Species Associations: The Kendall Coefficient of Concordance Revisited

Pierre LEGENDRE

The search for species associations is one of the classical problems of community ecology. This article proposes to use Kendall's coefficient of concordance (W) to identify groups of significantly associated species in field survey data. An overall test of independence of all species is first carried out. If the null hypothesis is rejected, one looks for groups of correlated species and, within each group, tests the contribution of each species to the overall statistic, using a permutation test. A field survey of oribatid mites in the peat blanket surrounding a bog lake is presented as an example. In the permutation framework, an *a posteriori* test of the contribution of each "judge" (species) to the overall W concordance statistic is possible; this is not the case in the classical testing framework. A simulation study showed that when the number of judges is small, which is the case in most real-life applications of Kendall's test of concordance, the classical χ^2 test is overly conservative, whereas the permutation test has correct Type I error; power of the permutation test is thus also higher. The interpretation and usefulness of the *a posteriori* tests are discussed in the framework of environmental studies. They can help identify groups of concordant species that can be used as indices of the quality of the environment, in particular in cases of pollution or contamination of the environment.

Key Words: *A posteriori* tests; Environmental studies; Kendall W; Oribatid mites; Permutation test; Power; Simulation study; Species associations; Type I error.

1. INTRODUCTION

The search for species associations is one of the classical problems of community ecology. Summarized in a seminal review article by Whittaker (1962), an unfinished debate raged during the 20th century between the tenants of the association-unit theory who believed that ecological communities are natural units that exist independently of human perception and statistical analysis (Pavillard 1912; Whittaker 1956), and those of the individualistic hypothesis who saw species associations as resulting from the coincident relationship of species to environmental forcing variables (Gleason 1926).

Pierre Legendre is Professor, Département de Sciences Biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec, Canada H3C 3J7 (E-mail: Pierre.Legendre@umontreal.ca).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 10, Number 2, Pages 226–245
DOI: 10.1198/108571105X46642

Ecologists are interested in species associations as a conceptual framework to synthesize environmental characteristics. When associations have been found, one can concentrate on finding the ecological requirements common to most or all members of an association instead of having to describe the biology and habitat of each species individually. In an inverse approach, species associations may be used to predict environmental characteristics. Associations may be better predictors of environmental conditions than individual species because they are less subject to sampling error.

Operational concepts of association refer to groups of species that are "significantly" found together, without this implying necessarily any positive biological interaction among them. In other words, an association is simply a group of species (or taxa pertaining to some other systematic category) recognized as a cluster following the application of a clearly stated set of rules. The search for associated groups of species can be based upon presence-absence (e.g., Fager and McGowan 1963; Jackson, Somers, and Harvey 1992) or abundance data. A review of statistical methods for the study of species associations was presented by Legendre and Legendre (1998, sec. 7.5 and 8.9).

Legendre and Legendre (1998, p. 292) described the difficulties involved in computing correlation coefficients based upon community composition (i.e., species abundance) data. The difficulties are due to the highly positively skewed frequency distributions of the species abundances across sites, a phenomenon exacerbated by the high frequency of zeros in many community composition data matrices. Legendre and Gallagher (2001) showed how to transform community composition data in such a way as to make them suitable for statistical analysis through principal component or canonical redundancy analysis which, for untransformed data, preserve Euclidean distances among the objects. The transformations also produce acceptable projections of the correlations among species in principal component space. They will be used as part of the methodology described in this article to search for species associations using quantitative community composition data.

This article proposes to use Kendall's coefficient of concordance (W) to identify significantly associated groups of species in field survey data. Starting with an appropriately transformed species abundance data table, the strategy is the following: (1) conduct an overall test of independence of all species in the study. (2) If the null hypothesis is rejected, look for groups of correlated species. (3) Within each group, test the contribution of each species to the overall statistic, using a permutation test. Results of numerical simulations will also be presented to estimate the Type I error and power of the classical and permutation tests of W, as well as for the new *a posteriori* tests of concordance of individual species.

2. ECOLOGICAL EXAMPLE: ORIBATID MITES

Oribatid mites (or beetle mites; order Acarina, suborder Oribatida) are one of the most abundant groups of arthropods in the world. They are very numerous in humus and moss and play a key role in the recycling of organic matter, which would be much slower without them. Their action has important consequences on the fertility of soils and the productivity of terrestrial areas of the planet. Oribatid mites are very small: .1 to 1.5 mm in length.

In June 1989, my colleague Daniel Borcard obtained 70 soil cores from a small, 10 × 2.6-meter area in the peat blanket surrounding a bog lake, going from the edge of the forest to the open water of the lake (see Figure 5, p. 242). Lac Geai is located on the territory of the Station de Biologie des Laurentides (46°N, 74°W). The mites were extracted, identified to species, and counted. Borcard identified 49 species and estimated their density to about 72,000 individuals per square meter. The spatial and environmental determinants of the mite community structure variation were analyzed in detail by Borcard, Legendre, and Drapeau (1992), Borcard and Legendre (1994), and Borcard, Legendre, Avois-Jacquet, and Tuomisto (2004). The species and environmental data are freely available to researchers on the author's lab Web site at http://www.bio.umontreal.ca/legendre/.

The spatial variation of the 35 species represented by more than a few individuals is re-examined here. The ecological questions are the following: Are the 35 species distributed independently of one another across the study area, or are they significantly and positively associated into recognizable groups of species? Can these species be combined to form indicators of the environmental conditions in the 70 soil cores? The above-mentioned articles had evidenced a major, nonlinear gradient in species composition along the 10-meter dimension of the sampling area (vertical direction in Fig. 5). The environmental conditions (abundance of shrubs, substratum microtopography, substratum density, and humidity) changed in that direction; they are controlled by the amount of water in the peat blanket, which increased from the edge of the forest to the open lake water.

Prior to concordance analysis, the mite abundance data were transformed using the four transformations proposed by Legendre and Gallagher (2001): the chord, chi-square, species profile, and Hellinger transformations were used. Because all results were very similar and led to the exact same species associations, the Hellinger transformation will be used for the results described in this article. The Hellinger transformation consists of two steps: (1) express each abundance value as a proportion with respect to the total sum of animals collected at a site, and (2) take the square root of that proportion. This transformation is such that the Euclidean distance computed among sites for the transformed data is equal to the Hellinger distance (Rao 1995) for the untransformed data. The Hellinger distance is an asymmetric measure of association; it is appropriate for community composition data containing many zeros (Rao 1995; Legendre and Legendre 1998; Legendre and Gallagher 2001). The purpose of the square root, which is the second step of the Hellinger transformation, is to reduce the importance of the very high species abundances. It does not affect the results of the Spearman correlations or the Kendall coefficient of concordance that will be computed in the following, but it does affect the results of principal component analysis. The Hellinger-transformed species abundance data file will be used in all calculations reported in Sections 9 and 10.

3. THE KENDALL COEFFICIENT OF CONCORDANCE (*W*)

Kendall's coefficient of concordance (W) is a measure of the agreement among several (p) judges who are assessing a given set of n objects. Depending on the application field,

the "judges" can be variables, characters, and so on. They are species in the present article. Simulations were carried out (Sections 6 and 7) to empirically compare the classical χ^2 test of the coefficient of concordance to a permutation procedure. When programming the permutation test for W, it was realized that different permutation strategies could be used. One of these strategies allows users to carry out *a posteriori* tests of the contributions of the individual judges to the overall concordance statistic. *A posteriori* tests are possible only in the permutation framework.

There is a close relationship between Friedman's two-way analysis of variance without replication by ranks and Kendall's coefficient of concordance. They address hypotheses concerning the same data table and they use the same χ^2 statistic for testing. They differ only in the formulation of their respective null hypothesis. Consider Table 1, which contains the data of the illustrative example of Section 9. In Friedman's test, the null hypothesis is that there is no real difference among the n objects (sites), which are the rows of the data table. Under H_0, they should have received random ranks from the various judges, so that their sums of ranks should be approximately equal. Kendall's test focuses on the p judges (species). If the null hypothesis of Friedman's test is true, this means that the judges have produced rankings that are independent of one another. This is the null hypothesis of Kendall's test.

- Friedman's H_0: The n objects (sites) are drawn from the same statistical population.
- Kendall's H_0: The p judges (species) produced independent rankings of the objects.

There are two ways found in textbooks for computing Kendall's W statistic (upper and lower forms of Equations (3.1) and (3.2)); they lead to the same result. S or S' is computed first from the row-marginal sums of ranks R_i received by the objects (Siegel 1956: p. 234; Siegel and Castellan 1988, p. 266):

$$S = \sum_{i=1}^{n} \left(R_i - \overline{R}\right)^2$$

or

$$S' = \sum_{i=1}^{n} R_i^2 = \text{SSR}. \tag{3.1}$$

S is a sum-of-squares statistic over the row sums of ranks R_i. $\overline{R}$ is the mean of the R_i values. Following that, Kendall's W statistic can be obtained from either of the following formulas:

$$W = \frac{12S}{p^2(n^3 - n) - pT}$$

or

$$W = \frac{12S' - 3p^2n(n+1)^2}{p^2(n^3 - n) - pT}, \tag{3.2}$$

where n is the number of objects, p the number of judges. T is a correction factor for tied

Table 1. Illustrative example. Upper panel: Hellinger-transformed abundances of four mite species at 10 sites selected along the long axis of Figure 5. The Hellinger transformation was computed for the full dataset (70 sites). Lower panel: the same data transformed into ranks (with ties); last column: sum of the ranks for each site.

	Hellinger-transformed abundances			
	Species 13	*Species 14*	*Species 15*	*Species 23*
Site 4	.25087	.40538	.24380	.08362
Site 9	.40324	.25503	.39303	.00000
Site 14	.26577	.47620	.27267	.06097
Site 22	.32350	.63337	.47003	.00000
Site 31	.26312	.29089	.39223	.08771
Site 34	.33675	.44836	.53727	.10153
Site 45	.07956	.19487	.19487	.11251
Site 53	.00000	.18570	.26261	.11744
Site 61	.00000	.15430	.15430	.00000
Site 69	.12769	.62987	.27584	.34578

	Ranks (species-wise)				*Sum of ranks*
	Species 13	*Species 14*	*Species 15*	*Species 23*	R_i
Site 4	5	6	3	5	19.0
Site 9	10	4	8	2	24.0
Site 14	7	8	5	4	24.0
Site 22	8	10	9	2	29.0
Site 31	6	5	7	6	24.0
Site 34	9	7	10	7	33.0
Site 45	3	3	2	8	16.0
Site 53	1.5	2	4	9	16.5
Site 61	1.5	1	1	2	5.5
Site 69	4	9	6	10	29.0

ranks (Siegel 1956, p. 234; Siegel and Castellan 1988, p. 266; Zar 1999, p. 446):

$$T = \sum_{k=1}^{m} \left(t_k^3 - t_k\right) \tag{3.3}$$

in which t_k is the number of tied ranks in each (k) of m groups of ties. The sum is computed over all groups of ties found in all p columns (judges) of the data table.

Kendall's W statistic is an estimate of the variance of the row sums of ranks R_i divided by the maximum possible value the variance can take; this occurs when all judges are in total agreement; hence $0 \leq W \leq 1$. To derive the formulas for W given above, one has to know that the sum of all ranks in the data table is $pn(n+1)/2$ and that the sum of squares of all ranks is $p^2n(n+1)(2n+1)/6$. Friedman's χ^2 statistic is obtained from W using the formula:

$$\chi^2 = p(n-1)W. \tag{3.4}$$

This quantity is asymptotically distributed like chi-square with $(n-1)$ degrees of freedom. This allows us to test W for statistical significance. When $n \leq 7$ and $p \leq 20$, Siegel and Castellan (1988, p. 270, 365) recommended using their table of critical values for W, which was obtained by the method of complete permutations.

There is a close relationship between Spearman's correlation coefficient r_S and Kendall's W statistic, which will prove useful in Section 5: W can be calculated directly from the mean $(\overline{r})$ of the pairwise Spearman correlations r_S using the following relationship (Siegel and Castellan 1988, p. 262; Zar 1999, p. 448):

$$W = \frac{(p-1)\overline{r}+1}{p}, \tag{3.5}$$

where p is the number of variables (or judges) among which Spearman's correlation coefficients are computed. For two variables (or judges) only, W is simply a linear transformation of r_S : $W = (r_S + 1)/2$. In that case, a permutation test of W for two variables is the exact equivalent of a permutation test of r_S for the same variables.

The relationship described by Equation (3.5) clearly limits the domain of application of the coefficient of concordance to data that are all meant to estimate the same general property of the objects: judges are only considered concordant if their Spearman correlations are positive. Two judges that give perfectly opposite ranks to a set of objects have a Spearman correlation of –1, hence $W = 0$ for these two judges; this is the lower bound of the coefficient of concordance. For two judges only, $r_S = 0$ gives $W = .5$. So the coefficient W applies well to rankings given by a panel of judges called in to assess overall performance in sports, or quality of wines or restaurants, or to rankings obtained from criteria used in quality tests of appliances or services by consumer organizations, and so on. It does not apply, however, to ordinary variables used in multivariate analysis. Zar (1999), for example, uses wing length, tail length and bill length of birds to illustrate the use of the coefficient of concordance. These data are appropriate for W because they are all indirect measures of a common property, the size of the birds. One should not conclude from that example that W should be used in morphometric or numerical taxonomic studies: in those fields, a negative and a positive correlation have equal importance and should play the same role in the analysis; this is not the case with W.

In ecological applications, one can use the abundances of various species as indicators of the good or bad environmental quality of the sampling sites. One should be careful that the study only includes sites that belong to the same type of environment, since different species often characterize different types of environment. If a group of species are used to produce a global index of the overall quality (good or bad) of the environment at a series of sites, only the species that are significantly associated and positively correlated to one another should be included in the index, since different groups of species may be associated to different environmental conditions. The example in Section 10 will show that the *a posteriori* tests of concordance, described in Section 5, can help identify the groups of species that are positively associated to one another along a dominant environmental gradient. Only the positively associated species can be used to construct indicator functions of the quality of the environment.

4. PERMUTATION TEST OF *W*

Does the classical chi-square test of significance of W have correct Type I error and good

power? Numerical simulations were used to answer the question empirically (Sections 6 and 7), comparing the classical chi-square test to a permutation test. The overall permutation test of W will allow us, in turn, to test the contribution of individual judges to the W statistic (Section 5); this type of *a posteriori* test is not available in the classical testing framework.

When testing hypotheses using Kendall's W statistics, the objects are the permutable units under H_0 (H_0 is stated in Section 3; the objects are sites in Table 1). For the global test of significance, the rank values in all judges are permuted at random, independently from judge to judge. The null hypothesis of this test is the independence of the rankings produced by all judges. The alternative hypothesis is that at least one of the judges is concordant with one, or with some of the other judges. The test is one-tailed because it only recognizes positive associations between vectors of ranks. This can be shown by considering two vectors with exactly opposite rankings: they produce a Spearman statistic of –1, hence a value or zero for W (Equation (3.5)). The testing procedure is the following:

1. Transform quantitative or semiquantitative data into ranks if necessary. Assign mean values to tied ranks.
2. Compute Kendall's W coefficient of concordance among the ranked vectors (Equation (3.2)). Transform W into Friedman's χ^2 statistic, which is a pivotal statistic appropriate for testing. This provides the reference statistic (χ^2_{ref}) for the test. Actually, within a given permutation test, the three statistics W, χ^2, and SSR, are monotonic to one another since n, p, as well as T, are constant within a given permutation test; thus they are equivalent statistics for testing, producing the same permutational probability.
3. Permute all vectors of ranked data at random, independently of one another. Compute a χ^{2*} (or W^*, or SSR*) value of the statistic under permutation.
4. Repeat Step 3 a large number of times to obtain an estimate of the distribution of the χ^2 (or W, or SSR) statistic under permutation. Add the reference value χ^2_{ref} (or W_{ref}, or SSR$_{\text{ref}}$) to the distribution (Hope 1968).
5. Calculate the one-tailed probability (P value) of the data under the null hypothesis as the proportion of values of χ^{2*} (or W^*, or SSR*) in the distribution that are larger than or equal to χ^2_{ref} (or W_{ref}, or SSR$_{\text{ref}}$). The test indicates that the set contains concordant judges if the P value is equal to or smaller than the preselected significance level (say, $\alpha = .05$).

If the null hypothesis (independence of all judges) is true, the reference value of the chi-square statistic, χ^2_{ref}, should not be distinguishable from the distribution of values χ^{2*} obtained under permutation. If the null hypothesis is false, one expects the reference value to be larger than most values obtained under permutation.

5. *A POSTERIORI* TESTS

If the overall null hypothesis is rejected, *a posteriori* tests can be computed, in the permutation framework, to determine which of the individual judges are concordant with

Table 2. Results of (a) the overall and (b) the *a posteriori* tests of concordance among the mite species. $P =$ permutational probability, based upon 9,999 random permutations. (c) Complementary Spearman correlation coefficients (r) with results of one-tailed tests of significance, and partial concordance statistics $\bar{r}_j$ and W_j for each species j described in Section 5. * Reject H_0 at $\alpha = .05$.

(a) Overall test of the W statistic. H_0: The four species are not concordant with one another			
Kendall's $W =$	.44160		
Friedman's chi-square $=$	15.89771	$P = .0448$*	Reject H_0

(b) A posteriori *tests*	*H_0: This species is not concordant with the other three*	
Species 13	$P = .0766$	Do not reject H_0
Species 14	$P = .0240$*	Reject H_0
Species 15	$P = .0051$*	Reject H_0
Species 23	$P = .7070$	Do not reject H_0

(c) Spearman correlation table		*H_0: $r = 0$; H_1: $r > 0$ (one-tailed test)*					
		Species 13	*Species 14*	*Species 15*	*Species 23*	$\bar{r}_j$	W_j
Species 13	*r*	1.0000	.5593	.8389	−.4185	.32657	.49493
	P	—	.0464	.0012	.8856		
Species 14	*r*	.5593	1.0000	.6242	.0061	.39655	.54741
	P	.0464	—	.0269	.4933		
Species 15	*r*	.8389	.6242	1.0000	–.0920	.45704	.59278
	P	.0012	.0269	—	.5998		
Species 23	*r*	–.4185	.0061	–.0920	1.0000	–.16813	.12391
	P	.8856	.4933	.5998	—		

one or several of the other judges. There is interest in several fields for identifying discordant variables or judges. This includes all fields that use panels of judges to assess the overall quality of the objects under study (sports, law, consumer protection, etc.). In other types of studies, scientists are interested to identify variables that agree in their estimation of a common property of the objects. This is the case in environmental studies where ecologists are interested in identifying groups of concordant species that are indicators of some property of the environment and can be combined into indices of its quality, in particular in cases of pollution or contamination; see the example in Section 10.

The contribution of individual judges to the W statistic can easily be assessed by a modified form of permutation test. The null hypothesis is the monotonic independence of the judge subjected to the test, with respect to all the other judges in the study. The alternative hypothesis is that this judge is concordant with other judges in the set under study, having similar rankings of values (one-tailed test). The statistic W can be used directly in *a posteriori* tests (see also next paragraph). Step 3 of the testing procedure, described in the previous section, is modified: only the judge under test is permuted. If the judge under test has values that are monotonically independent of the other judges, permuting its values at random should have little influence on the overall W^* statistic. If, on the contrary, it is concordant with one or several other judges, permuting its values at random should break the concordance and have a noticeable influence on W^*. The importance of the concordance between this judge and all the others is assessed in Steps 4 and 5 of the testing procedure.

Two specific partial concordance statistics could be used in *a posteriori* tests. The first

Table 3. Results of the tests of concordance involving mite group 1. The 24 species were ordered by the values of the partial concordance statistics, $\overline{r}_j$ and W_j, to facilitate interpretation. P = permutational probability based upon 9,999 random permutations. P_H = probability after Holm adjustment, computed using the 35 P values from Tables 3 and 4. * Reject H_0 at $\alpha = .05$.

(a) Overall test of the W statistic. H_0: The 24 species are not concordant with one another

Kendall's W = .30979 *P = .0001** *Reject H_0*

(b) A posteriori *tests*
H_0: This species is not concordant with most of the others

	$\overline{r}_j$	W_j	P	P_H
Species 2	.42581	.44974	.0001	.0035*
Species 27	.42176	.44585	.0001	.0035*
Species 17	.42122	.44533	.0001	.0035*
Species 20	.41807	.44232	.0001	.0035*
Species 13	.41773	.44200	.0001	.0035*
Species 11	.41385	.43827	.0001	.0035*
Species 21	.36234	.38891	.0001	.0035*
Species 4	.35906	.38576	.0001	.0035*
Species 14	.33012	.35803	.0001	.0035*
Species 26	.30162	.33072	.0001	.0035*
Species 7	.28333	.31319	.0001	.0035*
Species 28	.25770	.28863	.0001	.0035*
Species 19	.25748	.28842	.0002	.0035*
Species 5	.25055	.28178	.0002	.0035*
Species 10	.24447	.27595	.0001	.0035*
Species 30	.23014	.26222	.0001	.0035*
Species 24	.21882	.25137	.0012	.0156*
Species 15	.21854	.25110	.0002	.0035*
Species 1	.18512	.21907	.0025	.0300*
Species 6	.18022	.21437	.0045	.0360*
Species 12	.12638	.16278	.0325	.1938
Species 22	.12502	.16148	.0701	.1938
Species 29	.11080	.14785	.0435	.1938
Species 8	.09248	.13029	.0851	.1938

one is the mean, $\overline{r}_j$, of the pairwise Spearman correlations between judge j under test and all the other judges. The second statistic, W_j, is obtained by applying Equation (3.5) to $\overline{r}_j$ instead of $\overline{r}$, with p the total number of judges. These two statistics are shown in Tables 2–4 for the example data. $\overline{r}_j$ and W_j are clearly monotonic to each other since p is constant in a given permutation test. Within a given *a posteriori* test, W is also monotonic to W_j because, in the procedure described in the previous paragraph, only the values related to judge j are permuted when testing judge j. These three statistics are thus equivalent for *a posteriori* permutation tests, producing the same permutational probabilities. Like $\overline{r}_j$, W_j can take negative values; this was not the case of W.

There are advantages in performing a single *a posteriori* test for judge j, instead of $(p-1)$ tests of the Spearman correlation coefficients between judge j and all the other judges: the tests of the $(p-1)$ correlation coefficients would have to be corrected for multiple testing, and they could provide discordant information; a single test of the contribution of judge j to the W statistic has greater power and provides a single, clearer answer.

Table 4. Results of the tests of concordance involving mite group 2. The 11 species were ordered by the values of the partial concordance statistics, $\bar{r}_j$ and W_j, to facilitate interpretation. P = permutational probability based upon 9,999 random permutations. P_H = probability after Holm adjustment, computed using the 35 P values from Tables 3 and 4. * Reject H_0 at $\alpha = .05$.

(a) Overall test of the W statistic. H_0: The 11 species are not concordant with one another

Kendall's $W = .29119$ $P = .0001$ Reject H_0

(b) A posteriori *tests*
H_0: This species is not concordant with most of the others

	$\bar{r}_j$	W_j	P	P_H
Species 31	.34466	.40423	.0001	.0035*
Species 25	.33423	.39476	.0001	.0035*
Species 33	.31889	.38081	.0001	.0035*
Species 9	.27121	.33746	.0001	.0035*
Species 35	.24989	.31808	.0001	.0035*
Species 16	.19064	.26422	.0031	.0310*
Species 32	.18331	.25755	.0025	.0300*
Species 34	.17642	.25129	.0033	.0310*
Species 18	.13756	.21596	.0253	.1771
Species 23	.13424	.21295	.0323	.1938
Species 3	.12226	.20205	.0390	.1938

In order to preserve a correct or approximately correct experimentwise error rate, the probabilities of the *a posteriori* tests should be adjusted for multiple testing. Wright (1992) recommended the Holm (1979) procedure for sets of nonindependent tests such as we have here. This procedure is less conservative than an ordinary Bonferroni adjustment.

A posteriori tests are useful to identify the judges that are not concordant with the others, as will be seen in the examples, but they do not tell us if there are one or several groups of congruent judges among those for which the null hypothesis of independence is rejected. This information can be obtained by computing Spearman correlations among the judges and clustering the judges into groups of variables that are significantly and positively correlated. Because the alternative hypothesis of Kendall's test of concordance is one-tailed, one-tailed tests should also be used for the Spearman statistics.

6. SIMULATION METHOD

Simulations have been performed to compare the classical χ^2 test and the permutation test of concordance in terms of Type I error and power. Type I error concerns rejecting the null hypothesis of the test when the data conform to this hypothesis. To be valid, a test of significance should have a rate of rejection of the null hypothesis no larger than the nominal (α) significance level of the test (Edgington 1995, p. 37) when the null hypothesis (H_0) is true. On the other hand, a test of significance should be able to reject the null hypothesis when H_0 is false; the frequency of rejection of H_0 in these circumstances is referred to as the power of the test.

The simulations involved two types of judges: a number p_{IJ} of independently generated

judges (IJ), and a number p_{PJ} of partly similar judges (PJ). Independently generated judges were created by generating vectors of random standard normal deviates $N(0, 1)$. A group of partly similar judges was created by first generating a vector of random standard normal deviates $N(0, 1)$; for each judge of the group, random normal deviates with a preselected standard deviation were added to the values of this common vector to make the judges partly different. The standard deviation values used in the simulations were $\sigma = \{.5, 1.0, 2.0\}$. The values for each judge were transformed into ranks.

In the simulations to estimate Type I error, only independent judges (IJ) were generated. The simulations to estimate power, in which the alternative hypothesis of the test must be true (at least some of the judges must be concordant), involved various combinations of IJ and PJ judges. There were $n = \{5, 10, 20, 50, 100\}$ objects and $p_{IJ} = \{2, 3, 4, 5, 10, 20, 25, 30\}$ judges in the simulations for Type I error. There were $n = 20$ objects and $p = \{5, 10\}$ judges in the simulations for power, which involved all combinations of the two types of judges ($p_{IJ} + p_{PJ} = p$).

For each result, 10,000 replicate simulations were run; 999 random permutations were used for the permutation tests in each simulation. The rate of rejection of the null hypothesis was calculated, together with its 95% confidence interval. A simulation result consists of: the rate of rejection of the null hypothesis by the classical χ^2 test at the $\alpha = .05$ significance level together with its 95% confidence interval, and the same information for the permutation test as well as for the *a posteriori* tests of significance.

Additional simulations were carried out to determine if the method was able to identify several groups of correlated judges (e.g., species). A correlation matrix, containing positive correlations among the group members and null or negative correlations between groups, was read into the simulation program and subjected to Cholesky factorization, $\mathbf{R} = \mathbf{L}'\mathbf{L}$, where $\mathbf{L}$ is an upper triangular matrix. Vectors containing random standard normal deviates were generated and written to a work matrix $\mathbf{X}$. Matrix $\mathbf{W}$ containing the correlated vectors (judges) was obtained by computing $\mathbf{W} = \mathbf{XL}$. Justification of this procedure was given by Legendre (2000, sec. 5.1).

7. SIMULATION RESULTS

7.1 OVERALL TEST OF *W*

Figure 1 presents the empirical Type I error rates at $\alpha = .05$ for the classical χ^2 test and the permutation test of W. The classical χ^2 test is overly conservative, although it remains valid, having rejection rates well below the significance level ($\alpha = .05$ in these results). For 20 judges and more, the 95% confidence intervals of the rejection rates of the classical χ^2 test almost always included the significance value, $\alpha = .05$. The permutation test always had a correct rate of Type I error. Siegel and Castellan (1988, pp. 270, 365) recommended the use of a permutation-based table of critical values for W only when $n \leq 7$ and $p \leq 20$. For $n > 7$, they recommend using the χ^2 approximation. The simulation results presented here show that the classical χ^2 test remains too conservative for any sample size (n), when the number of judges p is smaller than 20.

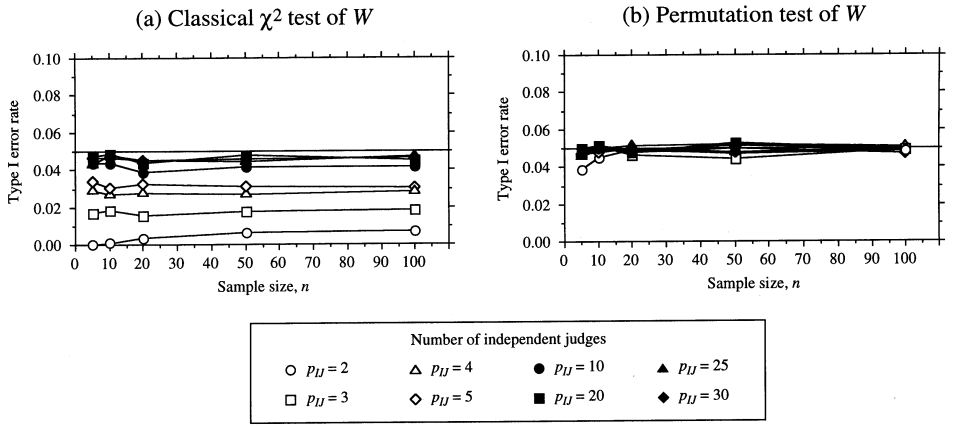

Figure 1. Type I error rate of (a) the classical χ^2 test and (b) the permutation test of W for different numbers of objects (sample size n) and independent judges (p_{IJ}, symbols). The $\alpha = .05$ significance level is materialized by a horizontal line. Each point (rejection rate) is the result of 10,000 simulations.

Simulation results also show that the power of the permutation test is higher than that of the classical χ^2 test (Figure 2). The differences in power are due to the differences in rates of Type I error between the two forms of test (Figure 1). The differences in power disappear asymptotically as the number of judges increases.

7.2 *A POSTERIORI* COMPARISONS

Simulations were also done to assess the *a posteriori* comparisons obtained by permuting a single judge at a time. The *a posteriori* tests on individual independent judges

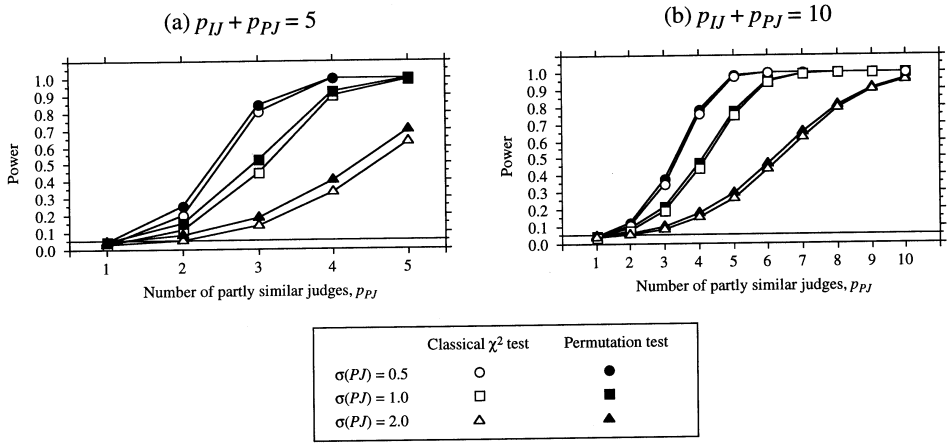

Figure 2. Power study: rejection rate of the classical χ^2 test (empty symbols) and the permutation test (filled symbols) of W, at $\alpha = .05$, for various numbers of partly similar judges (PJ). $\sigma(PJ)$ is the standard deviation of the random components of the partly similar judges. (a) 5 judges, (b) 10 judges in total; $n = 20$. For the leftmost point of each curve ($p_{PJ} = 1$), the null hypothesis of independence of the judges was true.

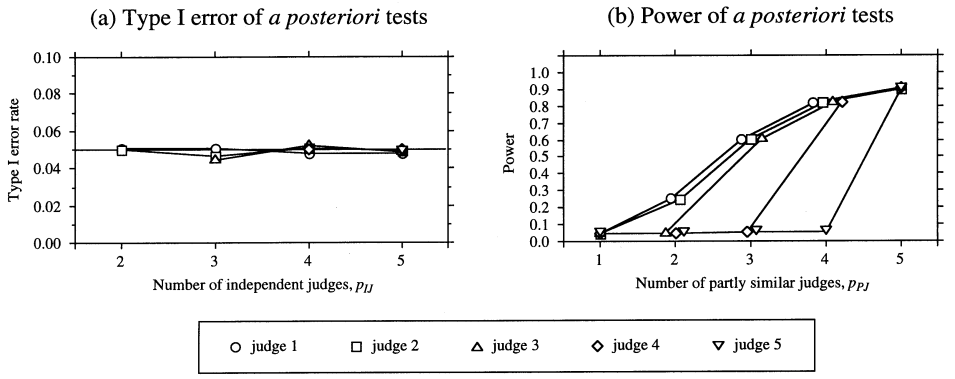

Figure 3. (a) Type I error rate of a posteriori *tests for n = 20 objects and p_{IJ} = 2-5 independent judges. The $\alpha = .05$ significance level is materialized by a horizontal line. The* a posteriori *test of judge k can only be done when the number of judges is larger than or equal to k. (b) Power of* a posteriori *tests for simulations with n = 20 objects and p = 5 judges. The standard deviation of the normal error of the partly similar judges (PJ) in the results shown was the intermediate value, $\sigma(PJ) = 1$. When $p_{PJ} = 2$, for example, it is the first two judges that were generated to be similar; likewise for $p_{PJ} = 3$, 4, or 5. All judges are independent when $p_{PJ} = 1$. Open circles represent the rejection rates of the null hypothesis in tests involving judge 1 only; similarly for the other symbols (judges). Some symbols have been moved sideways to improve the clarity of the graph.*

had, individually, correct rates of Type I error for all combinations of number of objects ($n = 5, 10, 20, 50$, and 100) and number of independent judges (p_{IJ}) in the study (Figure 3(a) shows results for $n = 20$ objects and 2 to 5 independent judges).

When H_0 was false by construct for the judge under test, the *a posteriori* test had a rejection rate higher than the significance level (Figure 3(b)), but when one of the independent judges was permuted, the rejection rate was at or near the α significance level, as it should. Power of the *a posteriori* tests also increased with n when H_0 was false (not shown). So, when the null hypothesis of the overall test is rejected and there is a single group of partly similar judges in the dataset, it should be possible to identify the concordant judges. This property will be illustrated in the real-case applications of Sections 9 and 10.

7.3 ADDITIONAL SIMULATIONS: TWO OR THREE GROUPS OF JUDGES

Simulations were also produced for 2 and 3 groups of correlated judges. For 2 groups of 10 correlated judges (within-group correlations $= \{.25, .50, .75\}$, between-group correlations $= 0$, for $n = 50$ and a total of 25 judges), the tests had very high power (rejection rate of $H_0 = 1.0$ for the global test and near 1.0 for the *a posteriori* tests). The *a posteriori* tests did not allow, however, distinguishing the members of the two groups. When negative correlations of –.3 were introduced between members of different groups of judges, the power of the overall test remained high (rejection of H_0 at or near 1.0) as long as the positive within-group correlations were higher than the negative among-group correlations. The power of the *a posteriori* tests decreased faster than that of the global test. Type 1 error of the *a posteriori* tests remained correct for the independent judges (IJ).

For three groups (10, 5, and 5 correlated judges, among 25 judges, $n = 50$) and correlations of 0 among the groups, power of the global test was very high (rejection rate of $H_0 = 1.0$). Power of the *a posteriori* tests was higher for the members of the largest group of 10 judges. With correlations of .5 within the groups and –.3 among the groups, the *a posteriori* tests were not able to detect the members of the two smallest groups of correlated judges, although they did identify correctly the members of the 10-judge group.

For three very small groups of correlated judges (3, 3, and 2 correlated judges, among 10 judges, $n = 50$) and correlations of 0 among the groups, power of the global test was slightly lower (rejection rate of $H_0 = .97$) than with larger groups of correlated judges. Power of the *a posteriori* tests was higher (rejection rate of H_0 about .80) for members of the two largest groups of three judges than for the group of two judges (rejection rate of H_0 about .32). When negative correlations of –.3 were imposed among the groups, even the global test failed to reject H_0.

8. PROCEDURE FOR THE IDENTIFICATION OF ASSOCIATED SPECIES

The simulation results reported in Section 7 indicate that an appropriate procedure for the identification of associated species requires a division of the dataset into groups of potentially concordant species, on the basis of their correlations. Starting with an appropriately transformed species abundance data table, the strategy is the following:

1. First, conduct an overall test of concordance using all species.
2. If that test is significant, look for groups of correlated species. This can be done in several ways. (a) A principal component analysis based upon standardized species vectors, with eigenvectors normalized to the square root of the respective eigenvalues, will produce a plot of the species with angles representing the correlations among species. Groups of species may be identifiable on that plot. (b) Use the full table of eigenvectors in a clustering or K-means partitioning of the species. (c) Standardize the species data and apply K-means partitioning to find groups of species. (d) Or, compute a Pearson or Spearman correlation matrix among the species and use agglomerative clustering to find groups of correlated species. Methods (b)–(d), which involve clustering, can help delineate several groups of species, whereas the PCA used in (a) is only adequate when two groups are present. These strategies will all be used in Section 10.
3. Submit each group of species to a separate analysis of concordance. The *a posteriori* tests will identify the species that are significantly associated.

9. ORIBATID MITES: ANALYSIS OF A SUBSET

A small subset of the oribatid mite data will be used to illustrate the calculations. Ten sites were selected along the long axis of the sampling area (rectangular area shown in

Figure 5, p. 242), one in every one-meter section. The site coordinates are given in a file included in the mite data package at http://www.bio.umontreal.ca/legendre/. Four species were selected: three from the largest group of associated species identified in section 10 (species #13, 14, and 15) plus one of the nonassociated species (#23). Table 1 presents the Hellinger-transformed species abundances as well as the sites ranked according to each species. The sum-of-squares statistic (S or S') was computed from the sums of ranks (column 10 of the Table) using Equation (3.1); that value was used in Equation (3.2) to compute Kendall's coefficient of concordance W.

A test of concordance for only 10 objects (sites) has little power. Nevertheless, the overall test involving the four mite species is significant at the $\alpha = .05$ level (Table 2(a)), and so are the *a posteriori* tests for species 14 and 15 (Table 2(b)). Species 13 does not reach the .05 significance level here, although it will be significant in the analysis of the full dataset (Section 10). Table 2(c) shows how the $\overline{r}_j$ statistic is computed as the mean of the Spearman correlations of species j with all the other species. W_j is obtained by applying Equation (3.5) to $\overline{r}_j$ instead of $\overline{r}$.

10. ORIBATID MITES: RESULTS

Concordance analysis involving all 35 mite species (data described in Section 2) indicated that one or some of the species were concordant with one or some of the other species (Kendall's $W = .06886$, $P = .0001$ after 9,999 permutations). This is not a very interesting conclusion, however, if one wants to use the species as indicators of the major environmental and spatial trend. Siegel (1956) suggested that the variables that produce a highly significant coefficient of concordance could be pooled into an overall index. Applying this suggestion, the 35 species abundances were summed. This new variable bore little relationship to the environmental variables: a multiple regression against the three environmental variables mentioned above, plus water content of the substratum, and a set of dummy variables representing seven types of substrate, led to a parsimonious model with $R^2 = .128$ (adjusted $R^2 = .088$) which included only three dummy variables representing three of the seven classes of substratum. A third-degree polynomial trend surface model of this index was not strongly related to the Y geographic direction of the map: an $R^2 = .092$ was obtained.

Let us explore whether groups of concordant species can produce better environmental and spatial models. The four procedures described in Section 8 were used to identify groups of positively correlated species:

(a) A principal component analysis (PCA) was computed after standardizing the Hellinger-transformed species vectors to means of 0 and variances of 1. The eigenvectors were normalized to the square roots of their respective eigenvalues, so that angles between the species vectors in the PCA plot are projections of their correlations. Two groups of species are easily identified in Figure 4: 24 species on the left

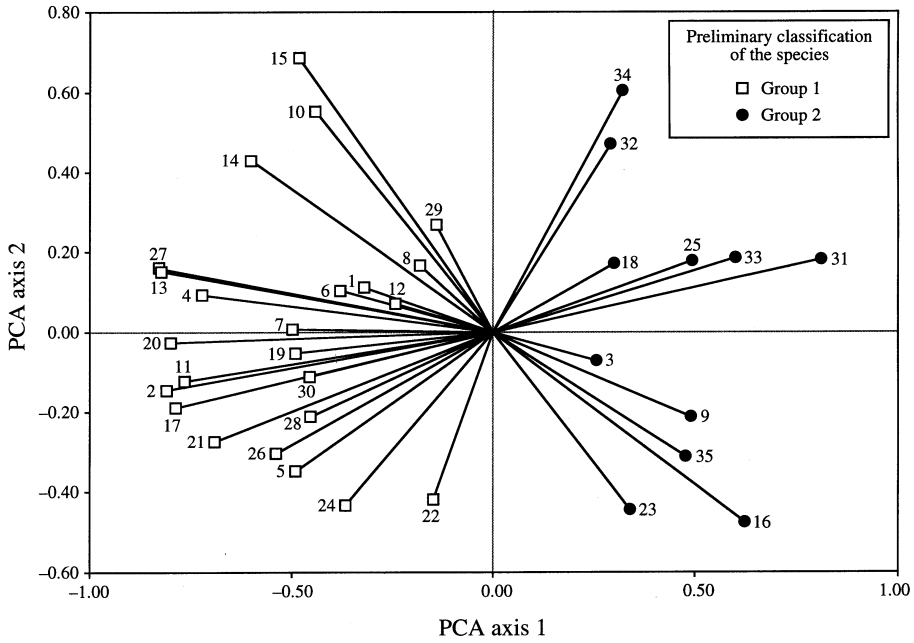

Figure 4. Principal component plot: the mite species vectors projected in the space of PCA axes 1 and 2. Total variance accounted for: 38% (axis 1: 28.8%, axis 2: 9.2%).

(group 1) and 11 species on the right (group 2).

(b) The full table of eigenvectors from the PCA was used in a K-means partitioning procedure, which produced partitions into two to ten groups. The Calinski and Harabasz (1974) criterion (C-H) was computed in order to decide which partition was the best. C-H is simply the F-statistic of multivariate analysis of variance and canonical analysis; it is the ratio of the mean square for the given partition divided by the mean square for the residuals. The partition for which C-H is maximum is the best one in the least-squares sense. A simulation study conducted by Milligan and Cooper (1985) showed that, among 30 such criteria, C-H was the best one to recover the correct number of clusters in multivariate datasets. For random data not structured into distinct groups, C-H decreases smoothly as the number of groups K increases, so that C-H is maximum for $K = 2$ groups (no C-H value can be computed for a single group). With the mite data, C-H was maximum for two groups, the same two groups as shown in Figure 4. Because the value of C-H was much larger for two than for three or more groups, the criterion was interpreted to indicate, in that case, the presence of two recognizable groups of species.

(c) The Hellinger-transformed species data were standardized to means of 0 and variances of 1. K-means partitioning of the species was applied directly to that table (in paragraph b, the partitioning procedure was applied to the PCA eigenvectors). The C-H criterion identified again the partition in two groups as the best one in the least-squares sense; the two groups were the same as in (a) and (b).

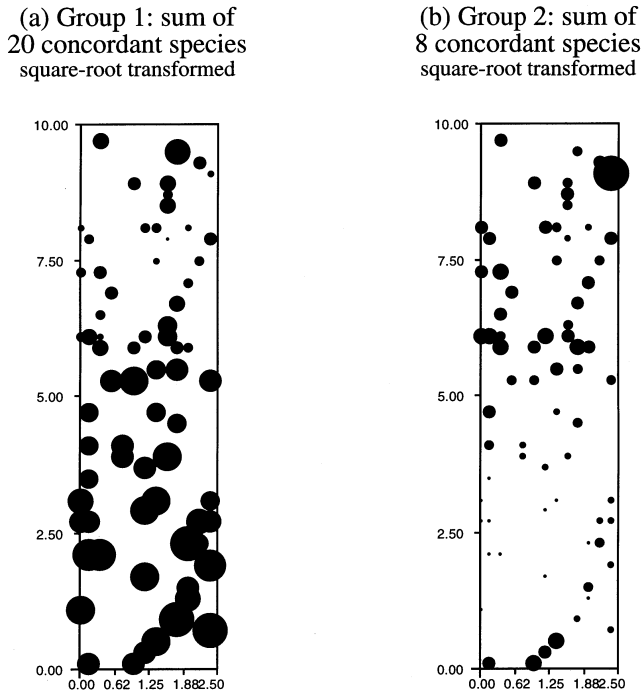

Figure 5. Bubble plot maps of the square-root-transformed sums (a) of the 20 concordant species belonging to group 1, (b) of the 8 concordant species pertaining to group 2. Bubble size is proportional to the value. The coordinates of the 70 soil cores are in meters. The edge of the forest is at the bottom of the figures, the open water at the top.

(d) A Spearman correlation matrix was computed among the Hellinger-transformed species vectors. The Spearman correlations were interpreted as similarity indices among the species and used in Ward's agglomerative clustering, which produced the same two groups as the previous methods did.

For the 24 species belonging to group 1, the *a posteriori* tests (Table 3(b)) indicated that 20 species are significantly concordant with most of the other species in that group. For group 2, the *a posteriori* tests (Table 4(b)) showed that 8 of the 11 species are significantly concordant. Seven mite species remain unattached to the two groups of concordant species. The global tests reported in Tables 3(a) and 4(a) were not necessary since the global test of Table 2(a) was significant.

The sum of abundances of the 20 concordant species from group 1 (Figure 5(a)) was used as the dependent variable in multiple regression modeling. This index produced a highly significant environmental model ($R^2 = .731$) featuring substratum microtopography (positive regression coefficient), density (positive coefficient), and water content (negative coefficient), and a positive association with two types of mosses, as well as a highly significant polynomial trend-surface spatial model ($R^2 = .664$). Likewise, the sum of abundances of the 8 concordant species from group 2 (Figure 5(b)) was used as the dependent variable

in multiple regression. This variable produced a highly significant environmental model ($R^2 = .304$) featuring a positive regression coefficient for water content and a negative coefficient, indicating avoidance, for the dummy variable coding for the bare peat substrate, as well as a highly significant polynomial trend-surface model ($R^2 = .375$) depicting a spatial structure different from that of the species from group 1. Detailed relationships would be revealed by analyzing, by canonical redundancy analysis, each group of concordant species with respect to the environmental variables and drawing biplots of the species and the environmental variables.

11. DISCUSSION

11.1 STATISTICAL ASPECTS

The Kendall coefficient of concordance can be used to assess the degree to which a group of species, or other types of judges, provide a common ranking of a set of sites or other types of objects. It should only be used to obtain a statement about variables that are all meant to measure a single general property of the objects, the same one for all the judges included in the analysis. It should not be used to analyze sets of variables in which the negative and positive correlations have equal importance for the interpretation.

The partial concordance coefficients and *a posteriori* tests of significance are essential complements of the overall test of concordance. In several fields, there is interest in identifying discordant variables or judges; this is the case in all fields that use panels of judges to assess the overall quality of the objects under study (sports, law, consumer protection, etc.). In other applications, one is interested in using the sum of ranks, or the sum of values, provided by several variables or judges, as an overall indicator of the response of the objects under study. The *a posteriori* tests, combined with the (rank) correlation coefficients and clustering results, allow scientists to determine which variables significantly belong to the same group, before summing their ranks or values into an overall index.

The simulation results summarized in Figure 1 (p. 237) show that, when the null hypothesis is true, permutation testing leads to correct Type I error in tests of significance of the Kendall coefficient of concordance. In the classical χ^2 test, however, Type I error is too low when the number of judges is smaller than 20, leading to tests that are overly conservative and, thus, have reduced power. Because in most real-life applications of the method the number of judges is small, permutation tests should be routinely used to test Kendall's W statistic.

The overall test of significance using permutations is interpreted as follows. The null hypothesis is that of independence of all judges. If the probability is smaller than or equal to the nominal significance level α, the null hypothesis should be rejected with a probability of Type I error equal to α. One concludes that the judges are not all independent of one another. There is at least partial concordance among them.

Siegel (1956, p. 237) as well as Siegel and Castellan (1988, p. 271) wrote: "A high or significant value of W may be interpreted as meaning that the observers or judges are applying essentially the same standard in ranking the N objects under study. Often their

pooled ordering may serve as a 'standard,' especially when there is no relevant criterion for ordering the objects." When H_0 is rejected, one cannot conclude that all judges are concordant with one another; only that at least one of the judges is concordant with one, or some of the others. The partial concordance coefficients and *a posteriori* tests help in identifying the groups of judges that ranked the objects in the same way. It is advisable to look, for instance by clustering, for one or several groups of judges that rank the objects broadly in the same way, and then carry out *a posteriori* tests on the putative members of each group, before pooling their values or ranks into an overall index used for ranking the objects.

The *a posteriori* tests are interpreted as follows. The null hypothesis is that of independence of a given judge with respect to all other judges.

- If the probability is smaller than or equal to the nominal significance level α, the null hypothesis should be rejected for this judge with a probability of Type I error equal to α. This judge is concordant with other judges in the set under study.
- If the probability is larger than the nominal significance level α, the null hypothesis cannot be rejected. One concludes that this judge differs from most or all the other judges. The strength of the evidence against the null hypothesis is measured by the probability (Baird 1988): the higher it is, the more different the corresponding judge is with respect to most of the other judges in the analysis.

11.2 ECOLOGICAL ASPECTS

For the mite data, the *a posteriori* tests allowed the identification of two separate groups of concordant species that can be used as indicators of two sets of environmental determinants. This result is ecologically quite interesting: none of our previous analyses (papers cited in Section 2) had shown, on statistical grounds, a separation of the mites into two groups. The groups correspond to the second view of species associations described by Whittaker's (1962): species are associated (positive correlations) as a result of their concordant abundances across the sites, without any implication of present or past biological interactions among them.

By applying the method described in Section 8, ecological community composition data can be used to produce indicators of environmental quality, including pollution and environmental contamination of various kinds. In Section 10, simple indices were constructed by summing the abundances of the species found to be concordant. More sophisticated indices could be obtained by canonical analysis (canonical redundancy analysis or canonical correspondence analysis) of the abundances of the species that belong to the concordant groups, against the environmental or contamination variables; see Legendre and Legendre (1998, chap. 11) for details on these asymmetric forms of canonical analysis.

ACKNOWLEDGMENTS

I am grateful to Daniel Borcard and to an anonymous reviewer who provided constructive comments on manuscript drafts of this article. This work was supported by NSERC grant OGP0007738 to P. Legendre.

[Received July 2004. Revised November 2004.]

REFERENCES

Baird, D. (1988), "Significance Tests, History and Logic," in *Encyclopedia of Statistical Sciences* (vol. 8), eds. S. Kotz and N. L. Johnson, New York: Wiley, pp. 466–471.

Borcard, D., and Legendre, P. (1994), "Environmental Control and Spatial Structure in Ecological Communities: An Example Using Oribatid Mites (Acari, Oribatei)," *Environmental and Ecological Statistics*, 1, 37–53.

Borcard, D., Legendre, P., and Drapeau, P. (1992), "Partialling Out the Spatial Component of Ecological Variation," *Ecology*, 73, 1045–1055.

Borcard, D., Legendre, P., Avois-Jacquet, C., and Tuomisto, H. (2004), "Dissecting the Spatial Structure of Ecological Data at Multiple Scales," *Ecology*, 85, 1826–1832.

Calinski, T., and Harabasz, J. (1974), "A Dendrite Method for Cluster Analysis," *Communications in Statistics*, 3, 1–27.

Edgington, E. S. (1995), *Randomization Tests* (3rd ed.), New York: Marcel Dekker.

Fager, E. W., and McGowan, J. A. (1963), "Zooplankton Species Groups in the North Pacific," *Science* (Washington DC), 140, 453–460.

Gleason, H. A. (1926), "The Individualistic Concept of the Plant Association," *Bulletin of the Torrey Botanical Club*, 53, 7–26.

Holm, S. (1979), "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 6, 65–70.

Hope, A. C. A. (1968), "A Simplified Monte Carlo Test Procedure," *Journal of the Royal Statistical Society*, Ser. B, 50, 35–45.

Jackson, D. A., Somers, K. M., and Harvey, H. A. (1992), "Null Models and Fish Communities: Evidence of Nonrandom Patterns," *The American Naturalist*, 139, 930–951.

Legendre, P. (2000), "Comparison of Permutation Methods for the Partial Correlation and Partial Mantel Tests," *Journal of Statistical Computation and Simulation*, 67, 37–73.

Legendre, P., and Gallagher, E. D. (2001), "Ecologically Meaningful Transformations for Ordination of Species Data," *Oecologia*, 129, 271–280.

Legendre, P., and Legendre, L. (1998), *Numerical Ecology* (2nd English ed.), Amsterdam: Elsevier Science BV.

Milligan, G. W., and Cooper, M. C. (1985), "An Examination of Procedures for Determining the Number of Clusters in a Dataset," *Psychometrika*, 50, 159–179.

Pavillard, J. (1912), "Essai sur la Nomenclature Phytogéographique," *Bulletin de la Société Languedocienne de Géographie*, 35, 165–176.

Rao, C. R. (1995), "A Review of Canonical Coordinates and an Alternative to Correspondence Analysis Using Hellinger Distance," *Qüestiió (Quaderns d'Estadística i Investigació Operativa)*, 19, 23–63.

Siegel, S. (1956), *Nonparametric Statistics for the Behavioral Sciences*, New York: McGraw-Hill.

Siegel, S., and Castellan, N. J., Jr. (1988), *Nonparametric Statistics for the Behavioral Sciences* (2nd ed.), New York: McGraw-Hill.

Whittaker, R. H. (1956), "Vegetation of the Great Smoky Mountains," *Ecological Monographs*, 26, 1–80.

——— (1962), "Classification of Natural Communities," *The Botanical Review*, 28, 1–239.

Wright, S. P. (1992), "Adjusted P Values for Simultaneous Inference," *Biometrics*, 48, 1005–1013.

Zar, J. H. (1999), *Biostatistical Analysis* (4th ed.), Upper Saddle River, New Jersey: Prentice Hall.