PERMANOVA

Permutational multivariate analysis of variance

A computer program
by Marti J. Anderson

Department of Statistics
University of Auckland
(2005)

DISCLAIMER

This FORTRAN program is provided without any explicit or implicit warranty of correct functioning. It has been developed as part of a university-based research program. If, however, you should encounter problems with this program, the author will be happy to help solve them. Researchers can use this program for scientific and research purposes, but intellectual property and copyright for the source code and program remains the property of Marti Jane Anderson. The program relies on several routines in FORTRAN from *Numerical Recipes* (Press et al. 1992). Namely, for finding the eigenvalues of a real symmetric distance matrix, using a householder reduction, it uses the routines "tred2" and "tqli". Also, for sorting, it uses a variation of the routine "piksr2". For random number generation, it uses "ran2". The program also uses several routines from *Applied Statistics*, namely: AS91 (ppchi2) for chi-squared deviates, AS111 (ppnd) for normal deviates, AS239 (gammad) for the incomplete gamma integral, AS66 (alnorm) for the tail area of a standard normal and AS245 (alngam) for the log of the gamma function.

Publications using this program should give credit to the method by referring to the following papers:

Anderson, M.J. 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26: 32-46.

McArdle, B.H. & Anderson, M.J. 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* 82: 290-297.

The present user's guide may be referred to as:

Anderson, M.J. 2005. PERMANOVA: a FORTRAN computer program for permutational multivariate analysis of variance. Department of Statistics, University of Auckland, New Zealand.

Author's contact details:

Dr Marti J. Anderson
Department of Statistics
University of Auckland
Private Bag 92019
Auckland, New Zealand
Tel: 64-9-373-7599 ext 85052
Fax: 64-9-373-7000
Email: mja@stat.auckland.ac.nz
Website: http://www.stat.auckland.ac.nz/~mja

Table of Contents

I. Introduction

PERMANOVA is a computer program for testing the simultaneous response of one or more variables to one or more factors in an ANOVA experimental design on the basis of any distance measure, using permutation methods. These notes for users assume knowledge of multi-factorial ANOVA, which has the same basic logic in multivariate as in univariate analysis, and an understanding of what it means to test a multivariate hypothesis. A more complete description of the method is given in Anderson (2001a) and McArdle & Anderson (2001). The program includes:

- choice of appropriate transformation and/or standardization of the data;
- choice of 19 distance (or dissimilarity) measures to use as the basis of the analysis;
- option to rank the distances in the distance matrix before the analysis;
- analysis and partitioning of the total sum of squares according to the full model, including appropriate treatment of factors that are fixed or random, crossed (orthogonal) or nested (hierarchical), and all interaction terms;
- correct calculation of an appropriate distance-based pseudo *F*-statistic for each term in the model, based on expected mean squares as in univariate ANOVA (Winer et al. 1991, Searle et al.1992);
- correct permutation procedures to obtain *P*-values for each term in the model, using the correct permutable units (Anderson & ter Braak 2003);
- choice of permutation method: raw data units or residuals under either a reduced or a full model (Anderson 2001b, Anderson & Legendre 1999, Anderson & Robinson 2001);
- correct *P*-values also obtained through Monte Carlo random draws from the asymptotic permutation distribution (Anderson & Robinson 2003);
- option to include one or more covariables (i.e., to perform ANCOVA or MANCOVA);
- pair-wise *a posteriori* comparisons of levels for single factors, including within individual levels of other factors in the case of significant interactions and the use of correct permutable units in each case.

II. Description of the test statistic

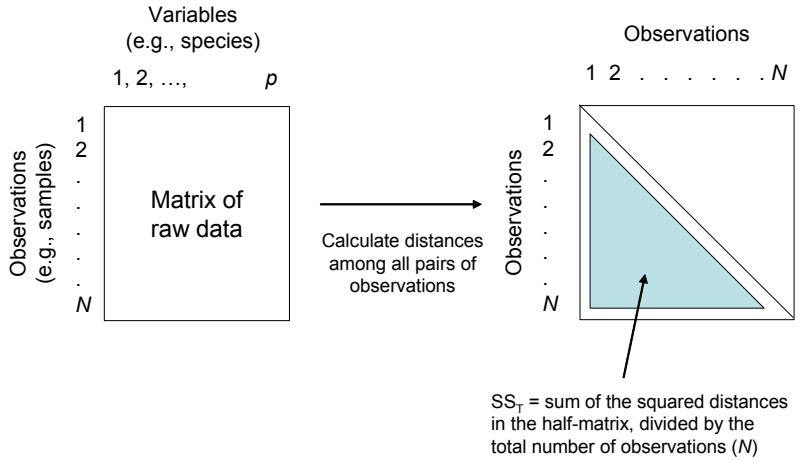
The program gives a partitioning of multivariate variation (defined by the distance measure used) according to individual factors in any fully balanced multi-way ANOVA design, with tests done by permutations. First, the program calculates the distances between each pair of observation units (sampling units) to obtain a distance matrix. It then calculates the test-statistics from this according to the relevant experimental design. Please refer to Anderson (2001a) and McArdle and Anderson (2001) for greater details describing the approach.

A. One-way Design

In the simplest situation of a one-way test (i.e. the test of a single factor), with *a* groups and *n* observation units (replicates) per group, let $N = an$ be the total number of observation units and let d_{ij} be the distance between observation *i* and observation *j*. Then, the total sum of squares is:

$$SS_T = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 \qquad (1)$$

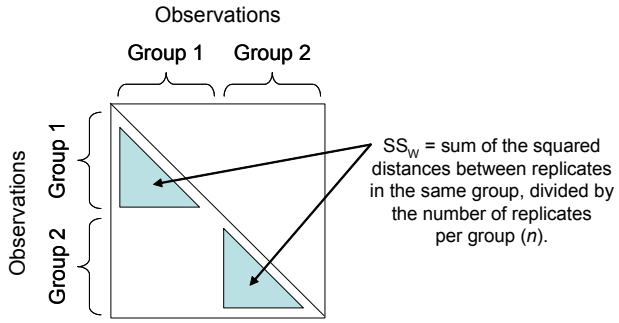
Thus, to calculate SS_T the program simply adds up the squares of all of the distances in the sub-diagonal half of the distance matrix and divides by the total number of observations (*N*), as illustrated below:



Next, the within-group sum of squares is:

$$SS_W = \frac{1}{n}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 \varepsilon_{ij} \tag{2}$$

where ε_{ij} takes the value of 1 if observation *i* and observation *j* are in the same group, otherwise it takes the value of zero. This amounts to adding up the squares of all the distances between observations that occur in the same group. Thus, schematically, for the case of two groups with equal sample sizes, this is:



Then, the among-group sum of squares is the difference: $SS_A = SS_T - SS_W$. A pseudo *F*-ratio associated with the test of this factor is then (Anderson 2001a):

$$F = \frac{SS_A/(a-1)}{SS_W/(N-a)} \tag{3}$$

where $(a-1)$ are the degrees of freedom associated with the factor and $(N-a)$ are the residual degrees of freedom.

Another way to describe the calculation of the test statistic (McArdle & Anderson 2001) is to consider ANOVA as a linear model (e.g., Neter et al. 1996) and apply this to a multivariate distance matrix after a transformation in the manner of Gower (1966). Let **Y** be the $(N \times p)$ matrix of *N* observation units by *p* variables. To do the analysis, the first step is to let $\mathbf{D} = (d_{ij})$ be an $(N \times N)$ distance matrix calculated from observation units of **Y**, using some chosen appropriate distance measure. Let $\mathbf{A} = (a_{ij}) = (-\frac{1}{2} d_{ij}^2)$, then calculate Gower's (1966) centered matrix (**G**) by centering the elements of **A**, i.e.,

$$\mathbf{G} = \left(\mathbf{I} - \tfrac{1}{N}\mathbf{1}\mathbf{1}'\right)\mathbf{A}\left(\mathbf{I} - \tfrac{1}{N}\mathbf{1}\mathbf{1}'\right) \tag{4}$$

where **1** is a column of 1's of length N and **I** is an $(N \times N)$ identity matrix.

In the one-way ANOVA case, let matrix **X** $(N \times (a-1))$ contain the design matrix, having $(a-1)$ orthogonal contrast vectors (i.e., which code for the a levels of the factor). Next, we calculate the "hat" or projection matrix $\mathbf{H} = \mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'$ (e.g. Johnson & Wichern 1992). Then, the test statistic used to test the null hypothesis of no difference among the a groups is the pseudo F-statistic (McArdle & Anderson 2001):

$$F = \frac{tr(\mathbf{HGH})/(a-1)}{tr[(\mathbf{I}-\mathbf{H})\mathbf{G}(\mathbf{I}-\mathbf{H})]/(N-a)} \tag{5}$$

where "*tr*" indicates the trace (sum of diagonal elements) of a matrix. This is the F-statistic that is used for the one-way case. ***Important:*** The F-statistics given in (3) and (5) above are equivalent. Also, in the special case of one variable and Euclidean distances having been used to produce **D**, either of these two equations (3 or 5) produces Fisher's traditional F-ratio.

B. Multi-factor Design

A good reason for considering the form of the statistic given in (5) is that it is very clear then how to partition the distance matrix according to a larger multi-factor design. Individual **X** design matrices can be coded for each term in the linear ANOVA model. Then, if a term to be tested has a numerator with design matrix $\mathbf{X}_n$ and a denominator with design matrix $\mathbf{X}_d$ (and corresponding "hat" matrices $\mathbf{H}_n$ and $\mathbf{H}_d$, respectively) then the test statistic is constructed as:

$$F = \frac{tr(\mathbf{H}_n\mathbf{G}\mathbf{H}_n)/df_n}{tr(\mathbf{H}_d\mathbf{G}\mathbf{H}_d)/df_d} \tag{6}$$

where df_n and df_d are the numerator and denominator degrees of freedom, respectively.

Once again, if there is only one variable in the analysis and one chooses to use Euclidean distances, then the resulting sums of squares and F-ratios are exactly the same as Fisher's univariate F-statistic in traditional ANOVA. Thus, although PERMANOVA was designed to do multivariate analysis on distance matrices, it can be used to do univariate ANOVA. However, there is a very important difference between PERMANOVA and traditional ANOVA: ***PERMANOVA calculates P-values using permutations***, rather than relying on tabled P-values, which assume normality. A nice way to familiarize oneself with the program is to do a traditional univariate ANOVA using another package and compare this with the outcome from the analysis of one variable (and based on Euclidean distance) using PERMANOVA.

III. Permutation tests

A. One-way Design

In general, the distribution of the multivariate F-statistic in (3) or (5) under the null hypothesis of no effect of the factor is unknown. Thus, to create a distribution of F under a true null hypothesis, a permutation procedure is used.

The idea of a permutation test is this: if there is no effect of the factor (that is, if the null hypothesis were true), then it is equally likely that any of the individual treatment labels could have been associated with any one of the observation units. We could have obtained the observations in any order by reference to the treatments, if the treatments did not matter. So, another possible value of the test-statistic under a true null hypothesis can be obtained by randomly shuffling the treatment labels onto different units. The random shuffling of labels is repeated a large number of times, and each time, a new value of F, which I will call F^{π}, is calculated. If the null hypothesis were true, then the F-statistic actually obtained with the real ordering of the data relative to the treatments will be similar to the values obtained under permutation. If, however, there is a significant effect of treatments, then the value of F obtained with the real ordering will appear large relative to the distribution of values obtained under permutation. In that case, the value of F for our data is unlikely to have been obtained if the null hypothesis were true.

The frequency distribution of the values of F^{π} is discrete: that is, the number of possible ways that the data could have been re-ordered is finite. The probability associated with the test statistic under a true null hypothesis is calculated as the proportion of the F^{π} values that are greater than or equal to the value of F observed for the real data. In this calculation of a P-value, we include the observed value as a member of the distribution. This is because one of the possible random orderings of the data is the ordering we actually got! So the P-value is

$$P = \frac{\left(No.\ of\ F^{\pi} \geq F\right)+1}{\left(Total\ no.\ of\ F^{\pi}\right)+1} \qquad (7)$$

For multivariate data, the observations (usually entered as rows) of the data matrix (raw data) are simply permuted randomly among the groups. Note that permutation of the raw data for multivariate analysis does not mean that values in the data matrix are shuffled just anywhere. A whole observation (e.g., an entire row) is permuted as a unit; the exchangeable units are the labels associated with rows of the data matrix (e.g., Anderson 2001b). For the one-way test, enumeration of all possible permutations (re-orderings of the data) gives an exact P-value associated with the null hypothesis. In practice, the possible number of permutations is very large in most cases. A practical strategy, therefore, is to perform the test using a large random subset of values of F^{π}, drawn randomly, independently and with equal probability from the distribution of F^{π} for all possible permutations. PERMANOVA does not systematically do all permutations, but rather draws a random subset of them, with the number to be done being chosen by the user. Such a test is still exact in the sense that the probability of Type I error is still equal to the *a priori* chosen significance level (Dwass 1957).

B. Multi-factor Design

As for the one-way case, the distribution of the test-statistic in (6) under a given null hypothesis is also generally unknown. Thus, a permutation test or some other test using re-sampling methods is desirable. The method used by the program is to permute the units identified by the denominator term of the F-ratio, as described in detail by Anderson & ter Braak (2003).

When there is more than one factor, situations commonly arise which prevent the possibility of obtaining an exact test of individual terms in the model using permutations. For example, there is no exact permutation test for an interaction (although Pesarin 2001 describes some special cases where they may be constructed). However, several good approximate permutation methods can be used instead to get accurate P-values (Anderson & Legendre 1999, Anderson & ter Braak 2003). PERMANOVA provides three general options regarding the method of permutation to be used: (1) permutation of raw data, (2) permutation of residuals under a reduced model, or (3) permutation of residuals under a full model. These methods are described in detail elsewhere (Manly 1997, Anderson & Legendre 1999, Anderson & ter Braak 2003). Although these methods do not give "exact" P-values, they are asymptotically exact and give very reliable results. An exact test is a test for which the Type I error is exactly equal to the *a priori* significance level (α) chosen for the test. For an asymptotically exact test, the Type I error asymptotically approaches α with increases in sample size.

In practice, these three approaches will give very similar results, so there is no need to agonise too much in making a choice here. Importantly, *all three* of the methods ensure that the correct permutable units are used for each individual test (e.g., Anderson & ter Braak 2003). The following summary of the known properties of these methods can be used as a guide.

1. Permutation of raw data

This is a good approximate test proposed for complex ANOVA designs by Manly (1997). It will generally have Type I error close to α, although with larger sample sizes it tends to be more conservative (less powerful) than the tests that permute residuals (Anderson & ter Braak 2003). However, this method does not need large sample sizes to work well (Gonzalez & Manly 1998). It is also, computationally, the fastest option.

2. Permutation of residuals under a reduced model

This method empirically gives the best power and the most accurate Type I error for complex designs in the widest circumstances (Anderson & Legendre 1999, Anderson & ter Braak 2003). Also, this method is theoretically the

closest to the conceptually exact test (Anderson & Robinson 2001). The approach was first described for linear models by Freedman & Lane (1983).

3. Permutation of residuals under a full model
This method was described by ter Braak (1992); see also the descriptions given by Manly (1997) and Anderson & Legendre (1999). The general idea is to obtain residuals of the full model by subtracting from each replicate the mean corresponding to its particular cell (where a "cell" means a particular combination of factor levels). These residuals are estimating the errors associated with each replicate. These are then permuted and the statistic is re-calculated *for the errors alone* under permutation. This method generally gives results highly comparable to method (2). It relies somewhat more than (2) on large within-cell sample sizes for accuracy. It has the advantage, however, of being faster than method (2) for the analysis of the entire design.

Methods (2) and (3) both require calculation of residuals by estimating means and subtracting the observations from these means. When sample sizes are small, these estimates of means are not very good (i.e. they are not very close to the "true" means), so the residuals we are permuting are not really very close to the "true" errors (Anderson & Robinson 2001). Thus, method (1) is to be recommended in the case of relatively small sample sizes. Note also that "less powerful" does not necessarily mean that using (1) will give you a smaller *P*-value than (2) or (3) for any particular data set. It means that, in repeated simulations, the empirical power (estimated probability of rejecting the null hypothesis when it is false) was in some cases less for method (1) than for either of the other two methods.

IV. Monte Carlo *P*-values

In some situations, there are not enough possible permutations to get a reasonable test. Consider the case of two groups, with three observations per group. There are a total of 6 observations, so the total number of possible permutations is 6! = 720. However, with a groups and n replicates per group, the number of distinct possible outcomes for the *F*-statistic in the one-way test is $(an)!/[a!(n!)^a]$ (e.g., Clarke 1993), which in this case is: $(6)!/[2!(3!)^2] = 10$ unique outcomes. This means that even if the observed value of *F* is quite large, the smallest possible *P*-value that can be obtained is $P = 0.10$. This is clearly insufficient to make statistical inferences at a significance level of 0.05.

One alternative is to use the result given in Anderson & Robinson (2003) regarding the asymptotic permutation of the numerator (or denominator) of the test statistic under permutation (see equations (1) and (4) on p. 305 of Anderson & Robinson 2003). It is demonstrated that $tr(\mathbf{HGH})$ has, under permutation, an asymptotic distribution that is a linear form in chi-square variables, where the coefficients are the eigenvalues from the singular value decomposition of matrix **G**. Thus, chi-square variables can be drawn randomly and independently, using Monte Carlo sampling, and these can be combined with the eigenvalues to construct the asymptotic permutation distribution for each of the numerator and denominator and, thus, for the entire *F*-statistic, in the event that too few actual unique permutations exist. This is a very useful tool.

PERMANOVA gives the permutation *P*-value and also the Monte Carlo asymptotic *P*-value for each test it performs. When there is a large number of possible permutations, then these two *P*-values should be very close to one another, essentially converging on the same answer. When, on the other hand, there are very few possible permutations, then the *P*-value associated with the permutation test may be quite different, because of this limitation, and so the Monte Carlo *P*-value should be used in preference.

In multi-factor designs, it is not always easy to calculate how many unique permutations there are for a given term in the model. (Believe me, I have tried!) It depends not only on the permutation method chosen, but may also depend on other terms in the model, how many levels they have and whether they may, under permutation, coincide with the term being tested in serendipitous ways. Therefore, for each test, PERMANOVA tells you how many permutable (exchangeable) units were being shuffled, and it also keeps track of how many unique values of the test statistic it encounters out of the total number of random permutations done. Armed with this information, it is then possible to judge how reasonable the permutation *P*-value given actually is, under the circumstances. For example, if you ask the program to perform 999 permutations and, for a given term, you see that 948 of these ended up being unique, then the permutation *P*-value is perfectly fine for obtaining a rigorous inference. On the other hand, if you do 999

permutations and only 10 of these turned out to be unique, you would be wise to choose to use the Monte Carlo *P*-value for that term instead.

V. Assumptions

Recall that for traditional one-way ANOVA, the assumptions are that the errors are independent, that they are normally distributed with a mean of zero and a common variance, and that the treatment effects are additive. In the case of a one-way analysis, the PERMANOVA test using permutations assumes only that ***the observation units are exchangeable under a true null hypothesis***. There are no explicit assumptions regarding the distributions of the original variables; they are certainly not assumed to be normally distributed. However, implicit in the notion of exchangeability is the notion of independence, for if observations are correlated with one another (e.g., temporally or spatially), then randomly shuffling them will destroy this kind of inherent structure, if it is there. Thus, in general, we would assume that the *observation units* are independent of one another. In contrast, we do not at all expect the individual variables which have been measured on the same observation units (in the multivariate case) to be independent of one another, and this is not assumed.

Although there is also no explicit assumption regarding the homogeneity of spread within each group, PERMANOVA, like ANOSIM (Clarke 1993), will be sensitive to differences in spread (variability) among groups. Thus, if a significant difference between groups is detected using PERMANOVA, then this could be due to differences in location, differences in spread, or a combination of the two. Perhaps the best approach is to perform a separate test for homogeneity (e.g., using the program PERMDISP) including pair-wise comparisons, as well as examining the average within and between-group distances and associated MDS plots. This will help to determine the nature of the difference between any pair of groups, whether it be due to location, spread, or a combination of the two.

The strategy of PERMANOVA is to apply the traditional ANOVA partitioning procedure to the distance matrix, with *P*-values obtained using permutations. If the user chooses to ***rank*** the values in the distance matrix before proceeding, then the partitioning is done on these ranks, instead of on the raw distances. Therefore, in the one-way case, when ranks are chosen, PERMANOVA is truly non-parametric and will give results effectively indistinguishable from those given by ANOSIM.

When there is more than one factor, however, PERMANOVA is no longer, strictly speaking, “non-parametric”, but rather is “semi-parametric”. The reason for this is that it fits the full linear additive model to the distance matrix. This is necessary if we wish to measure and test interaction terms. Thus, although tests are done using permutations, we do still estimate “parameters” in some sense by fitting such a model. In contrast, ANOSIM does not try to fit any parameters; it does not attempt to model or test interaction terms. Thus, to take the extra step of modeling more complex designs using PERMANOVA, we unfortunately need to give up a pure non-parametric framework. This is just like the situation in univariate statistics – there can be no purely non-parametric test for an interaction term. However, PERMANOVA retains as many of the good properties of the non-parametric approach as possible in terms of lack of assumptions and flexibility. PERMANOVA is still “distribution free”, like ANOSIM, even for complex designs, because it relies on permutation procedures to obtain *P*-values and can be based on any distance measure. However, due to the direct modeling of the distance matrix, the choice of distance measure used for PERMANOVA is therefore very important, and should be chosen with care. Ranking the distances before proceeding will tend to make the test more robust, but will still mean that a linear model is being fitted that is additive on these ranks.

VI. Input files

Two input files are required by the program:

1) a file containing the details of the ***experimental design***; and
2) a file containing the ***data*** (response variables or a distance matrix).

If you choose to include one or more covariables in the analysis (i.e. to perform ANCOVA or MANCOVA), then a ***third*** file containing the ***covariable(s)*** is also required.

A. Design file

The first file the program requires is the design file. This file contains all of the information that the program needs regarding the factors in the experimental design, whether they are fixed or random, crossed (orthogonal) or nested in one another. The file is very easily created by hand in a text editor (such as "Notepad" or "SimpleText"). It should be saved as an ASCII text (*.txt) file. The design file contains the following information:

1. In the first line is given an integer, which is the number of factors
2. Next, a line is given for each factor in the design, *in the order that they occur in the data file*, providing the following information separated by tabs:
    a. the name of the factor;
    b. the number of levels of the factor;
    c. whether the factor is crossed ("C") or nested ("N") in some other factor(s);
    d. whether the factor is fixed ("F") or random ("R");
    e. if the factor is nested, then the number(s) corresponding to the factor(s) within which it is nested must be listed, *with no spaces or tabs in between these numbers*. If the factor is not nested, then a zero is given here ("0").
3. The last line contains the number of replicate observation units per cell (sample size, *n*).

Note that the current version of the program will only handle balanced experimental designs: that is, (i) an equal number of observations per cell; and (ii) no cells (combination of factor levels) in the design can be missing.

Here is an example of a design file for a two-way fixed-factor ANOVA:

```
2
Position    2     C     F     0
Shade       3     C     F     0
4
```

The first line indicates that there are 2 factors. The second line indicates that the name of factor 1 is "Position", it has 2 levels, is crossed ("C"), fixed ("F"), and the zero ("0") at the end indicates that this factor is not nested in any other factors. The third line gives similar information, but now for factor 2, called "Shade", which has 3 levels, is crossed, fixed and is (therefore) also not nested in anything else (thus the zero "0" at the end). The last line gives the number of observations per cell, $n = 4$.

Notice that:

- The number of cells in the full design will be the product of the number of levels of the factors given. In this case, the number of cells is $2 \times 3 = 6$.
- The total number of observation units (either rows or columns) in the full data matrix will be equal to the product of the total number of cells times the number of observations per cell. In this case, the total number of observations is $N = 2 \times 3 \times 4 = 24$.

Here is another example of a design file. This one is for a hierarchical sampling design:

```
3
Location    4     C     R     0
Site        2     N     R     1
Area        2     N     R     12
5
```

Here, there are 3 factors. Factor 1 is "Location", it has 4 levels, is crossed ("C") and is random ("R"). Factor 2 is "Site", which has 2 levels and is also random; it is nested ("N") in factor 1 (notice the "1" at the end of this line). Factor 3 is "Area"; it has 2 levels, is random and is nested in factor 1 and in factor 2. Notice the "12" at the end of this line. This is not the number "twelve"! It is to indicate that the individual areas belong to (are nested within) sites *and* locations (factor 1 and factor 2). Finally, the last line of this design file indicates that there are $n = 5$ replicates

within each combination of levels of the factors. So, in this example, there are a total of $4 \times 2 \times 2 = 16$ cells and a total of $N = 4 \times 2 \times 2 \times 5 = 80$ observations.

IMPORTANT
To ensure that the program can read the design file correctly:
- Use capital letters only to designate "R", "F", "C" and "N".
- The names of factors should not contain any spaces in them and should not exceed 25 characters in length.

The current ***limitations*** to the ANOVA designs that can be analysed using PERMANOVA are:
- The number of factors must not be greater than 9.
- The design must have replication within cells (i.e. $n \geq 2$).
- The sample size must be equal within each cell (i.e. balanced designs only)
- All cells in the design must be filled: no cells (combinations of factors given) can be missing.

Thus, if you have an unbalanced design (some cells or some observations missing), if you have a randomized block or split-plot type of design, or if you have an asymmetrical design, then you must use the program DISTLM with appropriate design (**X**) matrices to implement tests of individual terms in your model one at a time.

B. Data file
For the data file, the program allows the user to input either a distance matrix (**D**) directly or a raw multivariate data matrix of response variables (**Y**). In either case, the file should be saved as ASCII text (*.txt), ***with no column or row headings***.

If you are using a Macintosh, then save the file (for example, from Excel) as "Text (Windows)". If you save it as a text file from Excel, it won't run unless you open up the file again (for example, using a text editor such as BBEdit) and manually add a carriage return ("Enter") to the end of the last line. This is a weird quirk of the way the Mac version of Excel saves text files that, for some reason unknown to me, causes my program to crash. You should not experience any such problem, however, if the file is saved as "Text (Windows)" or if the file was created using a different text editor.

If a file with a raw data matrix is input, either the rows or the columns may correspond to the variables for the analysis; the user will be given the option to choose one of these. Then, the user has several options for transformation, standardization and choice of distance measure.

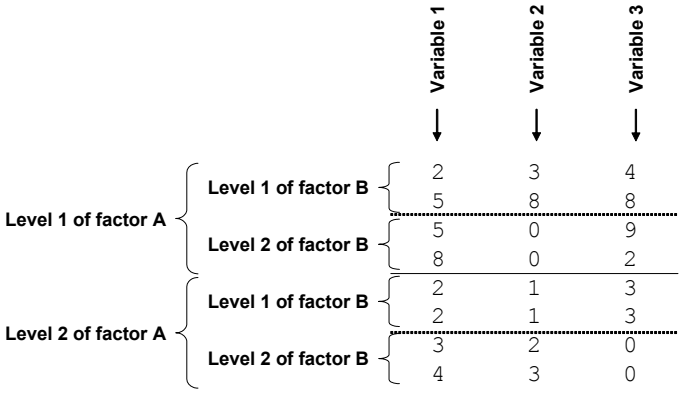
Perhaps the most important point about the data file is that ***the order of the observations it contains must match the order of the factors given in the design file***. For example, imagine that you have 3 response variables and you have the following design file:

```
2
A       2       C       F       0
B       2       C       F       0
2
```

Thus, there are a total of 8 observations. Your tab-delimited data file (*.txt) may look like this (where variables are columns and observation units are rows):

```
2       3       4
5       8       8
5       0       9
8       0       2
2       1       3
2       1       3
3       2       0
4       3       0
```

Note there are no column or row headers here. More importantly, note that the structure of this file is:



Notice that the first split in the data is associated with factor A, which is also the first factor that is listed in the design file. The next splits in the data (within each level of factor A) are associated with the levels of factor B, which is the factor listed next in the design file. This structuring, expected by the program, applies generally for as many factors as are present in the design: be sure that the order of the factors listed in the design file corresponds to the way they are structured in the input data file. The use of Excel's "Sort" tool under "Data" can be helpful here. You will also need to take careful note of the order of the individual levels within each factor, so that if you choose to do pair-wise comparisons after the main analysis, you will know the order of the groups and therefore which groups are being compared.

C. Covariable file

In the case of ANCOVA or MANCOVA, you will be asked to provide a file that contains the covariable(s). This file should be an ASCII text (*.txt) file with no headers on either the rows or the columns. The covariables can either be columns or rows in this input file. The most important thing is that the observation (sampling) units must be in exactly the same order as their corresponding order in the original data file. The coincidence of the ordering of samples in the file containing covariable(s) with that in the data input file should be double-checked carefully before proceeding.

D. To Run the Program

To avoid dealing with long file names and paths to locate files, place the relevant input file(s) in the same location on your computer (i.e. the same directory) as the PERMANOVA.exe (or PERMANOVA.mac) file, for use with the program. Double-click on the "PERMANOVA.exe" (or "PERMANOVA.mac") file to run the program.

The program uses dynamic memory allocation, and so (theoretically) does not have any limits on the sizes of matrices (numbers of rows or columns) that may be used for the input files. For PCs, the program has been compiled to allocate a stack memory at run-time equal to 100 Mb. This should be sufficient for most requirements. If you require greater memory for your analysis, then contact the author and a compilation of the program with a larger stack allocation can be created for you.

If you are using a Macintosh and you get an error that reads:

```
BUFFER allocation failed
REWIND(UNIT=*,...
```

then you need to increase the memory allocated to the program. To do this, click on the program's icon and type "i" while holding the apple key, then choose "Memory" (or choose "Get Info > Memory" from the File menu). Depending on the size of your matrices, you might even have to increase this value to something obnoxious, like 50000 or so. However, if you have a large input file and cannot seem to get the program to work even after doing this, then please contact the author.

VII. Questions asked by the program

The questions asked by the program are best demonstrated by examples. The one given here consists of data from an experiment examining the effect of shade and the effect of proximity to the seafloor on assemblages of subtidal invertebrates and algae on hard surfaces near marinas (courtesy of Dr Tim Glasby). For further details, see Glasby (1999).

The experiment was a two-way crossed (orthogonal) design with $n = 4$ observations per cell, having the following structure:
Factor 1 = Position (fixed, 2 levels: far or near to the seafloor)
Factor 2 = Shade (fixed, 3 levels: shade, a procedural control which was a clear plexiglass shade, and no shade)

Organisms colonising subtidal 15cm x 15cm sandstone settlement plates were counted and a total of 46 taxa were included in analyses. Organisms that occurred less than twice were not included. The data matrix (24 rows (observations) x 46 columns (taxa)) is contained in an ASCII text file called "Tim.txt."

The questions asked by the program are given in `Courier` font, while responses given for this data set are in ***Times bold italics***. Note that you can stop the program at any time by typing Ctrl-C (as for any DOS program) or, if you are a Mac user and this does not work, then try typing a period (a full stop) while holding down the apple key.

```
PERMANOVA v. 1.6
------------------
A program for analysing multivariate data
on the basis of any distance measure,
according to any linear ANOVA model,
using permutations.

by M.J. Anderson
Department of Statistics
University of Auckland (2005)

Type the name of the input file containing the design.
```
Timdesign.txt
Note: the design file for this analysis is shown in the middle of page 10 above.

```
Type the name of the input file containing your data
(either RESPONSE VARIABLES or a DISTANCE MATRIX).
```
Tim.txt

```
Type a name for the output file of results (*.txt)
```
Results.txt

```
Nature of the data in the input file:
1) raw data (N x p)
2) distance matrix (N x N)
```
1

```
Structure of the input file:
1) rows are samples and columns are variables
2) columns are samples and rows are variables
```
1

```
How many variables (columns) are there?
```
46

```
Choice of transformation:
1) none
```

```
2) square-root
3) fourth-root
4) ln(x)
5) ln(x+1)
6) log10(x)
7) log10(x+1)
5) presence/absence
3

Choice of standardisation:
1) none
2) standardise by row (sample) sums
3) standardise by column (variable) sums
4) double standardise by row and column sums
5) standardise each variable to z-scores (normalize)
6) standardise each variable by dividing by its s.d.
7) standardise each variable by dividing by its range
1

Choice of distance measure:
1)  Bray-Curtis dissimilarity
2)  square root of Bray-Curtis
3)  Euclidean distance
4)  Orloci's Chord distance
5)  Chi-square metric
6)  Chi-square distance
7)  Hellinger distance
8)  Gower dissimilarity
9)  Gower, excluding double zeros
10) Canberra distance
11) square root of Canberra distance
12) CY dissimilarity
13) Deviance based on the binomial
14) Deviance per observation (scale invariant)
15) Kulczynski dissimilarity
16) McArdle-Gower dissimilarity
17) Manhattan (city block) distance
18) Jaccard dissimilarity
19) Genetic distance (allozyme data)
1

Would you like to replace the distances with
their ranks before proceeding with the analysis?
1) No
2) Yes
1

Do you want to output the distance matrix?
1) No
2) Yes
1

Are there covariables in this design?
1) No
2) Yes
1

Do you wish to output the Monte Carlo mean squares
and the mean squares under permutation for each term?
1) No
2) Yes
1
```

Note: This is a useful option only if you wish to reconstruct individual *F*-ratios or to determine new *F*-ratios as linear combinations of particular mean squares for each permutation.

```
How many permutations do you want for the test?
(i.e. 99, 499, 999, 4999, etc.)
```
9999

Note: It has been recommended by Manly (1997) that at least 999 permutations should be used for tests at an α-level of 0.05, while at least 4999 permutations should be used for tests at an α-level of 0.01.

```
Which general method of permutation do you want?
1) Unrestricted permutation of raw data
2) Permutation of residuals under the reduced model
3) Permutation of residuals under the full model
```
1

```
Type an integer to be used as the seed
for the random permutations
```
5

Note: Any random integer value will do here in order to set the seed for the computer to choose a random subset of all possible permutations. The advantage to this is that if for some reason you wish to repeat the exact procedure you once used to produce a particular outcome, then you can do so by choosing the same integer here in a re-analysis of the data. A different choice for the random seed will give a different random subset of the possible permutations, thus leading to (very slightly) different *P*-values.

The program will then proceed to do the complete analysis and then to print the table of results simultaneously to the screen and to the output file. In this case, the information on the screen looks like this:

```
Please wait while I do a few preliminary calculations...

Calculating...
Please wait while I do the permutations...

There are 3 terms in the model
Working on term no.  1 of  3, perm. no.    1
Working on term no.  1 of  3, perm. no.    2
Working on term no.  1 of  3, perm. no.    3
.
.
.
Working on term no.  3 of  3, perm. no. 9997
Working on term no.  3 of  3, perm. no. 9998
Working on term no.  3 of  3, perm. no. 9999

--- Experimental Design ---
 Factor 1 is Position with  2 levels and is fixed
 Factor 2 is Shade    with  3 levels and is fixed
 The sample size (n)           =    4
 The total no. of observations =   24
 The total no. of variables    =   46

--- Results ---
 Permutational Multivariate Analysis of Variance

 Source             df        SS           MS          F      P(perm) P(MC)
------------------------------------------------------------------------------
 Po                  1     5595.3982    5595.3982   13.5360  0.0001  0.0001
 Sh                  2     3566.4372    1783.2186    4.3139  0.0006  0.0007
 PoxSh               2     1238.9368     619.4684    1.4986  0.1462  0.1544
 Residual           18     7440.6619     413.3701
 Total              23    17841.4341
------------------------------------------------------------------------------
```

```
 Data were transformed to fourth root
 No standardisation
 Analysis based on Bray-Curtis dissimilarities
 Unrestricted permutation of raw data using correct permutable units
 Integer used as seed     =    5
 No. of permutations used = 9999

Do you want to do pairwise comparisons?
1) No
2) Yes
2
```

Note: Each of the main effects is statistically significant, according to the above table, so pairwise comparisons of each of these would be good to do.

```
Which term would you like to investigate
with pairwise comparisons?
  1) Po
  2) Sh
  3) PoxSh
1

How many permutations would you like for the tests?
9999

Type an integer for the random seed for permutations
4

Now working on tests for set no.  1 of  1 sets

These results have been sent to the output file

Would you like to do another set of pairwise tests?
1) No
2) Yes
2

Which term would you like to investigate
with pairwise comparisons?
  1) Po
  2) Sh
  3) PoxSh
2

How many permutations would you like for the tests?
9999

Type an integer for the random seed for permutations
8

Now working on tests for set no.  1 of  1 sets

These results have been sent to the output file

Would you like to do another set of pairwise tests?
1) No
2) Yes
1

All results have been sent to the output file.
End of program.
Type q to quit.
q
```

VIII. Output file

After completing PERMANOVA, open the output file to see all of the results, including the *a posteriori* comparisons. The output file contains quite a lot of useful information, including:

- the names of files used;
- the details of the experimental design;
- the permutational MANOVA table of results;
- choices made for transformation, standardization, distance measure, integer seeds, permutation method, etc.;
- details of expected mean squares of the model, including the term used for the denominator MS for each test;
- the number of permutable units and the number of unique values in the permutation distribution for each term;
- results of pairwise comparisons.

The *F*-ratio and associated *P*-value for each term in the analysis can be interpreted in the same way that one would interpret the result of a univariate analysis of variance, but it is the multivariate hypothesis on the basis of the chosen distance measure that is actually being tested for each term in the model.

Results of pair-wise *a posteriori* tests are also given in the output file. Here, the multivariate version of the *t*-statistic (based on distances) is used. Once again, these values can generally be thought of and interpreted in the same manner as a univariate *t*-statistic, but by reference to the multivariate hypothesis.

It is generally a good idea to name the output file something with the extension *.txt on the end. That means that when one is working in Windows, double-clicking on the output file brings it up automatically in Notepad, for example, for easy examination and printing, if desired.

In the case of Tim's data, the file "Results.txt" looks like this:

```
 PERMANOVA v.1.6
-----------------
A program for analysing multivariate data on the basis of any distance measure,
according to any linear ANOVA model, using permutations.

by M.J. Anderson
Department of Statistics
University of Auckland (2005)


Input file of design information: Timdesign.txt
Input file of data: Tim.txt

--- Experimental Design ---
 Factor 1 is Position with  2 levels and is fixed
 Factor 2 is Shade    with  3 levels and is fixed
 The sample size (n)           =    4
 The total no. of observations =   24
 The total no. of variables    =   46

--- Results ---
 Permutational Multivariate Analysis of Variance

 Source              df         SS           MS          F      P(perm) P(MC)
-------------------------------------------------------------------------------
 Po                   1      5595.3982    5595.3982   13.5360  0.0001  0.0001
 Sh                   2      3566.4372    1783.2186    4.3139  0.0006  0.0007
 PoxSh                2      1238.9368     619.4684    1.4986  0.1462  0.1544
 Residual            18      7440.6619     413.3701
 Total               23     17841.4341
-------------------------------------------------------------------------------
 Data were transformed to fourth root
 No standardisation
 Analysis based on Bray-Curtis dissimilarities
```

```
 Unrestricted permutation of raw data using correct permutable units
 Integer used as seed     =    5
 No. of permutations used = 9999


--- Details of the expected mean squares (EMS) for the model ---

 Source                                Terms included in the EMS
-------------------------------------------------------------------------------
 Po                        = 1         R + 1
 Sh                        = 2         R + 2
 PoxSh                     = 12        R + 12
 Res                       = R         R
-------------------------------------------------------------------------------


                          #permutable  #unique vals   Term used for
 Source                      units      in perm dist  denom MS in F-ratio
-------------------------------------------------------------------------------
 Po                            24           9947       Res
 Sh                            24           9940       Res
 PoxSh                         24           9925       Res
-------------------------------------------------------------------------------


 --- Results ---
 Pair-wise a posteriori comparisons
-----------------------------------------------------------
 Term chosen: Po
 Name of the factor being tested: Position
 No. of sets of pairwise comparisons =       1
 No. of groups compared within each set =    2
 Total no. of tests done =                   1
 No. of raw observations per group =        12
 No. of permutable units per group =        12
 No. of permutations done =               9999
 Integer chosen for the random seed =        4
 Permutation of raw data using appropriate permutable units
-----------------------------------------------------------


 Tests among levels of the factor Position
-----------------------------------------------------------
   Groups         t        P_perm     P_MC    #unique vals
-----------------------------------------------------------
  ( 1, 2)       3.1705     0.0001     0.0001     9874
-----------------------------------------------------------

 Average dissimilarities within/between groups
        1       2
   1  29.528
   2  43.556  34.565


 Note: The pair-wise tests have not been corrected
       for multiple comparisons.


 --- Results ---
 Pair-wise a posteriori comparisons
-----------------------------------------------------------
 Term chosen: Sh
 Name of the factor being tested: Shade
 No. of sets of pairwise comparisons =       1
 No. of groups compared within each set =    3
 Total no. of tests done =                   3
 No. of raw observations per group =         8
 No. of permutable units per group =         8
 No. of permutations done =               9999
 Integer chosen for the random seed =        8
 Permutation of raw data using appropriate permutable units
```

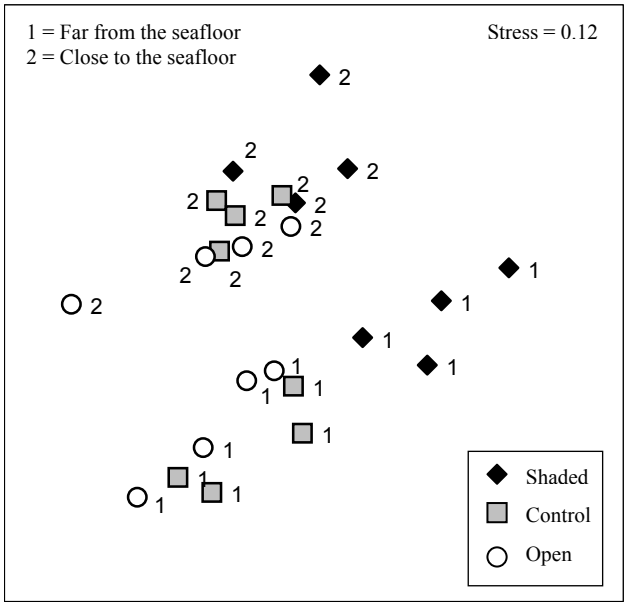
```
------------------------------------------------------------


 Tests among levels of the factor Shade
------------------------------------------------------------
   Groups         t        P_perm     P_MC     #unique vals
------------------------------------------------------------
  ( 1, 2)       1.7833     0.0151     0.0213     5052
  ( 1, 3)       1.9869     0.0021     0.0079     5054
  ( 2, 3)       0.8661     0.5626     0.5441     5042
------------------------------------------------------------

 Average dissimilarities within/between groups
        1       2       3
   1  36.041
   2  40.458  36.204
   3  41.610  35.080  35.062


 Note: The pair-wise tests have not been corrected
       for multiple comparisons.
```

The interpretation of the analysis is that there was no significant multivariate interaction between the factors of Position and Shade. Assemblages of organisms on settlement plates near the bottom were extremely different from those far away from the bottom. Shade also had a significant effect, with assemblages on shaded plates found to be significantly different from those in either the procedural control or on the unshaded plates, which themselves did not differ (i.e. Groups 2 and 3 above). These results are supported by a visual assessment of patterns in a non-metric MDS plot of fourth-root transformed data using Bray-Curtis distances, as shown below:



IX. Description of distance measures

There is a choice in PERMANOVA among 19 different possible distance (or dissimilarity) measures that one may use as the basis for the analysis. Due to the fact that many of these measures are defined in different ways by different authors and the potential confusion that may arise, this section provides an explicit mathematical description of each one, as calculated by this computer program, for reference. Further details on these and other distance measures can be found in the following references: Bray & Curtis (1957), Gower (1971), Hajdu (1981),

Gower & Legendre (1986), Faith et al. (1987), Clarke (1993), Cao (1997), Legendre & Legendre (1998), Legendre & Gallagher (2001) and Anderson & Millar (2004).

For the following, let $\mathbf{Y} = (y_{ik})$ be the $(N \times p)$ matrix of $i = 1, \ldots, N$ observation units (rows) by $k = 1, \ldots, p$ variables (columns). Also, dots will be used to denote summation over particular subscripts, thus: $y_{\bullet k} = \sum_{i=1}^{N} y_{ik}$ denotes the sum across all rows for column k, $y_{i\bullet} = \sum_{k=1}^{p} y_{ik}$ denotes the sum across all columns for row i, and the sum of all values in the matrix is $y_{\bullet\bullet} = \sum_{i=1}^{N} \sum_{k=1}^{p} y_{ik}$. Each dissimilarity or distance measure will be given as a value d_{ij} between observation units i and j, where j (like i) goes from 1 up to N.

1) Bray-Curtis dissimilarity

$$d_{ij} = \frac{\sum_{k=1}^{p} \left| y_{ik} - y_{jk} \right|}{\sum_{k=1}^{p} (y_{ik} + y_{jk})}$$

2) square root of Bray-Curtis

$$d_{ij} = \sqrt{\frac{\sum_{k=1}^{p} \left| y_{ik} - y_{jk} \right|}{\sum_{k=1}^{p} (y_{ik} + y_{jk})}}$$

3) Euclidean distance

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (y_{ik} - y_{jk})^2}$$

4) Orloci's Chord distance

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(\frac{y_{ik}}{\sqrt{\sum_{k=1}^{p} y_{ik}^2}} - \frac{y_{jk}}{\sqrt{\sum_{k=1}^{p} y_{jk}^2}} \right)^2}$$

5) Chi-square metric

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \frac{1}{y_{\bullet k}} \left(\frac{y_{ik}}{y_{i\bullet}} - \frac{y_{jk}}{y_{j\bullet}} \right)^2}$$

6) Chi-square distance

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \frac{y_{\bullet\bullet}}{y_{\bullet k}} \left(\frac{y_{ik}}{y_{i\bullet}} - \frac{y_{jk}}{y_{j\bullet}} \right)^2}$$

7) Hellinger distance

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(\sqrt{\frac{y_{ik}}{y_{i\bullet}}} - \sqrt{\frac{y_{jk}}{y_{j\bullet}}} \right)^2}$$

8) Gower dissimilarity

$$d_{ij} = \sum_{k=1}^{p} \left\{ \left| y_{ik} - y_{jk} \right| / R_k \right\}$$

where R_k = the range across the entire data set for variable *k*.

9) Gower, excluding double zeros

$$d_{ij} = \frac{\sum_{k=1}^{p} \left\{ w_k \left| y_{ik} - y_{jk} \right| / R_k \right\}}{\sum_{k=1}^{p} w_k}$$

where R_k is defined as in measure 8 and w_k is a weight for each variable:
$w_k = 0$ if both $y_{ik} = 0$ and $y_{jk} = 0$, else $w_k = 1$.

10) Canberra distance

$$d_{ij} = \sum_{k=1}^{p} \frac{\left| y_{ik} - y_{jk} \right|}{(y_{ik} + y_{jk})}$$

Note that the Bray-Curtis measure is a ratio of sums, while the Canberra measure is a sum of ratios.

11) square root of Canberra distance

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \frac{\left| y_{ik} - y_{jk} \right|}{(y_{ik} + y_{jk})}}$$

12) CY dissimilarity

$$d_{ij} = \frac{1}{s_{ij}} \sum_{k=1}^{p} \frac{1}{n_k} \left\{ n_k \log\left(\frac{1}{2}\right) - y_{ik} \log\left(\frac{y_{jk}}{n_k}\right) - y_{jk} \log\left(\frac{y_{ik}}{n_k}\right) \right\}$$

where $n_k = (y_{ik} + y_{jk})$ and where s_{ij} = the number of variables with non-zero values across the two observations *i* and *j*. Note that this measure requires the user to choose a value to replace zeros, wherever they occur, because the log of zero is not defined.

13) Deviance based on the binomial

$$d_{ij} = \sum_{k=1}^{p} \left\{ n_k \log\left(\frac{1}{2}\right) - y_{ik} \log\left(\frac{y_{ik}}{n_k}\right) - y_{jk} \log\left(\frac{y_{jk}}{n_k}\right) \right\}$$

where n_k is defined as in measure 12. Note that zero-replacement is not necessary here, as $y\log(y)$ is defined to be zero when $y = 0$.

14) Deviance per observation (scale invariant)

$$d_{ij} = \sum_{k=1}^{p} \frac{1}{n_k} \left\{ n_k \log\left(\frac{1}{2}\right) - y_{ik} \log\left(\frac{y_{ik}}{n_k}\right) - y_{jk} \log\left(\frac{y_{jk}}{n_k}\right) \right\}$$

where n_k is defined as in measures 12 and 13. Note that zero-replacement is not necessary here, as $y\log(y)$ is defined to be zero when $y = 0$.

15) Kulczynski dissimilarity

$$d_{ij} = 1 - \left(\frac{1}{2}\right)\left(\frac{M}{y_{i\bullet}} + \frac{M}{y_{j\bullet}}\right)$$

where $M = \sum_{k=1}^{p} \min(y_{ik}, y_{jk})$, and $\min(y_{ik}, y_{jk})$ is the minimum value for variable *k* out of the pair of values (y_{ik}, y_{jk}).

16) McArdle-Gower dissimilarity

First, the data are transformed to: $y' = \log_{10}(y^*) + 1$, where $y^* = y$ unless $y = 0$, in which case $y^* = 0.1$. Then the following distance is calculated on these transformed values:

$$d_{ij} = \frac{\sum_{k=1}^{p} \left\{ w_k \left| y'_{ik} - y'_{jk} \right| \right\}}{\sum_{k=1}^{p} w_k}$$

where $w_k = 0$ if both of the original values $y_{ik} = 0$ and $y_{jk} = 0$, else $w_k = 1$.

17) Manhattan (city block) distance

$$d_{ij} = \sum_{k=1}^{p} \left| y_{ik} - y_{jk} \right|$$

18) Jaccard dissimilarity

First, the data are transformed to presence/absence, as follows: $y^+ = 1$ if $y > 0$ and $y^+ = 0$ if $y = 0$. Then, the following distance measure is calculated on the transformed data:

$$d_{ij} = \frac{\sum_{k=1}^{p} \left| y_{ik}^+ - y_{jk}^+ \right|}{\sum_{k=1}^{p} w_k}$$

where $w_k = 0$ if both of the original values $y_{ik} = 0$ and $y_{jk} = 0$, else $w_k = 1$.

19) Genetic distance (allozyme data)

This distance measure has yet to be published but was recently invented by me with colleagues M. Virgilio & M. Abbiati (University of Bologna, Ravenna, Italy). It is unlike the others in that it requires a particular type of genetic data as input. Data must be provided as an ASCII text (*.txt) file, but where the alleles for each locus are the *p* variables (usually columns) and individuals are the *N* observation units (usually rows). Each organism is presumed to have only two alleles (being diploid and having two chromosomes) for each gene, but there may be several kinds of alleles possible for each locus. Then, for any particular allele, an organism may be recorded as having 0, 1, or 2. In addition, missing values for some genes for certain individuals are allowed for in these kinds of datasets. The user is asked to input the value which has been used to indicate missing information. For example, an input data file might look something like this:

```
2   0  99  99   1   0   1   2   0
1   1   1   1   0   1   1   0   2
0   2   0   2   0   0   2   1   1
99  99  1   1  99  99  99   0   2
```

The structure of this dataset is:

	locus 1		locus 2		locus 3			locus 4	
	A	**a**	**B**	**b**	**C**	**c**	**χ**	**D**	**d**
Individual 1	2	0	99	99	1	0	1	2	0
Individual 2	1	1	1	1	0	1	1	0	2
Individual 3	0	2	0	2	0	0	2	1	1
Individual 4	99	99	1	1	99	99	99	0	2

The first row contains information for individual 1, the second row for individual 2, and so on. The first two columns show the occurrence of the two alleles for each individual at the first locus. For example, individual 2 is heterozygous, having alleles **Aa** at locus 1. The next 2 columns code occurrences of each of the two alleles for the second locus, and so on. Notice that for this particular dataset, at locus three, it is

possible to have any two of three alleles: **C**, **c** or **χ**, and the sum of these three columns for any individual (provided data are not missing) will always be equal to two. There is no upper limit to the number of possible alleles that may be present at a given locus. Note also that the user has chosen to use the number "99" to indicate missing information. Thus, in the above data there is no available information for locus 2 for individual 1, and individual 4 is missing information at locus 1 and at locus 3.

Given genetic allozyme data in the above format, the distance measure calculated by the program is an average genetic distance between two individuals per locus, including only those loci for which information is available for both individuals, and is calculated as:

$$d_{ij} = \frac{\sum_{k=1}^{p} w_k \left| y_{ik} - y_{jk} \right|}{2 \min(n_i, n_j)}$$

where p = the total number of columns

n_i = the total number of loci in observation row i that are not missing

n_j = the total number of loci in observation row j that are not missing

w_k = 0 if either $y_{ik} = m$ or $y_{jk} = m$, where m = the value used to indicate missing information.

X. References

Anderson, M.J. 2001a. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26: 32-46.

Anderson, M.J. 2001b. Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences* 58: 626-639.

Anderson, M.J. & Legendre, P. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation* 62: 271-303.

Anderson, M.J. & Millar, R.B. 2004. Spatial variation and effects of habitat on temperate reef fish assemblages in northeastern New Zealand. *Journal of Experimental Marine Biology and Ecology* 305: 191-221.

Anderson, M.J. & Robinson, J. 2001. Permutation tests for linear models. *Australian and New Zealand Journal of Statistics* 43: 75-88.

Anderson, M.J. & Robinson, J. 2003. Generalised discriminant analysis based on distances. *Australian & New Zealand Journal of Statistics* 45: 301-318.

Anderson, M.J. & ter Braak, C.J.F. 2003. Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation* 73: 85-113.

Bray, J.R. & Curtis, J.T. 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs* 27: 325-49.

Cao,Y., Williams, W.P. & Bark, A.W. 1997. Similarity measure bias in river benthic Aufwuchs community analysis. *Water Environment Research* 69(1): 95-106.

Clarke, K.R. 1993. Non-parametric multivariate analysis of changes in community structure. *Australian Journal of Ecology* 18: 117-143.

Dwass, M. 1957. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* 28: 181-187.

Faith, D.P., Minchin, P.R & Belbin, L. 1987. Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio* 69: 57-68.

Glasby, T.M. 1999. Interactive effects of shading and proximity to the seafloor on the development of subtidal epibiotic assemblages. *Marine Ecology Progress Series* 190: 113-124.

Gower, J.C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53: 325-338.

Gower, J.C. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 23: 623-637.

Gower, J.C. & Legendre, P. 1986. Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification* 3: 5-48.

Hajdu, L.J. 1981. Graphical comparison of resemblance measures in phytosociology. *Vegetatio* 48: 47-59.

Johnson, R.A. & Wichern, D.W. 1992. *Applied multivariate statistical analysis, 3rd edition*. Prentice-Hall, Englewood Cliffs, New Jersey.

Legendre, P. & Legendre, L. 1998. *Numerical Ecology, 2nd English edition.* Elsevier, Amsterdam.

Legendre, P. & Gallagher, E.D. 2001. Ecologically meaningful transformations for ordination of species data. *Oecologia* 129: 271-280.

Manly, B.F.J. 1997. *Randomization, bootstrap and Monte Carlo methods in biology, 2nd edition*. Chapman and Hall, London.

McArdle, B.H. & Anderson, M.J. 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* 82: 290-297.

Neter, J., Kutner, M.H., Nachstein, C.J. & Wasserman, W. 1996. *Applied Linear Statistical Models, 4th edition*. Irwin, Chicago.

Pesarin, F. 2001. *Multivariate permutation tests with applications in biostatistics*. Wiley, New York.

Press, W.H., Teuklosky, S.A., Vetterling, W.T. & Flannery, B.P. 1992. *Numerical Recipes in FORTRAN, 2nd edition*. Cambridge University Press, Cambridge.

Searle, S.R., Casella, G. & McCulloch, C.E. 1992. *Variance components*. John Wiley & Sons, New York.

Winer, B.J., Brown, D.R. & Michels, K.M. 1991. *Statistical principles in experimental design, 3rd edition*. McGraw-Hill, New York.