Chapter 5

Measures of distance between samples: non-Euclidean

Euclidean distances are special because they conform to our physical concept of distance. But there are many other distance measures which can be defined between multivariate samples. These non-Euclidean distances are of different types: some still satisfy the basic axioms of what mathematicians call a metric, while others are not even metrics but still make very good sense as a measure of difference between samples in the context of certain data. In this chapter we shall consider several non-Euclidean distance measures that are popular in the environmental sciences: the Bray-Curtis dissimilarity, the L_1 distance (also called the city-block or Manhattan distance) and the Jaccard index for presence-absence data. We also consider how to measure dissimilarity between samples for which we have heterogeneous data.

Contents

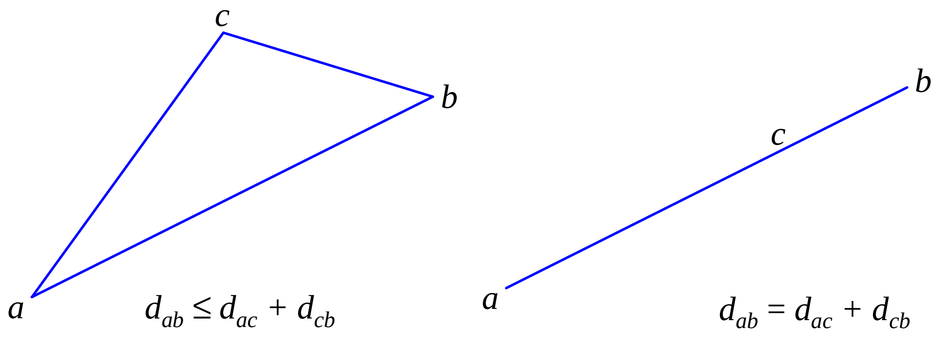
The axioms of distance

In mathematics, a true measure of distance, called a *metric*, obeys three properties. These metric axioms are as follows, where d_{ab} denotes the distance between objects *a* and *b*:

1. $d_{ab} = d_{ba}$
2. $d_{ab} \geq 0$ and $= 0$ if and only if $a = b$
3. $d_{ab} \leq d_{ac} + d_{ca}$ (5.1)

The first two axioms are trivial: the first says that the distance from *a* to *b* is the same is from *b* to *a*, in other words the measure is symmetric; the second says that distances are always positive except when the objects are identical, in which case the distance is necessarily 0. The third axiom, called the *triangle inequality*, may also seem intuitively obvious but is the more difficult one to satisfy. If we draw a triangle *abc* in our Euclidean world, for example in Exhibit 5.1, then it is obvious that the distance from *a* to *b* must be shorter than the sum of the distances via another point *c*, that is from *a* to *c* and from *c* to *b*. The triangle inequality can only be an equality if *c* lies exactly on the line connecting *a* and *b* (see right hand sketch in Exhibit 5.1).

Exhibit 5.1 Illustration of the triangle inequality for distances in Euclidean space.



But there many apparently acceptable measures of distance that do not satisfy this property: with those it would be theoretically possible to get a 'route' from *a* to some point *c* and then from *c* to *b* which is shorter than from *a* to *b* 'directly'. Because these are not true distances (in the mathematical sense) they are sometimes called *dissimilarities*.

Bray-Curtis dissimilarity

When it comes to ecological abundance data collected at different sampling locations, the Bray-Curtis dissimilarity is one of the most well-known ways of quantifying the difference between samples. This measure appears to be very reasonable way of achieving this goal but it does not satisfy the triangle inequality axiom, and hence is not a true distance (we shall discuss the implications of this in later chapters when we analyze Bray-Curtis dissimilarities). To illustrate its definition, we consider again the count data for the last two samples of Exhibit 1.1, which we recall here:

	a	b	c	d	e	sum
s29	11	0	7	8	0	26
s30	24	37	5	18	1	85

On of the assumptions of the Bray-Curtis measure is that the samples are taken from the same physical size, be it area or volume. This is because dissimilarity will be computed on raw counts, not on relative counts, so the fact that there is higher overall abundance at site s30 is part of the difference between these two samples – that is, 'size' and 'shape' of the count vectors will be taken into account in the measure[1].

The computation involves summing the absolute differences between the counts and

[1] In fact, the Bray-Curtis dissimilarity can be computed on relative abundances, as we did for the chi-square distance, to take into account only 'shape' differences – this point is discussed later.

dividing this by the sum of the abundances in the two samples:

$$b_{\mathrm{s29,s30}} = \frac{|11-24|+|0-37|+|7-5|+|8-18|+|0-1|}{26+85} = \frac{63}{111} = 0.568$$

The general formula for calculating the *Bray-Curtis dissimilarity* between samples *i* and *i′* is as follows, supposing that the counts are denoted by n_{ij} and that their sample (row) totals are n_{i+} :

$$b_{ii'} = \frac{\sum_{j=1}^{J} | n_{ij} - n_{i'j} |}{n_{i+} + n_{i'+}} \qquad (5.2)$$

This measure takes on values between 0 (samples identical: $n_{ij} = n_{i'j}$ for all *j*) and 1 (samples completely disjoint; that is, when there is a nonzero abundance of a species in one sample, then it is zero in the other: $n_{ij} > 0$ implies $n_{i'j} = 0$) – hence it is often multiplied by 100 and interpreted as a percentage. Exhibit 5.2 shows part of the Bray-Curtis dissimilarities between the 30 samples (the caption points out a violation of the triangle inequality):

Exhibit 5.2 Bray-Curtis dissimilarities, multiplied by 100, between the 30 samples of Exhibit 1.1, based on the count data for taxa ***a*** to ***e***. Violations of the triangle inequality can be easily picked out: for example, from s25 to s4 Bray-Curtis is 93.9, but the sum of the values ‘via s6’ from s25 to s6 and from s6 to s4 is 18.6+69.2 = 87.8, which is shorter!

	s1	s2	s3	s4	s5	s6	· · ·	s24	s25	s26	s27	s28	s29
s2	45.7												
s3	29.6	48.1											
s4	46.7	55.6	46.7										
s5	47.7	34.8	50.8	78.6									
s6	52.2	22.9	52.2	69.2	41.9								
s7	45.5	41.5	49.1	87.0	21.2	50.9							
·	·	·	·	·	·	·	·						
·	·	·	·	·	·	·	· ·						
·	·	·	·	·	·	·	· · ·						
s25	70.4	39.3	66.7	93.9	52.9	18.6	· · ·	46.4					
s26	69.6	32.8	60.9	92.8	41.7	15.2	· · ·	39.3	13.7				
s27	63.6	38.1	63.6	93.3	38.2	21.5	· · ·	42.6	16.3	22.6			
s28	32.5	21.5	50.0	57.7	31.9	29.5	· · ·	30.9	41.8	47.5	34.4		
s29	43.4	35.0	43.4	54.5	31.2	53.6	· · ·	39.8	64.5	58.2	61.2	34.2	
s30	60.7	36.7	58.9	84.5	48.0	21.6	· · ·	40.8	18.1	25.3	23.6	37.7	56.8

If the Bray-Curtis dissimilarity is subtracted from 100, a measure of *similarity* is obtained, called the Bray-Curtis index. For example, the similarity between sites s25 and s4 is 100 – 93.9 = 6.1%, which is the lowest amongst the values displayed in Exhibit 5.2; whereas the highest similarity is for sites s25 and s26: 100–13.7 = 86.3%. Checking back to the data in Exhibit 1.1 one can verify the similarity between sites s25 and s26, compared to the lack of similarity between s25 and s4.

Bray-Curtis dissimilarity versus chi-square distance

An ecologist would like some recommendation on whether to use Bray-Curtis or chi-square on a particular data set. It is not possible to make any absolute statement of which is preferable, but we can point out some advantages and disadvantages of each one. The advantage of the chi-square distance is that it is a true metric, while the Bray-Curtis dissimilarity violates the triangle inequality, which is slightly problematic when we come to analyzing them later. The advantage of Bray-Curtis is that the scale is easy to understand: 0 means the samples are exactly the same, while 100 is the maximum difference that can be observed between two samples. The chi-square, on the other hand, has a maximum which depends on the marginal weights of the data set, and it would be difficult to assign any substantive meaning to any particular value. Also, a zero chi-square means that the relative abundances are identical, not the original abundances. As pointed out in the footnote on page 5-2, we could calculate Bray-Curtis dissimilarities on the relative abundances (although conventionally the calculation is on raw counts), and in addition we could calculate chi-square distances on the raw counts, without 'relativizing' them (although conventionally the calculation is on relative counts). This would make the comparison between the two approaches fairer.

So we calculated Bray-Curtis on the relative counts and chi-square on the raw counts – Exhibit 5.3 shows parts of the four distance matrices, where the values in each triangular matrix have been strung out columnwise (the column 'site pair' shows which pair corresponds to the values in the rows). The scatterplots of the two comparable sets of measures are shown in Exhibit 5.4. Two features of these plots are immediately apparent: first, there is much better agreement between the two approaches when the counts have been relativized (plot (b)); and second, when the counts are in their raw form (plot (a)) one can obtain 100% dissimilarity for the Bray-Curtis corresponding to a whole range of chi-square distances, from approximately 5 to 16 (see points above the tic-mark of 100 on the axis *B-C raw*). This means that the measurement of shape is fairly similar in both measures, but the way they take size into account is quite different. A good illustration of this is the measure between samples s1 and s17, which have counts as follows (taken from Exhibit 1.1):

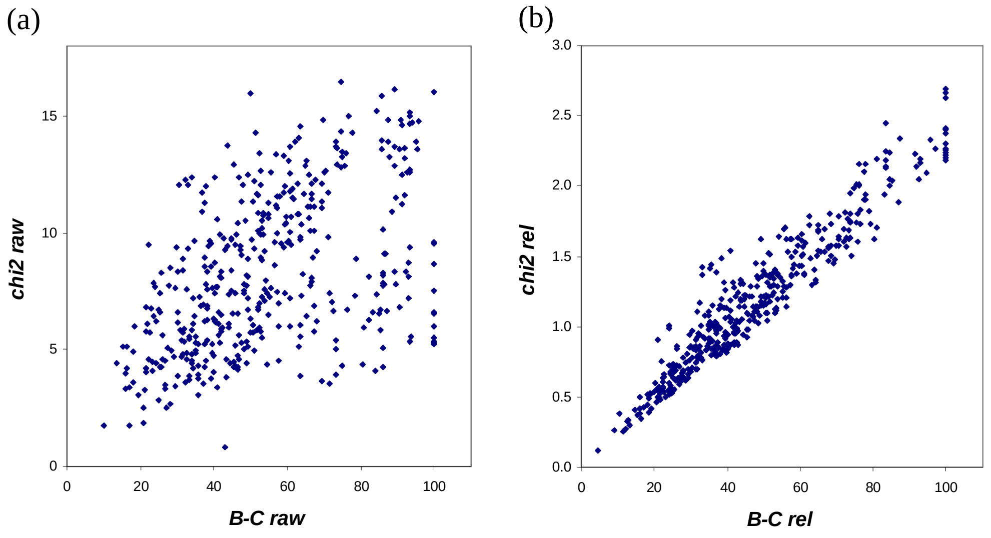
	a	b	c	d	e	sum
s1	0	2	9	14	2	27
s17	4	0	0	0	0	4

The Bray-Curtis dissimilarity is 100% because the two sets of counts are disjoint, whereas the chi-square distance is a fairly low 5.533 (see row (s17,s1) of Exhibit 5.3). This is because the absolute differences between the two sets are not large. If they were larger, say if we doubled both sets of counts, then the chi-square distance would increase accordingly whereas the Bray-Curtis would remain at 100%. It is by considering examples like these that researchers will obtain a feeling for the properties of these measures, in order to be able to choose the measure that is most appropriate for their own data.

Exhibit 5.3 Various dissimilarities and distances between pairs of sites (count data from Exhibit 1.1). ***B-C-raw***: Bray Curtis dissimilarities on raw counts (usual definition and usage), ***chi2 raw***: chi-square distances on raw counts, ***B-C rel***: Bray-Curtis dissimilarities on relative counts, ***chi2 rel***: chi-square distances on relative counts (usual definition and usage).

site pair	*B-C raw*	*chi2 raw*	*B-C rel*	*chi2 rel*
(s2,s1)	45.679	7.398	48.148	1.139
(s3,s1)	29.630	3.461	29.630	0.855
(s4,s1)	46.667	4.146	50.000	1.392
(s5,s1)	47.692	5.269	50.975	1.093
(s6,s1)	52.212	10.863	53.058	1.099
(s7,s1)	45.455	4.280	46.164	1.046
(s8,s1)	93.333	5.359	92.593	2.046
(s9,s1)	33.333	5.462	40.741	0.868
(s10,s1)	40.299	6.251	36.759	0.989
(s11,s1)	35.714	4.306	36.909	1.020
(s12,s1)	37.500	5.213	39.762	0.819
(s13,s1)	57.692	5.978	59.259	1.581
(s14,s1)	63.265	5.128	59.091	1.378
(s15,s1)	20.755	1.866	20.513	0.464
(s16,s1)	85.714	13.937	80.960	1.700
(s17,s1)	100.000	5.533	100.000	2.258
(s18,s1)	56.897	11.195	36.787	0.819
(s19,s1)	16.923	1.762	11.501	0.258
(s20,s1)	33.333	3.734	31.987	0.800
:	:	:	:	:
(s23,s22)	34.400	7.213	25.655	0.688
(s24,s22)	61.224	9.493	35.897	0.897
(s25,s22)	23.567	7.855	25.801	0.617
s(24,s23)	34.177	4.519	16.401	0.340
s(25,s23)	37.681	11.986	37.869	1.001
(s25,s24)	56.757	13.390	44.706	1.142

Exhibit 5.4 Graphical comparison of Bray-Curtis dissimilarities and chi-square distances for (a) raw counts, taking into account size and shape, and (b) relative counts, taking into account shape only.



L_1 distance (city-block)

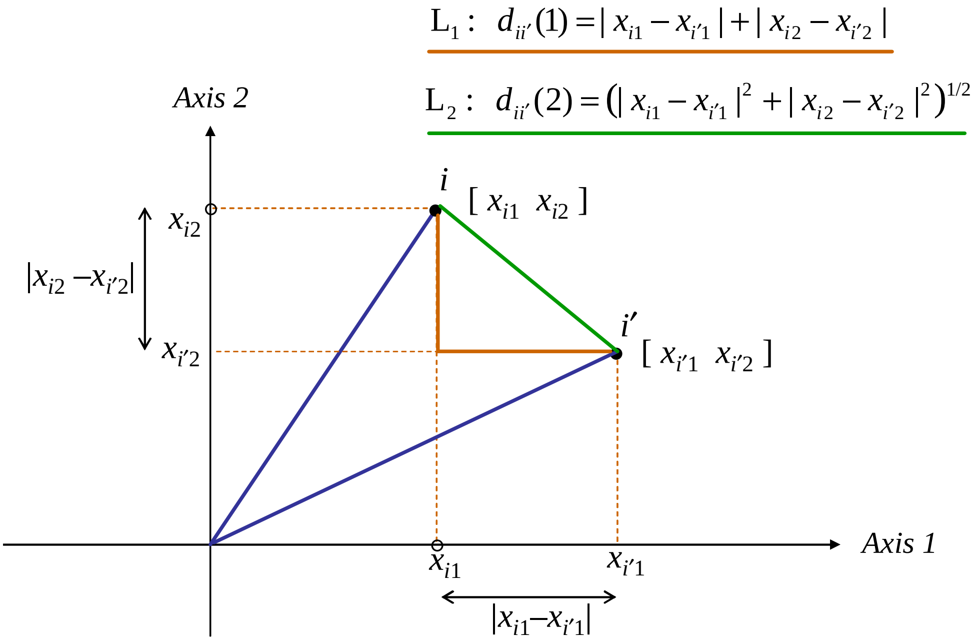
When the Bray-Curtis dissimilarity is applied to relative counts, that is, row profiles with values which can be denoted as $r_{ij} = n_{ij} / n_{i+}$, the row sums r_{i+} in the denominator of (5.2) are 1 for every row, so that the dissimilarity reduces to:

$$b_{ii'} = \frac{1}{2}\sum_{j=1}^{J} | r_{ij} - r_{i'j} | \qquad (5.3)$$

The sum of absolute differences between two vectors is called the L_1 distance, or city-block distance. This is a true distance function since it obeys the triangle inequality, and as can be seen in the right hand scatterplot of Exhibit 5.4, agrees fairly well with the chi-square distance for the data under consideration. The reason why it is called the city-block distance, and also as the Manhattan distance or taxicab distance, can be seen in the two-dimensional illustration of Exhibit 5.5. Going from a point A to a point B is achieved by walking ‘around the block’, compared to the Euclidean ‘straight line’ distance. The city-block and Euclidean distances are special cases of the L_p distance, defined here between rows of a data matrix **X** (the Euclidean distance is obtained for $p = 2$):

$$d_{ii'}(p) = \left(\sum_{j=1}^{J} | x_{ij} - x_{i'j} |^p \right)^{1/p} \qquad (5.4)$$

Exhibit 5.5 Two-dimensional illustration of the L_1 (city-block) and L_2 (Euclidean) distances between two points *i* and *i′*: the L_1 distance is the sum of the differences in the coordinates, while the L_2 distance is the square root of the sum of squared differences.



$L_1: \; d_{ii'}(1) = | x_{i1} - x_{i'1} | + | x_{i2} - x_{i'2} |$
$L_2: \; d_{ii'}(2) = (| x_{i1} - x_{i'1} |^2 + | x_{i2} - x_{i'2} |^2)^{1/2}$
Axis 2
Axis 1
i
$[\, x_{i1} \;\; x_{i2} \,]$
i'
$[\, x_{i'1} \;\; x_{i'2} \,]$
x_{i2}
$x_{i'2}$
$|x_{i2} - x_{i'2}|$
x_{i1}
$x_{i'1}$
$|x_{i1} - x_{i'1}|$

Dissimilarity measures for presence–absence data

In Chapter 4 we considered the matching coefficient and the chi-square distance for categorical data in general, but there is a special case which is often of interest to ecologists: presence–absence, or dichotomous, data. When categorical variables have only two categories, there are a host of coefficients defined to measure inter-sample difference (see Bibliographical Appendix for references to this topic). Here we consider one example which is an alternative to the matching coefficient.

Exhibit 5.6 gives some data that we shall use again (in Chapter 7), concerning the presence–absence of 10 species in 7 samples. The distance based on the matching coefficient is obtained either by counting the matches or mismatches between the two samples. For example, between samples A and B there are 6 matches and 4 mismatches. Usually expressed relative to the number of variables (species) this would give a similarity value of 0.6 and and a dissimilarity value of 0.4. But often in ecology it is possible to have very many species in the data set, up to 100 or more, and in each sample we find relatively few of these present. This makes the number of matches based on the co-absence of species very high compared to those based on co-presence. If co-absence is not really so important compared to co-presence, we can simply ignore the co-absences and calculate similarity in terms of co-presences. Furthermore, this co-presence count is expressed not relative to the total number of species but relative to the number of species present in at least one of the two samples under consideration. This is the definition of the *Jaccard index* for dichotomous data. Taking samples A and B of Exhibit 5.6 again, the number of co-presences is 4, we ignore the 2 co-absences, then we express 4 relative to 8, so the result is 0.5. In effect, the Jaccard index is the matching coefficient of similarity calculated for a pair of samples after eliminating all the species which are co-absent (0 and 0). The dissimilarity between two samples is – as before – 1 minus the similarity.

Exhibit 5.6 Presence–absence data of 10 species in 7 samples.

Samples	*Species*									
	sp1	*sp2*	*sp3*	*sp4*	*sp5*	*sp6*	*sp7*	*sp8*	*sp9*	*sp10*
A	1	1	1	0	1	0	0	1	1	1
B	1	1	0	1	1	0	0	0	0	1
C	0	1	1	0	1	0	0	1	0	0
D	0	0	0	1	0	1	0	0	0	0
E	1	1	1	0	1	0	1	1	1	0
F	0	1	0	1	1	0	0	0	0	1
G	0	1	1	0	1	1	0	1	1	0

Here's another example, for samples C and D. This pair has 4 co-absences, so we eliminate them. To get the dissimilarity we can count the mismatches – in fact, all the rest are mismatches – so the dissimilarity is 6/6 = 1, the maximum that can be attained. Using the Jaccard approach we would say that samples C and D are completely different, whereas the matching coefficient would lead us to a dissimilarity of 0.6 because of the 4 matched co-absences.

To formalize these definitions, the counts of matches and mismatches in a pair of samples are put into a 2×2 table as follows:

		Sample 2		
		1	0	
Sample 1	1	a	b	$a+b$
	0	c	d	$c+d$
		$a+c$	$b+d$	$a+b+c+d$

where a is the count of co-presences (1 and 1), b the count of mismatches where sample 1 has value 1 but sample 2 has value 0, and so on. The overall number of matches is $a+d$, and mismatches $b+c$. The two measures of distance/dissimilarity considered so far are thus defined as:

Matching coefficient distance: $$\frac{b+c}{a+b+c+d} = 1 - \frac{a+d}{a+b+c+d} \qquad (5.5)$$

Jaccard index dissimilarity: $$\frac{b+c}{a+b+c} = 1 - \frac{a+d}{a+b+c} \qquad (5.6)$$

To give one final example, the correlation coefficient can be used to measure the similarity between two vectors of dichotomous data, and can be shown to be equal to:

$$r = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \qquad (5.7)$$

Hence, a dissimilarity can be defined as $1-r$. Since $1-r$ has a range from 0 (when $bc=0$, no mismatches) to 2 (when $ad=0$, no matches), a convenient measure between 0 and 1 is ½ $(1-r)$.

Distances for heterogeneous data

When a data set contains different types of variables and it is required to measure inter-sample distance, we are faced with another problem of standardization: how can we balance the contributions of these different types of variables in an equitable way? We will demonstrate two alternative ways of doing this. Here's an example of mixed data (shown here are the data for four stations out of a set of 33 – we shall analyze the whole data set later in this book):

	Continuous variables			*Discrete variables*	
Station	Depth	Temperature	Salinity	Region	Substrate
s3	30	3.15	33.52	Ta	Si/St
s8	29	3.15	33.52	Ta	Cl/Gr
s25	30	3.00	33.45	Sk	Cl/Sa
⋮	⋮	⋮	⋮	⋮	⋮
s84	66	3.22	33.48	St	Cl

Apart from the three continuous variables, depth, temperature and salinity there are the categorical variables sampled region (with four regions: Tarehola, Skognes, Njosken and Storura), and substrate character (which can be any selection of clay, silt, sand, gravel or stone). The fact that more than one substrate category can be selected implies that each category is a separate dichotomous variable, so that substrate consists of five different variables.

The first way of standardizing the continuous against the discrete variables is called *Gower's generalized coefficient of dissimilarity*. First we express the discrete variables as dummies and calculate the means and standard deviations of all variables in the usual way:

	Continuous variables			*Sampled region*				*Substrate character*				
Station	*Depth*	*Temperature*	*Salinity*	*Tarehola*	*Skognes*	*Njosken*	*Storura*	*Clay*	*Silt*	*Sand*	*Gravel*	*Stone*
s3	30	3.15	33.52	1	0	0	0	0	1	0	0	1
s8	29	3.15	33.52	1	0	0	0	1	0	0	1	0
s25	30	3.00	33.45	0	1	0	0	1	0	1	0	0
:	:	:	:	:	:	:	:	:	:	:	:	:
s84	66	3.22	33.48	0	0	0	1	1	0	0	0	0
mean	58.15	3.086	33.50	0.242	0.273	0.242	0.242	0.606	0.152	0.364	0.182	0.061
s.d.	32.45	0.100	0.076	0.435	0.452	0.435	0.435	0.496	0.364	0.489	0.392	0.242

Notice that dichotomous variables (such as the substrate categories) are coded as a single dummy variable, not two, while polychotomous variables such as region are split into as many dummies as there are categories. The next step is to standardize each variable and multiply all the columns corresponding to dummy variables by $1/\sqrt{2} = 0.7071$, a factor which compensates for their 0/1 coding:

	Continuous variables			*Sampled region*				*Substrate character*				
Station	*Depth*	*Temperature*	*Salinity*	*Tarehola*	*Skognes*	*Njosken*	*Storura*	*Clay*	*Silt*	*Sand*	*Gravel*	*Stone*
s3	-0.868	0.615	0.260	1.231	-0.426	-0.394	-0.394	-0.864	1.648	-0.526	-0.328	2.741
s8	-0.898	0.615	0.260	1.231	-0.426	-0.394	-0.394	0.561	-0.294	-0.526	1.477	-0.177
s25	-0.868	-0.854	-0.676	-0.394	1.137	-0.394	-0.394	0.561	-0.294	0.921	-0.328	-0.177
:	:	:	:	:	:	:	:	:	:	:	:	:
s84	0.242	1.294	-0.294	-0.394	-0.426	-0.394	1.231	0.561	-0.294	-0.526	-0.328	-0.177

Now distances are calculated between the stations using either the L_1 (city-block) or L_2 (Euclidean) metric. For example, using the L_1 metric and dividing the sum of absolute differences by the total number of variables (12 in this example), the distances between the above four stations are given in the left hand table of Exhibit 5.8. Because the L_1

Exhibit 5.8 Distances between four stations based on the L_1 distance between their standardized and rescaled values, as described above. The distances are shown equal to the part due to the continuous variables plus the part due to the categorical variables.

$$\text{TOTAL DISTANCE} \quad = \quad \text{DISTANCE CONT. VARS} \quad + \quad \text{DISTANCE CAT. VARS}$$

$$\begin{array}{c|cccc} & s3 & s8 & \cdots & s25 \\ \hline s8 & 0.677 & & & \\ s25 & 1.110 & 0.740 & & \\ \vdots & \vdots & \vdots & \ddots & \\ s84 & 0.990 & 0.619 & \cdots & 0.689 \end{array} = \begin{array}{c|cccc} & s3 & s8 & \cdots & s25 \\ \hline s8 & 0.674 & & & \\ s25 & 0.910 & 0.537 & & \\ \vdots & \vdots & \vdots & \ddots & \\ s84 & 0.795 & 0.421 & \cdots & 0.386 \end{array} + \begin{array}{c|cccc} & s3 & s8 & \cdots & s25 \\ \hline s8 & 0.003 & & & \\ s25 & 0.200 & 0.203 & & \\ \vdots & \vdots & \vdots & \ddots & \\ s84 & 0.195 & 0.198 & \cdots & 0.303 \end{array}$$

distance decomposes into parts for each variable, we can show the part of the distance due to the continuous variables, and the part due to the categorical variables. Generally, the categorical variables are contributing more to the differences between the stations, but the differences in the continuous variables is actually small if one looks at the original data; except for the distance between s84 and s25, where there is a bigger difference in the continuous variables, then they contribute almost the same (0.303) as the categorical ones (0.386).

Exhibit 5.8 suggests the alternative way of combining different types of variables: first compute the distances which are the most appropriate for each set and then add them to one another. For example, suppose there are three types of data, a set of continuous variables, a set of categorical variables and a set of percentages or counts. Then compute the distance or dissimilarity matrices $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}_3$ appropriate to each set of homogeneous variables, and then combine these in a weighted average:

$$\mathbf{D} = \frac{w_1\mathbf{D}_1 + w_2\mathbf{D}_2 + w_3\mathbf{D}_3}{w_1 + w_2 + w_3} \qquad (5.8)$$

Weights are a subjective but convenient inclusion because there might be substantive reasons for down-weighting the distances for one set of variables, which might not be so important, or might suffer from high measurement error, for example.

SUMMARY: Measures of distance between samples: non-Euclidean

1. A well-defined distance function obeys the triangle inequality, but there are several justifiable measures of difference between samples which do not have this property: to distinguish these from true distances we often refer to them as dissimilarities.
2. The Bray-Curtis dissimilarity is frequently used by ecologists to quantify differences between samples based on abundance or count data. This measure is usually applied to raw abundance data, but can be applied to relative abundances just like the chi-square distance. The chi-square distance can also be applied to the original abundances to include overall size differences in the distance measure.
3. The sum of absolute differences, or L_1 distance (or city-block distance), is an alternative to the Euclidean distance: an advantage of this distance is that it decomposes into contributions made by each variable (for the L_2 Euclidean distance, we would need to decompose the squared distance).
4. A dissimilarity measure for presence–absence data is based on the Jaccard index, where co-absences are eliminated from the calculation, otherwise the measure resembles the matching coefficient.
5. Distances based on heterogeneous data can be computed after a process of standardization of all variables, using the L_1 or L_2 distances. Alternatively, distance matrices can be calculated for each set of homogeneous variables and then these matrices can be linearly combined, optionally with user-defined weights.