5. Statistical tests for multivariate data

Ecological data are difficult to handle when it comes to statistical testing. All the methods above, used as they are presented, are descriptive or explanatory, but as yet no statistical test has been presented to assess the significance of the relationships or structures. Here we shall present two tests, both in the general framework of **permutation testing**: the test on canonical axes of a canonical ordination and the Mantel test on distance matrices.

5.1 Parametric tests

Classical, parametric testing has many constraints and generally supposes that several conditions are fulfilled for the test to be valid. One fundamental assumption is that the observations must be independent from one another (i.e. the probability of obtaining a given value of the response variable in one observation is independent of the values found in other observations). Autocorrelated data violate this principle, their error terms being correlated across observations. This topic is especially important in the context of spatial analysis. Another frequent requirement of classical testing is the conformity of the distribution of the data to some well-known theoretical distribution, most often the normal distribution.

When the conditions of a given test are fulfilled, an auxiliary variable (for instance an *F* or *t*-statistic), constructed on the basis of one or several parameters estimated from the data, has a known behaviour under the null hypothesis. It is thus possible to ascertain whether the observed value of that statistic is likely or not to occur if H_0 is true. If the observed value is as extreme or more extreme than the value of the reference statistic for a pre-established probability level (usually $\alpha = 0.05$), then H_0 is rejected. If not, H_0 is not rejected (Figure 31).

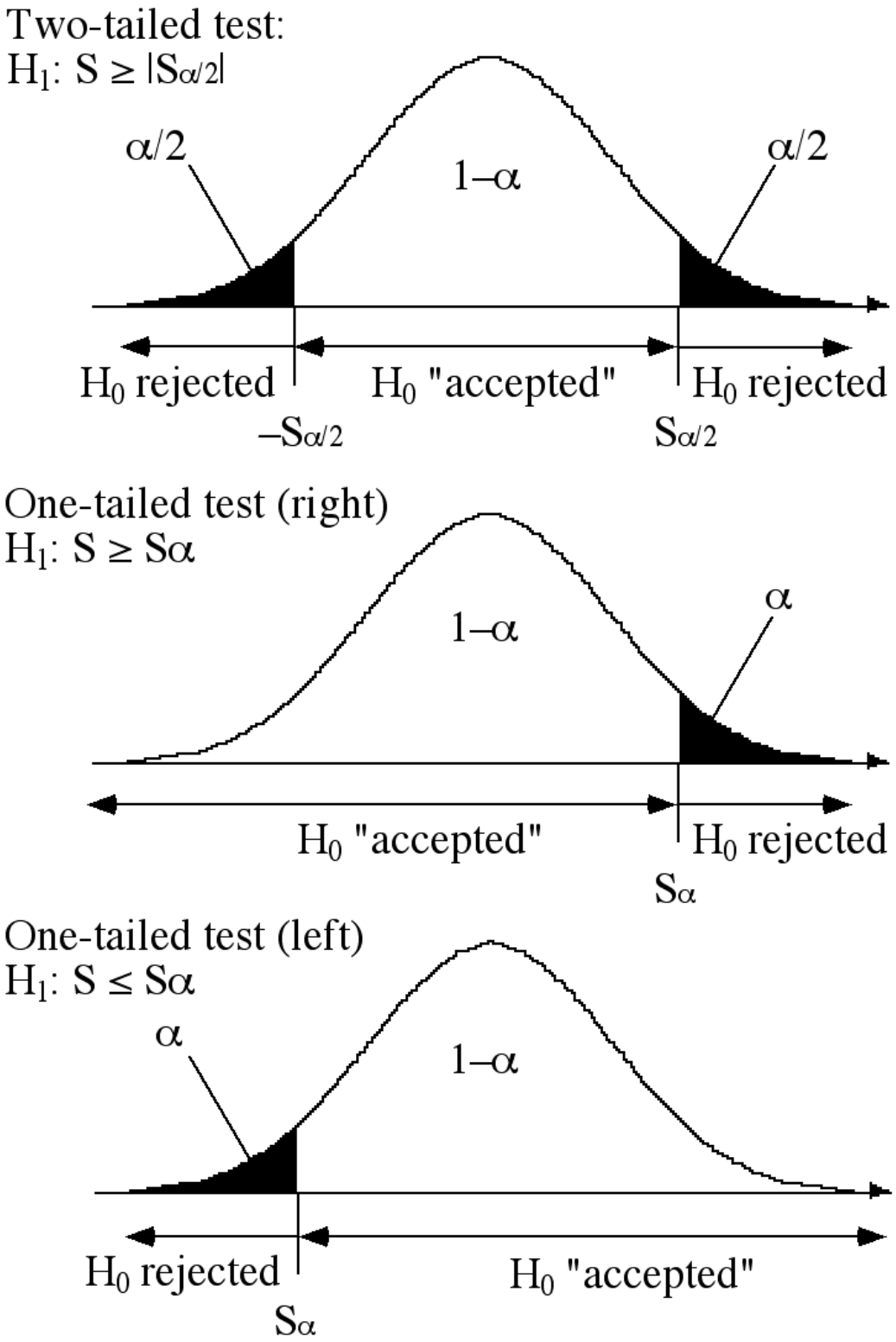

Figure 31 - Decision in statistical testing. S is some test statistic (e.g. Student's *t* statistic). Adapted from course notes by Pierre Legendre.

5.2 Permutation tests

The parametric procedure is rarely usable with ecological data, mainly because these data rarely fulfil the assumptions related to distribution. Furthermore, even data transformations often do not manage to normalize the data. In these conditions, another, very elegant but computationally more intensive approach is available: testing by random permutations.

Principle of permutation testing: if no theoretical reference distribution is available, then generate a reference distribution under H_0 from the data themselves. This is achieved by permuting the data randomly in a scheme that ensures H_0 to be true, and recomputing the test statistic. Repeat the procedure a large number of times. The observed test statistic is then compared to the set of test statistics obtained by permutations. If the observed value is as extreme or more extreme than, say, the 5% most extreme values obtained under permutations, then it is considered too extreme for H_0 to be true. H_0 is rejected.

An example can be construed based on the Pearson correlation coefficient between two quantitative variables (Table X):

Table X - Example of data for permutation test: Pearson's *r*

Var.1	Var.2		Perm.1	Perm.2	Perm.3	.....
1	4		4	10	11	
3	5		5	8	8	
2	3		8	4	14	
4	6		10	6	5	
3	8		11	3	3	
6	7		9	7	10	
7	9		3	5	6	
5	10		14	14	9	
8	11		7	9	7	
9	14		6	11	4	
Pearson *r*	**0.890**	*r**	-0.081	0.288	-0.474	.....

In Table X, the two leftmost columns represent the original, unpermuted data. 0.890 is the value of Pearson's *r* correlation coefficient between the two variables. Perm.1, Perm.2 and so on are permutations of Var.2. The *r** are the values of *r* between Var.1 and the permuted Perm.* columns. Since these columns have had their values permuted randomly, there is no expected relationship between Var.1 and Perm.1, Perm.*, and so on. These are thus realisations of H_0, the null hypothesis of the test: there is no linear relationship between the two variables.

In permutation testing, the observed (true) value must *a priori* be considered as belonging to the reference distribution. Therefore, it is customary to ask for 99, 999 or 9999 random permutations. It is then easy to verify the ranking of the observed value with respect to the permuted ones, and to transform this into a probability value (Figure 32):

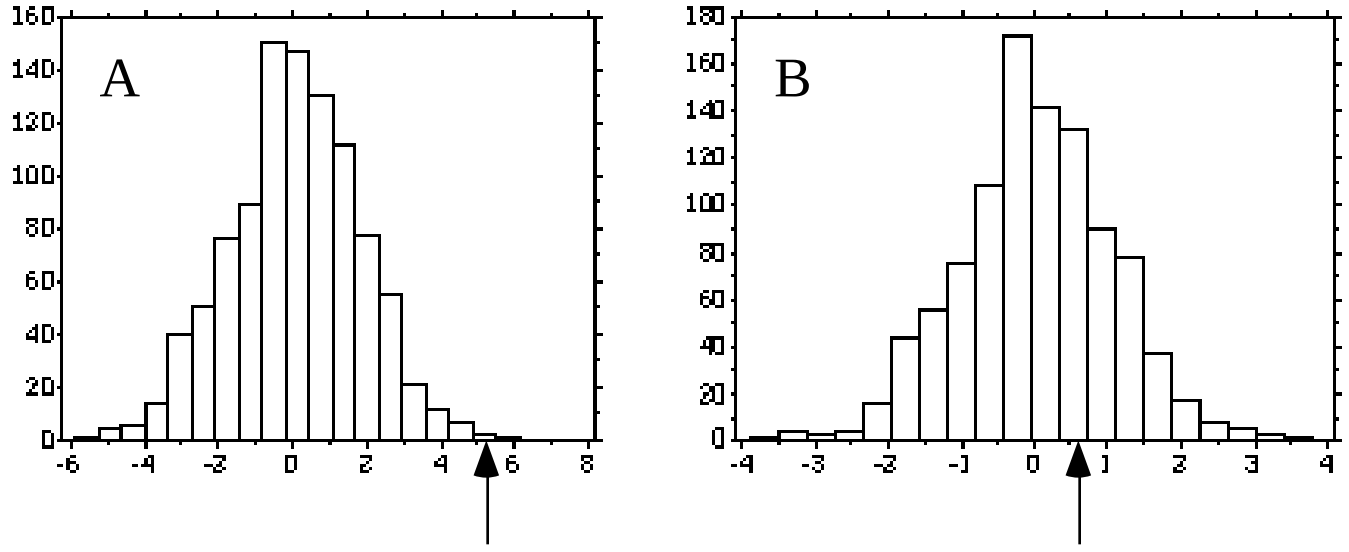

Figure 32 - Examples of comparison of true test values with reference distributions generated by random permutations. In A, the true value (arrow) is quite extreme: 5% or less random values are larger than the true one. H_0 would be rejected at the 5% one-tailed probability level. B: the true value is amidst the random ones. H_0 is not rejected. Adapted from course notes by Pierre Legendre.

If the test is two-tailed, H_0 is rejected at the 0.05 probability level when

$$P_{comp} = \frac{[\text{per} < -|\text{obs}|] + [\text{per} = -|\text{obs}|] + [\text{per} = |\text{obs}|] + [\text{per} > |\text{obs}|]}{\text{Nb. permutations} + 1} \leq 0{,}05$$

If the test is one-tailed (right), H_0 is rejected at the 0.05 probability level when

$$P_{comp} = \frac{[\text{per} = |\text{obs}|] + [\text{per} > |\text{obs}|]}{\text{Nb. permutations} + 1} \leq 0{,}05$$

Table XI gives some examples of numerical summaries of permutation tests, with the probability of H_0 derived from the results, for two- and one-tailed tests.

Table XI - Examples based on Pearson's *r*:

Two-tailed tests:

[per<-|obs|]	[per=-|obs|]	[-|obs|<per<|obs|]	[per=|obs|]	[per>|obs|]	$P(H_0)$
6	1	969	1	23	0.031
21	20	926	1	32	0.074
0	0	999	1	0	0.001
0	0	990	1	9	0.010
0	0	99	1	0	0.01
0	1	98	1	0	0.02

One-tailed test, right tail:

[per<-|obs|]	[per=-|obs|]	[-|obs|<per<|obs|]	[per=|obs|]	[per>|obs|]	$P(H_0)$
6	1	969	1	23	0.024
21	20	926	1	32	0.033
0	0	999	1	0	0.001
0	1	98	1	0	0.01

One-tailed test, left tail:

[per<-|obs|]	[per=-|obs|]	[-|obs|<per<|obs|]	[per=|obs|]	[per>|obs|]	$P(H_0)$
6	1	969	1	23	0.007
21	20	926	1	32	0.041
0	1	999	0	0	0.001
0	1	98	1	0	0.01

Words of caution about permutation tests

Elegant as it may seem, the method of permutations does not solve all the problems related to statistical testing.

1. Beyond simple cases like the one above, other problems may require different and more complicated permutation schemes than the simple random scheme applied here. It is, in particular, the case with the tests of the main factors of an ANOVA coded as proposed in Section 4.7.7, where the permutations for factor A must be limited within the levels of factor B, and vice versa.

2. Permutation tests do solve several, but not all, distributional problems. In particular, **they do not solve distributional problems linked to the hypothesis being tested**. For instance, permutational ANOVA does not require normality, but it still does require

homogeneity of variances, because this relates to the Behrens-Fisher problem linked to comparisons of means: actually two hypotheses are tested simultaneously in ANOVA, i.e. equality of the means and equality of the variances. This is true also for the two-sample *t*-test of comparison of means.

3. Contrary to popular belief, **permutation tests do not solve the problem of independence of observations**. This problem has still to be addressed by special solutions, differing from case to case, and often related to the correction of degrees of freedom.

4. Although many statistics can be tested directly by permutations (e.g. Pearson's *r* above), it is generally advised to use a *pivotal statistic* whenever possible (for Pearson's *r* it would be a Student's *t* statistic). A pivotal statistic has a distribution under the null hypothesis which remains the same for any value of the measured effect.

5. Observe that **it is not the statistic itself which determines if a test is parametric or not**: it is the reference to a theoretical distribution (which requires assumptions about the parameters of the statistical population from which the data have been extracted) or to permutations.

5.3 Tests of an RDA or CCA

5.3.1 Principle

Remember that the eigenvalue of a canonical axis represents the amount of variation of the response data explained by the axis. If one wants to test one single axis at a time the idea of the test is to verify whether an equal or larger eigenvalue can be obtained under the null hypothesis of no relationship between the response matrix and the explanatory matrix. But normally one first tests the significance of the analysis globally. The basis is then the sum of all canonical eigenvalues. The hypotheses are thus:

- H_0: there is no linear relationship between the response matrix and the explanatory matrix;

- H_1: there is a linear relationship between the response matrix and the explanatory matrix.

Originally, the test statistic was the eigenvalue or sum of canonical eigenvalues itself. Now, one uses a pivotal statistic instead, which is a "pseudo-*F*" statistic which is defined as

$$F = \frac{\text{sum of all canonical eigenvalues } / m}{\text{RSS}/(n - m - 1)}$$

where n is the number of objects, m is the number of explanatory variables and RSS is the residual sum of squares, i.e. the sum of non-canonical eigenvalues (after fitting the explanatory variables).

Partial canonical analyses and their axes can also be tested for significance. The *F* statistic then takes into account the covariables, i.e. the q variables of the **W** matrix that are held constant in the analysis:

$$F = \frac{\text{sum of all canonical eigenvalues } / m}{\text{RSS}/(n - m - q - 1)}$$

where the "sum of all canonical eigenvalues" is, this time, the value obtained when holding **W** constant. It is the (unadjusted) fraction [a] of the partitioning of variation.

5.3.2 Permutation procedures

The permutation procedures for these tests are not trivial (see Legendre & Legendre 1998, p. 607sq. for details). The main permutation types are the following:

a) Without covariables in the analysis

- permutation of raw data; the null hypothesis is that of exchangeability of the rows of **Y** with respect to the

observations in **X**. This is implemented by permuting the rows of **Y** (or, alternatively, the rows of **X**) at random and recomputing the redundancy analysis;

- permutation of residuals; here the residuals of a linear (or other) model are the permutable units. In canonical analysis, the null hypothesis is that of exchangeability of the residuals of the response variables after fitting the explanatory variables. Tests of significance using permutation of residuals have only asymptotically exact significance levels (i.e. as *n* becomes large).

b) With covariables in the analysis: two methods of permutation of residuals are used to test the significance of the sum of all canonical eigenvalues:

- permutation of residuals under a reduced (or null) model: the permutable units are the residuals of variables **Y** on **W**;
- permutation of residuals under a full model: the permutable units are the residuals of variables **Y** on **X** and **W** together.

Using Monte Carlo simulations, Anderson & Legendre (1999)[1] compared empirical type I error and power of various permutation techniques for a test of significance of a single partial regression coefficient. Their results are relevant to RDA because this method, using a single *y* variable, is equivalent to partial regression. Their main conclusions were:

- when the error in the data strongly departed from normality, permutation tests had more power than parametric *t*-tests;
- type I error and power were asymptotically equivalent for permutation of raw data or permutation of residuals under the full or reduced model;

[1] Anderson, M. J. & P. Legendre. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *J. Statis.Comput.Simulation* **62**: 271-303.

- when the covariable contained an extreme outlier, permutation of raw data resulted in unstable (often inflated) type I error. This was not the case with permutation of residuals. Thus, permutation of residuals are especially recommended when the matrix **W** of covariables contains outliers.

5.4 Mantel test: matrix correlation

Described in 1967 by the epidemiologist Nathan Mantel[2], the test of matrix correlation that bears his name has been increasingly used by ecologists in the eighties. Presently, however, with the advent of the powerful canonical ordination techniques, **the use of the Mantel test must be restricted to cases where the hypotheses and data themselves are naturally stated in terms of distances or similarities** rather than in terms of raw data. This is very important. See justification at the end of this chapter.

5.4.1 Principle of the test

The Mantel procedure tests the linear correlation between similarity or distance matrices. For example, one could use it to compare a matrix **Y** of Bray-Curtis distances among sites based on species abundances and a matrix **X** of Euclidean distances among the same sites, built on the basis of remote-sensing informations. The test will tell if the species-based distances are significantly, linearly correlated with the remote-sensing-based distances, In other words, it will answer a question of the type:

"Do pairs of sites that are similar in terms of species composition also tend to be similar in terms of spectral signatures?"

If it is the case, then one will have gained some confidence in the perspective of assessing variation in species composition by means of variation in remote sensing data. Such a conclusion must be made with

[2] Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27: 209-220.

caution, since the interpretation has to be made in the "world of distances" and not in the "world of raw data".

Formally, the hypotheses of the Mantel test can be stated as follows:

H_0: the distances (or similarities) among objects in matrix **Y** are not (linearly) correlated with the corresponding distances in matrix **X**.

H_1: the distances among objects in matrix **Y** are linearly correlated to the distances in **X**.

The original **Mantel *z* statistic**, i.e. the measure used to evaluate the resemblance between the two matrices, is:

$$z_M = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} x_{ij} y_{ij}$$

where *i* and *j* are row and column indices of the resemblance matrices.

However, nowadays the Mantel test is generally computed using the **standardized Mantel *r* statistic**, whose formula is the same as that of Pearson's *r* correlation coefficient:

$$r_M = \frac{1}{d-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{x_{ij} - \bar{x}}{s_x} \right) \left(\frac{y_{ij} - \bar{y}}{s_y} \right)$$

where *i* and *j* are as above, *x*-bar, *y*-bar, s_x and s_y are the means and standard deviations of the distance values of each matrix,

and $d = n(n-1)/2$ is the number of distance or similarity measures in one of the upper triangular matrices.

Variants of the Mantel test can be computed using rank correlations.

5.4.2. Example

Let us imagine two similarity matrices between 4 objects:

	2	3	4		2	3	4
1	0.25	0.43	0.55		0.43	0.41	0.47
2		0.17	0.39			0.22	0.60
3			0.66				0.71
	"Species" matrix				"Remote sensing" matrix		

Figure 33 - Two fictitious similarity matrices between 4 objects.

Mantel's *z* statistic is computed as follows:

$$z = (0.25 \times 0.43) + (0.43 \times 0.41) + (0.55 \times 0.47) + (0.17 \times 0.22) + + (0.39 \times 0.60) + (0.66 \times 0.71) = 1.2823$$

This value (1.2823) is the "true" (observed) value, that must then be compared to a reference distribution obtained by randomly permuting (99 or 999 or 9999 times) the rows and corresponding columns of one of the two similarity matrices. **Beware**: the values of the similarity matrices cannot be permuted completely at random. The permutation scheme is actually **equivalent to permuting the raw data** and recomputing the similarities.

Finally, the observed *z* value is compared to the reference distribution in the same way as in the Pearson correlation example of Section 5.2, using the one-tailed hypothesis count. Indeed, if one has used two similarity matrices or two distances matrices (**not** a similarity and a distance matrix!), then the only meaningful alternative hypothesis in ecology is that the distances or similarities are *positively* correlated. A negative correlation between distance measures would mean that, for instance, the sites would be more similar as perceived by the living community when they are *less* similar with respect to the remote

sensing variables. This illustrates the specificities of the interpretation of a Mantel test, which must be based on a reasoning on association measures and not on raw data.

Additional remarks:

- like Pearson's correlation coefficient, the Mantel test also has a partial form, where the matrix correlation r_M(**AB.C**) between two matrices **A** and **B** is tested while controlling for the effect of matrix **C**. r_M(**AB.C**) is computed in the same way as a partial Pearson correlation coefficient;

- the Mantel test has sometimes be used to detect linear geographical gradients. The **Y** matrix was as usual (e.g. Bray-Curtis distance on species data). The **X** matrix contained Euclidean distances computed from the geographical coordinates of the sites. Note, however, that much more powerful techniques, based on raw data, are available nowadays to detect spatial structures, so that this approach must now be avoided.

Words of caution (Mantel test)

The paragraph below is an exerpt from a text by Pierre Legendre. It warns users against misuses of the Mantel test.

"Empiricists who frown upon theoretical justifications should be interested in the fact that the R^2_M of a Mantel test or a regression on distance matrices is always much lower than the R^2 of a (multiple) regression or canonical analysis computed on the raw data, when it is possible to do so; this has often been noted by users of the Mantel test. This was one of the results reported by Dutilleul *et al.* (2000, Table 2)[3]; it can easily be verified using any data set. Legendre (2000, Table

[3] Dutilleul, P., J. D. Stockwell, D. Frigon, and P. Legendre. 2000. The Mantel-Pearson paradox: statistical considerations and ecological implications. *Journal of Agricultural, Biological, and Environmental Statistics 5: 131-150.*

II)[4] has also shown that the power of a Pearson correlation (i.e., its capacity to reject the null hypothesis when H_0 is false) is much higher than the power of a simple Mantel test computed on distance matrices derived from the same data (...). Hence, whenever possible, use statistical procedures based on tables of raw data, such as correlation, regression, or canonical analysis. Save the Mantel test and derived forms to test hypotheses formulated in terms of distances."

Another paper has been published recently by Legendre et al. (2005)[5], comparing the performances of tests based on raw data and Mantel tests computed on distance matrices derived from the same data. The theoretical developments and simulation results presented in this paper led to the following observations:

(1) The variance of a community composition table is a measure of beta diversity.

(2) The variance of a dissimilarity matrix among sites is **neither** the variance of the community composition table **nor** a measure of beta diversity; hence, partitioning on distance matrices **should not** be used to study the variation in community composition among sites.

(3) In all of the simulations, partitioning on distance matrices underestimated the amount of variation in community composition explained by the raw-data approach.

(4) The tests of significance in the distance approach had less power than the tests of canonical ordination. Hence, the proper statistical procedure for partitioning the spatial variation of community composition data among environmental and spatial components, and for testing hypotheses about the origin and maintenance of variation in community composition among sites, is **canonical partitioning**.

[4] Legendre, P. 2000. Comparison of permutation methods for the partial correlation and partial Mantel tests. *Journal of Statistical Computation and Simulation 67: 37-73.*

[5] Legendre, P., D. Borcard and P. R. Peres-Neto. 2005. Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecological Monographs 75: 435-450.*

5.5 The controversy about variation partitioning

In 1994 Legendre *et al.*[6] proposed an extension of the Mantel approach, called multiple regression on distance matrices. This extension was devised for a specific context, phylogeny, where some hypotheses are explicitly stated in terms of distance matrices and cannot be restated in terms of raw data, and where the aim was to assess the influence of several "explanatory" distance matrices on a "response" distance matrix. The distance matrices are unrolled and treated as in multiple regression, but the significance test requires one of several complex permutation schemes. A Mantel *R*-square (R_M^2) can be computed, but can be used only as a measure of fit of the model.

In recent years, several researchers have proposed to use multiple regression on distance matrices to compute distance-based variation partitioning (e.g. Duivenvoorden *et al.* 2002[7], Tuomisto *et al.* 2003[8]). In some papers this technique was applied to hypotheses proper to the distance world, but in others it has been applied as an equivalent to raw-data based variation partitioning (this latter using canonical ordination, as explained in Chapter 4b).

In our 2005 paper cited above (Legendre, Borcard & Peres-Neto) we demonstrated that (1) the distance approach to partitioning is **not** equivalent to the partitioning of raw data; (2) The restatement of raw-data based hypotheses into the distance world leads to tests that are much less powerful than their equivalent in the raw-data world; (3) Therefore, whenever one can state hypotheses in terms of raw data, one should test these in the world of raw data and by all means avoid to translate them into distance-based hypotheses. Furthermore, following a Comment by Tuomisto & Ruokolainen[9] accusing us of

[6] Legendre, P., F.-J. Lapointe, and P. Casgrain. 1994. Modeling brain evolution from behavior: a permutational regression approach. Evolution **48**:1487-1499.
[7] Duivenvoorden, J. F., J.-C. Svenning, and S. J. Wright. 2002. Beta diversity in tropical forests. Science **295**:636–637.
[8] Tuomisto, H., K. Ruokolainen, and M. Yli-Halla. 2003. Dispersal, environment, and floristic variation of western Amazonian forests. Science **299**:241–244.
[9] Tuomisto, H., and K. Ruokolainen. 2006. Analyzing or explaining beta diversity? Understanding the targets of different methods of analysis. Ecology **87**:2697-2708.

confusing the issues, we have submitted another paper[10] where we demonstrate by new simulations that, even in the cases where the distance approach would be appropriate in terms of ecological hypotheses, distance-based variation partitioning still remains highly suspect on the statistical side for several reasons. Among these reasons are: (1) while an unbiased estimation of a fraction of variation is available in the raw data world (adjusted R^2), no equivalent exist or can be proposed in the distance world; hence, no unbiased estimation of a fraction of variation can be computed. (2) The use of the Mantel *R*-square (R_M^2) as a measure of the fraction of explained variation, and following that as the basis for computation of the fractions in variation partitioning, is mathematically highly suspicious and its validity has never been demonstrated. (3) Raw-data based R^2 are additive (as shown in Chapter 4b), but in the distance world they are not: "We now know how to partition the variation of a response matrix **Y** with respect to several explanatory matrices **X** using RDA. In raw data partitioning, an identical total fraction of explained variation is obtained, whether all explanatory variables are put in a single table **X** or they are divided into any number of subtables (environmental, spatial, etc.). The effects of the explanatory variables are thus additive. This is not the case in partitioning on distance matrices: different total amounts of explained variation for the response **Y** are obtained if one includes all explanatory variables in a single distance matrix or if separate distance matrices are computed for the various explanatory variables." (Legendre *et al.* in review).

[10] Legendre, P., D. Borcard & P. R. Peres-Neto (*in review*): Analyzing or explaining beta diversity: Comment. *Ecology*.