COMMUNITY ECOLOGY 14(2): 153-163, 2013
1585-8553/$20.00 © The Author
DOI: 10.1556/ComEc.14.2013.2.5

How accurate and powerful are randomization tests in multivariate analysis of variance?

V. D. Pillar

Departamento de Ecologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil.
Phone: +55 51 33087101, Fax: +55 51 33087626, E-mail: vpillar@ufrgs.br

Keywords: Count data, Distance-based MANOVA, Distribution free, MRPP, Neyman-Pearson lemma, Permutation tests, Type I error, Type II error.

Abstract: Multivariate analysis of variance, based on randomization (permutation) tests, has become an important tool for ecological data analyses. However, a comprehensive evaluation of the accuracy and power of available methods is still lacking. This is a thorough examination of randomization tests for multivariate group mean differences. With simulated data, the accuracy and power of randomization tests were evaluated using different test statistics in one-factor multivariate analysis of variance (MANOVA). The evaluations span a wide spectrum of data types, including specified and unspecified (field data) distributional properties, correlation structures, homogeneous to very heterogeneous variances, and balanced and unbalanced group sizes. The choice of test statistic strongly affected the results. Sums of squares between groups (Q_b) computed on Euclidean distances (Q_b-*EUD*) gave better accuracy. Q_b on Bray-Curtis, Manhattan or Chord distances, the multiresponse permutation procedure (*MRPP*) and the sum of univariate ANOVA *F* produced severely inflated type I errors under increasing variance heterogeneity among groups, a common scenario in ecological data. Despite pervasive claims in the ecological literature, the evidence thus suggests caution when using test statistics other than Q_b-*EUD*.

Abbreviations: ANOSIM – Analysis of Similarity, ANOVA – Analysis of variance, CHOR – Chord distance, EUD – Euclidean distance, LR-IND – Likelihood-ratio test, assuming independence of variables, MAN – Mahattan distance, MANOVA – Multivariate analysis of variance, MRPP – Multiresponse permutation procedure, PERMANOVA – Permutational multivariate analysis of variance, Qb – Sums of squares between groups, Qw – Within-groups sum of squares, SUM-F – Univariate ANOVA F statistic summed over all variables

Introduction

Analysis of data from complex systems as ecosystems and ecological communities requires multivariate methods, for it may be impossible to understand complex behaviour by examining the system in pieces. Multivariate methods for exploratory analysis are well known in different contexts. However, multivariate methods for testing differences among group mean vectors and factor interactions (i.e., multivariate analysis of variance - MANOVA) are used much less frequently. MANOVA based on conventional statistics and normal theory (Hotelling 1931, Wilks 1932) is often hindered by the assumption that the variables have multivariate (joint) normal distribution of errors, which is unrealistic in many contexts (Bradley 1968, Edgington 1987, Orlóci 1993).

Randomization or permutation tests are much less limited by assumptions. They rely on distribution free statistics and on algorithms that use systematic or random data permutations to generate alternative outcomes for the chosen test statistic under a true null hypothesis and, based on these, find the probability that will or will not support rejection of the null hypothesis. As for the distinction between permutation and randomization tests, there is no consistency in the literature in the definition of the terms. It seems some authors simply prefer one term than another and thus I will use them interchangeably. The roots of randomization testing are found in the works of Fisher (1951), Kempthorne (1952, 1955), but the available computational environment in those days limited the use of the techniques and thus, under simplifying assumptions, standard statistical tables became a shortcut to finding the probabilities that the statistical tests required (Pillar and Orlóci 1996). It was only with the advent of fast and affordable personal computers that the methods became practical. Permutation tests have been described earlier (Bradley 1968, Edgington 1969a,b, 1987) but with emphasis on the analysis of experimental univariate data, and are reviewed in biological contexts (Crowley 1992, Potvin and Roff 1993, Manly 2007). Different tests have been devised for multivariate comparisons of groups of units in mensurative (observational) or manipulative experimental designs (Mantel and Valand 1970, Mielke et al. 1976, Clarke 1993, Pillar and Orlóci 1996, Legendre and Anderson 1999, Mielke and Berry 1999, Anderson 2001, McArdle and Anderson 2001, Mielke and Berry 2001, Manly 2007). Despite its increasing importance in ecological data analyses, with a few exceptions there is little information available in standard statistical textbooks and introductory statistical courses. Perhaps a healthy bit of suspicion is preventing acceptance, and this is an important reason why statistical accuracy and power of these methods need to be demonstrated.

A hypothesis test should be accurate, i.e., it should present a type I error equal or acceptably close to the test's a priori significance level when the null hypothesis (H_0) is true. In addition, it should have sufficient power to detect differences of groups when they do exist (when H_0 is false). How permutations of the observation vectors should be done in order to achieve an accurate test is straightforward for one-factor analysis of variance: The vectors are simply permuted among the units and their corresponding group labels, for under a true null hypothesis the association of a given observation vector to a given group is arbitrary, and thus any combination of observation vectors can theoretically be associated with any of the group labels (Edgington 1987, Pillar and Orlóci 1996, Manly 2007). For two-factor analysis of variance different permutation options involve unrestricted permutation of raw data, or permutation of some form of residuals, or restricted permutation (Anderson and ter Braak 2003, Torres et al. 2010). The accuracy and power of different solutions can be evaluated empirically by data simulation.

The choice of the test statistic may be critical for the accuracy and power of a randomization test. Also, some test statistics may be equivalent, such as the Fisher *F*-ratio and the sum of squares between groups or within groups, which in a randomization test for one-factor analysis of variance will give identical probabilities with the same data and permutations (Edgington 1987). For multivariate data, T^2 (Hotelling 1931) and Wilks' lambda statistics (Wilks 1932) have been used; these are influenced by correlations between the variables (Anderson 2001). Several other test statistics have been proposed for multivariate data, such as the sum of squares within groups in Ward's *E* statistics (Romesburg 1985), sum of squares between groups (Q_b, Pillar and Orlóci 1996), the pseudo *F*-ratio (Anderson 2001), the average rank of between-group dissimilarities in ANOSIM (Clarke 1993), or the weighted average Euclidean distance within groups in MRPP (Mielke and Berry 2001). Warton and Hudson (2004) suggested that for multivariate abundance data there was no advantage in using distance-based statistics (such as the Q_b) over scale-invariant statistics such as the univariate ANOVA *F* statistic summed over all variables (*SUM-F*) or the Wilks' lambda statistic derived assuming completely independent variables (*LR-IND*). The MRPP statistic has also been pointed out as being "more powerful" than sums of squares (Zimmerman et al. 1985, Mielke and Berry 2001). However, a thorough evaluation of the accuracy and power of these different choices for ecological data is still lacking.

The central purpose of this paper is to assess the accuracy and power of randomization tests in one-factor multivariate analysis of variance, using different test statistics and an extremely large amount of simulated data with a broad range of specified conditions. I address questions about the relative test statistic performance. In this respect, I compare scale-invariant statistics (*SUM-F* and LR-IND, Warton and Hudson 2004) to the weighted average of within-group distances (MRPP, Mielke and Berry 2001) and the sum of squares between groups (Q_b, Pillar and Orlóci 1996) derived from different distances measures. For that, I use continuous or count data, differing in the number of variables, sample size, group variance and size heterogeneity, correlation structure, mean abundances and levels of aggregation.

Material and methods

Test statistics in randomization testing

The methods are based on two sets of information. The first is the matrix of *n* units described by *q* variables and the second is one or more (here limited to a maximum of one) factors with discrete states defining *k* groups of units. In distance-based methods an *n*-by-*n* dissimilarity matrix obtained from the first matrix is needed for computing the test statistic according to the *k* groups. I restrict the assessment to evaluations using Euclidean, Chord, Manhattan and Bray-Curtis distances (for definition of these measures see, e.g., Legendre and Legendre 1998, Podani 2000). Resemblance matrices of Manhattan and Bray-Curtis distances may not hold Euclidean properties (Gower and Legendre 1986), however McArdle and Anderson (2001) demonstrated that sums of squares may still be validly computed for any type of dissimilarity irrespective of their Euclidean properties.

For distance-based MANOVA we evaluated accuracy and power using as randomization test criterion the sum of squares between groups (Q_b) computed according to (Pillar and Orlóci (1996):

$$Q_b = Q_t - Q_w$$

where

$$Q_t = \frac{1}{n}\sum_{h=1}^{n-1}\sum_{i=h+1}^{n} d_{hi}^2$$

is the total sum of squares of $n(n\text{-}1)/2$ pair-wise squared dissimilarities between *n* sampling units and

$$Q_w = \sum_{c=1}^{k} Q_{wc}$$

is the sum of squares within *k* groups, such that

$$Q_{wc} = \frac{1}{n_c}\sum_{h=1}^{n_c-1}\sum_{i=h+1}^{n_c} d_{hi|c}^2$$

where $d^2_{hi|c}$ are pair-wise squared dissimilarities of n_c units belonging to group *c*. In one-factor analysis, the *k* groups are given by the states of the factor being considered.

The partitioning of the total sum of squares based on the distance matrix is not novel and has been used before in the published literature (e.g., Edwards and Cavalli-Sforza 1965, Orlóci 1967). Computing sum of squares from a Euclidean distance matrix is equivalent to (but simpler than) doing it from group centroids based on coordinates and deviations from centroids as used by others (Romesburg 1985, Edgington 1987, Manly 2007). Further, computing sums of squares on distances is more flexible in that it allows choices

among different dissimilarity measures (Anderson 2001). In addition, it is not constrained by the kind of variables describing the units, for appropriate resemblance measures also exist for binary, qualitative and mixed data types.

The distinction made by Warton and Hudson (2004) between distance-based and variable-based statistics is in most cases only computational. For instance, within-groups sum of squares computed from Euclidean distances (Q_w) is equivalent (i.e., will give identical values) to Ward's E statistic (Romesburg 1985). As a matter of fact, with univariate data, Q_b and Q_w computed from Euclidean distances are identical to the between-groups and residual sum of squares calculated the usual way in classical ANOVA.

In one-factor analysis, since the total sum of squares is invariant over permutations, Q_b and the sum of squares within groups (Q_w or Ward's E) will give identical probabilities for the same permutations and therefore are equivalent statistics in randomization testing. Further, it can be demonstrated that the pseudo F-ratio described by Anderson (2001) as PERMANOVA's $F = Q_b/Q_w$ is equivalent to Q_b as a permutation test statistic in one-factor multivariate analysis of variance.

Also, Q_b is equivalent to the average distance within groups used in multiresponse permutation procedures (MRPP, Mielke et al. 1976, Mielke and Berry 2001) when the distance computed within MRPP is based on squared rather than non-squared distances. Evaluations here were performed with the MRPP test statistic (Mielke and Berry 2001) defined as the weighted average dissimilarity within k groups

$$\delta = \sum_{c=1}^{k} \frac{1}{n_c} \xi_c$$

where

$$\xi_c = \frac{2}{n_c(n_c-1)} \sum_{h=1}^{n_c-1} \sum_{i=h+1}^{n_c} d_{hi|c}$$

and $d_{hi|c}$ are pair-wise non-squared dissimilarities of n_c units belonging to group c. Only the Euclidean distance was used in this case. None of the data commensuration methods described in Mielke and Berry (2001) was involved in the computation of the MRPP test statistic.

Mielke and Berry (2001) have claimed that MRPP computed based on non-squared distances is more robust than on squared distances (and therefore Q_b) for data distributions with extreme values. One advantage of Q_b is that it can be partitioned among factors and interaction(s) in multifactorial designs or among orthogonal contrasts between groups (Pillar and Orlóci 1996). Though MRPP used in regression analysis of linear models can handle factorial designs (Mielke and Berry 2001), the procedure is less straightforward than it is with Q_b. Furthermore, Q_b is equivalent to the sum of squares computed on all canonical eigenvalues in distance-based redundancy analysis (RDA, Legendre and Anderson 1999) and therefore it is unnecessary, not to mention more computationally demanding, to use RDA (see also McArdle and Anderson 2001).

Evaluations also used scale-invariant statistics: the classical univariate ANOVA F statistic summed over all variables (*SUM-F*) and the Wilks' lambda likelihood-ratio statistic derived by Warton and Hudson (2004) assuming completely independent variables, that is,

$$LR - IND = \sum_{j=1;Q_{wj}>0}^{q} n \log\left(\frac{Q_{tj}}{Q_{wj}}\right)$$

where n is the number of units, q the number of variables, and Q_{tj} and Q_{wj} are respectively total sum of squares and within groups sum of squares for variable j. If $Q_{wj} = 0$, variable j was ignored in the computation of *LR-IND* and *SUM-F*.

Random permutations

If the null hypothesis (H_0) is true, the observation vector in a given unit is independent of the group to which the unit belongs. The observed data set is seen as one of the possible permutations of observation vectors among the units and their groups in the data matrix. The observation vectors contain q variables and are permuted intact, preserving the correlation structure in the data. Over permutations, when using Q_b as test criterion, Q_t remains constant and is only redistributed between Q_b and Q_w. For the same reason, there is no need to recalculate the dissimilarities for each permutation; the dissimilarity matrix is only rearranged according to the permutation, analogously to the permutations performed in a Mantel test.

Thus, the basis of randomization testing is to randomly permute the data according to H_0 and for each of these permutations compute the test criterion (e.g. Q_b^0), comparing it to the value of Q_b found in the observed data. The probability $P(Q_b^0 \geq Q_b)$ will be given by the proportion of permutations in which $Q^0_b \geq Q_b$. For the MRPP, since d measures within group dispersion, the probability is instead represented by $P(\delta^0 \leq \delta)$.

The collection of all possible permutations, the reference set, can be generated systematically, but depending on the number of units the computation demand may be too high to consider all possible permutations. A random but still large sample of this reference set is sufficient for generating the probabilities (Hope 1968, Edgington 1987). The larger the number of random permutations, the closer the P-values will be to the ones that would be obtained in complete systematic data permutation. The observed data set is included as one of these permutations (see Hope 1968), thus determining a minimum probability of $1/B$, where B is the number of permutations.

H_0 will be rejected if $P \leq \alpha$, where α is the significance level chosen for the test. If H_0 is rejected, we conclude that the groups differ, with an error probability of P. In practice however, P may be used as a measure of the plausibility of

H_0, instead of comparing it to an absolute α threshold (Lehmann 1993).

Evaluating the tests' accuracy and power

The randomization tests described above were evaluated using simulated data matrices varying with regard to specified group mean differences and several other data choices. The estimation of type I error or power for a given combination of simulated data choices was based on an extremely large number of randomization tests (1,000,000 in some cases; 20,000 in others), each with a new generated data set. The proportion of H_0 rejection estimated the type I error when using data sets with no specified mean differences between groups, and the power (1-β, where β is the probability of committing a type II error) when any difference was specified. In all cases the threshold used for H_0 rejection was α = 0.05. The randomization tests used 20 random permutations (B = 20), which was adequate for the purpose of evaluating accuracy and power at this α = 0.05 threshold. With B = 20 the minimum probability will be P = 1/20 = 0.05 when the test statistic value for the observed data is the most extreme among the B values. The true probability in this extreme case is $P \leq 0.05$, but it is enough for the rejection of H_0. The proportion of H_0 rejection (in this case over one million or 20,000 sets of data) is the only information needed in type I error and power evaluation, and it may be any value between 0 and 1. For this purpose $B \geq 1/\alpha$ (e.g., for α = 0.01, B should be at least 100). It is important to note that such a small B would not be adequate in ordinary use of randomization testing, when interpretations are based on the results of one test only and not on the basis of many thousands or more, as is the case here.

Data simulation with specified distributional properties

In one set of evaluations, continuous data with n units and q variables were simulated holding normally distributed variables with specified correlation structure, variance and mean in each group of units. It has been shown that the correlation structure, at least for bivariate data, affects the power of permutation tests with MRPP (Mielke and Berry 1999, 2001). The procedure for generating data with known correlation structure, adapted from Peres-Neto and Jackson (2001), is based on the Cholesky decomposition of the specified q-by-q correlation matrix for each group of n_c units; the resulting upper triangular matrix is then premultiplied by an n_c-by-q matrix with random values drawn from a normal distribution with zero mean and unit variance, yielding an n_c-by-q data matrix sought for the group. By concatenating these matrices, an n-by-q matrix is created, which is in fact a simulated random sample drawn from normally distributed, centered and standardized variables arranged in data sets with known correlation structure. The values in each group were then multiplied by the specified standard deviation (equal or heterogeneous between groups) and added to the specified group mean. Simulated data sets had the following correlation structures: (1) uncorrelated variables with expected pairwise correlation coefficient r = 0 for all variables; (2) variables positively correlated with pairwise r = 0.8 for all variables; (3) all variables highly correlated (r = 0.8) in one group of units and uncorrelated in another (r = 0); (4) variables forming two groups, with r = 0.8 for within-group correlations and r = –0.8 for between-group correlations. These correlation structures reflect different intrinsic dimensionalities: In (1) the data will have a larger number of intrinsic dimensions than in (2) and (4), for the same number of variables. In (2) unit group differences will be along the main axis while in (4) unit group differences will be orthogonal to the main axis of variation in the data space.

In another set of evaluations, count-data matrices were generated with values drawn at random from a negative binomial distribution, allowing different levels of data aggregation (variance larger than the mean). The negative binomial distribution has been widely used to model species abundances (White and Bennets 1996, McArdle and Anderson 2004, Warton et al. 2012). The computational function implemented for drawing random values from the negative binomial distribution (Galassi et al. 2003) requires two parameters: (1) a probability k, where $k < 1$ is the mean/variance ratio ($k = \mu/\sigma^2$, as k decreases the distribution becomes more aggregated); (2) a parameter $r = \mu k/(1-k)$. See Appendix 1 for examples of simulated data using these parameters. In power evaluations the parameter k was adjusted to keep constant the defined σ^2 for each group as specified mean differences between groups increased; in this way, possible confounding effects of mean and variance differences were overcome.

Using the above mentioned methods, type I error in one-factor designs, comparing two groups of units, was assessed with simulated data defined by combinations of choices in number of units (7-20), group sizes (balanced to very unbalanced), number of variables (1-30), group variances (homogeneous to very heterogeneous), error distribution (normal, negative binomial), correlation structure between variables (only for continuous data, see above), and mean abundances and levels of aggregation (only for count data). Mean abundances (μ) ranged from 1 to 8 and inverse levels of aggregation (k) from 0.1 to 0.9. With each data set the following test statistics were computed: Q_b on Euclidean (Q_b-*EUD*), Chord (Q_b-*CHOR*), Manhattan (Q_b-*MAN*) and Bray-Curtis distances (Q_b-*BRAY*), MRPP on Euclidean distances (*MRPP-EUD*), *SUM-F* and *LR-IND*. The test was deemed accurate if type I error was within 99.9% confidence limits (0.05 ± 0.00072 with 1,000,000 tests or 0.05 ± 0.00507 with 20,000 tests). Power was evaluated with data simulated with a selected range of values for some of the abovementioned data choices.

A relative centroid distance (c_{jk}) was used in power comparisons among simulation choices generating data with different scales. It was defined as $c_{jk} = b_{jk}/a_j$, where b_{jk} is the centroid distance between group mean vectors specified a priori for groups j and k, and a_j is the range of values in the simulated data for group j over all variables averaged over the tests. In all cases, group j was the group for which the

Table 1. Properties of grassland community data (Pillar et al. 1992) used in a posteriori power analysis. All species or a subset of the least frequent ones was used. Relief position was categorized in four groups (see b). Balanced group sizes were obtained by pooling units on the same gradient, within the same relief position. Variance homogeneity among groups was tested by randomization (Anderson 2006), a method analogous to Levene's test.

Data sets	Unbalanced 60 original stand units		Balanced 16 pooled stand units
(a) Species subset	All 60 species	Least frequent 30 species	All 60 species
Species presence (% stands)	7 to 80%	7 to 28%	13 to 88%
Median abundances across groups	0 to 5	0 to 0	0 to 4.25
Mean abundances across groups	0.1 to 4.4	0.1 to 1.1	0.1 to 3.8
Variances across groups	0.1 to 9.1	0.1 to 4.8	0.02 to 6.5
Mean/variance ratios	0.19 to 1.13 (1)	0.19 to 0.84	0.24 to 2.79 (2)
Correlation mean x variance	r = 0.86 (n=60)	r = 0.93 (n=30)	r = 0.85 (n=60)
(b) Groups by relief position (3)			
Group sizes	23, 17, 9, 11	23, 17, 9, 11	4, 4, 4, 4
Variance homogeneity	F=3.7 (P=0.01)	F=1.5 (P=0.23)	F=3.8 (P=0.03)

(1) Only two species with mean/variance ratio larger than 1.
(2) 33 species with mean/variance ratio larger than 1.
(3) Top, convex, concave, lowland.

specified group means for the variables were kept the same as for type I error evaluation.

Data simulation with distributional properties of field data

Simulated data sets for one-factor designs were also generated with distributional properties of grassland community data (Pillar et al. 1992). The original data comprised 60 0.25-m^2 stand units, described by abundances of 60 species (variables), located on four relief gradients and categorized according to relief position (top, convex, concave, lowland); group sizes were unbalanced and group variances were in some cases heterogeneous (see Table 1, for details). The following a posteriori power analysis, relevant only for abundance data, was then applied:

(1) Calculate for each variable i the mean $\bar{x}_{ij}$ and variance s_{ij}^2 within each group j and the mean $\bar{x}_i$ and variance s_i^2 across the groups, which in this case were defined by one factor (relief position, Table 1).

(2) Generate for each variable i and unit h a new observation $y_{hi|j} = b$, where b is a random value drawn from the negative binomial distribution, with parameter k set as $k = \bar{x}_{ij}/s_{ij}^2$ when within group variance $s_{ij}^2 > 0$, as $k = \bar{x}_i/s_i^2$ when $s_{ij}^2 = 0$, or $k = 0.99$ when these computed k values were not smaller than 1. Parameter r for drawing a random value from the negative binomial distribution was defined as $r = [\bar{x}_i + w(\bar{x}_{ij} - \bar{x}_i)]k(1-k)$, where w is the specified weight for testing type I error or power indicating the proportion of the existing effects in the original data; if $w = 0$ a data set with the conditions specified by H_0 is maintained; if $w > 0$ power is evaluated, and if $w = 1$ the expected group means in the simulated data will match the original data.

(3) Perform randomization tests with the data using the chosen statistics.

(4) Repeat steps (2) to (3) many times (in this case 20,000), recording the proportion of H_0 rejection.

In other power analyses with the same grassland data set, a balanced number of units in each group was obtained by pooling units on the same gradient, within same relief position, averaging accordingly each species abundance values and thereby generating a data matrix of 16 grassland units, in which mean/variance ratios tended to be higher (see Table 1). The same procedure described above was then applied.

Randomization tests evaluated in this paper were implemented in the software MULTIV written by the author (available at http://ecoqua.ecologia.ufrgs.br). A computer program in C++ was specifically written for the purpose of performing the simulations and additional tests. The functions *gsl_ran_gaussian* and *gsl_ran_negative_binomial* from the GNU Scientific Library (Galassi et al. 2003) were used for drawing random values from the normal and negative binomial distributions.

Results

In most cases with homogeneous variances (Table 2) the tests were accurate, with type I error within 99.9% confidence limits for $P = 0.05$. However, small inaccuracies, especially with *MRPP-EUD*, were discernible with continuous data when all variables were highly correlated in one group of units ($r = 0.8$) and uncorrelated in another ($r = 0$). With heterogeneous variances (standard deviation ratio 1:3) type I error inaccuracies were found with all test statistics, but severely inflated type I error occurred when using *MRPP-EUD* on continuous data, and with count data using scale invariant *SUM-F*, *LR-IND,* and all distance-based statistics except Q_b-*EUD* (see Table 2 for some of these and Appendices 2 and 3 for full results). Although it was generally more accurate than the other statistics, Q_b-*EUD* did show slight to moderate inaccuracies with single- or few-variable count data (note that these inaccuracies were much greater using the other test statistics). With unbalanced group sizes, relatively minor inaccuracies occurred with univariate count data, and with continuous data where the groups had sharply different correlation structures (Table 2). The results with unbalanced

Table 2. Type I error of randomization testing for unequal locations in one-factor designs, using different test statistics and data simulation choices. For data simulation random values were drawn from the normal (Table **a**) or negative binomial distributions (Table **b**), with units distributed in two groups under true H_0. Each estimation of type I error (alpha = 0.05) was found after 1,000,000 simulated data sets. Tests were inaccurate (in bold) when type I error was outside the 99.9% confidence limits (0.05 ± 0.00072). See main text for details. Full results for other choices are in Appendices 2 and 3.

(a) Normal distribution (continuous data)

Data and test options		Type I error									
		Balanced group sizes (10:10)							Unbalanced group sizes (14:6)		
		Homogeneous variances			Heterogeneous variances ¥				Homogeneous variances		
Correlation structure‡		r = 0	r = 0.8	r = 0.8-0	r = 0	r = 0.8	r = 0.8-0	r = ±0.8	r = 0	r = 0.8	r = 0.8-0
Statistics (20 units, 10 variables)	*Qb-EUD*	0.050	0.050	**0.058**	**0.057**	**0.059**	**0.057**	**0.058**	0.050	0.050	**0.034**
	Qb-MAN	0.050	0.050	**0.029**	**0.064**	**0.059**	**0.036**	**0.059**	0.050	0.050	**0.012**
	Qb-BRAY	**0.051**	0.050	**0.049**	**0.073**	**0.079**	**0.059**	**0.059**	0.050	0.050	**0.018**
	Qb-CHOR	0.050	0.050	**0.057**	0.050	0.050	**0.056**	0.050	0.050	0.050	**0.033**
	SUM-F	0.050	0.050	**0.053**	**0.086**	**0.062**	**0.083**	**0.062**	0.050	0.050	**0.032**
	LR-IND	**0.051**	0.050	**0.053**	**0.082**	**0.062**	**0.079**	**0.062**	0.050	0.050	**0.030**
	MRPP-EUD	0.050	0.050	**0.120**	**0.984**	**0.421**	**0.965**	**0.421**	0.050	0.050	**0.064**
No. of variables (20 units, Qb-EUD)	1	0.050	–	–	**0.059**	–	–	–	0.050	–	–
	3	0.050	0.050	**0.052**	**0.058**	**0.058**	**0.058**	**0.058**	0.050	0.050	**0.045**

(b) Negative binomial distribution (count data)

Data and test options		Balanced group sizes (10:10)						Unbalanced group sizes (14:6)			
		Homogeneous variances (k_1 = 0.2, k_2 = 0.2)		(k_1 = 0.5, k_2 = 0.5)		Heterogeneous variances ¥ (k_1 = 0.9 k_2 = 0.1)		Homogeneous variances k_1 = 0.2, k_2 = 0.2		k_1 = 0.5, k_2 = 0.5	
Mean		1	8	1	8	1	8	1	8	1	8
Statistics (20 units, 10 variables)	*Qb-EUD*	0.050	0.050	**0.049**	**0.049**	**0.051**	0.050	0.050	0.050	**0.049**	0.050
	Qb-MAN	0.050	0.050	0.050	0.050	**0.595**	**0.262**	0.050	0.050	0.050	0.050
	Qb-BRAY	0.050	0.050	0.050	0.050	**0.828**	**0.205**	0.050	0.050	0.050	0.050
	Qb-CHOR	0.050	0.050	0.050	0.050	**0.349**	**0.355**	0.050	0.050	0.050	0.050
	SUM-F	0.050	0.050	0.050	0.050	**0.621**	**0.255**	0.050	0.050	0.050	0.050
	LR-IND	0.050	0.050	0.050	0.050	**0.601**	**0.237**	0.050	0.050	0.050	0.050
	MRPP-EUD	0.050	0.050	0.050	0.050	**0.474**	**0.950**	0.050	0.050	0.050	0.050
Number of variables (20 units, *Qb-EUD*)	1	**0.027**	**0.045**	**0.030**	**0.043**	**0.138**	**0.082**	**0.035**	**0.047**	**0.036**	**0.046**
	3	**0.048**	0.050	**0.048**	0.050	**0.105**	**0.062**	**0.049**	0.050	0.050	0.050
	30	0.050	0.050	0.050	0.050	**0.035**	0.050	0.050	0.050	0.050	0.050

‡ Uncorrelated variables (r=0); all variables correlated in one (r=0.8) or in two negatively correlated groups of variables (r=±0.8); all variables highly correlated in one group of units and uncorrelated in another (r=0.8-0).

¥ Group standard deviations 1:3.

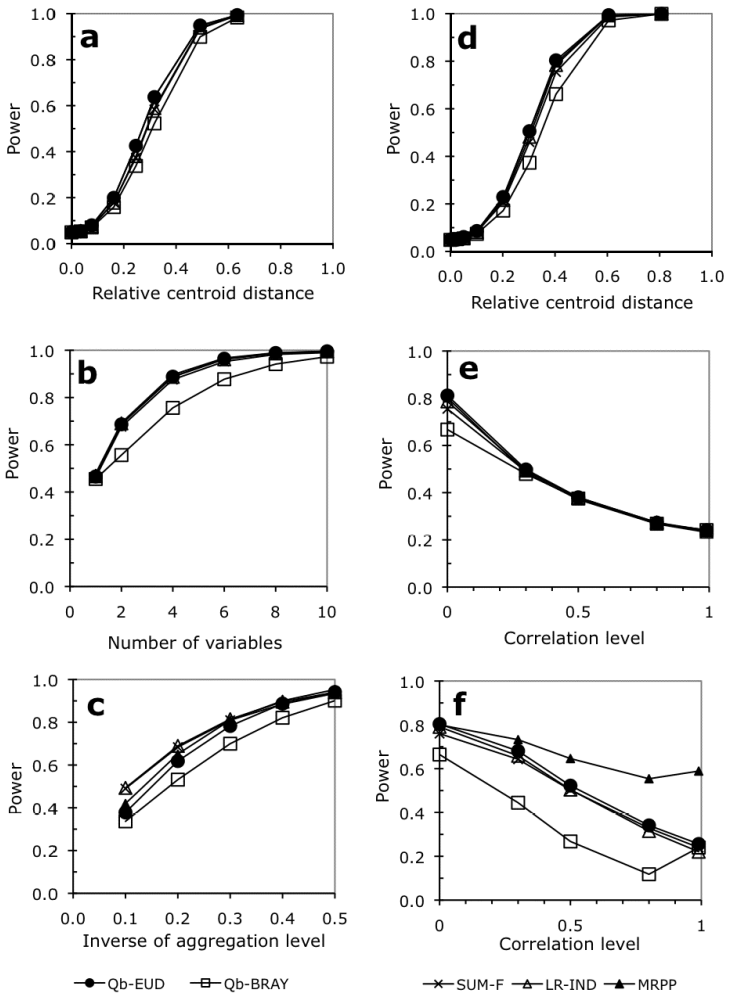

Figure 1. Power of randomization tests for group mean differences in one-factor MANOVA with count (**a-c**) or continuous data (**d-f**) and the effect of the level of aggregation, correlation and number of variables. Simulated data had 20 units and, except for (**b**) 10 variables, with increasing (a, d) or constant (**b-c, e-f**) differences between two equally sized groups. In (**a-c**) random counts were drawn from the negative binomial distribution with means ranging from rare (μ=0.1) to abundant (μ=8) in the same data set and inverse aggregation level k = 0.5 (**a-b**) or varying (**c**) for equal group means and adjusted to keep equal group variances according to the specified group means. In (**d-f**) random values were drawn from normal distribution with s = 1 in both groups. In (**a-d**) variables were uncorrelated, and in (**e-f**) they were set to increasing correlation levels in both groups, which were all positive in (**e**) and in (**f**) variables formed two groups, with increasing positive within-group and equally increasing negative between-group correlations.

group sizes combined with heterogeneous variances will be presented later.

The simulations showed that – irrespective of test statistic - with balanced group sizes and uncorrelated variables with homogeneous variances, power with contagious count data was similar to that with continuous data drawn from a normal distribution (Fig. 1a,d), and was affected by the levels of aggregation (Fig. 1c). Also, there was a clear effect of the number of uncorrelated variables on power when group differences in each variable were kept constant, which with more variables resulted in increased group centroid distance (Fig. 1b). Furthermore, the correlation structure of variables affected power: the more correlated the variables the lower the power, but when all correlations were positive the test statistics had similar results (Fig. 1e), while when the variables were arranged in two negatively correlated groups this tendency was more pronounced with the Q_b-*BRAY* and less with the *MRPP-EUD* test statistics (Fig. 1f). Additional results show that for all test statistics type I error was inaccurate (lower than α) with small sample sizes up to 12 units, both with count and continuous simulated data (Appendix 4 for Q_b-*EUD*; similar results not included were found using the other test statistics). As expected, the power of the test increased with larger sample sizes and this became more evident with smaller group differences (Appendix 5).

Further examining the effect of variance heterogeneity, inflated type I error increased sharply at standard deviation ratios beyond 1:2, reaching values higher than 0.6 (α = 0.05) with count data when *SUM-F*, *LR*-IND, Q_b-*BRAY* and *MRPP-EUD* were used (Fig. 2a). This effect was not evident with Q_b-*EUD* aside from slight inaccuracies at extremely high standard deviation ratios, and was intermediate for Q_b-*CHOR* and Q_b-*MAN*. Logarithm transformation of the data (Fig. 2b) did not correct the problem, and indeed inflated the type I errors even further, even with Q_b-*EUD*. With continuous data drawn from the normal distribution the effect of variance heterogeneity on type I error inflation was absent or less prominent, except for Q_b-*BRAY* and particularly for *MRPP-EUD*, the latter which produced extremely high (nearly 1) type I errors at standard deviation ratios beyond 1:2 (Fig. 2d,e). Similar results (Appendix 6) were found for continuous data with variables highly correlated (r=0.8) in one unit group and uncorrelated in another group. Variance heterogeneity reduced power with count and continuous data (Fig. 2c,f) and the effect was less pronounced with Q_b-*BRAY* and absent for *MRPP-EUD*.

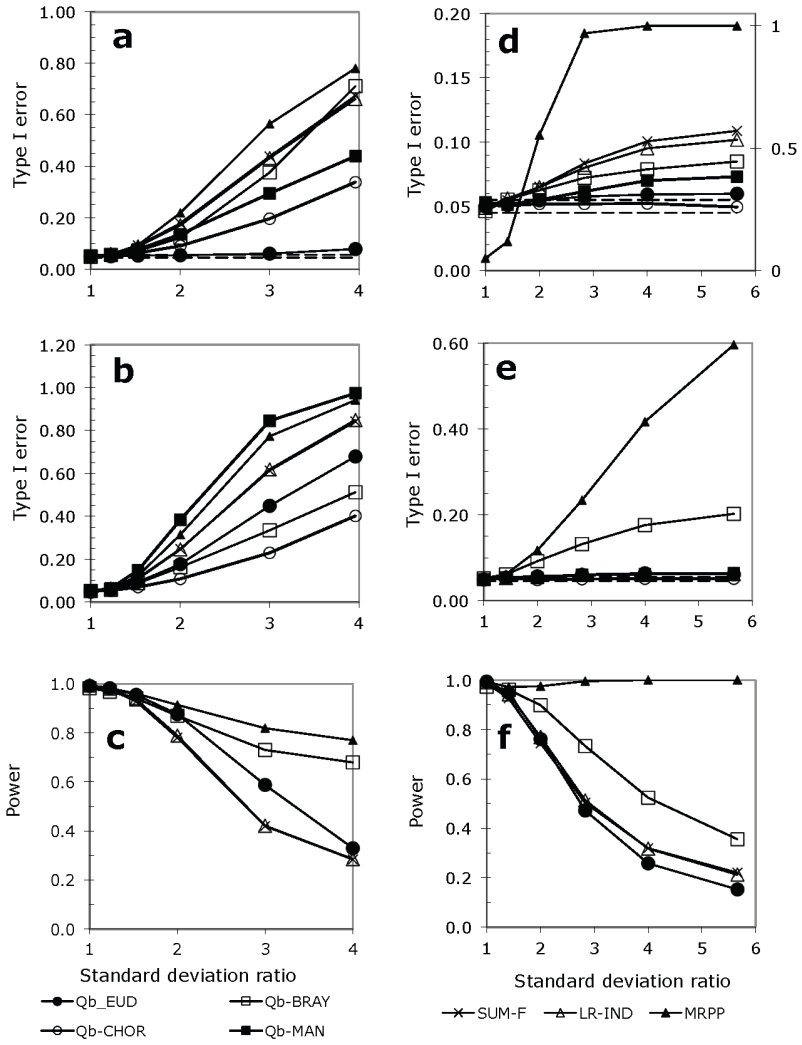

Figure 2. Effect of variance heterogeneity on type I error and power in one-factor MANOVA randomization tests for group mean differences with count (**a-c**) or continuous normal data (**d-f**). The secondary axis in (**d**) is for MRPP only. Variables were uncorrelated in (**a-d, f**) and highly correlated (r=0.99) in both groups in (**e**). In (**b**) data were log transformed (y = log(x +1). Simulated data had 20 units in 2 balanced groups, and 10 variables. Relative centroid distance for power was ca. 0.6 in (**c, f**). Counts drawn from negative binomial distribution varied 80-fold from rare to abundant variables in one data set and the parameter k in the groups was set from 0.5:0.5 for equal, to 0.94:0.06 for very unequal variances. Dashed lines indicate 99.9% confidence limits around 0.05.

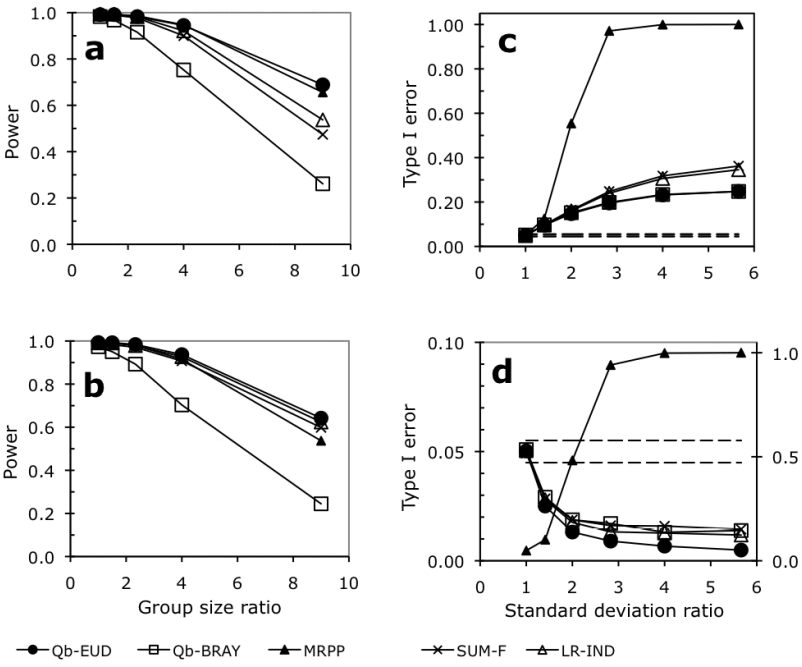

Figure 3. Power and type I error of randomization tests in one-factor MANOVA for group mean differences affected by increasing unbalanced group size (**a, b**) and the combined effect of unbalanced group size and increasing variance heterogeneity (**c, d**). The secondary axis in (**d**) is for MRPP only. Simulated data sets had 20 units, in 2 groups, and 10 random variables drawn from negative binomial (**a**) or normal distribution (**b-d**), with constant group differences for power and equal means for type I error. In (**c**) group sizes were 12 and 8 and in (**d**) were 8 and 12, the first group with fixed and the second with increasing variance. Dashed lines indicate 99.9% confidence limits around 0.05.

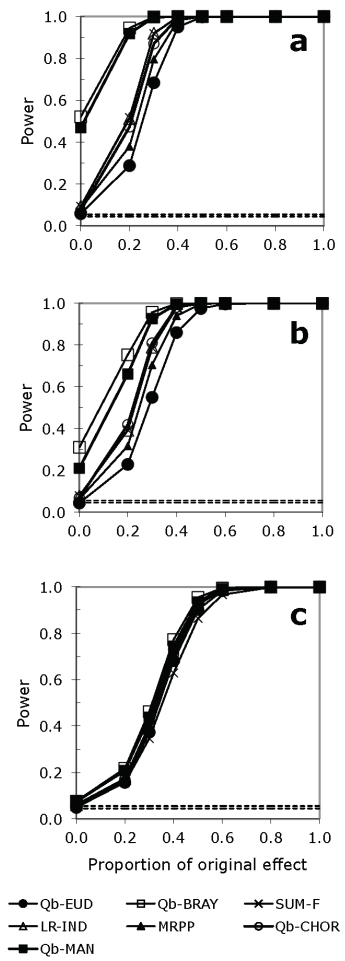

Figure 4. Power analysis for the detection of group mean differences with simulated data generated with distributional properties of the grassland community data with unbalanced (**a-b**) or balanced group sizes (**c**). Groups were defined by relief position. The 60 stand units were described by all 60 species (**a**) or the 30 least frequent ones (**b**), and in (**c**) a balanced design was obtained by pooling units, generating a data matrix of 16 stand units by 60 species. Proportion of original effect ranged from 0 (no effect, thus indicating type I error) to 1 (effect observed in the original data). See Table 1, for details on data sets.

The power of the test declined with increased inequality of group sizes (Fig. 3a,b); the effect was more distinct with Q_b-*BRAY* than for the other statistics for both count and continuous data. The combined effect of unbalanced group size and variance heterogeneity (Fig. 3c-d) resulted in severely inflated type I error when the smaller group was more dispersed, and the opposite when the larger group was more dispersed, irrespective of test statistic (except *MRPP-EUD*), with more pronounced type I error inflation with *SUM-F* and *LR-IND*, and extremely high inflation with *MRPP-EUD*.

The analysis based on the grassland data showed that the tests were most accurate using Q_b-*EUD*. Type I error was slightly inflated with Q_b-*CHOR, SUM-F*, *LR-IND* and *MRPP-EUD*, but severely inflated with Q_b-*BRAY* and Q_b-*MAN* (Fig. 4, full results in Appendix 7). All test statistics reached 100% power at similar proportions of the original effect (w) and the ones giving inflated type I errors showed higher power at lower w. Accuracy was improved by pooling units in order to get a balanced design out of an unbalanced one.

Discussion

The simulations indicated a strong influence of test statistic on the accuracy and power of randomization tests for group mean differences, particularly with aggregated count data, as opposed to normally distributed data. In general, the use of Q_b-*EUD* as test statistic results in higher accuracy than the other evaluated statistics, whether distance-based or scale-invariant (i.e., *SUM-F*, *LR-IND*). Excepting Q_b-*EUD*, the most significant problem found in the simulations is unacceptably severe inflation of type I error under increasing variance heterogeneity among groups with aggregated count data. That is, where test statistics other than Q_b-*EUD* are used, groups of sampling units with high between-group heterogeneity in variance - but with mean vectors differing only by chance - are likely to be declared as being different more often than the nominal significance of the test. Though small inaccuracies may not have major consequences, under heterogeneous variances for a nominal significance of $\alpha = 0.05$, type I error with these statistics in some cases were larger than 0.2, and reached values greater than 0.6 with count data. This problem is particularly pronounced with *MRPP-EUD*, with nearly all tests wrongly rejecting H_0 under high variance heterogeneity conditions with continuous and count data. Severely inflated type I error was also manifested for Q_b-*BRAY* and Q_b-*MAN* in the simulations based on the grassland data set. The poor performance of the statistics tested in this paper (with the notable exception of Q_b-*EUD*) is in sharp contrast with results reported by Warton and Hudson (2004), where none of the tested statistics, including the ones evaluated here, were found to be inaccurate (but see Warton et al. 2012). A possible explanation for this discrepancy may be the relatively "benign" range of conditions evaluated in the data sets evaluated in Warton and Hudson (2004), which may have been less "challenging" than the broad spectrum of aggregation levels and variance heterogeneities used in the present paper.

Results showed that the data correlation structure affects test accuracy (Table 2) and power (Fig. 1e-f), which has been demonstrated by Mielke and Berry (1999, 2001) for the MRPP with bivariate data. When all variables are positively correlated the group differences are along the main axis of variation and the effect of the correlation level on power is not influenced by the test statistic (Fig. 1e); when the variables are structured in two negatively correlated groups (Fig. 1f), the effect of the correlation level on power is more pronounced for *Qb-BRAY* and less pronounced for *MRPP-EUD* than for the other test statistics. Results also indicated an interaction between the effects of correlation structure and variance heterogeneity on test accuracy and power, especially for *MRPP-EUD* (see Figs. 2d-f), suggesting that this

test statistic is very sensitive to variance heterogeneity, with the strength of this effect depending on differences in correlation structure between groups. Though the simulations in this case involved only continuous, normally distributed data, by analogy we could expect the same effect with count data.

In general, inaccurate (lower than α) type I error seems to be associated with lower power. In several instances, the simulation results have shown that test statistics giving such inaccurate type I errors had lower power than those that were accurate under the same data simulation choices. The opposite was the case for inflated type I errors: a test with an inflated type I error may appear to have higher power when it, in reality, does not. Therefore, power and type I errors should not be interpreted separately and the inaccuracy indicated by type I error is necessarily also reflected in power. This relationship is reflected in the performance of Q_b-*BRAY* and *MRPP-EUD,* which showed higher power with increased variance heterogeneity due to the inherent bias of these two statistics (Fig. 2c,f). Based on these results, the higher power reported for *LR-IND* and *SUM-F* by Warton and Hudson (2004), and the supposed advantages (i.e., lower sensitivity to extreme values) of *MRPP-EUD* inferred by Zimmerman et al. (1985), Mielke and Berry (1999, 2001) should both be taken cautiously since power alone, without consideration of accuracy, may be misleading. Likewise the same caution is necessary with log transformation, which Warton and Hudson (2004) reported to give higher power, but which both this paper and McArdle and Anderson (2004) have found to produce inflated type I errors (Fig. 2b).

The results thus suggest avoiding the use of *SUM-F*, *LR-IND*, *MRPP-EUD* and the other distance-based statistics tested here - excepting Q_b-*EUD* - in permutation tests for group mean differences, for they may produce unacceptably inflated type I errors under scenarios that are common in ecological data, such as high variance heterogeneity and levels of aggregation. This conclusion contradicts recommendations made by Warton and Hudson (2004) regarding the use of *SUM-F*, *LR-IND* with abundance data. Furthermore, *MRPP-EUD* presented a much higher type I error inflation than the other statistics when increased variance heterogeneity was combined with unbalanced group sizes (see Fig. 3c-d).

In addition to problems with type I error inflation under heterogeneous variances, Q_b-*BRAY* showed lower power compared to Q_b-*EUD* in extremely unbalanced designs. The advantage of adopting Q_b as a test statistic is the freedom to choose the most appropriate distance measure for the data at hand (Anderson 2001). However, the use of the Bray-Curtis measure, very popular among ecologists, and Chord distances (with count data) should not be recommended for the computation of Q_b in distance-based MANOVA, unless group variances are homogeneous. Furthermore, the use of MRPP based on within-group averaged (non-squared) Euclidean distances (*MRPP-EUD*) should not be recommended except for "well-behaved" data sets, with homogeneous variances and homogeneous correlation structures across groups. Otherwise, *Qb-EUD* or an equivalent statistic (e.g., *F-ratio*) should be used. In summary, the results suggest that, in one-factor MANOVA, Q_b-*EUD* is the most robust of the tested statistics, combining both power and accuracy.

It is well known that the Euclidean distance will attribute a low dissimilarity to sites with low total abundances that are in fact very different in composition, with consequences in exploratory analysis (Orlóci 1978). This is likely the case in the simulations with highly aggregated count data and may explain the small to moderate inaccuracies when simulated count data contained a small number of variables (Table 2). Nevertheless, Q_b-*EUD* was the most robust of the tested statistics for one-factor MANOVA in spite of the known limitations of the Euclidean distance. A deeper examination of this apparent paradox would be out of the scope of this paper.

It may be argued that tests that are inaccurate with heterogeneous group variances are actually detecting an effect (unequal dispersion) that is indeed ecologically relevant. For instance, *MRPP-EUD* is recognized as an omnibus test (Mielke and Berry 2001) that could detect both unequal group means (locations) and unequal dispersions. Nevertheless, if the null hypothesis is rejected in a test of group mean differences, one expects that the differences result from unequal locations and not unequal dispersions. A specific test for unequal dispersions exists (Anderson 2006) which could be applied in combination with the tests that may confound these effects, but it will not tell anything about locations and thus will not solve the confounding problem when the inflated type I error in the test of unequal locations using the abovementioned statistics is caused by unequal dispersions.

Acknowledgements: The research leading to this paper has been financially supported by CNPq (Brazil) and was partly developed during short stays at the Department of Mathematics of the University of Rome "La Sapienza" and at the University of Göttingen, Germany. I am thankful to S. Camiz and H. Behling for the invitations and kind hospitality. Early versions of the paper received valuable comments from L. Orlóci, M. Dale, A. Melo, R. Peet, M. Anderson, H. Safford, B. Manly, V. Bastazini, and anonymous reviewers. I thank V. Bastazini for helping in the manuscript preparation.

References

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

Anderson, M.J. 2006. Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* 62: 245-253.

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

Bradley, J.V. 1968. *Distribution-Free Statistical Tests.* Prentice-Hall, Englewood Cliffs.

Clarke, K.R. 1993. Non-parametric multivariate analyses of changes in community structure. *Aust. J. Ecol.* 18: 117-143.

Crowley, P.H. 1992. Resampling methods for computation-intensive data analysis in ecology and evolution. *Annu. Rev. Ecol. Syst.* 23: 405-447.

Edgington, E.S. 1969a. Approximate randomization tests. *J. Psychol.* 72: 143-149.

Edgington, E.S. 1969b. *Statistical Inference: The Distribution-Free Approach.* McGraw-Hill, New York.

Edgington, E.S. 1987. *Randomization Tests*. Marcel Dekker, New York.

Edwards, A.W.F. and Cavalli-Sforza L.L. 1965. A method for cluster analysis. *Biometrics* 21: 362-375.

Fisher, R.A. 1951. *The Design of Experiments*. 6th ed. Oliver and Boyd, Edinburgh.

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Booth, M. and Rossi, F. 2003. *GNU Scientific Library Reference Manual (2nd Ed). Available also at* http://www.gnu.org/software/gsl/.

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

Hope, A.C.A. 1968. A simplified Monte Carlo significance test procedure. *J. R. Stat. Soc.* 30: 582-598.

Hotelling, H. 1931. The generalization of Student's ratio. *Annals of Math. Stat.* 2: 360-378.

Kempthorne, O. 1952. *The Design and Analysis of Experiments*. Wiley, New York.

Kempthorne, O. 1955. The randomization theory of experimental inference. *J. Amer. Statistical Assoc.* 50: 946-967.

Legendre, P. and Anderson, M.J. 1999. Distance-based redundancy analysis: testing multi-species responses in multi-factorial ecological experiments. *Ecol. Monogr.* 69: 1-24.

Legendre, L. and Legendre, P. 1998. *Numerical Ecology 2nd ed.* Elsevier, New York.

Lehmann, E.L. 1993. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *J. Amer. Statistical Assoc.* 88: 1242-1249.

Manly, B.F.J. 2007. *Randomization, Bootstrap, and Monte Carlo Methods in Biology*. Chapman & Hall/ CRC, Boca Raton.

Mantel, N. and Valand, R.S. 1970. A technique of nonparametric multivariate analysis. *Biometrics* 26: 547-558.

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

McArdle, B.H. and Anderson, M.J. 2004. Variance heterogeneity, transformations, and models of species abundance: a cautionary tale. *Can. J. Fish. Aquat. Sci.* 61: 1294-1302.

Mielke, P.W. and Berry, J.A. 1999. Multivariate tests for correlated data in completely randomized designs. *J. Educ. Behav. Stat.* 24: 109-131.

Mielke, P.W. and Berry, J.A. 2001. *Permutation Methods: A Distance Approach*. Springer-Verlag, New York.

Mielke, P.W., Berry, K.J. and Johnson, E.S. 1976. Multi-response permutation procedures for a priori classifications. *Commun. Stat. Theory Meth.* 5: 1409-1424.

Orlóci, L. 1967. An agglomerative method for classification of plant communities. *J. Ecol.* 55: 193-205.

Orlóci, L. 1978. *Multivariate Analysis in Vegetation Research*. Junk, The Hague.

Orlóci, L. 1993. The complexities and scenarios of ecosystem analysis. In: Patil, G.P. and Rao, C.R. (eds.) *Multivariate Environmental Statistics*, Elsevier, Amsterdam. pp. 423-432.

Peres-Neto, P.R. and Jackson, D.A. 2001. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia* 129: 169-178.

Pillar, V.D., Jacques, A.V.A. and Boldrini, I.I. 1992. Fatores de ambiente relacionados à variação da vegetação de um campo natural. *Pesqui. Agropecu. Bras.* 27: 1089-1101.

Pillar, V.D. and Orlóci, L. 1996. On randomization testing in vegetation science: multifactor comparisons of relevé groups. *J. Veg. Sci.* 7: 585-592.

Podani, J. 2000. *Introduction to the Exploration of Multivariate Biological Data*. Backuys Publishers, Leiden.

Potvin, C. and Roff, D.A. 1993. Distribution-free and robust statistical methods: viable alternatives to parametric statistics? *Ecology* 74: 1617-1628.

Romesburg, H.C. 1985. Exploring, confirming, and randomization tests. *Comput. Geosci.* 11: 19-37.

Torres, P.S., Quaglino, M.B. and Pillar, V.D. 2010. Properties of a randomization test for multifactor comparisons of groups. *J. Statist. Comput. Simulation* 80: 1131 - 1150.

Warton, D.I. and Hudson, H.M. 2004. A MANOVA statistic is just as powerful as distance-based statistics, for multivariate abundances. *Ecology* 85: 858-874.

Warton, D.I., Wright, S.T. and Wang, Y. 2012. Distance-based multivariate analyses confound location and dispersion effects. *Methods Ecol. Evol.* 3: 89-101.

White, G.C. and Bennets, R.E. 1996. Analysis of frequency count data using the negative binomial distribution. *Ecology* 77: 2549-2557.

Wilks, S.S. 1932. Certain generalizations in the analysis of variance. *Biometrika* 24: 471-494.

Zimmerman, G.M., Goetz, H. and Jr. P.W. Mielke. 1985. Use of an improved statistical method for group comparisons to study effects of prairie fire. *Ecology* 66: 606-611.

Received May 14, 2013
Revised July 15, 2013
Accepted July 19, 2013

Electronic Appendices 1-7

Data examples, details on data simulation, full simulation results and supplemental figures, all referenced in the paper. The file may be downloaded from the web site of the publisher at www.akademiai.com.