Multivariate Methods For Nondetects, Part 2

Dennis Helsel

Practical Stats

Methods for dealing with nondetects

- NOT Substitution! This adds invasive data
  - Farnham (2002): with as little as 20% NDs, substitution produced incorrectly computed PCA axes
- NOT Deleting variables with NDs
  - Reimann (2002) found PCA and other results change markedly when variables are not used
- NOT Deleting observations with NDs
  - Aruga (1997) found PCA after deleting as little as 5% NDs produced unacceptable results

Better ways to deal with nondetects

1. Binary methods on detect/ND using simple matching coefficient
2. Euclidean distances of ranks
3. Euclidean distance on uscores of multiply-censored data

(first 2 valid for 1 DL or censoring at highest DL)

After computing a resemblance matrix, use a nonparametric method

- ANOSIM (Nonparametric MANOVA)
- Test for seriation (Nonparametric Trend Test)
- Nearest Neighbor Cluster Analysis
- Nonmetric Multidimensional Scaling

Method 1: Binomial methods
detect/ND

- For each variable, 0=LT dl 1=GTE dl
- Apply simple matching coeff for resemblance matrix:

$$S_{jk} = 100 * {(a+d)}/{(a+b+c+d)}$$

where a= joint 1s, d= joint 0s, and b,c = (0,1) or (1,0)

Example: DDT metabolites in fish

Site	opDDD	ppDDD	opDDE	ppDDE	opDDT	ppDDT	Age
1	<5	<5	<5	14	<5	<5	Young
2	<5	42	8.4	130	<5	31	Mature
3	5.3	38	<5	250	<5	11	Mature
4	<5	12	<5	57	<5	<5	Mature
5	<5	<5	<5	16	<5	<5	Young
6	<5	<5	<5	<5	<5	<5	Young

. . .(first 6 of 32 rows)

Example: DDT metabolites in fish

Site	opDDD	ppDDD	opDDE	ppDDE	opDDT	ppDDT	Age
1	0	0	0	1	0	0	Young
2	0	1	1	1	0	1	Mature
3	1	1	0	1	0	1	Mature
4	0	1	0	1	0	0	Mature
5	0	0	0	1	0	0	Young
6	0	0	0	0	0	0	Young

0 = below 5 1 = equal to or >5

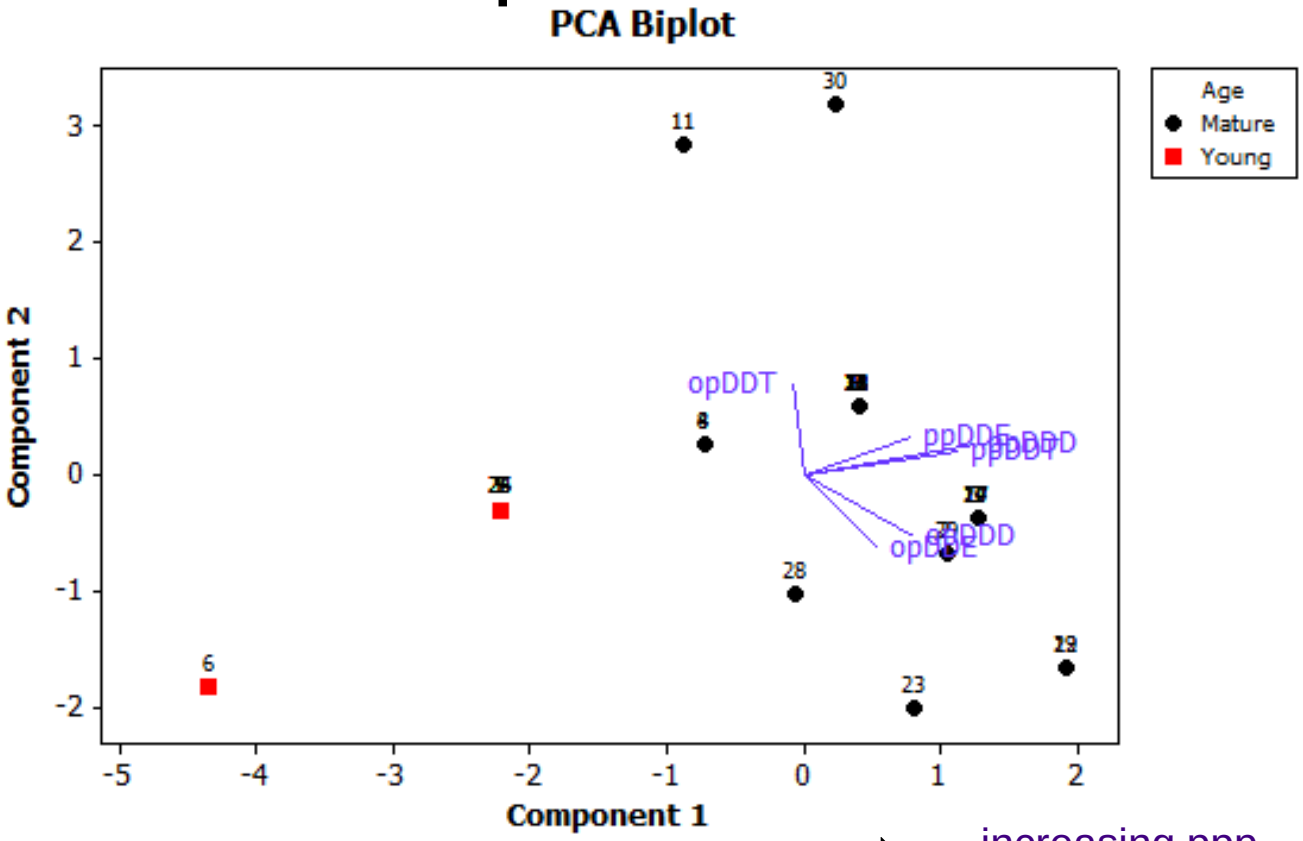
(first 6 of 32 rows)

Example: DDT metabolites in fish

Site	1	2	3	4	5	6
1	*					
2	50	*				
3	50	66.7	*			
4	83.3	66.7	66.7	*		
5	100	50	50	83.3	*	
6	83.3	33.3	33.3	66.7	83.3	*

simple matching coefficient
(first 5 rows of triangular matrix)

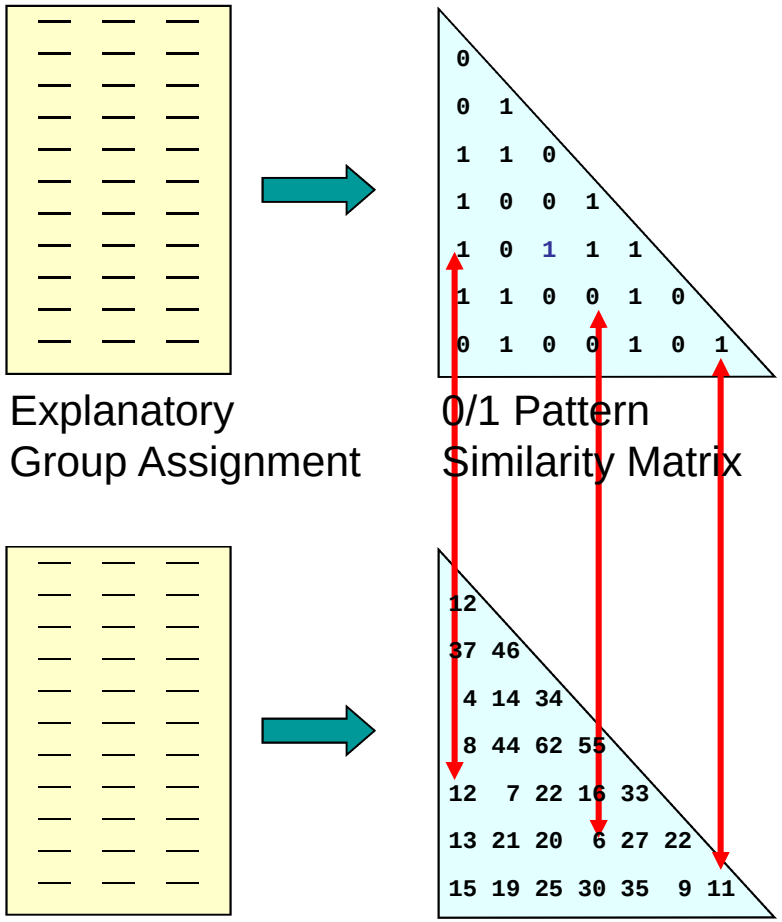
PCA biplot on 0/1 data



increasing ppp forms

sites 11, 30 have highest opDDT occurrence

ANOSIM



Explanatory Group Assignment

0/1 Pattern Similarity Matrix

Binary Data

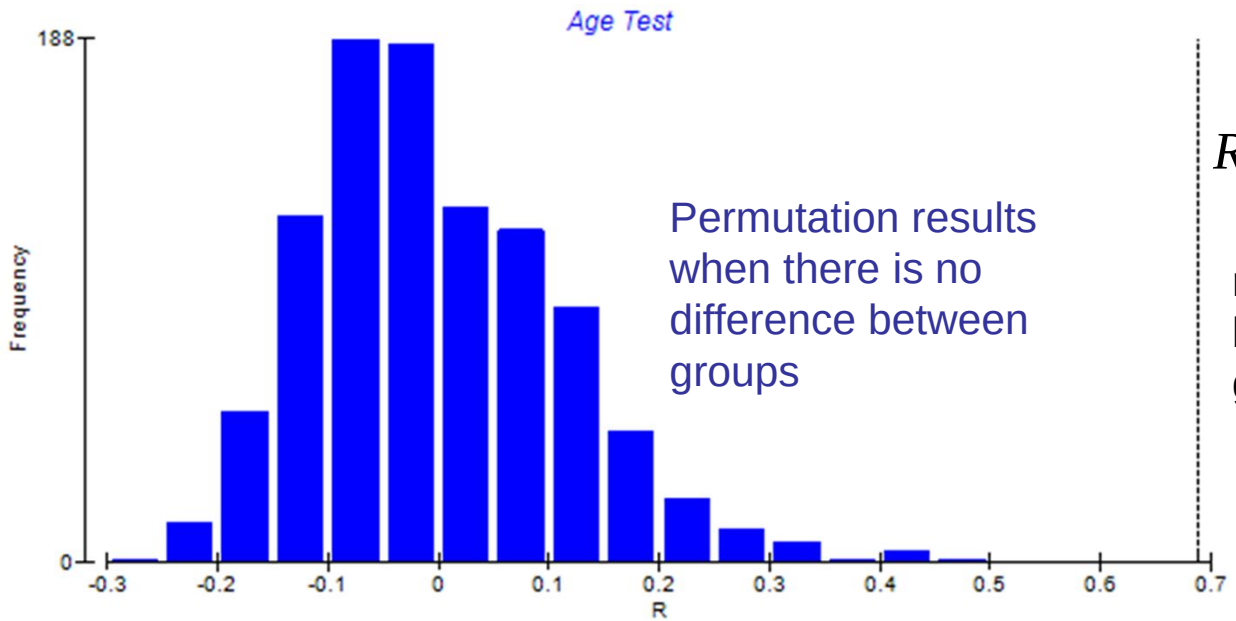
Simple matching similarity matrix

For tests between groups, the similarity matrix value is a 1 if two obs are in the same group, and a 0 otherwise.

Permutation p-values

Young vs. Mature fish

Observed test statistic
R=0.69



Permutation results when there is no difference between groups

$$R = \frac{\left(\bar{r}_{between} - \bar{r}_{within}\right)}{n(n-1)/4}$$

r is a rank similarity between or within groups

Observed test statistic higher than all 1000 results in permutations, so p-value = 0.001. Pattern of detections differs between Young and Mature fish.

Method 2: Euclidean distance of ranks (also 1 DL)

Site	opDDD	ppDDD	opDDE	ppDDE	opDDT	ppDDT	Age	Date
1	12.5	3.5	13.5	3	15.5	6	Young	1996
2	12.5	30	31	26	15.5	31	Mature	1990.5
3	28	27	13.5	32	15.5	20	Mature	1993.5
4	12.5	11.5	13.5	19	15.5	6	Mature	2001.5
5	12.5	3.5	13.5	5.5	15.5	6	Young	2000.5
6	12.5	3.5	13.5	1	15.5	6	Young	1999.5
7	12.5	14	13.5	17	15.5	22.5	Mature	1998

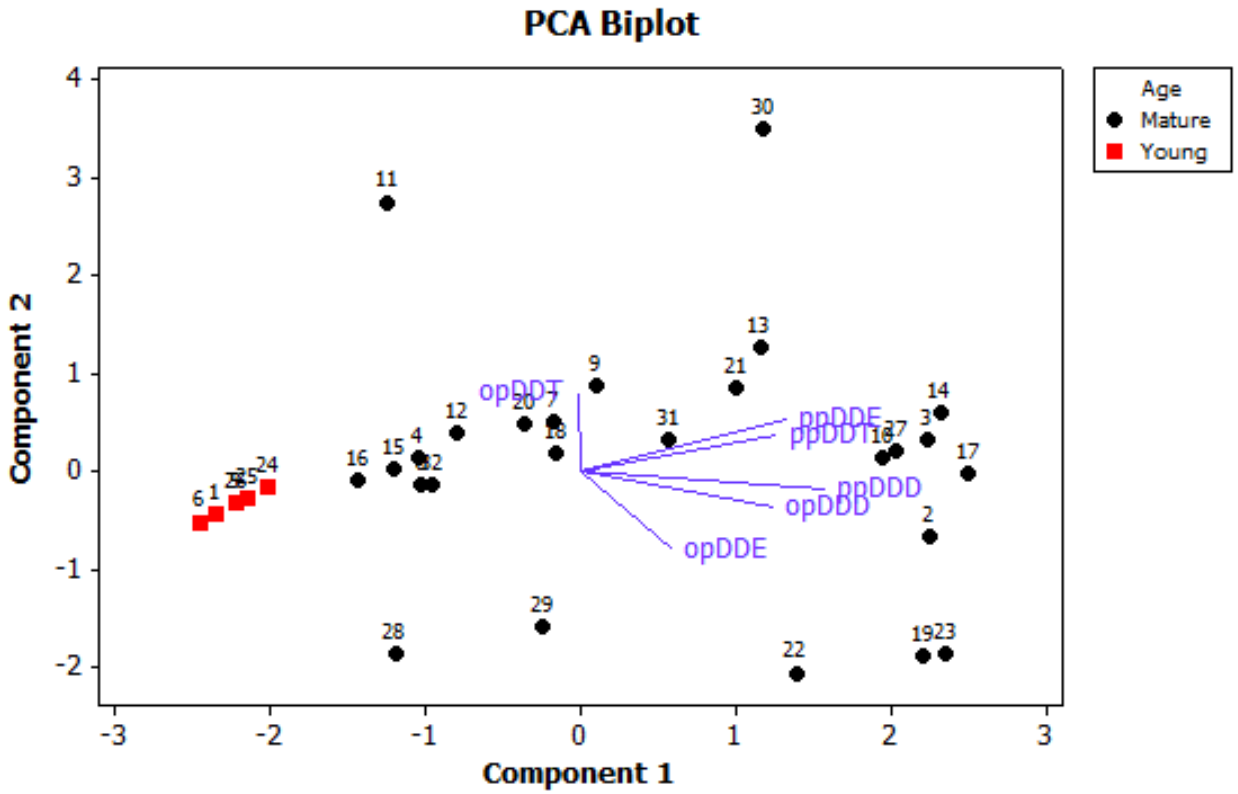
Showing the top 7 rows of ranks within each variable
(Also a Date variable has been added)

Euclidean distance of ranks (1 DL)

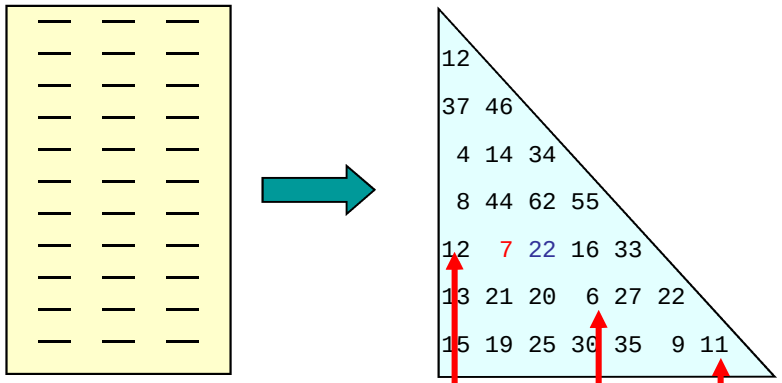
Site	1	2	3	4	5	6	Age	Date
1	*						Young	1996
2	46.5	*					Mature	1990.5
3	42.77	26.69	*				Mature	1993.5
4	17.89	30.37	29.08	*			Mature	2001.5
5	2.5	45.32	41.12	15.69	*		Young	2000.5
6	2	47.52	44.15	19.70	4.5	*	Young	1999.5
7	24.05	26.75	25.31	16.81	22.69	25.27	Mature	1998

top 7 rows of Eucliean distances between ranks

Biplot shows more detail than binary map did



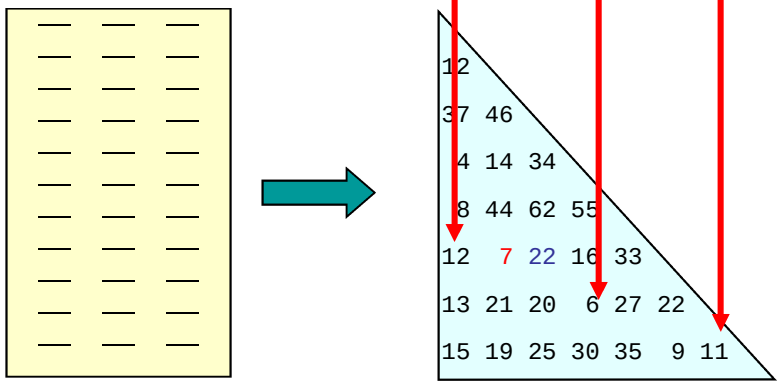
Test for seriation



Time of Sampling

Δ Time Steps

For trend, the explanatory matrix contains the number of time steps between measurements



Ranks of Environmental Data

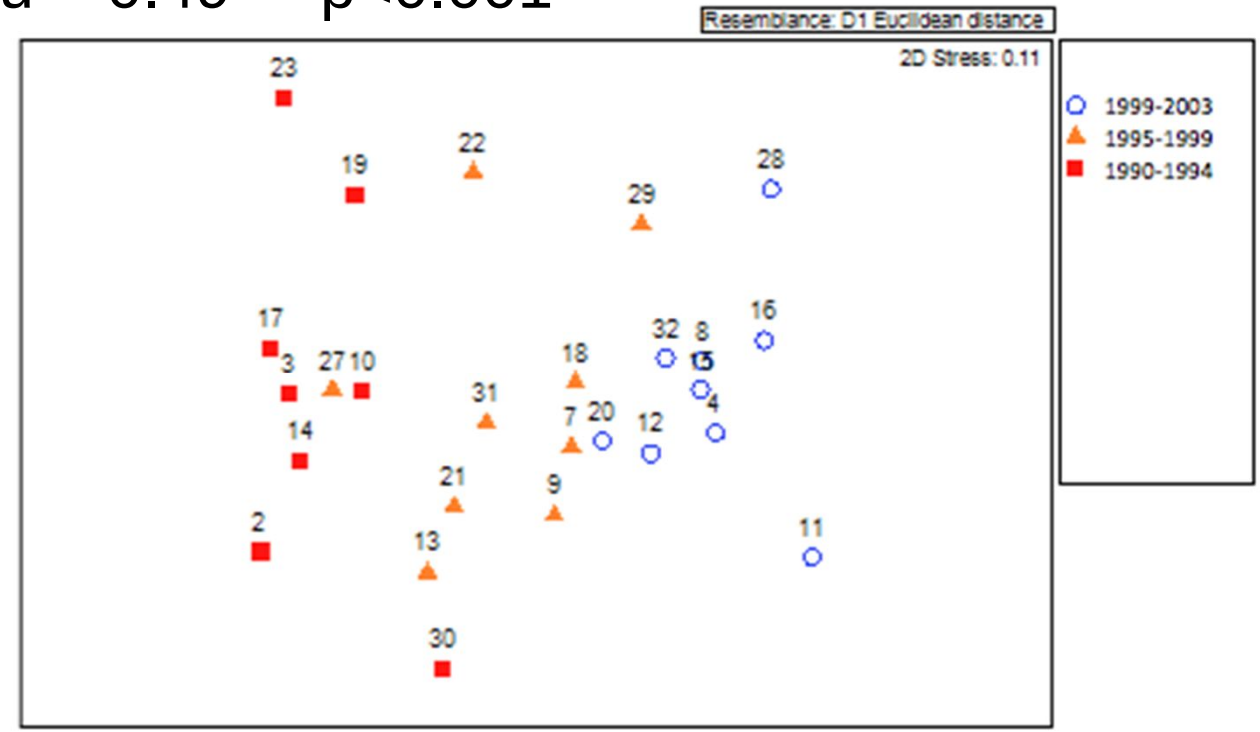
Kendall's tau similarity matrix

Kendall's tau between two triangular matrices computed as test statistic

Test for seriation

(using only mature fish)

tau = 0.49 p<0.001



trend: similar data are most similar in time

3. Methods for >1 RL

- Can account for <1, <5 and detected 2,3,4
- Can differentiate <DL from <QL
- Code data in interval format
- Compute uscore: the sum of the algebraic sign of differences comparing the ith observation to all other observations within the same variable

$$u_i = \sum_{i \neq k} sign(x_i - x_k)$$

uscore

- used in many nonparametric tests: sign test, Mann-Whitney, etc
- used in genomics to rate gene trait relationships
- a (0 to <2) is tied with a (0 to <5), but is lower than a (2 to <5)
- "remarked data" reported as "3J" or "4E" coded as (2 to <5)

DDT data with altered RL

Site	opDDD	ppDDD	opDDE	ppDDE	opDDT	ppDDT	Age	Date
1	0-5	0-5	0-5	14	0-5	0-5	Young	1996
2	0-5	42	8.4	130	0-5	31	Mature	1990.5
3	5.3	38	0-5	250	0-5	11	Mature	1993.5
4	0-2	12	0-2	57	0-2	0-2	Mature	2001.5
5	0-2	0-2	0-2	16	0-2	0-2	Young	2000.5
6	0-2	0-2	0-2	0-2	0-2	0-2	Young	1999.5
7	2-5	14	2-5	52	2-5	14	Mature	1998

After mid 1996 DL=2 used to report values instead of QL=5. Data coded (0-5) now is either (0-2) or (2-5)

uscores for the DDT data

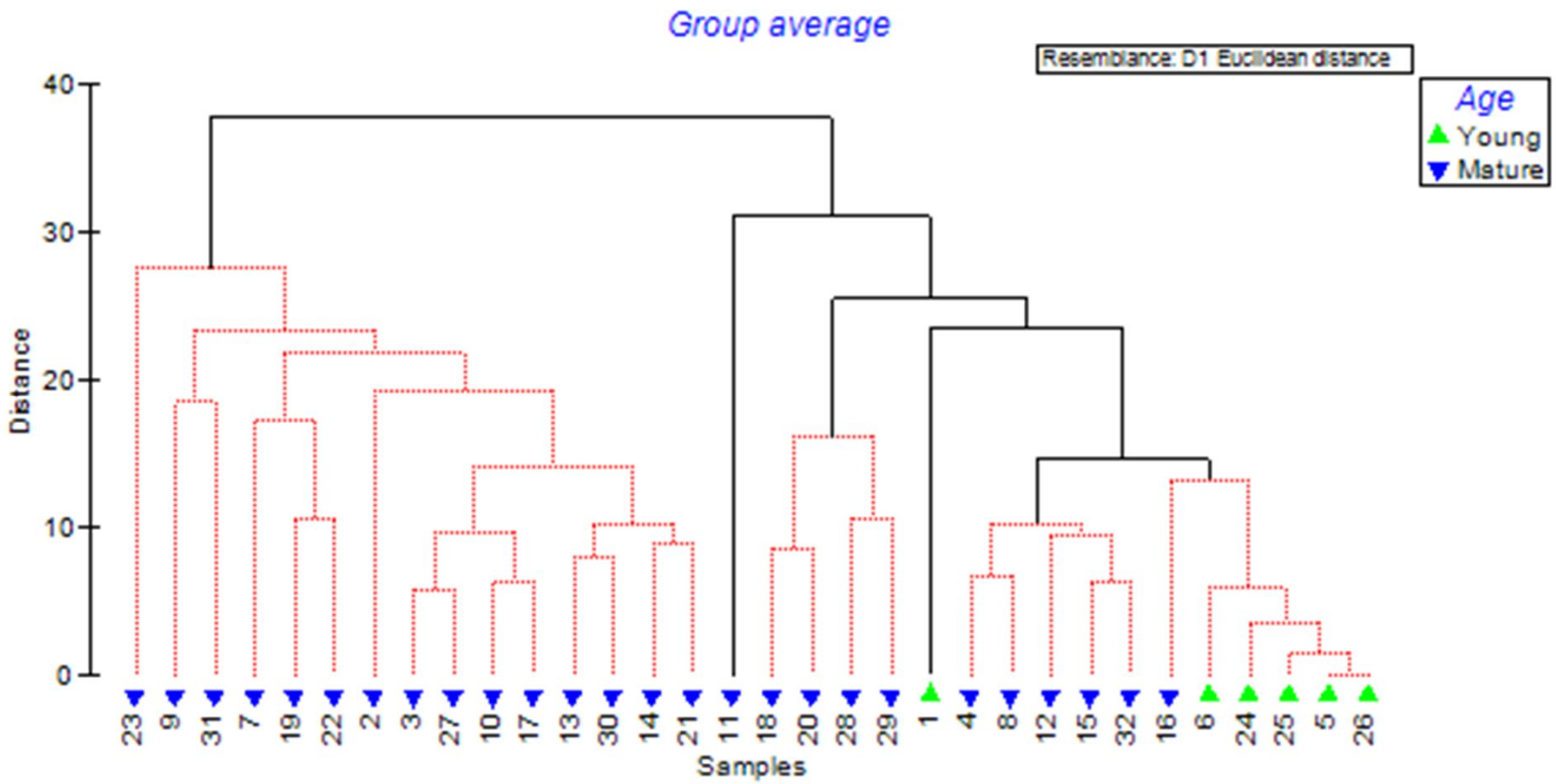
Site	uScore opDDD	uScore ppDDD	uScore opDDE	uScore ppDDE	uScore opDDT	uScore ppDDT	Age	Date
1	-8	-26	-6	-27	-2	-21	Young	1996
2	-8	27	29	19	-2	29	Mature	1990.5
3	23	21	-6	31	-2	7	Mature	1993.5
4	-11	-10	-12	5	-3	-21	Mature	2001.5
5	-11	-26	-12	-22	-3	-21	Young	2000.5
6	-11	-26	-12	-31	-3	-21	Young	1999.5
7	8	-5	5	1	15	12	Mature	1998

opDDD for site 1 (0-5) is known to be lower than only 8 other sites, but the site 4 (0-2) is known to be below 11 sites – is below the three (2-5) values

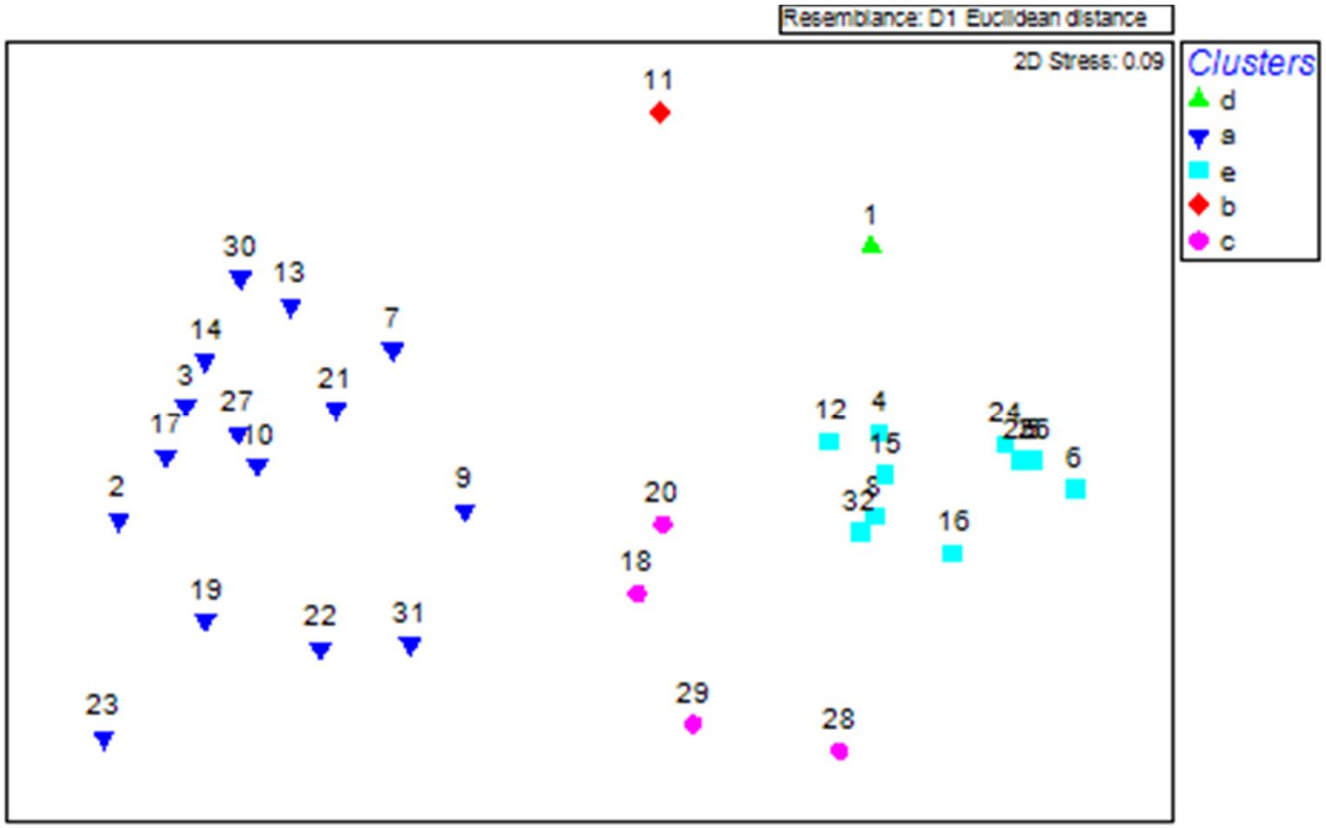
Resemblance matrix from uscores

- rank the uscores within each variable (to get all positive values)
- Euclidean distances on the ranks form the triangular matrix
- Run NMDS, ANOSIM, clustering, test of seriation etc. on the distance matrix

Clustering of sites using uscores



Clusters plotted on NMDS

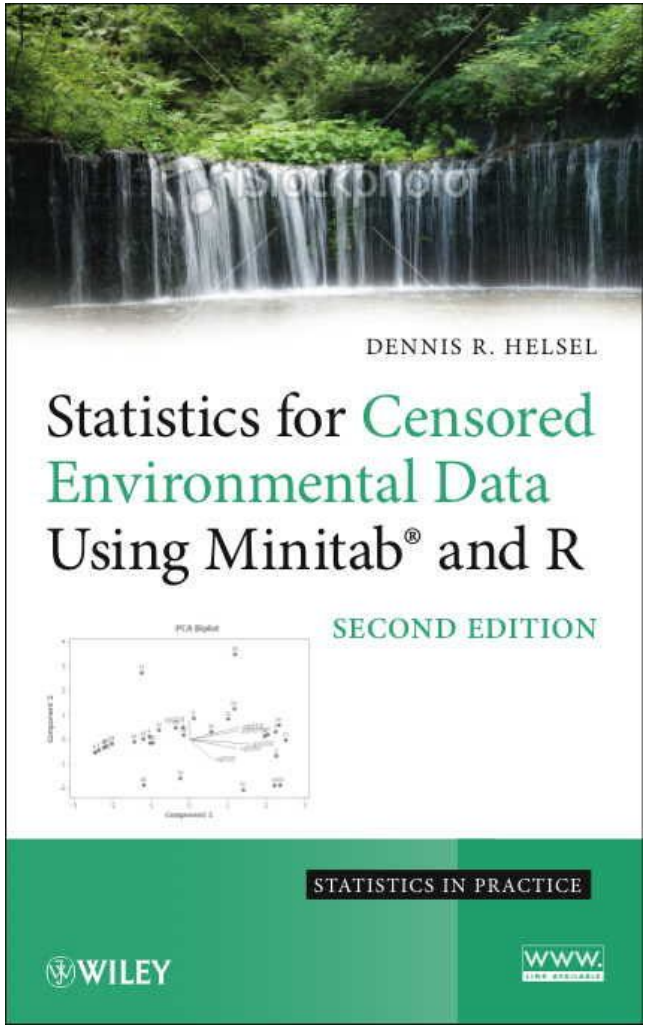


Remember !

- None of the three methods
  1. binary
  2. ranks
  3. uscores

  required bogus substitution of values for nondetects (such as ½ DL)

For more information:

by Dennis R. Helsel
Wiley (2012)

www.PracticalStats.com/nada