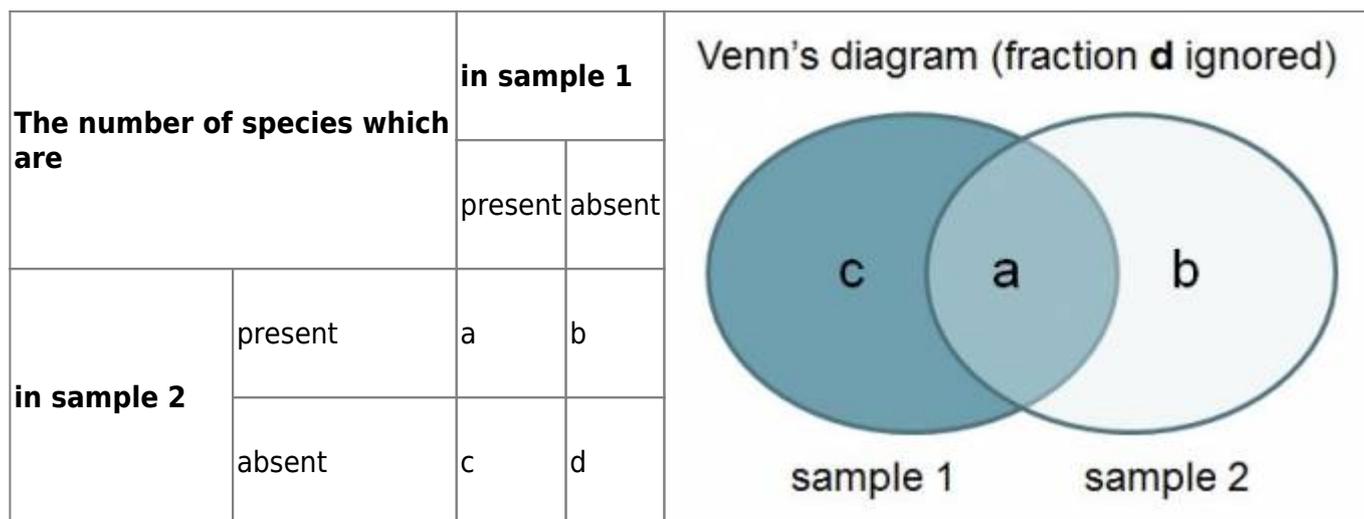


Table 2: The meaning of fraction a, b, c and d used in qualitative indices calculating similarity among two samples. In assymetric indices, the fraction d (double zero) is ignored.

Quantitative similarity indices (applied on quantitative abundance data) include **percentage similarity**, a quantitative version of Sørensen similarity index (if calculated on presence-absence data, it gives the same results as Sørensen similarity index). Note that *percentage difference*, calculated as 1-*percentage similarity*, is called Bray-Curtis distance index (see below).

Percentage similarity: $PS = \frac{2W}{A+B}$, where *W* is the sum of minimum abundances of various species; *A* and *B* each are sum of abundances of all species at each compared site:

	Species abundances						A	B	W
Site x₁	7	3	0	5	0	1	16		
Site x₂	2	4	7	6	0	3		22	
Minimum	2	3	0	5	0	1			11

$$PS = \frac{2W}{A+B} = \frac{2 \times 11}{16+22} = \frac{22}{38} = 0.579$$

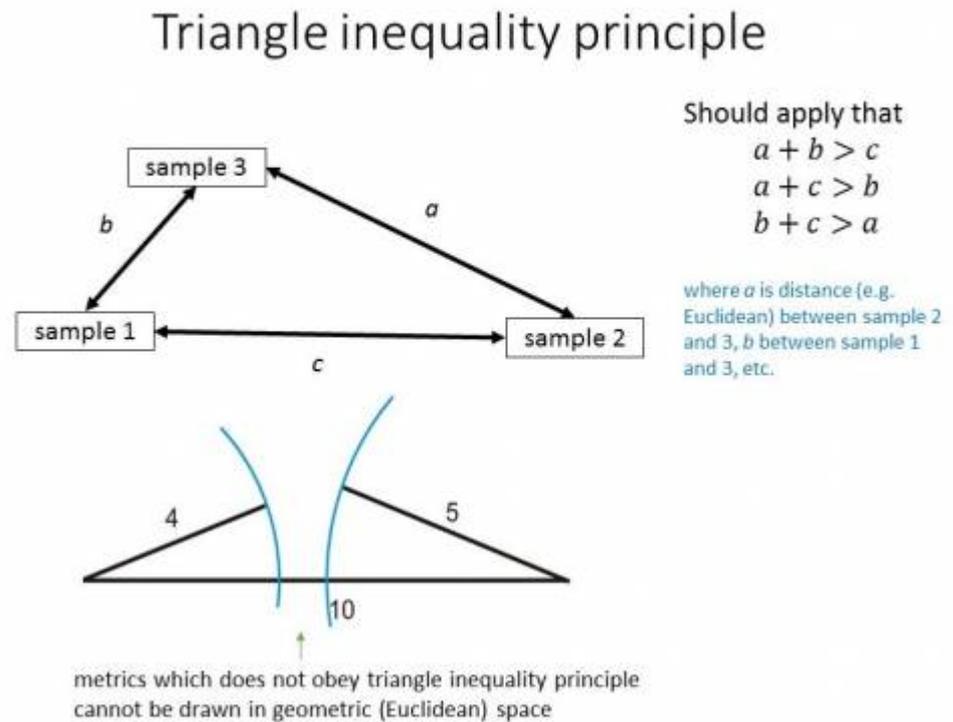
Distance indices

While similarity indices return the highest value in the case that both compared samples are identical

(maximally similar), distance indices are largest for two samples which do not share any species (are maximally dissimilar). There are two types of distance (or dissimilarity) indices²⁾:

1. those **calculated from similarity indices**, usually as $D = 1 - S$, where S is the similarity index; examples include Jaccard, Sørensen and Simpson dissimilarity for qualitative (binary) data and percentage difference (known also as Bray-Curtis distance) for quantitative data;
2. those **distances which have no analogue in the similarity indices**, e.g. Euclidean, chord, Hellinger or chi-square distance index.

Figure 3: Triangle inequality principle.



An important criterium is **whether the distance index is metric or not** (i.e. it is semi-metric or non-metric). The term “metric” refers to the distance indices that obey the following four metric properties: 1) minimum distance is zero, 2) distance is always positive (unless it is zero), 3) the distance between sample 1 and sample 2 is the same as the distance between sample 2 and sample 1, and 4) triangle inequality (see explanation in Fig. 3). Indices that obey the fourth, triangle-inequality principles, can be displayed in the orthogonal Euclidean space, and are sometimes labelled as having Euclidean property (note that Euclidean distance is just one of several distance indices which has such Euclidean property). Some distance indices calculated from similarities are metric (e.g. Jaccard dissimilarity), others are not (e.g. Sørensen dissimilarity and its quantitative version, called Bray-Curtis distance, are semi metric; some other distances may be nonmetric - they can reach negative values, which is nonsensible for ecological data). In the case of Sørensen and Bray-Curtis (and some others), this can be solved by calculating the dissimilarity as $D = \sqrt{1 - S}$ instead of the standard $D = 1 - S$ (where S is the similarity); resulting dissimilarity index is then metric. Indices which are not metric cause troubles in ordination methods relying on Euclidean space (PCoA or db-RDA) and numerical clustering algorithms which need to locate samples in the Euclidean space (such as Ward algorithm or K-means). For example, PCoA calculated using distances that are not metric creates axes with negative eigenvalues. In some analyses (e.g. in db-RDA), negative eigenvalues may result in virtually higher variation explained by explanatory variables than would reflect the data, and need to be carefully avoided.

Bray-Curtis dissimilarity or **percentage difference**³⁾ is one complement of *percentage similarity*

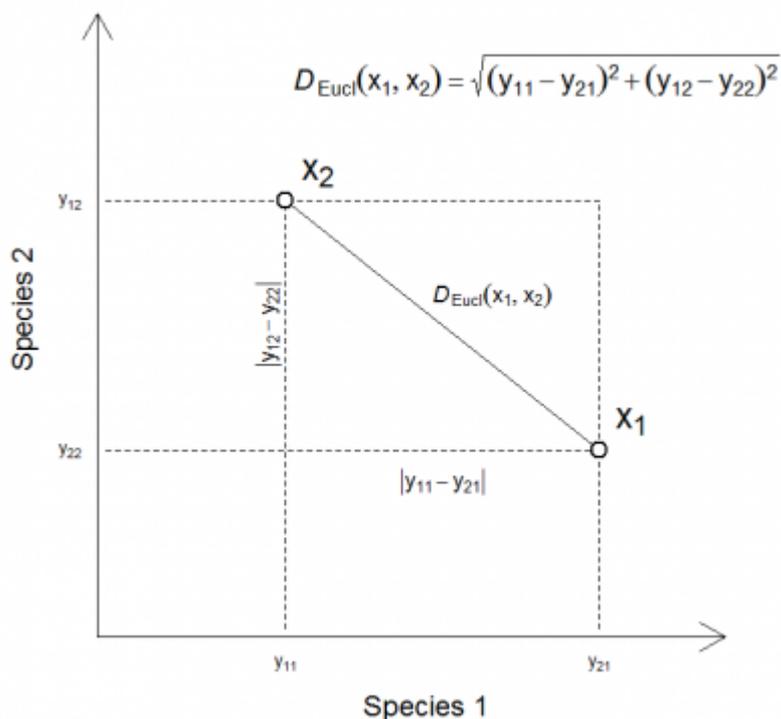
index described above. It is considered suitable for community composition data since it is asymmetrical (ignores double zeros), and it has a meaningful upper value equal to one (meaning complete mismatch between species composition of two samples, i.e. if one species in one sample is present and has some abundance, the same species in the other samples is zero, and vice versa). Bray-Curtis considers absolute species abundances in the samples, not only relative species abundances. The index is not metric, but the version calculated as $\sqrt{1-PS}$ (where PS is percentage similarity) is metric and can be used in PCoA.

Euclidean distance, although not suitable for ecological data, is frequently used in a multivariate analysis (mostly because it is the implicit distance for linear ordination methods like PCA, RDA and some clustering algorithms). Euclidean distance has no upper limit, and the maximum value depends on the data. The main reason why it is not suitable for compositional data is that it is a symmetrical index, i.e. it treats double zeros in the same way as double presences, and as a result, double zeros shrink the distance between two plots (the solution is to apply Euclidean distances on pre-transformed species composition data, e.g. using Hellinger, Chord or chi-square transformation; resulting distances are then asymmetrical). Another disadvantage of Euclidean distance is that it puts more emphasis on the absolute species abundances instead of species presences and absences in the samples; as a result, Euclidean distances between two samples not sharing any species may be smaller than between two samples sharing all species, but with the same species having large abundance differences between samples (Euclidean paradox). An example of calculating Euclidean distance between samples with only two species is on [Fig. 4](#).

Euclidean distance:
 where y_{1j} and y_{2j} are abundances of species j in sample 1 and 2, respectively

$$D_{Eucl} = \sqrt{\sum_{j=1}^p (y_{1j} - y_{2j})^2}$$

Figure 4: Euclidean distance between two samples with only two species.



Chord distance is Euclidean distance calculated on normalized species data. Normalization means that species vector in multidimensional space is of unit length; to normalize the species vector, one needs to divide each species abundance in a given sample by the square-rooted sum of squared abundances of all species in that sample. Chord distance is then the Euclidean distance between samples with normalized species data. An advantage of chord distance compared to Euclidean distance is that it is asymmetrical (ignores double zeros) and has the upper limit (equal to $\sqrt{2}$), while Euclidean distance has no upper limit.

Hellinger distance is Euclidean distance calculated on Hellinger transformed species data (and is the distance used in tb-PCA and tb-RDA if the species data are pre-transformed by Hellinger transformation). Hellinger transformation consists of first relativizing the species abundances in the sample by standardizing them to sample total (sum of all abundances in the sample); then, each standardized value is square-rooted. This puts the species abundances on the relative scale, and square-rooting lowers the dominant species' importance. Hellinger distance is asymmetrical (not influenced by double zeros) and has an upper limit of $\sqrt{2}$, which makes it a suitable method for ecological data with many zeros.

Chi-square distance is an asymmetrical distance which is rarely calculated itself, but is important since it is implicit for CA and CCA ordination.

Euclidean distance: abundance paradox

When comparing two samples, Euclidean distance puts more weight on differences in species abundances than on difference in species presences. As a result, two samples not sharing any species could appear more similar (with lower Euclidean distance) than two samples which share species but the species largely differ in their abundances (see the example below).

In the species composition matrix below, samples 1 and 2 does not share any species, while samples 1 and 3 share all species but differ in abundances (e.g. species 3 has abundance 1 in sample 1 and abundance 8 in sample 3):

	Species 1	Species 2	Species 3
Sample 1	0	1	1
Sample 2	1	0	0
Sample 3	0	4	8

Euclidean distance between sample 1 and 2 is lower than between sample 1 and 3, although samples 1 and 2 have no species in common, while sample 1 and 3 share all species. Distances, which are based on relative species abundances (i.e. those in which abundances of species in the sample are made relative e.g. by dividing each abundance of species in the sample by the sum of abundances for all species in that sample), do not have this problem (e.g. Hellinger distance, which is Euclidean distance applied on Hellinger-standardized data - the first step of Hellinger standardization converts absolute species abundances into relative ones).

Matrix of similarities/distances

The matrix of similarities or distances is squared (the same number of rows as columns), with the

values on diagonal either zeros (distances) or ones (similarities), and is symmetric - the upper right triangle is a mirror of values in lower left one (since it does not matter whether you calculate similarity/distance from sample A to sample B or from sample B to sample A; Fig. 5).

	1	2	3	4	5	6	7	8	9	10
1	0	87.85	26.99	28.15	60.87	60.01	59.2	59.46	17.39	34.63
2	87.85	0	74.39	101.5	75.22	77.78	76.07	74.92	79.92	71.65
3	26.99	74.39	0	42.47	45.83	44.32	46.36	45.47	21.09	22.2
4	28.15	101.5	42.47	0	80.5	80.14	79.05	79.12	29.28	55.91
5	60.87	75.22	45.83	80.5	0	34.71	38.11	35.27	54.3	39.63
6	60.01	77.78	44.32	80.14	34.71	0	28.93	26.35	56.33	33.38
7	59.2	76.07	46.36	79.05	38.11	28.93	0	23.34	55.06	36.84
8	59.46	74.92	45.47	79.12	35.27	26.35	23.34	0	55.05	37.38
9	17.39	79.92	21.09	29.28	54.3	56.33	55.06	55.05	0	30.27
10	34.63	71.65	22.2	55.91	39.63	33.38	36.84	37.38	30.27	0

Figure 5: Matrix of Euclidean distances calculated between all pairs of samples (a subset of 10 samples from Ellenberg's Danube meadow dataset used). Diagonal values (yellow) are zeros since the distance of two identical samples is zero.

Presence-absence vs quantitative species composition data

Species composition data (i.e. data about the occurrence of species in individual community samples) containing species quantities (abundances, covers, biomass, numbers of individuals) can always be transformed into species presences-absences. It is important to mention, however, that by transforming abundances into presences-absences, you are reducing the amount of information (information about species quantities are lost) and likely changing what data can tell about natural processes behind the community assembly.

Abundance data carry two types of information: 1) whether a species occurs in this community (or not), and 2) how much of this species occurs here; presence-absence data contain only the first type of this information. At the same time, whether a species occurs in a given community (or not) is driven by different ecological processes than the abundance of species in the community if the species already occurs there. The occurrence of a species in the site is often dependent not only on environmental suitability but also on dispersal limitation (whether the species can get there), random drift (species may go extinct due to stochastic processes) or the existence of biogeographical boundaries (two sites with similar environmental conditions may not share the same species simply because there is a river or mountain range between them). On the other hand, if the species already occurs in the sampled community, then its abundance is often driven by the suitability of environmental conditions (more abundant species are those for which the environment is favourable),

but also by biotic interaction (competition or mutualism with other species present in the community).

This also means that sometimes it may be meaningful to analyse parallelly both abundance and presence-absence-transformed species composition data, especially if we are about to quantify alternative processes. However, if the species composition data contain many zeros (meaning that they are very compositionally heterogeneous), then most of the information they represent is related to the first type (see above), even if they contain abundances of all non-absent species. Some studies (e.g. [Wilson 2012](#)) show that when analysing the relationship of species composition to environmental variables, transforming species composition data into presences-absences may actually improve results (variance explained by the environment in constrained ordination, or fit of environmental variables to unconstrained ordination axes).

1)

Percentage similarity (PS) is quantitative analog of Sørensen index; 1-PS is Percentage dissimilarity, also known as Bray-Curtis distance.

2)

Note that the use of “distance” and “dissimilarity” is somewhat not systematic; some authors call distances only those indices which are metric (Euclidean), i.e. can be displayed in metric (Euclidean) geometric space, and the other indices are called dissimilarities; but sometimes these two terms are simply synonyms.

3)

Note that according to P. Legendre, Bray-Curtis index should not be called after Bray and Curtis, since they have not really published it, only used it.

From:

<https://www.davidzeleny.net/anadat-r/> - **Analysis of community ecology data in R**

Permanent link:

<https://www.davidzeleny.net/anadat-r/doku.php/en:similarity>

Last update: **2021/03/03 19:49**