PBDB Intensive Summer Course 2007
Paleoecology Section – Thomas Olszewski

Measurement of Similarity

Foundations

Similarity index = a numerical index describing the similarity of two community samples in terms of their species content

Similarity matrix = a square, symmetrical matrix with the similarity value of every pair of samples, if Q-mode, or species, if R-mode, in the data matrix

The similarity matrix is the basis for all multivariate techniques depicting relationships among community samples or taxa, so the choices made at the initial stage of an analysis will strongly influence the results at the final stage

All similarity metrics can be converted to *difference* metrics by subtracting them from their maximum value (1.0 in all cases presented here); similarity is required for *cluster analysis*, whereas difference is required for *ordination analysis*

p = number of taxa

N = number of samples

x_{Ai} = abundance of species i in sample A

Transforming Data

Transforming data changes their properties to be more amenable to statistical analysis. In geometric terms, it shifts the relative positions of points in multivariate space in order to reveal an obscured pattern, impose a desired pattern, or hide an undesired one; in exploratory statistical analysis, revealing and obscuring pattern are opposite sides of the same coin. Nevertheless, data transformations are generally necessary in order to obtain interpretable results, so knowing how they influence results is critical for interpreting and evaluating multivariate analyses.

Converting numerical abundances to presence/absence values:

$$\text{If } (x_{Ai}) > 0, \text{ then } x^*_{Ai} = 1, \text{ else } x^*_{Ai} = 0 \qquad (1)$$

This transform makes all species equally important in characterizing a sample, regardless of their abundance.

Logarithmic transform

$$x^*_{Ai} = \log_b(x_{Ai} + k) \qquad (2)$$

b = base of logarithm (typically 2, 10, or e)

k = a constant necessary to prevent undefined log values when $x_{Ai} = 0$

This transformation converts log-normal abundance data into normally distributed (i.e., Gaussian) data. It reduces the influence of dominant species, but not as much as a conversion to presence/absence values. The need for a constant is not a desirable property because it has a disproportionately large influence on the contribution of rare species: adding 1 to 1 is 100% change whereas adding 1 to 10,000 is a 0.01% change.

A recent log-based transform that gets around the need to alter abundance values:

$$\text{if } x_{Ai} = 0,\ x^*_{Ai} = 0$$
$$\text{else } x^*_{Ai} = \log(x^*_{Ai}) + 1 \qquad (3)$$

This function has the advantage of rescaling that makes the classic log transform so useful, but it maintains the correct relationship between x_{Ai} values when they are subtracted from one another (as they are in most measures of similarity) regardless of their rarity.

Root transform

$$x^*_{Ai} = \sqrt[n]{x_{Ai}} \qquad (4)$$

Like all transformations, the root transform decreases the influence of dominant taxa. A "double square root" (i.e., $n = 4$) transform is quite common in ecology.

Arcsine-squareroot transform

$$x^*_{Ai} = \arcsin\sqrt{\frac{x_{Ai}}{\sum x_{Ai}}} \qquad (5)$$

This function increases the importance of low abundance taxa and decreases importance of high abundance taxa. Changes $x_{Ai}/\Sigma x_{Ai}$ values ranging from 0 to 1 to x^*_{Ai} values that range from 0 to 1.571.

Standardizing Data

Standardization weights samples or taxa so that they contribute a statistical analysis more equally – i.e., without starndardizing data, large samples or abundant taxa can overwhelm a subtle pattern.

Standardization to total

$$y_{Ai} = \frac{x_{Ai}}{\sum x_{Ai}} \qquad (6)$$

When applied to a sample, each taxon is represented by its proportion and every sample sums to 1.0; this is what is referred to as "relative abundance" data. Often multiplied by 100 to obtain taxon percentages.
If applied to each taxon, it emphasizes rare taxa and diminishes common taxa.

Standardization to maximum

$$y_{Ai} = \frac{x_{Ai}}{\max(x_{Ai})} \qquad (7)$$

When applied to a sample, the most common taxon is given a value of 1.0 and all the other taxa are scaled to it. The largest taxon in every collection is then equal.
If applied to each taxon, it "equalizes" influence of rare and common taxa (maximum abundance of every taxon equals 1.0).

Standardization to vector length

$$y_{Ai} = \frac{x_{Ai}}{\sqrt{\sum x_{Ai}^2}} \qquad (8)$$

If a site is considered a vector (i.e., the point representing it in a space defined by species axes is the head of a vector starting at the origin), this standardization gives the vector a length of 1.0 – all the sample points lie on a spheroid with radius equal to 1.

z-transform

$$z_{Ai} = \frac{x_{Ai} - \bar{x}_A}{\sigma_A} \qquad (9)$$

This transform is a common calculation in classical statistics – subtract the mean and divide by the standard deviation. The result is that the mean value of every sample is 0 (it is centered) and the standard deviation of its taxon abundances is 1 (it is unit length). This transformation is implicit when principal components analysis is applied to a *correlation matrix*. Note the similarity in the form of the z-transform to equation (8) – in effect, a z-transform scales a centered data series to be a vector of length one.

Two-way transform

It is a common procedure to standardize both taxa and samples. A common approach is to standardize taxa to their maximum and samples to their totals.

Correlation

The basic measure of correlation in classical statistics is Pearson's product-moment correlation coefficient, *r*:

$$r = \frac{\sum (x_{Ai} - \bar{x}_{Ai})(x_{Bi} - \bar{x}_{Bi})}{\sqrt{\sum (x_{Ai} - \bar{x}_{Ai})^2 (x_{Bi} - \bar{x}_{Bi})^2}} \qquad (10)$$

This is the covariance scaled by the products of the standard deviations of the two variables.

This coefficient plays a critical role in many techniques of parametric ordination, such as Principal Components Analysis, but as a measure of ecological similarity, it is very susceptible to sparse data (i.e., data matrices with lots of zero values indicating that many species do not occur in many samples). Note that *r* is the dot product of two z-transformed vectors of data.

Binary Coefficients of Association

Coefficients that do NOT incorporate species abundance information – i.e., "presence-absence" or "binary" coefficients

Contingency table of co-occurrence

		Sample B	
		1	0
Sample A	1	*a*	*b*
	0	*c*	*d*

a = number of taxa occurring in both samples; mutual presences
b = number of taxa occurring in A but not B
c = number of taxa occurring in B but not A
d = number of species in matrix absent from both samples; mutual absences

Simple Matching Coefficient

$$S_{SM} = \frac{a+d}{a+b+c+d} \qquad (11)$$

note that mutual absences contribute to similarity in this coefficient; size of denominator depends on total number of taxa in the data matrix

Jaccard Coefficient

$$S_J = \frac{a}{a+b+c} \qquad (12)$$

number of mutual presences divided by total number of taxa present in only the two samples being compared; independent of number of taxa in other samples, but susceptible to sample size (bigger samples => more species, but not linearly)

Sorensen Coefficient (Dice, Czekanowski)

$$S_S = \frac{2a}{2a+b+c} = \frac{2S_J}{1+S_J} \qquad (13)$$

number of mutual presences divided by the average number of taxa in the two samples being compared; less prone to extreme values than Jaccard, but otherwise monotonically related to Jaccard

Quantitative Coefficients of Association

Coefficients that incorporate abundance data; closely related to binary coefficients

Similarity Ratio

$$S_{SR} = \frac{\sum_{i=1}^{p} x_{Ai}x_{Bi}}{\left(\sum x_{Ai}^2 + \sum x_{Bi}^2 - \sum x_{Ai}x_{Bi}\right)} \qquad (14)$$

if used with presence-absence data, this reverts to the Jaccard coefficient

Percentage Similarity

$$S_{PS} = \frac{2\sum_{i=1}^{p}\min(x_{Ai}, x_{Bi})}{\sum x_{Ai} + \sum x_{Bi}} \qquad (15a)$$

if used with presence-absence data, this reverts to the Sorensen coefficient

the complement of percentage similarity is the percentage difference, also called the Bray-Curtis distance or the Lance & Williams metric; this metric is regarded as very good in retaining underlying ecological patterns:

$$1 - S_{PS} = \frac{\sum |x_{Ai} - x_{Bi}|}{\sum x_{Ai} + \sum x_{Bi}} \qquad (15b)$$

Metrics Based on Geometry

Species abundances in a sample can be thought of as x, y, z, etc. coordinates of a point in a multidimensional space; the sample is depicted as a point and the distances between points are related to their similarity/difference

Euclidean Distance

$$D_{AB} = \sqrt{\sum_{i=1}^{p}(x_{Ai} - x_{Bi})^2} \qquad (16a)$$

this metric is based on the Pythagorean Theorem
D_{AB} has several important properties:

1) $D_{AB} \geq 0$ (positive)
2) $D_{AB} = D_{BA}$ (symmetrical)
3) $D_{AC} \leq D_{AB} + D_{BC}$ (conforms to *triangular inequality*)
4) If A=B, $D_{AB} = 0$; if A≠ B, $D_{AB} > 0$

Value of Euclidean distance is dependent on number of taxa, so one way of scaling it is to divide by the total number of taxa:

$$D_{AB} = \sqrt{\frac{1}{p}\sum_{i=1}^{p}(x_{Ai} - x_{Bi})^2} \qquad (16b)$$

Euclidean distance is one of a general set of distance metrics called Minkowski Metrics:

$$D = \sqrt[Z]{\frac{1}{p}\sum_{i=1}^{p}(|x_{Ai} - x_{Bi}|)^Z} \qquad (17)$$

if $Z = 1$, then the Manhattan or City Block distance metric is obtained:

$$MCD = \frac{1}{p}\sum_{i=1}^{p}\left|x_{Ai} - x_{Bi}\right| \qquad (18)$$

Comments. Overall, Euclidean Distance (and more generally the Minkowski family of distance metrics) is not good for analyzing sparse data, which are typical of ecological data sets. However, this measure is fundamental to methods like Polar Ordination and Non-Metric Multidimensional Scaling and sees wide application in geometric morphometrics. In addition, several coefficients (e.g., chord distance) that have proven to be very useful in ecology are closely related to Euclidean distance.

Cos-theta or Ochiai Coefficient

Rather than evaluate taxonomic difference using multivariate distance, we can examine the *angle* between two sample points in a multidimensional space:

$$\cos\theta_{AB} = \frac{\sum_{i=1}^{p} x_{Ai}x_{Bi}}{\sqrt{\sum x_{Ai}^2 x_{Bi}^2}} \qquad (19)$$

when $\theta = 0$, $\cos\theta = 1$; when $\theta = 90$, $\cos\theta = 0$; when $\theta = 180$, $\cos\theta = -1$

Comments. If data are first z-transformed, then this metric reverts to Pearson's *r* – i.e., a geometric interpretation of *r* is as an *angle* between two vectors in ordination space. This metric has automatic vector length standardization. Like Euclidean distance, it is susceptible to sparse data.

Chord distance

$$CD = \sqrt{\sum_{i=1}^{p}\left(\frac{x_{Ai}}{\sqrt{\sum x_{Ai}^2}} - \frac{x_{Bi}}{\sqrt{\sum x_{Bi}^2}}\right)^2} \qquad (20)$$

Comments. This metric combines Euclidean distance and angles between points; it is equal to comparing samples standardized to unit vector length using Euclidean distance. This metric appears to perform well with ecological data.

References for Similarity Coefficients

Archer, A.W. and Maples, C.G., 1987, Monte Carlo simulation of selected binomial similarity coefficients (I): Effect of number of variables. Palaios, v. 2, p. 609-617.

Archer, A.W. and Maples, C.G., 1989, Response of selected binomial coefficients to varying degrees of matrix sparseness and to matrices with known data interrelationships. Mathematical Geology, v. 21, p. 741-753.

Cao, Y., Larsen, D.P., Hughes, R.M., Angermeier, P.L., and Patton, T.M., 2002, Sampling effort affects multivariate comparison of stream assemblages. Journal of the North American Benthological Society, v. 21, p. 701-714.

Cheetham, A.H. and Hazel, J.E., 1969, Binary (presence-absence) similarity coefficients. Journal of Paleontology, v. 43, p. 1130-1136.

Faith, D.P., 1983, Asymmetric binary similarity measures. Oecologia, v. 57, p. 287-290.

Faith, D.P., 1984, Patterns of sensitivity of association measures in numerical taxonomy. Mathematical Biosciences, v. 69, p. 199-207.

Faith, D.P., Minchin, P.R., and Belbin, L., 1987, Compositional dissimilarity as a robust measure of ecological distance. Vegetatio, v. 69, p. 57-68.

Gower, J.C. and Legendre, P., 1986, Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification, v. 3, p. 5-48.

Jackson, D.A., Somers, K.M., and Harvey, H.H., 1989, Similarity coefficients: measures of co-occurrence and association or simply measures of occurrence? The American Naturalist, v. 133, p. 436-453.

Maples, C.G. and Archer, A.W., 1988, Monte Carlo simulation of selected binomial similarity coefficients (II): Effect of sparse data. Palaios, v. 3, p. 95-103.

Ricklefs, R.E. and Lau, M., 1980, Bias and dispersion of overlap indices: results of some Monte Carlo simulations. Ecology, v. 61, p. 1019-1024.

Sepkoski, J.J., Jr., 1974, Quantified coefficients of association and measurement of similarity. Mathematical Geology, v. 6, p. 135-152.

Shi, G.R., 1993, Multivariate data analysis of palaeoecology and palaeobiogeography – a review. Palaeogeography, Palaeoclimatology, Palaeoecology, v. 105, p. 199-234.

Washington, H.G., 1984, Diversity, biotic and similarity indices: a review with special reference to aquatic systems. Water Resources, v. 18, p. 653-694.

Wolda, H., 1981, Similarity indices, sample size and diversity. Oecologia, v. 50, p. 296-302.