Chapter 7 *Ecological resemblance*

7.0 The basis for clustering and ordination

For almost a century, ecologists have collected quantitative observations to determine the resemblance between either the objects under study (sites) or the variables describing them (species or other descriptors). Objects and descriptors are defined in Section 1.4. Measuring the association (Section 2.2) between objects (Q mode) or descriptors (R mode) is the first, and sometimes the only step in the numerical analysis of ecological data. The various modes of analysis are discussed in Section 7.1. It may indeed happen that examining the association matrix suffices to elucidate the structure and thus answer the question at the origin of the investigation.

The present chapter provides a review of the main measures of association available to ecologists. Section 7.2 introduces the three types of association coefficients; the measures pertaining to each type — similarity, distance, and dependence — are described in Sections 7.3 to 7.5, respectively. In order to help ecologists choose from among this plurality of coefficients, Section 7.6 summarizes criteria for choosing a coefficient; the criteria are presented in the form of identification keys. Ecologists who do not wish to study the theory that underlies the measures of association may directly go to Section 7.6 after making themselves familiar with the terminology (Sections 7.1 and 7.2). When necessary, they may then refer to the paragraphs of Sections 7.3 to 7.5 describing the measures of interest.

Clustering

Ordination

In the next chapters, measures of resemblance between objects or descriptors will be used to cluster the objects or descriptors (Chapter 8) or to produce ordination diagrams in spaces of reduced dimensionality (Chapter 9). The clustering of objects (or descriptors) is an operation by which the set of objects (or descriptors) is partitioned in two or more subsets (clusters), using pre-established rules of agglomeration or division. Ordination in reduced space is an operation by which the objects (or descriptors) are represented in a space that contains fewer dimensions that in the original data set; the positions of the objects or descriptors with respect to one

another may also be used to cluster them. Both operations are often carried out on association matrices, which are computed as described in the following sections.

7.1 Q and R analyses

As noted by Cattell (1952), the ecological data matrix may be studied from two main viewpoints. One may wish to look at relationships among either the objects or the descriptors. The important point here is that these modes of analysis are based on different measures of association. The different types of coefficients are described in Section 7.2. Measuring the dependence between descriptors is done using coefficients like Pearson's r correlation coefficient (eq. 4.7, Section 4.2), so that studying the data matrix based on such coefficients is called *R analysis*. By opposition, studying the data matrix to uncover relationships among objects is called *Q analysis* (Fig. 2.1).

R mode
Q mode

Cattell (1966) had also observed that the *data box* (objects × descriptors × time instances; Fig. 7.1) may be looked at from other viewpoints than simply Q and R. He defined six modes of analysis:

O: among time instances, based on all observed descriptors (a single object);
P: among descriptors, based on all observed times (a single object);
Q: among objects, based on all observed descriptors (a single instance);
R: among descriptors, based on all observed objects (a single instance);
S: among objects, based on all observed times (a single descriptor);
T: among time instances, based on all observed objects (a single descriptor).

In the present chapter, the discussion of association coefficients will deal with the two basic modes only, i.e. Q measures (computed among objects) and R measures (computed among descriptors).

O-mode studies are conducted using Q measures; see, for example, Section 12.6. Similarly, P-mode studies are generally carried out with the usual R-type coefficients. When the data set forms a time series, however, P studies are based on special R-type coefficients that are discussed in Chapter 12: cross-covariance, cross-correlation, co-spectrum, coherence.

S- and T-mode studies mostly belong to autecology, i.e. studies involving a single species. S-mode comparisons among objects use the same coefficients as in P-mode analysis. Studying the relationship between "descriptor y observed at site x_1" and "the same descriptor y observed at site x_2" is analogous to the comparison of two descriptors along a time axis.

In T-mode studies, a variable is observed across several objects (sites, etc.) and different instances through time. Statistical tests of hypothesis for related samples are often applicable to these problems; see Table 5.2. In other cases, the two time instances

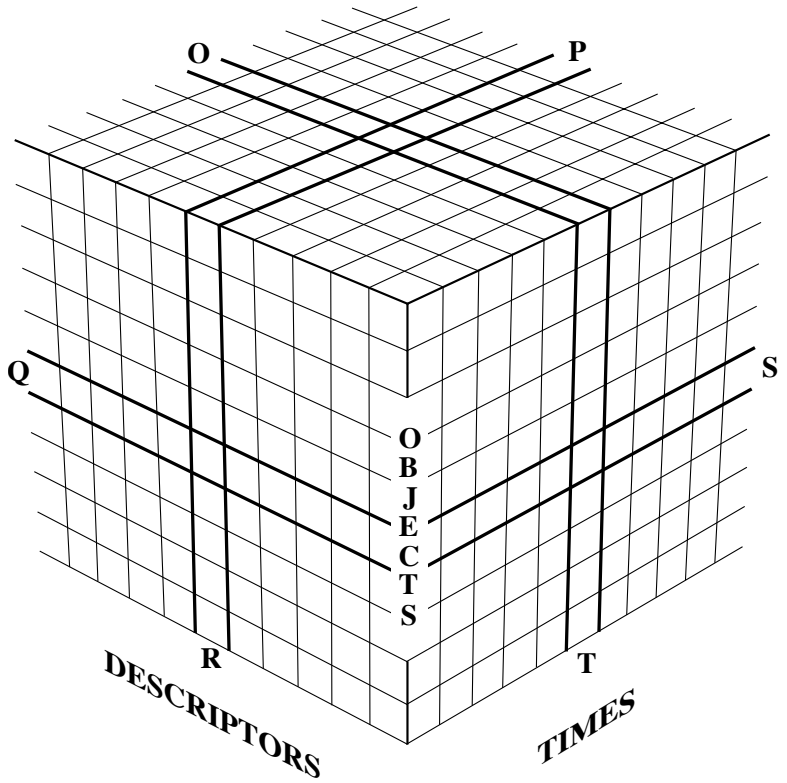

Figure 7.1 The three-dimensional data box (objects × descriptors × times). Adapted from Cattell (1966).

to be compared are considered to define two descriptors, as in the S mode, so that normal R-type measures may be used. Environmental impact studies form an important category of T-mode problems; ecologists should look at the literature on BACI designs when planning such studies (Before/After – Control/Impact: Green, 1979; Bernstein & Zalinski, 1983; Stewart-Oaten *et al*., 1986; Underwood, 1991, 1992, 1994).

Q or R?

It is not always obvious whether an analysis belongs to the Q or R mode. As a further complication, in the literature, authors define the mode based either on the association matrix or on the purpose of the analysis. Principal component analysis (Section 9.1), for instance, is based on a dispersion matrix among descriptors (R mode?), but it may be used for ordination of either the objects (Q mode?) or the descriptors (R mode?). In order to prevent confusion, in the present book, any study starting with the computation of an *association matrix among objects* is called a *Q analysis* whereas studies starting with the computation of an *association matrix among descriptors* are referred to as *R analyses*. In Chapter 9 for example, it is

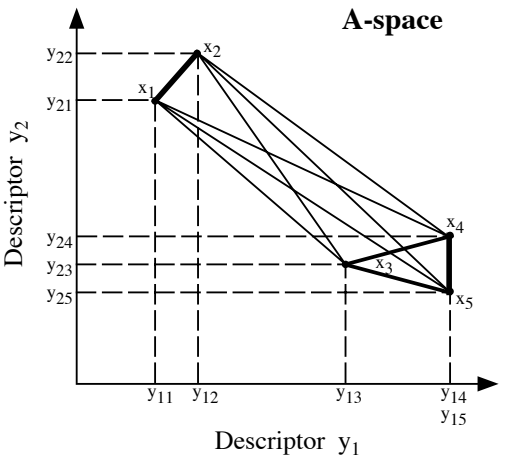

Figure 7.2 Scatter plot representation of five objects in an A-space with two descriptors. In this graph, the thickness of the lines that join the objects is proportional to their degree of resemblance with regard to the two descriptors, i.e. their proximity in the space.

possible to obtain an ordination of objects in low-dimension space using either the R method of principal component analysis (Section 9.1) or the Q method of principal coordinate analysis (Section 9.3). Interestingly, these two analyses lead to the same ordination of the objects when the principal coordinate analysis is computed from a Euclidean distance matrix (coefficient D_1, Section 7.4), although the results of Q and R analyses are not always reducible to each other.

The number of dimensions that can be represented on paper is limited to two or eventually three. Hence, one generally imagines distances among objects (Fig. 7.2) as embedded in a 2- or 3-dimensional space. Section 7.4 will show that such models can be extended to any number of dimensions. The descriptor, or *attribute space*, is called A-space. Distances and similarities computed in the present chapter will be based, in most instances, on measurements made in high-dimensional space.

A-space

Metric, Euclidean space

The A-space is called *metric* because the reference axes are quantitative, metric descriptors (Table 1.2). It is also called *Euclidean* because Euclide's geometry holds in that space. The qualifier *metric* will be used in different contexts in this book: metric variable (Table 1.2), metric space (here), metric properties of distances (beginning of Section 7.4), metric distances (Subsection 7.4.1); see also Plate 3.1, p. 142. Likewise, *Euclidean* will refer either to Euclidean space (here), to *the* Euclidean distance D_1 (Subsection 7.4.1), to the property of distances that can be embedded in Euclidean space (Tables 7.2 and 7.3), or to a property of ordination methods and plots (beginning of Section 9.3).

In addition to the methods described in the present and following chapters, there exist approaches allowing the analysis of the whole data box instead of subsets, as was the case in the six modes described above. Examples are found in Williams & Stephenson (1973), Williams *et al*. (1982), Cailliez & Pagès (1976), Marcotorchino & Michaud (1979), Kroonenberg (1983: three-way principal component analysis*), Carlier & Kroonenberg (1996: three-way correspondence analysis) and Kroonenberg (2008).

7.2 Association coefficients

The most usual approach to assess the resemblance among objects or descriptors is to first condense all (or the relevant part of) the information available in the ecological data matrix (Section 2.1) into a square matrix of association among the objects or descriptors (Section 2.2). In most instances, the association matrix is *symmetric*. Non-symmetric matrices can be decomposed into symmetric and skew-symmetric components, as described in Section 2.3; the components may then be analysed separately. Non-symmetric matrices can also be subjected to a special type of clustering called *seriation* (Section 8.10). In Chapters 8 and 9, objects or descriptors will be clustered or represented in reduced space after analysing an association matrix. It follows that *the structure resulting from a numerical analysis is that of the association matrix; the results of the analysis do not necessarily reflect all the information originally contained in the ecological data matrix*. This stresses the importance of choosing an appropriate measure of association. This choice determines the issue of the analysis. Hence, it must take into account the following considerations:

- The nature of the study (i.e. the initial question and the hypothesis) determines the kind of ecological structure to be evidenced through an association matrix, and consequently the type of measure of resemblance to be used.

- The various measures available have different mathematical constraints. The methods of analysis to which the association matrix will be subjected (clustering, ordination) may require measures of resemblance with specific mathematical properties.

- One must also consider the computational aspect, and thus preferably choose a measure available in a computer package or R function (Section 7.8), or one that can easily be programmed.

* Program 3WayPack (Kroonenberg & De Roo, 2010) for three-way principal component analysis and other three-way analyses is available from Pieter M. Kroonenberg, Leiden Institute of Education and Child Studies, Leiden University, Wassenaarseweg 52, NL-2333 AK Leiden, The Netherlands. Other three-mode software is described on the Web page: http://three-mode.leidenuniv.nl/. An overview of these methods can be found in Kroonenberg (2008).

Ecologists are, in principle, free to define and use any measure of association suitable for the ecological question under study; mathematics impose few constraints to this choice. This is why so many association coefficients are found in the literature. Some of them are of wide applicability whereas others have been developed to meet specific needs. Several coefficients have been rediscovered by successive authors and may be known under various names. Reviews of some coefficients can be found in Cole (1949, 1957), Goodman & Kruskal (1954, 1959, 1963), Dagnelie (1960), Sokal & Sneath (1963), Williams & Dale (1965), Cheetham & Hazel (1969), Sneath & Sokal (1973), Clifford & Stephenson (1975), Orlóci (1978), Daget (1976), Blanc *et al.* (1976), Prentice (1980), Gower (1985), and Gower & Legendre (1986).

1 — Similarity, distance, and dependence coefficients

In the following sections, *association* will be used as a general term to describe any measure or coefficient used to quantify the resemblance or difference between objects or descriptors, as proposed by Orlóci (1975). With *dependence* coefficients, used in the R mode, zero corresponds to no association. In Q-mode studies, *similarity* coefficients between objects will be distinguished from *distance* (or *dissimilarity*) coefficients. Similarities are *maximum* ($S = 1$) when the two objects are identical and *minimum* when the two objects are completely different; distances follow the opposite rule. Figure 7.2 shows the difference between the two types of measures: the length of the line between two objects is a measure of their distance, whereas its thickness, which decreases as the two objects get further apart, is proportional to their similarity. If needed, a similarity can be transformed into a distance, for example by computing its one-complement. For a similarity (S) measure, which takes values between 0 and 1, the corresponding distance (D) may be computed as:

Dependence
Similarity
Distance

$$D = 1 - S, \quad D = \sqrt{1 - S}, \quad \text{or} \quad D = \sqrt{1 - S^2}$$

We will see in Tables 7.2 and 7.3 and in Subsection 9.3.4 that the choice of a transformation instead of another may have consequences for the result of ordination analysis. Distances, which in several cases are not bound by a pre-determined upper value, can be normalized using eqs. 1.10 or 1.11:

$$D_{norm} = \frac{D}{D_{max}} \quad \text{or} \quad D_{norm} = \frac{D - D_{min}}{D_{max} - D_{min}}$$

where D_{norm} is the distance normalized in the interval [0, 1]; D_{max} and D_{min} are the maximum and minimum values taken by the distance coefficient, respectively. Normalized distances can be used to compute similarities by reversing the transformations given above:

$$S = 1 - D_{norm}, \quad S = 1 - D_{norm}^2, \quad \text{or} \quad S = \sqrt{1 - D_{norm}^2}$$

The following three sections describe the coefficients that are most useful with ecological data. Criteria to be used as guidelines for choosing a coefficient are

discussed in Section 7.6. Section 7.7 describes transformations for community composition data; each of these transformations is the first step in the calculation of an asymmetrical distance function described in Section 7.4. Computer programs and R functions are reviewed in Section 7.8.

2 — The double-zero problem

Niche theory (Hutchinson, 1957) states that species have ecological "preferences", meaning that they have evolved genetic adaptations to specific environmental conditions, including other species. Species are mostly found at locations where they encounter appropriate living conditions. The theory also predicts that species have *unimodal distributions* along environmental variables (Whittaker, 1967), like the Gaussian curves in Fig. 4.5: a species is found in greater abundance in some intervals along the gradients of major environmental variables or along composite axes*. The position of the mode of a species distribution along an environmental variable can be interpreted as the optimum value for the species along that variable. Along an environmental gradient, a species becomes rare and ultimately absent as one departs from its optimal conditions.

Unimodal distribution

As a consequence, community composition data sampled across a range of environmental conditions typically contain many zero values. This phenomenon is discussed in most texts of community and numerical ecology, in particular in Whittaker (1967), ter Braak (1987c) and ter Braak & Prentice (1988).

Comparison of sites is often based upon species abundance data. Species are important indicators of the apportioning of environmental resources among them. The division of resources should be reflected in the relative productivities of the species (Whittaker, 1972). The productivity of different species is not easily measured, however, and ecologists most often rely on other values of species importance such as number of individuals, biomass, coverage (for plants or corals) or basal area (for plants).

If a species is present at two sites, this is an indication of similarity between these sites since they both present conditions that are favourable or at least tolerable for the species. Likewise, the presence of a species at site 1 and not at site 2 is taken as an indication of difference in ecological conditions, notwithstanding sampling error†. However, if a species is absent from two sites, it may be because these sites have environmental conditions that are outside the niche of the species, and these conditions

* Composite environmental axes can be computed by ordination (Chapter 9), for instance as the principal components (PCA, Section 9.1) of a matrix of environmental variables.

† For a variety of reasons, species may not be observed at sites where they are present. Species may be inconspicuous, camouflaged, or hidden. With fungi for example, the carpophores of a species may not appear above ground at the time of a survey although the mycelium is present in the soil. When sampling is done "blindly", e.g. in a lake or the ocean with a plankton or fish net, many species may escape capture either randomly of by active avoidance of the sampling gear.

may be similar or very different at the two sites. Hence most ecologists do not consider that the absence of a species from two sites provides univocal or useful information. It is also understood, of course, that besides unimodal distributions and niche optimality, several reasons may explain the local absence of a species: the niche of the species may be present in one (or both) of two sites but be occupied by substitute species; absence may also be the result of the species dispersion, random local extinction, historical events, or other processes that cause stochastic variation.

The proportion of zeros in community composition data generally increases with the variability in environmental conditions among the sampling sites. If sampling has been conducted along one or several environmental axes, the species present are likely to differ at least partly from site to site. Including double zeros in the comparison between sites would result in high values of similarity for the many pairs of sites holding only a few species, these pairs presenting many double zeros; this would not provide a correct ecological assessment of the situation.

Double-zero problem

Because double zeros are not informative, their interpretation generates *the double-zero problem*: is the value of an association coefficient affected by inclusion of double zeros in its calculation? When choosing an association coefficient, ecologists must pay attention to the interpretation of double zeros: except in very limited cases (e.g. controlled experiments involving very few species and with small uncontrolled ecological variation), it is preferable to draw no ecological conclusion from the simultaneous absence of a species at two sites. In numerical terms, this means to skip double zeros when computing similarity or distance coefficients using species presence-absence or abundance data. Coefficients of this type are called *asymmetrical* because they treat double absences in a different way than double presences.

Asymmetrical coefficient

Symmetrical coefficient

In similarity coefficients (S), the handling of double zeros is clear in coefficient formulas (Section 7.3). Similarity coefficients all have a minimum value of 0 and a maximum value of 1. In *symmetrical* similarity coefficients, state zero for two objects is treated in exactly the same way as any other pair of values. This would be the correct way to handle double zeros in the case, for example, where two lakes are found to have 0 mgL^{-1} of dissolved oxygen in the hypolimnion in winter because this observation provides valuable information concerning their physical and chemical similarity and their capacity to support species. Coefficient S_{15}, for instance, would consider a double zero as an indication of resemblance between the lakes and include this information in the overall assessment of their similarity.

In distance coefficients (D, Section 7.4), however, one has to examine if the value computed for pairs of sites depends primarily on which species are present at each site (*asymmetrical coefficients*), or strictly on the numerical differences between species abundances (*symmetrical coefficients*). Symmetrical coefficients like the Euclidean distance (D_1) will be shown to lead to incorrect conclusions from an ecologist's viewpoint (see Fig. 7.8). The asymmetrical distance coefficients all have a fixed upper bound, which is either 1 or $\sqrt{2}$ in most cases.

Because ordination methods implicitly (PCA, CA, Sections 9.1 and 9.2) or explicitly (PCoA, Section 9.3) use a distance function as their metric to position objects with respect to one another in ordination space, it is important to make sure that the chosen distance coefficient is meaningful for the objects under study, especially when dealing with community composition data. By choosing an appropriate distance measure, an ecologist tries to appropriately model the relationships among the sites for the data at hand. The choice of a similarity or distance measure (Section 7.6) is an ecological, not a statistical decision.

7.3 Q mode: similarity coefficients

Similarities form a large group of coefficients in the literature. The similarity coefficients in the present section measure the association between *objects*. Similarities take values in the interval [0, 1], 1 being the similarity of two identical objects and of an object with itself. In contrast to most distance coefficients, similarity coefficients are never metric (definition at the beginning of Section 7.4) since it is always possible to find two objects, A and B, that are more similar than the sum of their similarities with another, more distant, object C. It follows that similarities cannot be used directly to position objects in a metric space (ordination; Chapter 9); they must be converted into distances using one of the transformations of Subsection 7.2.1. Which transformation to use is discussed at the beginning of Section 7.4 and in Tables 7.2 and 7.3. For clustering (Chapter 8), however, algorithms can be easily adapted to conduct the analysis on either a distance or a similarity matrix.

Similarity coefficients were first developed for binary descriptors, representing presence-absence data or answers to yes-no questions. They were later generalized to multi-state descriptors when computers made that possible. Another major dichotomy among similarity coefficients concerns how they deal with double-zeros or negative matches. This dichotomy was discussed in Subsection 7.2.2.

The remainder of this section distinguishes between binary and quantitative similarity coefficients and, for each type, those that use double-zeros or exclude them from the assessment of resemblance. Tables 7.4 and 7.5 summarize the use of the various similarity and distance coefficients in ecology.

1 — Symmetrical binary coefficients

In the simplest cases, the similarity between two sites is based on presence-absence data. Binary descriptors may describe the presence or absence of environmental

conditions (here) or species (next subsection). Observations may be summarized in a 2×2 frequency table:

Object $\mathbf{x}_1$ \ Object $\mathbf{x}_2$	1	0	
1	a	b	$a+b$
0	c	d	$c+d$
	$a+c$	$b+d$	$p=a+b+c+d$

where a is the number of descriptors for which the two objects are coded 1, d is the number of descriptors coding the two objects 0, whereas b and c are the numbers of descriptors for which the two objects are coded differently; and p is the total number of descriptors. An obvious way of computing the similarity between two objects is to count the number of descriptors that code the objects in the same way and divide this by the total number of descriptors:

$$S_1(\mathbf{x}_1, \mathbf{x}_2) = \frac{a+d}{p} \qquad \textbf{(7.1)}$$

Coefficient S_1* is called the *simple matching coefficient* (Sokal & Michener, 1958). When using this coefficient, one assumes that there is no difference between double-0 and double-1. This is the case, for instance, when any one of the two states of each descriptor could be coded 0 or 1 indifferently. A variant of this measure is the *coefficient of Rogers & Tanimoto* (1960) in which differences are given more weight than resemblances:

$$S_2(\mathbf{x}_1, \mathbf{x}_2) = \frac{a+d}{a+2b+2c+d} \qquad \textbf{(7.2)}$$

Sokal & Sneath (1963) proposed four other measures that include double-zeros. They have their counterparts in coefficients that exclude double-zeros, in the next subsection:

$$S_3(\mathbf{x}_1, \mathbf{x}_2) = \frac{2a+2d}{2a+b+c+2d} \qquad \textbf{(7.3)}$$

counts resemblances as being twice as important as differences;

* The numbers of the coefficients found in the first edition of this book (*Écologie numérique*, Masson, Paris, 1979) were not changed in subsequent editions because these numbers had rapidly been adopted by ecologists and used as coefficient identifiers in computer programs. Coefficients added since the 1983 edition have received sequential numbers.

$$S_4(\mathbf{x}_1, \mathbf{x}_2) = \frac{a+d}{b+c} \quad \textbf{(7.4)}$$

compares the resemblances to the differences, in a measure that goes from 0 to infinity;

$$S_5(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{4}\left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right] \quad \textbf{(7.5)}$$

compares the resemblances to the marginal totals;

$$S_6(\mathbf{x}_1, \mathbf{x}_2) = \frac{a}{\sqrt{(a+b)\,(a+c)}}\,\frac{d}{\sqrt{(b+d)\,(c+d)}} \quad \textbf{(7.6)}$$

is the product of the geometric means of the terms relative to a and d, respectively, in coefficient S_5.

Among the above coefficients, S_1 to S_3 are of more general interest for ecologists, but the others may occasionally prove useful to adequately handle special descriptors. Three additional measures are available in the NTSYSPC package (Section 7.8, footnote): the *Hamann coefficient*:

$$S = \frac{a+d-b-c}{p} \quad \textbf{(7.7)}$$

the *Yule coefficient*:

$$S = \frac{ad-bc}{ad+bc} \quad \textbf{(7.8)}$$

and *Pearson's* ϕ (phi):

$$\varphi = \frac{ad-bc}{\sqrt{(a+b)\,(c+d)\,(a+c)\,(b+d)}} \quad \textbf{(7.9)}$$

where the numerator is the value of the determinant of the 2×2 frequency table. φ is actually the square root of the X^2 (chi-square) statistic computed from 2×2 tables divided by n (eq. 7.61). In ecology, coefficients of this type are mostly used in R-mode analyses. These last indices are described in detail in Sokal & Sneath (1963).

2 — Asymmetrical binary coefficients

Coefficients that parallel those above can be used to compare sites based on species presence-absence data, where the comparison must exclude double-zeros. The best-known measure is Jaccard's (1900, 1901, 1908) *coefficient of community*. It is often referred to simply as *Jaccard's coefficient*:

$$S_7(\mathbf{x}_1, \mathbf{x}_2) = \frac{a}{a+b+c} \quad \textbf{(7.10)}$$

in which all terms have equal weights. The *Sørensen coefficient*[*] (1948) gives double weight to double presences:

$$S_8(\mathbf{x}_1, \mathbf{x}_2) = \frac{2a}{2a + b + c} \quad \textbf{(7.11)}$$

because (see above) one may consider that the presence of a species is more informative than its absence at one site. Absence of a species at one site may be due to various factors, as discussed in Subsection 7.2.2; it does not necessarily reflect differences in the environment. Double-presence, on the contrary, is a strong indication of resemblance. S_8 is the binary form of the Steinhaus similarity S_{17}, meaning that its value is equal to S_{17} applied to binary data. Before Sørensen, Dice (1945) had used S_8 under the name *coincidence index* in an R-mode study of species associations; this question is further discussed in Section 7.5.

The distance version of this coefficient, $D_{13} = 1 - S_8$, is a semimetric, as shown in the example that follows eq. 7.57. A consequence is that principal coordinate analysis (Section 9.3) of a S_8 or D_{13} resemblance matrix is likely to produce negative eigenvalues. Solutions to this problem are discussed in Subsection 9.3.4. The easiest solution is to base the principal coordinate analysis on square-root-transformed distances $D = \sqrt{1 - S_8}$ instead of $D = 1 - S_8$ (Table 7.2).

Another variant of S_7 gives triple weight to double presences:

$$S_9(\mathbf{x}_1, \mathbf{x}_2) = \frac{3a}{3a + b + c} \quad \textbf{(7.12)}$$

The asymmetrical counterpart to the coefficient of Rogers & Tanimoto (S_2 in the previous subsection) was suggested by Sokal & Sneath (1963). This coefficient gives double weight to differences in the denominator:

$$S_{10}(\mathbf{x}_1, \mathbf{x}_2) = \frac{a}{a + 2b + 2c} \quad \textbf{(7.13)}$$

Coefficients S_9, S_8, S_7 and S_{10} form a series S_w with weights $w = \{1/3, 1/2, 1, 2\}$ respectively; w is the weight of $(b + c)$ in the formulas, considering that a receives a weight of 1. Gower & Legendre (1986) have shown that the coefficients in this series are monotonically related, meaning that they produce the same results when used in order-invariant methods like single, complete and proportional-link linkage clustering (Section 8.5) or nonmetric multidimensional scaling (Section 9.4), which rely on the

[*] Some authors refer to this coefficient as having been first described by Czekanowski (1913). The Czekanowski (1913) paper, written in Polish, is about body part measurements and anthropology; it deals with quantitative measurement variables only. The index developed in that paper with the a, b, c symbols has nothing to do with the binary indices (Jaccard, Sørensen) described in the present section.

ordinal and not the absolute values of the similarities. Among the symmetrical binary coefficients, S_3, S_1 and S_2 form a similar series with weights $w = \{0.5, 1, 2\}$.

Russell & Rao (1940) suggested a measure that compares the number of double presences, in the numerator, to the total number of species found at all sites, including species that are absent (d) from the pair of sites considered:

$$S_{11}(\mathbf{x}_1, \mathbf{x}_2) = \frac{a}{p} \quad \textbf{(7.14)}$$

Kulczynski (1928) proposed a coefficient opposing double-presences to the sum of disagreements (presence in one site and absence in the other):

$$S_{12}(\mathbf{x}_1, \mathbf{x}_2) = \frac{a}{b+c} \quad \textbf{(7.15)}$$

Among their coefficients for presence-absence data, Sokal & Sneath (1963) mention the binary version of Kulczynski's coefficient S_{18} for quantitative data:

$$S_{13}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{2}\left[\frac{a}{a+b} + \frac{a}{a+c}\right] \quad \textbf{(7.16)}$$

where double-presences are compared to the marginal totals $(a + b)$ and $(a + c)$.

Ochiai (1957) used, as measure of similarity, the geometric mean of the ratios of a to the number of species in each site, i.e. the marginal totals $(a + b)$ and $(a + c)$:

$$S_{14}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\frac{a}{(a+b)}\,\frac{a}{(a+c)}} = \frac{a}{\sqrt{(a+b)\,(a+c)}} \quad \textbf{(7.17)}$$

Note the relationship to the X^2 (chi-square) statistic, where the expected value of the cell containing the value a is $E(a) = (a + b)(a + c)/n$. Coefficient S_{14} is the same as S_6 with the portion that concerns double-zeros (d) excluded. S_{14} is the binary form of both the chord (D_3) and Hellinger (D_{17}) distances (Section 7.4): when applied to binary data, these two distance coefficients produce values equal to $\sqrt{2}\sqrt{1 - S_{14}}$.

Faith (1983) suggested the following coefficient for community composition data, in which disagreements b and c (presence in one site and absence in the other) are given a weight opposite to that of double presences a. The value of S_{26} is not invariant but decreases with an increasing number of double-zeros:

$$S_{26}(\mathbf{x}_1, \mathbf{x}_2) = \frac{2a+d}{2p} = \frac{a-b-c}{2p} + \frac{1}{2} \quad \textbf{(7.18)}$$

3 — Symmetrical quantitative coefficients

Ecological descriptors often have more than two states. Researchers have sometimes extended the binary coefficients described in Subsection 7.3.1 to accommodate nonordered multi-state descriptors. For example, the simple matching coefficient can be used as follows with multi-state qualitative descriptors:

$$S_1(\mathbf{x}_1, \mathbf{x}_2) = \frac{\text{agreements}}{p} \qquad \textbf{(7.19)}$$

where the numerator contains the number of descriptors for which the two objects are in the same state. For example, if a pair of objects was described by the following 10 multi-state qualitative descriptors:

	Descriptors									
Object $\mathbf{x}_1$	9	3	7	3	4	9	5	4	1	6
Object $\mathbf{x}_2$	2	3	2	1	2	9	3	2	1	6
Agreements	0	+ 1	+ 0	+ 0	+ 0	+ 1	+ 0	+ 0	+ 1	+ 1 = 4

the value of S_1 computed for these data would be:

$$S_1(\mathbf{x}_1, \mathbf{x}_2) = 4 \text{ agreements}/10 \text{ descriptors} = 0.4$$

It is possible to extend in the same way the use of all binary coefficients to multi-state descriptors. However, coefficients of this type often result in a loss of valuable information, especially in the case of ordered descriptors for which two objects can be compared on the basis of the *amount of difference* between states.

Gower (1971a) proposed a general coefficient of similarity that can combine different types of descriptors and process each one according to its own mathematical type. Although the description of this coefficient may seem a bit complex, it can be easily translated into a short computer program. The coefficient initially takes the following form (see also the final form, eq. 7.21):

$$S_{15}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{p}\sum_{j=1}^{p} s_{12j}$$

Partial similarity

The similarity between two objects is the average, over the p descriptors, of the similarities calculated for all descriptors. For each descriptor j, the *partial similarity value* s_{12j} between objects $\mathbf{x}_1$ and $\mathbf{x}_2$ is computed as follows:

- For *binary* descriptors, $s_j = 1$ (agreement) or 0 (disagreement). Gower proposed two forms for this coefficient. The form used here is symmetrical, giving $s_j = 1$ to double-zeros. The other form, used in Gower's asymmetrical coefficient S_{19} (Subsection 7.3.4), gives $s_j = 0$ to double-zeros.

• *Qualitative* descriptors are treated following the simple matching rule stated above: $s_j = 1$ when there is agreement and $s_j = 0$ when there is disagreement. Double-zeros are treated as in the symmetrical form of the previous paragraph.

• *Semiquantitative descriptors* can be handled in various ways in the computation of S_{15}. In function ***gowdis()*** of the R package FD, three options are available for handling semiquantitative variables, which are called *ordered factors* in R.

1. With the "classic" option, eq. 7.20 (below) is used as if the values were quantitative.

2. With the "metric" option, the values are ranked across the set of objects under study. Tied values are replaced by their average ranks. Consider an example where the objects under study are a subset of a larger data base in which a semiquantitative variable has ten ordered states, but the subset only contains three of the ten states. With this option, the states are recoded as ranks 1, 2 and 3 (or by the average values of the tied ranks) before they are used in eq. 7.20. Function ***daisy()*** of package CLUSTER also recodes semiquantitative variables in that way before computing the Gower distance.

3. With the "podani" option, the states and tied ranks are recoded as in the "metric" option. Equation 7.20 (below) is modified to take tied ranks into account, as proposed by Podani (1999). The modified eq. 7.20 is shown in the documentation file of the ***gowdis()*** function.

With these options, one should make sure that distances between adjacent states are comparable in magnitude. For example, for a semiquantitative descriptor coded from 1 to 3, $|y_{1j} - y_{2j}|$ of eq. 7.20 makes sense only if the difference between states 1 and 2 can be thought of as similar to the difference between states 2 and 3. If there is too much difference, values $|y_{1j} - y_{2j}|$ are not comparable and the semiquantitative descriptor should be converted into an unordered factor.

• *Quantitative* descriptors (real numbers) are treated in an interesting way. For each descriptor, one first computes the difference between the states of the two objects $|y_{1j} - y_{2j}|$, as in the case of distance coefficients belonging to the Minkowski metric group (Section 7.4). This value is then divided by the largest difference (R_j) found for this descriptor across all sites in the study — or if one prefers, in a reference population*. Since this ratio is actually a normalized distance, it is subtracted from 1 to transform it into a similarity:

$$s_{12j} = 1 - [|y_{1j} - y_{2j}| / R_j] \qquad \textbf{(7.20)}$$

Missing values Kronecker delta

The Gower coefficient may be programmed to include an additional element of flexibility: no comparison is computed for descriptors where information is *missing* for one or the other object. This is obtained by a value w_j, called a *Kronecker's delta*, describing the presence or absence of information: $w_j = 0$ when the information about

* In most applications, the largest difference R_j is calculated for the data table under study. In epidemiological studies, for example, one may proceed to the analysis of a subset of a much larger database. To ensure consistency of the results in all the partial studies, it is recommended to calculate the largest differences (the "range" statistic of databases) observed throughout the whole database for each descriptor j and use these as values R_j when computing S_{15} or S_{19}.

y_j is missing for one or the other object, or both; $w_j = 1$ when information is present for the two objects. The final form of the *Gower coefficient* is the following:

$$S_{15}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{p} w_{12j} s_{12j}}{\sum_{j=1}^{p} w_{12j}} \qquad \textbf{(7.21)}$$

Coefficient S_{15} produces similarity values between 0 and 1 (maximum similarity).

One last touch of complexity, which was not suggested in Gower's paper but is added here, provides weighting to the various descriptors. Instead of *0 or 1*, one can assign to w_j a value *between 0 and 1* corresponding to the weight one wishes each descriptor to have in the analysis. Descriptors with weights close to 0 contribute little to the final similarity value whereas descriptors with higher weights (closer to 1) contribute more. Giving a weight of 0 to a descriptor is equivalent to removing it from the analysis. A missing value automatically changes the weight w_j to 0.

The following numerical example illustrates the computation of coefficient S_{15}. In the example, two sites are described by eight quantitative environmental descriptors. Values R_j (the range of values among all objects, for each descriptor y_j) given in the table have been calculated for the whole database prior to computing coefficient S_{15}. Weights w_{12j} are only used in this example to eliminate descriptors with missing values (Kronecker delta function):

	Descriptors j								Sum
Object $\mathbf{x}_1$	2	2	–	2	2	4	2	6	
Object $\mathbf{x}_2$	1	3	3	1	2	2	2	5	
w_{12j}	1	1	0	1	1	1	1	1	= 7
R_j	1	4	2	4	1	3	2	5	
$\lvert y_{1j} - y_{2j} \rvert$	1	1	–	1	0	2	0	1	
$\lvert y_{1j} - y_{2j} \rvert / R_j$	1	0.25	–	0.25	0	0.67	0	0.20	
$w_{12j} s_{12j}$	0	0.75	0	0.75	1	0.33	1	0.80	= 4.63

thus $S_{15}(\mathbf{x}_1, \mathbf{x}_2) = 4.63/7 = 0.66$.

Another general coefficient of similarity was proposed by Estabrook & Rogers (1966). The similarity between two objects is, as in S_{15}, the sum of the partial similarities by descriptors, divided by the number of descriptors for which there is information for the two objects. In their original publication, the authors used state 0 to

identify missing values, but any other convention is acceptable, like *NA* in R. The general equation of this coefficient is the same as for Gower's coefficient (eq. 7.21):

$$S_{16}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{p} w_{12j} s_{12j}}{\sum_{j=1}^{p} w_{12j}} \tag{7.22}$$

As in S_{15}, the w_j parameters may be used as weights (between 0 and 1) instead of only playing the roles of Kronecker deltas. The coefficient of Estabrook & Rogers differs from S_{15} in the computation of the partial similarities s_j.

Partial similarity

In the paper of Estabrook & Rogers (1966), the state values were positive integers and the descriptors were either ordered or unordered. The partial similarity between two objects for a given descriptor j is computed using a monotonically decreasing function of partial similarity. Among all possible functions of this type, the authors empirically chose the following function of two numbers, d and k:

$$\begin{aligned} s_{12j} &= f(d_{12j}, k_j) = \frac{2(k+1-d)}{2k+2+dk} && \text{when } d \le k \\ s_{12j} &= f(d_{12j}, k_j) = 0 && \text{when } d > k \end{aligned} \tag{7.23}$$

where d is the distance between the states of the two objects $\mathbf{x}_1$ and $\mathbf{x}_2$ for descriptor j, i.e. the same value as $|y_{1j} - y_{2j}|$ in eq. 7.20; k is a parameter, determined *a priori* by the user for each descriptor, that describes how far non-null partial similarities are permitted to go. Parameter k is equal to the largest difference d for which the partial similarity s_{12j} (for descriptor j) is allowed to be larger than 0. Values k for the various descriptors may be quite different from one another. For example, for a descriptor coded from 1 to 4, one might decide to use $k = 1$ for this descriptor; for another descriptor with code values from 1 to 50, $k = 10$ could be used. For qualitative descriptors, k is set to 0.

In order to fully understand the partial similarity function s_{12j} (eq. 7.23), readers are invited to compute s_{12j} by hand in the following numerical example. Values k, are provided for each descriptor in the table:

	Descriptors j						$S_{16}(\mathbf{x}_1, \mathbf{x}_2)$
Object $\mathbf{x}_1$	2	1	3	4	2	1	
Object $\mathbf{x}_2$	2	2	4	3	2	3	
k_j	1	0	1	2	1	1	
	↓	↓	↓	↓	↓	↓	
$s_{12j} = f(d_{12j}, k_j)$	1.0 + 0 + 0.4 + 0.5 + 1.0 + 0						= 2.9 / 6 = 0.483

Table 7.1 Values of the partial similarity function $f(d, k)$ in coefficients S_{16} and S_{20}, for the most usual values of k (adapted from Legendre & Rogers, 1972: 594).

k	d							
	0	1	2	3	4	5	6	7
0	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	1	0.40	0.00	0.00	0.00	0.00	0.00	0.00
2	1	0.50	0.20	0.00	0.00	0.00	0.00	0.00
3	1	0.55	0.29	0.12	0.00	0.00	0.00	0.00
4	1	0.57	0.33	0 18	0.08	000	0.00	0.00
5	1	0.59	0.36	0.22	0.13	0.05	0.00	0.00

Values taken by the partial similarity function for the first values of k are shown in Table 7.1. If $k = 0$ for all descriptors, S_{16} is identical to the simple-matching coefficient for multistate descriptors (eq. 7.19).

These same values of function $f(d, k)$ are shown in Fig. 7.3a, which illustrates how the function decreases with increasing d. It is easy to see that function $f(d, k)$, which was originally defined by Estabrook & Rogers for discontinuous descriptors (coded only with integers: 0, 1, 2, 3, …), can actually be used with real-number descriptors since the function only requires that d and k be differences, i.e. natural numbers. Figure 7.3a also raises the question: could $f(d, k)$ take negative values? To accomplish that, Legendre & Chodorowski (1977) proposed to simply leave out the second line of eq. 7.23 stating that $f(d, k) = 0$ when $d > k$. This is shown in Fig. 7.3b, where the function decreases over the interval $[0, \infty)$, taking negative values when $d > (k + 1)$; differences are subtracted from resemblances in this form of the coefficient. This contributes to further separate dissimilar objects when the similarity matrix is subjected to clustering (Chapter 8).

Partial similarity matrix

The major interest of S_{16} over all other coefficients is the possibility to use predefined partial similarity matrices for environmental descriptors. Estabrook & Rogers (1966) proposed this alternative for situations where function $f(d, k)$ does not adequately describe the relationships between objects for some descriptor j. The approach consists in providing the computer program with a "do-it yourself" matrix that describes the partial similarities between all states of descriptor j. This partial similarity matrix replaces in eq. 7.22 the values $s_{12j} = f(d_{12j}, k)$ computed by eq. 7.23

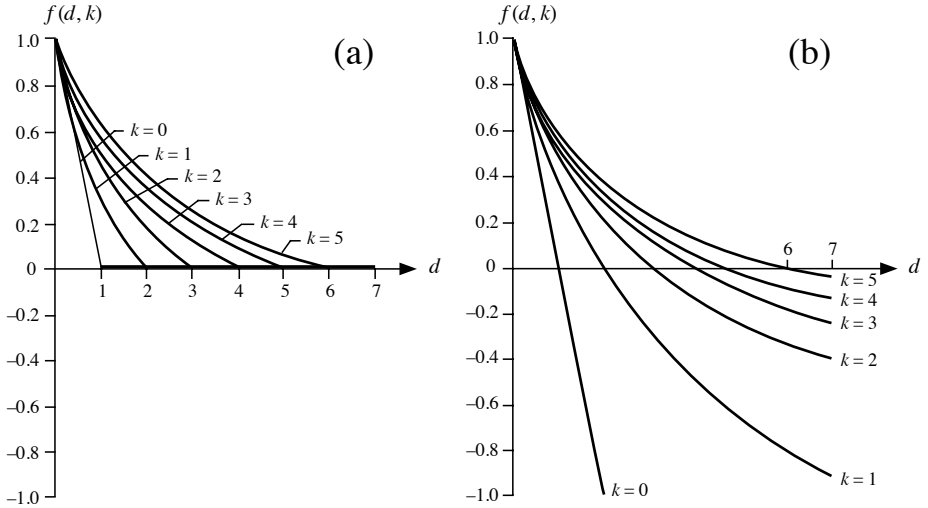

Figure 7.3 Coefficients S_{16} and S_{20}: change in $f(d, k)$ as a function of d, for six values of k, (a) under the condition $f(d, k) = 0$ when $d > k$; (b) without this condition. Adapted from Legendre & Chodorowski (1977).

for descriptor j; the partial similarity of the other descriptors is computed with eq. 7.23. Partial similarity matrices may be constructed for any or all descriptors if needed.

Ecological application 7.3a

Partial similarity matrix

In a study of terrestrial fauna, Legendre *et al*. (1978) used the following partial similarity matrix (matrix unpublished) for the descriptor "category of aquatic ecosystem", which was one of the descriptors of the ecosystems where tracks of animal had been recorded in the snow blanket:

	1	2	3	4	5	6	7	8	9
1. <5% water	1.0								
2. 5-15% water	0.4	1.0							
3. >15% water, lakes < 250 ha	0.0	0.4	1.0						
4. >15% water, lakes 250-500 ha	0.0	0.3	0.8	1.0					
5. >15% water, lakes 500-1000 ha	0.0	0.2	0.7	0.8	1.0				
6. >15% water, lakes 1000-2500 ha	0.0	0.2	0.6	0.7	0.8	1.0			
7. >15% water, lakes > 2500 ha	0.0	0.2	0.5	0.6	0.7	0.8	1.0		
8. >5% rivers	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	
9. Bordering large rivers	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	1.0

The upper triangle of the matrix is not given; it is symmetric to the lower one. The diagonal may also be left out because the partial similarity of a state with itself must be 1. It is shown here to indicate that the matrix contains similarities, not distances. The matrix means that a site from an area with less than 5% of its surface covered by water is given a partial similarity $s_j = 0.4$ with another site from an area with 5 to 15% of its surface covered by water. State 1 has partial similarity with state 2 only; lake systems only have partial similarities with other lake systems, the similarity decreasing as the difference in lake areas increases; and rivers only have partial similarities when compared to other rivers. Partial similarity matrices are especially useful with descriptors that are nonordered, or only partly ordered as is the case here.

Partial similarity matrices represent a powerful way of using unordered or partly ordered descriptors in multivariate data analyses. They are useful in the following cases:

- When, from the user's point of view, function $f(d, k)$ (eq. 7.23) does not adequately describe the partial similarity relationships.

- When the descriptor states are not fully ordered. For example, in a study on ponds, the various states of descriptor "water temperature" may be followed by state "dry pond", which is quite different from a lack of information.

- If some states are on a scale different from that of the other states. For example, 0-10, 10-20, 20-30, 30-40, and then 50-100, 100-1000, and >1000.

- With nonordered or only partly ordered descriptors (including "circular variables" such as directions of the compass card or hours of the day), if one considers that pairs of sites coded into different states are partly similar, as in Ecological application 7.3a.

4 — Asymmetrical quantitative coefficients

Subsection 7.3.3 started with an extension of coefficient S_1 to multi-state descriptors. In the same way, the binary coefficients described in Subsection 7.3.2 could be extended to accommodate multi-state species abundance data. For example, Jaccard's coefficient would become:

$$S_7(\mathbf{x}_1, \mathbf{x}_2) = \frac{\text{agreements}}{p - \text{double-zeros}}$$

where the 'agreement' quantity in the numerator is the number of species with *the exact same* abundance at the two sites. This form would obviously cause the loss of part of the information carried by species abundance data.

The classic indices of compositional similarity described in Subsection 7.3.2 are highly sensitive to sample size, especially for assemblages containing many rare species. Chao *et al*. (2005) developed new forms of the Jaccard (S_7) and Sørensen (S_8) indices, applicable to quantitative community composition data, that estimate and take into account the number of unseen shared species. A full description of these indices,

based on a probabilistic derivation, is found in the Chao *et al*. (2005) paper. Function ***vegdist()*** of the VEGAN library (with method = "chao") produces distances corresponding to the modified abundance-based Jaccard similarity index ($D = 1 - S$). Numerical simulations reported in the Chao *et al*. (2005) paper show that the new estimators of the Jaccard and Sørensen indices are considerably less biased than the corresponding classic indices (S_7, S_8) when a substantial proportion of the species are missing from the sample data.

Other measures are available for species abundance data. They are divided in two categories: the coefficients that can be used with either raw or normalized data and the measures whose application should be limited to normalized data.

• As discussed in Subsection 1.5.6 and Section 7.7, the distribution of abundances of a species across an ecological gradient may be strongly skewed. Normalization of species abundances often calls for square root, double square root, or logarithmic transformations. Another way to obtain approximately normal data is to use a scale of relative abundances with boundaries forming a geometric progression, for example a scale from 0 (absent) to 7 (very abundant). The Anderson *et al*. (2006) transformation (eq. 7.66) is an example of such a recoding method.

• Abundances thus normalized reflect the role of each species in the ecosystem better than the raw abundance data, since a species represented by 100 individuals at a site does not have a role 10 times as important in the ecological equilibrium as another species represented by 10 individuals, everything else being equal. The former is perhaps twice as important as the latter; this is the ratio obtained after applying a base-10 logarithmic transformation, and assuming that numbers 100 and 10 at the site are representative of true relative abundances in the population.

Some coefficients lessen the effect of the largest differences and may therefore be used with raw species abundances, whereas others compare the different abundance values in a more linear way and are thus better adapted to normalized data.

In the group of coefficients to be used with raw species abundances, the best-known is a coefficient attributed to the Polish mathematician H. Steinhaus by Motyka (1947) and Motyka *et al*. (1950). This measure has been rediscovered a number of times; its one-complement is known as the percentage difference, Odum, or Bray-Curtis coefficient (eq. 7.58; see note there). It is sometimes incorrectly attributed to anthropologist Czekanowski (1909 and 1913; Czekanowski's *mean character difference* coefficient is described in Subsection 7.4.1, eq. 7.45). The Steinhaus coefficient compares two sites ($\mathbf{x}_1$, $\mathbf{x}_2$) in terms of the minimum abundance of each species:

$$S_{17}(\mathbf{x}_1, \mathbf{x}_2) = \frac{W}{(A+B)/2} = \frac{2W}{(A+B)} \qquad \textbf{(7.24)}$$

where W is the sum of the minimum abundances of the various species, this minimum being defined as the abundance at the site where the species is the rarest. A and B are the sums of the abundances of all species at each of the two sites or, in other words, the

total number of specimens observed or captured at each site, respectively. Consider the following numerical example:

	Species abundances						A	B	W
Site $\mathbf{x}_1$	7	3	0	5	0	1	16		
Site $\mathbf{x}_2$	2	4	7	6	0	3		22	
Minimum	2	3	0	5	0	1			11

$$S_{17}(\mathbf{x}_1, \mathbf{x}_2) = \frac{2 \times 11}{16 + 22} = 0.579$$

This measure is closely related to the Sørensen coefficient (S_8): if presence-absence data are used instead of species counts, S_{17} becomes S_8 (eq. 7.11).

The distance version of this coefficient, $D_{14} = 1 - S_{17}$, is a semimetric, as shown in the example that follows eq. 7.58. A consequence is that principal coordinate analysis of a D_{14} resemblance matrix is likely to produce negative values. Solutions to this problem are discussed in Subsection 9.3.4. The easiest solution is to base the principal coordinate analysis on square-root-transformed distances $D = \sqrt{1 - S_{17}}$ instead of $D = 1 - S_{17}$ (Table 7.2).

The Kulczynski coefficient (1928) also belongs to the group of measures that are appropriate for raw abundance data. The sum of minima is first compared to the grand total at each site; then the two values are averaged:

$$S_{18}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{2}\left(\frac{W}{A} + \frac{W}{B}\right) \qquad \textbf{(7.25)}$$

For presence-absence data, S_{18} becomes S_{13} (eq. 7.16). For the numerical example above, coefficient S_{18} is computed as follows:

$$S_{18}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{2}\left(\frac{11}{16} + \frac{11}{22}\right) = 0.594$$

Coefficients S_{17} and S_{18} always produce values between 0 and 1, although Kulczynski (1928) multiplied the final value by 100 to obtain a percentage. Kulczynski's approach, which consists in computing the average of two comparisons, seems more arbitrary than Steinhaus' method, in which the sum of minima is compared to the mean of the two site sums. In practice, values of these two coefficients are almost monotonic.

The following coefficients belong to the group adapted to "normalized" abundance data, meaning here unskewed or not strongly skewed frequency distributions. These coefficients parallel S_{15} and S_{16} of the previous subsection. Concerning coefficient S_{19}, Gower (1971a) had initially proposed that his general coefficient S_{15} should exclude

double-zeros from the comparison (Subsection 7.3.3); this makes it well-suited for quantitative species abundance data. Since the differences between states are computed as $|y_{1j} - y_{2j}|$ and are thus linearly related to the measurement scale, this coefficient should be used with previously normalized data. The general form is:

$$S_{19}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{p} w_{12j} s_{12j}}{\sum_{j=1}^{p} w_{12j}}, \text{ where} \qquad \textbf{(7.26)}$$

Partial similarity

- $s_{12j} = 1 - [|y_{1j} - y_{2j}| / R_j]$, as in S_{15},
- and $w_{12j} = 0$ when $y_{1j} = y_{2j}$ = absence of the species, i.e. $(y_{1j} + y_{2j}) = 0$;
- while $w_{12j} = 1$ in all other cases.

For binary (presence-absence) species data, S_{19} is equivalent to the Jaccard coefficient S_7. The weights w_j could be made to vary between 0 and 1, either to reflect the biomasses or biovolumes of the different species, or to compensate for selective effects of the sampling gear.

Legendre & Chodorowski (1977) proposed an asymmetrical coefficient of similarity that parallels S_{16}. This measure uses a slightly modified version of the partial similarity function $f(d, k)$ (eq. 7.23), or else an imposed matrix of partial similarities as in Ecological application 7.3a. Since S_{20} processes all differences d in the same way, irrespective of whether they correspond to high or low values in the scale of abundances, it is better to use this measure with normalized abundance data. The only difference between S_{16} and S_{20} is in the way in which double-zeros are handled. The general form of the coefficient is the sum of the partial similarity values over all species, divided by the total number of species in the combined two sites:

$$S_{20}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{p} w_{12j} s_{12j}}{\sum_{j=1}^{p} w_{12j}}, \text{ where} \qquad \textbf{(7.27)}$$

Partial similarity

- $s_{12j} = f(d_{j12}, k_j)$ $\begin{cases} = \dfrac{2(k+1-d)}{2k+2+dk} & \text{when } d \le k \quad \text{(I)} \\ = 0 & \text{when } d > k \quad \text{(II)} \\ = 0 & \text{when } y_{j1} \text{ or } y_{j2} = 0 \text{ (i.e. } y_{j1} \times y_{j2} = 0) \quad \text{(III)} \end{cases}$

When comparing the presence of a species at a site with its absence at the other site, a similarity of 0 is imposed in point III to acknowledge a strong ecological difference.

• or else $s_{12j} = f(y_{1j}, y_{2j})$ is given by a partial similarity matrix, as in Ecological application 7.3a, in which $s_{12j} = 0$ when y_{1j} or $y_{2j} = 0$,

• and $w_{12j} = 0$ when y_{1j} or y_{2j} = absence of information, or
when $y_{1j} = y_{2j}$ = absence of the species, i.e. $(y_{1j} + y_{2j}) = 0$,

• while $w_{12j} = 1$ in all other cases. Else, w_{12j} may receive a value between 0 and 1, as explained above for S_{19}.

In summary, the properties of coefficient S_{20} are the following:

• when $d_j > k_j$, the partial similarity between sites is $s_{12j} = 0$ for species j (see $f(d, k)$, part II);

• when $d_j = 0$, then $s_{12j} = 1$ (see $f(d, k)$, part I), except when $y_{1j} = 0$ or $y_{2j} = 0$ (see $f(d, k)$, part III);

• $f(d, k)$ decreases with increasing d, for a given k;

• $f(d, k)$ increases with increasing k, for a given d;

• when $y_{1j} = 0$ or $y_{2j} = 0$, the partial similarity between sites is $s_{12j} = 0$ for species j, even if d_{12j} is not larger than k_j (see $f(d, k)$, part III);

• when $k_j = 0$ for all species j, S_{20} is the same as the Jaccard coefficient (S_7) for multi-state descriptors.

The above properties correspond to the opinion that ecologists may have on the problem of partial similarities between normalized (or at least not strongly skewed) species abundances. Depending on the scale chosen (0 to 5 or 0 to 50, for example), function $f(d, k)$ can be used to contrast to various degrees the differences between species abundances, by increasing or decreasing k_j, for each species j if necessary. An example of clustering using this measure of similarity is presented in Ecological application 8.2.

The last quantitative coefficient that excludes double-zeros is called the χ^2 *similarity*. It is the complement of the χ^2 metric (D_{15}; Section 7.4):

$$S_{21}(\mathbf{x}_1, \mathbf{x}_2) = 1 - D_{15}(\mathbf{x}_1, \mathbf{x}_2) \qquad \textbf{(7.28)}$$

The discussion of how species that are absent from two sites are excluded from the calculation of this coefficient is deferred to the presentation of D_{15}.

5 — Probabilistic coefficients

Probabilistic measures form a special category. These coefficients are based on statistical estimation of the significance of the relationship between objects.

Goodall's probabilistic coefficient (1964, 1966a) takes into account the frequency distribution of the various states of each descriptor in the whole set of objects. Indeed, it is less likely for two sites to both contain the same rare species than a more frequent species. In this sense, when estimating the similarity between sites, agreement for a rare species should be given more importance than for a frequent species. Goodall's probabilistic index, which had been originally developed for taxonomy, seems

especially meaningful for ecological classifications, because abundances of species in different sites are stochastic functions (Sneath & Sokal, 1973: 141). Orlóci (1978) suggested to use it for clustering sites (Q mode). The index has also been used in the R mode, for clustering species and identifying associations (Subsection 7.5.2).

The probabilistic coefficient of Goodall is based on the probabilities of the various states of each descriptor. The resulting measure of similarity is itself a probability, namely the complement of the probability that the resemblance between two sites is due to chance.

The probabilistic index, as formulated by Goodall (1966a), is a general taxonomic measure in which binary and quantitative descriptors can be used together. The coefficient as presented here follows the modifications of Orlóci (1978) and is limited to the clustering of sites based on species abundances. It also takes into account the remarks made at the beginning of Subsection 7.2.2 concerning double-zeros. The resulting measure is therefore a simplification of Goodall's original coefficient, oriented towards the clustering of sites. The computational steps are as follows:

(a) A partial similarity measure s_j is first calculated for all pairs of sites and for each species j. Because there are n sites, the number of partial similarities s_j to compute, for each species, is $n(n-1)/2$. If the species abundances have been normalized, one may choose either the partial similarity measure $s_{12j} = 1 - [|y_{1j} - y_{2j}| / R_j]$ from Gower's S_{19} coefficient or function s_{12j} from coefficient S_{20}, which were both described above. In all cases, double-zeros must be excluded. This is done by multiplying the partial similarities s_j by Kronecker delta w_{12j}, whose value is 0 upon occurrence of a double-zero. For raw species abundance data, Steinhaus' similarity S_{17}, computed for a single species at a time, may be used as the partial similarity measure. The chord and Hellinger distances, D_3 and D_{17}, could also be used. The outcome of this first step is a partial similarity matrix, containing as many *rows* as there are species in the ecological data matrix (p) and $n(n-1)/2$ *columns*, i.e. one column for each pair of sites; see the numerical example below.

(b) In a second table of the same size, for each species j and each of the $n(n-1)/2$ pairs of sites, one computes the proportion of partial similarity values belonging to species j that are larger than or equal to the partial similarity of the pair of sites being considered; the s_j value under consideration is itself included in the calculation of the proportion. The larger the proportion, the less similar are the two sites with regard to the given species.

(c) The above proportions or probabilities are combined into a site × site similarity matrix, using Fisher's method, i.e. by computing the product Π of the probabilities relative to the various species. Since none of the probabilities is equal to 0, there is no problem in combining these values, but one must assume that the probabilities of the different species are independent vectors. If there are correlations among species, one may use, instead of the original descriptors of species abundance (Orlóci, 1978: 62), a

matrix of component scores from a correspondence or principal coordinate analysis of the original species abundance data (Sections 9.2 and 9.3).

(d) There are two ways to define Goodall's similarity index. In the first approach, the products Π are put in increasing order. Following this, the similarity between two sites is calculated as the proportion of products that are larger than or equal to the product for the pair of sites considered:

$$S_{22}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{pairs} \mathrm{d}}{n(n-1)/2} \text{ where } \begin{cases} \mathrm{d} = 1 \text{ if } \Pi \geq \Pi_{12} \\ \mathrm{d} = 0 \text{ if } \Pi < \Pi_{12} \end{cases} \quad \textbf{(7.29)}$$

(e) In the second approach, the χ^2 value corresponding to each product is computed, under the hypothesis of independence of the products:

$$\chi^2_{12} = -2 \log_e \Pi_{12}$$

This χ^2-statistic has $2p$ degrees of freedom (p is the number of species). The similarity index is the complement of the probability associated with this χ^2, i.e. the complement of the probability that a χ^2 value taken at random exceeds the observed χ^2 value:

$$S_{23}(\mathbf{x}_1, \mathbf{x}_2) = 1 - \text{prob}\ (\chi^2_{12}) \quad \textbf{(7.30)}$$

It should be clear to the reader that the value of Goodall's index for a given pair of sites may vary depending on the sites included in the computation, since it is based on the rank of the partial similarity for that pair of sites among all pairs. This makes Goodall's measure different from the other coefficients discussed so far.

The following numerical example illustrates the computation of Goodall's index. In this example, five ponds are characterized by the abundances of eight zooplankton species. Data are on a scale of relative abundances, from 0 to 5 (data from Legendre & Chodorowski, 1977).

Species	Ponds					Range R_j
	212	214	233	431	432	
1	3	3	0	0	0	3
2	0	0	2	2	0	2
3	0	2	3	0	2	3
4	0	0	4	3	3	4
5	4	4	0	0	0	4
6	0	2	0	3	3	3
7	0	0	0	1	2	2
8	3	3	0	0	0	3

(a) Gower's matrix of partial similarities has 8 rows and $n(n-1)/2 = 10$ columns which correspond to the 10 pairs of ponds:

Species	Pairs of ponds									
	212	212	212	212	214	214	214	233	233	431
	214	233	431	432	233	431	432	431	432	432
1	1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	1	0	0
3	0.33	0	0	0.33	0.67	0.33	1	0	0.67	0.33
4	0	0	0.25	0.25	0	0.25	0.25	0.75	0.75	1
5	1	0	0	0	0	0	0	0	0	0
6	0.33	0	0	0	0.33	0.67	0.67	0	0	1
7	0	0	0.50	0	0	0.50	0	0.50	0	0.50
8	1	0	0	0	0	0	0	0	0	0

(b) In the next table, one computes, for each pair of sites and each row (species), the proportion of partial similarity values in the row that are larger than or equal to the partial similarity of the pair of sites being considered. The value under consideration is itself included in the proportion. For example, for the pair of ponds (214, 233), the third species has a similarity of 0.67. In the third row, there are 3 values out of 10 that are larger than or equal to 0.67. Thus the ratio associated with the pair (214, 233) in the table is 0.3.

Species	Pairs of ponds									
	212	212	212	212	214	214	214	233	233	431
	214	233	431	432	233	431	432	431	432	432
1	0.1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	0.1	1	1
3	0.7	1	1	0.7	0.3	0.7	0.1	1	0.3	0.7
4	1	1	0.7	0.7	1	0.7	0.7	0.3	0.3	0.1
5	0.1	1	1	1	1	1	1	1	1	1
6	0.5	1	1	1	0.5	0.3	0.3	1	1	0.1
7	1	1	0.4	1	1	0.4	1	0.4	1	0.4
8	0.1	1	1	1	1	1	1	1	1	1

(c) The next table is a site × site symmetric matrix, in which are recorded the products of the terms in each column of the previous table:

Ponds	Ponds				
	212	214	233	431	432
212	–				
214	0.00035	–			
233	1.00000	0.15000	–		
431	0.28000	0.05880	0.01200	–	
432	0.49000	0.02100	0.09000	0.00280	–

(d) The first method for computing the similarity consists in entering, in a site × site matrix, the proportions of the above products that are larger than or equal to the product corresponding to each pair of sites. For example, the product corresponding to pair (212, 431) is 0.28. In the table, there are 3 values out of 10 that are larger than or equal to 0.28, hence the similarity S_{22} (212, 431) = 0.3 (eq. 7.29).

Ponds	Ponds				
	212	214	233	431	432
212	–				
214	1.0	–			
233	0.1	0.4	–		
431	0.3	0.6	0.8	–	
432	0.2	0.7	0.5	0.9	–

(e) If the chosen similarity measure is the complement of the probability associated with χ^2 (eq. 7.30), the following table is obtained. For example, to determine the similarity for pair (212, 431), the first step is to compute χ^2 (212, 431) = $-2 \log_e(0.28) = 2.5459$, where 0.28 is the product associated with that pair in the table at step (d). The value of χ^2 (212, 431) is 2.5459 and the number of degrees of freedom is $2p = 16$, so that the corresponding probability is 0.9994. The similarity is the complement of this probability: S_{23} (212, 431) = 1 – 0.99994 = 0.00006.

Ponds	Ponds				
	212	214	233	431	432
212	–				
214	0.54110	–			
233	0.00000	0.00079	–		
431	0.00006	0.00869	0.08037	–	
432	0.00000	0.04340	0.00340	0.23942	–

Even though the values in the last two tables are very different, the differences are only in term of scale; measures S_{22} computed with eq. 7.29 and S_{23} computed with eq. 7.30 are monotonic to each other.

A probabilistic similarity coefficient among sites has been proposed by palaeontologists Raup & Crick (1979) for species presence-absence data; this is the type of data usually favoured in palaeoecology. Consider the number of species in common to sites h and i; this is statistic a_{hi} of the binary coefficients of Section 7.3. The null hypothesis of the test is H_0: there is no association between sites h and i. Two variants of that null hypothesis are described in steps 2a and 2b below.

Permutation test

The association between sites, measured by a_{hi}, is tested using permutations, and the p-value is used as a distance coefficient. There are two ways of testing the significance of statistic a_{hi} depending on the precise null hypothesis one wants to use.

1. Compute the value of the number of species in common, a_{hi}, for each pair of sites h and i. This is the reference value of the statistic used in step 3. Then go to permutation method 2a or 2b.

2a. The first method, which is actually a simulation rather than a permutation method, implements the null hypothesis (H_0) that each site has received a random subset of species from the species pool, which is either the regional pool or the set of species found in a whole sediment core, while preserving the original species richness at each site. Raup & Crick (1979) called this formulation of H_0 the *random sprinkling hypothesis*. Calculate the relative frequency of each species in the whole data matrix $\mathbf{Y}$, which represents the regional or whole-core species pool. These values will be used as species weights during permutations. Consider site $\mathbf{x}_1$, with species richness s_1. To construct a vector $\mathbf{x}^*_1$ under permutation, draw s_1 species at random from the regional species pool as follows, taking the computed species weights into account.

Random sprinkling hypothesis

- Imagine a stick of length 1 that represents the sum of the weights of all species in the regional pool. For example, species 1 may be very abundant in the region and occupy the first 10% of the stick. Species 2, which is rare, may occupy the next 0.2% of the stick. And so on.

- Draw a number at random from a uniform distribution in the [0, 1] interval. Find the species whose range includes that value on the stick. This is the first species selected in vector $\mathbf{x}^*_1$.

- Remove that species from the stick and rescale the remaining species so that they fully occupy the [0, 1] interval. Draw a new random number from the uniform distribution in the [0, 1] interval. The position of that number along the stick identifies the second species in vector $\mathbf{x}^*_1$.

- Repeat the species selection process until s_1 species have been selected at random from the regional species pool. That completes the construction of vector $\mathbf{x}^*_1$.

Repeat the random species selection process for each site, creating site vectors $\mathbf{x}^*_2$, $\mathbf{x}^*_3$, …, $\mathbf{x}^*_n$ under permutation. Compute the number of species in common, a^*_{hi}, for each pair of site vectors under permutation.

2b. The second permutation method implements the null hypothesis (H_0) that each species is distributed at random among the sites. A permutation under this hypothesis is obtained by permuting at random each vector of species occurrences (i.e. each column of **Y**) independently of the other species vectors. The number of occurrences of each species in the data set is preserved during permutations, but not the original species richness at each site (which is a measure of alpha diversity, Subsection 6.5.3). Compute the number of species in common, a^*_{hi}, for each pair of sites under permutation. This variant of the permutation method for the a_{hi}-statistic was described in McCoy *et al*. (1986).

3. Repeat step 2 (a or b) a large number of times, e.g. 999 or 9999 times, to obtain the null distribution of a^*_{hi}. Add the reference value a_{hi} to the distribution, i.e. the Hope (1968) correction, which is used here to agree with the description of permutation tests in Subsection 1.2.2.

4. For each pair of sites, compare a_{hi} to the reference distribution (obtained at step 3) and calculate the probability $p(a_{hi})$ that $a^*_{hi} \geq a_{hi}$ (one-tailed test), using the procedure described in Subsection 1.2.2.

The Raup-Crick coefficient is available in distance form in function ***raupcrick()*** of VEGAN. The p-values can be computed in several ways, including method 2a above.

Numerical simulations conducted while writing this chapter to check the type I error of the two variants of the Raup-Crick test described above showed that permutation method 2a produced tests that had extremely low levels of type I error, especially when the species had unequal probabilities of occurrence in the species pool. This resulted in a great loss of power when testing the association between sites, to the point that it made the test useless for the analysis of real data because it very seldom recognized significant site associations. Permutation method 2b produced tests that still had low levels of type I error, but not as low as with method 2a, also resulting in a test that had low power to detect significant associations of pairs of sites. As a result, the Raup-Crick test does not seem useful as a test of significance of the similarity of pairs of sites. Ecologists may, however, use the similarity or distance coefficients obtained from that test as they use any other resemblance coefficient among sites, i.e. as the basis for clustering or ordination, without giving them a strict significance test interpretation.

When the test is conducted in the upper tail of the distribution of a^*_{hi} (step 3), the probability $p(a_{hi})$ is expected to be near 0 for sites h and i showing high association, i.e. with more species in common than expected under the null hypothesis. A value near 0.5 indicates that the data support the null hypothesis. One could also test in the lower tail of the distribution, looking for pairs of sites that are significantly dissimilar. The probability would then be calculated as follows: $p(a^*_{hi} \leq a_{hi})$. Significantly dissimilar sites would suggest that some process may have influenced the selection of species, so that fewer species are common to the two sites than expected under the null hypothesis. Taking this approach one step further, Chase *et al*. (2011) rescaled the p-

value in the interval [–1, 1] by subtracting 0.5 and multiplying the result by 2. This modified index indicates whether local communities are more dissimilar (approaching 1), as dissimilar (approaching 0), or less dissimilar (approaching –1) than expected by chance, providing some indication of the possible underlying mechanisms of community assembly.

The probability computed in the upper tail of the distribution of a_{hi}^{*} behaves like a distance (Section 7.4) since p-values are small for similar sites. If necessary, the p-value can be turned into a probabilistic similarity measure of association between sites $\mathbf{x}_1$ and $\mathbf{x}_2$ as follows:

$$S_{27}(\mathbf{x}_1, \mathbf{x}_2) = 1 - \mathrm{p}(a_{12}) \qquad \textbf{(7.31)}$$

Vellend (2004) and Vellend *et al*. (2007) provided a new description of the Raup & Crick (1979) permutation method (paragraph 2a above) and used the coefficient to analyse forest plant communities.

7.4 Q mode: distance coefficients

Distance coefficients are functions that take their maximum values for two objects that are entirely different, and value 0 for two objects that are identical over all descriptors. Distances, like similarities, (Section 7.3), are used to measure the association between *objects*. Distance coefficients may be subdivided in three groups. The first group consists of *metrics* which share the following four properties:

Metric properties

1. minimum 0: if $a = b$, then $D(a, b) = 0$;

2. positiveness: if $a \neq b$, then $D(a, b) > 0$;

3. symmetry: $D(a, b) = D(b, a)$;

4. triangle inequality: $D(a, b) + D(b, c) \geq D(a, c)$. The sum of two sides of a triangle drawn in Euclidean space is necessarily equal to or larger than the third side.

Some authors prefer to restrict the use of *distance* to those coefficients that satisfy the four metric properties and use *dissimilarity* as the general term for all coefficients, i.e. metric, semimetric and nonmetric; see the following paragraphs.

The second group of distances are the *semimetrics* (or *pseudometrics*). These coefficients do not obey the triangle inequality, which is a theorem in Euclidean geometry. These measures cannot directly be used to order points in a *metric* or *Euclidean space* because, for three points (a, b and c), the sum of the distances from a to b and from b to c may be smaller than the distance between a and c. Numerical examples are given in Subsection 7.4.2.

Table 7.2 Some properties of distance coefficients calculated from the similarity coefficients presented in Section 7.3. These properties (from Gower & Legendre, 1986), which will be used in Section 9.3, strictly apply when there are no missing data.

Similarity coefficient	$D = 1 - S$ metric, etc.	$D = 1 - S$ Euclidean	$D = \sqrt{1 - S}$ metric	$D = \sqrt{1 - S}$ Euclidean
$S_1 = \frac{a+d}{a+b+c+d}$ (simple matching; eq. 7.1)	metric	No	Yes	Yes
$S_2 = \frac{a+d}{a+2b+2c+d}$ (Rogers & Tanimoto; eq. 7.2)	metric	No	Yes	Yes
$S_3 = \frac{2a+2d}{2a+b+c+2d}$ (eq. 7.3)	semimetric	No	Yes	No
$S_4 = \frac{a+d}{b+c}$ (eq. 7.4)	nonmetric	No	No	No
$S_5 = \frac{1}{4}\left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right]$ (eq. 7.5)	semimetric	No	No	No
$S_6 = \frac{a}{\sqrt{(a+b)\,(a+c)}} \frac{d}{\sqrt{(b+d)\,(c+d)}}$ (eq. 7.6)	semimetric	No	Yes	Yes
$S_7 = \frac{a}{a+b+c}$ (Jaccard; eq. 7.10)	metric	No	Yes	Yes
$S_8 = \frac{2a}{2a+b+c}$ (Sørensen; eq. 7.11)	semimetric	No	Yes	Yes
$S_9 = \frac{3a}{3a+b+c}$ (eq. 7.12)	semimetric	No	No	No
$S_{10} = \frac{a}{a+2b+2c}$ (eq. 7.13)	metric	No	Yes	Yes
$S_{11} = \frac{a}{a+b+c+d}$ (Russell & Rao; eq. 7.14)	metric	No	Yes	Yes
$S_{12} = \frac{a}{b+c}$ (Kulczynski; eq. 7.15)	nonmetric	No	No	No

The third group of distances consists of *nonmetrics*. These coefficients may take negative values, thus violating the property of positiveness of metrics. Only two such coefficients are described in this book: S_4 and S_{12}.

All similarity coefficient from Section 7.3 can be transformed into distances, as mentioned in Section 7.2. The metric and Euclidean properties of distance coefficients resulting from the transformations $D = (1 - S)$ and $D = \sqrt{1 - S}$ are shown in Table 7.2. These properties determine how to use them in principal coordinate analysis (PCoA, Section 9.3). Stating that a distance coefficient is *not* metric or Euclidean actually means that the coefficient is, sometimes or often, not metric or Euclidean; it does not mean that the coefficient is never metric or Euclidean. A coefficient is *likely* to be metric or Euclidean when the binary form of the coefficient, whose code name given in the table, is known by the proof of a theorem to be metric or Euclidean, and

Table 7.2 Continued.

Similarity coefficient	$D = 1 - S$ metric, etc.	$D = 1 - S$ Euclidean	$D = \sqrt{1 - S}$ metric	$D = \sqrt{1 - S}$ Euclidean
$S_{13} = \frac{1}{2}\left[\frac{a}{a+b} + \frac{a}{a+c}\right]$ (eq. 7.16)	semimetric	No	No	No
$S_{14} = \frac{a}{\sqrt{(a+b)\,(a+c)}}$ (Ochiai; eq. 7.17)	semimetric	No	Yes	Yes
$S_{15} = \sum w_j s_j / \sum w_j$ (Gower; eq. 7.21)	metric	No	Yes	Likely* (S_1)
$S_{16} = \sum w_j s_j / \sum w_j$ (Estabrook & Rogers; eq. 7.22)	metric	No	Yes	Likely* (S_1)
$S_{17} = \frac{2W}{A+B}$ (Steinhaus; eq. 7.24)	semimetric	No	Likely* (S_8)	Likely* (S_8)
$S_{18} = \frac{1}{2}\left[\frac{W}{A} + \frac{W}{B}\right]$ (Kulczynski; eq. 7.25)	semimetric	No	No* (S_{13})	No* (S_{13})
$S_{19} = \sum w_j s_j / \sum w_j$ (Gower; eq. 7.26)	metric	No	Yes	Likely
$S_{20} = \sum w_j s_j / \sum w_j$ (Legendre & Chodorowski; 7.27)	metric	No	Yes	Likely* (S_7)
$S_{21} = 1 - \chi^2\ metric$ (eq. 7.28)	metric	Yes	Yes	Yes
$S_{22} = 2\,(\sum d)\,/\,n\,(n-1)$ (Goodall; eq. 7.29)	semimetric	No	–	–
$S_{23} = 1 - p(\chi^2)$ (Goodall; eq. 7.30)	semimetric	No	–	–
$S_{26} = (a + d/2)\,/\,p$ (Faith, 1983; eq. 7.18)	metric	No	Yes	Yes

* These results follow from the properties of the corresponding binary coefficients (coefficient numbers given), when continuous variables are replaced by binary variables.
– Property unknown for this coefficient.

Euclidean coefficient

test runs using quantitative data have never turned up cases to the contrary. A coefficient is said to be Euclidean if the distances are fully embeddable in Euclidean space; principal coordinate analysis (Section 9.3) of such a distance matrix does not produce negative eigenvalues.

For ordered descriptors, distance functions are described in Subsection 7.4.1, in addition to those derived from similarity coefficients, found in Table 7.2. The metric and Euclidean properties of these distance coefficients are shown in Table 7.3. How to use the various distance coefficients is summarized in Tables 7.4 and 7.5.

Table 7.3 Some properties of the distance coefficients described in Section 7.4.

Distance coefficient	D metric, etc.	D Euclidean	$\sqrt{D}$ metric	$\sqrt{D}$ Euclidean
D_1 (Euclidean distance; eq. 7.32)	metric	Yes	Yes	Yes
D_2 (average distance; eq. 7.34)	metric	Yes	Yes	Yes
D_3 (chord distance; eqs. 7.35, 7.36)	metric	Yes	Yes	Yes
D_4 (geodesic metric; eq. 7.37)	metric	No	Yes	Yes
D_5 (Mahalanobis generalized distance; eq. 7.38)	metric	Yes	Yes	Yes
D_6 (Minkowski metric; eq. 7.43)	metric	*	–	–
D_7 (Manhattan metric; eq. 7.44)	metric	No	Yes	Yes
D_8 (mean character difference; eq. 7.45)	metric	No	Yes	Yes
D_9 (index of association; eqs. 7.47, 7.48)	metric	No	Yes	Yes
D_{10} (Canberra metric; eq. 7.49)	metric	No	Yes	Yes
D_{11} (coefficient of divergence; eq. 7.51)	metric	Yes	Yes	Yes
D_{12} (coefficient of racial likeness; eq. 7.52)	nonmetric	No	No	No
D_{13} (nonmetric coefficient; eq. 7.57)	semimetric	No	Yes	Yes
D_{14} (percentage difference; eq. 7.58)	semimetric	No	Yes	Yes
D_{15} (χ^2 metric; eq. 7.54)	metric	Yes	Yes	Yes
D_{16} (χ^2 distance; eq. 7.55)	metric	Yes	Yes	Yes
D_{17} (Hellinger distance; eq. 7.56	metric	Yes	Yes	Yes
D_{18} (distance between species profiles; eq. 7.53)	metric	Yes	Yes	Yes
D_{19} (modified mean character difference; eq. 7.46)	semimetric	No	No	No

* The result depends on the exponent r.
– Not tested for all exponents r.

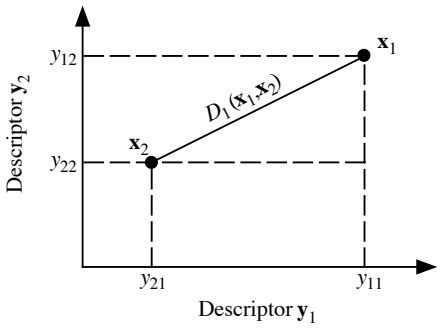

Figure 7.4 Computation of the Euclidean distance (D_1) between objects $\mathbf{x}_1$ and $\mathbf{x}_2$ in 2-dimensional space.

1 — Metric distances

Metric distances have been developed for quantitative descriptors, but they have occasionally been used with semiquantitative descriptors. Some of these measures (D_1, D_2, D_5 to D_8, D_{12}) should not be used, in general, with species abundances, as will be seen in the paradox described below, which results from the handling of double-zeros in the same way as any other value of the descriptors. Coefficients D_3, D_4, D_9 to D_{11} and D_{15} to D_{19} are, on the contrary, well adapted to species abundance data.

The most common metric measure is the *Euclidean distance*. It is computed using Pythagoras' formula from site-points positioned in a p-dimensional space called a *metric* or *Euclidean space*:

$$D_1(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2} \qquad \textbf{(7.32)}$$

When there are only two descriptors, this expression becomes the measure of the hypotenuse of a right-angled triangle (Fig. 7.4; Section 2.4):

$$D_1(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(y_{11} - y_{21})^2 + (y_{12} - y_{22})^2}$$

The square of D_1 may also be used for clustering purpose. One should notice, however, that D_1^2 is a semimetric, which makes it less appropriate than D_1 for ordination:

$$D_1^2(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{p} (y_{1j} - y_{2j})^2 \qquad \textbf{(7.33)}$$

The Euclidean distance does not have an upper limit, its value increasing indefinitely with the number of descriptors. The value also depends on the scale of each descriptor: changing the scale of some or all descriptors changes the values of D_1 in a non-monotonic way. The latter problem may be avoided by using standardized variables (eq. 1.12) instead of the original data, or by restricting the use of D_1 and other distances of the same type (D_2, D_6, D_7 and D_8) to dimensionally homogeneous data matrices (Chapter 3).

Species abundance paradox

The Euclidean distance may lead to the following paradox when it is used as a measure of resemblance among sites based on species abundances: sites without any species in common may be at smaller distance than other sites sharing species. This would be incorrect from an ecologist's point of view. This paradox is illustrated by a numerical example also used in Fig. 7.8 (data modified from Orlóci (1978: 46):

Sites	Species		
	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$
$\mathbf{x}_1$	0	4	8
$\mathbf{x}_2$	0	1	1
$\mathbf{x}_3$	1	0	0

From these data, the following distances are calculated among the sites:

Sites	Sites		
	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\mathbf{x}_1$	0	7.6158	9.0000
$\mathbf{x}_2$	7.6158	0	1.7321
$\mathbf{x}_3$	9.0000	1.7321	0

The Euclidean distance between sites $\mathbf{x}_2$ and $\mathbf{x}_3$, which have no species in common, is smaller than the distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ which share species $\mathbf{y}_2$ and $\mathbf{y}_3$. From an ecologist's point of view, this is an incorrect assessment of the relationships among sites. For environmental descriptors on the contrary, double zeros may well be a valid basis for comparing sites. D_1 should therefore not be used for comparing sites based on species abundance data. The main difficulty in ecology concerning the Euclidean distance arises from the fact that frequently used ordination methods, i.e. principal component and redundancy analyses, order objects in the multidimensional space of descriptors using D_1. The ensuing problems are discussed in Sections 7.7 and 9.1.

Various modifications of D_1 have been proposed. First, the effect of the number of descriptors may be tempered by computing an *average distance*:

$$D_2^2(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{p}\sum_{j=1}^{p}(y_{1j} - y_{2j})^2 \text{ or } D_2(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{D_2^2} \qquad \textbf{(7.34)}$$

While it is difficult to show that D_1 is sensitive to double zeros because D_1 has no upper bound, that demonstration is easy for D_2: because of the division by p, D_2 has a maximum value of 1 for presence-absence data. Consider the following example:

	Species												Sum
Object $\mathbf{x}_1$	1	1	0	1	1	1	0	1	0	0	0	0	
Object $\mathbf{x}_2$	0	0	1	1	1	0	1	1	0	0	0	0	
$(y_{1j} - y_{2j})^2$	1	1	1	0	0	1	1	0	0	0	0	0	= 5

With the first 8 columns of the data table ($p = 8$), $D_2^2 = 5/8 = 0.625$ ($D_2 = 0.79057$), whereas with all 12 columns ($p = 12$), $D_2^2 = 5/12 = 0.41667$ ($D_2 = 0.64550$). Adding double zeros has reduced the distance value; this effect would also be demonstrated with abundance data. D_2 is then a symmetrical coefficient in the sense of Subsections 7.2.2 and 7.3.1. This conclusion also applies to D_1.

Orlóci (1967b) proposed to use the *chord distance* to analyse community composition data. That distance, which is also widely used in genetics (Cavalli-Sforza & Edwards, 1967), has a maximum value of $\sqrt{2}$ for sites with no species in common and a minimum of 0 when two sites share the same species *in the same proportions* of the site vector lengths, without it being necessary for these species to be represented by the same *numbers of individuals* at the two sites. This measure is the Euclidean distance computed after scaling the site vectors to length 1 (normalization of a vector, eq. 2.7). After normalization, the Euclidean distance computed between two objects (sites) is equivalent to the length of a chord joining two points within a segment of a sphere or hypersphere of radius 1. If there are only two species involved, the normalization places the sites on the circumference of a 90° sector of a circle with radius 1 (Fig. 7.5). The chord distance may also be computed directly from non-normalized data through the following formulas:

$$D_3(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{2\left(1 - \frac{\sum_{j=1}^{p} y_{1j} y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2 \sum_{j=1}^{p} y_{2j}^2}}\right)} = \sqrt{\sum_{j=1}^{p} \left(\frac{y_{1j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2}} - \frac{y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{2j}^2}}\right)^2} \quad \textbf{(7.35)}$$

The right-hand formula is a modified form of the Euclidean distance formula. The inner part of the left-hand form is the cosine of the angle (θ) between the two site vectors (eq. 2.9). So the chord distance formula can be written:

$$D_3 = \sqrt{2\,(1 - \cos\theta)} \quad \textbf{(7.36)}$$

The chord distance is maximum when the species found at two sites are completely different. In such a case, the normalized site vectors are at 90° from each other on the

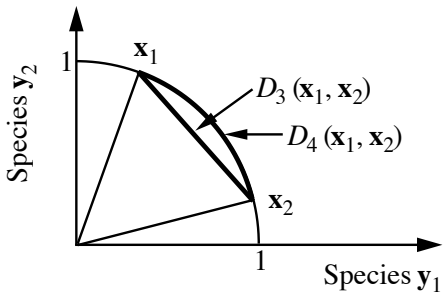

Figure 7.5 Computation of the chord distance D_3 and geodesic metric D_4 between sites $\mathbf{x}_1$ and $\mathbf{x}_2$.

circumference of a 90° sector of a circle (when there are only two species), or on the surface of a segment of a hypersphere (for p species), and the distance between the two sites is $\sqrt{2}$. This measure solves the problem caused by sites having different total abundances of species as well as the paradox explained above for D_1. Indeed, with D_3, the distances between pairs of sites for the numerical example are:

Sites	Sites		
	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\mathbf{x}_1$	0	0.3204	1.4142
$\mathbf{x}_2$	0.3204	0	1.4142
$\mathbf{x}_3$	1.4142	1.4142	0

The chord distance is an Euclidean metric since it is computed with the Euclidean distance formula (eq. 7.35 right). Adding any number of double zeros to a pair of sites does not change the value of D_3, which is thus an asymmetrical coefficient in the sense of Subsections 7.2.2 and 7.3.4. Since double zeros do not influence the chord distance, it can be used to compare sites described by species abundances.

A transformation of the previous measure, known as the *geodesic metric*, measures the length of the arc at the surface of the hypersphere of unit radius (Fig. 7.5):

$$D_4\,(\mathbf{x}_1, \mathbf{x}_2) \;=\; \arccos\left[1 - \frac{D_3^2\,(\mathbf{x}_1, \mathbf{x}_2)}{2}\right] \qquad \textbf{(7.37)}$$

In the numerical example, pairs of sites ($\mathbf{x}_1$, $\mathbf{x}_3$) and ($\mathbf{x}_2$, $\mathbf{x}_3$), with no species in common, are at an angle of 90°, whereas sites ($\mathbf{x}_1$, $\mathbf{x}_2$), which share two of the three species, are separated by a smaller angle (18.4°).

Mahalanobis (1936) developed a generalized distance that takes into account the covariances among descriptors; it produces identical results for variables that are standardized or not. This measure computes the distance between two points in a space whose axes are not necessarily orthogonal, in order to take into account the correlations among descriptors. The formula for the *Mahalanobis generalized distance* between two objects $\mathbf{x}_1$ and $\mathbf{x}_2$ from a data table $\mathbf{X}$ is the following:

$$D_5^2(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{d}_{12}\mathbf{V}^{-1}\mathbf{d}'_{12} \quad \textbf{(7.38)}$$

In this equation, $\mathbf{d}_{12}$ is the row vector (length = p) of the absolute value differences between two objects over the p variables and $\mathbf{V}$ is the covariance matrix over all objects in the group (matrix $\mathbf{X}$). The Mahalanobis generalized distance is the square root of $D_5^2(\mathbf{x}_1, \mathbf{x}_2)$. The principal component analysis framework (Section 9.1) will provide a geometric interpretation of Mahalanobis distances among objects (eq. 9.14). The Mahalanobis distance is also the distance preserved among group means in a canonical space of linear discriminant functions (Section 11.3).

The Mahalanobis distance is also used for comparing *groups of objects*, $\mathbf{w}_1$ and $\mathbf{w}_2$, containing n_1 and n_2 objects respectively, that are described by the same p variables. The square of the generalized distance is given by the following formula in that case:

$$D_5^2(\mathbf{w}_1, \mathbf{w}_2) = \overline{\mathbf{d}_{12}}\mathbf{V}^{-1}\overline{\mathbf{d}'_{12}} \quad \textbf{(7.39)}$$

In this equation, $\overline{\mathbf{d}_{12}}$ is the row vector (length = p) of the absolute value differences between the *means* of the p variables in the two groups of objects. $\mathbf{V}$ is the pooled within-group dispersion matrix of the two groups of objects, estimated from the matrices of sums of squares and cross products among descriptors *centred within each of the two groups*, then added up term by term and divided by $(n_1 + n_2 - 2)$, as in discriminant analysis (Table 11.8) and in multivariate analysis of variance:

$$\mathbf{V} = \frac{1}{n_1 + n_2 - 2}[(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2] \quad \textbf{(7.40)}$$

$\mathbf{S}_1$ and $\mathbf{S}_2$ are the dispersion matrices (eq. 4.6) of the two groups, so that $\mathbf{V}$ takes into account the within-group covariances among descriptors. This formula can also be used to calculate the distance between a single object and a group.

If one wishes to test D_5 for significance, the within-group dispersion matrices must be homogeneous (homoscedasticity, Box 1.4). Homoscedasticity of matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ can be tested using Kullback's test (eq. 11.41) or through the multivariate generalization of Levene's test of homogeneity of variances proposed by Anderson (2006); the latter is available in function ***betadisper()*** of VEGAN. The test of significance also assumes multinormality of the within-group distributions (Sections 4.3 and 4.6) although the generalized distance tolerates some degree of deviation from this condition.

To perform the test of significance, the generalized distance is transformed into Hotelling's T^2 (1931) statistic, using the following equation:

$$T^2 = \frac{n_1 n_2}{(n_1 + n_2)} D_5^2 \tag{7.41}$$

The F-statistic is computed as follows:

$$F = \frac{n_1 + n_2 - (p + 1)}{(n_1 + n_2 - 2)\, p} T^2 \tag{7.42}$$

with p and $[n_1 + n_2 - (p + 1)]$ degrees of freedom. Statistic T^2 is a generalization of Student's t-statistic to the multidimensional case. It allows one to test the hypothesis that two groups originate from populations with similar centroids. The final generalization to several groups, called Wilks Λ (lambda), is discussed in Section 11.3 (eq. 11.42).

The Euclidean distance D_1 is the second degree ($r = 2$) of the *Minkowski metric*:

$$D_6(\mathbf{x}_1, \mathbf{x}_2) = \left[\sum_{j=1}^{p} \left| y_{1j} - y_{2j} \right|^r \right]^{1/r} \tag{7.43}$$

Forms of this metric with $r > 2$ are seldom used in ecology because powers higher than 2 give too much importance to the largest differences $\left| y_{1j} - y_{2j} \right|$. For the exact opposite reason, exponent $r = 1$ is used in many instances. The basic form,

$$D_7(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{p} \left| y_{1j} - y_{2j} \right| \tag{7.44}$$

is known as the *Manhattan metric*, *taxicab metric*, or *city-block metric*. This refers to the fact that, for two descriptors, the distance between two objects is the distance on the abscissa (descriptor y_1) plus the distance on the ordinate (descriptor y_2), like the distance travelled by a taxicab around blocks in a city with an orthogonal plan like Manhattan. This metric presents the same problem for double-zeros as in the Euclidean distance and thus leads to the same paradox.

The *mean character difference* ("durchschnittliche Differenz", in German), proposed by anthropologist Czekanowski (1909),

$$D_8(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{p} \sum_{j=1}^{p} \left| y_{1j} - y_{2j} \right| \tag{7.45}$$

has the advantage over D_7 of not increasing with the number of descriptors p. Distance D_8 is metric, since it is the Manhattan metric divided by p which is constant for a given

data matrix, but it is not Euclidean. $\sqrt{D_8}$ is, however, metric and Euclidean. When applied to presence-absence data, D_8 becomes the one-complement of the simple matching coefficient $(1 - S_1)$.

Equation 7.45 can be used with species abundances if one modifies it to exclude double-zeros from the calculations. This is done by replacing p by pp, which is the number of pairs of values in site vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ that are not double zeros:

$$D_{19}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{pp} \sum_{j=1}^{p} |y_{1j} - y_{2j}| \qquad \textbf{(7.46)}$$

For abundance data, this *modified mean character difference* coefficient has no fixed upper limit. Neither D_{19} nor $\sqrt{D_{19}}$ are metric or Euclidean, which limits their use as the basis for principal coordinate ordination (Section 9.3). When applied to species presence-absence data, eq. 7.46 becomes the one-complement of the Jaccard coefficient $(1 - S_7)$, which is metric but not Euclidean, whereas $\sqrt{1 - S_7}$ is both metric and Euclidean (Table 7.2).

Before computing D_{19}, species abundance data must be transformed to reduce their distribution skewness. In that respect, Anderson *et al*. (2006) redescribed distance D_{19} as a modified form of the asymmetrical Gower coefficient (S_{19}). They applied it to abundance data transformed following eq. 7.66 and called this combination the *modified Gower dissimilarity.*

Whittaker's *index of association* (1952) is well adapted to species abundance data, because each species is first transformed into a fraction of the total number of individuals at the site before the subtraction. Empty data rows (e.g. sites), where no species were found, must be excluded from the calculation. The complement of this index is the following distance, which can be seen as a Manhattan-type (D_7) version of the distance between species profiles D_{18}:

$$D_9(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{2} \sum_{j=1}^{p} \left| \frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right| \qquad \textbf{(7.47)}$$

where y_{1+} is the sum of values in row $\mathbf{x}_1$ and y_{2+} is the sum of values in row $\mathbf{x}_2$. The difference is zero for a species when its *relative abundances* are identical at the two sites. An equivalent formula is to compute, over all species, the sum of the smallest relative abundances at the two sites:

$$D_9(\mathbf{x}_1, \mathbf{x}_2) = \left[1 - \sum_{j=1}^{p} \min \left(\frac{y_{1j}}{y_{1+}}, \frac{y_{2j}}{y_{2+}} \right) \right] \qquad \textbf{(7.48)}$$

D_9 takes values between 0 and 1. The metric and Euclidean properties of D_9 were checked over several community composition data tables: D_9 seems to always be

metric, but it is not Euclidean. $\sqrt{D_9}$ seems, however, to always be metric and Euclidean.

The Australians Lance & Williams (1967a) developed several variants of the Manhattan metric, including their *Canberra metric* (Lance & Williams, 1966c):

$$D_{10}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{p} \left[\frac{|y_{1j} - y_{2j}|}{(y_{1j} + y_{2j})} \right] \quad \textbf{(7.49)}$$

which must exclude double-zeros in order to avoid indetermination. This measure has no fixed upper limit. It can be shown that in D_{10}, a given difference between abundant species contributes less to the distance than the same difference between rarer species (Section 7.6). D_{10} is a non-Euclidean metric whereas $\sqrt{D_{10}}$ is both metric and Euclidean.

The Canberra metric is implemented in the ***vegdist()*** function of the VEGAN package with division of D_{10} by pp, which is the number of pairs of values that are not double zeros in the computation of a given D_{10} value. A metric coefficient taking values between 0 and 1 is thus obtained. Like D_{10}, D_{10}/pp is a non-Euclidean metric, which becomes Euclidean when taking its square root. As an ecological *similarity* measure, Stephenson *et al*. (1972) and Moreau & Legendre (1979) used the one-complement of the Canberra metric with division by pp:

$$S(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{1}{pp} D_{10} \quad \textbf{(7.50)}$$

Clark's (1952) *coefficient of divergence* is a modification of D_{10} that uses the Euclidean distance formula:

$$D_{11}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\frac{1}{pp} \sum_{j=1}^{p} \left(\frac{y_{1j} - y_{2j}}{y_{1j} + y_{2j}} \right)^2} \quad \textbf{(7.51)}$$

D_{11} is a metric and Euclidean coefficient with a maximum value of 1. Because, in D_{11}, the difference for each descriptor is first expressed as a fraction, before squaring the values and summing them, this coefficient is appropriate for species abundance data. Double-zeros must be excluded from the computation to avoid indetermination. This coefficient was first described for multivariate taxonomic analysis, where division was by the number of characters (p) included in the calculation. For community composition analysis, division must be by pp, the number of non double-zero species included in the calculation, as in eq. 7.50.

Another coefficient related to D_{11} was developed by Pearson (1926) for anthropological studies under the name *coefficient of racial likeness*. Using this coefficient, it is possible to measure a distance between groups of objects, as with the

Mahalanobis generalized distance D_5, but without taking into account the within-group covariances among descriptors. The formula is:

$$D_{12}(\mathbf{w}_1, \mathbf{w}_2) = \sqrt{\frac{1}{p}\sum_{j=1}^{p}\left[\frac{(\bar{y}_{1j} - \bar{y}_{2j})^2}{(s_{1j}^2/n_1) + (s_{2j}^2/n_2)}\right] - \frac{2}{p}} \qquad \textbf{(7.52)}$$

for two groups of objects $\mathbf{w}_1$ and $\mathbf{w}_2$ containing respectively n_1 and n_2 objects; $\bar{y}_{ij}$ is the mean of descriptor j in group i and s_{ij}^2 is the corresponding variance. Equation 7.52 can produce negative distances when the square-root portion of the equation is smaller than $2/p$, so the coefficient is not a metric in that particular case.

Other metrics are available to calculate the distance among sites using species abundances or other types of frequency data; no negative value is allowed in the data. The first of these coefficients is the distance between species profiles. It is the Euclidean distance (D_1) computed between relative frequencies (e.g. species relative abundances computed by rows) in a frequency table. In the following example,

$$\mathbf{Y} = \begin{bmatrix} 45 & 10 & 15 & 0 & 10 \\ 25 & 8 & 10 & 0 & 3 \\ 7 & 15 & 20 & 14 & 12 \end{bmatrix} \begin{bmatrix} y_{i+} \\ 80 \\ 46 \\ 68 \end{bmatrix} \rightarrow \left[y_{ij}/y_{i+}\right] = \begin{bmatrix} 0.563 & 0.125 & 0.188 & 0.000 & 0.125 \\ 0.543 & 0.174 & 0.217 & 0.000 & 0.065 \\ 0.103 & 0.221 & 0.294 & 0.206 & 0.176 \end{bmatrix}$$

y_{i+} is the sum of the frequencies in row i. After transformation of the data into relative frequencies, the distance between pairs of rows of the right-hand matrix is computed using the Euclidean distance formula D_1 (eq. 7.32). The equation of the *distance between species profiles* is then:

$$D_{18}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p}\left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}}\right)^2} \qquad \textbf{(7.53)}$$

With this equation, the most abundant species contribute predominantly to the sum of squares. This coefficient is asymmetrical, meaning that it is insensitive to double-zeros; the upper limit of D_{18} is $\sqrt{2}$ for sites that have no species in common.

The χ^2 *metric* D_{15} is a weighted form of the distance between species profiles D_{18}. In the calculation of the sum of squares, each squared difference between relative

frequencies of a pair of rows is weighted by the inverse of the frequency of its column j, y_{+j}, computed across the whole table, as shown in the following example:

$$\mathbf{Y} = \begin{bmatrix} 45 & 10 & 15 & 0 & 10 \\ 25 & 8 & 10 & 0 & 3 \\ 7 & 15 & 20 & 14 & 12 \end{bmatrix} \begin{bmatrix} 80 \\ 46 \\ 68 \end{bmatrix} \quad \rightarrow \begin{bmatrix} y_{ij}/y_{i+} \end{bmatrix} = \begin{bmatrix} 0.563 & 0.125 & 0.188 & 0.000 & 0.125 \\ 0.543 & 0.174 & 0.217 & 0.000 & 0.065 \\ 0.103 & 0.221 & 0.294 & 0.206 & 0.176 \end{bmatrix}$$

$$\begin{bmatrix} y_{+j} \end{bmatrix} = \begin{bmatrix} 77 & 33 & 45 & 14 & 25 \end{bmatrix} \quad 194$$

The χ^2 metric is computed using the following equation:

$$D_{15}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} \qquad \textbf{(7.54)}$$

Thus D_{15}, as well as the chi-square distance (D_{16}, next distance), give higher weights to the rare than to the common species in the calculation of the distance. This distance is recommended when the rare species are considered to be good indicators of special ecological conditions.

For the numerical example, computation of D_{15} between the first two sites (rows) gives:

$$D_{15}(\mathbf{x}_1, \mathbf{x}_2) =$$

$$\left[\frac{(0.563 - 0.543)^2}{77} + \frac{(0.125 - 0.174)^2}{33} + \frac{(0.188 - 0.217)^2}{45} + \frac{(0 - 0)^2}{14} + \frac{(0.125 - 0.065)^2}{25} \right]^{1/2}$$

$$= 0.015$$

The fourth species, which is absent from the first two sites, cancels itself out. This is how the χ^2 metric excludes double-zeros from the calculation.

The upper limit of D_{15} is $\sqrt{2}$. This value is only obtained when there is a single species presence (with an abundance of 1) in the sites producing this value and each species has a total abundance of 1 in the data table; there may be, or not, multiple species with abundances of 1 at other sites than those producing values of $D_{15} = \sqrt{2}$. In all other situations, the distances among sites are smaller than $\sqrt{2}$. To avoid indetermination, absent species (with total abundances of 0) must be eliminated from the data table before the coefficient is computed. D_{15} is asymmetrical since it has an upper limit and its value is not affected by double-zeros.

The χ^2 metric D_{15} can be calculated either among the rows or among the columns of a frequency table. If it is computed among the rows (Q-mode analysis), the relative

frequencies y_{ij}/y_{i+} are computed across the values of each object (e.g. site). If it is computed among columns (R-mode analysis), the relative frequencies y_{ij}/y_{+j} are computed across the values of each column (e.g. species), interchanging rows for columns in eq. 7.54. For example, D_{15} was used by Roux & Reyssac (1975) to calculate distances among sites described by species abundances.

A related measure is called the χ^2 *distance* (Lebart & Fénelon, 1971). It differs from the χ^2 metric in that the terms of the sum of squares are divided by the *relative frequency* of each column in the overall table instead of its absolute frequency. In other words, it is identical to the χ^2 metric multiplied by $\sqrt{y_{++}}$ where y_{++} is the sum of all frequencies in the data table:

$$D_{16}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}/y_{++}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} = \sqrt{y_{++}} \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} \quad \textbf{(7.55)}$$

Since D_{16} is simply D_{15} multiplied by a constant, it shares the property of asymmetry of D_{15}. The maximum value of D_{16} is $\sqrt{2y_{++}}$. The χ^2 distance is the distance preserved in correspondence analysis (Section 9.2). More generally, it is used to compute the association between the rows or columns of a frequency table.

The data used above to illustrate the paradox obtained when the Euclidean distance was computed over species abundances are used again here to contrast D_{16} with D_1.

$$\begin{matrix} & & \left[y_{i+} \right] \\ \mathbf{Y} = \begin{bmatrix} 0 & 4 & 8 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} & \begin{bmatrix} 12 \\ 2 \\ 1 \end{bmatrix} & \rightarrow \left[y_{ij}/y_{i+} \right] = \begin{bmatrix} 0 & 0.333 & 0.667 \\ 0 & 0.500 & 0.500 \\ 1 & 0 & 0 \end{bmatrix} \end{matrix}$$

$$\left[y_{+j} \right] = \begin{bmatrix} 1 & 5 & 9 \end{bmatrix} 15$$

Computing D_{16} between rows (sites) 1 and 3 gives:

$$D_{16}(\mathbf{x}_1, \mathbf{x}_2) = \left[\frac{(0-1)^2}{1/15} + \frac{(0.333-0)^2}{5/15} + \frac{(0.667-0)^2}{9/15} \right]^{1/2} = 4.0092$$

The distances between all pairs of sites are:

Sites	Sites		
	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\mathbf{x}_1$	0	0.3600	4.0092
$\mathbf{x}_2$	0.3600	0	4.0208
$\mathbf{x}_3$	4.0092	4.0208	0

Comparison with results obtained for D_1 (after eq. 7.33) shows that the problem caused with D_1 by the presence of double-zeros does not exist here. Distance D_{16} can therefore be used directly with sites described by species abundances, contrary to D_1.

Another coefficient related to the distance between species profiles (D_{18}) is the Hellinger distance, described by Rao (1995). The formula of the Hellinger distance is:

$$D_{17}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left[\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right]^2} \qquad \textbf{(7.56)}$$

Some of its properties are discussed near the end of Section 7.6. Like D_3 and D_{18}, this distance is asymmetrical (i.e. it is insensitive to double-zeros), and its upper limit is $\sqrt{2}$. The Hellinger distance is actually the chord distance D_3 computed on square-root-transformed frequencies (e.g. species abundances). It is highly recommended for clustering or ordination of species abundance data (Prentice, 1980*; Rao, 1995). Rao (1995) recommended this measure as the basis for a new ordination method; one can obtain the same ordination by computing D_{17} among the objects and carrying out principal coordinate analysis (PCoA, Section 9.3) of the resulting distance matrix. When applied to presence-absence data, the chord (D_3) and Hellinger (D_{17}) distances are both related to the Ochiai similarity (S_{14}) as follows:

$$D_3 = D_{17} = \sqrt{2}\sqrt{1 - S_{14}}$$

2 — Semimetrics

Some distance measures do not follow the fourth property of metrics, i.e. the triangle inequality theorem described at the beginning of the present section. As a consequence, they do not allow a proper ordination of sites in a full Euclidean space. They may, however, be used for ordination by principal coordinate analysis after correction for negative eigenvalues (Subsection 9.3.4) or by nonmetric multidimensional scaling (Section 9.4). These measures are called *semimetrics* or *pseudometrics*. Some semimetrics derived from similarities are identified in Table 7.2. Other such measures are presented here.

The distance corresponding to Sørensen's coefficient S_8 was described by Watson *et al*. (1966) under the name *nonmetric coefficient*:

$$D_{13}(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{2a}{2a + b + c} = \frac{b + c}{2a + b + c} \qquad \textbf{(7.57)}$$

* Prentice (1980) called D_{17} the "chord distance". He gave D_3 the name "cosine theta distance".

a, b and c were defined at the beginning of Subsection 7.3.1. The following numerical example shows that D_{13} does not obey the triangle inequality theorem:

Sites	Species				
	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$
$\mathbf{x}_1$	1	1	1	0	0
$\mathbf{x}_2$	0	0	0	1	1
$\mathbf{x}_3$	1	1	1	1	1

Distances between the three pairs of sites are:

$$D_{13}(\mathbf{x}_1, \mathbf{x}_2) = \frac{3+2}{0+3+2} = 1.00$$

$$D_{13}(\mathbf{x}_1, \mathbf{x}_3) = \frac{0+2}{(2\times 3)+0+2} = 0.25$$

$$D_{13}(\mathbf{x}_2, \mathbf{x}_3) = \frac{0+3}{(2\times 2)+0+3} = 0.43$$

Hence $0.25 + 0.43 < 1.00$, which violates the triangle inequality theorem.

Among the measures for species abundance data, the coefficients of Steinhaus S_{17} and Kulczynski S_{18} are semimetrics when transformed into distances (Table 7.2). In particular, $D_{14} = 1 - S_{17}$ was first described by Odum (1950) in distance form, who called it the *percentage difference*. It was also used by Bray & Curtis (1957) in a study of Wisconsin upland forest vegetation*. The percentage difference formula is:

$$D_{14}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{p} |y_{1j} - y_{2j}|}{\sum_{j=1}^{p} (y_{1j} + y_{2j})} = 1 - \frac{2W}{A+B} \qquad \textbf{(7.58)}$$

* It is unclear why some computer packages refer to D_{14} as the Bray-Curtis distance. The main objective of the Bray & Curtis (1957) paper was to describe a new ordination method, which is known as the Bray-Curtis ordination. The authors never pretended that they were proposing a new S or D coefficient. They used the similarity form of the D_{14} coefficient (S_{17}, eq. 7.24) in their paper (p. 329), transforming it into a distance (p. 332) before computing their new ordination method. Bray & Curtis referred to Motyka *et al*. (1950) for the origin of this similarity coefficient, which is also described by Motyka (1947) who stated that the formula of coefficient S_{17} had first been proposed by Professor H. Steinhaus. In their study, Bray & Curtis (1957) made a very restricted application of the Steinhaus coefficient: because they analysed relative tree abundances at sampling sites, the quantities A and B were both equal to a constant and the similarity between stands was thus equal to W in their study.

The value of D_{14} does not change when double zeros are added to a pair of sites, because the values of A, B and W remain unchanged. Hence this coefficient is asymmetrical in the sense of Subsection 7.2.2. For species relative abundances $y_{ij} = y_{ij}/y_{i+}$, where y_{i+} is the row sum, $D_{14} = D_9$ where D_9 is Whittaker's index of association; D_{14} is also equal to $D_7/2$ where D_7 is the Manhattan distance. D_9 and D_7 are both metric but not Euclidean, but $\sqrt{D_9}$ and $\sqrt{D_7}$ are Euclidean. For binary data, D_{14} is equal to D_{13} or $(1 - S_8)$, which is neither metric nor Euclidean, but $\sqrt{D_{13}}$ is Euclidean.

Contrary to the Canberra metric D_{10}, differences between abundant species contribute the same to D_{14} as differences between rare species. This may be seen as a desirable property, for instance when using normalized species abundance data. Bloom (1981) compared the Canberra metric, the percentage difference and other indices to a theoretical standard. For these data, he showed that only D_{14} (or S_{17}) accurately reflected the true resemblance along its entire 0 to 1 scale, whereas D_{10}, for example, underestimated the resemblance over much of its 0 to 1 range.

The following numerical example, from Orlóci (1978: 59), shows that D_{14} does not obey the triangle inequality theorem and is thus not a metric distance:

Quadrats	Species				
	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$
$\mathbf{x}_1$	2	5	2	5	3
$\mathbf{x}_2$	3	5	2	4	3
$\mathbf{x}_3$	9	1	1	1	1

The distances between the three pairs of sites are:

$$D_{14}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1 + 0 + 0 + 1 + 0}{17 + 17} = 0.059$$

$$D_{14}(\mathbf{x}_1, \mathbf{x}_3) = \frac{7 + 4 + 1 + 4 + 2}{17 + 13} = 0.600$$

$$D_{14}(\mathbf{x}_2, \mathbf{x}_3) = \frac{6 + 4 + 1 + 3 + 2}{17 + 13} = 0.533$$

hence $0.059 + 0.533 < 0.600$, which violates the triangle inequality theorem. Coefficient D_{14} is thus not a metric distance. Table 7.3 shows that D_{14}, which is equal to $(1 - S_{17})$, is also not Euclidean. The consequence is that ordination of sites by principal coordinate analysis (PCoA, Section 9.3) based upon D_{14} is likely to produce negative eigenvalues and complex axes. The way to obtain a distance matrix that is both metric and Euclidean before PCoA is to take the square root of D_{14} (Table 7.3).

7.5 R mode: coefficients of dependence

The main purpose of R-mode analysis is to investigate the relationships among *descriptors*; dependence (R-mode) matrices may also be used, in some cases, as the computational basis for the ordination of *objects*, e.g. in principal component or linear discriminant analyses (Sections 9.1 and 11.3). Following the classification of descriptors in Table 1.2, dependence coefficients will be described for quantitative, semiquantitative, and qualitative descriptors. This will be followed by special measures to assess the dependence between species, to be used for the identification of biological associations (Section 8.9).

Most dependence coefficients are amenable to *statistical testing*. For such coefficients, it is thus possible to associate a matrix of probabilities with the dependence matrix, if required by subsequent analyses. While it is not always legitimate to apply statistical tests of significance, it is never incorrect to compute a dependence coefficient among variables. For example, there is no objection to computing a Pearson correlation coefficient for any pair of metric variables, but these same variables must be normally distributed (Sections 4.2 and 4.3) and the sites must be independent realizations (Sections 1.1 and 1.2) to legitimately test the significance of the coefficient using the standard parametric test; a permutation test (Section 1.2) can, however, be used with non-normal data. A test of significance only allows one to reject, or not, a specific null hypothesis concerning the value of the statistic (here, the coefficient of dependence), whereas the coefficient itself measures the intensity of the relationship between descriptors. Table 7.6 summarizes the use of R-mode coefficients with ecological variables.

1 — Descriptors other than species abundances

Measures of resemblance in the present subsection, which summarizes the coefficients described in Chapters 4, 5 and 6, are used for comparing physical, chemical, geological, and other environmental variables. Measures adapted for species presence-absence and abundance data are described in the next subsection.

The resemblance between *quantitative descriptors* can be computed using parametric measures of dependence, i.e. measures based on parameters of the frequency distributions of the descriptors. These measures are the covariance and the Pearson correlation coefficient; they were described in Chapter 4. They are only adapted to descriptors whose relationships are *linear*.

The *covariance* s_{jk} between descriptors j and k is computed from centred variables $(y_{ij} - \overline{y_j})$ and $(y_{ik} - \overline{y_k})$ (eq. 4.4). The range of values of the covariance has no *a priori* upper or lower limits. The variances and covariances among a group of descriptors form their dispersion matrix **S** (eq. 4.6).

The Pearson *correlation coefficient* r_{jk} is the covariance of standardized descriptors j and k (eqs. 1.12 and 4.7). Coefficients of correlations computed among a group of descriptors form a correlation matrix **R** (eq. 4.8). Correlation coefficients range in values between –1 and +1. The significance of individual coefficients is tested using eq. 4.13, the null hypothesis being generally H_0: $r = 0$, whereas eq. 4.14 is used to test the hypothesis of complete independence among all descriptors. Pearson correlation coefficients should not be computed in Q mode (Box 7.1).

The resemblance between *semiquantitative descriptors* and, more generally between any pair of *ordered* descriptors whose relationship is *monotonic* may be determined using nonparametric measures of dependence (Chapter 5). Since *quantitative* descriptors are ordered, nonparametric coefficients may be used to measure their dependence, as long as they are monotonically related.

Two *nonparametric correlation coefficients* have been described in Section 5.3: Spearman's r and Kendall's τ (tau). In Spearman's r (eq. 5.3), quantitative values are replaced by ranks before computing Pearson's r formula. Kendall's τ (eqs. 5.5 to 5.7) measures the resemblance in a way that is quite different from Pearson's r. Values of Spearman's r and Kendall's τ range between –1 and +1. The significance of individual coefficients (the null hypothesis being generally H_0: $r = 0$) is tested using eq. 5.4 (Spearman's r) or 5.8 (Kendall's τ).

As with Pearson's r above, rank correlation coefficients should not be used in the Q mode. Indeed, even if quantitative descriptors are standardized, the same problem arises as with Pearson's r, i.e. the Q measure for a pair of objects is a function of all objects in the data set. In addition, in most biological sampling units, several species are represented by small numbers of individuals. Because these small numbers are subject to large stochastic variation, the ranks of the corresponding species are uncertain in the reference ecosystem. As a consequence, rank correlations between sites would be subject to important random variation because their values would be based on large numbers of uncertain ranks. This is equivalent to giving preponderant weight to the many poorly sampled species.

The importance of *qualitative descriptors* in ecological research is discussed in Section 6.0. The measurement of resemblance between pairs of such descriptors is based on two-way contingency tables (Section 6.2), whose analysis is generally conducted using X^2 (chi-square) statistics. Contingency table analysis is also the major approach available for measuring the dependence between *quantitative* or *semiquantitative* ordered descriptors that are not monotonically related. The minimum value of X^2 is zero, but it has no *a priori* upper limit. Its formulae (eqs. 6.5 and 6.6) and test of significance are explained in Section 6.2. If all qualitative descriptors have the same number of states, X^2 values can be transformed into contingency coefficients (eqs. 6.19 and 6.20), whose values are in the range [0, 1].

Two-way contingency tables may also be analysed using coefficients derived from information theory. In that case, the amounts of information (B) shared by two

Q-mode correlation Box 7.1

Can one compute Pearson correlation coefficients among rows, i.e. in Q mode? There are at least six objections to that.

• Because the same physical dimensions are present in the numerator and denominator of Pearson's r computed in the R mode (eq. 4.7), the resulting coefficient has no physical dimension, i.e. it is dimensionless (Chapter 3). On the contrary, correlations computed between objects (Q mode) have complex and non-interpretable physical dimensions when the descriptors are not dimensionally homogeneous. Furthermore, in Q mode, the row means $\bar{y}_i$ do not make sense for variables that have different physical dimensions so that the differences $(y_{ij} - \bar{y}_i)$ in eq. 4.7 cannot be computed.

• Physical descriptors are often expressed in arbitrary units (e.g. mm, cm, m, or km are all equally correct length measures). In R mode, the value of r remains unchanged after any arbitrary linear change of units, whereas in Q mode the same operation can dramatically change the values of correlations computed between objects, in unpredictable fashion.

• In order to avoid the two previous problems, it has been suggested to standardize the descriptors (eq. 1.12) before computing correlations in the Q mode. Consider two objects $\mathbf{x}_1$ and $\mathbf{x}_2$: their similarity should be independent of the other objects in the study; removing objects from the data set should not change the value of their similarity. Any change in object composition of the data set would, however, change the standardized variables, and so it would affect the value of the correlation computed between $\mathbf{x}_1$ and $\mathbf{x}_2$. Hence, standardization does not solve the problems because the resulting correlation between two objects would be a function of the values of all the other objects in the data set.

• In the R mode, the central limit theorem (Section 4.3) predicts that, as the number of objects increases, the means, variances, and covariances (or correlations) converge towards their values in the statistical population. Computing these same parameters in the Q mode is likely to have the opposite effect since the addition of new descriptors to the estimation of these parameters is likely to change their values in major and non-trivial ways.

• If correlation coefficients could be used as a general measure of resemblance in Q mode, they should be applicable to the simple case of the description of the proximities among sites, computed from their geographic coordinates X and Y on a map; the correlations obtained from this calculation should reflect in some way the distances among the sites. This is not the case: correlation coefficients computed among sites from their geographic coordinates are all +1 or –1. As an exercise, readers are encouraged to compute an example of their own.

• Correlation coefficients can be tested by the method of permutations, as shown in Subsection 1.2.3. In the R mode, permuting the values of a variable within a column makes physical sense: under H_0, each value could be found at any one site. In the Q mode, however, permuting values within a row of the data matrix does not make sense because, in the real world, these values could not belong to different variables. As an illustration, it would not make sense to move a salinity of 35 psu to the pH column.

Conclusion: coefficients designed for R-mode analysis should not be used in the Q mode. Sections 7.3 and 7.4 describe several Q-mode coefficients whose properties and dimensions are known or easy to determine.

descriptors j and k and exclusive to each one (A and C) are first computed. These quantities may be combined into similarity measures, such as $S(j, k) = B/(A + B + C)$ (eq. 6.15; see also eqs. 6.17 and 6.18), or into distance coefficients such as $D(j, k) = (A + C)/(A + B + C)$ (eq. 6.16). The analysis of multiway contingency tables (Section 6.3) is based on the Wilks X^2-statistic (eq. 6.6).

A *qualitative* descriptor (including a *classification*; Chapter 8) can be compared to a *quantitative* descriptor using the F-statistic of *one-way analysis of variance* (one-way ANOVA; Table 5.2 and text). The classification criterion for this ANOVA is the qualitative descriptor. As long as the assumptions underlying analysis of variance are met (i.e. normality of within-group distributions and homoscedasticity, Box 1.4), the significance of the relationship between the descriptors can be tested using the F–distribution. If the quantitative descriptor does not obey the within-group normality assumption, a permutation test of F can be used. If the comparison is between a *qualitative* and a *semiquantitative* descriptor, *nonparametric one-way analysis of variance* (Kruskal-Wallis H-test; Table 5.2) can be used.

2 — Species abundances: biological associations

The search for species associations is one of the classical problems of community ecology. How to conduct that search using clustering methods is discussed in Section 8.9. The present subsection focuses on the dependence coefficients that are appropriate for the study of species interrelationships. Measures that can be used for presence-absence data are discussed first, followed by measures for quantitative data.

1. *Species presence-absence data*. — There are several approaches in the literature for measuring the association between species based on presence-absence data. Indeed, biological associations may be defined on the basis of the *co-occurrence of species*, instead of the co-fluctuations in abundances in the quantitative approaches described below. Indeed, the definition of association may refer to the sole concept of co-occurrence, as suggested by Fager (1957) who pointed out that associations must group species that are almost always part of one another's biological environment. The reason is that quantitative data may not accurately reflect the proportions of the various species in the environment, because of problems with sampling, preservation, identification or counting, or simply because the concept of individuality is not clear (e.g. plants multiplying through rhizomes; colonial algae or animals), or because the comparison of numbers of individuals does not make ecological sense (e.g. the baobab and the surrounding herbaceous plants). The spatio-temporal aggregration of organisms may also obscure the true quantitative relationships among species, as in the case of plankton patches or reindeer herds. It follows that associations are often defined on the sole basis of the presence or absence of species.

There are many approaches in the literature for measuring the association between species on the basis of presence and absence data. These coefficients are based on the following 2 × 2 contingency table:

Species $\mathbf{y}_2$ \ Species $\mathbf{y}_1$	presence	absence	
presence	a	b	$a + b$
absence	c	d	$c + d$
	$a + c$	$b + d$	$n = a + b + c + d$

where a and d are the numbers of sites where both species are present and absent, respectively, whereas b and c are the numbers of sites where only one of the two species is present; n is the total number of sites. The measures of association between species always exclude the number of double absences, d.

Among the many binary coefficients that exclude double-zeros (Subsection 7.3.2), some have been used for assessing association between species. Jaccard's *coefficient of community* (eq. 7.10) has been used by Reyssac & Roux (1972) in the R mode:

$$S_7(\mathbf{y}_1, \mathbf{y}_2) = \frac{a}{a + b + c}$$

The corresponding distance has been used by Thorrington-Smith (1971) for the same purpose:

$$D = 1 - S_7(\mathbf{y}_1, \mathbf{y}_2) = \frac{b + c}{a + b + c} \qquad \textbf{(7.59)}$$

The Sørensen coefficient (eq. 7.11)

$$S_8(\mathbf{y}_1, \mathbf{y}_2) = \frac{2a}{2a + b + c}$$

was originally defined under the name *coincidence index* for studying species associations (Dice, 1945). The Ochiai coefficient (S_{14}) can also be used in R mode for analysing species presence-absence data.

When used in the R mode, the Sørensen (S_8) coefficient can be tested for significance using random permutations of the observations in one of the species vectors (for the test of association of two species only) or in all species vectors (for simultaneous tests of association among all species). The basic co-occurrence statistic is a, the number of sites where both species are present. For statistic a computed in R mode, a test using permutation method 2b (permutation of one or both columns), described for the Raup & Crick coefficient (eq. 7.31), is equivalent to a permutation test of the Sørensen statistic since the denominator of S_8, $(a + b) + (a + c) = 2a + b + c$,

is invariant under permutation of the values in the columns, and a permutation test of a produces the same permutational probability as a test of $2a$.

Ecological application 7.5

Clua *et al*. (2010) studied the ecology and residence patterns of a group of photo-identified adult sicklefin lemon sharks, *Negaprion acutidens*, at a shark-feeding site monitored by divers during 44 months. An objective of the study was to delineate groups of sharks that were present together and formed recognizable behavioural groups. From the observation data (presence-absence of 29 sharks during 949 dives), the authors computed co-occurrence statistics a among all pairs of sharks and tested their significance using permutation method 2b described for the Raup & Crick coefficient (S_{27}). The p-values of 0.0001, obtained after 9999 random permutations, had a corrected experimentwise p-value of 0.0406 after Holm correction for multiple testing (406 simultaneous tests; Box 1.3). The 52 edges corresponding to these p-value smaller than 0.05 were drawn on a principal coordinate ordination diagram (PCoA, Section 9.3). Five behavioural groups of sharks were recognized on the plot.

An elaborate coefficient was proposed by Fager & McGowan (1963):

$$S_{24}(\mathbf{y}_1, \mathbf{y}_2) = \frac{a}{\sqrt{(a+b)\,(a+c)}} - \frac{1}{2\sqrt{a + \max(b,c)}} \qquad \textbf{(7.60)}$$

Coefficient S_{24} replaced a probabilistic coefficient proposed earlier by Fager (1957). The first part of coefficient S_{24} is the same as the Ochiai coefficient S_{14}, i.e. the geometric mean of the proportions of co-occurrence for each of the two species; the second part is a correction for small sample size.

Krylov (1968) proposed to use the probability associated with the X^2 (chi-square) statistic of the above 2×2 contingency table to test the null hypothesis that two species are distributed independently of each other among the sites. Rejecting H_0 gives support to the alternative hypothesis of association between the two species. In the case of a 2×2 contingency table, and using Yate's correction factor for small samples, the X^2-formula is:

$$X^2 = \frac{n\,[\,|ad - bc| - (n/2)\,]^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)} \qquad \textbf{(7.61)}$$

The number of degrees of freedom for the test of significance is ν = (no. rows – 1) × (no. columns – 1) = 1 (eq. 6.7). The X^2-statistic could also be tested by permutation (Section 1.2). Given that associations should be based on positive relationships between species (negative relationships reflecting competition), Krylov proposed to set $S(\mathbf{y}_1, \mathbf{y}_2) = 0$ when the expected value of co-occurrence, $E = (a + b)(a + c)/n$, is larger than or equal to the observed frequency ($E \geq a$). Following the test, two species are considered associated if the probability (p) computed for their X^2-statistic is

smaller than a pre-established significance level, for example $\alpha = 0.05$. The similarity measure between species is the complement of that probability:

$$S_{25}(\mathbf{y}_1, \mathbf{y}_2) = 1 - p(X^2), \text{ with } \nu = 1, \quad \text{when } E(a) = (a+b)(a+c)/n < a$$
$$S_{25}(\mathbf{y}_1, \mathbf{y}_2) = 0 \quad \text{when } E(a) = (a+b)(a+c)/n \geq a \quad \textbf{(7.62)}$$

The p-value itself can be used as a distance. When the number of sites n is smaller than 20 or a, b, c or d are smaller than 5, Fisher's exact probability formula should be used to compute the p-value instead of a test of X^2. The formula can be found in most textbooks of statistics.

The same formula can be derived from Pearson's ϕ (phi) (eq. 7.9), given that $X^2 = n\phi^2$. Pearson's ϕ is also called the *point correlation coefficient* because it is the general correlation coefficient (eq. 5.1) computed from presence-absence data.

2. *Species abundance data*. — For quantitative species abundance or biomass data, parametric or nonparametric correlation coefficients (Pearson r and Spearman r) can be used to assess the relationships among species (Greig-Smith, 1983; O'Connor & Aarssen, 1987; Myster & Pickett, 1992). When looking for species associations, Legendre (2005) suggested to transform the species abundances through one of the transformations described in Section 7.7 to control for differences in total abundance (or total biomass for biomass data) among sites before computing the correlations among species, in order to linearize the relationships (for Pearson r) or make them more monotonic (for Spearman r).

If the correlations must be transformed into distances before hierarchical clustering or ordination, use the transformations $D = (1 - r)$ or $D = \sqrt{1 - r}$ that were used to transform S into D in Subsection 7.2.1: two species that are identically distributed across the sites have a correlation $r = 1$, hence $D = 0$, which would be the appropriate distance measure in that case. For negative correlations, if any, these transformations produce distances larger than 1, which would cause no problem in clustering or ordination. To induce the clustering or ordination methods to put together species that are either positively or negatively correlated, use $D = (1 - r^2)^{0.5}$ to transform correlations into distances.

The chi-square metric (D_{15}) and the chi-square distance (D_{16}) are appropriate in both the Q and R modes. Both distances can be computed among species before clustering. The D_{16} matrix is obtained by transposing the data matrix so that species are now the rows; then, apply the chi-square distance transformation (eq. 7.70) and compute the Euclidean distance (D_1) on the transformed data.

Whittaker (1972) proposed a coefficient called *SC* (species correlation), constructed like his index of association (distance D_9). Despite its name, this coefficient has no relationship with the Pearson and Spearman correlation formulas. Each abundance value y_{ij} is first transformed into a relative abundance (eq. 7.68) by dividing it by the corresponding row sum y_{i+}, then the coefficient is computed using

the transformed data. As in coefficient D_9 (eqs. 7.47 and 7.48), there are two algebraic forms for the computation of *SC* between species (columns) $\mathbf{y}_1$ and $\mathbf{y}_2$:

$$SC(\mathbf{y}_1, \mathbf{y}_2) = 1 - \frac{1}{2}\sum_{i=1}^{n}\left|\frac{y_{i1}}{y_{1+}} - \frac{y_{i2}}{y_{2+}}\right| = \sum_{i=1}^{n}\min\left(\frac{y_{i1}}{y_{1+}}, \frac{y_{i2}}{y_{2+}}\right) \quad \textbf{(7.63)}$$

where y_{1+} is the sum of values in row $\mathbf{x}_{i=1}$ and y_{2+} is the sum of values in row $\mathbf{x}_{i=2}$. *SC* takes values between 0 and 1. Before clustering, *SC* can be transformed into a distance coefficient by computing

$$D_{20}(\mathbf{y}_1, \mathbf{y}_2) = (1 - SC) \quad \textbf{(7.64)}$$

Probabilistic association

Goodall's probabilistic coefficient (S_{22} or S_{23}, eqs. 7.29 and 7.30) can also be applied to species abundances in the R mode. An example is found in Legendre (1973). This probabilistic coefficient allows one to set an objective limit to species associations; indeed, one may then use a probabilistic definition of an association, such as: "all species that are related at a probability level $(1 - p) \geq 0.95$ are members of the association". Goodall's coefficient has the following meaning in R mode: given *p* species and *n* sites, the similarity of a pair of species is defined as the complement (1 – p) of the probability that any pair of species chosen at random would be as similar as, or more similar than the two species under consideration. Goodall's similarity coefficient is computed as in Subsection 7.3.5, with species interchanged with sites. In step (a), if the species data have been normalized (for example using the transformation $y' = \log(y + 1)$ in eq. 7.65, or eq. 7.66), the partial similarity of Gower's coefficient S_{19} (eq. 7.26)

$$s_{i12} = 1 - [|y_{i1} - y_{i2}| / R_i]$$

may be used to describe the similarity between species y_1 and y_2 at site i. R_i is the range of variation of the normalized species abundances at site i; R_i scales the differences between species for each site.

7.6 Choice of a coefficient

Criteria for choosing a coefficient are summarized in Tables 7.4 to 7.6. In these tables, the coefficients are identified by the names and numbers used in Sections 7.3 to 7.5. The three tables distinguish between coefficients appropriate for species (or frequency) descriptors, and those for other types of descriptors.

Levels 4 and 6 of Table 7.4 require some explanation. Coefficients differentiated in these levels are classified with respect to two criteria, i.e. (a) standardization (or not) of each object-vector prior to the comparison and (b) relative importance given by the coefficient to the abundant or rare species. This defines various types of coefficients.

Type 1 coefficients. Consider two objects, each represented by a vector of species abundances, to be compared using a Q-mode measure. With type 1 coefficients, if there is a given difference between sites for some abundant species and the same difference for a rare species, the two species contribute equally to the similarity or distance between sites. A small numerical example illustrates this property for the percentage difference (D_{14}), which is the complement of Steinhaus' similarity (S_{17}):

Species:	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$
Site $\mathbf{x}_1$	100	40	20
Site $\mathbf{x}_2$	90	30	10
$\lvert y_{1j} - y_{2j} \rvert$	10	10	10
$(y_{1j} + y_{2j})$	190	70	30

Using eq. 7.58 shows that each of the three species contributes 10/290 to the total distance between the two sites. With some coefficients (D_3, D_4, D_9, D_{17}, D_{18}), the standardization of the site-vectors, which is automatically done prior to the computation of the coefficient, may make the result unclear as to the importance given to each species. With these coefficients, the property of "equal contribution" is found only when the two site-vectors are equally important, the importance being measured in different ways depending on the coefficient (see the footnote of Table 7.4).

Type 2a coefficients. — With coefficients of this type, a difference between values for an abundant species contributes less to the distance (and, thus, more to the similarity) than the same difference for a rare species. The *Canberra metric* (D_{10}) belongs to this type. For the above numerical example, calculation of D_{10} (eq. 7.49) shows that species $\mathbf{y}_1$, which is the most abundant, contributes 10/190 to the distance, $\mathbf{y}_2$ contributes 10/70, whereas the contribution of $\mathbf{y}_3$, which is the rarest species, is the largest of the three (10/30). The total distance is $D_{10} = 0.529$. The *coefficient of divergence* (D_{11}; eq. 7.51) also belongs to this type.

Type 2b coefficients. — Coefficients of this type behave similarly to the previous ones, except that the importance of each species is calculated with respect to the whole data set instead of the two site-vectors that are compared. The χ^2 *metric* (D_{15}) is representative of this. In eq. 7.54 and accompanying example, the squared difference between conditional probabilities, for a given species, is divided by y_{+j} which is the total number of individuals belonging to this species at all sites. If this number is large, it reduces the contribution of the species to the total distance between two rows (sites) more than would happen in the case of a rarer species. *Gower's coefficient* (S_{19}; eq 7.26) has the same behaviour (unless special weights w_{12j} are used for some species), since the importance of each species is determined from its range of variation over all sites. The coefficient of Legendre & Chodorowski (S_{20}; eq 7.27) also belongs to this type when parameter k in the partial similarity function s_{12j} for each species is made proportional to its range of variation over all sites.

Legendre *et al*. (1985) suggested that it is more informative to compare dominant or well-represented species than rare taxa, because the latter are generally not well

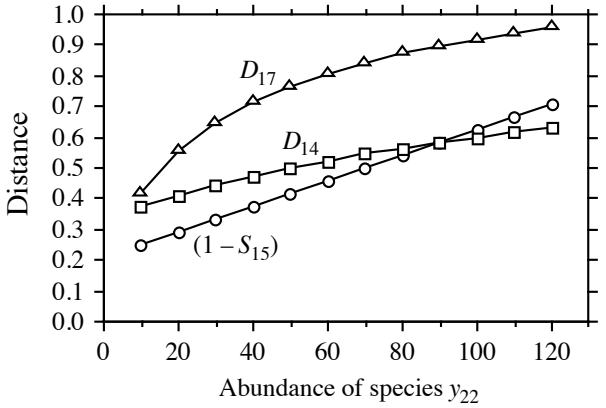

Figure 7.6 Results of an *ordered comparison case series* (OCCAS) where species y_{22} abundance varies from 10 to 120 by steps of 10. The values taken by coefficients $(1 - S_{15})$, D_{14}, and D_{17} are shown.

sampled. This provides an approach for choosing a coefficient. In immature communities, most of the species are represented by small numbers of individuals, so that only a few species are well sampled, whereas, in mature communities, several species exhibit intermediate or high abundances. When calculating similarities between species from immature communities, a reasonable approach may thus be to give more weight to the few well-sampled species (type 2 coefficients) whereas, for sites from mature communities, type 1 coefficients may be more appropriate.

Another way of choosing a resemblance coefficient is to construct an artificial data set representing contrasting situations that the similarity or distance measure should be able to differentiate. Computing several candidate coefficients for the test data will indicate which coefficient is the most appropriate for data of that type. In that spirit,

OCCAS

Hajdu (1981) constructed series of test cases, called *ordered comparison case series* (OCCAS), corresponding to linear changes in the abundances of two species along different types of simulated environmental gradients. The results are distances between sites, computed using different coefficients, for linearly changing species composition.

To illustrate the method, consider one of Hajdu's OCCAS with two species. For these species, site 1 had frequencies $y_{11} = 100$ and $y_{12} = 0$; site 2 had frequency $y_{21} = 50$ whereas y_{22} varied from 10 to 120. Figure 7.6 shows the results for three coefficients: $(1 - S_{15})$ has a completely linear behaviour across the values of y_{22}, D_{14} is not quite linear, and D_{17} is strongly curvilinear.

An ideal coefficient should change linearly when plotted against a series of test cases corresponding to a linear change in species composition, as simulated in OCCAS

Non-linearity

runs. Hajdu (1981) proposed a measure of *non-linearity*, defined as the *standard deviation of the changes* in values of distance between adjacent test cases along the series. A good distance coefficient should also change substantially along the series

and reach its maximum value when the difference in species composition is maximum.

Resolution

Resolution was defined as the *mean change* occurring in distances between adjacent test cases along the series. High linearity is desirable in ordination methods whereas high resolution is desirable in cluster analysis. The ratio of non-linearity over resolution defines a coefficient of variation that should be small for a "good" overall resemblance coefficient.

Resolutions are only comparable among coefficients that are bounded in the interval [0, 1] or [0, $\sqrt{2}$]; as a consequence, this measure should not be used to compare coefficients, such as D_1, D_2, and D_{10}, which do not have an upper bound. Non-linearity near 0 is always a good property, but, again, higher values are only comparable for coefficients that are bounded. Coefficients of variation are comparable because the scale of variation of each specific coefficient is taken into account in the calculation.

Gower & Legendre (1986) used Hajdu's OCCAS to study the behaviour of several similarity and distance coefficients and to make recommendations about their use. They studied 15 coefficients for binary data (all of which are described in the present chapter) and 10 coefficients for quantitative data (5 of them are described here). Among the binary coefficients, S_{12} (eq. 7.15) and the coefficient of Yule (eq. 7.8) were strongly non-linear and should be avoided; all the other coefficients in that study (S_1, S_2, S_3, S_5, S_6, S_7, S_8, S_{10}, S_{13}, S_{14}, as well as eqs. 7.7 and 7.9) behaved well.

The coefficients for quantitative data included in that study were S_{15}, $D_{14} = 1 - S_{17}$, D_2, D_{10} and D_{11}. Coefficients D_2 and S_{15}, which are adapted to physical descriptors (Table 7.5), behaved well. D_2 is a standardized form of the Euclidean distance D_1; they both have the same behaviour.

All coefficients adapted to species abundance data (Table 7.4) that were included in the study (D_{10}, D_{11}, D_{14}) behaved well and are recommended. Coefficients S_{15} and D_{10} had perfect linearity in all specific OCCAS runs; they are thus the best of their kinds for principal coordinate analysis (PCoA, Section 9.3), which is a metric ordination method based on distances.

A later analysis of coefficient $D = \sqrt{D_{14}} = \sqrt{1 - S_{17}}$ showed that its non-linearity was very similar to that of $D_{14} = 1 - S_{17}$; the resolution of $\sqrt{D_{14}}$ was slightly lower than that of D_{14}. Both forms are thus equally suitable for ordination whereas D_{14} may be marginally preferable for clustering purposes. The square root transformation of D_{14}, used in the latter part of Numerical example 1 (continued) in Subsection 9.3.5, offers a simple way to avoid negative eigenvalues in principal coordinate ordination.

Another comparative analysis involving the chi-square metric and related forms (D_{15}, D_{16}, D_{17} and D_{18}) showed that the best of this group for metric ordination (PCoA) is the Hellinger distance (D_{17}), which has the lowest coefficient of variation (best compromise between linearity and resolution), despite the fact that it is strongly non-linear. Other properties of resemblance coefficients have been investigated by Bloom (1981), Wolda (1981) and Hubálek (1982).

Table 7.4 Choice of an association measure among objects (Q mode), to be used with species descriptors (asymmetrical coefficients). For explanation of levels 4 and 6, see the accompanying text.

1) Descriptors: presence-absence or ordered classes on a scale of relative abundances (no partial similarities computed between classes) **see 2**

   2) Metric coefficients: *coefficient of community* (S_7) and variants (S_{10}, S_{11})

   2) Semimetric coefficients: variants of the coef. community (S_8, S_9, S_{13}, S_{14})

   2) Nonmetric coefficient: Kulczynski (S_{12}) (non-linear: not recommended)

   2) Probabilistic coefficient: S_{27}

1) Descriptors: quantitative or semiquantitative (states defined in such a way that partial similarities can be computed between them) **see 3**

   3) Coefficients for raw or normalized abundance data **see 4**

      4) No standardization by object; the same difference for either abundant or rare species, contributes equally to the similarity between sites: *coefficients of Steinhaus* (S_{17}) and *Kulczynski* (S_{18}), *percentage difference* (D_{14}), $\sqrt{D_{14}}$

      4) Standardization by object-vector; if objects are of equal importance*, same contributions for abundant or rare species to the similarity or distance between sites: *chord distance* (D_3), *geodesic metric* (D_4), *index of association* (D_9), *Hellinger dist.* (D_{17}), *dist. between profiles* (D_{18})

      4) Standardization by object-vector*; differences for abundant species (in the whole data set) contribute more than differences between rare species to the similarity (less to the distance) between sites: χ^2 *similarity* (S_{21}), χ^2 *metric* (D_{15}), χ^2 *distance* (D_{16})

   3) Limited to normalized abundances (species distributions not strongly skewed). [Normalization of species abundance data: Sections 1.5.6 and 7.7] **see 5**

      5) Coefficients without associated probability levels **see 6**

         6) Differences for abundant species (for two sites under consideration) contribute more than differences between rare species to the similarity (less to the distance) between sites: *Canberra metric* (D_{10}), *coefficient of divergence* (D_{11}). Both have low resolution: not recommended for clustering

         6) Differences for abundant species (in the whole data set) contribute more than differences between rare species to the similarity (less to the distance) between sites: *asymmetrical Gower coefficient* (S_{19}), *coefficient of Legendre & Chodorowski* (S_{20})

         6) Differences for abundant and rare species contribute the same to the distance between sites: *modified mean character difference* or *modified Gower dissimilarity* (D_{19})

      5) Probabilistic coefficient: *Goodall coefficient* (S_{23})

* D_3 and D_4: importance quantified relative to the length of the row vector $\sqrt{\sum_t y_{ij}^2}$

D_9, D_{15} to D_{18}: importance relative to the sum of individuals in the row vector $\sum_t y_{ij}$

Table 7.5 Choice of an association measure among objects (Q mode), to be used with chemical, geological physical, etc. descriptors (symmetrical coefficients, using double-zeros).

1) Association measured between individual objects **see 2**

  2) Descriptors: presence-absence or multistate (no partial similarities computed between states) **see 3**

    3) Metric coefficients: *simple matching* (S_1) and derived coefficients (S_2, S_6)

    3) Semimetric coefficients: S_3, S_5

    3) Nonmetric coefficient: S_4

  2) Descriptors: multistate (states defined in such a way that partial similarities can be computed between them) **see 4**

    4) Descriptors: quantitative and dimensionally homogeneous **see 5**

      5) Differences enhanced by squaring: *Euclidean distance* (D_1) and *average distance* (D_2)

      5) Differences mitigated: *Manhattan metric* (D_7), *mean character difference* (D_8)

    4) Descriptors: not dimensionally homogeneous; weights (equal or not, according to values w_j used) given to each descriptor in the computation of association measures **see 6**

      6) Descriptors are qualitative (no partial similarities computed between states) and quantitative (partial similarities based on the range of variation of each descriptor): *symmetrical Gower coefficient* (S_{15})

      6) Descriptors are qualitative (possibility of using matrices of partial similarities between states) and semiquantitative or quantitative (partial similarity function for each descriptor): *coefficient of Estabrook & Rogers* (S_{16})

1) Association measured between groups of objects

  7) Removing the effect of correlations among descriptors: *Mahalanobis generalized distance* (D_5)

  7) Not removing the effect of correlations among descriptors: *coefficient of racial likeness* (D_{12})

Table 7.6 Choice of a dependence measure among descriptors (R mode).

1) Descriptors: species abundances	**see 2**
2) Descriptors: presence-absence	**see 3**
3) Coefficients without associated probability levels: S_7, S_8, S_{14}, S_{24}	
3) Probabilistic coefficient: S_{25}	
2) Descriptors: multistate	
4) Data are raw abundances: χ^2 *similarity* (S_{21}), χ^2 *metric* (D_{15}), χ^2 *distance* (D_{16}), *Whittaker's SC* (D_{20})	**see 4**
4) Data are abundances in linear or monotonic relationships	**see 5**
5) Coefficients without associated probabilities: *covariance*, *Pearson r*, *Spearman r*, Pearson or Spearman correlations among chord-transformed or Hellinger-transformed data	
5) Probabilistic coefficients: *probabilities associated to Pearson r* or *Spearman r*, *Goodall coefficient* (S_{23})	
1) Descriptors: chemical, geological, physical, etc.	**see 6**
6) Coefficients without associated probability levels	**see 7**
7) Descriptors are quantitative and linearly related: *covariance*, *Pearson r*	
7) Descriptors are ordered and monotonically related: *Spearman r*, *Kendall* τ	
7) Descriptors are qualitative or ordered but not monotonically related: χ^2, *reciprocal information coefficient*, *symmetric uncertainty coefficient*	
6) Probabilistic coefficients	**see 8**
8) Descriptors are quantitative and linearly related: *probabilities associated to Pearson r*	
8) Descriptors are ordered and monotonically related: *probabilities associated to Spearman r* or *Kendall* τ	
8) Descriptors are qualitative or ordered but not monotonically related: *probabilities associated to* χ^2	

7.7 Transformations for community composition data

In communities sampled over fairly homogeneous environmental conditions, e.g. short environmental gradients, the species composition data contain few zeros, and symmetric association coefficients, including the Euclidean distance D_1, can be used for clustering or ordination. Frequency histograms of individual species may, however, display asymmetric distributions because species tend to have exponential growth when conditions are favourable. This well-known fact has been embedded in the theory of species-abundance models; see He & Legendre (1996, 2002) for a synthetic view of these models. To reduce the asymmetry of the species distributions, a species abundance variable **y** may be transformed to **y**' by taking the square root or the fourth root (equivalent to taking the square root twice), or by using a log transformation:

$$y' = y^{0.5} \quad \text{or} \quad y' = y^{0.25} \quad \text{or} \quad y' = \log(y + c) \qquad \textbf{(7.65)}$$

where y is the species abundance and c is a constant. Usually, $c = 1$ in species abundance log transformations; in this way, an abundance $y = 0$ is transformed into $y' = \log(0 + 1) = 0$ for any logarithmic base. These transformations represent the series of exponents $\gamma = 0.5, 0.25$ and 0 of the Box-Cox transformation (eq. 1.15).

Another interesting transformation that reduces the asymmetry of heavily skewed abundance data is the one proposed by Anderson *et al*. (2006). The abundance data y_{ij} are transformed as follows to a logarithmic scale that makes allowance for zeros:

$$y'_{ij} = \log_{10}(y_{ij}) + 1 \qquad \text{when } y_{ij} > 0 \qquad \textbf{(7.66)}$$

$$\text{or } y'_{ij} = 0 \qquad \text{when } y_{ij} = 0.$$

Hence, for $y_{ij} = \{0, 1, 10, 100, 1000\}$, the transformed values y'_{ij} are $\{0, 1, 2, 3, 4\}$. Note that this is *not* the $\log(y_{ij} + 1)$ transformation. This transformation is available in the ***decostand()*** function of VEGAN (method = "log") where users can choose the base of the logarithm. Changing the base of logarithms in eq. 7.65 (right) produces a linear change among the y'_{ij} values, so it does not induce any change in the relationships among the transformed values. With eq. 7.66 on the contrary, the transformations produced by different bases of logarithms are not perfectly linearly related.

Community composition data sampled over variable environmental conditions, e.g. along long environmental gradients, typically contain many zero values because species are known to generally have unimodal distributions along environmental gradients (ter Braak & Prentice, 1988) and to be absent from sites far from their optimal living conditions. The proportion of zeros is greater when the environmental conditions are more variable across the sampling sites. For association coefficients, this situation generates the double-zero problem that was discussed in Subsection 7.2.2 and leads to the selection of an asymmetrical similarity or distance coefficient for clustering or ordination.

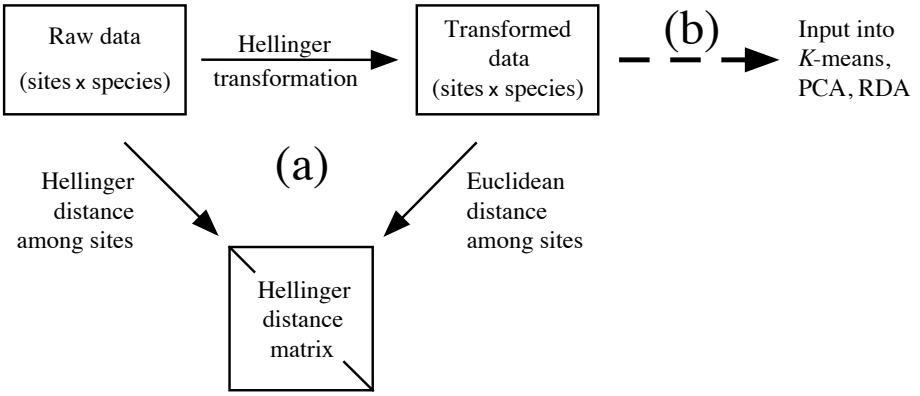

Figure 7.7 (a) Calculation of a distance matrix either directly from the raw data (left diagonal arrow) or through a two-step approach in which the raw data are transformed (horizontal arrow) before computation of the distance matrix (right diagonal arrow). The example shown here uses the Hellinger transformation to obtain the Hellinger distance matrix (D_{17}). The same approach can be used to obtain the chord (D_3), species profile (D_{18}), chi-square metric (D_{15}) and chi-square distance (D_{16}) matrices, as summarized in Fig. 7.8. (b) The transformed species data can also be used as input (dashed arrow) into linear methods of analysis, in particular PCA, RDA, and *K*-means partitioning. Modified from Legendre & Gallagher (2001).

An alternative method of computation for the asymmetrical distance coefficients D_3, D_{15}, D_{16}, D_{17} and D_{18} was proposed by Legendre & Gallagher (2001). The method consists of a transformation of the community composition data followed by the calculation of Euclidean distances (D_1) among sites. These two steps produce the distance function corresponding to the name of the transformation (Fig. 7.7). Data subjected to one of these transformations can also be used directly as input into linear methods of analysis that carry out computations in Euclidean space, such as *K*-means partitioning, PCA, and RDA (Sections 8.8, 9.1, 11.1). This approach is called transformation-based PCA (tb-PCA), transformation-based RDA (tb-RDA), and transformation-based *K*-means partitioning (tb-*K*-means).

1 — Transformation formulas

The following transformations, found in the vertical rectangle in the centre of Fig. 7.8, can be used to obtain the distance coefficients found on their left. The effect of these transformations is to remove the differences in total abundance (for abundance data) or total biomass (for biomass data) from the data, keeping the variations in relative species composition among sites. The chord and Hellinger transformations described below have been in use in community ecology and palaeoecology for a long time (e.g. Noy-Meir *et al.*, 1975; Prentice, 1980). Legendre & Gallagher (2001) showed

Species abundance paradox data ⇒
(3 sites, 3 species)

	Species 1	Species 2	Species 3
Site 1	0	4	8
Site 2	0	1	1
Site 3	1	0	0

Transformations ⇓

$$D_1(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2} \qquad \mathbf{D}_1 = \begin{bmatrix} 0.0000 & 7.6158 & 9.0000 \\ 7.6158 & 0.0000 & 1.7321 \\ 9.0000 & 1.7321 & 0.0000 \end{bmatrix}$$

$$D_3(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left(\frac{y_{1j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2}} - \frac{y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{2j}^2}} \right)^2} \qquad y'_{ij} = \frac{y_{ij}}{\sqrt{\sum_{j=1}^{p} y_{ij}^2}} \qquad \mathbf{D}_3 = \begin{bmatrix} 0.0000 & 0.3204 & 1.4142 \\ 0.3204 & 0.0000 & 1.4142 \\ 1.4142 & 1.4142 & 0.0000 \end{bmatrix}$$

$$D_{18}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} \qquad y'_{ij} = \frac{y_{ij}}{y_{i+}} \qquad \mathbf{D}_{18} = \begin{bmatrix} 0.0000 & 0.2357 & 1.2472 \\ 0.2357 & 0.0000 & 1.2247 \\ 1.2472 & 1.2247 & 0.0000 \end{bmatrix}$$

$$D_{17}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left[\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right]^2} \qquad y'_{ij} = \sqrt{\frac{y_{ij}}{y_{i+}}} \qquad \mathbf{D}_{17} = \begin{bmatrix} 0.0000 & 0.1697 & 1.4142 \\ 0.1697 & 0.0000 & 1.4142 \\ 1.4142 & 1.4142 & 0.0000 \end{bmatrix}$$

$$D_{16}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{y_{++}} \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} \qquad y'_{ij} = \sqrt{y_{++}} \frac{y_{ij}}{y_{i+}\sqrt{y_{+j}}} \qquad \mathbf{D}_{16} = \begin{bmatrix} 0.0000 & 0.3600 & 4.0092 \\ 0.3600 & 0.0000 & 4.0208 \\ 4.0092 & 4.0208 & 0.0000 \end{bmatrix}$$

Figure 7.8 Species abundance paradox data, modified from Orlóci (1978). The paradox is that the Euclidean distance between sites 2 and 3, which have no species in common, is smaller than that between sites 1 and 2 which share species 2 and 3. This results in an incorrect assessment of the ecological relationships among sites. With the other coefficients in this figure, which are asymmetrical, the distance between sites 2 and 3 is larger than that between sites 1 and 2, and the distance between sites 1 and 3 is the same as between sites 2 and 3, or very nearly so. Distance matrix $\mathbf{D}_{15}$ (not shown) is equal to $\mathbf{D}_{16}/\sqrt{y_{++}} = \mathbf{D}_{16}/\sqrt{15}$.

that these transformations were the first step towards the calculation of one of the asymmetrical distances that are appropriate for Q-mode analysis of community data. Only five of the coefficients discussed in this chapter can be computed by the two-step procedure described in Fig. 7.7, i.e. $D_3, D_{15}, D_{16}, D_{17}$ and D_{18}.

Chord transformation

1) *Chord transformation*. — The species abundances from each object (sampling unit) are transformed into a vector of length 1 using the following equation:

$$y'_{ij} = \frac{y_{ij}}{\sqrt{\sum_{j=1}^{p} y_{ij}^2}} \qquad \textbf{(7.67)}$$

where y_{ij} is the abundance of species j in object i. This equation, called the "chord transformation" in Legendre & Gallagher (2001), is available in the program CANOCO (Centring and standardisation for "samples": *Standardise by norm)* and in the ***decostand()*** function of VEGAN (method = "normalize"). If one computes the Euclidean distance (D_1) between two rows of the transformed data table, the resulting value is identical to the chord distance (D_3, eq. 7.35) computed between the rows of the original (untransformed) species abundance data table; this is how the chord distance can be computed through the two-step calculation shown in Fig. 7.7a. As a consequence, after a chord transformation, the community composition data are suitable for PCA or RDA, as well as other methods of analysis that preserve the Euclidean distance among the objects (Fig. 7.7b).

Species profile transformation

2) *Species profile transformation*. — The data can be transformed into profiles of relative species abundances through the following equation:

$$y'_{ij} = \frac{y_{ij}}{y_{i+}} \qquad \textbf{(7.68)}$$

This is a method of data standardisation that is often used prior to analysis, especially when the sampling units are not all of the same size. Data transformed in that way are called *compositional data*. In community ecology, the species assemblage is considered to represent a response of the community to environmental, historical, or other types of forcing; the variation of any single species has no clear interpretation. Compositional data are used because ecologists feel that the vectors of relative proportions of species can lead to meaningful interpretations. Relative abundances can be transformed into percentages by multiplying the values y'_{ij} by 100. Computing Euclidean distances among rows of a data table transformed in this way produces distances among species profiles (D_{18}, eq. 7.53). The transformation to relative abundance profiles is available in the ***decostand()*** function of VEGAN (method = "total"). Statistical criteria investigated by Legendre and Gallagher (2001) show that this is not the best transformation and that the Hellinger transformation (next paragraph) is often preferable.

Abundance data transformed into profiles by eq. 7.68 have the following property: centring the data by columns to means of 0 automatically centres the rows to means of 0. Make sure that the raw abundance data contain no row that sums to 0, though.

Hellinger transformation

3) *Hellinger transformation*. — A modification of the species profile transformation produces the Hellinger transformation:

$$y'_{ij} = \sqrt{\frac{y_{ij}}{y_{i+}}} \qquad \textbf{(7.69)}$$

Computing Euclidean distances among objects of a data table transformed in this way produces a matrix of Hellinger distances among sites (D_{17}, eq. 7.56; Fig. 7.7). The Hellinger distance has good statistical properties as assessed by the criteria of R^2 and monotonicity used by Legendre and Gallagher (2001) in their comparison of transformation methods. The Hellinger transformation is available in the ***decostand()*** function of VEGAN (method = "hellinger").

Chi-square distance transformation

4) *Chi-square distance transformation*. — A more complex modification of the species profile transformation is the chi-square distance transformation:

$$y'_{ij} = \sqrt{y_{++}}\frac{y_{ij}}{y_{i+}\sqrt{y_{+j}}} \qquad \textbf{(7.70)}$$

where y_{ij} is a species presence or abundance value, y_{i+} is the sum of values over row (object) i, y_{+j} is the sum of values over column (species) j, and y_{++} is the sum of values over the whole data table. Euclidean distances computed among the rows of the transformed data table $[y'_{ij}]$ are equal to chi-square distances (D_{16}, eq. 7.55) among the rows of the original, untransformed data table. The chi-square distance transformation is available in the ***decostand()*** function of VEGAN (method = "chi.square").

The chi-square distance transformation equation reduces the value of an abundant species more than that of a rare species. Hence this transformation is interesting when one wants to give more weight to rare species; this is the case when the rare species are considered to be good indicators of special ecological conditions.

Chi-square metric transformation

5) *Chi-square metric transformation*. — The *chi-square metric* (D_{15}) only differs from the *chi-square distance* (D_{16}) by the constant $\sqrt{y_{++}}$ found in eq. 7.70. It can be obtained by the simplified transformation:

$$y'_{ij} = \frac{y_{ij}}{y_{i+}\sqrt{y_{+j}}} \qquad \textbf{(7.71)}$$

followed by calculation of the Euclidean distance. Data transformed using eq. 7.71 are smaller than the same data transformed using eq. 7.70 by a constant factor of $\sqrt{y_{++}}$.

Before applying the transformations described in the previous paragraphs, any of the standardizations investigated by Noy-Meir *et al*. (1975), Prentice (1980), and Faith *et al*. (1987) may be used if the study justifies it. These include species adjusted to equal maximum abundances or equal standard deviations, sites standardised to equal totals, or both. In particular, one may apply a square root or log transformation to the species abundances in order to reduce the asymmetry of the species distributions.

The chord and Hellinger transformations appear to be the best for general use. Legendre & Gallagher (2001) showed that the values of the corresponding distances are monotonically increasing across a simulated ecological gradient and are maximally related (R^2) to the spatial distances along the geographic gradient. Other asymmetrical distances, like D_{14}, that are useful for the analysis of community composition data cannot be obtained through the two-step process of a transformation followed by calculation of the Euclidean distance illustrated in Fig. 7.7. The chord and Hellinger transformations are closely related: chord-transformed abundance data are equal to squared abundance data that are then Hellinger-transformed.

The five transformations described above can be applied to presence-absence data. In that situation, the chord and Hellinger transformations produce identical results, and the corresponding distances, D_3 and D_{17}, are both equal to $\sqrt{2}\sqrt{1-\frac{a}{\sqrt{(a+b)\ (a+c)}}}$ where $\frac{a}{\sqrt{(a+b)\ (a+c)}}$ is the Ochiai similarity coefficient for binary data (S_{14}). Correspondence analysis, which preserves the chi-square distance, has long been used with species presence-absence data; hence the chi-square transformation can also be applied to this type of data.

2 — Numerical example

The modified Orlóci paradox data set was used in Subsection 7.4.1 to shows that the Euclidean distance function may produce misleading results when applied to assemblage composition data. Asymmetrical similarity and distance functions, which were specifically designed for the analysis of community composition data, do not have this drawback. Figure 7.8 (right-hand side) shows, for five distance functions, the distance matrices obtained for these data. From a community ecologist viewpoint, the identity of the species present at two sites is more important for assessment of the differences among these sites than their abundances. Following that conception, sites 1 and 2, which share two species, are more similar to each other than either of them is to site 3, which harbours a single species not found at sites 1 and 2.

Instead of that, Euclidean distances (D_1) show that sites 1 and 2 (D = 7.6158) are more dissimilar than sites 2 and 3 (D = 1.7321). This assessment would be considered incorrect by most community ecologists although the calculations are mathematically correct. In contrast, the four other distance matrices in Fig. 7.8 indicate that the two less dissimilar sites are 1 and 2, an answer that would be considered a correct

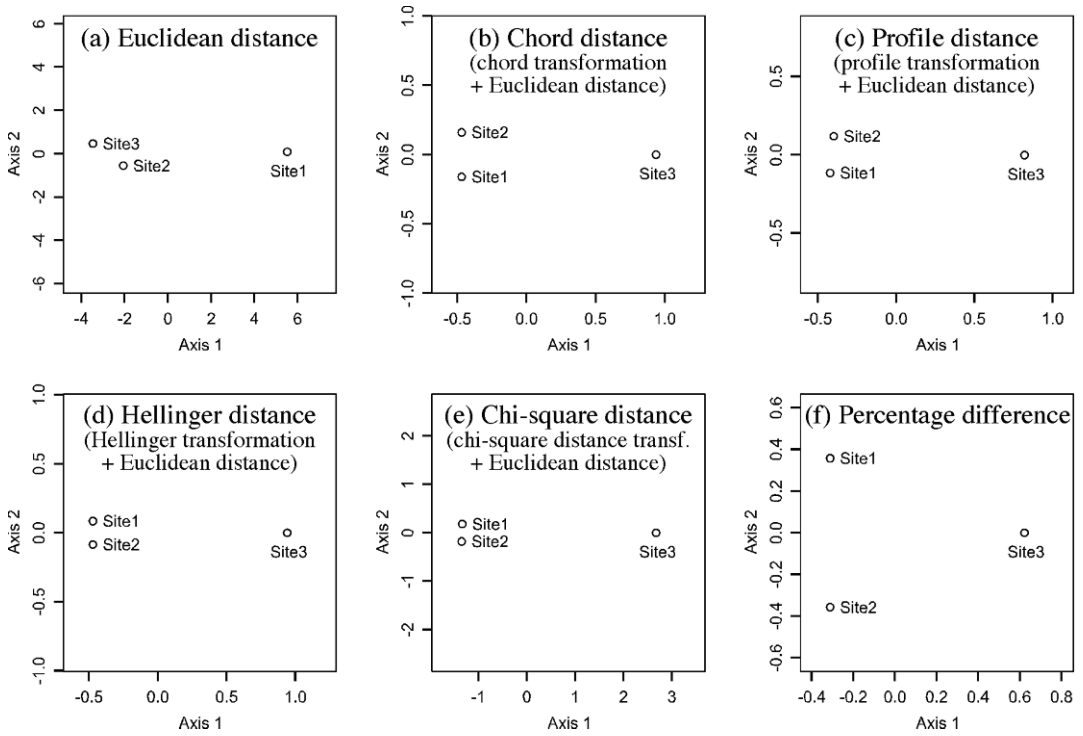

Figure 7.9 Principal coordinate ordination plots (PCoA, Section 9.3) of the distance matrices computed in Fig. 7.8: (a) D_1, (b) D_3, (c) D_{18}, (d) D_{17}, (e) D_{16}, and (f) a PCoA plot of the percentage difference (Steinhaus/Odum/Bray-Curtis) distance matrix (D_{14}) computed for the same data.

assessment of community similarity by most ecologists. Observe also that the chord and Hellinger distances produce a value of $\sqrt{2} = 1.4142$ between sites that have no species in common; this is the maximum value attainable by these distance functions, as noted in Subsection 7.4.1.

Figure 7.9 presents principal coordinate ordination plots (PCoA, Section 9.3) computed from the distance matrices in Fig. 7.8, plus a PCoA plot of the percentage difference matrix (D_{14}) computed for the same data. In the Euclidean distance ordination (Fig. 7.9a), sites 2 and 3 are the closest among the three sites, which may be seen as incorrect for the data under consideration. In all five other ordination plots (Fig. 9b-f), sites 1 and 2 are the closest. The plots also display the interesting property that the different asymmetrical distance functions deal with the differences among sites differently: sites 1 and 2 are the closest to each other in Fig. 7.9e and the farthest in Fig. 7.9f (percentage difference). The distance between sites 1 and 2 would be even larger if PCoA had been computed from square-rooted D_{14} values, which is recommended before PCoA to make percentage difference matrices Euclidean.

3 — Beals smoothing

Beals smoothing is a multivariate transformation designed for species presence/absence community data containing noise and/or many zeros. This transformation replaces the observed presence/absence values of a species by predicted probabilities of occurrence, on the basis of the co-occurrences of that species with the other species in the data set (Beals, 1984; McCune, 1994). The transformed values can be used as input in multivariate analyses. De Cáceres & Legendre (2008) studied the statistical and ecological bases underlying the Beals smoothing function and explored the factors that may affect the reliability of the transformed values using simulated data. They showed that Beals predictions are only reliable for target species that are closely related to the overall ecological structure displayed by the data set. They developed a statistical test to determine when the observed presence/absence values can be replaced with Beals smoothing predictions.

7.8 Software

Only the largest general-purpose commercial statistical packages, such as SAS, SPSS, SYSTAT, JMP, and STATISTICA, offer clustering among their methods for data analysis (Section 8.15), and functions to compute some resemblance coefficients. The smaller commercial packages offer no such facilities. Among the Q-mode coefficients found in the larger packages, one always finds the Euclidean distance. The squared Euclidean, Manhattan, Chebychev* and Minkowski distances may also be found, as well as the simple matching coefficient for multistate nominal data (eq. 7.19). For R-mode analyses, one finds Pearson's r in most packages, or related measures such as the cosine of the angle between variables, dot product, or covariance. Nonparametric correlation coefficients, as well as chi-square, uncertainty and contingency coefficients may also be found. In addition, for Q-mode analysis, SYSTAT offers several binary coefficients and some coefficients for quantitative data (Bray-Curtis, Kulczynski).

Packages written for ecological or taxonomic analysis emphasize resemblance coefficients and clustering methods. They are: NTSYSPC†, developed by F. J. Rohlf, originally for numerical taxonomy studies; CLUSTAN‡, developed by D. Wishart;

---

* In R, the Chebychev distance, $D_{\text{Chebychev}}(\mathbf{x}_1, \mathbf{x}_2) = \max_j |x_{1j} - x_{2j}|$, is computed by function ***dist()*** with method = "maximum". $D_{\text{Chebychev}}$ is a metric. This distance function does not seem to have been used in community ecology. It is described here because it is found in computer packages and in an R function, hence readers may wonder what its equation is.

† NTSYSPC is available from Exter Software Inc., 47 Route 25A, Suite 2, Setauket, New York 11733-2870, USA; http://www.exetersoftware.com.

‡ The CLUSTAN package may be ordered from CLUSTAN Limited, 16 Kingsburgh Road, Edinburgh EH12 6DZ, Scotland. See also the Web page http://www.clustan.com/.

PATN[*], developed by L. Belbin; PC-ORD[†] written under the direction of B. McCune; and SYN-TAX 2000[**] written by J. Podani. In the R language,

1. The most inclusive functions to compute distances are: ***dist()*** in STATS, ***vegdist()*** in VEGAN, ***dist.binary()*** in ADE4, and ***daisy()*** in CLUSTER. In addition, ***gowdis()*** in FD offers a complete set of options to compute the Gower distance ($1 - S_{15}$). Function ***mahalanobis()*** in STATS computes Mahalanobis distances between the objects in a data table and a vector, which can be the multivariate mean vector of the same data table. Function ***raupcrick()*** in VEGAN computes the Raup-Crick distance ($1 - S_{27}$). Function ***is.euclid()*** of ADE4 checks the Euclidean nature of distance matrices; see Tables 7.2 and 7.3.

2. In the R mode, package STATS offers functions ***var()*** and ***cov()*** to compute covariance matrices and ***cor()*** to compute correlation matrices. ***cor.test()*** is used to test the significance of correlation coefficients. The Pearson, Spearman, and Kendall correlation coefficients are available as options in both ***cor()*** and ***cor.test()***. ***chisq.test()*** of STATS provides chi-square tests of significance. ***pf()*** of STATS computes the parametric p-value associated with F-statistics.

3. Transformations for community composition data described in Section 7.7 are available in the VEGAN function ***decostand()***. Multivariate homogeneity of variances is tested by VEGAN's function ***betadisper()***. Beals smoothing is available in the VEGAN function ***beals()***. The test of significance to determine when species presence/absence values can be replaced with Beals smoothing predictions is conducted by function ***BSS.test()***[‡].

[*] PATN is available from Blatant Fabrications Pty Ltd, Carlton, Tasmania, Australia. Technical information is available on the Web page http://www.patn.com.au, or from Lee Belbin at <lee@blatantfabrications.com>.

[†] Availability: see Section 11.7 (footnote).

[‡] R code and documentation file available with the Beals smoothing function on the Web page http://sites.google.com/site/miqueldecaceres/.