4
Multivariate comparison



4.1 Resemblance in multivariate space

When talking about resemblance we address two types of measurement: *similarity*, where high values signify a high proportion of common features, and *distance*, where high values signify dissimilarity. As long as sampling units are described by one species or one site factor only, comparison is straightforward and the operational rules discussed in Section 3.1 on data types are valid. When more attributes exist, the technique is no longer trivial and several questions need to be answered in advance of data analysis:

- Are the attributes of the same type or is treatment necessary?
- Do the attributes have the same weight or is transformation necessary?

Data Analysis in Vegetation Ecology Otto Wildi
© 2010 John Wiley & Sons, Ltd

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

- Are the attributes measured on the same scale or have the scales to be adjusted by transformation?

- Are some of the attributes correlated and therefore partly carrying the same information?

Due to the multivariate nature of data, several attributes or sampling units have to be taken into account simultaneously. A first and really illustrative approach to resemblance is the geometric, where attributes function as axes in a scatter diagram. Attribute scores are therefore the coordinates of the sampling units, which are points located in space (Figure 4.1).

A second way of measuring resemblance is a statistical one, specifically suited for species lists where presence and absence are major issues. The joint occurrences can be counted and statistical measures will help to decide whether the frequency assessed is higher or lower than expected compared to a random situation.

Probably the most common technique is the use of product moments, among which the better known are correlation and covariance. If much of the variance of two sampling units is shared then covariance is high and they are considered similar.

Of course, there are more approaches to the comparison of sampling units or species, such as measures relying on information theory (Rényi 1961, Orlóci 1978). I will not discuss these in the following sections.

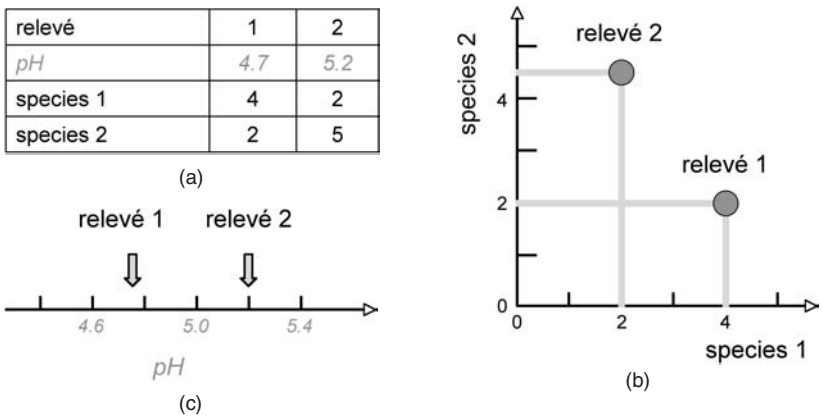

Figure 4.1 Presentation of data in the Euclidean space. The data are shown in (a). In (b), the biological attributes are used to represent the relevés in two-dimensional (biological) space. (c) shows the one-dimensional environmental space.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

4.2 Geometric approach

Multivariate similarity can easily be related to geometry, because geometry considers dimensionality. Geometrical space may be one-dimensional (a straight line), two-dimensional (a surface) or three-dimensional (a volume). The dimensions can also be extended to any number, say four or a hundred.

In practice, there are at least two constraints. First, it is assumed that the dimensions (i.e. the axes) are based on the same scale. Second, the weight of the axes is the same. The latter is *only* the case if the attributes (and hence the axes) are uncorrelated. If the attributes are perfectly correlated then they carry identical information. The use of the multivariate Euclidean space is only justified if the attributes are equally scaled, as is the case when using the Braun-Blanquet code, for example Table 3.3. The principle is shown in Figure 4.1, where the data are given in (a). It is assumed that the pH values are part of the environmental space, whereas the species scores form the biological space. In (b) the two-dimensional biological space is shown, where the relevés are points in the scatter diagram. Whenever comprehensive species lists are used, the biological space is extremely high-dimensional, with each species forming its own dimension. In (c), however, it can be seen that a space may also be one-dimensional only. The relevés are still points, but on a one-dimensional vector, in this case pH.

Resemblance of any two sampling units in Euclidean space is most easily measured as a distance. If the distance is short then any two relevés are similar. If the distance is long, the relevés involved diverge in many possible ways. There are different methods of calculating distance, as shown in Figure 4.2. A straightforward measure is *Euclidean distance*. The Euclidean distance between relevé 1 and relevé 2 is calculated by:

$$De_{1,2} = \sqrt{\sum_{j=1}^{p} (x1_j - x2_j)^2} \tag{4.1}$$

In the left-hand side of Figure 4.2, this is the direct distance between the corresponding data points. Equation (4.1) is written for p species and is therefore valid for any number of dimensions. The lower bound of $De_{1,2}$ is zero for identical relevés; the upper bound has no limit. When the number of dimensions (species) increases, the Euclidean distance tends to become larger.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

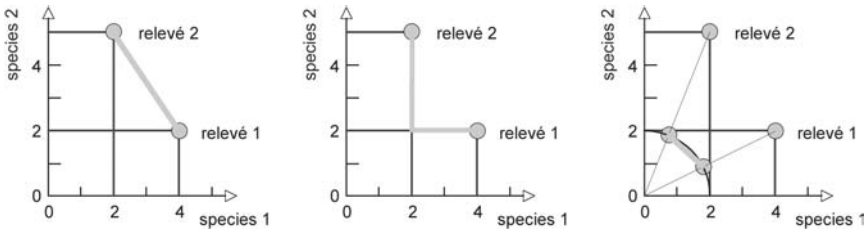

Figure 4.2 Three ways of measuring distance. Left: Euclidean distance. Centre: Manhattan distance. Right: Chord distance.

A second possible measure is *Manhattan distance*, which is the sum of the differences of the scores calculated on all axes, that is:

$$Dm_{1,2} = \sum_{j=1}^{p} |x1_j - x2_j| \tag{4.2}$$

The Manhattan distance (Equation 4.2) has similar properties to the Euclidean. As shown in Figure 4.2, centre, the Manhattan distance is generally somewhat longer than the Euclidean distance.

In some cases, methods differ by the intrinsic transformation applied. *Chord distance* is an example. It is identical to the Euclidean distance, but after normalizing the vectors. Combining these two operations yields the corresponding formula:

$$Dc_{1,2} = \sqrt{2\left(1 - \frac{\sum_{j=1}^{p} x1_j x2_j}{\sqrt{\sum_{j=1}^{p} x1_j^2 \sum_{j=1}^{p} x2_j^2}}\right)} \tag{4.3}$$

Chord distance has a lower bound of zero (for identical relevés or species vectors). Unlike the previous measures, there is now a maximum value of square root of two; that is, 1.414213. This is the case when relevés have no species in common. It is difficult to decide whether the normalizing involved is ideal for applications: when transformation is really needed, many researchers prefer standardization (adjusting vector length and variance) to normalization (adjusting vector length only). This idea will be discussed in the context of the product moment measures (Section 4.4).

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

4.3 Contingency testing

Contingency testing is a statistical approach, focusing on the joint occurrence of objects. In the case of relevés, these are common species. If there are many, one assumes that the relevés are similar. From a statistical point of view the question arises whether the number of common species is above, equal to or below expectation. Hence, we will have to deal with the meaning of 'expectation'.

The standard setup for this type of measurement is the contingency table, as shown in Table 4.1. This explains how relevés are compared. For each species, common occurrence is counted in cell a. When a species occurs in one relevé only, it is counted in either cell b or cell c. If a species occurs in neither relevé 1 nor 2, it contributes to cell d.

The row and column sums yield useful numbers as well. The sum in row $a+b$ is the total number of species found in relevé 2; the sum in column $a+c$ for relevé 1. The sum in row $c+d$ is the number of species that do not occur in relevé 1 and the sum of column $b+d$ the number that do not occur in relevé 2. The grand total Σ is the total number of species considered for calculations, including those occurring in neither relevé 1 nor 2.

Using such counts from contingency tables, an almost unlimited number of coefficients can be calculated. Many of these are listed in Legendre & Legendre (1998), pp. 275−276. They differ in their properties and some are related to other types of resemblance measure. Four of them are shown in Table 4.2.

The *Jaccard coefficient* S_J is the oldest, published in 1901. It counts the number of common species and the total number of species present in either of the two relevés. The range is from zero (no species in common) to one (all species in common). When 50% of the species are common, $S_J = 0.50$.

Table 4.1 Notations in contingency tables. *a*, *b*, *c* and *d* are frequency counts.

	relevé 1		
relevé 2	+	−	
+	a	b	$a+b$
−	c	d	$c+d$
	$a+c$	$b+d$	Σ

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Table 4.2 Resemblance measures using the notations in Table 4.1.

Name	Formula	Distance measure	property
Jaccard	$S_J = \frac{a}{a+b+c}$	$D_J = 1 - S_J$	metric
Soerensen	$S_S = \frac{2a}{2a+b+c}$	$D_S = 1 - S_S$	semimetric
Simple maching	$S_{SM} = \frac{a+d}{a+b+c+d}$	$D_{SM} = 1 - S_{SM}$	metric
Chi squared	$\chi^2 = \left(\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}\right)^2$	$D_{\chi^2} = 1 - \chi^2$	metric

The second is the *Soerensen coefficient* S_S. This differs from the Jaccard coefficient in that common species have double weight. The range is also zero to one, but when 50% of the species are in common, $S_S = 0.667$. The derived distance measure (the complement) is called semimetric, because it may happen that the distance configuration of three or more relevés cannot be presented in Euclidean space (i.e. the triangular unequality is violated), limiting its application in some methods.

In the *Simple maching coefficient* S_{SM}, frequency d is used as well. When analysing a sample, such as a synoptic table (Section 6.6), the total number of species considered remains the same for all pairs of relevés. However, when using different lists of species, S_{SM} differs for the same pair of relevés.

The fourth coefficient is the *Chi squared* (χ^2), as known from statistics. This is the sum of squared differences from the expected frequencies when independence is assumed. The probability distribution of the χ^2 can be found in most statistical textbooks. This allows it to be used for significance tests – as long as data are based on statistical sampling. When analysing vegetation data the χ^2 is rarely used in the statistical sense, but rather as yet another similarity measure with a lower bound of zero and no finite upper bound.

4.4 Product moments

Product moments are a flexible group of measures. They express the degree to which vectors point in the same direction. This conforms with the basic concept of variance (the variance within one vector) and covariance (the variance shared by two vectors). Four related measures that differ in their implicit transformation only are listed in Table 4.3.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Table 4.3 Product moments. Types differ in the mode of implicit data transformation.

Name	Formula	Transformation
Scalar product	$S_{jk} = \sum_{h=1}^{p} A_{hj} A_{hk}$	$A_{hj} = X_{hj}$
Centred scalar product	$S_{jk} = \sum_{h=1}^{p} A_{hj} A_{hk}$	$A_{hj} = X_{hj} - \overline{X}_h$
Covariance	$S_{jk} = \sum_{h=1}^{p} A_{hj} A_{hk}$	$A_{hj} = (X_{hj} - \overline{X}_h)/\sqrt{n-1}$
Correlation	$S_{jk} = \sum_{h=1}^{p} A_{hj} A_{hk}$	$A_{hj} = \frac{(X_{hj} - \overline{X}_h)}{\left(\sum_{e=1}^{n} X_{he} - \overline{X}_h\right)^{1/2}}$

The *scalar product* is the vector product with no further transformation involved. If all scores are positive it ranges from zero to infinity. The more attributes involved, the larger the scalar product.

The *centred scalar product* involves centring of the observational vectors; thus the mean of any vector will be zero. On average, half of the coefficients will be negative with no upper and lower bound.

Covariance does the same thing as the centred scalar product, but in addition it offers a correction for the number of elements, n. It is used in analysis of variance. Note that $n - 1$ corrects for the underestimation of variance in small samples n.

The *product moment correlation coefficient* (termed *correlation* in Table 4.3) standardizes the observational vectors implicitly. Their mean is zero and the standard deviation is equal to one. This has the practical advantage that there are fixed upper and lower bounds: $-1 \leq r \leq +1$. This is shown in Figure 4.3 in the form of a geometrical interpretation. When two vectors show the same trend but in the opposite direction, correlation approaches $cos\alpha \approx -1$. When they are independent, it is around zero. When they point in the same direction it approaches $cos\alpha \approx +1$.

Many standard statistical packages use the correlation coefficient as a default measure for the majority of methods. Thus, measurements taken at different scales become comparable, and the variance is adjusted. If this is not desirable because one expects important information from variance differences then another option should be considered. An example where this frequently is suggested is the comparison of species-rich relevés versus species-poor relevés.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

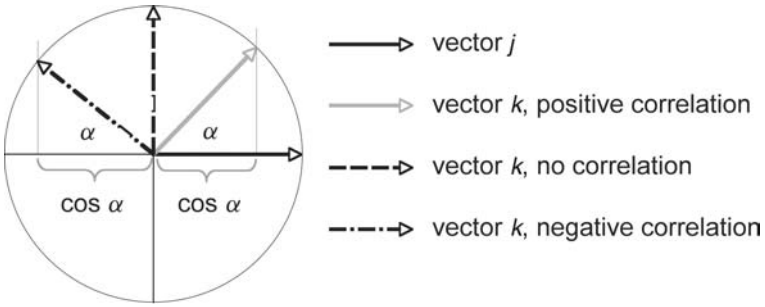

Figure 4.3 The correlation of vector *j* with vector *k*. The correlation coefficient is the cosine of the angle α between any two observational vectors *j* and *k*.

4.5 The resemblance matrix

Whereas pairwise comparison of observational vectors like relevés or species is useful for many purposes, assessing the pattern of an entire sample involves the computation of a resemblance matrix. This is done by comparing all possible pairs of sampling units, resulting in an $n * n$ matrix of resemblance coefficients. Such a matrix (Figure 4.4) is generally symmetric and only the lower-left triangle (or the upper-right) has to be considered. Depending on the resemblance measure used, the diagonal elements, the self-similarity of the sampling units, may be of interest or not. When using Euclidean distance, for example, they are all zero; when using the correlation coefficient they all equal 1.0. When using covariance, however, they carry the variances of the sampling units and these usually vary.

Table 1

	r1	r2	r3
s1	10	15	18
s2	20	30	25
s3	4	5	15
s4	6	8	12

Distance matrix 1

0		
11.4	0	
15.7	12.2	0

$\bar{d} = 13.1$

Table 2

	r1	r2	r3
s1	4	14	20
s2	11	31	41
s3	2	4	24
s4	5	9	17

Distance matrix 2

0		
22.8	0	
42.2	24.4	0

$\bar{d} = 29.8$

Figure 4.4 The average distance of a distance matrix is a perfect measure for homogeneity of a sample. Left: high homogeneity. Right: low homogeneity.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

Resemblance matrices may become very large. When computing the triangular matrix only, without the diagonals, the number of elements is $(n * (n - 1))/2$. This is far too great for immediate interpretation. The matrix therefore has to be processed further with the aim of pattern recognition, by component analysis (Chapter 5), cluster analysis (Chapter 6) or ranking (Section 5.6), for example.

A simple and yet most useful application is shown in Figure 4.4, lower part. The aim is to determine the *homogeneity* of a sample. From the data matrices in the upper row, the distance matrices are calculated and the mean Euclidean distance is computed. This is as a measure of distance or dissimilarity of the total set of relevés. Table 1, with a relatively low average distance, is hence more homogeneous than Table 2.

4.6 Assessing the quality of classifications

Under specific circumstances a resemblance matrix can be used to evaluate group patterns, as shown in Figure 4.5. This is a graphical representation of the similarities within and between 71 forest vegetation types in Switzerland, distinguished by Ellenberg & Klötzli (1972). The underlying data have been reconstructed from the original notes of the authors and the relevés found in the literature (Keller *et al*. 1998). From these 2533 relevés we know the corresponding classification used for definition of the forest types. The coefficients in Figure 4.5 are not just pairwise similarities, but average similarities between all relevés of the 71 groups involved. The diagonal elements are the average similarities within the groups and thus a measure of homogeneity, as explained in Figure 4.4.

Let us first look at some findings concerning the *diagonal* elements. There are examples of vegetation types exhibiting high internal homogeneity: the average similarity of relevés is high and therefore the symbol is large. Typical examples are forest types 49 (Equiseto-Abietetum), 56 (Sphagno-Piceetum typicum) and 70 (Rhododendro ferruginei-Pinetum montanae). The opposite is true for forest types 11 (Aro-Fagetum), 44 (Carici elongatae-Alnetum Glutinosae) and 64 (Cytiso-Pinetum silvestris). When inspecting all diagonal elements it becomes clear that the internal homogeneity of the different vegetation types varies amazingly: large symbols, indicating homogeneous groups, alternate with small symbols, indicating heterogeneity. In practice this means that there are types that are easy to recognize in the field (homogeneous ones) and others that are difficult to recognize (heterogeneous ones).

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)

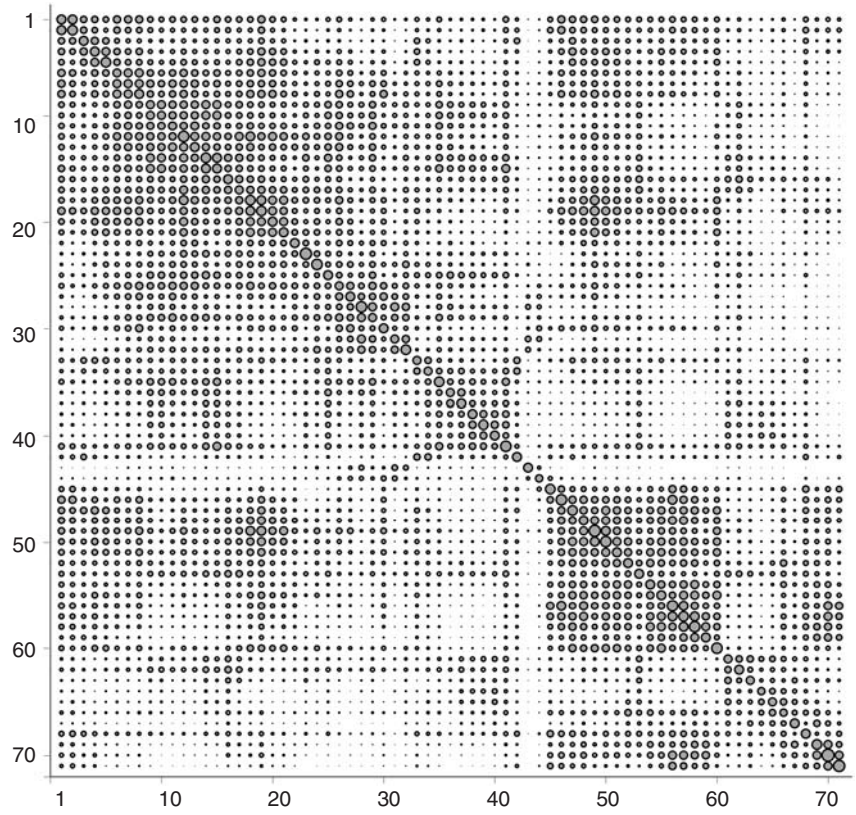

Figure 4.5 Similarities within and between the forest types of Switzerland according to Ellenberg & Klötzli (1972) based on the revision of Keller *et al.* (1998).

The *off-diagonal* elements show which of the vegetation types are difficult to distinguish from others (large symbols) and which are easily differentiated (small symbols). Forest types 1−21 form a block with large off-diagonal symbols. These are beech (*Fagus sylvatica*) forests. Differences in species composition between these types are minor and careful inspection of the species lists is required for proper identification. A similar example is seen in spruce and fir forests (*Picea abies* and *Abies alba*), forest types 45−60. Interestingly, there are also certain forest types which bridge the two blocks when taking species composition into account (types 19 and 49). As can be seen from this example, a similarity matrix presented graphically is an excellent tool for predicting problems in practical applications of classifications such as vegetation mapping. A real-world example is shown in Section 11.5, where the quality of a phytosociological classification system is evaluated.

You created this PDF from an application that is not licensed to print to novaPDF printer (http://www.novapdf.com)