STATISTICAL ECOLOGY

A PRIMER ON METHODS AND COMPUTING

John A. Ludwig
CSIRO Division of Wildlife and Ecology
Deniliquin, NSW, Australia

James F. Reynolds
San Diego State University
San Diego, California

A WILEY-INTERSCIENCE PUBLICATION
JOHN WILEY & SONS
NEW YORK • CHICHESTER • BRISBANE • TORONTO • SINGAPORE

Library of Congress Cataloging in Publication Data:

Ludwig, John A.
Statistical ecology: a primer on methods and computing / John A. Ludwig, James F. Reynolds.
p. cm.
"A Wiley-Interscience publication."
Bibliography: p.
Includes index.
ISBN 0-471-83235-9
1. Ecology—Statistical methods. I. Reynolds, James F., 1946– II. Title.
QH541.15.S72L83 1988
574.5'24'015195—dc 19 87-26348
CIP

Printed in the United States of America

10 9 8 7 6 5

CHAPTER 9

Background

Ecological communities are composed of a number of coexisting species. Some communities may have a large number of species (e.g., a tropical forest); others may have just a few (e.g., a polluted river). In Chapters 7 and 8 we described some empirical models for quantifying the relationships between the total number of species in a community and some measure of their abundances (e.g., total numbers). In this part of the book, we are interested in examining the affinities of coexisting species. How do coexisting species utilize common resources?

Consider, for example, a species-rich lake that has four dominant fish, all about the same size. Are they in direct competition for food and space? Do some species feed exclusively in the surface waters, while others feed on the lake bottom? When we spatially locate species *A*, are we likely more often than not to find species *B* there as well? In a broad sense, we can define such interspecific interactions as the degree of affinity between species.

One measure of affinity is the degree to which species *overlap* in their utilization of common resources. This overlap is defined in terms of various portions of the species niche that is shared by other species. Niche studies are based on such species attributes as diet, microhabitat preference, and timing of activities (e.g., foraging). Measures of niche overlap are presented in Chapter 10.

In Chapter 11 we cover the topic of interspecific association. In this instance, we are concerned only with measuring how often two species are found together in the same location. This affinity (or lack of it) for coexistence is tested by examining if the occurrence of the species [in a series of sample units (SUs)] is greater than or less than what would be expected if they were

independent. If either positive or negative association is detected, we can measure the strength of this association with indices.

Association is based solely on presence/absence data. If a sample contains quantitative measures of species abundance, we can determine the covariation in abundances between species. This may lead to questions concerning species affinities. For example, if the abundance of one species always decreases when the other species increases, is there some type of causal negative interaction? Measures of interspecific covariation are presented in Chapter 12.

Each of these approaches is intended to help the ecologist detect patterns in species interactions. Of course, nothing about the underlying causes of a pattern can be inferred simply from its detection, although we hope that pattern detection will lead directly to testable hypotheses.

9.1 MATRIX VIEW

Cattell (1952) noted that the ecological data matrix could be studied from two distinct viewpoints: (1) down columns (the SUs) or (2) across rows (the species). Depending on which of these options is chosen, certain measures of *resemblance* are available. A taxonomy of these resemblance functions is given in Figure 9.1. It is important to recognize that the appropriate choice of

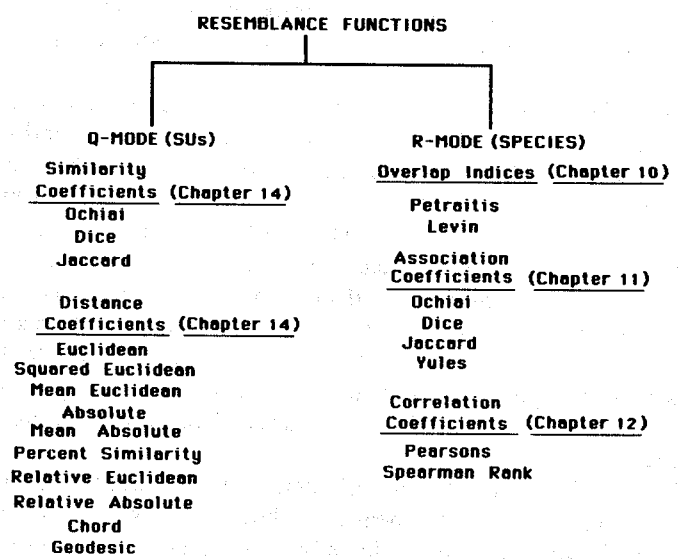

Figure 9.1 *Resemblance functions applicable to Q-mode analysis (similarity, distance) and R-mode analysis (overlap, association, correlation).*

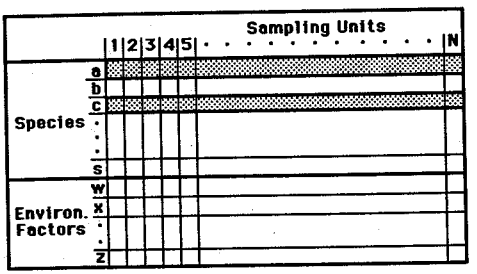

Figure 9.2 *The shaded area indicates the form of the ecological data matrix for measuring species affinity. Interest is in the occurrences or abundances of species across sampling units.*

a function is related to the fact that, in the ecological data matrix, we consider the species (rows) to be *dependent* on one another, whereas the SUs (columns) are *independent* samples (Legendre and Legendre 1983).

From our background discussion, the student will recognize that our interest in Chapters 10–12 is in species affinity, that is, measuring pairwise species resemblance based on data across rows in the ecological data matrix (Figure 9.2). Ecologists refer to this as *R*-mode analysis (Legendre and Legendre 1983).

The *R*-mode resemblance functions (Figure 9.1) are divided into overlap indices (Chapter 10), association coefficients (Chapter 11), and covariation coefficients (Chapter 12). These *R*-mode indices measure the dependence or intensity of the affinity between species. Intuitively, this makes sense because we are measuring the resemblance of species that occur together in SUs.

Q-mode resemblance functions measure the *similarity* or *dissimilarity* between SUs in terms of their species composition (i.e., down columns). Again, this terminology makes sense since we are comparing how similar or dissimilar *independent* samples are. Parts V through VII of this book are largely concerned with *Q*-mode analyses. *Q*-mode resemblance functions are presented in Chapter 14. A graphical representation of *R*- and *Q*-mode approaches is given in Figure 13.2.

Whereas the scheme presented in Figure 9.1 seems straightforward, there are many cases in the ecological literature where *R*- and *Q*-mode indices have been interchanged. For example, the association coefficients we describe in Chapter 11 have also been used to measure similarity between SUs (Chapter 14). Because of the nature of some of these particular coefficients, such usage is acceptable; however, using an *R*-mode coefficient like the correlation coefficient to measure *Q*-mode resemblance is not recommended (Orloci 1972, 1978).

association.) Depending on the size and shape of the SU, it is possible to influence the outcome of association. This dependence can be lessened if the selection of the SU is made relative to the size, shape, and spatial distribution of the species under study. The SU must be large enough to potentially include at least one individual of each species and yet not so large that one of these species is included in every SU (Greig-Smith 1983).

11.2.1 Test of Association (Two-Species Case)

STEP 1. DATA SUMMARY. For each pair of species, A and B, we obtain the following:

a = the number of SUs where both species occur
b = the number of SUs where species A occurs, but not B
c = the number of SUs where species B occurs, but not A
d = the number of SUs where neither A nor B are found
N = the total number of SUs ($N = a + b + c + d$)

This information is conveniently summarized in the form of a 2 × 2 table (Figure 11.1). Both the test and measures of association presented below are based on these data.

The expected frequency of occurrence of species A in the SUs, which we will represent as $f(A)$, is given by

$$f(A) = \frac{a + b}{N} \tag{11.1a}$$

and, for species B, by

$$f(B) = \frac{a + c}{N} \tag{11.1b}$$

We assume that both species have occurred in at least one SU in the collection, that is, $f(A)$ and $f(B)$ are greater than 0.

		Species B: present	Species B: absent	
Species A	present	a	b	m=a+b
	absent	c	d	n=c+d
		r=a+c	s=b+d	N=a+b+c+d

Figure 11.1 *2 × 2 contingency or species association table.*

STEP 2. STATE HYPOTHESIS. The null hypothesis is that the species are independent (i.e., there is no association).

STEP 3. COMPUTE TEST STATISTIC. The 2 × 2 table contains *observed* values for each of the cells (a, b, c, and d) from the sample of size N. To test for association, we compute what the *expected* values for each cell would be if the occurrences of species A and B are, in fact, independent and compare them to the observed values. A chi-square test statistic can be used to test the null hypothesis of independence in the 2 × 2 table. The chi-square test statistic is computed as

$$\chi_t^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \tag{11.2}$$

which is a summation over the four cells of the 2 × 2 table.

The expected value for cell a is given by

$$E(a) = \frac{(a+b)(a+c)}{N} = \frac{rm}{N} \tag{11.3}$$

or, from Eq. (11.1),

$$E(a) = f(B)(a+b) = f(A)(a+c) \tag{11.4}$$

In words, Eq. (11.4) states that of the total number of SUs where species A was present (i.e., $a + b$), we expect that if A and B were independent, species B should also be present in proportion to its overall frequency in the SUs, that is, $f(B)$; and vice versa for species A's presence in SUs where species B is present.

Similarly, the expected values for cells b, c, and d are, respectively,

$$E(b) = \frac{ms}{N}, \quad E(c) = \frac{rn}{N}, \quad \text{and} \quad E(d) = \frac{sn}{N} \tag{11.5}$$

The chi-square test statistic [Eq. (11.2)] is now given as

$$\chi_t^2 = \frac{[a - E(a)]^2}{E(a)} + \cdots + \frac{[d - E(d)]^2}{E(d)} \tag{11.6}$$

A mathematically equivalent, but certainly simpler equation, which may be used instead of Eq. (11.6), is

$$\chi_t^2 = \frac{N(ad - bc)^2}{mnrs} \tag{11.7}$$

Actually, in addition to being simpler to use, since Eq. (11.7) does not require the computation of expected values, nor the differences between observed and expected values, the associated rounding errors are avoided.

The significance of the chi-square test statistic is determined by comparing it to the theoretical chi-square distribution. The 2 × 2 contingency table has one degree of freedom, since a contingency table with r rows and c columns has $(r - 1)$ times $(c - 1)$ degrees of freedom (Zar 1974). The theoretical chi-square value for 1 df at the 5% probability level is 3.84. If $\chi_t^2 > 3.84$, we reject the null hypothesis that the co-occurrence of species A and B is independent and conclude that they are associated.

There are two types of associations:

1. Positive—if observed $a > E(a)$, that is, the pair of species occurred together more often than expected if independent.
2. Negative—if observed $a < E(a)$, that is, the pair of species occurred together less often than expected if independent.

This comparison of observed a to $E(a)$, that is,

$$a - E(a) = (ad - bc)/N \tag{11.8}$$

results in the quantity $ad - bc$ appearing in the numerator of all χ^2-like formulations, such as Eq. (11.7).

If any cell in the 2 × 2 table has an expected frequency < 1 or if more than two of the table cells have expected frequencies < 5, then the resulting chi-square test statistic will be biased (Zar 1974). A corrected chi-square is used to avoid biased values resulting from low cell expectations. In such cases, a continuity correction is applied to ensure a closer approximation to the theoretical, continuous chi-square distribution. This is achieved by using Yates's correction formula:

$$\chi_t^2 = \frac{N[|(ad) - (bc)| - (N/2)]^2}{mnrs} \tag{11.9}$$

11.2.2 Measures of Association (Two-Species Case)

Hubalek (1982) reviewed the properties of 43 indices that have been used to measure the degree of association between pairs of species. To sort through this plethora of indices, Hubalek identified five "admission" conditions. Indices

that failed to satisfy any one of these conditions were deemed inadmissible and dropped from further consideration. The remaining admissible indices were then compared against eight optional criteria in order to help select the best association indices. Janson and Vegelius (1981) conducted a similar study where the characteristics of 20 association indices were examined over six "admission" conditions. The details of all these admission conditions are beyond the scope of our presentation, but five important conditions are listed here.

CONDITION 1. Each association index should reach its minimum value at $a = 0$, that is, when the two species are never found together.

CONDITION 2. The maximum value of the index should be when both species always occur together, that is, when $b = c = 0$.

CONDITION 3. The association index should be symmetric, that is, the value of the index should be the same regardless of which species is designated "A" or "B" (Figure 11.1).

CONDITION 4. The index should be able to discriminate between positive and negative associations. Formally, this means that the value of the index when $a > E(a)$ is always greater than when $a < E(a)$.

CONDITION 5. The index should be independent of d, that is, the number of joint absences. There has been much debate as to whether the joint absence of species has any ecological meaning (Clifford and Stephenson 1975, Goodall 1978b, Sneath and Sokal 1973). We agree with Hubalek (1982) that indices using values of d are limited in ecology. For example, in a study of leaf miners on oak leaves by Bultman and Faeth (1985), average values for their 2×2 tables used to test for association were $a = 24$, $b = 875$, $c = 1{,}140$, and $d = 134{,}650$! Any index using d would be "swamped" by the magnitude of d (joint absences).

Hubalek (1982) found six association measures to satisfy his admission conditions, and Janson and Vegelius (1981) found three that generally performed well. Three measures recommended by both studies—the Ochiai, Dice, and Jaccard indices—are presented below. These indices are equal to 0 at "no association" and 1 at "maximum association." The Ochiai and Dice measures are means of the ratios a/m and a/r, that is, the number of joint occurrences of the two species compared to the total occurrences of species A and B, respectively (Figure 11.1).

OCHIAI INDEX (OI). The Ochiai (1957) index is based on the geometric mean of a/m and a/r, that is,

$$\text{OI} = \frac{a}{\sqrt{a+b}\sqrt{a+c}} \tag{11.10}$$

DICE INDEX (DI). The Dice (1945) index is based on the harmonic mean of a/m and a/r, that is,

$$\text{DI} = \frac{2a}{2a+b+c} \tag{11.11}$$

JACCARD INDEX (JI). This index is the proportion of the number of SUs where both species occur to the total number of SUs where at least one of the species is found:

$$\text{JI} = \frac{a}{a+b+c} \tag{11.12}$$

To determine the sampling properties of a number of association measures, Goodall (1973) took repeated samples from a population with known species frequencies (a, b, c, and d) and computed the mean and variance of each index. Jaccard's index was found to be generally unbiased, even at small ($N = 10$) sample sizes. The Dice index tended to underestimate the true population values at small samples, but performed well at $N = 20$. Goodall did not test the Ochiai index.

11.2.3 Interspecific Association (Multiple-Species Case)

Usually, the association of more than a single pair of species is of interest; we may be interested in from 5, to perhaps 50 or more species. The number of all possible pairwise species associations or combinations that may be computed increases rapidly according to the equation $S(S-1)/2$, where S is the number of species. For example, with five species there are $5(4)/2 = 10$ combinations; for 10 species, there are $10(9)/2 = 45$. Obviously, there is the problem of representing all the pairwise association index values in such a way as to ease interpretation. There are two ways of diagramming these multiple-species associations.

DIAGRAM 1. SPECIES ASSOCIATION COMPARISON MATRIX. All possible pair combinations of species associations can be displayed in a matrix of the form shown in Figure 11.2. To aid in interpretation, the species positions in the matrix can be reordered in such a way as to place species with highly significant positive index values along the diagonal of the matrix.

CHAPTER 13

Background

Given a set of objects and some measure of their resemblance to each other, we can define classification as the grouping or clustering of these objects based on their resemblance. Classification plays a fundamental role in many areas of science in the search for what Sokal (1974) terms the "natural" system. A natural system might be viewed as a reflection of those various processes that have led to the observed arrangement of the objects. For example, in ecology this "natural" system could be the end result of evolutionary processes.

The first step in the classification of ecological communities involves sampling. Artificial or natural sampling units (SUs) are used, and various types of data, both qualitative and quantitative, are obtained. These data may include lists of species present or some indication of their abundance (density, frequency, cover, biomass). Next, some measure of ecological resemblance between all pairs of SUs is computed in order to quantify their similarity or dissimilarity (what we term a Q-mode analysis, see Chapter 9). Finally, the SUs (objects) are grouped according to their resemblances; SUs in each group should have a number of common characteristics that set them apart from the SUs of other such groups. The objective is to demonstrate the relationships of the SUs to each other and, it is hoped, to simplify these relationships in order to be able to make general statements about the classes of objects that exist.

A study by Able and Noon (1976) is a good example of a potential classification problem. Their objective was to describe avian community structure along an elevational gradient in the Adirondack Mountains of New York. They found 44 species of birds distributed along a census belt transect that ranged from 400 to 1400 m elevation. Some species of birds were found in all

SUs along the gradient; others were found only within narrow limits of elevation. At several locations (ecotones) along the gradient, major changes in vegetation structure occurred, resulting in some natural upper and lower distribution limits for some of the bird species. A classification of the SUs in their study, based on species-abundance data, would reflect both the continuous and discontinuous distributions of all 44 bird species. The "natural" system of interest here, in a broad sense, is the grouping of the SUs that reflects the abundances (and amplitudes of abundances) of the major bird species along the elevational gradient.

Of course, homogeneous communities are not amenable to classification. Also, since some of the techniques presented in Chapters 15 and 16 will "classify" even a random data set, we have to be careful in our interpretation of these results. Simply because it may be possible to classify a data set, a spurious classification will not yield a meaningful ecological interpretation. Various philosophies of classification theory have been reviewed by Goodall (1970), Ratliff and Pieper (1982), Sokal (1974), and Whittaker (1978a, b); we highly recommend that students seriously interested in classification read these papers.

In the early days of ecology, classification of communities was largely intuitive, based on subjective decisions and qualitative descriptions. More recent trends have been toward objective methods of classification based on quantitative data. In the next few chapters, we will examine some of these objective methods. It is inevitable, however, that a degree of subjectivity remains in all classification studies; whereas a given classification method will yield unique results, an alternative method may yield different results and, consequently, subjective decisions must be made.

Some terminology used in the following chapters is briefly defined below:

1. Classifications may be either *hierarchical* or *reticulate*. As the name implies, in a hierarchical classification, groups at any lower level of a classification are exclusive subgroups of those groups at higher levels. In a reticulate classification (which we will not consider), groups are defined separately and, rather than hierarchically ordered, are linked together in a weblike network.

2. Classifications may be either *divisive* or *agglomerative*. In a divisive classification, the entire collection of SUs is divided and redivided, based on SU similarities, to arrive at the final groupings (i.e., picture an inverted tree). In an agglomerative classification, as its name implies, individual SUs are combined and recombined successively to form larger groups of SUs (the tree).

3. Classifications may be either *monothetic* or *polythetic*. In a monothetic classification the similarity of any two SUs or groups is based on the value of

a single variable, for example, the presence or absence of a single species. In a polythetic classification the similarity of any two SUs or groups is based on their overall similarity as measured by numerous variables, for example, species abundances.

In the subsequent treatment of classification methodologies, we make use of both qualitative data (e.g., presence–absence or two-state character data) and quantitative data (e.g., abundance or ordered multistate character data). In Chapter 14 indices of ecological resemblance pertaining to *Q*-mode analyses will be presented. In Chapter 15, the technique of normal association analysis is described; this is a monothetic, divisive classification model based on the presence–absence of species in SUs. In Chapter 16 a polythetic, agglomerative classification technique generally known as *cluster analysis* is described; this method is based on quantitative abundance data of species in SUs.

13.1 MATRIX VIEW

In studies of species affinity (Chapters 10, 11, and 12), the ecological data matrix was viewed across rows (Figure 9.2), that is, a *R*-mode analysis. In classification studies, the ecological data matrix may either be viewed across rows (Chapter 15) or down columns, that is, a *Q*-mode analysis (Figure 13.1). In either case, *the objective is the same: to classify SUs.*

Another way to view the relationship between *R*-mode (species) and *Q*-mode (SU) analysis is to conceptualize these relationships in a geometric way; this has lead to the term *hyperspace* (Williams and Dale 1965) when referring to *R*- and *Q*-mode studies. *Species hyperspace* is conceptualized as being *S*-dimensional, that is, one dimension for each species in the sample of *S* species. (Obviously, it is impossible for us to diagram the *S*-space beyond

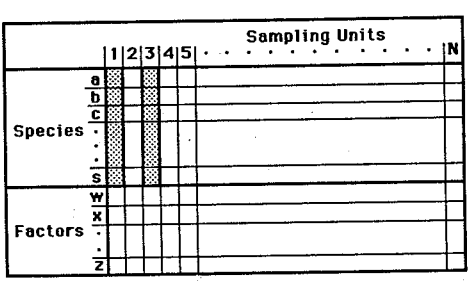

Figure 13.1 *The shaded area indicates the form of the ecological data matrix for measuring Q-mode resemblance. Interest is in pairwise SU similarities.*

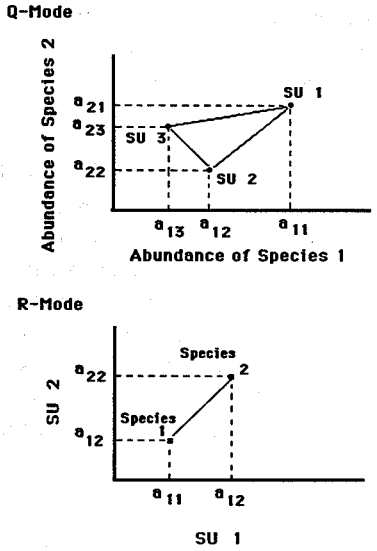

Figure 13.2 *Q- and R-mode analyses viewed geometrically. Q-mode is the representation of SUs in species space and R-mode is the representation of species in SU space. Note that $a_{i,j}$ is the abundance of the* i*th species in the* j*th SU. Adapted from Legendre and Legendre (1983).*

$S = 3$.) SUs are then positioned within this S-space based on the relative abundance of each species in a SU. The distance between the SUs in this S-space represents their similarity (or, alternatively, their dissimilarity; Chapter 14) to one another. An example is given in Figure 13.2, showing the location of three SUs (Q-mode) in the space of two species.

SU-hyperspace, on the other hand, is conceptualized as being N-dimensional, one dimension for each of the N SUs in the sample. The species are then positioned within this N-space in relation to their abundances; the closer two species are within this space, the more similar are their respective abundances in the SUs. In Figure 13.2, the position of two species (R-mode) in two SU-space is diagrammed. This type of spatial representation is much more artificial, since the values on the SU axes are the abundances of the species in the SUs (Legendre and Legendre 1983).

Finally, recall the distinction between a SU and a sample. A sample consists of a *collection* of SUs (examples given in Table 1.1). Because of the tremendous diversity of ecological communities, the student should be aware that the columns of the ecological data matrix may represent either individual SUs or samples. This is best illustrated by some examples from the literature.

In a study of benthic communities in Moreton Bay, Queensland, Australia,

Stephenson et al. (1970) conducted extensive dredging of the aquatic substrate. Their SU was a dredge that had a mouth 84 × 29 cm, a 2.5 m bag with 75 cm of mesh, and cutting edges inclined at 25°. These specifications are important because variations in any of these parameters would, most likely, affect the collection of bottom-dwelling animals. The speed of their boat was maintained at 0.6 km/hr and dredging was done for 2 minutes. The dredge catches were sorted to determine the macrobenthos species present. They collected 355 species in 400 dredge stations or locations throughout the bay. Their ecological data matrix of presence–absence data was 355 rows (species) by 400 columns (dredges or SUs), and both *R*-mode and *Q*-mode analyses were performed on these data.

In the next example, the columns of the data matrix represent samples rather than individual SUs. Huhta (1979) examined the changes in composition of soil arthropod communities in undisturbed and clear-cut forests north of Helsinki, Finland, from 1962 through 1965 and again in 1968. In each forest, samples were taken bimonthly. A sample consisted of four 25 × 25 cm quadrats of soil and litter, which were taken to the laboratory, and the total number of arthropod species and their relative abundances determined. These bimonthly data were combined into yearly samples for analysis. Thus, the data matrix in Huhta's study consisted of rows as soil arthropod species and columns representing the yearly samples (5 years); the entries in the matrix were number of individuals. Huhta proceeded to conduct a *Q*-mode analysis of these data to ascertain how the years differed in terms of the composition of the arthropod communities.

TABLE 13.1 *Selected literature for examples of community classification studies using either association analysis (AA) or cluster analysis (CA)*

Location	Community	Method	Reference
England	Chalk grassland	AA	Gittins 1965
NWT, Canada	Benthos	AA	Vilks et al. 1970
Virginia	Deciduous forest	AA	Madgwick and Desrochers 1972
Nigeria	Savanna	AA	Kershaw 1973
Australia	Forest–woodland	AA	Ashton 1976
Arizona	Desert grassland	AA	Fish 1976
North Sea	Marine benthos	CA	Stephenson et al. 1972
Australia	Rain forest	CA	Dale and Clifford 1976
Atlantic	Marine algae	CA	Lawson 1978
Puerto Rico	Rain forest	CA	Crow and Grigal 1979
New York	Deciduous forest	CA	Gauch and Stone 1979
England	Peat-bog	CA	Clymo 1980
NWT, Canada	Arctic tundra	CA	Thompson 1980
Australia	Rangelands	CA	Foran, et al. 1986

CHAPTER 14

Resemblance Functions

The ecologist is often faced with the task of making comparisons of plant and/or animal samples when addressing questions of community structure. These samples may be (1) obtained over various locations in the landscape, such as Able and Noon's (1976) study of bird distributions along an elevational gradient, or (2) obtained from the same location but at differing times, such as Livingston's (1976) comparisons of December and June fish catch data. In this chapter we describe some resemblance functions that quantify the similarity or dissimilarity between samples. The more similar samples are in species composition and quantity, the greater their resemblance, that is, the closer their ecological distance.

14.1 GENERAL APPROACH

Resemblance functions, as broadly defined by Sneath and Sokal (1973), quantify the similarity or dissimilarity between two objects based on observations over a set of descriptors. The objects of interest to the ecologist are SUs (sampling units or samples, see Chapter 13) and the descriptors are measures of species abundance (e.g., density, biomass). Thus, as defined, these resemblance functions involve a *Q*-mode analysis, that is, between SUs.

The distinction between *Q*-mode and *R*-mode analysis was made in Chapter 13 and the various resemblance functions used in these different modes was illustrated in Figure 9.1. In general, two types of *Q*-mode resemblance functions are distinguished: (1) similarity coefficients and (2) distance coefficients. Similarity coefficients vary from a minimum of 0 (when a pair of SUs

are completely different) to 1 (when the SUs are identical). On the other hand, distance coefficients are the opposite; they assume a minimum value of 0 when a pair of SUs are identical and have some maximum value (in some cases infinity) when the pair of SUs are completely different. Hence, distance coefficients are also referred to as *dissimilarity coefficients*. In fact, a similarity index may always be represented as a distance, even if just by a simple transformation such as 1 − similarity (Legendre and Legendre 1983). Thus, distance may be thought of as the complement of similarity (Sneath and Sokal 1973).

Needless to say, the number of resemblance functions is large. In this chapter we limit our treatment to some of the more common similarity and distance measures used in *Q*-mode studies. However, this does not imply that some of the statistical and probability indices proposed to measure *Q*-mode resemblance (Legendre and Legendre 1983) might not be equally good, or, perhaps, even better with certain data sets (see Section 14.7). Distance coefficients are perhaps the most popular among community ecologists and, in our view, are the most straightforward in concept and application to community data.

14.2 PROCEDURES

14.2.1 Similarity Coefficients

Similarity coefficients are, by far, the most prolific indices in the ecological literature (Legendre and Legendre 1983). These indices are based solely on presence (indicated with a 1) or absence (indicated with a 0) data. Consider, for example, the presence–absence of 3 species in 3 SUs:

	SU		
Species	(1)	(2)	(3)
A	1	1	0
B	1	1	0
C	1	0	1

In a *Q*-mode analysis, we are interested in the degree of similarity in species composition *between* each pair of SUs (columns of the data matrix). The more species two SUs share relative to their total species complements, the greater their ecological similarity. In this example, SU(2) contains two of the three species also found in SU(1), but has no species in common with SU(3).

Recall that in Section 11.2.2 we presented three indices (Ochiai, Dice, and Jaccard) based on presence–absence data that were used to measure the

degree of association between species (an *R*-mode analysis, i.e., across the rows of the data matrix). These same three indices can also be used to compute a *Q*-mode similarity between SUs. The student should note that these are the *only types* of functions that we use to measure *both* *Q*-mode (sample similarity) and *R*-mode (species association) resemblance. The similarity between SU(1) and SU(2) in the above example, as measured by the Ochiai index [OI, Eq. (11.10)], Dice index [DI, Eq. (11.11)], and Jaccard index [JI, Eq. (11.12)] are:

$$\mathrm{OI}_{1,2} = \frac{2}{\sqrt{2}\sqrt{3}} = 0.82$$

$$\mathrm{DI}_{1,2} = \frac{4}{4+0+1} = 0.80$$

$$\mathrm{JI}_{1,2} = \frac{2}{2+0+1} = 0.67$$

Since we covered the use of these indices in Chapter 11, they will not be presented again in this chapter.

14.2.2 Distance Coefficients

Three groups of distance measures are distinguished below: (1) *E*-group (the Euclidean distance coefficients), (2) BC-group (the Bray–Curtis dissimilarity index), and (3) RE-group (the relative Euclidean distance measures).

The following matrix notation is used in the equations presented below: X_{ij} represents the abundance of the ith species in the jth SU. For example, $X_{4,3}$ would be the abundance of the 4th species in the 3rd SU. As before, the community data matrix is composed of S species and N SUs.

14.2.2.1 E-Group Distances

DISTANCE 1. EUCLIDEAN DISTANCE (ED). This measure is the familiar equation for calculating the distance between two points SU_j and SU_k in Euclidean space:

$$\mathrm{ED}_{jk} = \sqrt{\sum_{i=1}^{S} (X_{ij} - X_{ik})^2} \tag{14.1}$$

ED emphasizes the larger differences in abundances of species between SUs, since each species difference is squared and then summed. The final distance

value is scaled down by taking the square root of the sum. The value of ED ranges from zero to infinity, as do all of the *E*-group measures.

DISTANCE 2. SQUARED EUCLIDEAN DISTANCE (SED). This measure is simply the square of ED:

$$\mathrm{SED}_{jk} = \sum_{i=1}^{S} (X_{ij} - X_{ik})^2 \tag{14.2}$$

DISTANCE 3. MEAN EUCLIDEAN DISTANCE (MED). MED is similar to ED, but the final distance is on a smaller scale since the mean difference is used:

$$\mathrm{MED}_{jk} = \sqrt{\frac{\sum_{i=1}^{S} (X_{ij} - X_{ik})^2}{S}} \tag{14.3}$$

DISTANCE 4. ABSOLUTE DISTANCE (AD). This measure is the sum of the absolute abundance differences taken over the *S* species:

$$\mathrm{AD}_{jk} = \sum_{i=1}^{S} |X_{ij} - X_{ik}| \tag{14.4}$$

AD places less emphasis on larger differences than the previous three measures since differences in abundance are summed, but not squared. Thus, smaller differences are given relatively greater weight in the final distance. This distance measure is known as character difference in numerical taxonomy (Sneath and Sokal 1973).

DISTANCE 5. MEAN ABSOLUTE DISTANCE (MAD). The MAD is similar to AD but a mean distance is used rather than an absolute distance:

$$\mathrm{MAD}_{jk} = \frac{\sum_{i=1}^{S} |X_{ij} - X_{ik}|}{S} \tag{14.5}$$

MAD is equivalent to mean character difference used in numerical taxonomy (Sneath and Sokal 1973).

14.2.2.2 BC-Group Distances. This group is represented by a single index first introduced into the ecological literature by Bray and Curtis (1957). This index remains very popular among ecologists. The first step is to compute the percent similarity (PS) between SUs *j* and *k* as

$$\mathrm{PS}_{jk} = \left(\frac{2W}{A + B}\right)(100) \tag{14.6a}$$

where

$$W = \sum_{i=1}^{S} [\min(X_{ij}, X_{ik})]$$

$$A = \sum_{i=1}^{S} X_{ij} \quad \text{and} \quad B = \sum_{i=1}^{S} X_{ik}$$

Thus, PS between the jth SU and the kth SU is a numerator of twice the sum of the minimum (min) of the paired observations X_{ij} and X_{ik} (the "shared" species abundance between each pair of SUs) divided by a denominator of the total of all the species abundances for the two SUs. For any pair of SUs with identical species abundances, their similarity is complete, that is, PS = 100%.

The distance complement of PS is percent dissimilarity (PD), computed as

$$\text{PD} = 100 - \text{PS} \tag{14.6b}$$

PD may also be computed on a 0–1 scale as

$$\text{PD} = 1 - [2W/(A + B)] \tag{14.6c}$$

which is useful since it is more in line with the range of values assumed by many of the other distance indices. We will use PD as computed in Eq. (14.6c) in our calculations below.

14.2.2.3 RE-Group Distances. This group contains distance indices that are expressed on standardized or relative scales.

DISTANCE 7. RELATIVE EUCLIDEAN DISTANCE (RED). This measure incorporates species abundance totals within each SU so that the final distance measure is standardized relative to differences in *total* SU abundances:

$$\text{RED}_{jk} = \sqrt{\sum_{i=1}^{S} \left[\left(\frac{X_{ij}}{\sum_i^S X_{ij}} \right) - \left(\frac{X_{ik}}{\sum_i^S X_{ik}} \right) \right]^2} \tag{14.7}$$

This equation is derived by applying Whittaker's (1952) relative transformation for absolute distance [Eq. (14.8)] to Euclidean distance as suggested by Orloci (1978). RED ranges from 0 to $\sqrt{2}$.

DISTANCE 8. RELATIVE ABSOLUTE DISTANCE (RAD). This measure applies Whittaker's (1952) relative abundance correction to AD (in the same sense that relative Euclidean distance "corrects" Euclidean distance):

$$\text{RAD}_{jk} = \sum_{i=1}^{S} \left| \left(\frac{X_{ij}}{\sum_i^S X_{ij}} \right) - \left(\frac{X_{ik}}{\sum_i^S X_{ik}} \right) \right| \tag{14.8}$$

RAD has a range from 0 to 2.

DISTANCE 9. CHORD DISTANCE (CRD). This measure puts greater importance on the relative proportions of species in SUs and correspondingly less importance on their absolute quantities. Technically, this is done by projecting the SUs onto a circle of unit radius through the use of direction cosines. The measure is then the *chord* distance between the two SUs after such a projection. We refer the student to Pielou (1984, p. 48) for a geometric illustration. Chord distance is given by

$$\text{CRD}_{jk} = \sqrt{2(1 - \text{ccos}_{jk})} \tag{14.9a}$$

where the chord cosine (ccos) is computed from

$$\text{ccos}_{jk} = \frac{\sum_{i=1}^{S} (X_{ij} X_{ik})}{\sqrt{\sum_i^S X_{ij}^2 \sum_i^S X_{ik}^2}} \tag{14.9b}$$

Note that, in the case of presence–absence data, this ccos is identical to Ochiai's coefficient. CRD, like RED, ranges from 0 to $\sqrt{2}$.

DISTANCE 10. GEODESIC DISTANCE (GDD). This measure is the distance along the *arc* of the unit circle (rather than the chord distance) after projection of the SUs onto a circle of unit radius:

$$\text{GDD}_{jk} = \arccos(\text{ccos}_{jk}) \tag{14.10}$$

GDD has a range from 0 to $\pi/2$ (i.e., 0 to 1.57).

To summarize the distances computed between all possible pairs of SUs based on any of the similarity or distance measures described previously, it is convenient to create a SU × SU matrix of distance (or similarity) values. Examination of this matrix quickly reveals the distance between any two SUs of interest. It is on this distance matrix that the clustering strategies of community classification operate (Chapter 16). We give an example of a distance matrix in Section 14.6.

14.3 EXAMPLE: CALCULATIONS

To illustrate the computations for the distance measures, the data in Table 14.1 will be utilized. From this simple data matrix of abundances for three

TABLE 14.1 Community data matrix composed of three SUs with abundance data for three species (Spp)

	SUs		
Spp	(1)	(2)	(3)
(1)	20	15	0
(2)	10	0	6
(3)	17	0	0

TABLE 14.2 Differences (DIF), sums (SUM), and sums of squares (SSQ) needed for computing the distance between SUs 1 and 3

	SU		DIF	DIF^2	SUM
Spp	(1)	(3)	(1 − 3)	$(1-3)^2$	(1 + 3)
(1)	20	0	20	400	20
(2)	10	6	4	16	16
(3)	17	0	17	289	17
SUM =	47	6	41	705	
SSQ =	789	36			

species in three SUs, the computations for the distances between SUs 1 and 3 will be illustrated. For the computations, the differences, sums, and sums of squares within and between the three species in SUs 1 and 3 are needed (Table 14.2).

DISTANCE 1. EUCLIDEAN DISTANCE [Eq. (14.1)]:

$$ED_{1,3} = \sqrt{[(20-0)^2 + (10-6)^2 + (17-0)^2]}$$
$$= \sqrt{(400 + 16 + 289)} = \sqrt{705} = 26.6$$

DISTANCE 2. SQUARED EUCLIDEAN DISTANCE [Eq. (14.2)]:

$$SED_{1,3} = (400 + 16 + 289) = 705$$

DISTANCE 3. MEAN EUCLIDEAN DISTANCE [Eq. (14.3)]:

$$MED_{1,3} = \sqrt{(705/3)} = \sqrt{235} = 15.3$$

When comparing these three related measures, note that the value of SED is much larger than that for ED and MED because the squared differences are summed. The species with the largest differences between SUs 1 and 3 (i.e., Spp. 1) receive the greatest weighting in the final distance value (e.g., 400 for Spp. 1 versus 289 for Spp. 3 and 16 for Spp. 2).

DISTANCE 4. ABSOLUTE DISTANCE [Eq. (14.4)]:

$$AD_{1,3} = |20 - 0| + |10 - 6| + |17 + 0| = 20 + 4 + 17 = 41$$

DISTANCE 5. MEAN ABSOLUTE DISTANCE [Eq. (14.5)]:

$$MAD_{1,3} = 41/3 = 13.7$$

When contrasting these two related measures with ED, SED, and MED, note that since differences are not squared, less relative importance is given to those species with the larger abundance differences (e.g., Spp. 1).

DISTANCE 6. BRAY–CURTIS DISSIMILARITY [Eq. (14.6c)]:

$$PD_{1,3} = 1 - [(2)(0 + 6 + 0)/(47 + 6)] = 1 - (12/53) = 0.77$$

DISTANCE 7. RELATIVE EUCLIDEAN DISTANCE [Eq. (14.7)]:

$$\begin{aligned} RED_{1,3} &= \sqrt{[(20/47) - (0/6)]^2 + \cdots + [(17/47) - (0/6)]^2} \\ &= \sqrt{(0.426 - 0)^2 + \cdots + (0.362 - 0)^2} \\ &= \sqrt{0.181 + 0.619 + 0.131} = \sqrt{0.931} = 0.96 \end{aligned}$$

DISTANCE 8. RELATIVE ABSOLUTE DISTANCE [Eq. (14.9)]:

$$\begin{aligned} RAD_{1,3} &= |(20/47) - (0/6)| + \cdots + |(17/47) - (0/6)| \\ &= |0.426 - 0| + \cdots + |0.362 - 0| \\ &= (0.426 + 0.787 + 0.362) = 1.57 \end{aligned}$$

DISTANCE 9. CHORD DISTANCE. First determine the cosine of the chord distance (ccos) using Eq. (14.9b):

$$\begin{aligned} ccos_{1,3} &= [(20)(0) + (10)(6) + (17)(0)]/\sqrt{(789)(36)} \\ &= (0 + 60 + 0)/\sqrt{28{,}404} = 60/168.5 = 0.356 \end{aligned}$$

Then the chord distance [Eq. (14.9a)]:

$$\mathrm{CRD}_{1,3} = \sqrt{[2(1.0 - 0.356)]} = \sqrt{(2)(0.64)} = 1.13$$

DISTANCE 10. GEODESIC DISTANCE [Eq. (14.10)]:

$$\mathrm{GDD}_{1,3} = \arccos[0.356] = 1.21$$

14.4 EVALUATION OF DISTANCE FUNCTIONS

In the previous section we presented 10 common distance functions. It is obvious that some of these are very similar to each other, while others seem to be quite different. In this section we have some examples of how these 10 functions perform on different data sets.

All possible (j, k) distances for the data set in Table 14.1 are shown in Table 14.3. Although SUs 2 and 3 share no species, the first five distance measures (the *E*-group) actually indicate that these two SUs are more similar (lower distance value) than either SUs 1 and 2 or SUs 1 and 3, which have one species in common. The final five distance measures (the BC- and RE-groups) do not give this unreasonable result. In fact, PD, RED, RAD, CRD, and GDD each give the same ranking of distances, that is, SUs 1 and 2 are the most similar and SUs 2 and 3 are the least similar, a more realistic result.

A first glance at Table 14.1 might intuitively suggest that the distance between SUs 1 and 2 would be larger than the distance between SUs 1 and 3.

TABLE 14.3 Computed values for each of the 10 distance measures described in the text based on the data given in Table 14.1

Distance Group	Distance Measure	Equation	SU(j,k) (1,2)	(1,3)	(2,3)
E	ED	14.1	20.4	26.6	16.2
	SED	14.2	414.	705.	261.
	MED	14.3	11.7	15.3	9.3
	AD	14.4	32.	41.	21.
	MAD	14.5	10.7	13.6	7.0
BC	PD	14.6	0.52	0.77	1.00
RE	RED	14.7	0.71	0.96	1.41
	RAD	14.8	1.15	1.57	2.00
	CRD	14.9	0.76	1.14	1.41
	GDD	14.10	0.78	1.21	1.57

For the species in common, the absolute difference in abundance between SUs 1 and 2 is 5 (Spp. 1), whereas this difference for SUs 1 and 3 is 4 (Spp. 2). However, the relative weights involved in computing these indices produces the result that the distance between SUs 1 and 2 is smaller than between SUs 1 and 3. Although the data set in Table 14.1 is artificial and simple, it does help to illustrate some of the difficulties when zero data are present in community data (the usual case) and the student should always proceed with caution.

RED and RAD express species abundances relative to the total abundance across all of the SUs. The effect of the RED and RAD expressions is to more equalize the importance of species relative to SUs with high and low total abundances. Two SUs with species in approximately the same proportions will tend to be more similar (i.e., have a close distance). Thus, if one is interested in measuring SU resemblance where species of high abundance in SUs with high total abundance will tend to be equally weighted with species of low abundance in SUs with low total abundance, the RED or RAD measures could be used.

Chord (CRD) and geodesic (GDD) distances compare species abundances relative to the abundance sums of squares for the SUs. Thus, as with RED and RAD, two SUs with species abundances in approximately the same proportions will be close in distance.

To illustrate further the performances of these resemblance functions, the data sets in Tables 11.4*a* and 14.4 were used to compute values for each of

TABLE 14.4 *Percentage abundance data for 11 species (A–K) in seven SUs. These data were used to compute rank correlations between the 10 distance indices discussed in this section*

	SUs						
Species	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A	100	0	50	100	5	0	85
B	90	10	50	40	0	0	0
C	80	20	50	20	5	0	65
D	70	30	50	10	0	5	0
E	60	40	50	5	5	0	75
F	50	50	50	0	0	5	0
G	40	60	50	5	5	5	0
H	30	70	50	10	0	0	65
I	20	80	50	20	5	0	85
J	10	90	50	40	0	5	0
K	0	100	50	100	5	5	0

TABLE 14.5 Spearman rank correlations between the 10 distance measures. Correlations above and below the diagonal are based on the data in Tables 11.4a and 14.4, respectively

Group		ED	SED	MED	AD	MAD	PD	RED	RAD	CRD	GDD
E	ED	1.0	0.99	0.99	0.99	0.99	0.27	−.25	−.24	−.22	−.22
	SED	1.0	1.0	0.99	0.99	0.99	0.27	−.25	−.24	−.22	−.22
	MED	1.0	1.0	1.0	0.99	0.99	0.27	−.25	−.24	−.22	−.22
	AD	0.86	0.86	0.86	1.0	1.0	0.27	−.26	−.24	−.21	−.21
	MAD	0.86	0.86	0.86	1.0	1.0	0.27	−.26	−.24	−.21	−.21
BC	PD	0.48	0.48	0.48	0.47	0.47	1.0	0.64	0.67	0.66	0.66
RE	RED	0.36	0.36	0.36	0.16	0.16	0.56	1.0	0.99	0.97	0.97
	RAD	0.31	0.31	0.31	0.20	0.20	0.63	0.92	1.0	0.97	0.96
	CRD	0.46	0.46	0.46	0.28	0.28	0.57	0.96	0.87	1.0	0.96
	GDD	0.46	0.46	0.46	0.28	0.28	0.57	0.96	0.87	1.0	1.0

these 10 distance functions over all pairwise combinations of SUs. Then, using Eq. (12.1), Spearman rank correlation coefficients were computed between all pairwise combinations of the 10 distance functions (Table 14.5). Distances based on functions *within* the *E*-group (ED, SED, MED, AD, and MAD) and the RE-group (RED, RAD, CRD, and GDD) are highly correlated; on the other hand, the correlations *between* these groups are low. The Bray–Curtis PD index had low correlations with the *E*-group, but fairly high correlations (0.56–0.67) with the RE-group.

From these evaluations, we note the following:

1. In spite of the widespread popularity of the *E*-group distance measures, we do not recommend their use. It is clear from our results (Table 14.3) that spurious results can occur. Wolda (1981) reached a similar conclusion.
2. Any of the RE-group functions appear to perform reasonably well. There seems little advantage in choosing any one over another (given their high correlation, Table 14.5), but we have found chord distance [Eq. (14.9)] to perform very satisfactorily over a diverse set of ecological data sets.
3. PD offers an alternative to the RE-group. This coefficient has been highly recommended by Beals (1984), based on his successful use of PD over a wide range of ecological studies.

14.5 EXAMPLE: PANAMANIAN COCKROACHES

This BASIC microcomputer program SUDIST.BAS (see accompanying disk) was used to compute the distances between six Panamanian localities based

TABLE 14.6 Distance measures between six Panamanian localities based on abundances for five cockroaches using SUDIST.BAS

Locality		Distance Measure									
		E-Group					BC-Group	RE-Group			
(*j*)	(*k*)	ED	MED	SED	AD	MAD	PD	RED	RAD	CRD	GDD
1	2	52.2	23.3	2722	94	18.8	0.50	0.49	0.83	0.74	0.76
1	3	74.6	33.4	5571	115	23.0	0.97	0.74	1.28	1.10	1.16
1	4	75.0	33.5	5626	116	23.2	0.98	1.08	1.76	1.28	1.38
1	5	64.2	28.7	4127	101	20.2	0.76	0.25	0.47	0.34	0.34
1	6	54.5	24.4	2966	94	18.8	0.66	0.47	0.84	0.44	0.44
2	3	47.1	21.1	2219	69	13.8	0.95	0.60	0.89	0.87	0.90
2	4	47.2	21.1	2226	70	14.0	0.97	0.62	0.93	0.64	0.66
2	5	38.6	17.3	1489	55	11.0	0.63	0.40	0.57	0.56	0.57
2	6	38.5	17.2	1486	48	9.6	0.50	0.77	1.18	0.89	0.92
3	4	1.0	0.4	1	1	0.2	0.33	0.71	1.00	0.77	0.79
3	5	11.4	5.1	130	14	2.8	0.78	0.85	1.38	1.18	1.27
3	6	24.1	10.8	579	27	5.4	1.00	1.19	2.00	1.41	1.56
4	5	11.4	5.1	131	15	3.0	0.88	1.02	1.50	1.15	1.22
4	6	24.0	10.8	578	26	5.2	1.00	1.39	2.00	1.41	1.56
5	6	13.7	6.1	187	19	3.8	0.46	0.38	0.63	0.36	0.36

on the abundances of five cockroaches (data in Table 11.4a). The results are given in Table 14.6. Note the great differences in scale for the various distance measures. For example, squared Euclidean distance (SED) ranges up into the thousands, but also as low as one. Recall that SED and the other *E*-group measures (ED, MED, AD, and MAD) range from zero to infinity; this is because they increase as the number of species (S) increases. In contrast, recall that relative Euclidean distance (RED) and chord distance (CRD) have an upper limit of only $\sqrt{2} = 1.41$, relative absolute distance (RAD) has an upper limit of 2, and the geodesic distance (GDD) has an upper limit of $\pi/2 = 1.57$.

14.6 EXAMPLE: WISCONSIN FORESTS

Using the data matrix for eight trees in 10 upland forest sites, southern Wisconsin (Table 11.6*a*), the program SUDIST.BAS (see accompanying disk) was used to calculate all pairwise combinations of chord distances (CRD) between the 10 sites (Table 14.7). Recalling that the maximum value of CRD is 1.41 (for maximum dissimilarity), it is obvious that SUs 1 and 2 are the most similar, followed by SUs 9 and 10.

Often, these results are used for subsequent analyses, such as cluster analysis (Chapter 16). For such analyses, distances are conveniently used in the form

TABLE 14.7 Chord distances between ten upland forest sites, southern Wisconsin in SU × SU matrix form (above diagonal)

SUs	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1)	0.15	0.61	0.50	0.83	1.17	1.02	1.22	1.35	1.30
	(2)	0.64	0.50	0.85	1.16	1.02	1.22	1.33	1.29
		(3)	0.45	0.94	1.14	1.12	1.18	1.38	1.25
			(4)	0.57	0.95	0.90	1.05	1.28	1.18
				(5)	0.79	0.67	0.95	1.16	1.07
					(6)	0.74	0.41	0.80	0.80
						(7)	0.63	0.62	0.61
							(8)	0.52	0.53
								(9)	0.31

of a SU × SU comparison matrix (as illustrated by the 10 × 10 matrix of CRD distances in Table 14.7).

14.7 ADDITIONAL TOPICS ON RESEMBLANCE FUNCTIONS

Hubalek (1982) judged the "admissibility" of 43 similarity coefficients as *Q*-mode resemblance functions for presence–absence data based on five major "conditions." Hubalek suggested that only four "generally worked well" on a set of test data: (1) Jaccard's coefficient of community, (2) Dice's coincidence index, (3) Kulczynski's coefficient, and (4) Ochiai's coefficient. The Jaccard, Dice, and Ochiai coefficients were also highly recommended in an independent critical review of 20 similarity measures by Janson and Vegelius (1981).

Wolda (1981) examined the effects of sample size and species diversity on 22 measures of ecological resemblance, including product-moment and rank correlation coefficients and various measures based on information content. Wolda did not recommend either of the correlation coefficients as similarity indices. Of the information measures he examined, Wolda highly recommended Morisita's (1959) index because it proved to be independent of both sample size and diversity. When the data require prior log transformation, Wolda recommended Horn's (1966) simplified version of Morisita's index and Horn's index of overlap. However, Bloom (1981) found these two indices of Horn's to "diverge greatly from one another and from the theoretical standard" (the standard being based on a table of the area of a normal curve).

Ecologists have not made much use of the probability measures of resemblance, such as the indices described by Goodall (1964, 1966) and Feoli and Lagonegro (1983). This is due in part to the relatively complex and lengthy computations involved. The principal advantages of Goodall's index are: (1) it is on a scale from zero to one, (2) it is linear, and (3) it is applicable to both

abundance and presence–absence data (Orloci 1978). The main disadvantage is that the probabilities for the similarity of any two SUs are based on all the SUs in the data set and, consequently, if SUs are either added to or taken from a data set, the probability-based similarity between the two given SUs will change.

Ecologists also have made little use of information indices, for example, Horn's (1966) index of overlap, for measuring the resemblance between SUs. Orloci (1978) describes some other information measures for SU resemblance. For excellent reviews of different resemblance functions, we refer the student to Boesch (1977), Campbell (1978), Clifford and Williams (1976), Goodall (1978b), Hubalek (1982), Orloci (1972, 1978), Pielou (1984), and Williams and Dale (1965).

Many ecological data sets are a mixture of quantitative information on the density, frequency, cover, biomass, and so forth, of species in SUs. Some species may be dominant in several SUs while absent in others. Also, there are species that may be only rarely found in the entire sample. To avoid the risk of overemphasizing the dominant species in the data analysis, ecologists often employ numerous standardizations or transformations of the data before computing ecological resemblances (Jensen 1978). A large number of types of transformations are possible, some of which we demonstrated above in Section 14.2.2.3. Chardy et al. (1976) used a logarithmic transformation before analyzing high-diversity plankton communities dominanted by only a few abundant species. Other transformations may also have been appropriate (e.g., angular and square-root transforms). Gauch (1982), Greig-Smith (1983), Jensen (1978), Noy-Meir (1973), Noy-Meir et al. (1975), and Orloci (1978) provide excellent overviews on the relative merits of ecological data transformations. Hajdu (1981), in a graphical comparison of 16 resemblance measures, found that standardization by SUs had undesirable effects on an ordered series.

Of course, the search for new (and the comparison of old) ecological resemblance functions will continue. For example, new similarity functions based on presence–absence (binary) data have been proposed and tested (Faith 1983, 1984), as well as a new way to compute distance (Bradfield and Kenkel 1987). The real ecological value will come from gained understanding of which measures are most robust when applied to classification and ordination procedures.

14.8 SUMMARY AND RECOMMENDATIONS

1. Two types of *Q*-mode (SU) resemblance functions may be distinguished: similarity coefficients and distance coefficients (Section 14.1).

2. We recommend the Ochiai, Dice, and Jaccard similarity coefficients for computing SU resemblance when the data consists of species presence–absence data (Section 14.2.1).

3. Three groups of distance functions based on abundance data may be distinguished: the *E*-group (Euclidean distance indices), the BC-group (represented by the Bray–Curtis dissimilarity index), and the RE-group (relative Euclidean distance indices) (Section 14.2.2).

4. We do not generally recommend the use of the *E*-group distance indices despite their widespread popularity. Although these indices have been of great heuristic value in ecology, there are various pitfalls in their use (Section 14.4).

5. For computing SU resemblances when the data consist of quantitative abundance data, we recommend chord distance from the RE-group of distance functions.

CHAPTER 16

Cluster Analysis

Cluster analysis (CA) is a classification technique for placing similar entities or objects into groups or "clusters." The cluster analysis models we present in this chapter are used to place similar samples into clusters, which are arranged in a hierarchical treelike structure called a *dendrogram.* These clusters or groups of SUs may delimit or represent different biotic communities.

16.1 GENERAL APPROACH

Given a set of objects and some measure of their resemblance to each other, we defined classification (Chapter 13) to be the "sorting" of these objects into groups or clusters. Cluster analysis is a technique that accomplishes this sorting. The objects of concern here are ecological samples or SUs (e.g., plots, transects, quadrats). CA is actually a general term that refers to a large number of algorithms that differ mainly in their treatment of cluster formation.

We first present the general approach to CA before detailing procedures. Initially, we must compute a Q-mode resemblance between the SUs. Although numerous resemblance functions could be used, we restrict our coverage to distance measures (see Chapter 14), because of their heuristic value in CA (Sneath and Sokal 1973). The distances between all pairwise combinations of SUs in a collection are summarized into a SU $\times$ SU distance or D matrix (e.g., Table 14.7) and the various CA strategies operate on this D matrix.

The CA models we describe in this chapter are agglomerative (Chapter 13): they begin with a collection of N individual SUs and progressively build groups or clusters of similar SUs. During each clustering cycle, only *one pair*

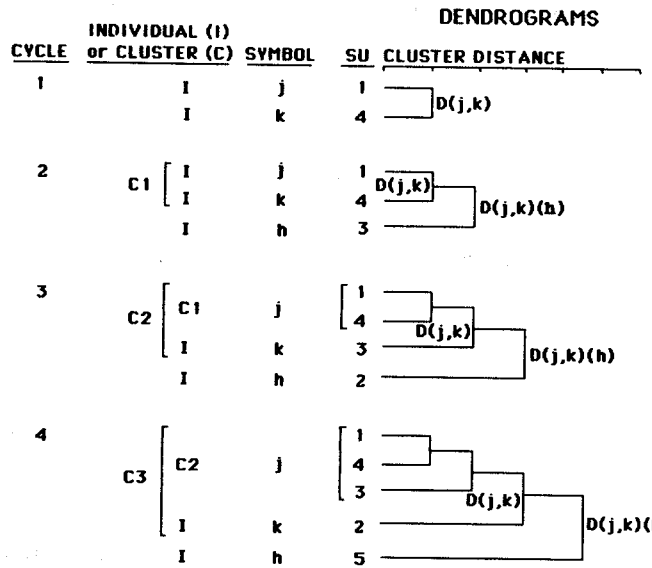

Figure 16.1 *Illustration of the Lance and Williams (1967) combinatorial clustering method on five SUs [see Eq. (16.1)].*

of entities may be joined to form a new cluster. This pair may be (1) an individual SU with another individual SU, (2) an individual with an existing cluster of SUs, or (3) a cluster with a cluster. Hence, the term *pair-group CA* is applied.

A general example of the pair-group approach is illustrated in Figure 16.1. For this example we use five SUs and, hence, 10 pairwise $[N(N-1)/2]$ distance values form the D matrix. The first step in all pair-group CA strategies involves searching the D matrix for the smallest distance value between two individual SUs. In Figure 16.1, this is shown to be between SUs 1 and 4, represented by the symbols j and k, respectively. Hence, the first cluster is formed at a distance $D(j,k)$ and this can be diagrammed using a dendrogram (Figure 16.1, Cycle 1). The initial collection of five SUs is now reduced to one cluster (C_1 = SUs 1 and 4 joined) and three individual SUs (2, 3, and 5). The distance between this cluster and each of these three remaining SUs must now be computed. Special equations have been developed for this type of computation and a general one by Lance and Williams (1967), called the *linear combinatorial equation*, is given below.

The linear combinatorial equation takes the form

$$D(j,k)(h) = \alpha_1 D(j,h) + \alpha_2 D(k,h) + \beta D(j,k) \qquad (16.1)$$

TABLE 16.1 *Parameter values for α_1, α_2, and β in the Lance and Williams' combinatorial equation [Eq. (16.1)] for different hierarchical clustering strategies. Names of the strategies follow Sneath and Sokal (1973)/Lance and Williams (1967). The number of SUs in the jth and kth groups are $t(j)$ and $t(k)$, respectively, and the number of SUs in the combined group (j,k) is $t(j,k)$*

Strategy	α_1	α_2	β
Centroid (unweighted)/ centroid	$t(j)/t(j,k)$	$t(k)/t(j,k)$	$-t(j)t(k)/t(j,k)$
Centroid (weighted)/ median	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$
Group mean/unweighted pair-grouping method	$t(j)/t(j,k)$	$t(k)/t(j,k)$	0
Flexible	0.625	0.625	-0.25^a

[a] β is flexible with this strategy under the constraint that $\alpha_1 + \alpha_2 + \beta = 1$ and that $\alpha_1 = \alpha_2$.

where the distance between the new cluster (j,k) formed from the jth and kth SUs and a third hth SU or group of SUs can be calculated from the known distances $D(j,k)$, $D(j,h)$, and $D(k,h)$ and the parameters α_1, α_2, and β. For example, the distance between SU 3 and the cluster represented by SUs 1 and 4 (Figure 16.1, Cycle 2) is given by

$$D(1,4)(3) = \alpha_1 D(1,3) + \alpha_2 D(4,3) + \beta D(1,4) \tag{16.2}$$

The different clustering strategies differ only in their values for α_1, α_2, and β (Table 16.1), which are the weights for determining the new distances (more on this below).

The relationships between the SUs and the formation of new clusters, as given in Eq. (16.1), are depicted in Figure 16.1 [the j's, k's and h's are those in Eq. (16.1)]. From Figure 16.1:

1. Given N SUs in a collection, there are $N-1$ cycles in CA. In this example, there are four cycles.
2. In cycle 1, two individual SUs (represented by I's) are joined to form a cluster. The distance at which SU 1 (symbol j) and SU 4 (symbol k) form a cluster is given by $D(j,k)$, the value from the D matrix.
3. In cycle 2, SU 3 (symbol h) joins the cluster formed in cycle 1 (symbol C_1). The j and k are SUs 1 and 4, respectively, and the cluster distance between SU 3 and C_1 is $D(j,k)(h)$.
4. In cycle 3, SU 2 (symbol h) joins the cluster formed in cycle 2 (symbol C_2). Note that j now represents cluster C_1, and k is the latest SU to join C_1.

5. In cycle 4, SU 5 (symbol h) joins cluster C_3. In Eq. (16.1), j is cluster C_2 (SUs 1, 3, 4) and k is SU 2.

As previously mentioned, the α's and β in Eq. (16.1) determine the "weighting" of the distances. Depending on the weighting scheme used, the resultant cluster formation will vary. In some cases, the differences are dramatic. Four specific weighting schemes, the centroid (unweighted and weighted), group-average, and flexible, are given in Table 16.1.

The concept of weighting is probably best illustrated by an example. Returning to the example in Figure 16.1, for group-mean weighting the cluster distance between cluster C_3 (where j = SUs 1, 4, and 3 and k = SU 2) and SU 5 is given by [Eq. (16.1)]

$$D(1,4,3;2)(5) = \tfrac{3}{4}D(1,4,3;5) + \tfrac{1}{4}D(2,5) \tag{16.3}$$

where $\alpha_1 = \frac{3}{4}$, $\alpha_2 = \frac{1}{4}$, and $\beta = 0$ (from Table 16.1).

The group mean clustering strategy (the unweighted pair-group method with arithmetic averages), effectively computes the mean of all distances between SUs of one group to the SUs of another and, hence, is unweighted. On the other hand, for the weighted centroid strategy (Table 16.1), the combinatorial equation is

$$D(1,4,3;2)(5) = \tfrac{1}{2}D(1,4,3;5) + \tfrac{1}{2}D(2,5) - \tfrac{1}{4}D(1,4,3;2) \tag{16.4}$$

which weights all fused groups as coequal regardless of differences in the number of SUs in each group. Also, note that in the centroid strategy, once a group is formed, it is replaced by its mean and intercluster distances are those distances between these means or centroids.

16.2 PROCEDURES

The various clustering procedures operate on the D matrix of all possible pairwise combinations of distances between SUs (e.g., Table 14.7). Any of the distance measures presented in Chapter 14 could be used. It is assumed that there are a total of N SUs in the collection.

STEP 1. OBTAINING THE INITIAL GROUP. The $N \times N$ D matrix is searched for the smallest distance value between a pair of SUs. This pair represents the two most similar SUs in the collection. These two SUs are joined (e.g., Figure 16.1).

STEP 2. REDUCTION OF THE D MATRIX. There are now $N - 1$ entities in the collection, in other words, one group composed of two SUs and the remaining $N - 2$ individual SUs. The distances between the new group and these remaining SUs is computed from Equation 16.1. A new reduced D' matrix is formed which is now $(N - 1) \times (N - 1)$.

STEP 3. SEARCH THE REDUCED D' MATRIX. Just as in Step 1, the new D' matrix is searched for the lowest distance value in order to identify the next new group to form.

STEP 4. REPEAT STEPS 2 AND 3 UNTIL ALL SUs ARE JOINED INTO ONE GROUP. This will take a total of $N - 1$ cycles since only a single pair (SU–SU, cluster–SU, or cluster–cluster) may be clustered during any one computational cycle. The number of entities present at the beginning of any cycle (Step 2) is $N - C$, where C is the cycle number.

The final problem in CA is one of identifying specific groups or communities once the clustering is completed. The dendrogram, as shown in Figure 16.1, can be examined for groupings of SUs. While this is largely a subjective decision, there have been some recent attempts to render this decision somewhat more objective (e.g., Hill 1980, Popma et al. 1983, Ratliff and Pieper 1981, Rohlf 1974, 1982). These objective procedures will be discussed in Section 16.6. A general guideline is that one does not divide so finely that one ends up with a large number of fragmentary and uninterpretable groups.

16.3 EXAMPLE: CALCULATIONS

The Lance and Williams (1967) combinatorial linear model [Eq. (16.1)] is illustrated by its application to a D matrix of Euclidean distances (Table 16.2). These distances were computed from the contrived data for abundances of three species in five SUs (Table 11.3*a*). The reconstructions of D after each clustering fusion are also given in Table 16.2. The flexible CA strategy is illustrated.

STEP 1. OBTAIN THE INITIAL GROUP. The smallest Euclidean distance in the D matrix is 1.41 between SUs 2 and 3. Hence, these two SUs are the first group formed and this can be depicted in a dendrogram as shown in Figure 16.2, clustering Cycle 1.

STEP 2. REDUCTION OF THE D MATRIX. The distance between this new group (2, 3) and the three remaining SUs is computed using Eq. (16.1) as

$$D(2,3)(1) = (0.625)(4.69) + (0.625)(5.10) - (0.25)(1.41)$$
$$= 2.93 + 3.19 - 0.35 = 5.77$$

TABLE 16.2 The D matrix of Euclidean distances between five SUs based on the data in Table 11.3a. Only the upper-right triangle is shown: (a) original D matrix, and (b)–(d) reduced D matrices after successive SU fusions

		Sampling Unit (SU)			
(*a*)		(2)	(3)	(4)	(5)
	(1)	4.69	5.10	3.00	2.24
$D =$	(2)		1.41	2.24	5.74
	(3)			3.00	5.92
	(4)				3.74
(*b*)			(2, 3)	(4)	(5)
	(1)		5.77	3.00	2.24
$D' =$	(2, 3)			2.93	6.94
	(4)				3.74
(*c*)				(2, 3)	(4)
$D'' =$	(1, 5)			7.39	3.66
	(2, 3)				2.93
(*d*)					(2, 3, 4)
$D''' =$	(1, 5)				6.18

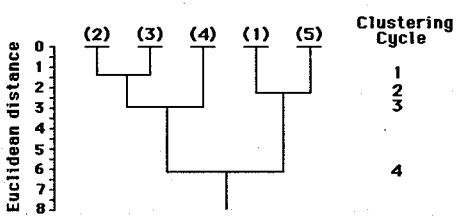

Figure 16.2 *Dendrogram of the clustering of five sampling units using Euclidean distance and the flexible strategy ($\beta = -0.25$).*

$$D(2,3)(4) = (0.625)(2.24) + (0.625)(3.00) - (0.25)(1.41)$$
$$= 1.40 + 1.88 - 0.35 = 2.93$$
$$D(2,3)(5) = (0.625)(5.74) + (0.625)(5.92) - (0.25)(1.41)$$
$$= 3.59 + 3.70 - 0.35 = 6.94$$

The reduced D' matrix is shown in Table 16.2*b*. Note that the distances between the unclustered SUs remain unchanged.

STEP 3. SEARCH THE REDUCED D′ MATRIX. The smallest distance in D' is 2.24 between SUs 1 and 5. Hence, these two SUs are the next cluster formed as shown in Figure 16.2, clustering cycle 2.

STEP 4. REDUCTION OF THE D′ MATRIX. The distance between this new cluster and the remaining SU (4) and group (2, 3) is computed as

$$\begin{aligned} D''(1,5)(2,3) &= (0.625)(5.77) + (0.625)(6.94) - (0.25)(2.24) \\ &= 3.61 + 4.34 - 0.56 = 7.39 \\ D''(1,5)(4) &= (0.625)(3.00) + (0.625)(3.74) - (0.25)(2.24) \\ &= 1.88 + 2.34 - 0.56 = 3.66 \end{aligned}$$

Note that the reduced D' matrix from the previous cycle is used to obtain all new distances. This next reduced D'' matrix is shown in Table 16.2*c*.

STEP 5. SEARCH THIS REDUCED D″ MATRIX. The smallest distance value in D'' is 2.93 between the group represented by SUs 2 and 3 and SU 4. Hence, these three SUs are joined to form a new cluster at a distance of 2.93 as shown in Figure 16.2, clustering cycle 3.

STEP 6. REDUCTION OF THIS D″ MATRIX. The distance between this new cluster of three SUs and the only remaining entity, a group composed of SUs 1 and 5, is computed as

$$\begin{aligned} D'''(2,3;4)(1,5) &= (0.625)(7.39) + (0.625)(3.66) - (0.25)(2.93) \\ &= 4.62 + 2.29 - 0.73 = 6.18 \end{aligned}$$

STEP 7. The final reduced matrix D''' is shown in Table 16.2*d* and the final fusion joins all of the SUs together at a Euclidean distance of 6.18. This is illustrated in Figure 16.2, clustering cycle 4.

16.4 EXAMPLE: PANAMANIAN COCKROACHES

The BASIC program CLUSTER.BAS (see accompanying disk) was used to compute a CA on the Panamanian cockroach data set of Table 11.4*a*. There are five species of cockroaches in six locations, and relative Euclidean distance (RED, see Section 14.2.2.3) is the resemblance function used along with the flexible clustering strategy (see Table 16.1). A summary of the output from the BASIC program is given in Table 16.3. Note that a cluster is referenced by the SU with the lowest numerical value (for example, at cycle 2, the cluster consisting of SUs, 1, 5, and 6, is referred to as "cluster 1").

TABLE 16.3 Program CLUSTER.BAS results giving (a) distances between the six Panamanian localities (SUs), and (b) clustering of the localities

(a) Relative Euclidean distances (D matrix)

$D =$

SUs	(2)	(3)	(4)	(5)	(6)
(1)	0.49	0.74	1.08	0.25	0.47
(2)		0.60	0.62	0.40	0.77
(3)			0.71	0.85	1.19
(4)				1.02	1.39
(5)					0.38

(b) Clustering by the flexible strategy with $\beta = -0.25$

Clustering Cycle	No. of Groups	Clustering Level	Reference SU[a]	SUs in the Group
1	5	0.25	1	5
2	4	0.47	1	5, 6
3	3	0.60	2	3
4	2	0.68	2	3, 4
5	1	1.43	1	All SUs form one group

[a]The lowest numerical value of SUs in group.

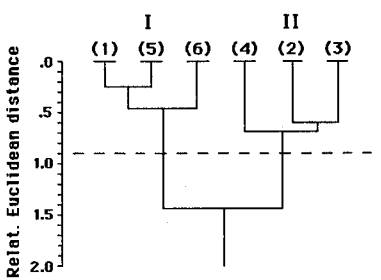

Figure 16.3 *Dendrogram for the cluster analysis of six Panamanian localities using RED and the flexible strategy. The horizontal dashed line shows the arbitrary division line for defining clusters I and II.*

The pattern of clustering for the six locations (SUs) is summarized in the dendrogram in Figure 16.3. To illustrate how these results might be interpreted, we have arbitrarily used a "cutoff" distance of 0.9 (shown as a horizontal dashed line in Figure 16.3). At this level of resemblance there are two distinct clusters: I (SUs 1, 5, and 6) and II (SUs 2, 3, and 4). Referring to Table 11.4*a*, it can be seen that SUs 1, 5, and 6 are largely dominated by a

single species, *Latindia dohrniana.* Of course, this is a simplified data set and used only for illustration (see Section 16.6 for a further discussion of interpreting CA results).

16.5 EXAMPLE: WISCONSIN FORESTS

The Wisconsin forest community data in Table 11.6*a* was used with the BASIC program CLUSTER.BAS. The results of CA on these 10 upland forest sites (with eight trees) using each of the four strategies are given in Table 16.4. Chord distance [CRD, Eq. (14.9)] was used.

The results for the flexible strategy are summarized in a dendrogram (Figure 16.4). The two arbitrary dashed lines, at chord distances of 1.0 and 1.5, can be used as reference points for identifying clusters. At a distance of 1.0, three clusters emerge: *I* (SUs 1, 2, 3, and 4), *II* (SUs 5 and 7), and *III* (SUs 6, 8, 9, and 10). At the higher chord distance of 1.5, clusters *II* and *II* fuse, forming a single cluster. Thus, the four sites dominated by bur and black oak (SUs 1–4) form a cluster distinct from the remaining six sites (SUs 5–10), which are characterized by basswood and sugar maple (see Table 11.6*a*).

From a comparison of each CA strategy (Table 16.4), it can be seen that the results for the centroid methods and the group average are essentially identical to those previously described for the flexible strategy. The major differences are in the clustering of SUs 5 and 7. These SUs are, in fact, somewhat intermediate between Clusters *I* and *III* (Figure 16.4), that is, they have species characteristic of both clusters (see Table 11.6*a*).

These results illustrate an important point. Our experience suggests that

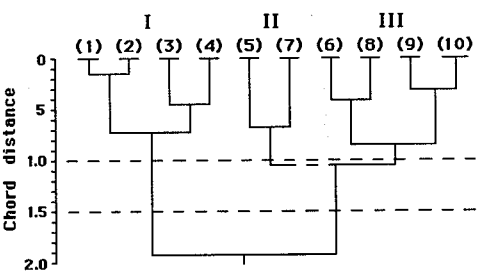

Figure 16.4 *Dendrogram for cluster analysis of 10 upland forest sites, southern Wisconsin, using CRD and the flexible strategy. The horizontal dashed lines represent reference points for delimiting clusters I, II, and III.*

TABLE 16.4 *Program CLUSTER.BAS results giving (a) chord distances between the 10 Wisconsin forest sites (SUs), and clustering of 10 sites by the (b) weighted centroid, (c) unweighted centroid, (d) group-average, and (e) flexible strategies*

(a) Chord distances (D matrix)

SUs	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1)	0.15	0.61	0.50	0.83	1.17	1.02	1.22	1.35	1.30
(2)		0.64	0.50	0.85	1.16	1.02	1.22	1.33	1.29
(3)			0.45	0.94	1.14	1.12	1.18	1.38	1.25
(4)				0.57	0.95	0.90	1.05	1.28	1.18
(5)					0.79	0.67	0.95	1.16	1.07
(6)						0.74	0.41	0.80	0.80
(7)							0.63	0.62	0.61
(8)								0.52	0.53
(9)									0.31

(b) Clustering by the centroid (weighted) strategy

Clustering Cycle	No. of Groups	Clustering Level	Reference SU[a]	SUs in the Group
1	9	0.15	1	2
2	8	0.31	9	10
3	7	0.41	6	8
4	6	0.45	3	4
5	5	0.41	1	2, 3, 4
6	4	0.48	6	8, 9, 10
7	3	0.44	6	7, 8, 9, 10
8	2	0.62	1	2, 3, 4, 5
9	1	0.52	1	All SUs form one group

(c) Clustering by the centroid (unweighted) strategy

Clustering Cycle	No. of Groups	Clustering Level	Reference SU[a]	SUs in the Group
1	9	0.15	1	2
2	8	0.31	9	10
3	7	0.41	6	8
4	6	0.45	3	4
5	5	0.41	1	2, 3, 4
6	4	0.48	6	8, 9, 10
7	3	0.44	6	7, 8, 9, 10
8	2	0.62	1	2, 3, 4, 5
9	1	0.65	1	All SUs form one group

TABLE 16.4 (*continued*)

(d) Clustering by the group-average strategy

Clustering Cycle	No. of Groups	Clustering Level	Reference SU[a]	SUs in the Group
1	9	0.15	1	2
2	8	0.31	9	10
3	7	0.41	6	8
4	6	0.45	3	4
5	5	0.56	1	2, 3, 4
6	4	0.61	7	9, 10
7	3	0.67	6	7, 8, 9, 10
8	2	0.79	1	2, 3, 4, 5
9	1	1.13	1	All SUs form one group

(e) Clustering by the flexible strategy with $\beta = -0.25$

Clustering Cycle	No. of Groups	Clustering Level	Reference SU[a]	SUs in the Group
1	9	0.15	1	2
2	8	0.31	9	10
3	7	0.41	6	8
4	6	0.45	3	4
5	5	0.67	5	7
6	4	0.72	1	2, 3, 4
7	3	0.84	6	8, 9, 10
8	2	1.04	5	6, 7, 8, 9, 10
9	1	1.93	1	All SUs form one group

[a] Lowest numerical value of SUs in group.

these CA strategies will usually give very similar results when somewhat well-defined groups exist in the D matrix. The Wisconsin forest data set illustrates this quite well; because of the simplicity of the data in Table 11.6*a*, the patterns are obvious and the CA results are consistent with these observations. Where problems arise is when the data sets are large and complex, and with patterns that are not obvious *a priori* to the ecologist. The different strategies may give somewhat different results and, because well-defined cluster patterns may not necessarily emerge in any of the strategies, caution is urged. Note that these CA results differ slightly from the classification of SUs by association analysis (Section 15.4) where SUs 1–5 grouped separately from SUs 6–10 (see Figure 15.2)

16.6 ADDITIONAL TOPICS IN CLUSTER ANALYSIS

In this chapter four hierarchical clustering methods were presented that operate under the combinatorial linear model of Lance and Williams (1967). These methods are computationally efficient once the distance matrix (D) is calculated, since it contains all the information needed to cluster the SUs. Other strategies require the repeated use of the original D matrix at each cycle, which becomes very tedious and is computationally inefficient. Extensive reviews of various CA methods are given in Anderberg (1973), Gauch (1982), Goodall (1978a), Orloci (1978), Pielou (1977, 1984), Romesburg (1984), Sneath and Sokal (1973), and Whittaker (1978b).

Another hierarchical clustering method that is popular with ecologists is *minimum variance clustering* (also known as *Ward's method*—see Everitt 1974, Hartigan 1975, and Orloci 1967a). This method has great intuitive appeal because it is based on the simple underlying principle that at each stage of clustering the variance *within* clusters is minimized with respect to the variance *between* clusters. The within-group variance is defined as the sums of squares of the distances between SUs within the cluster and the centroid of the cluster. At each clustering cycle, the two SU clusters whose fusion results in the smallest (minimum) increase in variance (relative to the variances within each cluster taken separately) are joined. The computations required by this method can be done through the use of SAS (Statistical Analysis Systems, Ray 1982) under the cluster procedure option for Ward's method.

As briefly mentioned in Section 16.2, delimiting homogeneous communities from the information provided by the clustering process is usually done subjectively. The ecologist usually has some feeling about the number of communities (groups of SUs) expected from the given data set, and it is a simple matter to "cut the stems" in the dendrogram (e.g., Figure 16.4) at the clustering level that gives this number of SU groups or communities.

There are, however, numerous objective methods that can be used along with the intuition of the ecologist. One of the earliest methods proposed for evaluating dendrograms was *cophenetic correlation* (Sneath and Sokal 1973), where the distances between SUs implied from the dendrogram are compared to the original SU $\times$ SU distance (D) matrix. Proceeding through each cluster cycle, as larger and larger groups are formed (ceasing when all SUs are joined), the correlation between the original D matrix distances and the dendrogram distances will drop. A large drop in this correlation from one cycle to another would suggest stopping fusion at the previous cycle. For example, if the cophenetic correlation for the CA shown in Figure 16.4 was 0.80 at the chord distance 1.0 (upper dashed line) but dropped to 0.50 at distance 1.5 (lower dashed line), then perhaps the clusters formed by cutting the stems at 1.0 should be accepted as homogeneous communities.

Orloci (1967a) suggested that significance levels could be determined for delimiting homogeneous groupings during the minimum variance clustering procedure. Goodall (1978a) formalized an approximate variance ratio for testing whether the increase in variance caused by the fusion of an SU or SU group with another SU group is within an "acceptable" (e.g., $P = 0.05$) level. Ratliff and Pieper (1981) generalized this analysis-of-variance approach to include a test of the hypothesis that the mean intracluster distance is not significantly different from the mean intercluster distance. Their procedure begins with applying the test at the *two-group* level, that is, testing for differences in mean distances between all SU as one group versus being split into two groups. This follows the procedure for Hill's (1980) *stopping rule*, which is a similar approach. For other recent developments in evaluating classifications, the student is referred to the reviews by Archie (1984) and Rohlf (1974) and studies by Duncan and Estabrook (1976), Popma et al. (1983), and Rohlf (1982).

The CA strategies shown in Table 16.1 (centroid, group average, and flexible) are considered *space conserving*; the clustering of SUs at the various levels (distances) introduces relatively little distortion when these clustering distances are compared to the original SU $\times$ SU D matrix distances.

We have not included all the CA strategies that operate under the Lance and Williams linear combinatorial model. For example, the single linkage and complete linkage strategies (Sneath and Sokal 1973) are omitted. These two CA methods are strongly *space distorting*, either by contracting distances with "single" linkages or by dilating distances with "complete" linkages (Pielou 1977). That is, after fusion of clusters the reconstructed distance matrix differs greatly from the original distance matrix (D).

The centroid strategies frequently result in what are termed *reversals*: the distance between centroids of some pair may be less than that between another pair merged at an earlier cycle. If the sum of the parameters α_1, α_2, and β equals 1, then successive hierarchical joinings will be monotonic and reversals will not occur. The flexible strategy is, therefore, by definition monotonic since the sum of the parameters is constrained to equal 1. The flexible strategy's chief feature is that by varying β, which controls the space-conserving properties of the clustering strategy, the space can be made to either dilate or contract. A β near -0.25 tends to be space-conserving, but as β becomes more negative, the distortion is toward dilating and as β becomes more positive, the distortion is toward contracting. We suggest the student refer to Sneath and Sokal (1973) for further details on the flexible strategy.

In this chapter we have presented cluster analysis, assuming the data were from SUs randomly dispersed over the landscape and with species abundances or presence–absence observations. However, clustering SU data from a time sequence can be used to examine ecological succession models (Legendre

et al. 1985). Other types of observations can also be used for clustering, for example, forest tree size classes or soil profile data (Faith et al. 1985).

16.7 SUMMARY AND RECOMMENDATIONS

1. Cluster analysis (CA) is a technique that accomplishes the sorting of objects (SUs) into groups or clusters based on their overall resemblance to one another. Similar SUs will form clusters distinct from other clusters of SUs. *Cluster analysis* is a general term that refers to a large number of such algorithms that differ mainly in their treatment of cluster formation.

2. The Lance and Williams (1967) general algorithm for CA is a linear combinatorial equation [Eq. (16.1)]. By selecting various values for the parameters of Equation 16.1 (as shown in Table 16.1), four CA strategies can be accomplished: the centroid (both weighted and unweighted), group-average, and flexible.

3. The results of CA are conveniently summarized in a dendrogram (e.g., Figure 16.2). The identification of specific groups or communities from this dendrogram is somewhat subjective. As a general guideline, we recommend not dividing so finely that you end up with a large number of fragmented clusters. Some objective methods have been proposed by several researchers (Section 16.6). Although it is helpful for the student to use methods such as CA as an aid in interpreting data, it is a mistake to place strong emphasis on results of a single analysis.

4. Most of the CA strategies usually give very similar results when the basic data set being analyzed is, in fact, one that is characterized by some relatively obvious patterns (such as the example in Section 16.5). However, in cases when the data set is large, somewhat complex in nature, and with no obvious patterns, the results of the various CA strategies can often vary (in some cases, substantially). In the latter case, we recommend that alternative strategies be explored and their results compared; such comparisons often help identify logical clusters (in view of the underlying ecological knowledge of the data).