3. Multivariate Resemblance

FISH 560: Applied Multivariate Statistics for Ecologists

Topics

- Calculate similarities and dissimilarities among objects based on a set of descriptors

R Libraries: vegan, simba, cluster, ecodist, gclus
R Source: biostats, coldiss

BACKGROUND

Ecologists have long collected quantitative observations to determine the resemblance between either objects under study or the variables describing them. Measuring the association between objects or descriptors is the first, and sometimes the only step in the numerical analysis of ecological data. The most common approach to assess resemblance is to first condense all (or the relevant part of) the information available in the data matrix into a square matrix of association among the objects or descriptors. In most instances the matrix is symmetric.

It follows that the structure resulting from the numerical analysis (i.e., the association matrix) does not necessarily reflect all the information originally contained in the ecological data matrix. This stresses the importance of choosing an appropriate measure of association! The choice determines the issue of the analysis. Therefore, the following facts must be considered:

- The nature of the study determines the structure of the data to be evaluated with an association matrix.
- The various measures available are subject to different mathematical constraints (depends upon whether one continues with ordination or clustering).
- Computational aspects such as what measures are available or can be programmed in particular software.

The analysis of similarities among objects (rows) is designated as Q mode analysis, whereas when relationships among variables (columns) are the focus of the study, this is referred to as R mode analysis (Legendre and Legendre 1998). Noticeably, the two modes of analysis do not generally use the same association coefficients. Although it is not possible to give a full review of all association coefficients here, it is useful to know that, for comparing objects (rows) based on their column attributes in Q mode analysis, coefficients may be chosen as a function of data type (quantitative, qualitative, ordinal, or mixed data, normalized data, presence/absence), importance given to rare species, weight given to each object, and calculation of associated probability levels.

In the following chapter, we will explore a large group of association coefficients based on the notion of similarity and dissimilarity. Throughout, the term “association coefficient” is used to describe any measure used to quantify the resemblance or difference between objects or descriptors.

SET-UP

First, set-up your R work session by setting the current work directory to your folder of choice and load the vegan, simba, ecodist, cluster libraries and source the BIOSTATS file from the *File* pull-down menu. You can also do this using the functions **`setwd`**, **`library`** and **`source`**.

Second, import the MAHA environment dataset, MAHA species abundance dataset and MAHA fish trait dataset by typing:

```
envdata <- read.csv('MAHA_environment.csv',header=TRUE, row.names=1)
```

```
speabu <- read.csv('MAHA_speciesabu.csv',header=TRUE, row.names=1)
spetrait <- read.csv('MAHA_speciestrait.csv',header=TRUE, row.names=1)
```

Remember to screen your data for errors, outliers, missing values, etc … and conduct any data transformations or standardizations (see previous chapter). The MAHA dataset should look pretty good because it was pre-screened prior to being released to you. *Therefore, for the purposes of this computer lab you can use the raw environmental and species dataset.*

CALCULATING COEFFICIENTS OF SIMILARITY FOR BINARY DATA (SPECIES PRESENCE/ABSENCE)

Let's calculate the similarity among sites according to their fish community composition based on species presence/absence. First, let's transform the species abundances into presence/absence (i.e., binary transformation) using the power method with an exponent equal to zero, by typing:

```
speocc <- data.trans(speabu,method='power',exp=0,plot=F)
```

Remember to include **plot=F** so that you are not required to press enter for each plot. If you do not you must press enter for all plots before **speocc** is created. Alternatively you could directly import the data (assuming that you have prepared it already), by typing:

```
speocc <- read.csv('MAHA_speciesocc.csv',header=TRUE, row.names=1)
```

There are two very useful functions to compute similarity (**sim** function in the **samba** library) and dissimilar coefficients (**vegdist** function in the **vegan** library). I would highly recommend that you explore both!

Now, let's use the **sim** function in the **simba** library to calculate a series of (dis)similarity coefficients to describe the resemblance of sites according to their species composition. This library will compute 56 dis(similarity) measures for binary data, in this case species presence/absence.

Its usage is:

```
sim(x, coord=NULL, method = "soer", dn=NULL, normalize = FALSE,
listin = FALSE, listout = FALSE, ...)
```

The relevant arguments are:

1. x – the data matrix to be analyzed
2. method – Binary Similarity index including "soerensen", "jaccard", "ochiai", "mountford", "whittaker", "lande", "wilsonshmida", "cocogaston", "magurran", "harrison", "cody", "williams", "williams2", "harte", "simpson", "lennon", "weiher", "ruggiero", "lennon2", "rout1ledge", "rout2ledge", "rout3ledge", "sokal1", "dice", "kulcz1insky", "kulcz2insky", "mcconnagh", "manhattan", "simplematching", "margaleff", "pearson", "roger", "baroni", "dennis", "fossum", "gower", "legendre", "sokal2", "sokal3", "sokal4", "stiles", "yule", "michael", "hamann", "forbes", "chisquare", "peirce", "eyraud", "simpson2", "legendre2", "fager", "maarel", "lamont", "johnson", "sorgenfrei", "johnson2".
3. Etc …

For additional documentation type **?sim** at the R prompt.

Let's try calculating Jaccard's similarity matrix, by typing:

```
sp.jac <- sim(speocc, method = "jaccard")
```

Type **sp.jac**. The first upper left cells in the similarity matrix should be:

```
           S1          S2          S3          S4
S2  0.29411765
S3  0.61538462 0.46153846
S4  0.14285714 0.16666667 0.18181818
S5  0.56250000 0.27777778 0.46666667 0.30769231
```

Now calculate a series of different similarity matrices based on the Simple Matching coefficient and Sørenson's coefficient (note that Sørenson is incorrectly spelt in **sim** library … huh!). For example, you can type:

```
sp.sim <- sim(speocc, method = "simplematching")
sp.sor <- sim(speocc, method = " soerensen")
```

How do these matrices compare to each other? You should be able to note and explain the differences. We can do this graphically by plotting the pair-wise similarities based on two different coefficients against one another, by typing:

```
plot(sp.jac,sp.sim,xlab="Jaccard's coefficient",ylab="Simple Matching
coefficient")
```

We can add a 1:1 line to help with the comparisons, by typing:

```
abline(0,1,col="darkgray")
```

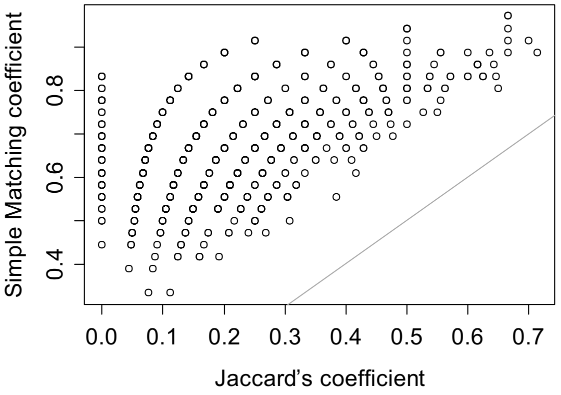
This should produce a plot resembling the one presented on the bottom left. Why do many sites exhibit 0% similarity based on Jaccard's coefficient but range between 45 and 85% similarity based on the Simple Matching coefficient?

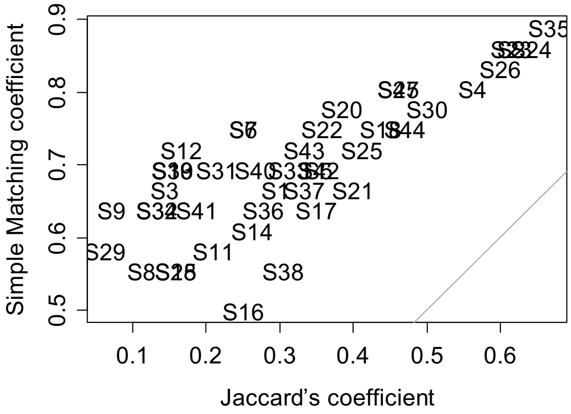
Next, let's just compare the similarities between site S1 and all others (recall that the dataset contains 45 sites (objects)) and add site labels instead of points, by typing:

```
plot(sp.jac[1:45],sp.sim[1:45],xlab="Jaccard's coefficient",ylab="Simple
Matching coefficient",type="n")

text(sp.jac[1:45],sp.sim[1:45],row.names(speocc))

abline(0,1,col="darkgray")
```

This should produce a plot resembling the one presented on below. How might we explain the large discrepancy for site S29?

Repeat the above comparisons for site similarities based on Sørenson's coefficient and other coefficients of interest. Have fun exploring and explaining these differences. Use this information and your knowledge of the statistical properties of the coefficients to select an appropriate similarity index for examining patterns of community composition based on species presence/absence.





CALCULATING COEFFICIENTS OF (DIS)SIMILARITY FOR CONTINUOUS SPECIES DATA (SPECIES ABUNDANCE, DENSITY)

Let's calculate the (dis)similarity among sites according to their fish community composition based on species abundance. From the **vegan** library we can use the function **vegdist** which can calculate a number of similarity/dissimilarity coefficients commonly used in ecology.

Its usage is:

```
vegdist(x, method="bray", binary=FALSE, diag=FALSE, upper=FALSE,
     na.rm = FALSE, ...)
```

The relevant arguments are:

4. x – the data matrix to be analyzed
5. method - Dissimilarity index, partial match to "manhattan", "euclidean", "canberra", "bray", "kulczynski", "jaccard", "gower", "morisita", "horn", "mountford", "raup" , "binomial" or "chao".
6. binary - Perform presence/absence standardization before analysis using decostand. True or false (default)

For additional documentation type "?vegdist" at the R prompt.

Let's try calculating a dissimilarity (distance) matrix based on the Bray-Curtis coefficient:

```
sp.bray <- vegdist(speabu, method="bray")
```

Type **sp.bray**. The first upper left cells in the Bray-Curtis distance matrix should be:

```
      S1        S2        S3
S2    0.8418879
S3    0.8635236 0.7457213
S4    0.9900867 0.9854015 0.8963415
```

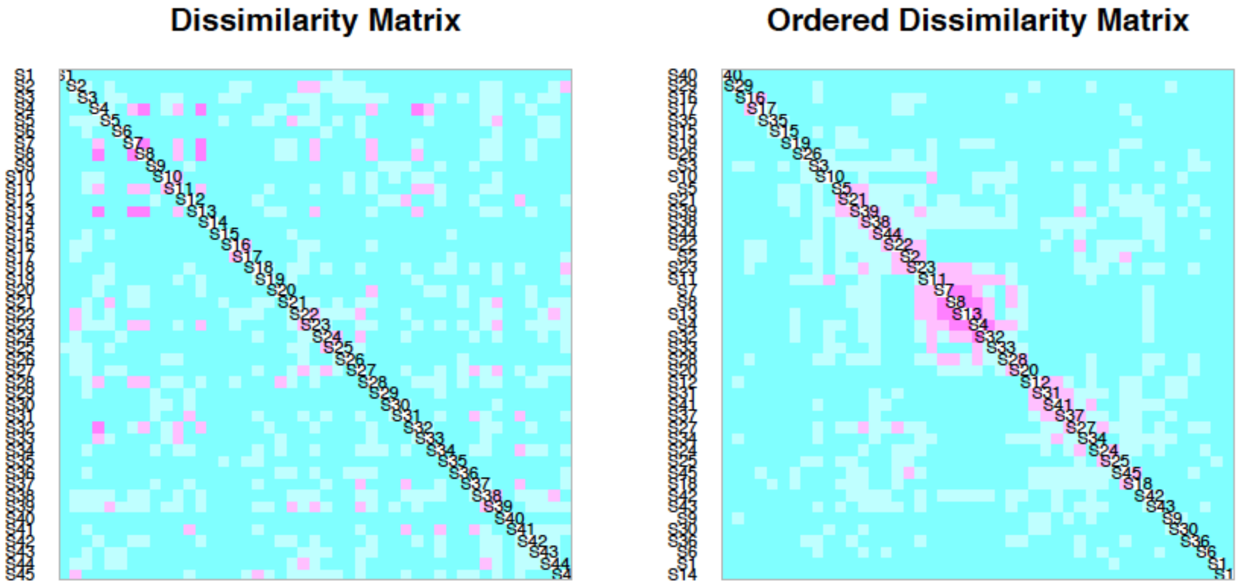
Recall that these are distance values, therefore a value of 1 indicates two sites that have zero similarity in species abundance.

When you have many objects in your dataset it may be more useful to display them in a way that emphasizes their main features.

Borcard et al. (2011) provides a nice function called "coldiss.R" that color codes distances in a heat map with (or without) re-ordering the records by dissimilarity. The function calls upon other functions in the gclus library. So, remember to source the coldiss.R file and install the gclus library!

```
coldiss(sp.bray,nc=4,byrank=FALSE,diag=TRUE)
```

The resulting plot (below) displays the raw and ordered dissimilarity matrices, where magenta are dissimilarities close to 0 and cyan are dissimilaries close to 1. The number of colors is set using “nc”, byrank=T refers to equal-sized categories and byrank=F refers to equal-length categories. If diag=T then object labels are displayed on the diagonal.



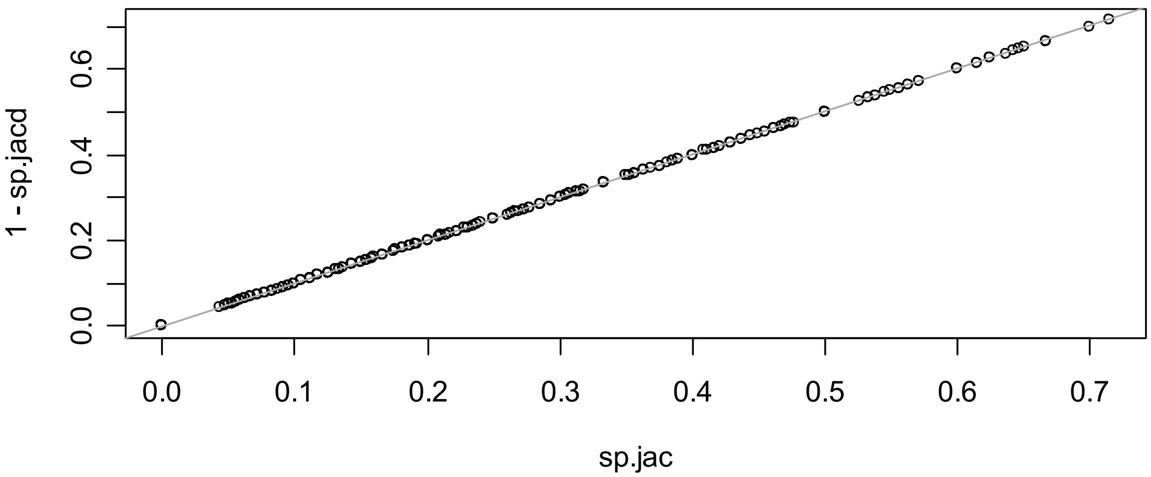
Also note that you can use the **vegdist** function to calculate many of the similarity coefficients that we previously explored. For example, you can calculate a distance matrix based on Jaccard’s coefficient, by typing:

```
sp.jacd<-vegdist(speocc, method=”jaccard”)
```

You should find that **sp.jacd** = 1- **sp.jac** because the only difference is that the first matrix expresses the values as distances and the second matrix contains similarity values. Quickly plot the results to confirm this, by typing:

```
plot(sp.jac,1-sp.jacd)
abline(0,1,col="darkgray")
```

This will return the following in the graphic window:



There are a number of alternative dissimilarity and distance metrics that are appropriate for asymmetrical species data (abundance) that are worth exploring. For example, we can use **Chord distance** or **Hellinger distance** on the raw species abundance datasets by first transforming the dataset and then calculating the Euclidean distance matrix. This is discussed in detail in the lecture.

For Chord's distance:

```
speabu.norm<-decostand(speabu, method='nor')
sp.chordd<-dist(speabu.norm)
```

For Hellinger's distance:

```
speabu.hel<-decostand(speabu, method='hel')
sp.held<-dist(speabu.hel)
```

CALCULATING COEFFICIENTS OF SIMILARITY FOR MIXED DATA TYPES

In many instances you will be interested in assessing the similarity among objects that are described by a suite of mixed descriptor types, including continuous, nominal and ordinal. When faced with this challenge, I suggest that you seriously consider Gower's similarity coefficient (see lecture notes).

Using the fish species trait data, let's explore the utility of Gower's similarity coefficient. First, let's take a look at the trait data by typing:

```
str(spetrait)
```

This will confirm that the data matrix contains both continuous and categorical data. Next, let's calculate Gower's similarity using the **daisy** function in the **cluster** library. NOTE: You will not be able to use the **sim** or **vegdist** functions because the trait matrix contains non-numeric data. Also, the Gower calculations in these libraries do not handle multiclass variables and do not handle missing values correctly (at least this was the case the last time I checked). The **daisy** function computes all the pairwise dissimilarities (distances) between observations in the data set.

Its usage is:

```
    daisy(x, metric = c("euclidean", "manhattan", "gower"),
    stand = FALSE, type = list())
```

The relevant arguments are:

7. x – the data matrix to be analyzed
8. metric - character string specifying the metric to be used. The currently available options are "euclidean" (the default), "manhattan" and "gower".

For additional documentation type "?daisy" at the R prompt.

Let's try calculating a dissimilarity (distance) matrix based on Gower's coefficient:

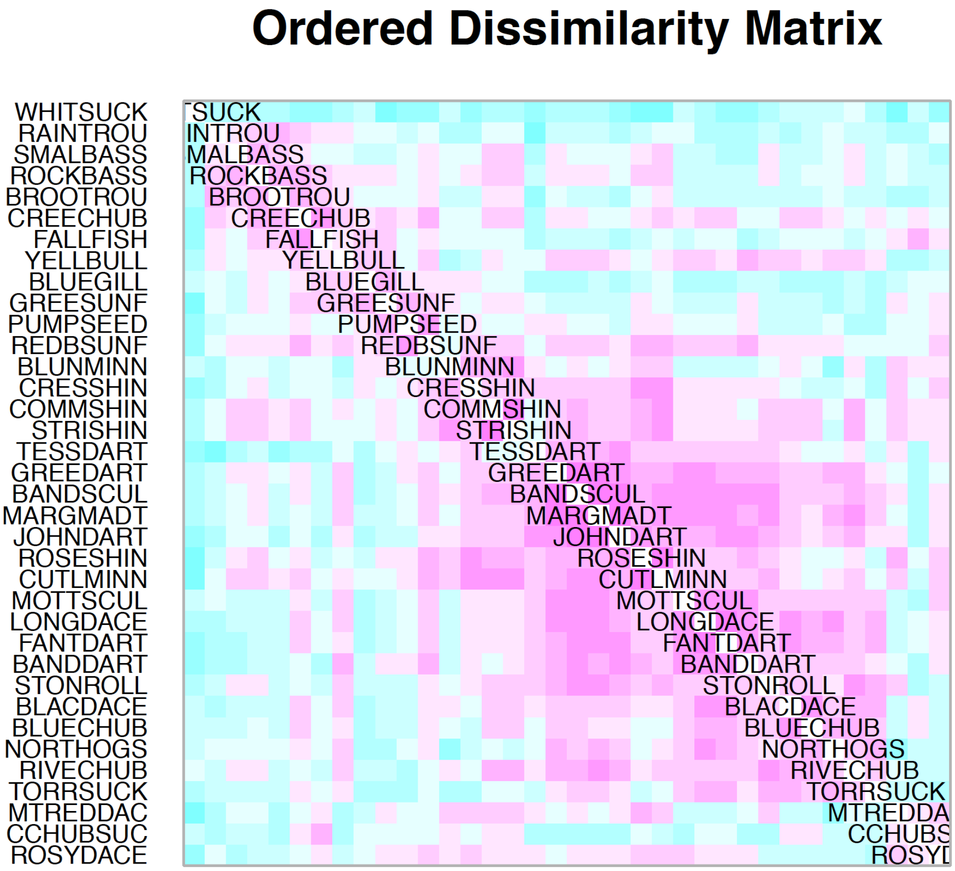
```
sptr.gower <- daisy(spetrait,metric="gower")
```

Type **sptr.gower**. The first upper left cells in the Gower distance matrix should be:

```
         BANDDART   BANDSCUL   BLACDACE   BLUEGILL
BANDSCUL 0.17464214
BLACDACE 0.28047433 0.28189602
BLUEGILL 0.52564155 0.61766608 0.62016722
BLUECHUB 0.32943933 0.32146386 0.10316266 0.67120222
```

We can quickly produce a heat map to look at similarities/diferences between species according to their traits. Here is just the ordered heat map.

```
coldiss(sptr.gower,nc=4,byrank=FALSE,diag=TRUE)
```



Makes sense. Taxonomically similar species have low distances (or high trait similarities) as indicated by the magenta color.

Special note: **`gowdis()`** function in the **`FD`** library is the most complete function to compute Gower's coefficient. It computes the distance for mixed variables, including sasymmetrical binary variables. Variable weights can be specified! Check it out.

CALCULATING COEFFICIENTS OF (DIS)SIMILARITY FOR CONTINUOUS DATA

For continuous data in which double-zeros are meaningful (i.e., environment data, species measurements or traits, etc …) we can calculate dissimilarities between objects based on any number of distance coefficients. Let's calculate environmental dissimilarity among the MAHA sites according to Euclidean distance (the most commonly used coefficient). Start by typing:

```
env.euc <- vegdist(envdata, method="euclidean")
```

Type **`env.euc.`** The first upper left cells in the Euclidean distance matrix should be:

```
      S1          S2          S3
S2    905.84678
S3    197.51321   720.34388
S4    193.87537   1086.18669  386.26245
```

To calculate pairwise Mahalanobis distances between sites is slightly more involved. Lots of solutions on the Internet. I suggest you check the following if you are interested in this distance metric.

http://stats.stackexchange.com/questions/65705/pairwise-mahalanobis-distance

CONVERTING SIMILARITY TO DISSIMILARITY (DISTANCE) OR VICE VERSA

Remember that when dealing with matrices, it is possible to change a similarity matrix (S) into a dissimilarity matrix (D) by applying the following transformations:

$$D = 1 - S \qquad D = \sqrt{(1 - S)} \qquad D = \sqrt{(1 - S^2)}$$

Note that we did this previous using Jaccard's coefficient.

CONVERTING CORRELATION TO DISTANCE

At times you might want to convert a correlation matrix to a distance matrix in order to perform subsequent multivariate analysis. Of course, you cannot use the transformations listed above because correlation values will vary between -1 and 1, thus resulting in negative distances when R<0. Not possible! Therefore I recommend the following transformation:

$$D = \sqrt{(2 - 2*\text{correlation value})}$$

Let's do this using the environmental data by typing:

```
env.dis <- sqrt(2-2*cor(envdata))
```

OPTIONAL READINGS

Baroni-Urbani, C., and M.W. Buser. 1976. Similarity of binary data. Syst. Zool. 25:251-259.

Bloom, S.A. 1981. Similarity indices in community studies: potential pitfalls. Marine Ecology – Progress Series 5: 125-128.

Bray, J.R., and J.T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecological Monographs 27: 325-349.

Chao, A., R.L. Chazdon, R.K. Colwell, and T-J Shen. 2005. A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecology Letters 8: 148-159.

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

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

Pavoine, S., Vallet, J., Dufour, A.B., Gachet, S., and Daniel, H. 2009. On the challenge of treating various types of variables: application for improving the measurement of functional diversity. Oikos 118: 391-402.

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

EXERCISE

Purpose

In this exercise, we will explore a large group of association coefficients based on the notion of similarity and dissimilarity.

Tasks

1. Compute the similarity among the objects of your data using an appropriate suite of association coefficients
   a. What is your ecological interpretation of associations among objects according to the descriptors (e.g., among stream sites according to their fish communities)
   b. How do the similarity matrices differ? Can you interpret these differences based on what you know about the equations underpinning the coefficients?
2. Compute the (dis)similarity among the descriptors of your data using an appropriate suite of association coefficients
   a. What is your ecological interpretation of associations among descriptors?
   b. How do the similarity matrices differ? Can you interpret these differences based on what you know about the equations underpinning the coefficients?
3. Compare association matrices for a single coefficient using raw and transformed data.
   a. How do your results differ? Think about simple graphically and numerical ways to present these comparisons.
4. Summarize patterns of similarity in your own dataset using the approaches described in Chapter 2.
   a. Can you use your results to identify potential outliers or influential variables?

Properties of distance coefficients calculated from a variety of similarity coefficients. Reprinted from Table 7.2. and Table 7.3 of Legendre and Legendre (2012).

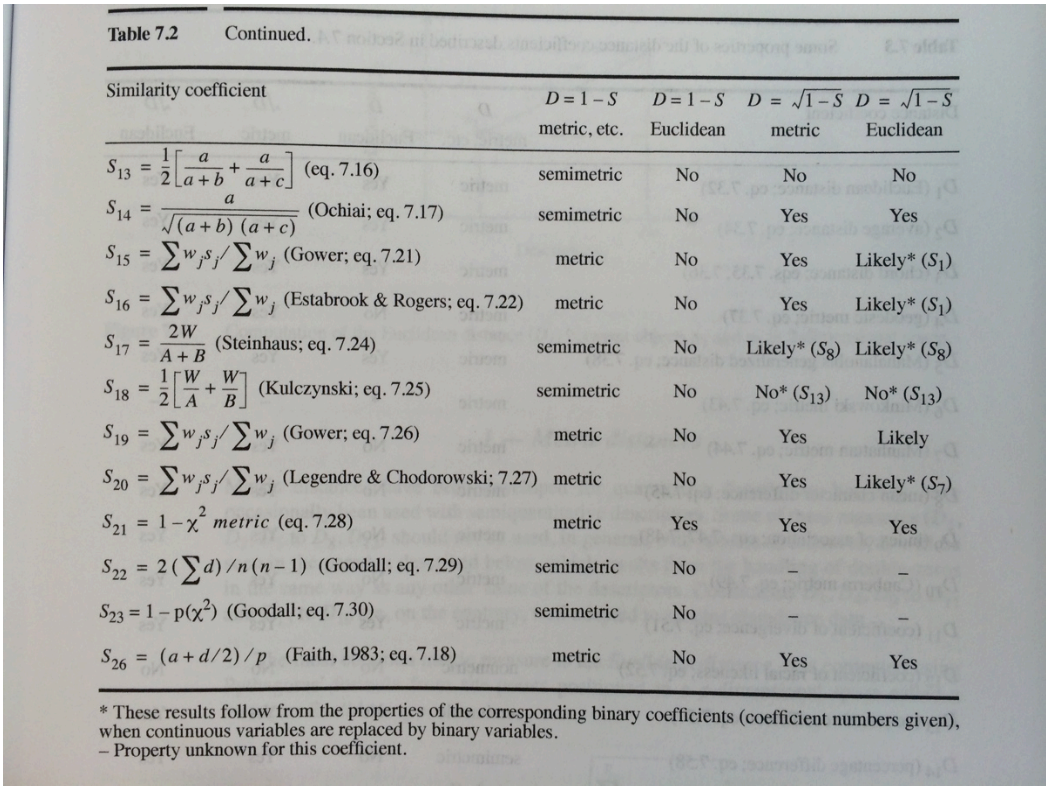


Table 7.2 Continued.

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.

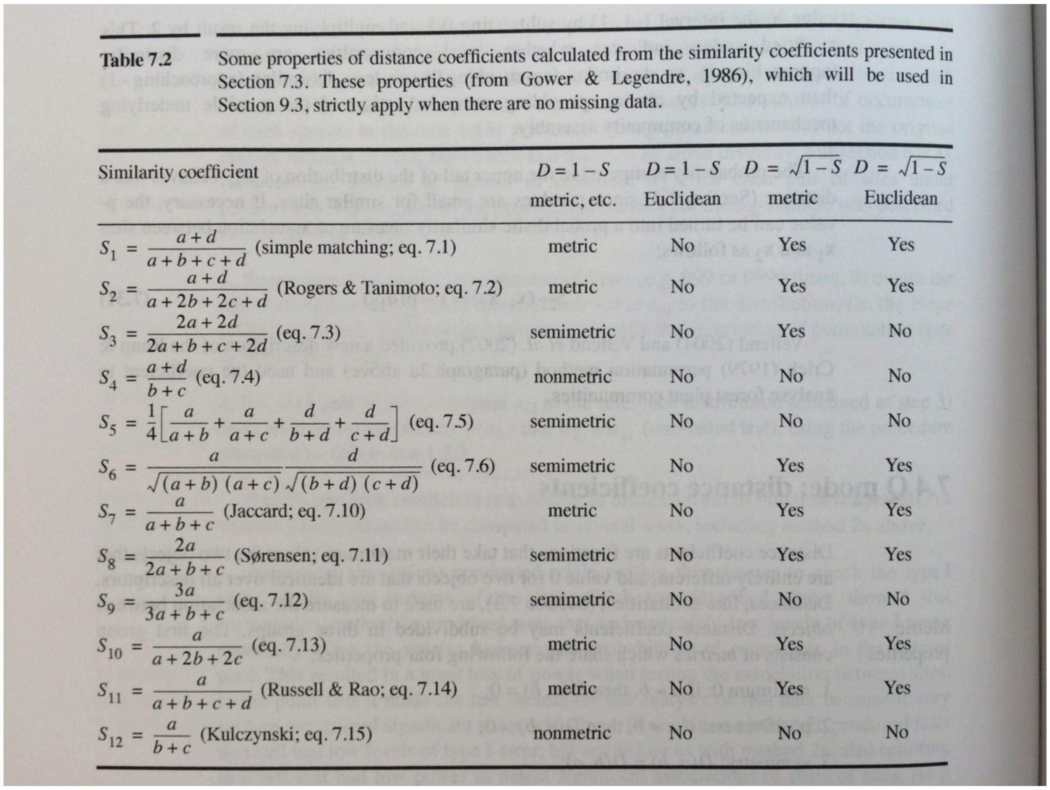


Table 7.2 Some properties of distance coefficients calculated from the similarity coefficients presented in Section 7.3. These properties (from Gower & Legendre, 1986), which will be used in Section 9.3, strictly apply when there are no missing data.

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

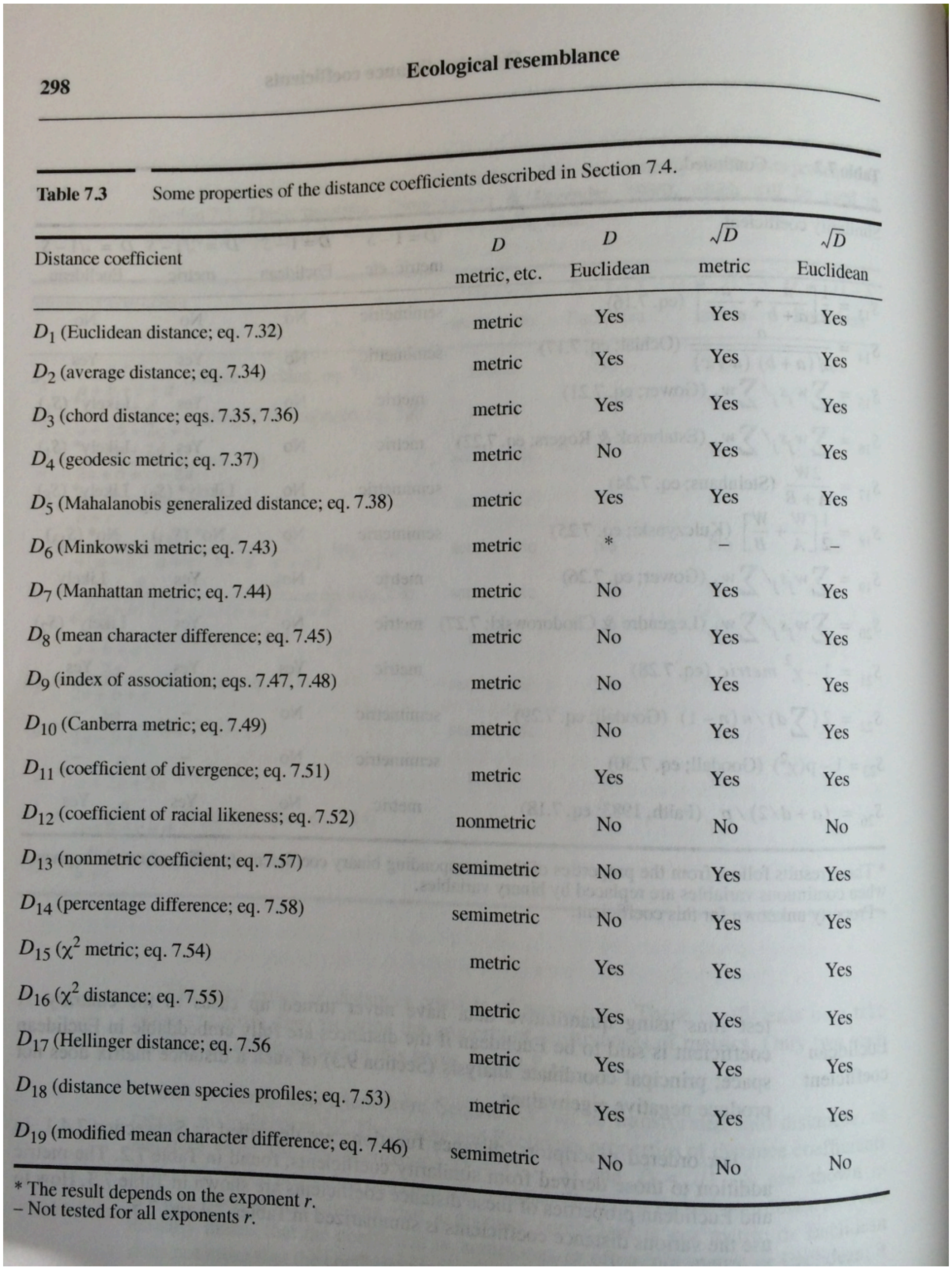


Table 7.3 Some properties of the distance coefficients described in Section 7.4.

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.