10 Measures of association

10.1 Introduction

In a multivariate dataset, more than one response variable can be analysed at the same time. In Chapter 4, we used Argentinean zoobenthic data where multiple species were measured at multiple sites with several explanatory variables measured at each sites. Possible underlying questions are as follows:

1. What are the relationships between the response variables (species)?
2. What are the relationships between response variables and explanatory variables?
3. What are the relationships between the explanatory variables?
4. What are the relationships between the observations (e.g., sites)? Are there differences between groups of observations (e.g., differences between Autumn and Spring data, or between the transects)?
5. Are there groups of species behaving similar?

Ordination and clustering are typically used to answer these questions. The reason we start by discussing measures of association is that both ordination and clustering techniques start by calculating a measure of association between the observations, or between the response variables. There is a wide range of choices of measures of association (see for example Legendre and Legendre 1998 or Jongman et al. 1995) and whichever measure of similarity is used will strongly affect the outcome of the analysis.

Once you have read this chapter, we strongly advise that you read Chapter 7 in Legendre and Legendre (1998), as we will closely follow it in Section 10.2. However, their chapter is more detailed (and technical) and has more measures of association. In Chapters 4 and 28 of this book, a zoobenthic dataset from Argentina is used. Here, we will use the same data to illustrate measures of association, but to keep the numerical output simple, we use totals per transect. The resulting data are given in Table 10.1. Three transects were sampled in Spring and Autumn giving six rows of data, but in this chapter we ignore the seasonal information. Formulated differently, we ignore the fact that sites 1 and 4, 2 and 5 and 3 and 6 are physically the same. See Chapter 28 for a detailed (and proper) statistical analysis of these data.

Table 10.1. Totals per transect for the Argentinean zoobenthic dataset. Transects were sampled in Autumn (labeled as 1, 2 and 3) and Spring (labeled as 4, 5 and 6).

Transect	Laeonereis acuta	Heteromastus similis	Uca uruguayensis	Neanthes succinea
1	407	79	0	0
2	769	139	87	1
3	44	429	0	22
4	654	108	0	0
5	563	189	110	17
6	84	327	0	63

Throughout this chapter, we discuss the statistical techniques from an ecological point of view. For example, we will talk about species measured at sites. However, all the methods can equally well be applied on other data, for example, financial or medical data. The best way to visualise the data is to imagine a spreadsheet where the rows correspond to observations (sites) and the columns to variables (species). In most of the ecological examples in this chapter, the species are the response variables and the sites are observations. In Table 10.1 we do not have sites but transects, but to avoid confusion we will just call them sites.

The first fundamental question that you have to address is whether you are interested in relationships between species or sites. All the measures of association that are to come can be divided into so-called Q and R analysis. Q analysis is used to define association between sites (objects, observations) and R analysis between species (descriptors, variables). This is a bit of an ecological thing; in some scientific fields, researchers may never have heard of it, or yet in other fields you may receive a lot of criticisms if you apply a Q analysis to define association between species, or worse, an R analysis to define association between sites. We start with Q analysis.

10.2 Association between sites: Q analysis

Once you have decided that interest is on similarity between sites, the next question is equally important: What about the zeros, and especially the double zeros? It is particularly important to know how a chosen technique treats double zeros and larger (or extreme) values. Double zeros refer to the situation in which there are many zeros in two rows. Suppose that observations on 10 species were made at two sites (S_1 and S_2). An artificial example is given below:

S_1: 0 0 0 0 0 0 0 2 2 1
S_2: 0 0 0 0 0 4 5 0 0 1

In some measures of association, the joint absence (double zeros) contributes to similarity. Other measures of association ignore the double zeros and instead focus on the joint presence. This is quite a crucial difference! A species not present at two sites may be due to environmental stress, and in this case, the sites should be

similarly labelled. On the other hand, rare species or poor experimental design will result in lots of zeros and you do not want to say that two sites are similar because a rare species is not present in any of them. You need to decide (based upon your ecological knowledge) what to do with this. Once you have decided whether double zeros should have an influence, you need to choose a measure of association that indeed achieves this. This choice requires the knowledge of symmetrical and asymmetrical coefficients.

In a symmetrical coefficient, the zeros and non-zeros are treated in the same way. Three examples are the simple matching coefficient (S_1), the coefficient of Rogers and Tanimoto (S_2) and a coefficient that gives more weight to joint presence and joint absence (S_3). The original references can be found in Chapter 7 in Legendre and Legendre (1998). To avoid confusion we used the same notation.

Suppose we want to define the association between sites 1 and 2, sites 1 and 3 and sites 2 and 3 in Table 10.1. The starting point of these three coefficients is a simple 2-by-2 matrix showing the number of joint presence (a), observation unique to site 1 (b), unique to site 2 (c), and the joint absence (d) at both sites for each combination of sites. The values for a, b, c and d for various sites are given in Table 10.2.

The indices S_1, S_2 and S_3 between two sites are defined by:

$$S_1 = \frac{a+d}{a+b+c+d} \qquad S_2 = \frac{a+d}{a+2b+2c+d} \qquad S_3 = \frac{2a+2d}{2a+b+c+2d}$$

The simple matching coefficient S_1 between sites 1 and 2 is $2/(2 + 0 + 2 + 0) = 2/4$. Hence, the simple matching coefficient considers the data as presence–absence, and takes into account joint absence (it is used in the formula via d). Computer software can be used to calculate an index for every possible combination of sites.

S_1, S_2 and S_3 are similarity coefficients; the larger the value the more similar. Software for dimension reduction techniques like non-metric multidimensional scaling (Chapter 15) is typically applied on dissimilarity coefficients. A conversion can be used to change a similarity coefficient into a dissimilarity coefficient, for example:

$$D = 1 - S \quad \text{or} \quad D = \sqrt{1 - S}$$

We applied the first option. Table 10.3 shows the simple matching coefficient (S_1) and coefficient of Rogers and Tanimoto (S_2) for the Argentinean zoobenthic data. There is not much difference in the ecological conclusions obtained by these two coefficients (at least not for these data).

Table 10.2. Values unique to sites, and joint presence and absence for the Argentinean data.

		Site 2	
		present	absent
Site 1	present	2 (=a)	0 (=b)
	absent	2 (=c)	0 (=d)

		Site 3	
		present	absent
Site 1	present	2 (=a)	0 (=b)
	Absent	1 (=c)	1 (=d)

		Site 3	
		present	absent
Site 2	present	3 (=a)	1 (=b)
	absent	0 (=c)	0 (=d)

Table 10.3. Simple matching coefficient and coefficient of Rogers and Tanimoto. The smaller the value the more similar are the two sites.

	Simple Matching Coefficient						Coefficient of Rogers and Tanimoto					
	1	2	3	4	5	6	1	2	3	4	5	6
1	0	0.50	0.25	0.00	0.50	0.25	0	0.67	0.40	0.00	0.67	0.40
2		0	0.25	0.50	0.00	0.25		0	0.40	0.67	0.00	0.40
3			0	0.25	0.25	0.00			0	0.40	0.40	0.00
4				0	0.50	0.25				0	0.67	0.40
5					0	0.25					0	0.40
6						0						0

The crucial point with S_1, S_2 and S_3 is that double zeros contribute towards similarity. For example, both S_1 and S_2 indicate that sites 1 and 3 and 2 and 3 are equally similar; $S_1 = 0.25$ for both combinations and for S_2 we have 0.4. Yet, sites 2 and 3 have three species in common and sites 1 and 3 only two. It is the joint zero of *U. uruguayensis* that is causing this. If this makes ecological sense for your data, then you are OK.

We now discuss asymmetrical coefficients. In these coefficients double zeros do not contribute towards similarity. The Jaccard coefficient is defined by

$$S_7 = \frac{a}{a+b+c}$$

Again, we used the same notation as in Legendre and Legendre (1998). The Jaccard index is also called the coefficient of community. A slightly modified coefficient, the Sørensen coefficient (S_8) can be defined by giving more weight to joint presence by using the following formula:

$$S_8 = \frac{2a}{2a+b+c}$$

The Jaccard index and the Sørensen coefficient treat the data as presence–absence data. The Sørensen coefficient is also called the Dice index, critical success index and, meteorology, the threat score. Note that both S_7 and S_8 do not use the joint zeros (d), hence the name asymmetrical. Table 10.4 gives the Jaccard and Sørensen coefficients for the zoobenthic species. There are minor differences between them. However, S_7 (and S_8) for sites 1 and 3, and sites 2 and 3 are now different. The latter combination is more similar, which is what you would expect based on ecology (three species in common versus two).

Table 10.4. Jaccard coefficient and Sørensen coefficient. The smaller the value, the more similar are the two sites.

	Jaccard Coefficient						Sørensen Coefficient					
	1	2	3	4	5	6	1	2	3	4	5	6
1	0	0.50	0.33	0.00	0.50	0.33	0	0.33	0.20	0.00	0.33	0.20
2		0	0.25	0.50	0.00	0.25		0	0.14	0.33	0.00	0.14
3			0	0.33	0.25	0.00			0	0.20	0.14	0.00
4				0	0.50	0.33				0	0.33	0.20
5					0	0.25					0	0.14
6						0						0

Both the Jaccard and Sørensen coefficients treat the data as presence/absence data. We now discuss a series of asymmetrical coefficients that take into account the quantitative aspect of the data. The first one is the similarity ratio (SR). Its mathematical formulation between two species Y and X is given by

$$SR(X,Y) = \frac{\sum_k Y_k X_k}{\sum_k Y_k^2 + \sum_k X_k^2 - \sum_k Y_k X_k}$$

where the index k refers to species. Note that double zeros ($Y_k = X_k = 0$) do not contribute towards the coefficient (the product of two zeros is zero), hence why it is asymmetrical. For presence/absence data, the similarity ratio gives exactly the same results as the Jaccard coefficient. Yet another one is the percentage similarity index. For 0/1 data, the percentage similarity is identical to the Sørensen or Dice coefficient. For other types of data, it takes into account the quantitative aspect of data. Its mathematical formulation between two sites Y and X is given by

$$S_{17} = 2 \times \frac{\sum_k \min(Y_k, X_k)}{\sum_k Y_k + \sum_k X_k}$$

where the index k refers to species. Other names for this coefficient are the Bray–Curtis coefficient and the Czekanowski coefficient. Just like the similarity ratio, it ignores double zeros (hence it is asymmetrical). If two sites have no species in common, then the coefficient is equal to zero. Table 10.5 shows the similarity and the Bray–Curtis coefficients for the six sites of the Argentinean data. There are no spectacular differences between them, but there are small differences in ecological conclusions compared with the Jaccard and Sørensen coefficients.

Table 10.5. Similarity ratio and Bray–Curtis distance for the Argentinean data. The smaller the value the more similar are the two sites.

	Similarity Ratio						Bray–Curtis					
	1	2	3	4	5	6	1	2	3	4	5	6
1	0	0.30	0.83	0.18	0.17	0.74	0	0.34	0.75	0.22	0.29	0.66
2		0	0.87	0.04	0.09	0.82		0	0.75	0.13	0.16	0.70
3			0	0.86	0.76	0.09			0	0.76	0.64	0.19
4				0	0.07	0.81				0	0.18	0.69
5					0	0.70					0	0.57
6						0						0

We now move towards a more controversial measure of association, the Euclidean distances. Suppose we have counts for three species (A, B and C) from five sites (1-5); see Table 10.6. If we consider the species as axes, the five samples can be plotted in a three-dimensional space (Figure 10.1). The Euclidean distance between two sites i and j is calculated by

$$D_1 = \sqrt{\sum_{k=1}^{3}(X_k - Y_k)^2}$$

This is just Pythagoras. Y_k is the abundance of species k at site Y. It is easy to show that the Euclidean distance between sites one and two is the square root of 24, and between sites three and four it is the square root of 13. The smaller the D_1, the more similar the two sites. Hence, D_1 indicates that sites three and four are more similar than sites one and two, and this does make sense if you look at the three-dimensional graph. However, sites three and four do not have any species in common, whereas sites one and two have at least one species in common: species A. This shows that for certain types of data, the Euclidean distance function is not the best tool to use, unless you want to identify species or sites with large values or outliers as these will cause a large D_1.

The Euclidean distances between the six sites for the Argentinean data are given in Table 10.7. Again, we obtain a different ecological interpretation. The Euclidean distance between site 3 and all other sites (except for 6) are among the highest and this is because of the high values of *L. acuta*. Actually, all values in Table 10.8 are mainly driven by abundances of *L. acuta* and in lesser context *H. similis* due to the high values of these species, large variation and definition of D_1. The Euclidean distance is sensitive to large values and outliers.

Table 10.6. Artificial example of numbers of three species (A, B and C) measured at five sites (1-5).

	1	2	3	4	5
A	1	5	0	0	3
B	0	2	3	0	2
C	2	0	0	2	3

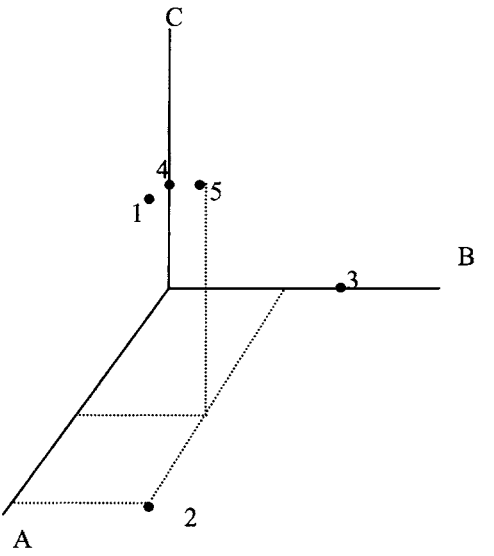

Figure 10.1. Three-dimensional graph of abundance at five sites of three species.

Table 10.7. Euclidean distances between the six sites for the Argentinean data. The smaller the value, the more similar are the two sites.

	1	2	3	4	5	6
1	0	377.11	504.73	248.70	220.96	412.07
2		0	785.96	147.50	213.83	718.32
3			0	689.66	582.31	116.98
4				0	165.02	613.87
5					0	512.54
6						0

As shown in the previous paragraph, absolute numbers influences the Euclidean distance function. To reduce this effect, the Orchiai coefficient (S_{14}) or Chord distance (D_3) can be used. To obtain the Orchiai coefficient, imagine a line from the origin to each site in Figure 10.1. The Orchiai coefficient between two sites is then the angle between the lines of the corresponding sites and can be calculated using simple geometry. The Chord distance between two sites is obtained by drawing a unit (= length 1) sphere around the origin in Figure 10.1, and calculating distances between the intersect points (these are the points where the sphere and the lines intercept). The Orchiai and Chord distances for the Argentinean data are presented in Table 10.8.

Table 10.8. Orchiai coefficient and Chord distance for the Argentinean data. The smaller the value the more similar are the two sites.

	Orchiai Coefficient						Chord Distance					
	1	2	3	4	5	6	1	2	3	4	5	6
1	0	0.29	0.18	0.00	0.29	0.18	0	0.11	1.19	0.03	0.23	1.08
2		0	0.13	0.29	0.00	0.13		0	1.20	0.11	0.16	1.09
3			0	0.18	0.13	0.00			0	1.22	1.09	0.20
4				0	0.29	0.18				0	0.24	1.10
5					0	0.13					0	0.97
6						0						0

We now look at two more measures of association that we use in later chapters: the Manhattan distance (also called taxicab or city-block distance function) and Whittaker's index of association. The Manhattan distance between two sites X and Y containing M species is defined by

$$D_7 = \sum_{k=1}^{M} |X_k - Y_k|$$

This is the sum of the absolute difference between X_k and Y_k, and it has the same problems as the Euclidean distance. The names Manhattan, taxicab and city-block indicate that this function has something to do with the real distance that a taxi would make if it drives around a city-block (distance in street 1 plus the distance in street 2, plus the distance in street 3, etc.). Indeed, this is how it calculates distance between two sites: the difference between species 1 and 2 at site A, plus the difference between species 1 and 2 at site B, etc.

The Whittaker index of association between two sites X and Y is given by

$$D_9 = \frac{1}{2} \sum_{k=1}^{M} \left| \frac{X_k}{\sum_k X_k} - \frac{Y_k}{\sum_k Y_k} \right|$$

With this index two sites are compared with each other, using the differences in proportions. The proportions are taken with respect of the total species at a site. To illustrate the mechanics of this index, suppose we have the abundances of five species at two sites:

X: 1 4 2 2 1
Y: 0 0 5 5 10

And we wish to calculate the Whittaker index of association between the two sites. The totals at the two sites are 10 and 20. The proportions and differences are:

X:	0.1	0.4	0. 2	0.2	0.1
Y:	0.0	0.0	0.25	0.25	0.5
|X − Y|:	0.1	0.4	0.05	0.05	0.4

Adding up the absolute differences, and multiplying by 0.5 gives $D_9 = 0.5$. It is well suited for species abundance (Legendre and Legendre 1998). The Manhattan and Whittaker index for the Argentinean data are given in Table 10.9. Note that the Manhattan distance is driven by abundant species.

Table 10.9. Manhattan distance and Whittaker index for the Argentinean data. The smaller the value, the more similar are the two sites.

	Manhattan Distance						Whittaker Index					
	1	2	3	4	5	6	1	2	3	4	5	6
1	0	510	735	276	393	634	0	0.09	0.75	0.02	0.20	0.66
2		0	1123	234	295	1022		0	0.77	0.09	0.13	0.68
3			0	953	874	183			0	0.77	0.68	0.18
4				0	299	852				0	0.22	0.68
5					0	773					0	0.59
6						0						0

10.3 Association among species: R analysis

We now discuss ways of defining association between two variables. Technically, we have to write that there are population variables y and x, and a sample of N (paired) observations $Y_1, X_1, Y_2, X_2, \ldots, Y_N, X_N$ is taken. We will define the association between y and x, and use the Y_i and X_i to calculate it. Again, there is the problem of double zeros.

Assume a sample of N observations with two variables y and x, for example the number of the zoobenthic species *L. acuta* (= Y) and mud content (= X). The structure of the sample data is as follows:

	Y	X
Observation 1	Value	Value
Observation 2	Value	Value
...	...	...
...	...	...
Observation N	Value	Value

The question we now wish to address is whether there is a *linear* relationship between y and x. Two obvious tools to analyse this are the covariance and correlation coefficients. Both determine how much the two variables covary (vary together): If the first variable increases/decreases, does the second variable increase/decrease as well? Mathematically, the (population) covariance between two random variables y and x is defined as

$$\text{cov}(y,x) = E[(y - E[y])(x - E[x])]$$

where $E[]$ stands for expectation. The (population) variance of y is defined as the covariance between y and y. If we take a sample of N observations $Y_1, X_1,\ldots,Y_N, X_N$, the sample covariance is calculated by:

$$cov(Y,X)=\frac{1}{N-1}\sum_{j=1}^{N}(Y_i-\bar{Y})(X_i-\bar{X})$$

The bars above Y and X indicate mean values. As an example, we have calculated the covariance among the four zoobenthic species from the Argentinean data (Table 10.1). The diagonal elements in the left side of Table 10.10 are the (sample) variances. Note that *L. acuta* has a rather large variance, which makes the comparison of covariance terms difficult. Although not relevant here, another problem with the covariance coefficient is that it depends on the original units. If for example, a weight variable is expressed in grams instead of kilos, one will find a larger covariance. The correlation coefficient standardises the data and takes values between –1 and 1, and is therefore a better option to use. The (population) correlation coefficient between two random variables y and x is defined by

$$cor(y,x)=\frac{\mathrm{cov}(y,x)}{\sigma_y\sigma_x}$$

where σ_y and σ_x are the population standard deviations of y and x, respectively. If we have a sample of N observations $Y_1, X_1,\ldots, Y_N, X_N$, the sample correlation is calculated by

$$cor(Y,X)=\frac{1}{N-1}\sum_{i=1}^{N}\frac{(Y_i-\bar{Y})}{s_Y}\frac{(X_i-\bar{X})}{s_X} \qquad (10.1)$$

s_y and s_x are the sample standard deviations of Y and X respectively. The correlation coefficients among the same four zoobenthic species are given in the right part of Table 10.10; *L. acuta* and *H. similis* have the highest (negative) correlation.

Table 10.10. Covariance and correlation coefficients among the four zoobenthic species from the Argentinean data. The abbreviations LA, HS, UU and NS refer to *L. acuta*, *H. similis*, *U. uruguayensis* and *N. succinea*. The higher the covariance and correlation coefficient, the more similar are the two species.

	Covariance Coefficients				Correlation Coefficients			
	LA	HS	UU	NS	LA	HS	UU	NS
LA	90289	−34321	9212	−5335	1	−0.83	0.6	−0.73
HS		18935	−1770	2314		1	−0.25	0.69
UU			2640	−285			1	−0.23
NS				595				1

In this dataset there are lots of zeros. For example, the percentage of observations with zero abundance for each of the four species in the original data is as fol-

lows: L. *acuta*: 10%, *H. similis*: 28%, *U. uruguayensis*: 75% and *N. succinea*: 67%. The problem with the correlation and covariance coefficients is that they may classify two species, which are absent at the same sites, as highly correlated.

The formula in equation (10.1) is the so-called Pearson (sample) correlation coefficient. In most textbooks, this is just called *the (sample) correlation coefficient.* Sometimes, the phrase product-moment is added because it is a product of two terms and involves the first and second moment (mean and variance). The Pearson correlation coefficient measures the strength of the linear relationship between two random variables y and x. To estimate the population Pearson correlation coefficient, observed data Y_1, X_1,..., Y_N, X_N are used, and therefore the estimator is called the *sample correlation coefficient*, or the Pearson sample correlation coefficient. It is a statistic and has a sample distribution. If you repeat the experiment m times, you end up with m estimations of the correlation coefficient. The most commonly used null hypothesis for the Pearson population correlation coefficient is: H_0: $\text{cor}(y,x) = 0$. If H_0 is true, there is no linear relationship between y and x. The correlation can be estimated from the data using equation (10.1), and a t-statistic or p-value can be used to test H_0. The underlying assumption for this test is that y and x are bivariate normally distributed. This means that both y and x need to be normally distributed. If either y or x is not normally distributed, the joint distribution will not be normally distributed neither. Graphical exploration tools can be used to investigate this. There are three options if non-normality is expected: Transform one or both variables, use a more robust measure of correlation, or do not use a hypothesis test.

More robust definitions for the correlation coefficient can be used if the data are non-normal, non-bivariate normal, a transformation does not help, or if there are non-linear relationships. One robust correlation function is the Spearman rank correlation coefficient, which is applied on rank transformed data. The process is explained in Table 10.11. The first two columns show data of two artificial variables Y and X. In the last two columns, each variable has been ranked. Originally the variable Y had the values 6 2 8 3 1. The first value, 6, is the fourth smallest value. Hence, it rank value is 4. Ranking all values results in 4 2 5 3 1. The same process is applied on X. Spearman's rank correlation coefficient is obtained by calculation the Pearson correlation coefficient between the ranked values of Y and X.

The correlation and covariance coefficients only detect monotonic relationships and not non-monotonic, non-linear relationships. It is therefore useful to inspect the correlation coefficients before and after a data transformation.

Table 10.11. Artificial data for *Y* and X. The first two columns show the original data and the last two columns the ranked data.

	Original Data			Ranked Data	
Sample	*Y*	*X*	Sample	*Y*	*X*
1	6	10	1	4	4
2	2	7	2	2	3
3	8	15	3	5	5
4	3	3	4	3	5
5	1	−5	5	1	1

Chi-square distance

Suppose we are interested in the numbers of parasites on fish from four different locations in three different years (Table 10.12). The underlying questions are whether there is any association between location and years (e.g., are there particular areas and years where more parasites were measured), and whether a Chi-square test is the most appropriate test. The Chi-square statistic is calculated by:

$$X^2 = \sum \frac{(O-E)^2}{E}$$

where O represents the observed value and E the expected value. For the data in Table 10.12, the following steps are carried out to calculate the Chi-square statistic. First the null hypothesis is formulated:

H_0: there is no association between the rows and columns in Table 10.12.

If rows and columns are indeed independent, the expected number of parasites in area A in 1999 is 1254 × (272/1254) × (567/1254) = 122.909. Table 10.13 shows all the observed (in normal font) and expected values (in italic font). The $(O - E)^2/E$ values are also given in this table. These are the values in italic font and bold. The Chi-square statistic is equal to 20.77 (the sum of all values in bold in Table 10.13). The degrees of freedom for the statistic is equal to the number of rows minus one, multiplied with the number of columns minus 1. This information can be used to work out a p-value, which is $p = 0.002$. Hence, the null hypothesis is very unlikely. Based on the values in Table 10.13, the highest contribution to the Chi-square test was from area A in 2001, and area A in 1999. This information can be used to infer that parasites in area A were different in those two years. In later chapters, correspondence analysis is used to visualise this.

Table 10.12. Numbers of parasites in 4 areas and 3 years.

Area	1999	2000	2001	Total
A	147	55	70	272
B	98	50	107	255
C	183	75	157	415
D	139	50	123	312
Total	567	230	457	1254

Table 10.13. Observed (normal font), expected values (italic font) and contribution of each individual cell to the Chi-square statistic (italic font in bold) for the parasites in fish.

Area	1999			2000			2001		
A	147	*122.99*	***4.69***	55	*49.89*	***0.52***	70	*99.12*	***8.56***
B	98	*115.30*	***2.60***	50	*46.77*	***0.22***	107	*92.93*	***2.13***
C	183	*187.64*	***0.11***	75	*76.1 2*	***0.02***	157	*151.24*	***0.22***
D	139	*141.07*	***0.03***	50	*57.22*	***0.91***	123	*113.70*	***0.76***

The Chi-square statistic can be used to define similarity among variables (e.g., species). The Chi-square distance between two variables Z_1 and Z_2 is defined by

$$D(Z_1,Z_2)=\sqrt{Z_{++}}\sqrt{\sum_{j=1}^{M}\frac{1}{Z_{+j}}\left(\frac{Z_{1j}}{Z_{1+}}-\frac{Z_{2j}}{Z_{2+}}\right)^2}$$

Z_{ij} is the abundance of the first species at site j, and a '+' refers to row or column totals. We have relaxed the mathematical notation with respect to populations and samples here. Let us have a look at what this formula is doing. Table 10.14 shows artificial data for two species measured at four sites. First of all, the two rows of data are changed into row-profiles by dividing each of them by row-totals Z_{1+} and Z_{2+}. This gives:

A: 0.5 0.1 0 0.4
B: 0.27 0.13 0.2 0.4

The profiles are subtracted from each other, and the differences are squared:

$(A-B)^2$: 0.05 0.0009 0.04 0

In the final step, the weighted average is calculated (weights are given by the site totals 18, 6, 6 and 20); take the square root and multiply with a constant:

$$D(Z_1,Z_2)=\sqrt{50}\sqrt{\frac{0.05}{18}+\frac{0.0009}{6}+\frac{0.04}{6}+\frac{0}{20}}=0.69$$

The lower this value, the more similar the two species. Note that a site with a high total (e.g., 18) is likely to have less influence on the distance, but sites with low totals may have more influence. If the data contain more than two species, the Chi-square distance can be calculated for any combination of two species. The Chi-square *metric* is identical to the Chi-square *distance*, except for the multiplication of the square root of Z_{++} (total of Y and X)[1]. Table 10.15 shows the Chi-square

[1] A distance function is used to calculate association between two observations or two variables a and b. A metric has the following properties: (i) if $a = b$, then the distance between them is 0; (ii) if $a \neq b$, the distance between them is larger than 0; (iii) the distance between a and b is equal to the distance between b and a,

distances among the four zoobenthic species from the Argentinean data. Note that *U. uruguayensis* (UU) and *N. succinea* (NS) have the highest Chi-square distances. Hence, these species are the most dissimilar, as judged by the Chi-square distance function. The higher the Chi-square distance the greater the difference.

The disadvantage of the Chi-square distance function is that it may be sensitive to species that were measured at only a few sites (patchy behaviour) with low abundance. The underlying measure of association in correspondence analysis and canonical correspondence analysis is the Chi-square distance function. Patchy species (with low abundance) will almost certainly dominate the first few axes of the ordination diagram.

Table 10.14. Artificial data for two species (A and B) at 4 sites (1-4).

	Sites				
Species	1	2	3	4	Total
A	10	2	0	8	20
B	8	4	6	12	30
Total	18	6	6	20	50

Table 10.15. Chi-square distances among the four zoobenthic species from the Argentinean data. The abbreviations LA, HS, UU and NS refer to *L. acuta*, *H. similis*, *U. uruguayensis* and *N. succinea*. The smaller the values, the more similar are two species.

	LA	HS	UU	NS
LA	0	1.3	1.1	2
HS		0	1.7	1.2
UU			0	2.3
NS				0

10.4 Q and R analysis: Concluding remarks

As explained in Section 10.1, Legendre and Legendre (1998) grouped the measures of association into Q and R analyses. The Q analysis is for relationships among observations (e.g., sites) and the R analysis for relationships among variables (e.g., species). The correlation coefficient is an R analysis, and the Jaccard and related methods is a Q analysis. Legendre and Legendre (p289-290, 1998) gave five objections for not using the correlation as a tool to calculate association among observations. We discuss some of their objections using this example:

---

and (iv) the distance between *a* and *b* plus the distance between *b* and *c* is equal to or larger than the distance between *a* and *c* (Legendre and Legendre 1998).

	A	B	C	D	E	F
1:	2	4	2	3	3	99
2:	1	3	1	2	0	90
3:	0	2	3	3	1	95
4:	3	5	2	3	1	94

The four rows are sites and the six columns are species. Assume we want to calculate the correlation among the four sites. In this case, the correlation between any two rows will be close to 1 because F is causing a large contribution to the correlation (its value is far above the row average). This situation can arise if one species is considerably more abundant than the others, or if the variables are for example physical variables and one of them has much higher values (which could also be due to different units). One option to avoid this problem is to standardise each variable (column) before calculating the correlation between rows. However, this means that the correlation between rows one and two depend on the data in other rows as these are needed for the standardisation.

Another argument is that correlation between variables will standardise the data and therefore the data are without units. However, if the variables are in different units, and if the correlation is applied as a Q analysis in which the variables have different units, we get similar problems as in the artificial example above.

So, what about an R analysis to quantify relationships between species that contain many (double) zeros? Legendre and Legendre (p. 292, 1998) advise to (i) remove the species with lots of zeros, (ii) consider zeros as missing values, and (iii) eliminate double zeros from the computation of the calculations of the association matrix. Aggregation to a higher taxa or taking totals per transect/beach/area may help as well. This is what we did in Table 10.1, and the same principle (totals per pasture) is applied in Chapter 32 and Chapter 12 (totals per taxa). Note that the suggestions made by Legendre and Legendre (1998) are not a golden rule. Some people would argue that joint absence of species at sites is important information that should be taken into account. In this case, you do not want to aggregate, remove species with lots of zeros, etc. Whichever route you choose, you should be prepared to defend your choice. The suggestion from Legendre and Legendre (1998) that the Chi-square distance copes better with double zeros is valid, but then this one is sensitive to patchy species. The alternative is to use an asymmetrical coefficient in an R analysis (p. 294 Legendre and Legendre 1998) like the Jaccard index or the Sørensen coefficient, and focus on the question of which species co-occur. For example, the Jaccard indices among the four zoobenthic species are given in Table 10.16.

The message here is to think carefully about the underlying ecological question, and what the chosen measure of association is actually doing with the data.

Table 10.16. Jaccard index among the four zoobenthic species from the Argentinean data. The abbreviations LA, HS, UU and NS refer to *L. acuta*, *H. similis*, *U. uruguayensis* and *N. succinea*. The smaller the value, the more similar are the two species.

	LA	HS	UU	NS
LA	0	0	0.67	0.33
HS		0	0.67	0.33
UU			0	0.5
NS				0

Some measures of association cannot deal with negative numbers. The Jaccard index, for example, is typically designed for count data. Others have problems when the total of a row (and/or column) is equal to zero, for example the Chi-square distance. The correlation coefficient cannot deal with variables that have the same value at each site (standard deviation is null). So, again the message here is to know the technical aspects of the chosen measure of association.

It is also possible to transform (square root, log, etc.) the data prior to calculation of the measure of association. Some authors (e.g., Clarke and Warwick 1994) even advocate choosing only one measure of similarity, i.e., the Bray–Curtis coefficient, and then use different data transformations. This is tempting as it simplifies the number of steps to consider, but it is a lazy option. Bray–Curtis may well work fine for marine benthic data (the original focus of PRIMER), but for different application fields, other measures of association may be more appropriate. We recommend knowing and understanding the chosen measure of association, and trying different ones as necessary (depending on the data and questions). Remembering that it is possible to combine some measures of association with a data transformation. Although this approach sounds a bit more challenging, it forces the researcher to know and understand what he/she is doing, and why.

Which measure of association to chose?

Legendre and Legendre (1998) give around 30 other measures of similarity, and this obviously raises the question of when to use which measure of association. Unfortunately, there is no easy answer to this. A cliché answer is: 'it depends'. It depends on the underlying questions, the data itself, and the characteristics of the association function. The underlying question itself should guide you to a Q or R analysis. From there onwards it is the double zeros, outliers and your underlying question that determine which measure is appropriate. If you are interested in the outliers, the Euclidean distance function is a good tool. But for ordination and clustering purposes, the Sørensen or Chord distance functions seem to perform well in practice, and certainly better than the correlation coefficient and Chi-Square functions.

Jongman et al. (1995) carried out a simulation study in which they looked at the sensitivity (sample total, dominant species, species richness) of nine measures of association. The Jaccard index, coefficient of community and the Chord distance

performed reasonably well. However, the (squared) Euclidean distance, similarity ratio and percentage similarity functions were sensitive to dominant species and sample totals.

Further guidance in the choice of a coefficient can be found in Section 7.6 in Legendre and Legendre (1998). The Argentinean data showed how important it is to choose the most appropriate measure of association. The Jaccard and Sørensen indices gave similar values, but other measures of association suggest different ecological interpretations. This in itself is useful, as long as the interpretation is done in the context of the characteristics of the different measure of association.

10.5 Hypothesis testing with measures of association

The measures of association discussed in the previous section result in an N-by-N matrix of similarities or dissimilarities, where N is the number of either observations or variables. We now discuss two techniques in where we want to link the (dis-)similarity matrix with external information. This will be done with ANOSIM (Clarke and Ainsworth 1993), the Mantel test (Sokal and Rohlf 1995) and the partial Mantel test (Legendre and Legendre 1998). In this section, we use real data.

The Data

The data are unpublished bird radar data (Central Science Laboratory, York, UK). During three days in October 2004, a new radar installation was used to measure birds. The radar gave X-band (vertical direction) and S-band data. The S-band data are used here. The radar turns 360 degrees in the horizontal plane and records latitude, longitude and time (among many other variables) for each bird. A certain amount of data preparation was carried out to avoid counting the same bird twice. Figure 10.2 shows the spatial distribution of the observations for the three sampling days. We use these data to illustrate ANOSIM and the Mantel test, but it should be stressed that the approach presented here is only one of several possible approaches. Later, we use specific spatial statistical methods on these data, and another valid approach would be to use a generalised additive (mixed) model.

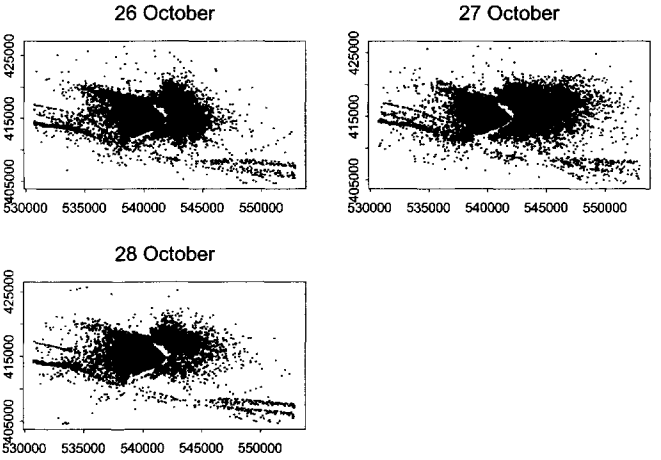

Figure 10.2. Observed values and spatial distribution for the three sampling days. Each dot represents an observation. The horizontal and vertical axes represent longitude and latitude, respectively (in metres).

To prepare the data for the ANOSIM and Mantel tests, totals per time unit per spatial unit were calculated. An arbitrary, but sensible choice for the time unit is one hour. Based on the spatial distribution, we decided to use data observed between a longitude of 535000 and 545000 and latitude between 410000 and 420000. This was to avoid too many cells with zero observations. The size of the spatial grid is again an arbitrary choice. The most convenient choice is to divide the axes in each panel in Figure 10.2 into M cells. We used $M = 10$, $M = 15$ and $M = 20$. The first option results in cells of 1000-by-1000 m, the second option in cells of 666-by-666 m, and the third value in cells of size 500-by-500 m. Figure 10.3 shows the spatial grid if $M = 15$ is used. Both the horizontal and the vertical axes are split up into 15 blocks, and as a result a total of 225 cells are used. $M = 10$, gives us a 100 cells, and $M = 20$ gives us 200 cells. Total birds per cell per hour were calculated and used in the analyses. The first question to address is whether there is a difference in the spatial distribution and abundance of birds across the three time periods, and the Mantel test is used for this.

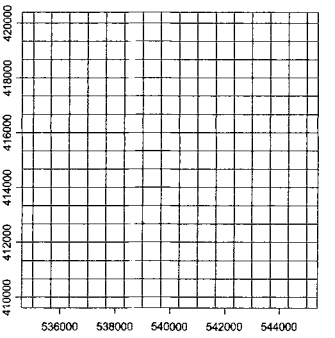

Figure 10.3. Example of a spatial grid using $M = 15$. All grids are of the same size (including grids on the boundary).

Spatial differences and the Mantel test

The data matrix is of the form:

$$\mathbf{D} = [\mathbf{D}_1 \quad \mathbf{D}_2 \quad \mathbf{D}_3] \quad \text{and} \quad \mathbf{S} = [\mathbf{X} \quad \mathbf{Y}]$$

If $M = 20$, then $\mathbf{D}_1$ is a 400-by-H_1 matrix, where H_1 is the number of hours in day one. We have $H_1 = 23$, $H_2 = 24$ and $H_3 = 23$ hours, which adds up to a total number of 70 hours. Hence, $\mathbf{D}$ is of dimension 400-by-70, and each row contains 70 bird observations in a particular cell. The vectors X and Y contain the longitude and latitude values for each grid, and $\mathbf{S}$ is of dimension 440-by-2.

The first ecological question we address is whether the relationships among cells (in terms of bird observations) is related to spatial distances; cells adjacent to each other might have a similar pattern over time. To test whether this is indeed the case, the Mantel test (Legendre and Legendre 1998; Sokal and Rohlf 1995) can be used. One way of applying this method is as follows:

- Calculate the similarity among the 400 cells in terms of observed number of birds. Call this matrix $\mathbf{F}_1$. The matrix $\mathbf{D}$ is used for this.
- Calculate the geographical similarity among the 400 grid cells. Use real distances for this. Call this matrix $\mathbf{F}_2$. The matrix $\mathbf{S}$ is used for this.
- Compared these two (dis-)similarity matrices $\mathbf{F}_1$ and $\mathbf{F}_2$ with each other using a correlation coefficient.
- Assess the statistical significant of the correlation coefficient.

The Mantel test calculates two distance matrices $\mathbf{F}_1$ and $\mathbf{F}_2$. Both matrices are of dimension 400-by-400. $\mathbf{F}_1$ represents the dissimilarity among the 400 grid points, and $\mathbf{F}_2$ the geographical (=Euclidean) distances among the cells. The Mantel test compares the two matrices $\mathbf{F}_1$ and $\mathbf{F}_2$ with each other. It does this by calculating a correlation between the (lower diagonal) elements of the two matrices. A

permutation test is then used to assess the significance of the correlation. This test tells us whether there is a significant relation among bird numbers across the grids and the geographical distances of those grids.

A bit more detail on the Mantel test and the permutation

We now explain the Mantel test and permutation test in more detail. Recall that the 400-by-70 matrix of bird abundance was used to calculate $\mathbf{F}_1$, and the geographical coordinates for $\mathbf{F}_2$. A sensible measure of association for $\mathbf{F}_1$ is the Jaccard index, and Euclidean distances can be used to define the geographical distances in $\mathbf{F}_2$. Figure 10.4 shows a schematic outline of the Mantel test. Both matrices $\mathbf{F}_1$ and $\mathbf{F}_2$ are of dimension 400-by-400. These matrices are symmetric; the Jaccard index between cells 1 and 2 is the same as between 2 and 1. The diagonal elements are irrelevant as they represent the Jaccard index between cell 1 and cell 1. The same holds for $\mathbf{F}_2$; hence, we only have to concentrate on the part above (or below) the diagonal matrices.

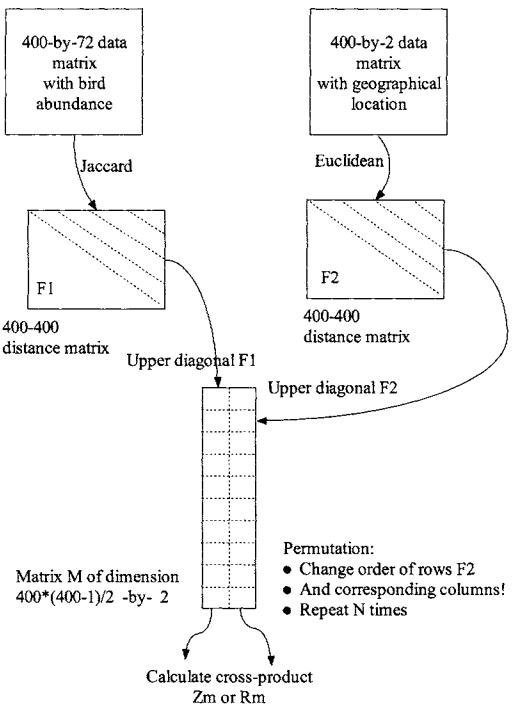

Figure 10.4. Schematic outline of the Mantel test. The data matrices are converted into distance matrices F_1 and F_2, and the upper diagonal matrices of these two distance matrices are compared using, for example, a correlation coefficient. A permutation test is used to assess its statistical significance.

To compare the upper diagonal elements in $\mathbf{F}_1$ and $\mathbf{F}_2$, they can be extracted and stored *side-by-side* in a new matrix **M** that has 400 × (400 − 1)/2 rows and 2 columns. With side-by-side we mean that the element $\mathbf{F}_{1,ij}$ is placed next to $\mathbf{F}_{2,ij}$ in **M** (see the lower part of Figure 10.4). To quantify how similar are the (dis-)similarity matrices $\mathbf{F}_1$ and $\mathbf{F}_2$, the cross-product of $M_{i,1}$ and $M_{i,2}$ can be calculated in two ways:

$$Z_m = \sum_{i=1}^{J} M_{i,1} M_{i,2}$$

$$R_m = \frac{1}{J-1} \sum_{i=1}^{J} \frac{(M_{i,1} - \overline{M}_1)}{s_{M_1}} \frac{(M_{i,2} - \overline{M}_2)}{s_{M_2}}$$

J is the number of rows in the matrix **M**, which is 400 × (400 − 1)/2 in this case. Z_m is the Mantel statistic, and R_m is the standardised Mantel statistic. In the latter, each column of **M** is mean deleted and divided by its standard deviation. The advantage of R_m is that it is rescaled to be between –1 and 1; it is a correlation coefficient.

Assume we find a certain value of R_m. The question now is what does it mean and is it significant? Recall that R_m measures the association among the elements of two (dis-)similarity matrices. So we cannot say anything in terms of the original variables, merely in terms of similarity among the matrices $\mathbf{F}_1$ and $\mathbf{F}_2$! In this case, the question is whether similarities between 400 sites in terms of the Jaccard index are related to the Euclidean distances (reflecting geographical distances in this case). The problem is that we cannot use conventional significance levels for the correlation coefficient R_m as the cells are not independent. Therefore, a permutation test is used. The hypotheses are as follows:

- H_0: The elements in the matrix $\mathbf{F}_1$ are not linearly correlated with the elements in the matrix $\mathbf{F}_2$. Or formulated in terms of the original data, the association among the 400 cells containing bird numbers is not (linearly) related to the spatial distances.
- H_1: There is a linear correlation between the *association* in the two original data matrices.

Under the null hypothesis, rows in one of the original data matrices can be permutated a large number of times, and each time the Mantel statistic, denoted by R_m^*, can be calculated. The number of times that R_m^* is larger than the original R_m, is used to calculate the p-value.

Results of the Mantel test

Several measures of similarity were used: the Jaccard index, the Chord distance, Whittaker index of association and Euclidean distances. The Chord distance function, the Jaccard index and Whittaker index are similar in the sense that the larger values are treated in the same way as the smaller values. The Jaccard index treats the data as presence–absence. The Whittaker index of association is using

proportions. The results are presented in Table 10.17. The Jaccard index, the Chord distances and the Whittaker index indicate a significant relationship between bird abundance in grid cells and geographical distances. However, taking into account the absolute value of the data (Euclidean distance), there is no relationship between the bird numbers and geographical distances. This means that cells with large bird numbers are not necessarily close to each other. However, if we consider the data as presence–absence of birds in grid cells (Jaccard index) or "down-weight" the larger values (Chord), then there is a relationship.

Table 10.17. Results of the Mantel test. The number of permutations was 9999.

Measure of Association	Statistic	p-value
Chord	0.463	<0.001
Jaccard	0.415	<0.001
Euclidean	−0.161	1
Whittaker	0.405	<0.001

Extensions: The Partial Mantel test

The Mantel test identifies whether there is a correlation between the elements of two matrices. The way it was applied above was to compare a dissimilarity matrix for the bird counts in grid cells with a matrix representing geographical distances. Now suppose that the study area can be divided into two parts, let us call them area A and B. An artificial scenario is sketched in Figure 10.5. The question is whether there are any differences between the black cells (area A) and the white cells (area B). A possible scenario is that area A is close to a wind farm or airport. The question is now whether area A and B are similar in terms of bird observations, and the partial Mantel test (Legendre and Legendre 1998) can be used for this.

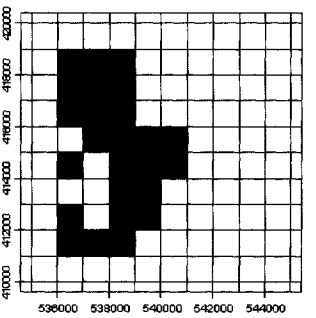

Figure 10.5. Scenario for partial Mantel test. Black cells form area A, and white cells area B.

The partial Mantel test uses three matrices. The first matrix is $\mathbf{F}_1$ and represents the dissimilarity among the 400 grid cells using the bird observations. The Jaccard index, or any other measure of association, can be used. The matrix $\mathbf{F}_2$ represents the difference between the two areas A and B. This sounds complicated, but all we need is a vector of length 400 with zeros (if a cell is from A) and ones (if a cell is from B). The Manhattan distance will ensure that the similarity among all cells from A are 0, and that the similarity among all cells from B are also 0. This is the within-area similarity. The between-area similarity contains only ones. The format of the resulting distance matrix $\mathbf{F}_2$ for a simple artificial example is sketched below:

$$\mathbf{F}_2 = \begin{array}{c} \\ A \\ \\ \\ B \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} A & & & & B & & \end{array} \\ \begin{bmatrix} 0 & & & & & & \\ 0 & 0 & & & & & \\ 0 & 0 & 0 & & & & \\ 1 & 1 & 1 & 0 & & & \\ 1 & 1 & 1 & 0 & 0 & & \\ 1 & 1 & 1 & 0 & 0 & 0 & \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

So $\mathbf{F}_2$ is only used to quantify whether cells are from the same area, yes (0) or no (1). The third matrix, $\mathbf{F}_3$, represents again the geographical distances among the 400 grid cells.

The partial Mantel test then compares $\mathbf{F}_1$ (similarity of bird abundance in the grids) with the design matrix $\mathbf{F}_2$, while partialling out similarities due to geographical distances. In other words, the association between bird abundance at the grid cells is compared with areas A and B in Figure 10.5, while taking into account (and removing) the effect of geographical distances. The idea of partialling out information is also discussed in Chapters 5 and 16. Further technical details can be found in Legendre and Legendre (1998).

Testing for differences in years using ANOSIM

The ANalyis Of SIMilarities (ANOSIM) method (Clarke 1993; Legendre and Legendre 1998) works in a similar way as the Mantel test. The starting point is the matrix:

$$\mathbf{D} = [\mathbf{D}_1 \quad \mathbf{D}_2 \quad \mathbf{D}_3]$$

If $M = 20$, the matrix $\mathbf{D}$ is of dimension 400-by-70. The matrix $\mathbf{D}_1$ contains the 400-cell-by-23 hour numbers. The question now is whether there is any difference in the association among the three days. Instead of comparing similarities among cells, we now calculate similarities among hours. This gives a 70-by-70 matrix $\mathbf{F}$

(sampling took place for 70 hours in total), using an appropriate measure of association. Each element in the matrix **F** represents the similarity between two hours. The 70-by-70 symmetric matrix **F** is of the form:

$$\mathbf{F} = \begin{pmatrix} \mathbf{F}_1 & & \\ \mathbf{F}_{12} & \mathbf{F}_2 & \\ \mathbf{F}_{13} & \mathbf{F}_{23} & \mathbf{F}_3 \end{pmatrix}$$

$\mathbf{F}_1$ represent the (dis-)similarities among the hours of day one, and the same holds for $\mathbf{F}_2$ (day two) and $\mathbf{F}_3$ (day three). The matrices $\mathbf{F}_{12}$, $\mathbf{F}_{13}$ and $\mathbf{F}_{23}$ represent the (dis-)similarities of hours from two different days.

Now let us try to understand how the ANOSIM method can help us, and which underlying questions it can address. Assume bird behaviour is the same on day one, day two and day three. If we then compare patterns between hour *i* and *j* on day 1, we would expect to find similar spatial patterns on day two, and also on day three. As described above, the matrix $\mathbf{F}_1$ contains a comparison among the (spatial) patterns in each hour. A high similarity between two hours means that the 400 spatial observations are similar (as measured by for example the Jaccard index). $\mathbf{F}_2$ and $\mathbf{F}_3$ represent the same information for days two and three, respectively. So what about $\mathbf{F}_{12}$? The information in this matrix tells us the relationship among the spatial observations in hour *i* on day one and hour *j* on day two. And a similar interpretation holds for day three. If spatial relationships are the same in all three days, we would expect to find similar dissimilarities in all the sub-matrices in **F**.

The ANOSIM method uses a test statistic that measures the difference between the within group variation (in $\mathbf{F}_1$, $\mathbf{F}_2$, and $\mathbf{F}_3$), and the between group variation (in $\mathbf{F}_{12}$, $\mathbf{F}_{13}$ and $\mathbf{F}_{23}$). In fact, the test statistic uses the ranks of all elements in **F**. If the difference between the *within* group similarity and *between* group similarity is large, then there is evidence that the association among the three time periods is different. The test statistic is given by

$$R = \frac{\bar{r}_b - \bar{r}_w}{n(n-1)/4}$$

where $\bar{r}_b$ and $\bar{r}_w$ are the between and within mean values of the ranked elements and *n* is the total number of observations. A permutation test similar to the one for the Mantel test can be used to assess its significance. The underlying null hypothesis is that there are no differences among the groups in terms of association. The approach summarised here is a so-called one-way ANOSIM test. The ANOSIM procedure can be extended to two-way ANOSIM. These methods are non-parametric multivariate extensions of ANOVA; see Clarke (1993) and Legendre and Legendre (1998) for further technical details.

Results ANOSIM

To define the matrix F, we used the Jaccard index, Chord distances, Euclidean distances and the Whittaker index of association, and the following results were obtained. For the Euclidean distances, we have $R = 0.192$ ($p < 0.001$). This means that there is a significant difference between the three time periods. The method also gives more detailed information like pair wise comparisons between days one and two ($R = 0.265$, $p < 0.001$), days one and three ($R = 0.346$, $p < 0.001$) and days two and three ($R = 0.016$, $p = 0.151$). Using the Jaccard index, the overall R is 0.234 ($p < 0.001$). The conclusions are the same as for the Euclidean distance. The same holds for the Chord distance ($R = 0.405$, $p < 0.001$) and the Whittaker index of association ($R = 0.431$, $p < 0.001$). With more groups, care is needed with the interpretation of p-values for the individual group-by-group comparisons; there is no such thing as a Bonferroni correction of p-values due to multiple comparisons in ANOSIM. However, a large p-value indicates we cannot reject the null-hypothesis that two time periods are the same, even if many multiple comparisons are made.

Instead of dividing the time into periods of three days, we could have chosen any other division, as long as there are not too many classes. The same principle as in ANOVA applies: If there are too many treatment levels (which are the three time periods here) with rather different effects, the null hypothesis will be rejected regardless of any real differences.