CHAPTER 6

Distance Measures

Background

The first step of most multivariate analyses is to calculate a matrix of distances or similarities among a set of items in a multidimensional space. This is analogous to constructing the triangular "mileage chart" provided with many road maps. But in our case, we need to build a matrix of distances in hyperspace, rather than the two-dimensional map space. Fortunately, it is just as easy to calculate distances in a multidimensional space as it is in a two-dimensional space.

This first step is extremely important. If information is ignored in this step, then it cannot be expressed in the results. Likewise, if noise or outliers are exaggerated by the distance measure, then these unwanted features of our data will have undue influence on the results, perhaps obscuring meaningful patterns.

Distance concepts

Distance measures are flexible:

- Resemblance can be measured either as a distance (dissimilarity) or a similarity.
- Most distance measures can readily be converted into similarities and vice-versa.
- All of the distance measures described below can be applied to either binary (presence-absence) or quantitative data.
- One can calculate distances among either the rows of your data matrix or the columns of your data matrix. With community data this means you can calculate distances among your sample units (SUs) in species space or among your species in sample space.

Figure 6.1 shows two species as points in sample space, corresponding to the tiny data set below (Table 6.1). We can also represent sample units as points in species space, as on the right side of Figure 6.1, using the same data set.

There are many distance measures. A selection of the most commonly used and most effective measures are described below. It is important to know the domain of acceptable data values for each distance measure (Table 6.2). Many distance measures are not compatible with negative numbers. Other distance measures assume that the data are proportions ranging between zero and one, inclusive.

Table 6.1. Example data set. Abundance of two species in two sample units.

<table>
<thead>
<tr>
<th>Sample unit</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 6.1. Graphical representation of the data set in Table 6.1. The left-hand graph shows species as points in sample space. The right-hand graph shows sample units as...
## Chapter 6

### Table 6.2: Reasonable and acceptable domains of input data, $x$, and ranges of distance measures, $d = f(x)$.

<table>
<thead>
<tr>
<th>Name (synonyms)</th>
<th>Domain of $x$</th>
<th>Range of $d = f(x)$</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sørensen (Bray &amp; Curtis; Czekanowski)</td>
<td>$x \geq 0$</td>
<td>$0 \leq d \leq 1$ (or $0 \leq x \leq 100%$)</td>
<td>proportion coefficient in city-block space; semimetric</td>
</tr>
<tr>
<td>Relative Sørensen (Kulczynski; Quantitative Symmetric)</td>
<td>$x \geq 0$</td>
<td>$0 \leq d \leq 1$ (or $0 \leq x \leq 100%$)</td>
<td>proportion coefficient in city-block space; same as Sørensen but data points relativized by sample unit totals; semimetric</td>
</tr>
<tr>
<td>Jaccard</td>
<td>$x \geq 0$</td>
<td>$0 \leq d \leq 1$ (or $0 \leq d \leq 100%$)</td>
<td>proportion coefficient in city-block space; metric</td>
</tr>
<tr>
<td>Euclidean (Pythagorean)</td>
<td>all</td>
<td>non-negative</td>
<td>metric</td>
</tr>
<tr>
<td>Relative Euclidean (Chord distance; standardized Euclidean)</td>
<td>all</td>
<td>$0 \leq d \leq \sqrt{2}$ for quarter hypersphere, $0 \leq d \leq 2$ for full hypersphere</td>
<td>Euclidean distance between points on unit hypersphere; metric</td>
</tr>
<tr>
<td>Correlation distance</td>
<td>all</td>
<td>$0 \leq d \leq 1$</td>
<td>converted from correlation to distance, proportional to arc distance between points on unit hypersphere; cosine of angle from centroid to points; metric</td>
</tr>
<tr>
<td>Chi-square</td>
<td>$x \geq 0$</td>
<td>$d \geq 0$</td>
<td>Euclidean but doubly weighted by variable and sample unit totals; metric</td>
</tr>
<tr>
<td>Squared Euclidean</td>
<td>all</td>
<td>$d \geq 0$</td>
<td>metric</td>
</tr>
<tr>
<td>Mahalanobis</td>
<td>all</td>
<td>$d \geq 0$</td>
<td>distance between groups weighted by within-group dispersion; metric</td>
</tr>
</tbody>
</table>

Distance measures can be categorized as metric, semimetric, or nonmetric. A **metric** distance measure must satisfy the following rules:

1. The minimum value is zero when two items are identical.
2. When two items differ, the distance is positive (negative distances are not allowed).
3. Symmetry: the distance from objects A to object B is the same as the distance from B to A.
4. Triangle inequality axiom: With three objects, the distance between two of these objects cannot be larger than the sum of the two other distances.

Kulczynski distances. Semimetrics are extremely useful in community ecology but obey a non-Euclidean geometry. **Nonmetrics** violate one or more of the other rules and are seldom used in ecology.

### Distance measures

The equations use the following conventions. Our data matrix $A$ has $q$ rows, which are sample units and $p$ columns, which are species. Each element of the matrix, $a_{ij}$, is the abundance of species $j$ in sample unit $i$. Most of the following distance measures can also be used on binary data ($1$ or $0$ for presence or absence). In each of the following equations, we are calculating the distance between sample units $i$ and $h$. 

...
**Distance Measures**

**Euclidean distance**

$$ED_{i,k} = \sqrt{\sum_{j=1}^{p} (a_{i,j} - a_{k,j})^2}$$

This formula is simply the Pythagorean theorem applied to $p$ dimensions rather than the usual two dimensions (Fig. 6.2).

**City-block distance** (= Manhattan distance)

$$CB_{i,k} = \sum_{j=1}^{p} |a_{i,j} - a_{k,j}|$$

In city-block space you can only move along one dimension of the space at a time (Fig. 6.2). By analogy, in a city of rectangular blocks, you cannot cut diagonally through a block, but must walk along either of the two dimensions of the block. In the mathematical space, size of the blocks does not affect distances in the space. Note also that many equal-length paths exist between two points in city-block space.

Euclidean distance and city-block distance are special cases (different values of $k$) of the Minkowski metric. In two dimensions:

$$Distance = \sqrt{x^k + y^k}$$

where $x$ and $y$ are distances in each of two dimensions. Generalizing this to $p$ dimensions, and using the form of the equation for $ED$:

$$Distance_{i,k} = \sqrt{\sum_{j=1}^{p} (a_{i,j} - a_{k,j})^k}$$

Note that $k = 1$ gives city-block distance, $k = 2$ gives Euclidean distance. As $k$ increases, increasing emphasis is given to large differences in individual dimensions.

**Correlation**

The correlation coefficient ($r$) is cosine $\alpha$ (third panel in Fig. 6.2) where the origin of the coordinate system is the mean species composition of a sample unit in species space (the "centroid"; see Fig. 6.2). If the data have not been transformed and the origin is at $(0,0)$, then this is a noncentered correlation coefficient.

For example, if two sample units lie at 180° from each other relative to the centroid, then $r = -1 = \cos(180°)$. Two sample units lying on the same radius from the centroid have $r = 1 = \cos(0°)$. If two sample units form a right angle from the centroid then $r = 0 = \cos(90°)$.

The correlation coefficient can be rescaled to a distance measure of range 0-1 by

$$r_{distance} = (1 - r)^2$$

**Proportion coefficients**

Proportion coefficients are city-block distance measures expressed as proportions of the maximum distance possible. The Sorensen, Jaccard, and QSK coefficients described below are all proportion coefficients. One can represent proportion coefficients as the overlap between the area under curves. This is easiest to visualize with two curves of species abundance on an environmental gradient (Fig. 6.3). If $A$ is the area under one curve, $B$ is the area under the other.
and \( w \) is the overlap (intersection) of the two areas, then the Sorensen coefficient is \( 2w/(A+B) \). The Jaccard coefficient is \( w/(A+B-w) \).

Written in set notation:

**Sorensen similarity**

\[
\text{Sorensen similarity} = \frac{2(A \cap B)}{(A \cup B) - (A \cap B)}
\]

**Jaccard similarity**

\[
\text{Jaccard similarity} = \frac{A \cap B}{A \cup B}
\]

Proportion coefficients as distance measures are foreign to classical statistics, which are based on squared Euclidean distances. Ecologists latched onto proportion coefficients for their simple, intuitive appeal despite their falling outside of mainstream statistics. Nevertheless, Roberts (1986) showed how proportion coefficients can be derived from the mathematics of fuzzy sets, an increasingly important branch of mathematics. For example, when applied to quantitative data, Sorensen similarity is the intersection between two fuzzy sets.

**Sorensen similarity** (also known as "BC" for Bray-Curtis coefficient) is thus shared abundance divided by total abundance. It is frequently known as "2w/(A+B)" for short. This logical-seeming measure was also proposed by Czekanowski (1913). Originally used for binary (0/1) data, it works equally well for quantitative data. It is often called the "Bray-Curtis coefficient" (Faith et al. 1987) when applied to quantitative data, as in Bray and Curtis (1957). Rewriting the equation as a dissimilarity (or distance) measure, dissimilarity between items \( i \) and \( h \) is:

\[
D_{ih} = \frac{\sum j=1^p a_j - a_{jh}}{\sum j=1^p a_j + \sum j=1^p a_{jh}}
\]

Another way of writing this, where MIN is the smaller of two values:

\[
D_{ih} = 1 - \frac{2 \sum j=1^p \text{MIN}(a_j, a_{jh})}{\sum j=1^p a_j + \sum j=1^p a_{jh}}
\]

One can convert this dissimilarity (or any of the following proportion coefficients) to a percentage dissimilarity (PD):

\[
\text{Sorensen distance} = BC_{ih} = PD_{ih} = 100 \cdot D_{ih}
\]

Jaccard dissimilarity is the proportion of the combined abundance that is not shared, or \( w = (A \cdot B - w) \) (Jaccard 1901):

\[
JD_{ih} = \frac{2 \sum j=1^p a_j - a_{jh}}{\sum j=1^p a_j + \sum j=1^p a_{jh}}
\]

**Quantitative symmetric dissimilarity** (also known as the Kulczynski or QSK coefficient; see Faith et al. 1987):

\[
\text{QSK}_{ih} = 1 - \frac{1}{2} \left[ \frac{\sum j=1^p \text{MIN}(a_j, a_{jh})}{\sum j=1^p a_j} + \frac{\sum j=1^p \text{MIN}(a_j, a_{jh})}{\sum j=1^p a_{jh}} \right]
\]

Although Faith et al. (1987) stated that this measure has a "built-in standardization," it is not a standardized city-block distance in the same way that relative Euclidean distance is a standardized measure of Euclidean distance. In contrast with the "relative Sorensen distance" (below), the QSK coefficient gives...
different results with raw data versus data standardized by SU totals (Ch. 9). After such relativization, however, QSK gives the same results as Sorensen, relative Sorensen, and city-block distance.

Relative Sorensen (also known as relativized Manhattan coefficient in Faith et al. 1987) is mathematically equivalent to the Bray-Curtis coefficient on data relativized by SU total. This distance measure builds in a standardization by sample unit totals, each sample unit contributing equally to the distance measure. Using this relativization shifts the emphasis of the analysis to proportions of species in a sample unit, rather than absolute abundances.

\[
D_{\text{rel}} = 1 - \frac{\sum_{j=1}^{p} \text{MIN} \left( \frac{a_{hj}}{\sum_{j=1}^{p} a_{hj}}, \frac{a_{bj}}{\sum_{j=1}^{p} a_{bj}} \right)}{\sum_{j=1}^{p} a_{hj} \sum_{j=1}^{p} a_{bj}}
\]

An alternate version, using an absolute value instead of the MIN function, is also mathematically equivalent to Bray-Curtis coefficient on data relativized by SU total:

\[
D_{\text{ab}} = \frac{1}{2} \sum_{j=1}^{p} \left| \frac{a_{hj}}{\sum_{j=1}^{p} a_{hj}} - \frac{a_{bj}}{\sum_{j=1}^{p} a_{bj}} \right|
\]

Note that with standardization or relativization of the data before application of the distance measure, many of the city-block distance measures become mathematically equivalent to each other. After relativization by sample unit totals, PD (Bray-Curtis) = CB = QSK = Relative Sorensen.

**Relative Euclidean distance (RED)**

RED is a variant of ED that eliminates differences in total abundance (actually totals of squared abundances) among sample units. RED ranges from 0 to the square root of 2 when all abundance values are nonnegative (i.e., a quarter hypersphere). Visualize the SUs as being placed on the surface of a quarter hypersphere with a radius of one (Fig. 6.4). RED builds in a standardization. It puts differently scaled variables on the same footing, eliminating any signal other than relative abundance. Note that the correlation coefficient also accomplishes this standardization, but \(\arccos(r)\) gives the arc distance on the quarter hypersphere, not the chord distance. Also, with the correlation coefficient the surface is a full hypersphere, not a quarter hypersphere, and the center of the hypersphere is the centroid of the cloud of points, not the origin.

**Chi-square distance**

The chi-square distance measure is used in correspondence analysis and related ordination techniques (Chardy et al. 1976, Minchin 1987a). Although Faith et al. (1987) found that this distance measure performed poorly, it is important to be aware of it, since it is the implicit distance metric in some of the more popular ordination techniques (detrended correspondence analysis and canonical correspondence analysis).

Let

\[
a_{bh} = \text{total for sample unit } h \text{ (i.e., } \sum_{j=1}^{p} a_{bj})
\]
then the chi-square distance (Chardy et al. 1976) is
\[ X_{ij}^2 = \sum_{j=1}^{p} \frac{1}{a_{ij}} \left( \frac{a_{ij} - a_{ij}}{a_{ij}} \right)^2 \]

Note that this distance measure is similar to Euclidean distance, but it is weighted by the inverse of the species totals. If the data are prerelativized by sample unit totals (i.e., \( b_j = a_0 a_{ij} \)), then the equation simplifies to:
\[ X_{ij}^2 = \sum_{j=1}^{p} \left( \frac{b_j - b_j}{a_{ij}} \right) \]

The numerator is the squared difference in relative abundance. It is expressed as a proportion of the species total (the denominator) and summed over all species.

Minchin (1987a) offered the following critique of this distance measure:

The appropriateness of Chi-squared distance as a measure of compositional dissimilarity in ecology may be questioned (Faith et al. 1987). The measure accords high weight to species whose total abundance in the data is low. Conversely, it de-emphasizes abundant species. It thus tends to exaggerate the distinctiveness of samples containing several rare species. Unlike the Bray-Curtis coefficient and related measures, Chi-squared distance does not reach a constant, maximal value for sample pairs with no species in common, but fluctuates according to variations in the representation of species with high or low total abundances. These properties of Chi-squared distance may account for some of the distortions observed in DCA ordinations.

**Mahalanobis distance (D^2)**

Mahalanobis distance \( D_{fh}^2 \) is used as a distance measure between two groups (\( f \) and \( h \)). It is commonly used in discriminant analysis and in testing for outliers. If \( \bar{a}_f \) is the mean for the \( f \)th variable in group \( f \), and \( w_i \) is an element from the inverse of the pooled within-groups covariance matrix representing variables \( i \) and \( j \), then

\[ D_{fh}^2 = (n - g) \sum_{i=1}^{p} \sum_{j=1}^{p} w_j \left( \bar{a}_f - \bar{a}_h \right) \left( \bar{a}_f - \bar{a}_h \right) \]

where \( n \) is the number of sample units, \( g \) is the number of groups, and \( i \neq j \). Note that differences are weighted more heavily by \( w_i \) when variables \( i \) and \( j \) are uncorrelated. Thus, Mahalanobis distance corrects for the correlation structure of the original variables (the dimensions of the space). The built-in standardization means that it is independent of the measurement units of the original variables.

In which case in Figure 6.5 are groups \( f \) and \( h \) more distant? Because the Mahalanobis distance inversely weights the distance between centroids by the variance, the two groups are more distant in Case B, even though the centroids are equidistant in the two cases.

Note the conceptual similarity to an \( F \)-ratio of between- to within-group variance. Indeed, the Mahalanobis distance can be used to calculate an \( F \)-test for multivariate differences between groups. Similarly, it can be used to test for outliers by calculating the distance between each point and the cloud of remaining points.

**Performance of distance measures**

**Loss of sensitivity with heterogeneity**

Performance of distance measures can be evaluated by comparing the relationship between environmental distance (distance along an environmental gradient, such as elevation) vs. sociological distance (the difference in communities as reflected by the distance in species space). This method of
evaluating distance measures was used by Beals (1984), Faith et al. (1987), Death (1999a), and Boyce and Ellison (2001). If species respond noislessly to environmental gradients and the environmental gradients are known, then we seek a perfect linear relationship between distances in species space and distances in environmental space. Any departure from that relationship represents a partial failure of our distance measures.

Two examples help clarify the variability in the relationship between distance in species space and environmental space. These examples are based on synthetic data sets with a known underlying structure and noisless responses of species to two environmental gradients.

The first example is an "easy" data set, consisting of 25 sample units and 16 species. It is easy because the beta diversity is fairly low (average Sorensen distance among SUs = 0.59; 1.3 half changes), the sample unit totals are fairly even (coefficient of variation (CV) of SU totals = 17%), and the species are all similarly abundant (CV of species totals = 37%).

Despite this being an easy data set, all of the distance measures show a curvilinear relationship with environmental distance (Fig. 6.6). Specifically, we see the loss in sensitivity of our distance measures at large environmental distances. The problem is least apparent in the Sorensen, chi-square, and Jaccard distances. The problem is worst with the correlation distance, where the curve not only flattens at high environmental distances, but starts to decline at the highest distances. The drop in the curve for the correlation coefficient (actually \((1 - r) / 2\) which converts \(r\) into a distance rather than a similarity measure) is due to interpreting shared zeros (0.0) as positive association.

The second example is a more difficult data set, consisting of 100 sample units and 25 species. The difficulty has nothing to do with the size of the data set. Rather its beta diversity is higher (average Sorensen distance among SUs = 0.79; 2.3 half changes), the sample unit totals vary more widely than in the previous example (CV of SU totals = 40%), and the species vary realistically in abundance (CV of species totals = 183%).

Again, all of the distance measures lose sensitivity with increasing environmental distance (Fig. 6.7). This loss is greatest for distance based on the correlation coefficient. Euclidean distance not only loses sensitivity at high distances, but introduces considerable error even at moderate distances. Note also that Euclidean distance shows no fixed upper bound for sample units that have nothing in common.

Sorensen distance loses sensitivity over a distance about half the length of the environmental gradients. The flat top on the Sorensen scatterplot results because it has a fixed maximum for SUs having no species in common. Many ecologists consider this a desirable, intuitive property for species data. Chi-square distance performs reasonably well at small environmental distances but misinterprets many distant SUs as being close in species space.

Transformation of the data to binary (presence-absence) in both examples results in a more linear relationship with most distance measures. This was shown for Euclidean distance and Sorensen distance with real data from an elevation gradient (Beals 1984).

Given the apparently poor performance of all of the distance measures, it is remarkable that multivariate analysis is able to extract clear, sensible patterns (Fig. 6.8). We are rescued by the redundancy in the data — all ordination and classification techniques benefit from this redundancy in the data. Where two species fail to be informative of a difference, another two species are informative. Some ordination techniques, nonmetric multidimensional scaling (NMS) in particular, are able to linearize the relationship between distance in species space and distance in a reduced ordination space. NMS has an advantage over other ordination techniques: it is based on ranked distances, which improves its ability to extract information from the nonlinear relationships illustrated in the two examples.
Figure 6.6. Relationship between distance in species space for an “easy” data set, using various distance measures and environmental distance. The graphs above are based on a synthetic data set with noiseless species responses to two known underlying environmental gradients. The gradients were sampled with a 5 x 5 grid. This is an “easy” data set because the average distance is reasonably small (Sorensen distance = 0.59, 1.3 half changes), all species are similar in abundance (CV of species totals = 37%), and sample units have similar totals (CV of SU totals = 17%).
Figure 6.7. Relationship between distance in species space for a "more difficult" data set, using various distance measures, and environmental distance. The graphs above are based on a synthetic data set with noiseless species responses to two known underlying environmental gradients. The gradients were sampled with a 10 x 10 grid. This is a "more difficult" data set because the average distance is rather large (Sorensen distance = 0.79, 2.3 half changes), species vary in abundance (CV of species totals = 183%), and sample units have moderately variable totals (CV of SU totals = 40%).
Chapter 6

Availability

The Sorensen (Bray-Curtis) index has repeatedly been shown to be one of the most effective measures of sample or species similarity, yet it is not widely available in general statistical software.

Compatibility

The primary disadvantage of city-block distance measures is that they are not compatible with many standard multivariate analyses (e.g., discriminant analysis, canonical correlation, and canonical correspondence analysis). This becomes less and less important, however, as non-Euclidean alternatives become increasingly available. Certain methods of classification and ordination are effective with ecological data (e.g., Bray-Curtis ordination and non-metric multidimensional scaling), in part because they are amenable to distance measures that perform well with ecological data.

Theoretical basis

Beyond the choice of a proportion coefficient or not (Box 6.1) and the choice of a relativized distance measure or not, there is little basis in ecological theory for selecting one distance measure over another. The rationale for our choice is primarily empirical: we should select measures that have shown superior performance, based on the other criteria listed here. One important theoretical difference between Euclidean and city-block distance is, however, apparent. Long Euclidean distances in species space are measured through an uninhabitable portion of species space — in other words the straight-line segments tend to pass through areas of impossibly species-rich and overly full communities. In contrast, city-block distances are measured along the edges of species space — exactly where sample units lie in the dust bunny distribution.

Intuitive criteria

Does Euclidean or city-block distance better match our intuition on how community distances should be measured? Consider the following examples.

In city-block space, the importance of a gradient is proportional to the number of species responding to it. For example, assume that 20 species respond only to gradient X and 4 species only to gradient Y. In city-block space, gradient X is 5 times as important as gradient Y. In Euclidean space, gradient X is square-
Distance Measures

root-of-5 times as important as gradient Y. Which space matches your intuition?

With Euclidean distance, large differences are weighted more heavily than several small differences (Box 6.2). This results in greater sensitivity to outliers with Euclidean distance than with city-block distance measures. For example, assume we have four species and three sample units, A, B, and C. The data and differences in abundance of each species for each pair of sample units are listed in Box 6.2.

**Geodesic distance**

Performance of all of the traditional distance measures declines as distances in species space increase (Figs. 6.7 and 6.8). An innovative solution to the problem of measuring long distances in nonlinear structures is the geodesic distance (Tenenbaum et al. 2000). This concept is similar to the “shortest path” adjustments to a distance matrix (Williamson 1978, 1983; Clymo 1980; Bradfield & Kenkel 1987; De’ath 1999a). Williamson summed distances between sample unit pairs representing the shortest path between two distant SUs, but only applied this to SUs with no species in common. Bradfield and Kenkel (1987) added flexibility by varying the threshold for the number of species in common. Bradfield and Kenkel found better results with a lower threshold; i.e., adjusting a larger proportion of the distance matrix. De’ath (1999a) further extended the method by using city-block distance measures, changing the threshold to a quantitative dissimilarity value, and allowing multiple-step paths between very distant SUs.

A geodesic distance between two points is measured by accumulating distances between nearby points. Tenenbaum et al. (2000) used Euclidean distances, but geodesic distances can be built from other distance measures. “Nearby” can be defined as a fixed radius or as the $n$ nearest neighbors. Geodesic distances should be able to find effectively the curvature of compositional gradients in species space. Geodesic distances are one of the most promising new methodological developments. A key issue will be objectively defining “nearby” to optimize the recovery of patterns in ecological communities.

The difference between Tenenbaum’s geodesic distance and the ecologists’ shortest path (SP) methods can be visualized with an analogy to crossing a stream dotted with stepping stones. We want to find the shortest route from a particular point on one bank to a particular point on the opposite bank. The SP method must find a single stepping stone that gets us across the stream in the two shortest possible leaps (one to the stepping stone and one to the far bank). The geodesic method, however, defines a comfortably small step, then seeks the shortest series of steps without ever exceeding that small step length. The geodesic method thus considers the whole array of stepping stones, while the SP method can consider only one stone at a time.

A problem with the SP method is that if the stream is broader than two leaps, then no single stepping stone will work. This corresponds to two SUs so different that there is no third SU that shares species with both of them. De’ath (1999a) solved this problem by allowing multiple passes of the SP method, in essence allowing multiple stepping stones.

Despite the excellent ordinations in Bradfield and Kenkel (1987), Boyce and Ellison (2001), and De’ath (1999a), the geodesic distances and related methods have not been widely adopted, probably because they have not been included in popular software packages. Whether Tenenbaum et al.’s (2000) geodesic distances offer further improvements over the SP methods used by ecologists remains to be seen.
Box 6.1. Comparison of Euclidean distance with a proportion coefficient (Sørensen distance). Relative proportions of species 1 and 2 are the same between Plots 1 and 2 and Plots 3 and 4.

<table>
<thead>
<tr>
<th>Data matrix containing abundances of two species in four plots</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Plot 1</td>
</tr>
<tr>
<td>Plot 2</td>
</tr>
<tr>
<td>Plot 3</td>
</tr>
<tr>
<td>Plot 4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example calculations of distance measures for Plots 3 and 4. ED = Euclidean distance. PD = Sørensen distance as percentage</th>
</tr>
</thead>
</table>
| $ED_{3,4} = \sqrt{(10-10)^2 + (10-0)^2} = 10$  
| $PD_{3,4} = \frac{100(|10-10| + |10-0|)}{10+20} = 33.3\%$ |

<table>
<thead>
<tr>
<th>Sørensen Distance matrix, expressed as percentages.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plot 1</td>
</tr>
<tr>
<td>Plot 1</td>
</tr>
<tr>
<td>Plot 2</td>
</tr>
<tr>
<td>Plot 3</td>
</tr>
<tr>
<td>Plot 4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Euclidean distance matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plot 1</td>
</tr>
<tr>
<td>Plot 1</td>
</tr>
<tr>
<td>Plot 2</td>
</tr>
<tr>
<td>Plot 3</td>
</tr>
<tr>
<td>Plot 4</td>
</tr>
</tbody>
</table>

The Sørensen distance between Plots 1 and 2 is 0.333 (33.3%), as is the Sørensen distance between Plots 3 and 4, as illustrated below. In both cases the shared abundance is one third of the total abundance. In contrast, the Euclidean distance between Plots 1 and 2 is 1, while the Euclidean distance between Plots 3 and 4 is 10. Thus the Sørensen coefficient expresses the shared abundance as a proportion of the total abundance, while Euclidean distance is unconcerned with proportions.
Box 6.2. Example data set comparing Euclidean and city-block distances, contrasting the effect of squaring differences versus not.

Hypothetical data: abundance of four species in three sample units (SU).

<table>
<thead>
<tr>
<th>Sp</th>
<th>SU 1</th>
<th>SU 2</th>
<th>SU 3</th>
<th>SU 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>C</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Sample units A,B: species differences $d = 1, 1, 1, 9$ for each of the four species.
Sample units A,C: species differences $d = 3, 3, 3, 3$

<table>
<thead>
<tr>
<th>Distance Measure</th>
<th>Pair of SUs</th>
<th>Euclidean</th>
<th>City-block</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>9.165</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>AC</td>
<td>6</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

The simple sum of differences (city-block distance) is the same for AB and AC. Euclidean distance sums the squared differences, so that difference of 9 is given more emphasis with Euclidean distance than with city-block distance. Thus, the Euclidean distance between A and B is larger than the distance between A and C. The city-block distance between A and B is the same as that between A and C. Which distance measure matches your intuition?