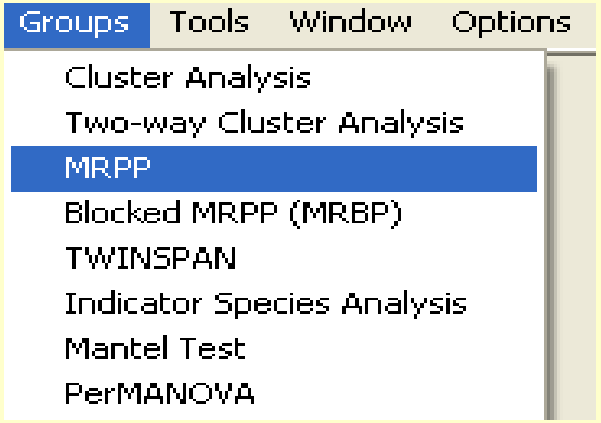
MRPP & Indicator Species

- *Objectives:*

Discuss general approaches of these two methods

Go over settings and results for these two methods

MRPP – Applications



➢ Multi-response Permutation Procedure (MRPP) is a non-parametric approach for testing the hypothesis of no differences between two or more groups of entities (species, variables)

➢ These pre-existing groups are defined using groups of samples on the basis of categorical levels (discrete groupings):

- Categories of environmental variables (e.g., early vs. late; water masses)

- The presence / absence of given species

➢ Recommendation This procedure yields a p value. If results significant, interpretation requires further exploration

MRPP – Pros / Cons

➢ Advantages:

• Ideal for evaluating specific hypotheses – differences between groups of samples

➢ Disadvantages:

• Cannot investigate interaction terms
(one grouping variable only; no variable correlations)

• Interpretation – difficult to determine what species are contributing to differences in community composition – requires additional exploration of the data

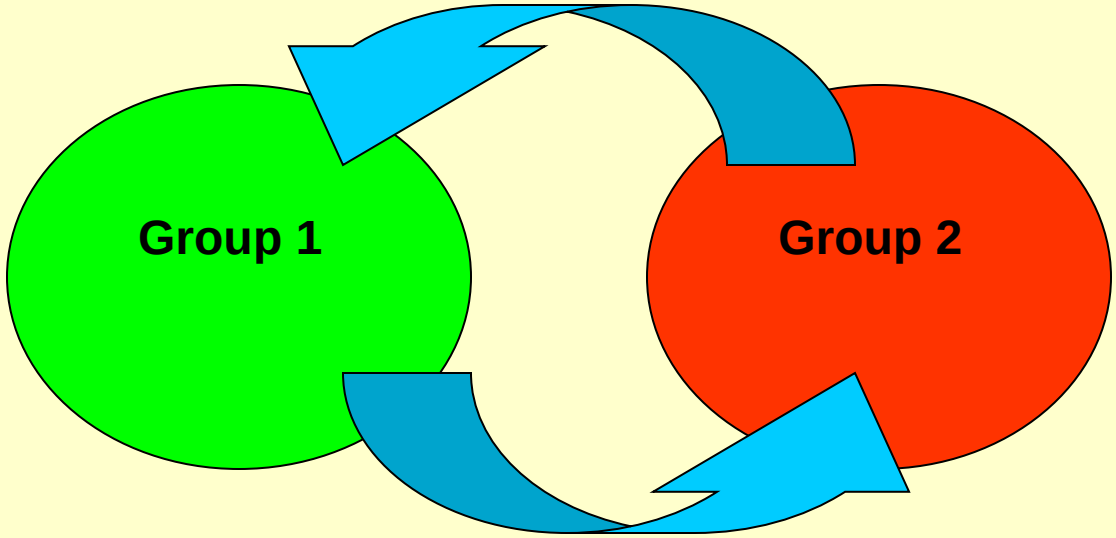
• Recommend: Follow-up with Indicator Species Analysis

MRPP – How it Works

➢ Setting Up:

- Define a Grouping Variable:
  Species Presence / Absence - Main Matrix
  Environmental Categorical Variable – Second Matrix

- Select a distance measure (Sorensen / Relative Sorensen) and calculate matrix of distances (D) between all pairs of points within each of the pre-defined groups you are testing



- Shuffle data and recalculate distances, for all possible arrangements of samples into groups

MRPP – How it Works

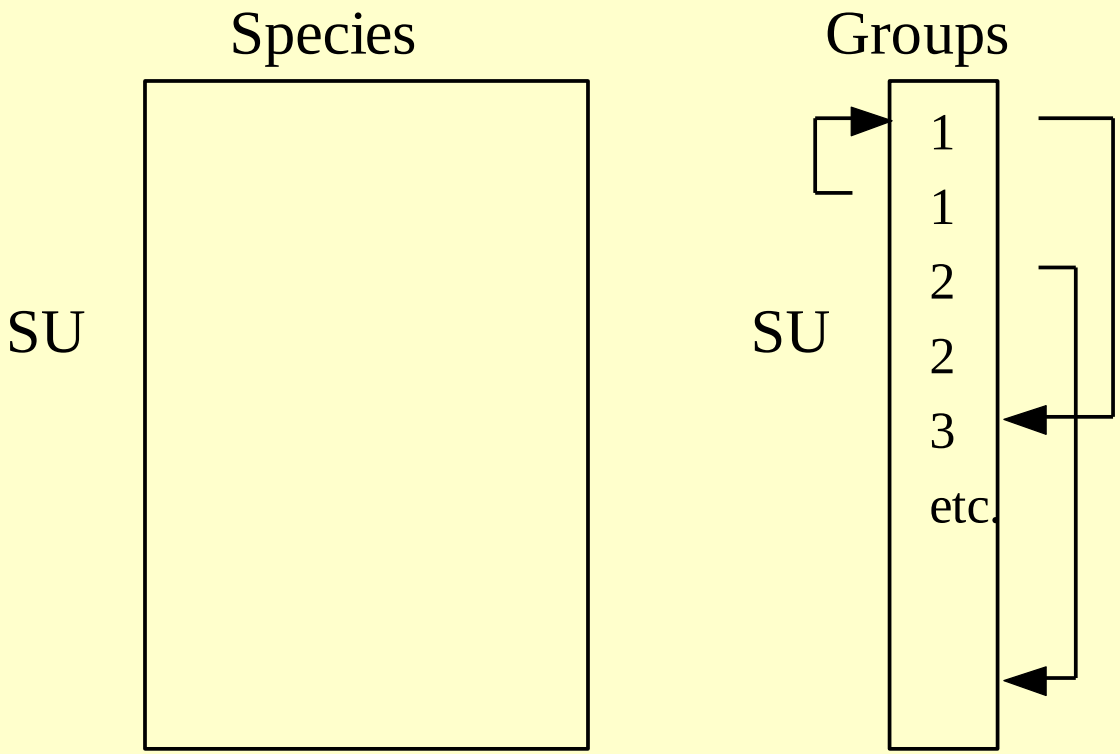
- Calculate distance matrix, **D**
- Calculate average distance x_i within each group i
- Calculate delta (weighted mean within-group distance)

Note: For g groups, recommended weight is delta: where C depends on the number of items in the groups ($C_i = n_i / N$, where n_i is the number of items in group i and N is the total number of items)

$$delta = \delta = \sum_{i=1}^{g} C_i x_i$$

MRPP – How it Works

- Permutations: $M = N! / (n_1! * n_2!)$
- Determine probability of a δ this small or smaller

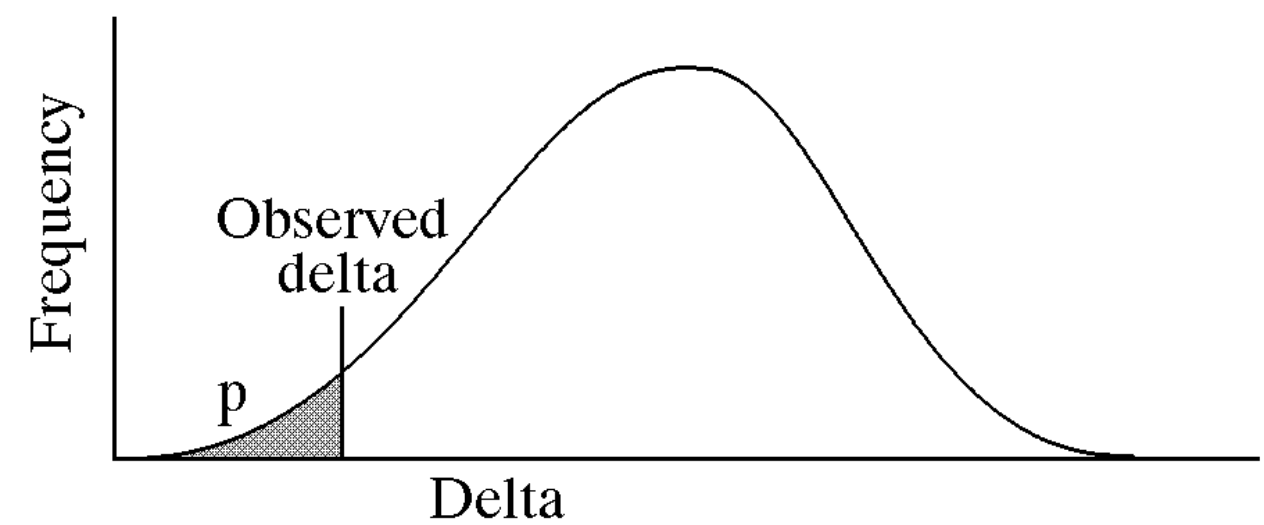


MRPP – How it Works

- Calculating the p value:

- Determine probability of a δ as small or smaller

$$p = \frac{1 + \text{no. smaller deltas}}{\text{total no. possible partitions}}$$



MRPP – How it Works

- Output:

- Test Statistic T: measures effect size

$$T = \frac{\text{observed } \delta - \text{expected } \delta}{\text{s.dev. of expected } \delta}$$

- A statistic: within-group agreement

$$A = 1 - \frac{\delta}{m_\delta} = 1 - \frac{\text{observed } \delta}{\text{expected } \delta}$$

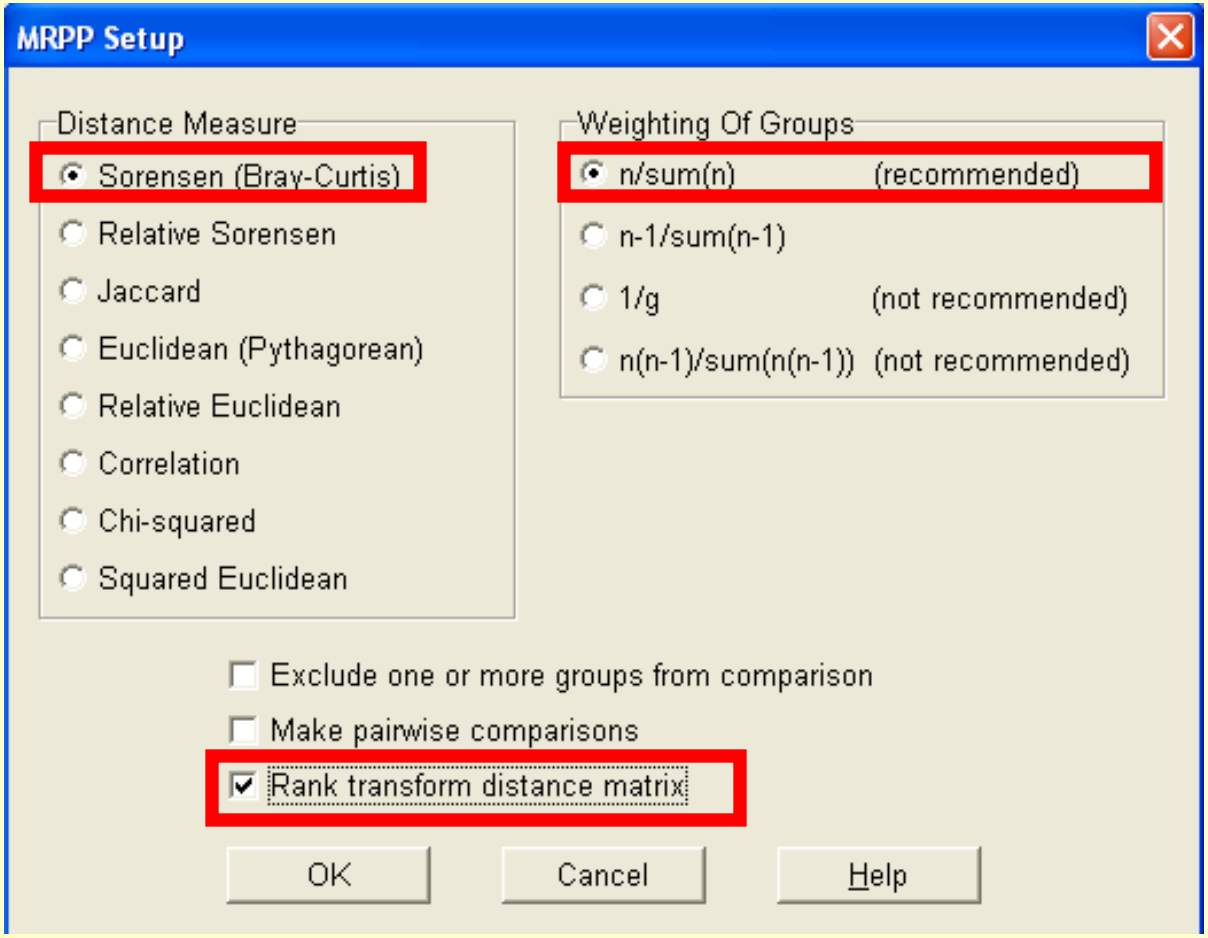
BEWARE:
DO NOT
over-interpret
T and A

Ongoing
Discussion

- P-value: Null Hypothesis:

within-group distance the same as amongst-group distances

MRPP – Suggested Procedure: Step1



➢ First, pick distance measure

- Distance: Sorensen

➢ Second, select Weights of Groups

- Recommend:

  n / sum (n)

➢ Third, use Ranks

- Useful for very heterogeneous data
- More comparable to NMS

Why use the Distance Ranking ?

- MRPP allows user to rank transform the distance matrix. This approach can be applied to any distance measure.

- In practice with community data, the test statistic, skewness of the test statistic under the null hypothesis, and the resulting p-value are often similar, whether the data are ranked or not.

- The chance-corrected within-group agreement, however, is often higher after the distance measure is converted to ranks.

How does the Distance Ranking Work ?

- The ranking procedure operates as follows:

Ties are assigned average rank of the tied elements. For example, values 1, 3, 3, 9, 10 receive ranks 1, 2.5, 2.5, 4, 5.

After elements assigned initial ranks, they are adjusted by subtracting the rank of the zero distance. This results in all raw distances of zero being assigned a rank distance of zero.

For example:

Five zero distances in the matrix would each be assigned a rank of 3, taking into account the five-way tie.

Then 3 is subtracted from each element in the matrix, thus recoding these zero distances into 0s.

What does the Distance Ranking do ?

- The rank transformation:

  - helps to correct the loss of sensitivity of distance measures as community heterogeneity increases.
  - makes the MRPP results more analogous to non-metric multidimensional scaling.

But, it also changes null hypothesis from "average within-group distance no smaller than expected by chance" to "no difference in average within-group rank of distances."

Note: MRPP on ranked distances with Sorensen distance - is similar to ANOSIM (analysis of similarity).

MRPP – Results

- Examine Results.txt file: Distribution of samples into groups

```
Data Distributions

        Groups were defined by values of: time
        Input data has:       20 years     by      20 Vars
        Weighting option: C(I) = n(I)/sum(n(I))
        Distance measure: Sorensen (Bray-Curtis)
        Distance matrix was rank transformed.

GROUP:     1
Code:      1
Size:     10       0.53328042     = Average distance
Members:
 YR85      YR86      YR87      YR88      YR89      YR90      YR91      YR92
 YR93      YR94

GROUP:     2
Code:      2
Size:     10       0.32428572     = Average distance
Members:
 YR97      YR98      YR99      YR00      YR01      YR02      YR03      YR04
 YR05      YR06
```

MRPP – Results

- Examine Results.txt file: T & A Statistics

```
Test statistic: T =       -4.1790420
   Observed delta =       0.42878307
   Expected delta =       0.50000000
Variance of delta =       0.29041098E-03
Skewness of delta =       -1.5010703

Chance-corrected within-group agreement, A =     0.14243386
  A = 1 - (observed delta/expected delta)
  Amax = 1 when all items are identical within groups (delta=0)
  A = 0 when heterogeneity within groups equals expectation by chance
  A < 0 with more heterogeneity within groups than expected by chance

Probability of a smaller or equal delta  p =     0.00361764
```

Smaller observed delta

A > 0 (more similar within groups)

Significant result: p < 0.05

- Fairly small Output: NO bi-plots, NO variance explained

MRPP – What to Report

- Distance Metric Used (Sorensen / Relative Sorensen)
- How groups were defined – Relate back to Hypothesis
- Chance corrected within-group agreement (A)
- Associated p value

MRPP + Indicator Species Analysis

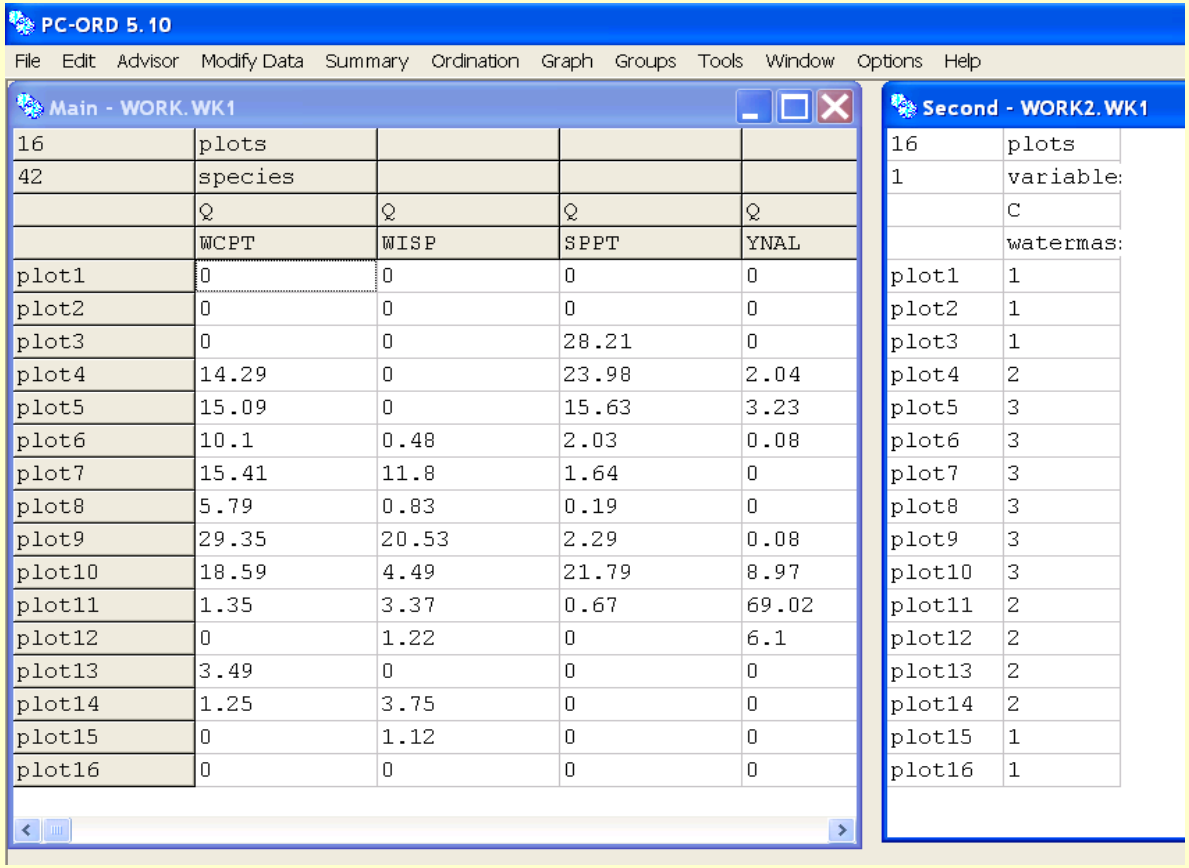
- Distance Metric Used (Relative Sorensen)
- Group Definitions – Relate back to Hypothesis

3 Watermasses:

- 1) Tropical:
  SST > 20 deg. C.

- 2) Subtropical:
  20 < SST < 18

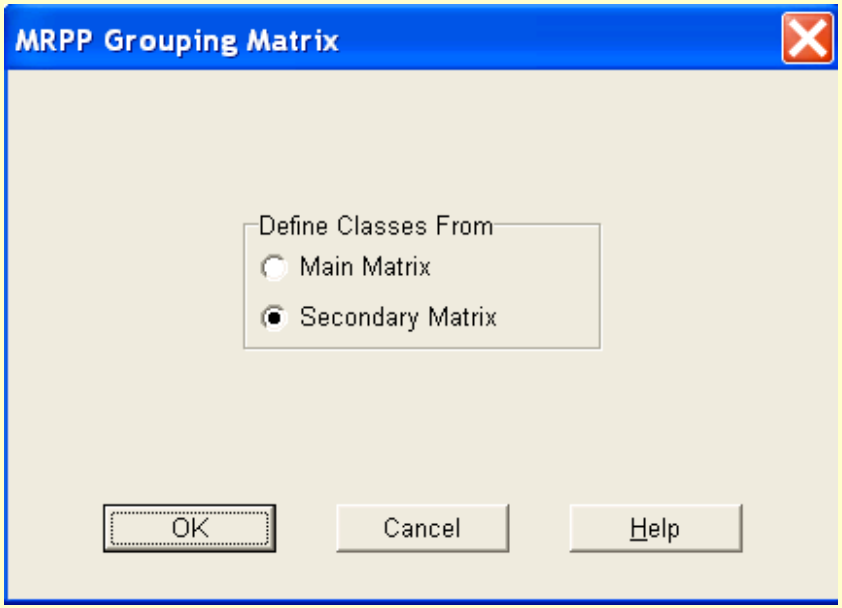
-3) Transition:
  18 > SST

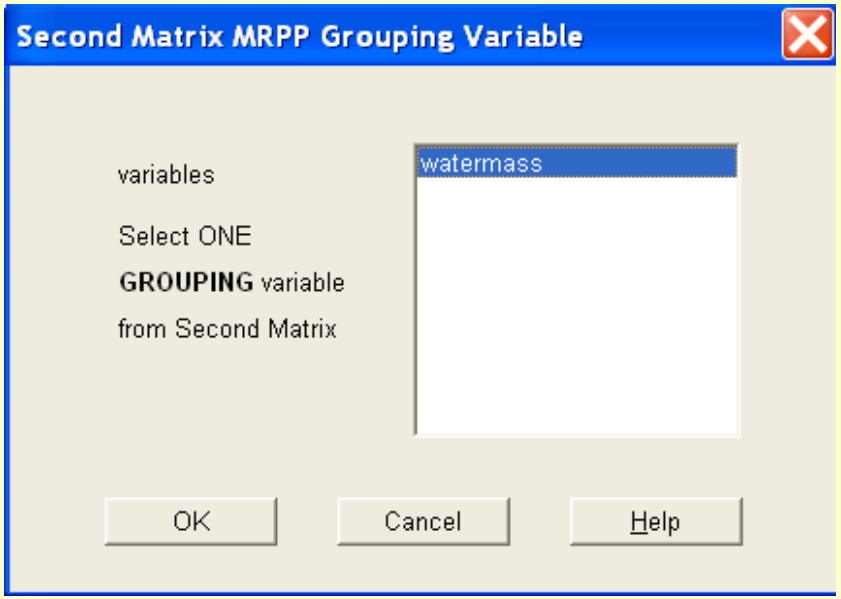


16	plots			
42	species			
	Q	Q	Q	Q
	WCPT	WISP	SPPT	YNAL
plot1	0	0	0	0
plot2	0	0	0	0
plot3	0	0	28.21	0
plot4	14.29	0	23.98	2.04
plot5	15.09	0	15.63	3.23
plot6	10.1	0.48	2.03	0.08
plot7	15.41	11.8	1.64	0
plot8	5.79	0.83	0.19	0
plot9	29.35	20.53	2.29	0.08
plot10	18.59	4.49	21.79	8.97
plot11	1.35	3.37	0.67	69.02
plot12	0	1.22	0	6.1
plot13	3.49	0	0	0
plot14	1.25	3.75	0	0
plot15	0	1.12	0	0
plot16	0	0	0	0

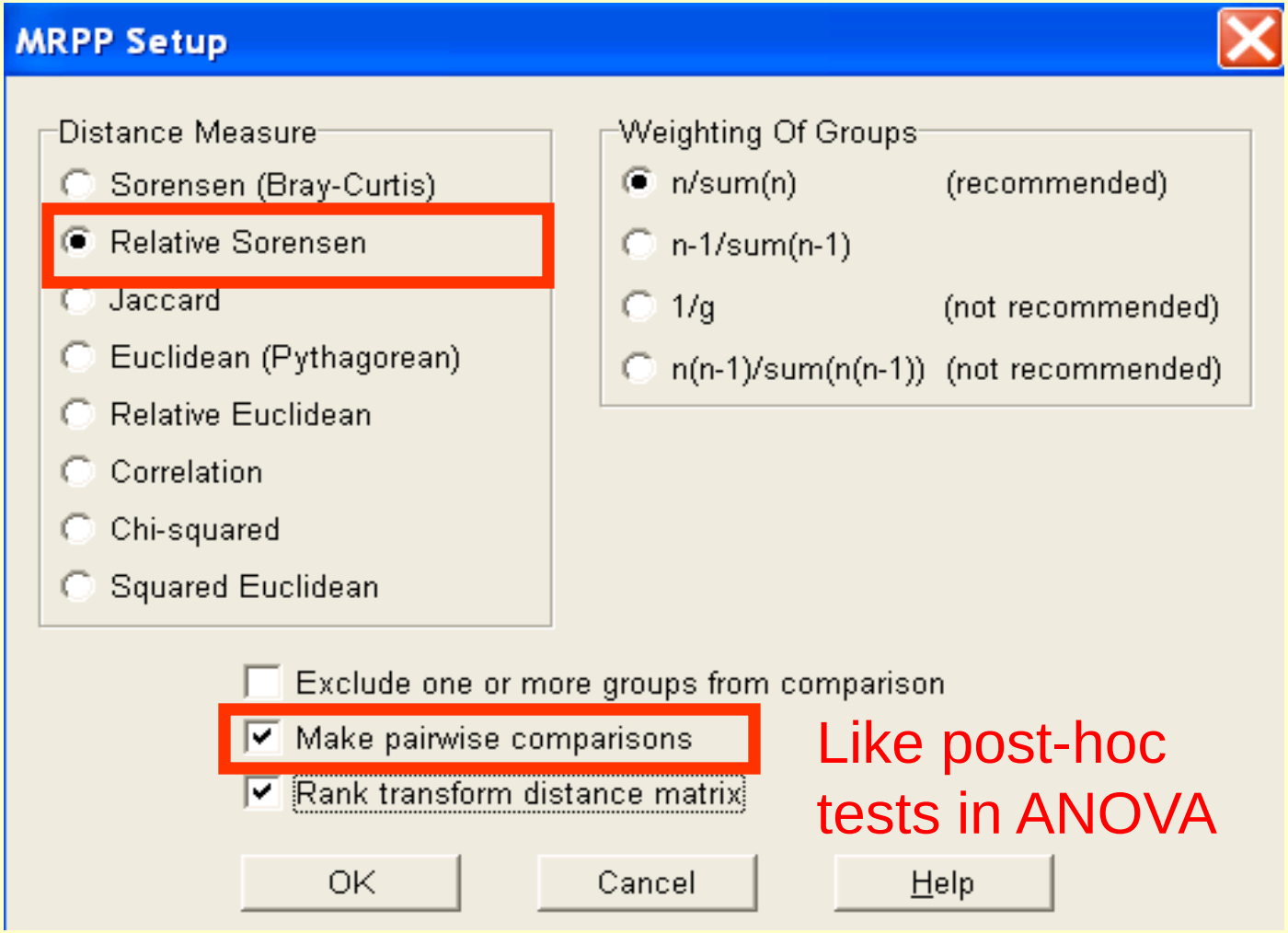
16	plots
1	variable:
	C
	watermas:
plot1	1
plot2	1
plot3	1
plot4	2
plot5	3
plot6	3
plot7	3
plot8	3
plot9	3
plot10	3
plot11	2
plot12	2
plot13	2
plot14	2
plot15	1
plot16	1

Selecting Grouping Variable





Selecting Grouping Variable



Like post-hoc tests in ANOVA

MRPP – Results

```
****************** Multi-Response Permutation Procedures (MRPP)

IndianOceanBirds_Groups

        Groups were defined by values of: watermas
        Input data has:      16 plots    by     42 species
        Weighting option: C(I) = n(I)/sum(n(I))
        Distance measure: Relative Sorensen
        Distance matrix was rank transformed.

GROUP:     1
Code:      1
Size:      5       0.42463235     = Average distance
Members:
 plot1    plot2    plot3    plot15   plot16

GROUP:     2
Code:      2
Size:      5       0.34338235     = Average distance
Members:
 plot4    plot11   plot12   plot13   plot14

GROUP:     3
Code:      3
Size:      6       0.18504902     = Average distance
Members:
 plot5    plot6    plot7    plot8    plot9    plot10


        Test statistic: T =       -6.0294581
           Observed delta =       0.30939798
           Expected delta =       0.50000000
        Variance of delta =       0.99930586E-03
        Skewness of delta =      -0.76504516


        Chance-corrected within-group agreement, A =     0.38120404
          A = 1 - (observed delta/expected delta)
          Amax = 1 when all items are identical within groups (delta=0)
          A = 0 when heterogeneity within groups equals expectation by chance
          A < 0 with more heterogeneity within groups than expected by chance

        Probability of a smaller or equal delta, p =     0.00003053
```

Reject the Null Hypothesis.

But… which groups are different

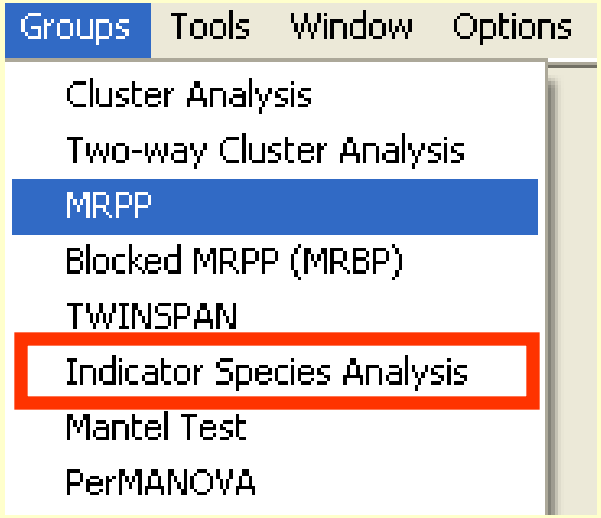
MRPP – Results

PAIRWISE COMPARISONS
Note: p values not corrected for multiple comparisons.

Group Codes Compared		T	A	p
1 vs.	2	-2.56062667	0.16363638	0.01809289
1 vs.	3	-5.69119563	0.36515151	0.00070369
2 vs.	3	-4.49992683	0.32506886	0.00309954

All three groups are different from each other

But… which species are responsible for the differences?



What Traits make a Good Indicator Species

Ecological Monographs, 67(3), 1997, pp. 345–366

(Dufrêne & Legendre 1997)

SPECIES ASSEMBLAGES AND INDICATOR SPECIES: THE NEED FOR A FLEXIBLE ASYMMETRICAL APPROACH

MARC DUFRÊNE[1] AND PIERRE LEGENDRE[2]

[1]*Unité d'Écologie et de Biogéographie, Université catholique de Louvain, Croix du Sud, 5, B-1348 Louvain-la-Neuve, Belgium*
[2]*Département de Sciences Biologiques, Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal, Québec, Canada H3C 3J7*

Abstract. This paper presents a new and simple method to find indicator species and species assemblages characterizing groups of sites. The novelty of our approach lies in the way we combine a species relative abundance with its relative frequency of occurrence in the various groups of sites. This index is maximum when all individuals of a species are found in a single group of sites and when the species occurs in all sites of that group; it is a symmetric indicator. The statistical significance of the species indicator values is evaluated using a randomization procedure. Contrary to TWINSPAN, our indicator index for a given species is independent of the other species relative abundances, and there is no need to use pseudospecies.

The new method identifies indicator species for typologies of species relevés obtained by any hierarchical or nonhierarchical classification procedure; its use is independent of the classification method. Because indicator species give ecological meaning to groups of sites, this method provides criteria to compare typologies, to identify where to stop dividing clusters into subsets, and to point out the main levels in a hierarchical classification of sites.

Species can be grouped on the basis of their indicator values for each clustering level, the heterogeneous nature of species assemblages observed in any one site being well preserved. Such assemblages are usually a mixture of eurytopic (higher level) and stenotopic species (characteristic of lower level clusters). The species assemblage approach demonstrates the importance of the "sampled patch size," i.e., the diversity of sampled ecological combinations, when we compare the frequencies of core and satellite species. A new way to present species–site tables, accounting for the hierarchical relationships among species, is proposed. A large data set of carabid beetle distributions in open habitats of Belgium is used as a case study to illustrate the new method.

ISA: Pros & Cons

Index is maximum when all individuals of a species found in a single group of sites and when the species occurs in all sites of that group.

It is a symmetric indicator:
% occurrence and % abundance have the same weight

Contrary to TWINSPAN, the ISA index for a given species is independent of the other species relative abundances; does not require use of pseudospecies.

Species only indicate one “community” or “habitat”.
Is this ecologically-realistic ?

(Dufrêne & Legendre 1997)

Indicator Species

- Good indicator species should be found mostly in a single group and be present at most of sites belonging to that group.

IndVal method proposed by Dufrêne and Legendre (1997):

IndValGroup *k, Species j* = *100 x A k,j x B k,j*

In that equation, A *k,j* = *Specificity* *Bk, j* = *Fidelity*

IndVal Species *j* = *max [IndVal k,j]*

ISA – How it Works

Calculate proportional abundance of species in particular group relative to abundance of that species in all groups.

Let **A** = sample unit × species matrix

a_{ijk} = abundance of species *j* in sample unit *i* of group *k*

n_k = number of sample units in group *k*

g = total number of groups

First calculate the mean abundance x_{kj} of species *j* in group *k*:

$$x_{kj} = \frac{\sum_{i=1}^{n_k} a_{ijk}}{n_k}$$

ISA – How it Works

Then calculate the relative abundance RA_{kj} of species *j* in group *k* (this measures specificity or exclusiveness, the concentration of abundance into a particular group):

$$RA_{jk} = \frac{x_{kj}}{\sum_{k=1}^{g} x_{kj}}$$

ISA – How it Works

Calculate the proportional frequency of the species in each group (fidelity, or the proportion of sample units in each group that contain that species).

First transform **A** to a matrix of presence-absence, **B**

Then calculate relative frequency RF_{kj} of species *j* in group *k*:

$$RF_{kj} = \frac{\sum_{i=1}^{n_k} b_{ijk}}{n_k}$$

ISA – How it Works

Combine the two proportions calculated in two previous steps, by multiplying them. Express the result as a percentage, yielding an indicator value (IV_{kj}) for each species *j* in each group *k*.

The highest indicator value (IV_{max}) for a given species across all groups is the indicator value of that species.

Evaluate statistical significance of IV_{max} by randomly reassigning SUs to groups 1000 times.

Each time, calculate IV_{max}.

H_0: IV_{max} is no larger than would be expected by chance (i.e., the species has no indicator value).

ISA -2 Step Process

RELATIVE ABUNDANCE in group
(% of perfect indication)

average abundance of species in a given group of plots, compared to the average abundance of that species in all plots, expressed as %

				Group		
			Sequence:	1	2	3
			Identifier:	1	2	3
			Number of items:	5	5	6
	Column	Avg	Max MaxGrp			
1	WCPT	33	79 3	0	21	79
2	WISP	33	77 3	3	20	77
3	SPPT	33	41 3	32	28	41

Possible Values: 0 - 100

RELATIVE FREQUENCY in group
(% of perfect indication)

% of plots in given group where given species is present

				Group		
			Sequence:	1	2	3
			Identifier:	1	2	3
			Number of items:	5	5	6
	Column	Avg	Max MaxGrp			
1	WCPT	60	100 3	0	80	100
2	WISP	54	83 3	20	60	83
3	SPPT	53	100 3	20	40	100

Possible Values: 0 - 100

ISA -2 Step Process

INDICATOR VALUES
(% of perfect indication)

based on combining the above values
for relative abundance and relative frequency

				Group		
			Sequence:	1	2	3
			Identifier:	1	2	3
			Number of items:	5	5	6
	Column	Avg	Max	MaxGrp			
1	WCPT	32	79	3	0	16	79
2	WISP	26	64	3	1	12	64
3	SPPT	19	41	3	6	11	41

Possible Values: 0 - 100

ISA – Randomizations

MONTE CARLO test of significance of observed maximum indicator value for species
999 permutations.
Random number seed: 5807

Column	Maxgrp	Observed Indicator Value (IV)	IV from randomized groups Mean	S.Dev	p *
1 WCPT	3	79.4	37.8	10.98	0.0020
2 WISP	3	64.2	40.2	14.33	0.0821
3 SPPT	3	40.7	38.1	13.01	0.3594

Expected Distribution

* proportion of randomized trials with indicator value equal to or exceeding the observed indicator value.

p = (1 + number of runs >= observed)/(1 + number of randomized runs)

Maxgrp = Group identifier for group with maximum observed IV

ISA – Looking at the Species

RELATIVE ABUNDANCE in group (% of perfect indication)

average abundance of species in a given group of plots, compared to the average abundance of that species in all plots, expressed as %

species	mean	max	group	g1	g2	g3
BRPT	33	100	1	100	0	0
AUSH	33	100	1	100	0	0
MSPT	33	100	1	100	0	0
BUPT	33	100	1	100	0	0
WTTR	33	100	1	100	0	0
WNPT	33	100	1	100	0	0
DKTE	33	100	1	100	0	0
WTSH	33	100	1	100	0	0
LISH	33	58	1	58	10	32
LTJA	33	42	1	42	39	18
GWPT	33	65	2	35	65	0
BBSP	33	51	2	20	51	29
GRSH	33	100	2	0	100	0
COSH	33	100	2	0	100	0
FFSH	33	100	2	0	100	0
JFPT	33	100	2	0	100	0
ROPN	33	96	2	0	96	4
YNAL	33	88	2	0	88	12
SOSH	33	87	2	0	87	13
SOAL	33	81	2	0	81	19

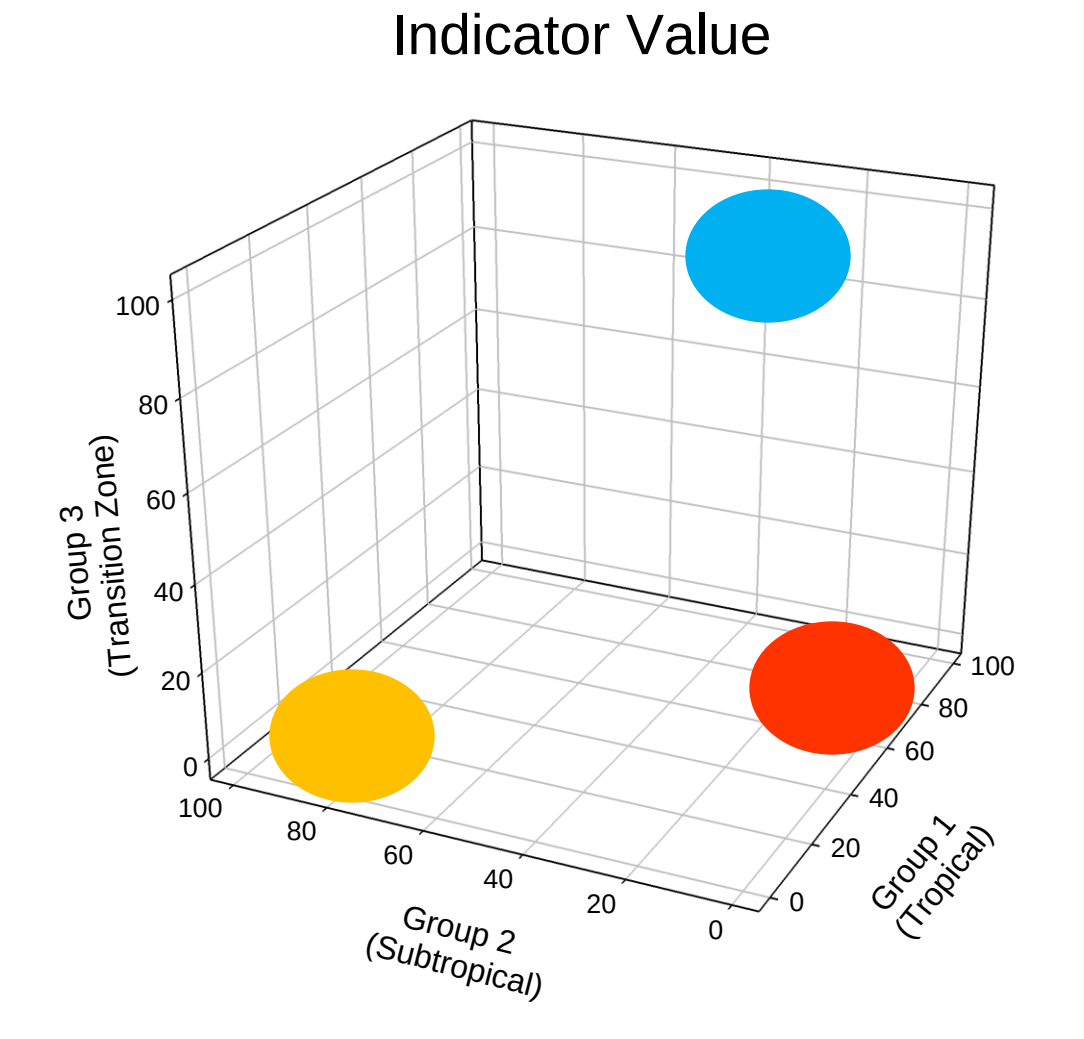
ISA – Looking at the Species

RELATIVE FREQUENCY in group (% of perfect indication)

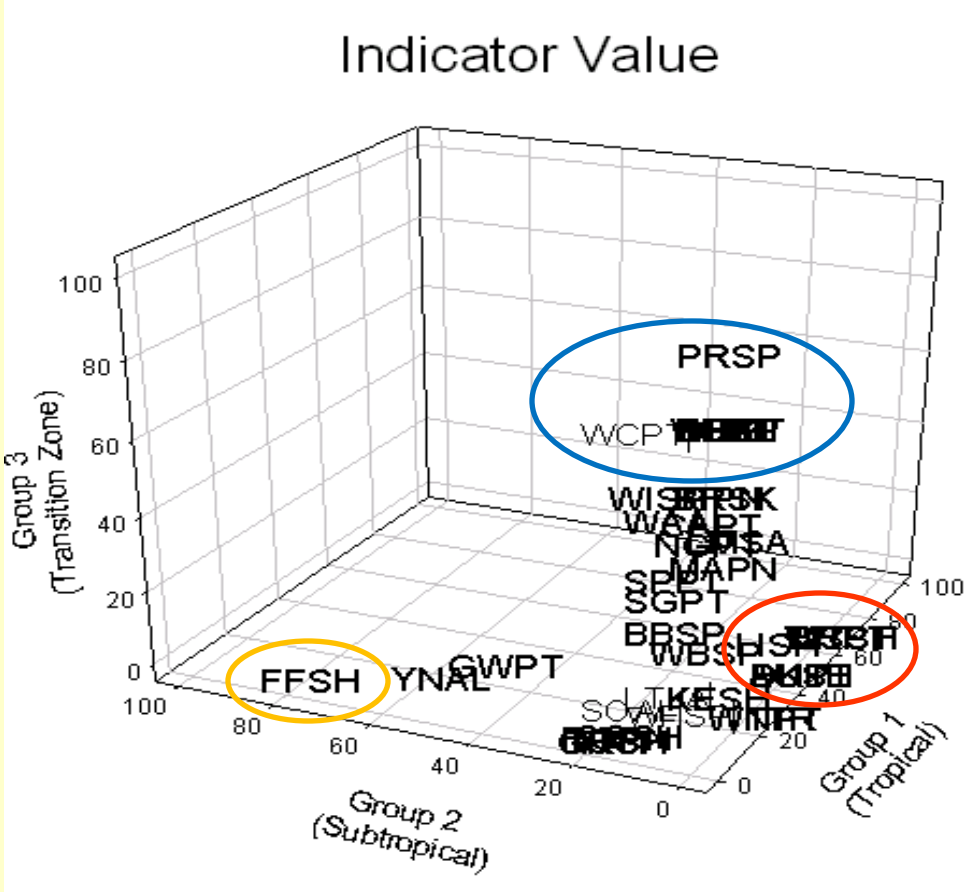
% of plots in given group where given species is present

species	mean	max	group	g1	g2	g3
BRPT	20	60	1	60	0	0
MSPT	20	60	1	60	0	0
WTSH	20	60	1	60	0	0
AUSH	13	40	1	40	0	0
BUPT	13	40	1	40	0	0
DKTE	13	40	1	40	0	0
WTTR	7	20	1	20	0	0
WNPT	7	20	1	20	0	0
GWPT	47	80	2	60	80	0
LISH	38	60	1	60	20	33
LTJA	32	40	1	40	40	17
FFSH	27	80	2	0	80	0
GRSH	7	20	2	0	20	0
COSH	7	20	2	0	20	0
JFPT	7	20	2	0	20	0
SOSH	12	20	2	0	20	17
WFSP	12	20	2	0	20	17

ISA – Looking at the Species



ISA – Looking at the Species



Species	Group	Observed_IV	Mean_IV	STD_IV	p value
PRSP	3	100	30.3	12.5	0.001
WCPT	3	79.4	37.8	10.98	0.002
BBAL	3	83.3	32.4	14.58	0.002
BLPT	3	83.3	29.3	13.71	0.003
GHAL	3	83.3	29.2	13.76	0.003
KEPT	3	83.3	31.3	15.18	0.003
WHPT	3	83.3	30.7	14.55	0.003
GBSP	3	83.3	31.2	15.32	0.003
DPSP	3	83.3	32.8	14.75	0.003
FFSH	2	80	26.1	12.41	0.006
KIPN	3	66.7	27.2	13.51	0.011
BRSK	3	66.7	27.6	13.1	0.014
WTSH	1	60	25.8	12.09	0.032
CAPT	3	60.5	29.1	13.15	0.032
BRPT	1	60	23.3	12.37	0.038
MSPT	1	60	27.2	12.32	0.038
NGPT	3	55.4	27.4	12.2	0.043

Only 17 species (out of 42) are “significant” indicators

Hierarchical Indicator Species

Novelty of approach lies in the combination of species relative abundance and relative frequency of occurrence in various groupings of sites (samples).

Statistical significance of the species indicator values evaluated using a randomization procedure.

Method can be used reiteratively, in conjunction with clustering.

(Dufrêne & Legendre 1997)

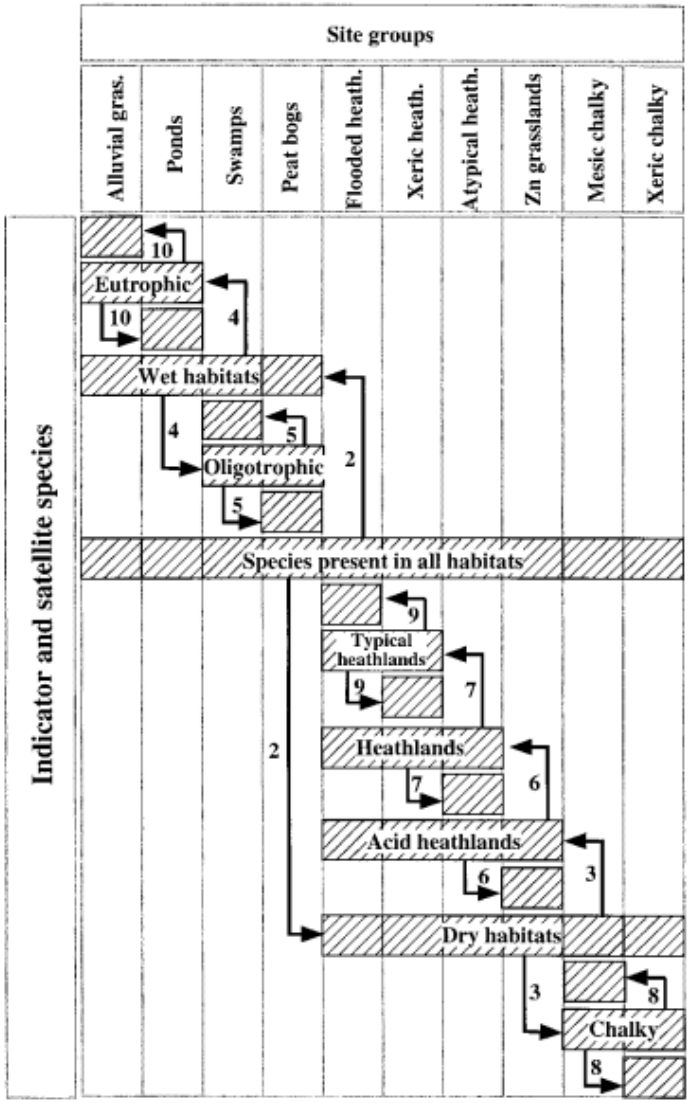
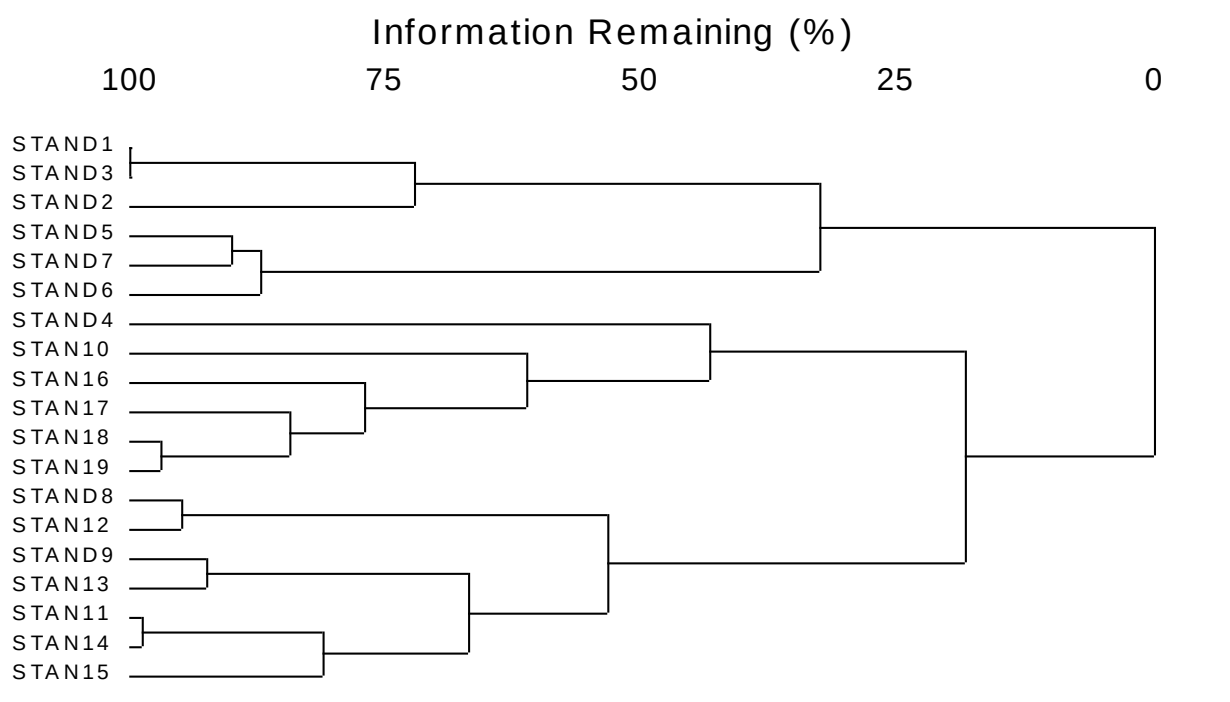


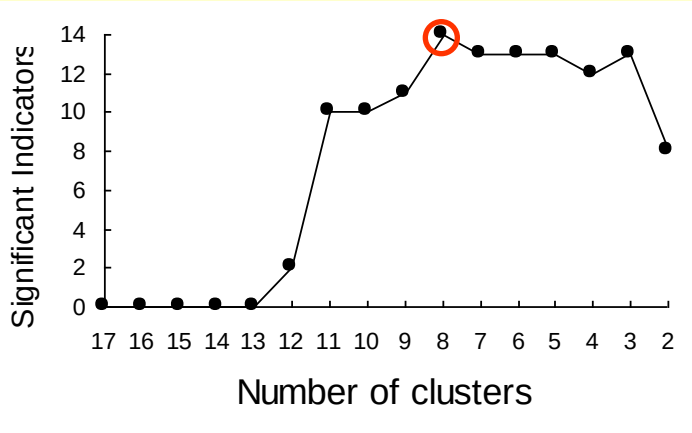
FIG. 13. Steps that are followed to build a two-way table from the hierarchical clusters indicator values. The first species group (center of figure) contains species that are common in all habitats (i.e., having their indicator value maximum when all sites are pooled in one group). At the next step, two species groups are created: one with species dominating in all wet habitats, and the other one with species that are common in all dry habitats. The procedure is repeated for each site cluster.

Indicator Species Applications



ISA provides an objective criterion for pruning a dendrogram.

Number of significant species with $p \leq 0.05$ for each step of clustering.



Best Result:
14 significant indicators, organized into 8 clusters.

Indicator Species Applications

- Classical problem in community ecology and biogeography:

Species are the best indicators we have for particular environmental conditions.

- The identification of characteristic or indicator species is traditional in ecology and biogeography.

- Field studies describing sites or habitats usually mention one or several species that characterize each habitat.

- In long-term environmental conservation and ecological management, researchers are looking for bioindicators of habitat types to preserve or rehabilitate.

MRPP & ISA – References

- MRPP:

Mielke, P. W., Jr. 1991. The application of multivariate permutation methods based on distance functions in the earth sciences. Earth-Science Reviews 31:55-71.

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

- Indicator Species Analysis:

Dufrene, M. & P. Legendre. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecological Monographs 67:345-366.

McGeoch, M. A. and S. L. Chown. 1998. Scaling up the value of bioindicators. Trends in Ecology & Evolution 13: 46-47.