Ecology, 85(3), 2004, pp. 858–874

A MANOVA STATISTIC IS JUST AS POWERFUL AS DISTANCE-BASED STATISTICS, FOR MULTIVARIATE ABUNDANCES

DAVID I. WARTON[1,3] AND H. MALCOLM HUDSON[2]

[1]*Department of Biological Sciences, Macquarie University, Sydney, New South Wales 2109 Australia*
[2]*Department of Statistics, Macquarie University, Sydney, New South Wales 2109 Australia*

Abstract. There is now quite an extensive literature based on analysis of multivariate abundances, which have often been collected according to a MANOVA design. Many test statistics have been proposed specifically for this case, yet remarkably the power of these methods has not previously been compared. In this paper, the power of distance-based statistics (e.g., Mantel, analysis of similarities) is compared to variable-based statistics (e.g., redundancy analysis, the sum of ANOVA *F* statistics), when using permutation tests to assess significance of all statistics. Different choice of transformation, standardization, and distance measure were considered.

For 19 data sets taken from the literature, *P* values for the different statistics were compared. Power simulations were then conducted, where data were generated to mimic the properties of each of the 19 data sets.

For transformed data, using different distance measures (Euclidean, Manhattan, Bray-Curtis) and different distance-based statistics had little effect on power. Overall, statistics based on multivariate analysis of variance (MANOVA) were at least as powerful as others, although particular data sets gave different results. The distance-based statistics most commonly used in the literature do not standardize abundances, so these were more powerful when effects are present in taxa that are more variable (on the transformed scale), and less powerful otherwise.

There are several reasons to prefer a statistic based on MANOVA to others (e.g., interpretability, generalization to more complex designs), and so we generally recommend that the MANOVA-based statistics used here be preferred to distance-based statistics.

Key words: community composition data; empirical power comparison; hypothesis tests; MANOVA, with permutation tests; multivariate analysis; permutation test; randomization test; statistical methods, comparisons.

INTRODUCTION

One of the most widely used types of ecological data is multivariate abundance data (abundance in each sample is recorded for many taxa, then each taxon treated as a variable). Often multivariate abundances are collected in several groups of samples, with the main purpose being to test for differences in abundance among these groups. For example, to test for community-level effects of insecticide on zooplankton, several treatments of insecticide could be applied to experimental ponds, and abundance of different types of zooplankton compared across treatments, as in Kreutzweiser et al. (2002). Alternatively, in attempts to describe the effects of sewage outfalls on benthic invertebrates, invertebrate abundances from samples near several outfalls might be compared to controls, as in Morris and Keough (2002). The procedures required for such research will be referred to in this paper as requiring a ''MANOVA test,'' as these require multivariate versions of procedures for which analysis of variance (ANOVA) is commonly used in the univariate case.

Several statistics have been suggested specifically for conducting MANOVA tests for this type of data (Smith et al. 1990, Clarke 1993, Pillar and Orloci 1996, Anderson 2001). Clearly it is important to understand whether these statistics all perform equally well, or if one should be preferred over others. In this paper we review the different types of statistics used for MANOVA tests of multivariate abundances, and compare their power.

The various statistics currently used for conducting a MANOVA test of multivariate abundances fall into two distinct categories: distance-based and variable-based statistics. Distance-based statistics are a function of distances between samples (most commonly, the Bray-Curtis distance). An example statistic is the average of all within-group distances (Mielke et al. 1976). Variable-based statistics are a function of summary statistics that have been produced for each variable. An example is the sum across all taxa of logarithms of ANOVA *F* statistics calculated for each of the taxa (Edgington 1995:188).

Manuscript received 15 July 2002; revised 9 May 2003; accepted 27 May 2003; final version received 25 June 2003. Corresponding Editor: N. C. Kenkel.

[3] Present address: Department of Statistics, School of Mathematics, University of New South Wales, NSW 2052 Australia. E-mail: dwarton@maths.unsw.edu.au

Should distance-based statistics be preferred to variable-based, or vice versa?

The researchers who suggested distance-based statistics found the assumptions of common procedures using variable-based statistics to be inappropriate for analyzing multivariate abundances. More specifically, it has been noted that the assumptions of multivariate analysis of variance (MANOVA) are unsatisfactory (abundances do not follow a multivariate normal distribution), and that simulations show the Euclidean distance is inappropriate for analysis of (untransformed) multivariate abundances (Clarke and Green 1988, Smith et al. 1990, Anderson 2001).

However, there remain the possibilities of using a MANOVA statistic on transformed abundances, or another variable-based statistic. It is known that the properties of ANOVA procedures are robust to modest violations of assumptions, particularly for balanced data (Miller 1986). If abundances are transformed to reduce skew, a MANOVA statistic may have desirable properties. In addition there is the little-explored possibility of using variable-based statistics other than MANOVA.

There are two important reasons for preferring a variable-based statistic to a distance-based statistic, if one is found that has similar performance to distance-based statistics:

1) *Taxon-level effects.* Usually it is of interest to identify the taxa that differ most strongly in abundance across groups. Summary statistics for each taxon are calculated in constructing variable-based statistics, so these can be used to determine which taxa most strongly express differences among groups of samples (for example, if the test statistic is a function of ANOVA F statistics, the F statistics can be compared across taxa). For distance-based statistics, in contrast, additional, more complicated methodology must be used, such as the SIMPER procedure (Clarke 1993) or a leave-one-out procedure (Smith 1998). On the strength of this argument, some previous authors have preferred a redundancy-analysis approach (van den Brink and ter Braak 1998) or Procrustes analysis (Peres-Neto and Jackson 2001) to distance-based statistics.

2) *Transparency of method.* When using a distance-based statistic, it is unclear what is being tested about the abundance data (e.g., are we testing if mean abundance differs among groups, median abundance, or something else?). Hence it is unclear what types of effects are more easily detected, or under what conditions a particular distance-based statistic is appropriate. In contrast, most variable-based statistics have an underlying model with explicit assumptions, e.g., the sum of ANOVA F statistics will be appropriate when typical ANOVA assumptions hold. An important consequence of failure of assumptions is the possibility of poor power properties (Staudte and Sheather 1990), so even if using a permutation test to ensure valid inference, if the sum of ANOVA F statistics is used, it is desirable that typical ANOVA assumptions hold to ensure good power properties. Similarly, it may also be desirable that ANOVA assumptions hold when using a distance-based statistic (to ensure good power properties), but simulations would be required to investigate this.

There are further reasons to prefer a variable-based statistic that is a simple generalization of univariate ANOVA statistics:

3) *Simplicity.* When only one abundance variable is of interest, data are routinely transformed, and univariate ANOVA applied. So why not use a generalization of this method when analyzing multivariate abundances, rather than a completely different approach?

4) *Analysis of complex designs.* Extensions of ANOVA methods to more complex designs are well-known, e.g., multi-factor designs or when covariates are present (Underwood 1997), or repeated-measures designs (von Ende 2001). For multivariate abundances, if the chosen test statistic is a function of univariate ANOVA statistics, its extension to complex designs is relatively straightforward.

A particular example of a statistic based on ANOVA is the following:

$$\text{LR-IND} = \sum_{j=1}^{p} N \log\left(\frac{\text{SS}_{j,0}}{\text{SS}_{j,1}}\right)$$

where N is the total number of samples, p the total number of taxa sampled, and $\text{SS}_{j,0}$ and $\text{SS}_{j,1}$ denote the univariate residual sums of squares (see Eq. 1 of the Appendix for more details) for the jth variable under H_0 and H_1 respectively. This statistic is referred to here as ''LR-IND'' for easy reference, and is of particular interest because it is a special case of a common MANOVA statistic, Wilk's Λ (Anderson 1984:293). Wilk's Λ is the likelihood-ratio test for a MANOVA model, and it simplifies to LR-IND if it is assumed that all variables are independent. The assumption of independence of variables is made in all test statistics considered in this paper, it is a necessary assumption so that test statistics can be calculated even when there are many more variables than samples.

Like all other statistics considered in this paper, it is best to use LR-IND only once abundances have been transformed to reduce skew (so that test statistics do not have low power because of the influence of a few large abundances), and only if permutation tests are used (so that inference is robust to failure of the assumption of independent variables).

Specific aims

This paper compares the power of statistics used to conduct MANOVA tests of multivariate abundance data. This subject has never previously been addressed.

Twenty multivariate abundance data sets are used in this study, to be representative of the multivariate abundance data typically encountered in practice. If simulations had been conducted without reference to a col-

TABLE 1. Distances considered in analyses, and their expressions in terms of (possibly transformed and standardized) abundances **y**.

Distance	Expression	Reference
Euclidean distance	$d_{ii'} = \left[\sum_{j=1}^{p} (y_{ij} - y_{i'j})^2\right]^{1/2}$	Legendre and Legendre (1998)
Bray–Curtis distance	$d_{ii'} = \sum_{j=1}^{p} \frac{\lvert y_{ij} - y_{i'j}\rvert}{\sum_{j=1}^{p} y_{ij} + y_{i'j}}$	Bray and Curtis (1957)
Manhattan distance	$d_{ii'} = \sum_{j=1}^{p} \lvert y_{ij} - y_{i'j}\rvert$	Legendre and Legendre (1998)

lection of real data sets, results might not be representative of the multivariate abundance data sets typically encountered. The use of many real data sets in this study ensures that conclusions are relevant to the analysis of multivariate abundances in general.

All *P* values are calculated by permutation, which provides an exact test of the hypotheses of interest (or nearly exact, with some error introduced if only a random sample of permutations is used). Hence Type I error is appropriately controlled by all statistics considered in this paper, and the remaining issue is Type II error (or power) when there are known differences in means among the groups being compared. We address the following, for the types of multivariate abundance data typically encountered:

a) Are distance-based statistics necessary, or are alternative statistics as powerful or more powerful in general?

b) Do the distance-based statistics all have similar power?

c) To what extent does transformation, standardization, or choice of distance measure affect conclusions?

This work is also the first to compare different distance measures for different transformations and standardizations, in more than one real data set. Previous studies (Faith et al. 1987, Jackson 1993, Thorne et al. 1999) have assessed the importance of transformation, standardization, or choice of distance measure, usually for the purpose of ordination. Some of these studies (Jackson 1993, Thorne et al. 1999) were based on one data set, which obviously limits their generality. Simulation studies previously conducted (Faith et al. 1987) only considered untransformed data.

METHODS

Procedures to be compared

There are four different stages in conducting a test of multivariate data, performed in the following sequence: Transformation; Standardization (if using a scale-dependent statistic); Choice of distance measure (if using a distance-based statistic); and Choice of test statistic.

The following sections describe the choices considered in this paper for each of the above steps.

Transformation.—Transformation of abundances reduces the influence of outliers, and can remove heteroscedasticity. The transformations considered in this study are: (1) **y**, i.e., untransformed data; (2) $\mathbf{y}^{1/4}$; and (3) $\log(\mathbf{y}/a + 1)$ (abbreviated as ''log-transformation'' in this paper) where a is the minimum possible non-zero abundance for a taxon.

The transformations to $\mathbf{y}^{1/4}$ and $\log(\mathbf{y} + 1)$ are the most commonly applied for abundance data, although $\log(\mathbf{y} + 1)$ has been criticized as scale dependent (Field et al. 1982). On the other hand, the transformation $\log(\mathbf{y}/a + 1)$ is scale invariant, and in the case of counted data it reduces to $\log(\mathbf{y} + 1)$, since 1 is the minimum possible non-zero counted abundance (so $a = 1$). In this study, some abundances were measured as percnt cover with minimum values of 0.1 or 0.01, in which case the log transformation does not reduce to $\log(\mathbf{y} + 1)$.

Standardization.—Standardization of variables ensures that more-variable taxa do not dominate analyses.

In this study, only distance-based statistics and redundancy analysis (RDA) required standardization. The other statistics considered are scale invariant (unaffected by standardization of variables) or implicitly contain a standardization of variables.

Distance-based statistics were calculated both for standardized and unstandardized data. The method of standardization ensures the same average contribution of each taxon to distance (or to the numerator of distance, for Bray-Curtis). When Euclidean distances were used (or in redundancy analysis), data were standardized by sample standard deviation. In the case of Manhattan and Bray-Curtis distances, an appropriate standardization for the jth variable is proportional to the sum of absolute deviations, as in Mielke and Berry (2001:50):

$$s_j = \sum_{i=1}^{N} \sum_{i'=1}^{i} \lvert y_{ij} - y_{i'j}\rvert.$$

Distance measures considered.—Table 1 describes the three distance measures considered in this paper.

TABLE 2. Test statistics to be compared.

Statistic†	Reference
a) Distance-based	
Smith	Smith et al. (1990)
Pillar-Orloci	Pillar and Orloci (1996)
ANOSIM	Clarke (1993)
b) Variable-based	
RDA	ter Braak and Smilauer (1998)
CCA	ter Braak and Smilauer (1998)
LR-IND	this paper
$\sum F$	Edgington (1995: 188)

† ANOSIM = analysis of similarities; CCA = canonical correspondence analysis; LR-IND = likelihood-ratio test, assuming independence of variables; RDA = redundancy analysis; ΣF = the sum of ANOVA F statistics.

These distance measures were chosen either because of previous recommendations, or because the distance has a simple form. The Euclidean distance has been recommended for this type of analysis as the most interpretable measure (Mielke and Berry 2001). For multivariate abundances, however, the Bray-Curtis distance is the most common choice in the literature. The Manhattan distance has a simpler form than the Bray-Curtis distance, and also is a function of absolute values of distances, and so is expected to be robust to outliers (Gower and Legendre 1986).

In one of the data sets analyzed in this paper, no taxa were present in two observations. This presented a computational problem for the Bray-Curtis, because the denominator equalled zero for the distance between these two observations. In this case the distance was set to 1 because for the Bray-Curtis, the interpretation of $d_{ii'} = 1$ is that no taxa are present in both observations i and i'. Observations with 0 total abundance were not excluded from the data set, because this would have biased estimates of mean abundance.

Test statistics.—We considered the test statistics in Table 2. The expressions for calculation of each of these statistics are given in Appendix A, as are details on other papers that suggest distance-based statistics equivalent to these. In particular, the Smith and Pillar-Orloci statistics are specific cases of MRPP (multiresponse permutation procedure; Mielke et al. 1976), or equivalently, of Mantel (Mantel 1967) statistics.

None of these statistics accounts for correlation of abundance between taxa, and RDA and distance-based statistics do not account for the different variability of abundance in different taxa. This can be seen by looking at the expressions for the distance measures (Table 1) and variable-based statistics (see Appendix A) considered: none of these expressions include a measure of correlation between variables, and in RDA and distance-based statistics, abundance of each taxon is not weighted by some measure of variability. Only LR-IND and ΣF are scale-invariant statistics.

The ANOSIM statistic, as defined in this paper (see Table A.1), has a similar form to the Smith statistic. The only difference between these is that the ANOSIM (analysis of similarities) statistic is a function of ranks of distances, whereas the Smith statistic is a function of the distances themselves.

The RDA and ΣF statistics are rare cases in which it is possible for a variable-based statistic to be reexpressed as a distance-based statistic. The RDA statistic is equivalent to the Pillar-Orloci statistic calculated on Euclidean distances, and ΣF is equivalent to RDA, where data are standardized to equal within-group variance (and this standardization is reapplied to each resampled data set). In this paper, the standardization used for RDA and distance-based statistics was to standardize by some measure of total variability. This can be compared to standardizing by a measure of within-group variability by comparing RDA calculated on standardized data to ΣF.

The LR-IND and ΣF statistics will be referred to as ''MANOVA statistics,'' because they can be derived from the respective MANOVA statistics Wilk's Λ and Hotelling-Lawley trace (Anderson 1984:323), by assuming independence of variables.

Power comparison

The power of different test statistics was compared in two ways: (1) Comparison of P values, and (2) Comparison of power estimated from simulations.

This work used 20 data sets that were taken from applications (Table 3). These will be referred to in the following as ''reference data sets,'' a notation necessary to simplify explanations in this section.

All work was conducted on Matlab version 5 (MathWorks 1998).

In all instances where a test statistic was calculated, any variable with only one (or no) non-zero element was excluded. Such a variable (on its own) could provide no useful information for the designs considered. On the other hand, variables with two non-zero elements were included, because such variables could potentially provide reasonable evidence against H_0 for many of the designs considered.

The significance of all test statistics was assessed using permutation tests. For most reference data sets, P values were determined by permuting samples within the groups defined under H_0 (hence among the groups defined under H_1). This is known as ''restricted permutation'' or ''restricted randomization'' (Manly 1997: 127), attributed to Edgington (1995). For reference data sets that included nested factors not being tested (indicated in Table 3), blocks of samples representing different levels of the nested factor were permuted. This method enabled testing of the hypotheses of interest while controlling for the effects of other nested factors.

Some of the reference data sets considered (as indicated in Table 3) were not sampled as MANOVA designs, but sample descriptions were available or continuous variables were measured for each sample. In these cases, factors were derived from the available

TABLE 3. Reference data sets, and their properties.

Source	Abundance	N	p	Subsampled†	No. of factors‡	Balanced?§
N. Andrews	count	120	13	time	1^n	yes
A. Pik	count	49	24	···	1^n	no!
Moulton (1982)	count	14	36	obs., time	1^n	no
Moulton (1982)	count	16	28	time	3	yes
I. Lunt	count	20	44	time	1	yes
B. Rice	% cover	41	135	treatments	1^d	no!
J. Overton	% cover	38	139	···	2^d	no!
B. Rice	% cover	13	76	treatments, time	1^d	no
B. Rice	% cover	39	293	*excluded*	2	no
Clements (1980)	% cover	24	22	treat., var., time	2	yes
A. Pik	count	134	21	···	1^n	yes
Warwick et al. (1990*b*)	count	12	17	treatments	1	yes
T. H. Pearson and J. Blackstock‖	biomass	12	46	···	2^d	no!
Gray et al. (1990)	count	22	113	treatments	1^d	no
Warwick et al. (1988)	count	12	43	treatments	1	yes
Gee et al. (1985)	count	12	39	···	1	yes
Warwick et al. (1990*a*)	count	16	11	···	2	yes
van Dobben et al. (1999)	count	32	77	···	2	yes
van den Brink et al. (1996)	count	12	31	variables, time	1	no!
van der Aart and Smeenk-Enserink (1970)	count	28	12	···	2^d	no!

Notes: Data are unpublished if a name is given instead of a citation. Size of data set analyzed is $N \times p$.

† ''Subsampled'' refers to whether a data set had been reduced in size by selecting a subset of all treatments, variables, and observations, or by analyzing only one of several sampling times.

‡ ''No. of factors'' is the number of factors involved in the hypothesis test considered here; superscript *n* means there was an additional nested factor in the design, so that blocks of samples were permuted rather than samples; superscript *d* means data were not originally collected according to a MANOVA design, and factors for a MANOVA test later derived from information on samples.

§ ''Balanced?'' refers to whether or not sampling was balanced, and ''no!'' indicates sample sizes in different groups differed by a factor of 2 or more.

‖ Their 1984 report (''Garroch head sludge dumping ground survey'') is available from Dunnstaffnage Research Laboratory, Oban, Scotland.

data, in a manner that kept designs as balanced as possible.

For each reference data set, a hypothesis test was chosen such that most statistics suggested there was some evidence against the null, but not overwhelming evidence against the null ($0.001 < P < 0.1$). It was important that there be some evidence against the null ($P < 0.1$ for most statistics) to ensure that in most cases H_1 was true, so that a powerful test statistic could be expected to have small P values in most cases. However, if there was overwhelming evidence against the null ($P \leq 0.001$ for most statistics), then P values for all test statistics would be similar and their comparison uninformative. In multi-factor designs, there were several possible hypothesis tests, from which one was chosen to satisfy $0.001 < P < 0.1$ (for most statistics) while also being a hypothesis of some biological interest (e.g., testing for an effect of latitude was preferred to testing for an effect of study site within latitudes). When there was strong evidence against H_1 for all possible hypothesis tests, a subset of the data set was used (as indicated in Table 3) to ensure that $0.001 < P < 0.1$ for most statistics. One data set (Table 3: ninth entry) was excluded because of lack of evidence for H_1 (all P values were large, $P > 0.1$ for most statistics).

Abundances were measured at several times in some data sets (indicated in Table 3). Abundances at only one sampling time were considered here, so that abundances within a taxon were not correlated (which would complicate analysis).

Comparison of P values.—As a rough measure of power, P values were calculated for the 19 reference data sets (Table 3). A statistic that is generally more powerful than others in practice will have generally lower P values, and very different P values for a given data set suggest statistics have very different power in particular instances.

All P values were calculated exactly or close to exactly, by permutation. When the total number of possible permutations was less than 10 000, these were systematically sampled and the test was exact. In other cases, P values were estimated from 10 000 random permutations, and the test was no longer exact because of the introduction of Monte Carlo error (from the random selection of permutations). However, this error was negligible (for example, the standard error of a P value of 0.05 is only 0.002).

P values were interpreted on a proportional (logarithmic) scale (using geometric means rather than arithmetic means), and truncated at 0.001. This was done to remain consistent with the way P values are interpreted in practice: P values of 0.01 and 0.1 are considered to be as different as P values of 0.001 and 0.01, and all P values smaller than 0.001 are usually interpreted the same way.

Power simulation.—More intensive analyses were conducted to directly estimate power under a variety of controlled situations. These simulations were very computationally intensive (requiring over three weeks of total computation time).

The power of a statistic was estimated as the proportion of times it was significant at the 0.05 level, for 1000 sets of data generated from the same distribution. For each of these data sets, statistical significance was assessed using 1000 permutations.

For each of the 19 reference data sets, three simulations were conducted, in which data were generated to mimic the properties of the reference data set. The negative binomial distribution was used to generate data (using the ''nbinrnd'' function included in Matlab's Statistics toolbox 2.1 [The MathWorks 1998]), as abundances in the reference data sets were noticed to have similar properties to this distribution (in terms of mean–variance relationship and frequency of zeros). Abundances in different taxa were generated independently. The design, number of variables, and samples was kept as in the reference data set, and the parameters of the negative binomial distribution (the mean μ and a nuisance parameter ϕ, such that the variance is $V(\mu) = \mu + \phi\mu^2$) were chosen to match sample estimates from the reference data set. For each simulation, variables were chosen to be either ''null'' (no difference in distribution of abundances across groups being compared) or ''effect'' variables (the distribution of abundances differs across the groups being compared). The means of null and effect variables were chosen to match sample estimates under H_0 and H_1, respectively, and a single nuisance parameter (ϕ) was chosen for each variable to equal the sample (moments) estimate under H_0 and H_1 for null and effect variables, respectively.

Independent variables were generated because it was not feasible to obtain sample estimates of parameters for a model accounting for correlation between variables. Correlated data could be generated through the use of random effects, the most appropriate model being a multivariate Poisson-lognormal model. However, estimation of parameters for this model is computationally demanding (McCulloch and Searle 2001:Chapter 10), and is not possible when p is at least as large as N ($N < p$ in many of the reference data sets, see Table 3). It is unlikely that abundances would be uncorrelated in practice, so assuming independence does not provide a realistic model of correlations between variables. However, as none of the test statistics considered in this paper account for correlation between variables, there is no reason why response to correlation would differ between statistics.

The three simulations conducted for each reference data set used different methods of choosing effect and null variables:

1) *No effect variables.* All variables were null variables. This simulation was used to demonstrate that restricted permutation testing provides exact significance levels (so power ≈ 0.05). Given computational demands, this simulation was only conducted for seven statistics (LR-IND and the Pillar-Orloci statistics for log-transformed data).

2) *Few effect variables.* A few effect variables were selected to have very different means and variances across groups. Effect variables were chosen from the reference data set as those whose univariate ANOVA statistics gave $P < 0.05$ for log-transformed data. The criterion $P < 0.05$ was used in choosing effect variables so that there were substantial differences among population means in simulations. For some reference data sets this provided too many or too few effect variables, so that power was very high or very low for all statistics being compared. In such cases, effect variables were selected using slightly different criteria (as those with $P < 0.1$, $P < 0.01$, or $P < 0.002$) to ensure (if possible) that $0.1 <$ power < 0.9 for all statistics.

3) *Many effect variables.* Many effect variables were selected to have very small differences in means and variances across groups. Effect variables were chosen as those with $P > 0.5$, when conducting a univariate ANOVA following log-transformation, on each variable from the reference data set. Again this criterion was modified (to $P > 0.6$ or $P > 0.4$), to ensure (if possible) that $0.1 <$ power < 0.9 for most statistics.

Power has been presented on an arithmetic scale. The average power of a statistic across all reference data sets then has an interpretation as the overall proportion of times the statistic was significant at the 0.05 level.

Results

Most differences in results could be attributed to whether or not a reference data set was very unbalanced (i.e., sample size of groups varied over a factor of 2 or more), and to the relative magnitude of the variance of effect variables. Consequently, all figures and tables distinguish very unbalanced data (those for which sample sizes in some groups differ by a factor of 2 or more) from other data sets, and where appropriate, figures use different marker sizes according to the how large the variance of effect variables is compared to the variance of null variables. The index used for marker size was the ratio of average variance of effect variables to average variance of null variables (of appropriately transformed abundance).

Computation time was not an issue for significance testing of a given data set. The CCA (canonical correspondence analysis) statistic was the most computationally demanding, yet for the largest data set considered here, it took only one minute to evaluate its significance with 10 000 permutations. The raw results (P values for the 19 data sets and power from simulations of these data sets) can be found in Appendix B.

Comparison of P values

Using geometric mean P value as a measure of power (Fig. 1a), there was similar power for different trans-

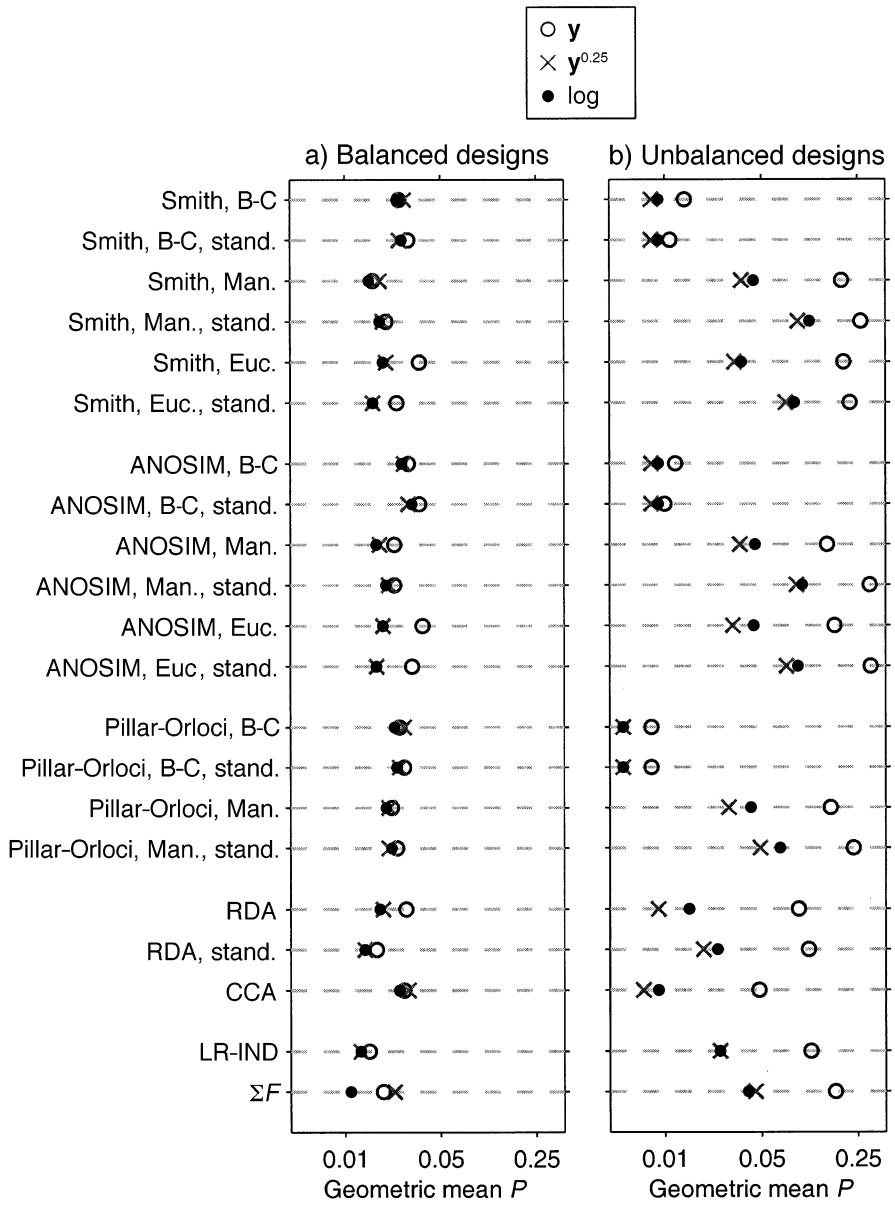

FIG. 1. Comparison of geometric mean *P* value of different statistics. Plotted points are geometric mean *P* value across (a) 13 data sets with balanced (or close-to-balanced) designs, and (b) six data sets with very unbalanced designs (where sample sizes of some groups differ by a factor of 2 or more). Note that the *x*-axis scales are logarithmic. Abbreviations: ANOSIM = analysis of similarities, B-C = Bray-Curtis distance measure, CCA = canonical correspondence analysis, Euc. = Euclidean distance measure, LR-IND = likelihood-ratio test assuming independence of variables, Man. = Manhattan distance measure, Pillar-Orloci = the statistic due to Pillar and Orloci (1996), Smith = the statistic due to Smith et al. (1990), stand. = standardization of data, RDA = redundancy analysis, and ΣF = the sum of ANOVA *F* statistics. See Table 2 for additional citation information.

formations of data sets with balanced (or close to balanced) designs. In very unbalanced designs (Fig. 1b), however, *P* values were generally much lower when calculated on transformed abundances. This pattern was less obvious when the Bray-Curtis distance was used.

Only for unbalanced designs were *P* values generally lower when using the Bray-Curtis distance than for other approaches (Fig. 1). The opposite was observed for balanced designs (*P* values being relatively large for the Bray-Curtis distance), so that for transformed data, the geometric mean *P* value across all data sets was similar for all Pillar-Orloci statistics and all variable-based statistics.

Standardizing data did not substantially change the geometric mean *P* value for transformed abundances.

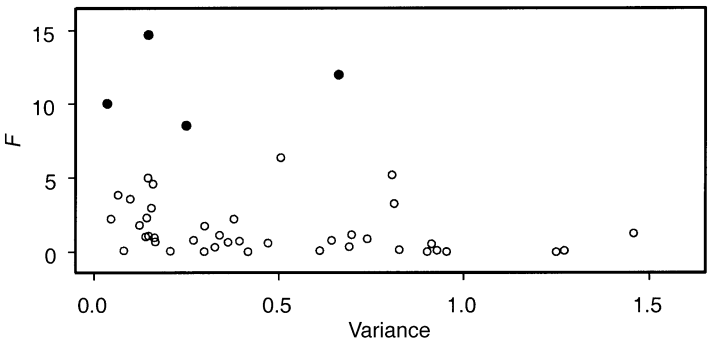

FIG. 2. ANOVA F statistic vs. sample variance for log-transformed abundance of each taxonomic group, for a reference data set where taxa with substantial differences in abundance between groups (i.e., high F [$F > 7$ and $P < 0.01$], plotted with solid circles) also have low variances.

However, substantial effects were occasionally observed. For example, for the RDA (redundancy analysis) statistic calculated on log-transformed abundances, one P value changed from 0.13 to 0.003 with standardization. Closer consideration of this data set revealed that a strong treatment effect was only observed in taxa whose abundances were less variable on the log-transformed scale (Fig. 2).

Power simulations

In the simulation with no effect variables, as expected, the average power of no statistic was significantly different from 0.05. In fact average power equalled 0.05 to two decimal places for all seven of the statistics for which a simulation with no effect variables was conducted. This was expected because using permutation tests ensures exact (or nearly exact) significance levels, when testing across several groups of observations for a difference in the distribution of abundances.

In the simulation with many effect variables, most data sets had 5–20 effect variables. In the simulation with few effect variables, there were 1–5 effect variables in most cases, with more on only two occasions.

Considering separately the simulation results for each data set, a variable-based statistic had the highest power in the majority of cases. A variable-based statistic had highest power in 24 of the 32 simulations for which there was not a tie. This was so despite there being almost 3 times as many distance-based statistics as there were variable-based ones, when considering all possible choices of standardization, transformation, and choice of distance.

In individual simulations and overall, power was usually higher for the MANOVA-based statistics (LR-IND and ΣF) than for most distance-based statistics (Fig. 3). In fact, average power was higher for ΣF than for any other statistic, in simulations with few effect variables, for transformed abundances (Fig. 3a). For transformed abundances in simulations with many effect variables, average power was similar for Pillar-Orloci statistics and MANOVA-based statistics, but lower for other distance-based statistics (Fig. 3b). Distance-based statistics had much lower power than MANOVA-based statistics in simulations with few effect variables, except when distance-based statistics were calculated on unstandardized log-transformed abundances, in which case average power was similar to MANOVA-based statistics (Fig. 3a). Even in this case, distance-based statistics in individual simulations rarely had substantially higher power than MANOVA-based statistics (Fig. 4). Substantial differences in power between MANOVA-based and unstandardized distance-based statistics occurred only when effect variables had very high or very low variance, compared to null variables. In three cases, for log-transformed data, the distance-based statistics had substantially higher power than LR-IND, when effect variables all had high variance (large symbols in Fig. 4b,e). In one case distance-based statistics had substantially lower power, when effect variables all had low variance (small symbol in Fig. 4c,f).

Choice of distance measure had little effect on the power of a test statistic, for transformed data. This was true of power overall (Fig. 3), of power in individual cases (Fig. 5), and of P values (Fig. 5a). This suggests that for log-transformed data, RDA could generally be used in place of distance-based statistics with no loss of power. This is in direct contradiction to what has been assumed to be the case in the past, it generally being considered inappropriate for analyses to be based on Euclidean distances. For untransformed abundances, statistics using the Euclidean distance had lowest power, and the Bray-Curtis distance should be preferred in this case (Fig. 3).

Choice of distance-based statistic had a small but consistent effect on average power (Fig. 3). The differences in power among statistics usually kept the rank order (ANOSIM $<$ Smith $<$ Pillar-Orloci) consistent with results for P values. However, the effect on power of choice of distance-based test statistic was small compared to the effects of transformation and standardization.

With few exceptions, statistics had greatest power when $\log(\mathbf{y}/a + 1)$ transformed, and least power when untransformed. This was true in most individual simulations (Fig. 6), not just in averages (Fig. 3). As was the case for P values, differences in power were small for balanced designs, but considerable for very unbal-

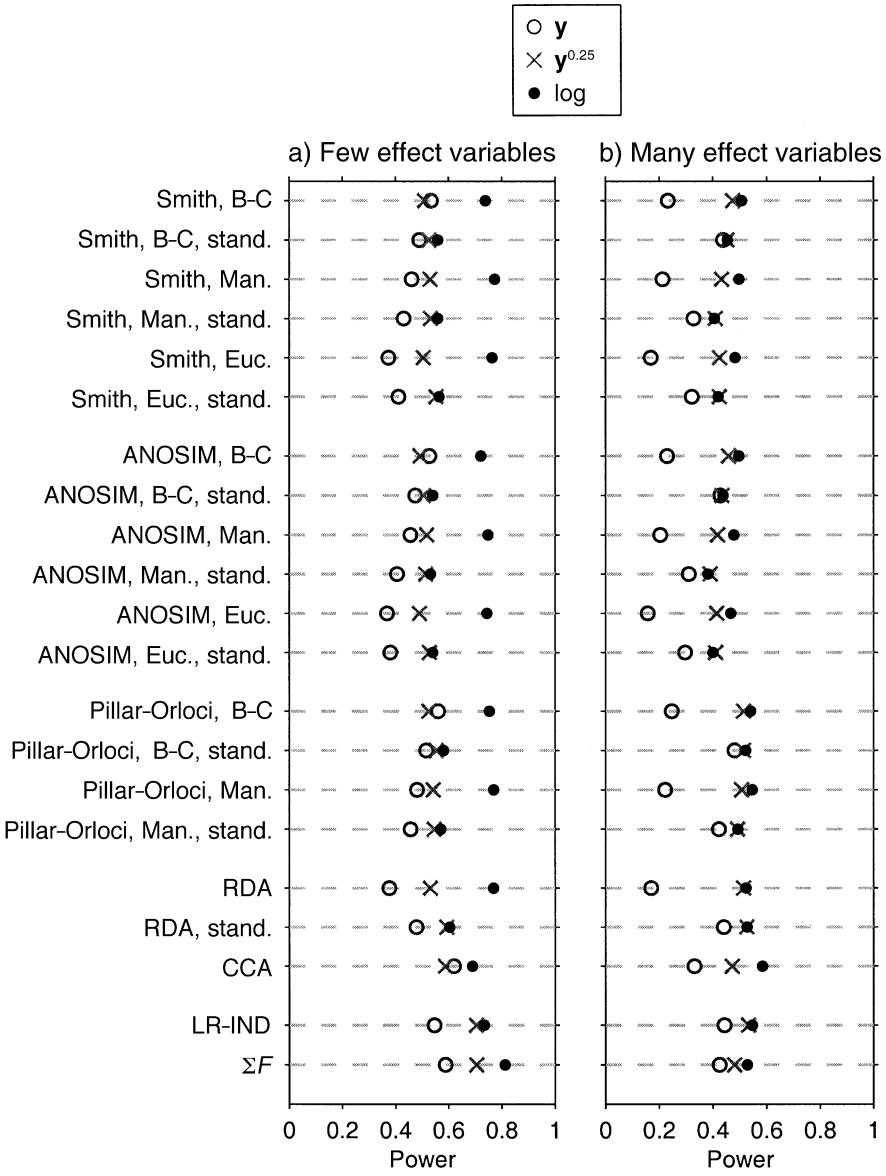

FIG. 3. Comparison of the overall power of different statistics. Plotted points are power averaged across the 19 data sets, for simulations (a) with few effect variables and (b) with many effect variables. Different symbols are used for different transformations of data. The RDA (redundancy analysis) statistic is equivalent to the Pillar-Orloci statistic for Euclidean distances. For transformed abundance, the standard error of average power was in the range 0.05–0.07, but it was 0.05–0.09 for untransformed abundance. See the Fig. 1 legend for the key to abbreviations.

anced designs. There was only a small difference in average power between log-transformed and $\mathbf{y}^{0.25}$-transformed abundance in most cases; however, the difference was substantial for unstandardized distance-based statistics in the simulation with few effect variables. This suggests that on the $\log(\mathbf{y}/a + 1)$ scale, the variance of effect variables was large for some data sets, whereas this was not the case on the $\mathbf{y}^{0.25}$ scale.

Standardization of RDA and distance-based statistics had substantial effects on power, which was not evident on consideration of *P* values. For log-transformed abundance, standardization was accompanied by a marked decrease in average power (Fig. 3a) when there were few effect variables. In such simulations, there was frequently a substantial loss of power with standardization, and rarely a substantial gain (Fig. 7b). The loss of power occurred only in cases when the effect variables had high variances, hence their influence on the test statistic was greater for unstandardized data. Standardization had little effect on power in simulations with many effect variables, except in one case where power was far higher for standardized data (0.84) than for unstandardized data (0.29), because all effect variables had small variances on the log-transformed

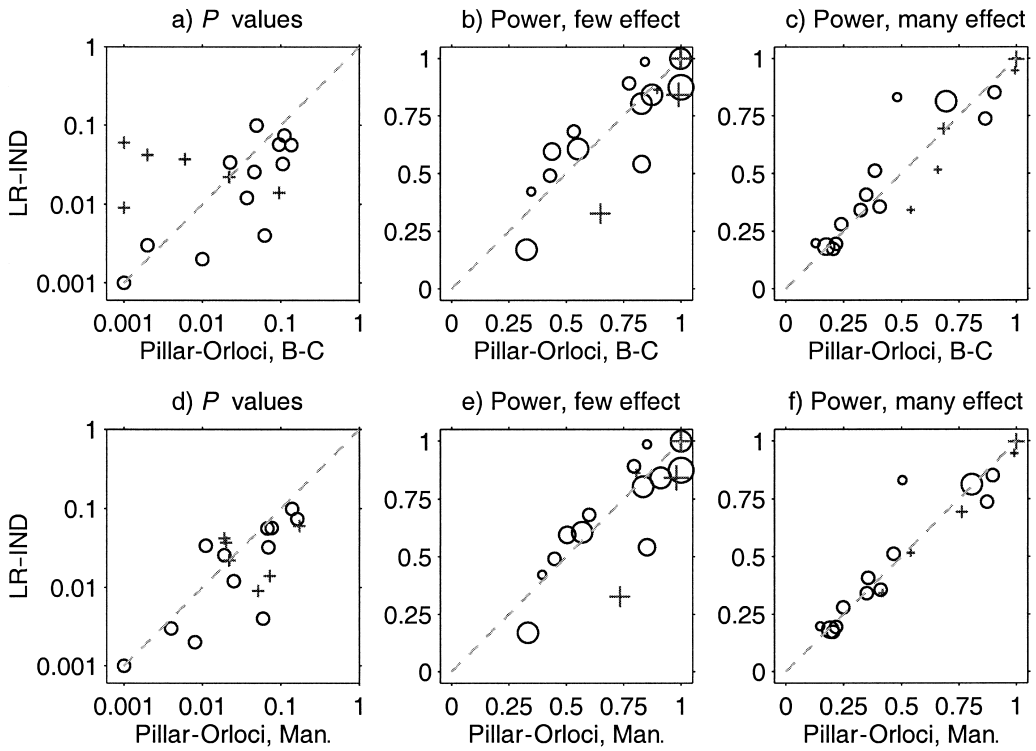

FIG. 4. Comparison of the power of a distance-based and a MANOVA statistic. For log-transformed data, the LR-IND statistic (likelihood-ratio test assuming independence of variables) is plotted against the Pillar-Orloci statistic calculated on unstandardized data using (a–c) the Bray-Curtis (B-C) distance or (d–f) the Manhattan (Man.) distance. ''Few effect'' and ''many effect'' indicate the relative number of effect variables. A ''+'' indicates very unbalanced data (where sample sizes in some groups differ by a factor of 2 or more). In panels b, c, e, and f, the circular marker size is larger when the average variance of effect variables is larger.

scale. For statistics based on untransformed abundances, such dramatic differences in power were common, and average power was substantially higher for standardized data when there were many effect variables (Fig. 3b).

The contrast between ΣF and the standardized RDA statistic demonstrates that there is a distinct power advantage in standardizing to equal within-group variance, rather than to equal total variance. The ΣF statistic is equivalent to an RDA statistic standardized by within-group variance (for each permuted data set). Power was usually similar for these two statistics. However, in several simulations with few effect variables, the ΣF statistic had substantially higher power than the standardized RDA statistic (Fig. 8), while the reverse never occurred.

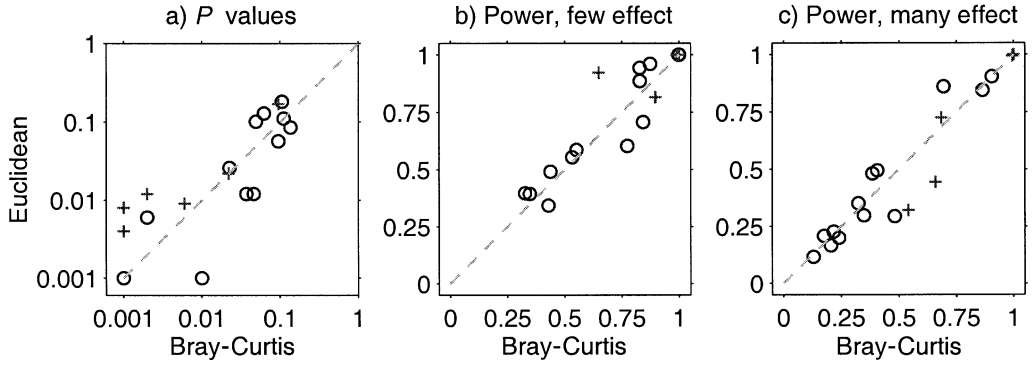

FIG. 5. The effect on power of using a different distance measure, for transformed abundances. ''Few effect'' and ''many effect'' indicate the relative number of effect variables. This figure shows the Euclidean vs. Bray-Curtis distance, for the Pillar-Orloci statistic calculated on unstandardized log-transformed data. Note that the Pillar-Orloci statistic for the Euclidean distance is equivalent to the RDA statistic. A ''+'' indicates very unbalanced data.

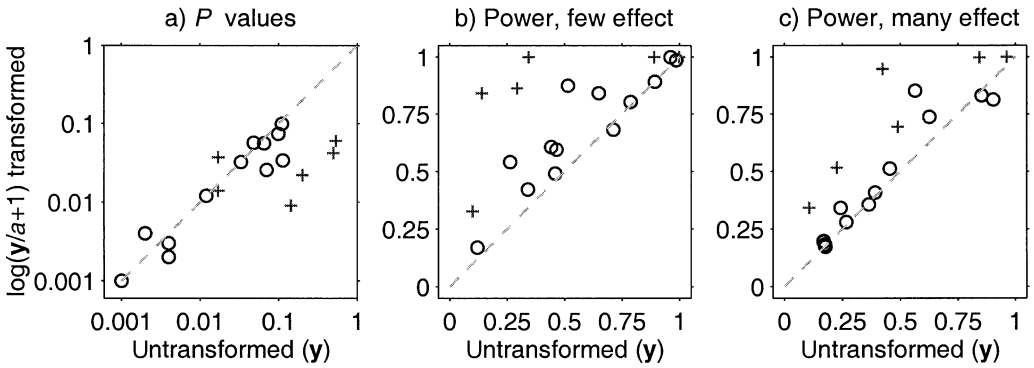

FIG. 6. The effect of transforming data on power. ''Few effect'' and ''many effect'' indicate the relative number of effect variables. Log-transformed vs. untransformed abundances are compared for the LR-IND statistic. A ''+'' indicates very unbalanced data.

DISCUSSION

These results provide no evidence that distance-based approaches are necessary when analyzing multivariate abundances. When data had been transformed to reduce skew, choice of distance had little influence on power, and variable-based statistics such as RDA (redundancy analysis), LR-IND (likelihood ratio test, assuming independence of variables), and ΣF had comparable or greater power than other statistics in most power simulations conducted. This means that, for a given data set, one is at least as likely to detect differences among groups of multivariate abundances when using a variable-based statistic as when using a distance-based statistic.

It is recommended that a MANOVA-based statistic (ΣF or LR-IND) be used when conducting MANOVA tests of multivariate abundances (of the statistics considered here). It has been established in power simulations that these statistics usually have relatively high power, although some other statistics had similar overall power. Advantages of using MANOVA-based statistics were mentioned in the introduction, and these provide a basis for preferring these statistics, given that no other statistic had a distinct power advantage.

There are a number of more peripheral issues arising from these results, which we consider in the following.

Standardization of data

Analyzing unstandardized data using the distance-based statistics considered here will emphasize effects in taxa whose transformed abundances are more variable, and ignore effects in taxa whose transformed abundances are less variable. This explained the few cases with a substantial difference in power between LR-IND and the Pillar-Orloci statistic for log-transformed unstandardized data (Fig. 4), but was more clearly seen in a direct comparison of the RDA statistic calculated on unstandardized vs. standardized data (Fig. 7). The trend was more extreme when there were few effect variables, because in this case it was more likely for all effect variables to have high variances or for all to have low variances. The effect on power was exaggerated in favor of unstandardized data in Fig. 7, because of a failure of the method of standardization, as discussed below.

It is recommended that data be standardized to equal within-group variability, not to equal total variability. In this study, distance-based statistics and RDA were

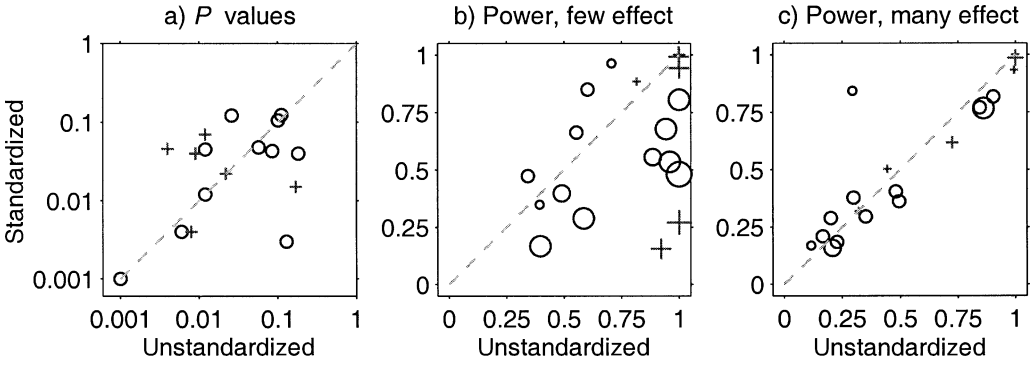

FIG. 7. The effect of standardizing data on power. ''Few effect'' and ''many effect'' indicate the relative number of effect variables. Power is compared for standardized vs. unstandardized log-transformed data, when using the RDA (redundancy analysis) statistic. A ''+'' indicates very unbalanced data. In (b) and (c) the marker size is larger when the average variance of effect variables is larger.

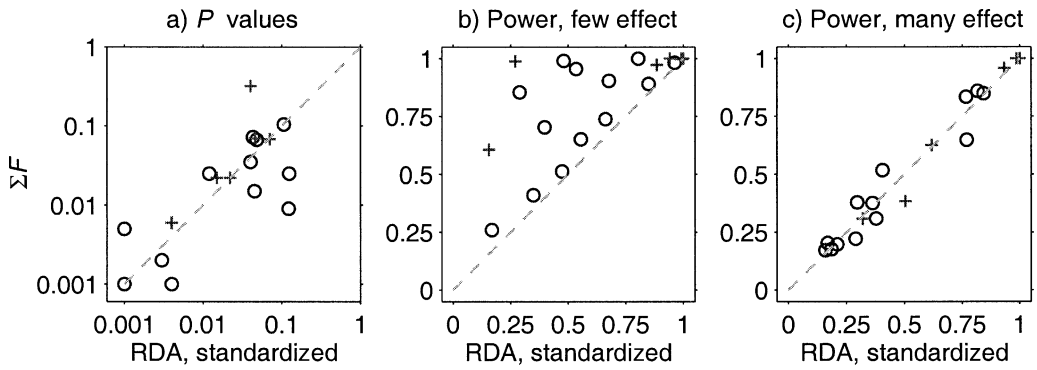

FIG. 8. The effect on power of standardizing by within-group variance rather than by total variance. ''Few effect'' and ''many effect'' indicate the relative number of effect variables. Power is compared for ΣF vs. standardized RDA, for log-transformed data. The RDA statistic was standardized to equal total variance, whereas the ΣF statistic is equivalent to an RDA standardized to equal within-group variance. A ''+'' indicates very unbalanced data.

standardized by a measure of total variability, as previously recommended (Mielke and Berry 2001). However, total variability is a function of among-group variability, so if standardized by total variability, taxa with larger differences among groups are given less relative weighting in analysis, and so power is often lower (Fig. 8). Despite the power advantage of standardizing by a measure of within-group variability, it is not routinely done in practice, and would be difficult to implement for distance-based statistics. This would require recalculation of the distance matrix for each permutation, because within-group variability changes with each permutation. On the other hand, if the variable-based statistics recommended from this study were used, there would be no need to standardize data prior to analysis (nor to choose a distance measure).

Whereas it has previously been stated that often it is desirable to give greater weight to abundant taxa in analyses (Clarke and Green 1988, for example), it should be noted that not standardizing data does not always give higher weight to abundant species. Failing to standardize transformed data can reduce the influence of abundant taxa, rather than the opposite. For example, the taxon with the highest F statistic in Fig. 2 is also the most abundant, despite having one of the lowest variances on the transformed scale. In this case, statistics calculated on unstandardized data did not detect the strong group effect, even though it was expressed in the most abundant variable.

If it is considered desirable to weight abundant species more than rare ones, it is recommended that the weights for each taxon be determined a priori, and that the test statistic be modified to incorporate these. For example given weights w_j, the test statistic could be $\Sigma_{j=1}^{p} w_j F_j$, where F_j is the ANOVA F statistic for the jth variable. This approach ensures the researcher controls the weightings given to each taxon, rather than (possibly incorrectly) assuming that not standardizing data gives the desired weightings.

Correlated abundances in different taxa

Correlation of abundances between taxa is expected to occur in practice, and research is required to find effective ways to account for correlation in data with many variables. None of the test statistics in this study attempted to account for correlation between variables, because of difficulties doing so when there are more variables than samples. However, it is desirable to account for correlation between taxa, just as it is desirable to account for different variances in different taxa. This is because the effect on power of not accounting for correlation is known to be similar to the effect of not standardizing variables. If differences in variance across taxa are not accounted for, there is high power at detecting differences among means in more variable taxa, and low power for differences among means in less variable taxa. Similarly, when correlation is not accounted for, there is high power at detecting differences in means along principal component axes with high variance (i.e., along axes of high correlation), but relatively lower power for other changes in means (Mielke and Berry 2001:53–63). Alternatives to assuming independence are currently under investigation.

When there are many fewer variables than samples, a standard MANOVA statistic such as Wilk's Λ (the statistic from which LR-IND was derived) could be used, possibly with permutation testing to ensure a valid test. A common example of when there are many fewer variables than samples is when sampling invertebrates in pitfall traps (usually many samples), and only sorting them to order (few taxa). Apart from the MANOVA-based statistics, none of the statistics considered in this study can be easily generalized to account for correlated abundances in different taxa.

Transformation

The importance of data transformation was highlighted in the often-large increases in power with transformation (Fig. 3 and 6). The transformations consid-

ered here have two effects relevant to power: to reduce the skew of data, and to reduce differences in variance of different groups and different taxa. The latter of these effects explains the marked increase in power with transformation of MANOVA-based statistics for unbalanced designs (Fig. 6b and c), as ANOVA statistics are known to be sensitive to unequal variances when sample sizes are unequal (Miller 1986). The fact that a similar pattern was observed for distance-based statistics suggests they share this sensitivity to heteroscedasticity in unbalanced designs. The methods least affected by transformation were those that have previously been reported to be robust to strongly skewed data: the Bray-Curtis or Manhattan distance (Gower and Legendre 1986), and the statistic based on ANOVA when excluding very unbalanced designs (Miller 1986).

Previous work

Results comparing distance measures were consistent with previous simulation work. Faith et al. (1987) conducted simulations on untransformed abundances, and also found that the Bray-Curtis distance should be preferred to the Euclidean and Manhattan for analysis of untransformed abundances. They did not, however, consider the case where data had been transformed to reduce skew and the influence of outliers.

This study highlights the importance of using multiple data sets when comparing different methods of analysis. In the past a single data set has often been analyzed several ways, and results published as a comparison of different standardizations (Cao et al. 1999), transformations (Jackson 1993, Thorne et al. 1999), or of data classified to different taxonomic resolutions (Mistri and Rossi 2001). Results were often very different for different data sets in Fig. 4–8, which could lead to very different conclusions concerning which statistic is most powerful. Clearly results from analysis of a single data set have little generality.

Recommendations for related problems

Only the analysis of quantitative abundances has been considered in this paper, but recommendations can be made for the analysis of presence–absence data and semi-quantitative abundances (e.g., the Braun-Blanquet scale, where abundances are scored from 0 to 5). Jongman et al. (1987) suggested treating semi-quantitative abundances as if they were transformed data, and analyzing them on the scale on which they were recorded. This could readily be done using either of the MANOVA-based statistics considered here. Presence–absence data, being binary, are appropriately analyzed using logistic regression models, and indeed have been analyzed in this manner in the univariate case since at least Austin et al. (1984). A multivariate test statistic using logistic regression could be defined as the sum across all taxa of the change in deviance of each taxon. This is the likelihood-ratio statistic for presence–absence data, assuming independence of abundances across taxa, and it shares most of the advantageous properties described here for LR-IND. If data were a mixture of both presence–absence and quantitative abundances, the appropriate likelihood-ratio test statistic would be a sum of logistic regression and LR-IND statistics.

Whereas only MANOVA tests were considered here, the same principles can be applied to analyses relating continuous environmental variables to multivariate abundances. At present the methods most commonly used are distance-based (Clarke and Ainsworth 1993) or CCA (ter Braak and Smilauer 1998), which assumes a linear relationship between each environmental variable and the transformation of abundances described in Appendix A. Using linear regression methods for log-transformed abundance in each taxon and either of the MANOVA-based statistics provides a conceptually simple alternative, which preserves the merits of the MANOVA-based approach previously outlined. Using nonparametric methods of line fitting (Efron and Tibshirani 1991) rather than linear regression would have the additional benefit of not requiring the linearity assumption.

Conclusions

There was no apparent advantage in power if using distance-based statistics rather than a variable-based approach on transformed data, yet there are many reasons to prefer a variable-based approach. There are many alternative variable-based statistics that could be considered, for example those designed for counted data (McCullagh and Nelder 1989:chapter 6), and modifications specifically suggested for abundance data (Welsh et al. 1996). We are presently investigating such methods.

Acknowledgments

Thanks to all who contributed their data to this study. The reference data set obtained from I. Lunt was collected in collaboration with P. Foreman and M. Titcumb, who would like to acknowledge funding from Parks Victoria, the Johnstone Centre, and Charles Sturt University. For advice on the manuscript, thanks to Mark Westoby, Peter Vesk, Jessica Gurevitch, Norm Kenkel, and two anonymous reviewers.

Literature Cited

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

Anderson, T. W. 1984. An introduction to multivariate statistical analysis. Second edition. John Wiley and Sons, New York, New York, USA.

Austin, M. P., R. B. Cunningham, and P. M. Fleming. 1984. New approaches to direct gradient analysis using environmental scalars and statistical curve fitting procedures. Vegetatio **55**:11–27.

Bray, J. R., and J. T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecological Monographs **27**:325–349.

Cao, Y., D. D. Williams, and N. E. Williams. 1999. Data transformation and standardization in the multivariate anal-

ysis of river water quality. Ecological Applications **9**:669–677.
Clarke, K. R. 1993. Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology **18**:117–143.
Clarke, K. R., and M. Ainsworth. 1993. A method of linking multivariate community structure to environmental variables. Marine Ecology Progress Series **92**:205–219.
Clarke, K. R., and R. H. Green. 1988. Statistical design and analysis for a ''biological effects'' study. Marine Ecology Progress Series **46**:213–226.
Clements, A. 1980. The vegetation of bushland in the northern Sydney area. Thesis. Macquarie University, New South Wales, Australia.
Edgington, E. S. 1995. Randomization tests. Third edition. Marcel Dekker, New York, New York, USA.
Efron, B., and R. Tibshirani. 1991. Statistical data analysis in the computer age. Science **253**:390–395.
Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131**:479–491.
Faith, D. P., P. R. Minchin, and L. Belbin. 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio **69**:57–68.
Field, J. G., K. R. Clarke, and R. M. Warwick. 1982. A practical strategy for analysing multispecies distribution patterns. Marine Ecology Progress Series **8**:37–52.
Gee, J. M., R. M. Warwick, M. Schaanning, J. A. Berge, and W. G. Ambrose, Jr. 1985. Effects of organic enrichment on meiofaunal abundance and community structure in sublittoral soft sediments. Journal of Experimental Marine Biology and Ecology **91**:247–262.
Gower, J. C., and W. J. Krzanowski. 1999. Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. Applied Statistics **48**: 505–519.
Gower, J. C., and P. Legendre. 1986. Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification **3**:5–48.
Gray, J. S., K. R. Clarke, R. M. Warwick, and G. Hobbs. 1990. Detection of initial effects of pollution on marine benthos: an example from the Ekofisk and Eldfisk oilfields, North Sea. Marine Ecology Progress Series **66**:285–299.
Jackson, D. A. 1993. Multivariate analysis of benthic invertebrate communities: the implication of choosing particular data standardisations, measures of association, and ordination methods. Hydrobiologia **268**:9–26.
Jongman, R. H. G., C. J. F. ter Braak, and O. F. R. van Tongeren. 1987. Data analysis in community and landscape ecology. Centre for Agricultural Publishing and Documentation, Wageningen, The Netherlands.
Kreutzweiser, D. P., R. C. Back, T. M. Sutton, D. G. Thompson, and T. Scarr. 2002. Community-level disruptions among zooplankton of pond mesocosms treated with a neem (azadirachtin) insecticide. Aquatic Toxicology **56**: 257–273.
Legendre, P., and L. Legendre. 1998. Numerical ecology. Second English edition. Elsevier Science, Amsterdam, The Netherlands.
Manly, B. F. J. 1997. Randomization, bootstrap and Monte Carlo methods in biology. Second edition. Chapman and Hall, London, UK.
Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research **27**:209–220.
MathWorks. 1998. Matlab, version 5. The MathWorks, Natick, Massachusetts, USA.
McCullagh, P., and J. A. Nelder. 1989. Generalized linear models. Second edition. Chapman and Hall, London, UK.
McCulloch, C. E., and S. R. Searle. 2001. Generalized, linear, and mixed models. John Wiley and Sons, New York, New York, USA.
Mielke, P. W., Jr. 1978. Clarification and appropriate inferences for Mantel and Valand's nonparametric multivariate analysis technique. Biometrics **34**:277–282.
Mielke, P. W., Jr., and K. J. Berry. 2001. Permutation methods: a distance function approach. Springer-Verlag, New York, New York, USA.
Mielke, P. W., Jr., K. J. Berry, and E. S. Johnson. 1976. Multiresponse permutation procedures for *a priori* classifications. Communications in Statistics Theory and Methods **5**:1409–1424.
Miller, R. G., Jr. 1986. Beyond ANOVA, basics of applied statistics. John Wiley and Sons, New York, New York, USA.
Mistri, M., and R. Rossi. 2001. Taxonomic sufficiency in lagoonal ecosystems. Journal of the Marine Biological Association of the United Kingdom **81**:339–340.
Morris, L., and M. J. Keough. 2002. Organic pollution and its effects: a short-term transplant experiment to assess the ability of biological endpoints to detect change in a soft sediment environment. Marine Ecology Progress Series **225**:109–121.
Moulton, T. P. 1982. The effect of prescribed burning and simulated burning on soil and litter arthropods in open forest at Cordeaux, N.S.W., Australia. Dissertation. Macquarie University, Australia.
Peres-Neto, P. R., and D. A. Jackson. 2001. How well do multivariate datasets match? The advantages of a Procrustean superimposition approach over the Mantel test. Oecologia **129**:169–178.
Pillar, V. D. P., and L. Orloci. 1996. On randomization testing in vegetation science: multifactor comparisons of relevé groups. Journal of Vegetation Science **7**:585–592.
Romesburg, H. C. 1985. Exploring, confirming, and randomization tests. Computers and Geosciences **11**:19–37.
Smith, E. P. 1998. Randomization methods and the analysis of multivariate ecological data. Environmetrics **9**:37–51.
Smith, E. P., K. W. Pontasch, and J. Cairns, Jr. 1990. Community similarity and the analysis of multispecies environmental data: a unified statistical approach. Water Research **24**:507–514.
Staudte, R. G., and S. J. Sheather. 1990. Robust estimation and testing. John Wiley and Sons, New York, New York, USA.
ter Braak, C. J. F., and P. Smilauer. 1998. CANOCO reference manual and user's guide to CANOCO for Windows: software for canonical community ordination (version 4). Microcomputer Power, New York, New York, USA.
Thorne, R. S., W. P. Williams, and Y. Cao. 1999. The influence of data transformations on biological monitoring studies using macroinvertebrates. Water Research **33**:343–350.
Underwood, A. J. 1997. Experiments in ecology—their logical design and interpretation using analysis of variance. Cambridge University Press, Cambridge, UK.
van den Brink, P. J., and C. J. F. ter Braak. 1998. Multivariate analysis of stress in experimental ecosystems by principal response curves and similarity analysis. Aquatic Ecology **32**:163–178.
van den Brink, P. J., R. P. A. van Wijngaarden, W. G. H. Lucassen, T. C. M. Brock, and P. Leeuwangh. 1996. Effects of the insecticide Dursban 4E (active ingredient chlorpyrifos) in outdoor experimental ditches. II. Invertebrate community responses and recovery. Environmental Toxicology and Chemistry **15**:1143–1153.
van der Aart, P. J. M., and N. Smeenk-Enserink. 1970. Correlations between distribution of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. Netherlands Journal of Zoology **25**:1–45.

van Dobben, H. F., C. J. F. Ter Braak, and G. M. Dirkse. 1999. Undergrowth as a biomonitor for deposition of nitrogen and acidity in pine forest. Forest Ecology and Management **114**:83–95.

von Ende, C. N. 2001. Repeated-measures analysis: growth and other time-dependent measures. Pages 134–157 *in* S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments. Second edition. Oxford University Press, Oxford, UK.

Warwick, R. M., M. R. Carr, K. R. Clarke, J. M. Gee, and R. H. Green. 1988. A mesocosm experiment on the effects of hydrocarbon and copper pollution on a sublittoral soft-sediment meiobenthic community. Marine Ecology Progress Series **46**:181–191.

Warwick, R. M., K. R. Clarke, and J. M. Gee. 1990*a*. The effect of disturbance by soldier crabs, *Mictyris platycheles* H. Milne Edwards, on meiobenthic community structure. Journal of Experimental Marine Biology and Ecology **135**: 19–33.

Warwick, R. M., H. M. Platt, K. R. Clarke, J. Agard, and J. Gobin. 1990*b*. Analysis of macrobenthic and meiobenthic community structure in relation to pollution and disturbance in Hamilton Harbour, Bermuda. Journal of Experimental Marine Biology and Ecology **138**:119–142.

Welsh, A. H., R. B. Cunningham, C. F. Donnelly, and D. B. Lindenmeyer. 1996. Modelling the abundance of rare species: statistical methods for counts with extra zeros. Ecological Modelling **88**:297–308.

APPENDIX A

Calculation of Test Statistics

Distance-based statistics

Table A1 presents distance-based statistics that have been suggested in the multivariate abundance literature or elsewhere. In each case, it was proposed that permutation tests be used for inference.

The Mantel and MRPP (multiresponse permutation procedure) statistics are conceptually equivalent (Mielke 1978), through appropriate choice of distance. Also, T_i is invariant under permutation of samples, so the Smith statistic is equivalent to a Mantel statistic (where all $c_k = 1$), and the AMOVA (analysis of molecular variance), Gower-Krzanowski, Pillar-Orloci, and NP-MANOVA (nonparametric multivariate analysis of varianace) statistics are equivalent. In fact, all statistics except ANOSIM (analysis of similarities) are equivalent to the MRPP statistic, for different choices of c_k and r, the Smith statistic requiring

$$c_k \propto \binom{n_k}{2}$$

and $r = 1$, and the Pillar-Orloci statistic requiring $c_k \propto n_k - 1$ and $r = 2$.

These statistics were generalized to multi-factor designs as suggested in the literature (Mielke et al. 1976, Pillar and Orloci 1996, ter Braak and Smilauer 1998, Gower and Krzanowski 1999, Anderson 2001). For example, consider testing for an effect of factor B (which has b levels) while controlling for the effect of factor A (which has a levels). To find statistics in this case, distance terms in the numerator and denominator are pooled across all a levels; e.g., in the Pillar-Orloci statistic, $W_{k,2}/n_k$ is summed across all ab groups, and T_2/N is calculated within and summed across each of the a levels.

Although the ANOSIM and Smith statistics can be used in multi-factor designs (Smith et al. 1990, Clarke 1993), main effects are not defined, and so interaction terms cannot be tested. This is not necessarily a disadvantage, because for distance-based statistics it is unclear what an interaction means for taxon abundances anyway.

The ANOSIM statistic, as defined here, only differs from the Smith statistic in its use of ranks of distances. The ANOSIM statistic was originally defined as the expression in Table A1 rescaled so that its maximum value is 1, and only for comparing two groups (Clarke 1993). The suggested generalization to the g-group case was to average across all pairwise comparisons of groups. We have used the definition in Table A1, because it has a simpler form, and its relationship to the Smith statistic allows consideration of the effect of using ranks of distances. For balanced data, our approach is equivalent to Clarke (1993).

Variable-based statistics

Table A2 presents variable-based statistics that have been suggested in the multivariate abundance literature or elsewhere. In each case, it was proposed that permutation tests be used for inference.

All the statistics in Table A2 are functions of the residual sum of squares for the jth variable, which as usual is defined as

$$\mathrm{SS}_{j,a} = \sum_{i=1}^{N} (y_{ij} - \hat{\mu}_{ij,a})^2$$

where y_{ij} is the abundance in the ith sample and the jth variable, and $\hat{\mu}_{ij,a}$ is the estimated mean for y_{ij} according to an ANOVA model that assumes H_a. The estimates $\hat{\mu}_{ij,a}$ are least-squares estimates in the sense that they minimize $\mathrm{SS}_{j,a}$. For ANOVA models without interaction terms, the $\hat{\mu}_{ij,a}$ are sample means across all samples in the same group as the ith sample, for the grouping structure defined under H_a.

For a couple of variables, $\mathrm{SS}_{j,1} = 0$, because in each group, abundances of all samples were the same (usually 1 or 0). This caused computational problems for LR-IND (likelihood-ratio test assuming independence of variables) and ΣF, because these respectively involve the logarithm and the inverse of each $\mathrm{SS}_{j,1}$, and both are undefined if $\mathrm{SS}_{j,1} = 0$. This problem was solved for LR-IND by setting zero values of $\mathrm{SS}_{j,1}$ to $N/2\pi e$, because this value ensured that the log likelihood of the jth variable was 0 under H_1. For ΣF, $\mathrm{SS}_{j,1}$ was recalculated for the denominator of the F statistic as if one of the zero abundances were actually equal to the smallest non-zero number.

The RDA (redundancy analysis) and CCA statistics used here are referred to as tests of ''all canonical axes'' by ter Braak and Smilauer (1998:49). The transformation and weightings for CCA are motivated by assuming abundances come from Poisson distributions, which is a poor assumption for strongly skewed abundances, and in particular for transformed abundances.

Note that RDA is equal to Edgington3 (Edgington 1995: 89–90). Because $\mathrm{SS}_{j,0}$ is invariant under permutation, RDA is also equivalent to Ward's E statistic (Romesburg 1985).

As mentioned previously, ΣF is equivalent to an RDA statistic calculated on data standardized to equal within-group variance, where the standardization is recalculated in each permutation during testing. The CANOCO package (ter Braak and Smilauer 1998) offers this choice of standardization, although in CANOCO the standardization is not reapplied for each permutation of data (C. J. F. ter Braak, *personal communication*) and so the test is not valid.

The $\Sigma \log F$ statistic suggested by Edgington (1995) was not considered in power comparisons, because $\log F$ is left skewed, so the statistic is expected to be dominated by occasional small F values and not by the size of the larger ones.

TABLE A1. Distance-based statistics for one-factor MANOVA tests.

Name†	Statistic	Reference
a) From literature outside ecology		
Mantel‡	$\sum_{k=1}^{g} c_k W_{k,1}$	Mantel (1967)
MRPP§	$\sum_{k=1}^{g} c_k \frac{W_{k,r}}{\binom{n_k}{2}}$	Mielke et al. (1976)
AMOVA	$\sum_{k=1}^{g} \frac{W_{k,2}}{n_k}$	Excoffier et al. (1992)
Gower-Krzanowski	$\frac{T_2}{N} - \sum_{k=1}^{g} \frac{W_{k,2}}{n_k}$	Gower and Krzanowski (1999)
b) From ecology literature		
Smith	$\frac{T_1 - \sum_{k=1}^{g} W_{k,1}}{\sum_{k=1}^{g} W_{k,1}}$	Smith et al. (1990)
ANOSIM	$\frac{R(T)_1 - \sum_{k=1}^{g} R(W)_{k,1}}{\frac{(N - n_k)}{2}} - \frac{\sum_{k=1}^{g} R(W)_{k,1}}{\binom{n_k}{2}}$	Clarke (1993)
Pillar-Orloci	$\frac{T_2}{N} - \sum_{k=1}^{g} \frac{W_{k,2}}{n_k}$	Pillar and Orloci (1996)
NPMANOVA	$\frac{(N - g)\left(\frac{T_2}{N} - \sum_{k=1}^{g} \frac{W_{k,2}}{n_k}\right)}{(g - 1) \sum_{k=1}^{g} \frac{W_{k,2}}{n_k}}$	Anderson (2001)

Notes: All statistics are functions of $W_{k,r}$, the sums within group k of distances raised to the power of r, and T_r, the sum across all samples of distances to the power of r. $R(W)$ and $R(T)$ refer to sums of ranks of distances. There are n_k samples in the kth group.

† All acronyms were suggested in their original references: MRPP = multiresponse permutation procedure, AMOVA = analysis of molecular variance, ANOSIM = analysis of similarities, NPMANOVA = nonparametric multivariate analysis of variance.

‡ The Mantel statistic requires a second distance (or similarity) matrix, chosen in this case to take the value 0 between samples of different groups, and c_k between samples in the kth group.

§ For MRPP, values of c_k and r must be chosen to calculate a test statistic. It is recommended in Mielke and Berry (2001) to use $r = 1$ and $c_k \propto n_k$ or $c_k \propto n_k - 1$.

TABLE A2. Variable-based statistics for one-factor MANOVA tests. Statistics from (a) literature outside ecology and (b) the ecology literature.

Name	Statistic	Reference
a) From literature outside ecology		
$\sum \log F$	$\sum_{j=1}^{p} \log\left[\frac{(N-g)(SS_{j,0}-SS_{j,1})}{(g-1)SS_{j,1}}\right]$	Edgington (1995: 188)
$\sum F$	$\sum_{j=1}^{p} \frac{(N-g)(SS_{j,0}-SS_{j,1})}{(g-1)SS_{j,1}}$	Edgington (1995: 188)
Edgington3	$\frac{(N-g)\sum_{j=1}^{p}(SS_{j,0}-SS_{j,1})}{(g-1)\sum_{j=1}^{p}SS_{j,1}}$	Edgington (1995: 189–190)
Ward's E	$\sum_{j=1}^{p} SS_{j,1}$	Romesburg (1985)
b) From ecology literature		
RDA and CCA†	$\frac{(N-g)\sum_{j=1}^{p}(SS_{j,0}-SS_{j,1})}{(g-1)\sum_{j=1}^{p}SS_{j,1}}$	ter Braak and Smilauer (1998: 47)
LR-IND‡	$\sum_{j=1}^{p} N \log\left(\frac{SS_{j,0}}{SS_{j,1}}\right)$	this paper

Notes: These are all functions of $SS_{j,0}$ and $SS_{j,1}$, the residual sum of squares of the jth variable, under H_0 and H_1, respectively (see Eq. 1).

† RDA = redundancy analysis; CCA = canonical correspondence analysis. The test statistic for CCA uses the residual sums of squares from weighted least squares on the transformed variable $y'_{ij} \propto y_{ij}/(y_{i.}\sqrt{y_{.j}})$, with the weight of the ith sample being y_i (where $y_{i.} = \Sigma_{j=1}^{p} y_{ij}$, $y_{.j} = \Sigma_{i=1}^{N} y_{ij}$).

‡ LR-IND = likelihood-ratio test, assuming independence of variables.

APPENDIX B

Tables showing *P* values and power-simulation results for each of the 19 data sets are available in ESA's Electronic Data Archive: *Ecological Archives* E085-023-A1.