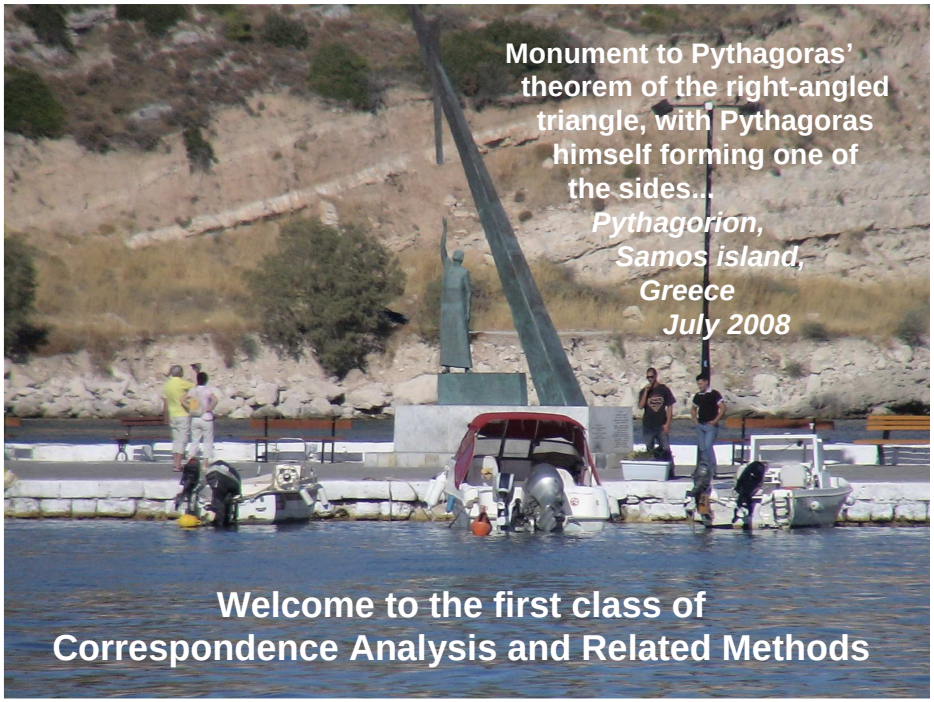


Correspondence Analysis & Related Methods

Department of Statistics, Stanford University

Michael Greenacre
Universitat Pompeu Fabra
Barcelona

michael@upf.es michael.greenacre@gmail.com
www.econ.upf.es/~michael www.globalsong.net

Web of course:
www.econ.upf.edu/~michael/stanford

Information about the course structure

- **Classes: Tuesdays & Thursdays 12h50-14h05, Sequoia 200.**
- **Some reading and/or homework every week.**
- **Homework mainly consists of applying R functions to data sets and interpreting the results-- Naras's course on *Computational Tools for Statistics* is highly recommended. The accent in this course is on learning about tools that can be applied to practical problems. However, some fairly simple theoretical problems will also be included in homework from time to time.**
- **Homework counts 20% towards final grade.**
- **Final exam counts 70% towards final grade.**
- **Class attendance & participation counts 10% towards final grade.**
- **My office hours for seeing students are … (to be negotiated at the first class) and my office is Sequoia 128.**
- **You can use my regular email addresses for communication but I actually prefer to channel all emails about teaching through this address:**

  mmr.upf@gmail.com

  so please try to use that address if possible – this will be more efficient.

Correspondence Analysis & Related Methods

Michael Greenacre

SESSION 1:

UNIVARIATE AND BIVARIATE SUMMARIES

MEASUREMENT SCALES, TRANSFORMATION, STANDARDIZATION

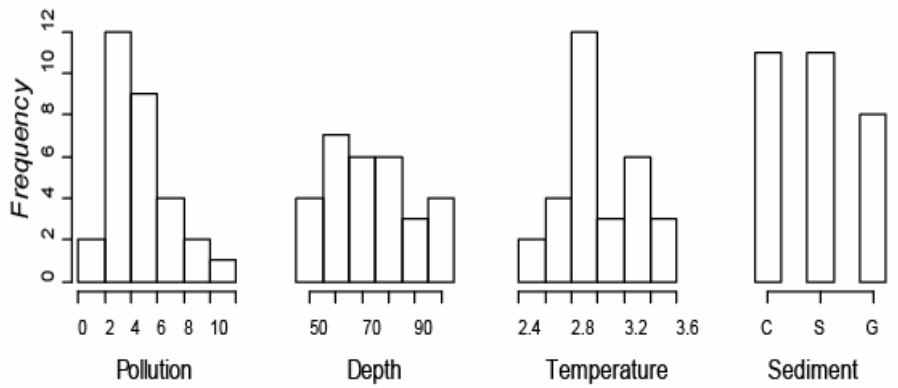
MULTIVARIATE DISTANCE BETWEEN SAMPLING UNITS

Four ecological variables (n=30 sites)

3 continuous variables, 1 categorical variable

SITE NO.	ENVIRONMENTAL VARIABLES			
	Pollution	Depth	Temperature	Sediment
s1	4.8	72	3.5	S
s2	2.8	75	2.5	C
s3	5.4	59	2.7	C
s4	8.2	64	2.9	S
s5	3.9	61	3.1	C
s6	2.6	94	3.5	G
s7	4.6	53	2.9	S
s8	5.1	61	3.3	C
s9	3.9	68	3.4	C
s10	10.0	69	3.0	S
s11	6.5	57	3.3	C
s12	3.8	84	3.1	S
s13	9.4	53	3.0	S
s14	4.7	83	2.5	C
s15	6.7	100	2.8	C
s16	2.8	84	3.0	G
s17	6.4	96	3.1	C
s18	4.4	74	2.8	G
s19	3.1	79	3.6	S
s20	5.6	73	3.0	S
s21	4.3	59	3.4	C
s22	1.9	54	2.8	S
s23	2.4	95	2.9	G
s24	4.3	64	3.0	C
s25	2.0	97	3.0	G
s26	2.5	78	3.4	S
s27	2.1	85	3.0	G
s28	3.4	92	3.3	G
s29	6.0	51	3.0	S
s30	1.9	99	2.9	G

Distributions



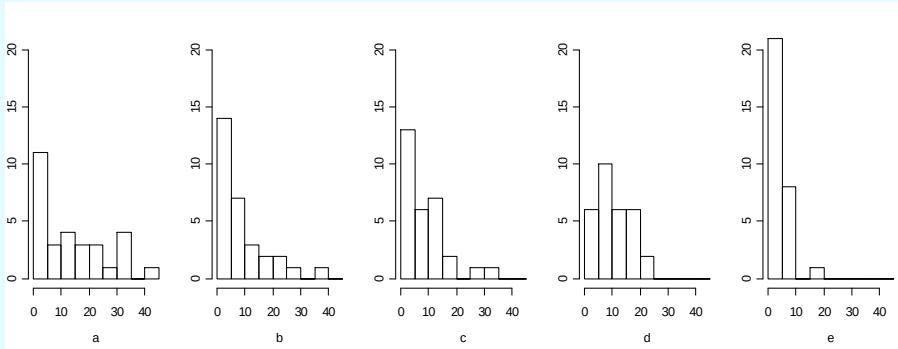
- Histograms of continuous variables: Pollution and Temperature look skew; Depth looks uniform
- Bar-chart of categorical variable: Bars are shown separated for 3 categories of Sediment.

Five species

Count variables (a measurement scale between categorical and continuous)

SITE NO.	SPECIES COUNTS				
	a	b	c	d	e
s1	0	2	9	14	2
s2	26	4	13	11	0
s3	0	10	9	8	0
s4	0	0	15	3	0
s5	13	5	3	10	7
s6	31	21	13	16	5
s7	9	6	0	11	2
s8	2	0	0	0	1
s9	17	7	10	14	6
s10	0	5	26	9	0
s11	0	8	8	6	7
s12	14	11	13	15	0
s13	0	0	19	0	6
s14	13	0	0	9	0
s15	4	0	10	12	0
s16	42	20	0	3	6
s17	4	0	0	0	0
s18	21	15	33	20	0
s19	2	5	12	16	3
s20	0	10	14	9	0
s21	8	0	0	4	6
s22	35	10	0	9	17
s23	6	7	1	17	10
s24	18	12	20	7	0
s25	32	26	0	23	0
s26	32	21	0	10	2
s27	24	17	0	25	6
s28	16	3	12	20	2
s29	11	0	7	8	0
s30	24	37	5	18	1

Five species



- Histograms of count variables: all the distribution are ‘skew to the right’

Questions

SITE	SPECIES COUNTS					ENVIRONMENTAL VARS			
NO.	a	b	c	d	e	Polln	Depth	Temp.	Sedmnt
s1	0	2	9	14	2	4.8	72	3.5	S
s2	26	4	13	11	0	2.8	75	2.5	C
s3	0	10	9	8	0	5.4	59	2.7	C
s4	0	0	15	3	0	8.2	64	2.9	S
s5	13	5	3	10	7	3.9	61	3.1	C
s6	31	21	13	16	5	2.6	94	3.5	G
s7	9	6	0	11	2	4.6	53	2.9	S
s8	2	0	0	0	1	5.1	61	3.3	C
s9	17	7	10	14	6	3.9	68	3.4	C
s10	0	5	26	9	0	10.0	69	3.0	S
s11	0	8	8	6	7	6.5	57	3.3	C
s12	14	11	13	15	0	3.8	84	3.1	S
s13	0	0	19	0	6	9.4	53	3.0	S
s14	13	0	0	9	0	4.7	83	2.5	C
s15	4	0	10	12	0	6.7	100	2.8	C
s16	42	20	0	3	6	2.8	84	3.0	G
s17	4	0	0	0	0	6.4	96	3.1	C
s18	21	15	33	20	0	4.4	74	2.8	G
s19	2	5	12	16	3	3.1	79	3.6	S
s20	0	10	14	9	0	5.6	73	3.0	S
s21	8	0	0	4	6	4.3	59	3.4	C
s22	35	10	0	9	17	1.9	54	2.8	S
s23	6	7	1	17	10	2.4	95	2.9	G
s24	18	12	20	7	0	4.3	64	3.0	C
s25	32	26	0	23	0	2.0	97	3.0	G
s26	32	21	0	10	2	2.5	78	3.4	S
s27	24	17	0	25	6	2.1	85	3.0	G
s28	16	3	12	20	2	3.4	92	3.3	G
s29	11	0	7	8	0	6.0	51	3.0	S
s30	24	37	5	18	1	1.9	99	2.9	G

1. How are pollution, depth, temperature and sediment inter-related? ... *correlations, y = f(x), x = f(y) … environmental classes, environmental gradients …*

1. How are species interrelated? ... *associations (correlations?) between species … species classes, species gradients …*

2. How are species related to environmental variables? ...

   (a,b,c,d,e) = f(p,d,t,s)

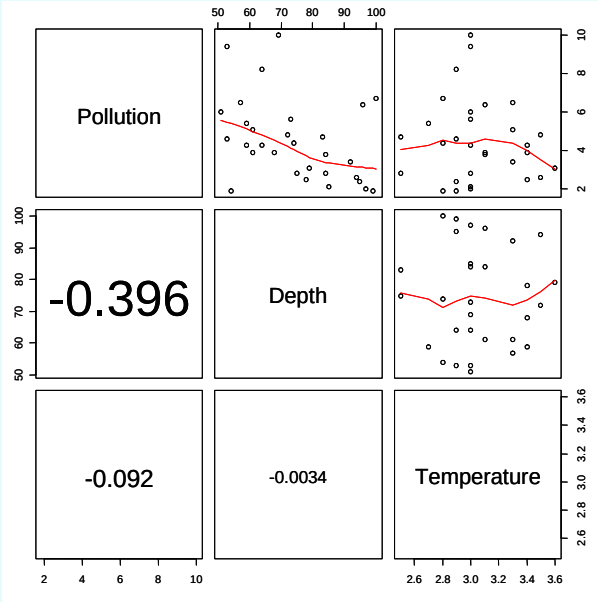
4. How can we cope with many species and many environmental variables?

Scatterplots: continuous variables

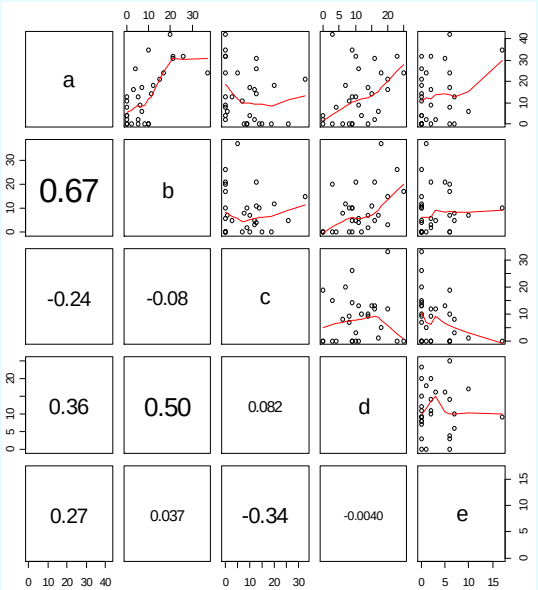
Showing smoothed relationship between the variables ("scatterplot smoother").

The correlations are shown opposite their scatterplot. Assuming normality, the critical value at 1% level (two-sided test) is ±0.347

pairs function in R

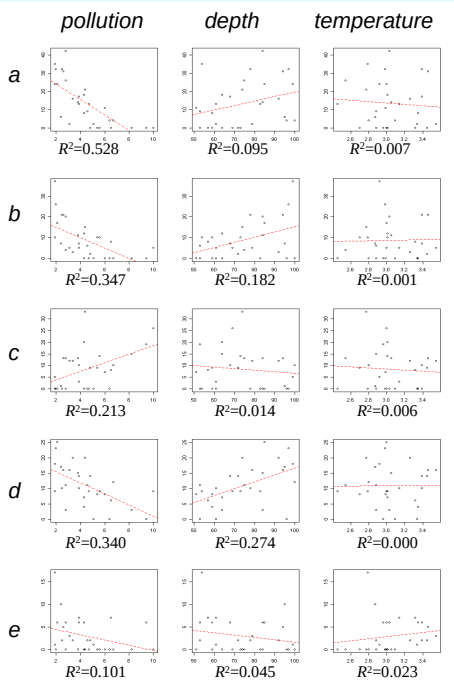


Scatterplots: count variables

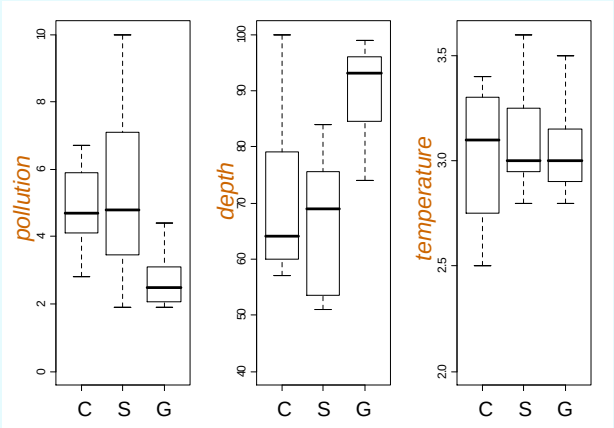


Scatterplots: count vs continuous

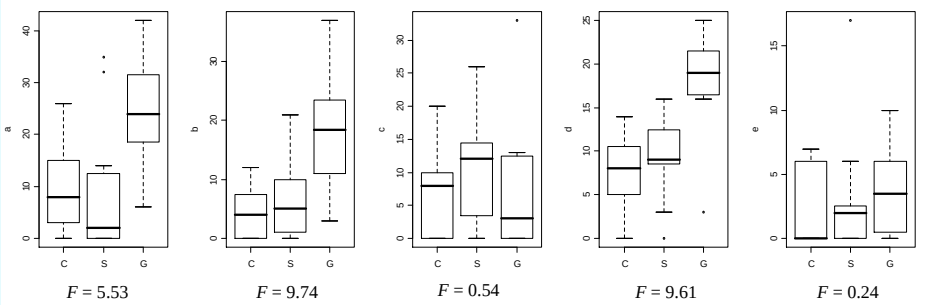
Showing coefficients of determination (measure of explained variance, usually expressed as a percentage) – in this simple linear regression case R^2 is the square of r, the correlation coefficient. The sign of r can be deduced by the slope of the regression line (if 0.347 was the critical point at 1% level for a significant r, then $0.347^2 = 0.120$ is the critical point for R^2.



Box-and-whisker plots: continuous vs categorical



Box-and-whisker plots: count vs categorical

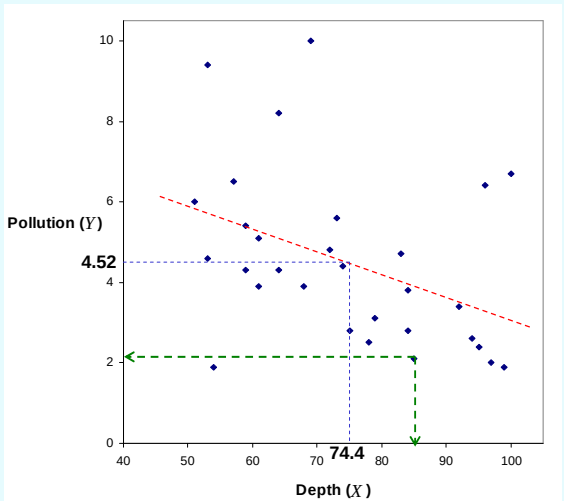


$F = 5.53$ $F = 9.74$ $F = 0.54$ $F = 9.61$ $F = 0.24$

Showing F statistics for the analysis of variance (ANOVA), which tests the difference in group means. In this case the critical point of the test at the 1% level is 5.49.

Relationship between gradients

Polln	Depth
4.8	72
2.8	75
5.4	59
8.2	64
3.9	61
2.6	94
4.6	53
5.1	61
3.9	68
10.0	69
6.5	57
3.8	84
9.4	53
4.7	83
6.7	100
2.8	84
6.4	96
4.4	74
3.1	79
5.6	73
4.3	59
1.9	54
2.4	95
4.3	64
2.0	97
2.5	78
2.1	85
3.4	92
6.0	51
1.9	99



Remember that the regression line is fitted by minimizing the sum of squared distances between the points and the line, parallel to the response variable.

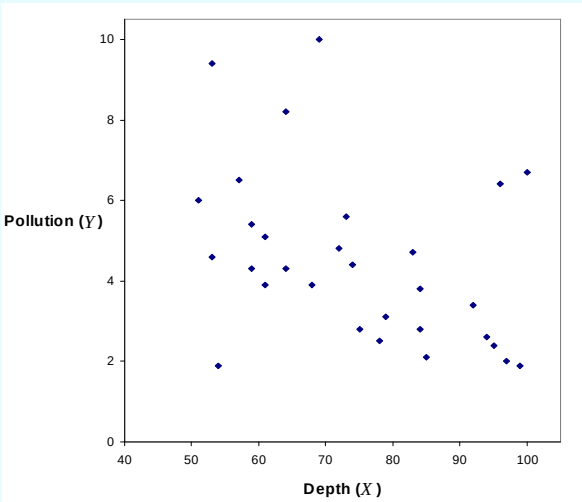
The means of the two variables lie exactly on the regression line.

$r = -0.396$ $y = 8.554 - 0.0542\,x$ $R^2 = 0.156$ $P = 0.022$

$x = 87.46 - 2.884\,y$

Scale and variance

Polln	Depth
4.8	72
2.8	75
5.4	59
8.2	64
3.9	61
2.6	94
4.6	53
5.1	61
3.9	68
10.0	69
6.5	57
3.8	84
9.4	53
4.7	83
6.7	100
2.8	84
6.4	96
4.4	74
3.1	79
5.6	73
4.3	59
1.9	54
2.4	95
4.3	64
2.0	97
2.5	78
2.1	85
3.4	92
6.0	51
1.9	99



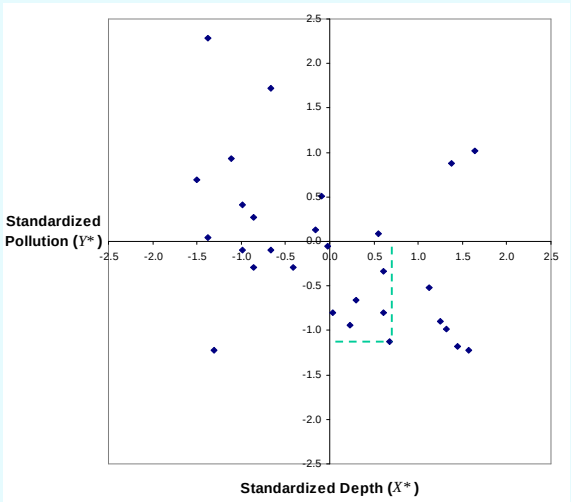
Since multivariate methods rely heavily on the concept of ***distance***, we need to be aware (and beware!) of the ***scales*** of variables. In this example the distance between points would be totally dominated by the variable 'depth'

Standardizing data

y	x	y*	x*
4.8	72	0.132	-0.156
2.8	75	-0.802	0.036
5.4	59	0.413	-0.988
8.2	64	1.720	-0.668
3.9	61	-0.288	-0.860
2.6	94	-0.895	1.253
4.6	53	0.039	-1.373
5.1	61	0.272	-0.860
3.9	68	-0.288	-0.412
10.0	69	2.561	-0.348
6.5	57	0.926	-1.116
3.8	84	-0.335	0.613
9.4	53	2.281	-1.373
4.7	83	0.086	0.549
6.7	100	1.020	1.637
2.8	84	-0.802	0.613
6.4	96	0.880	1.381
4.4	74	-0.054	-0.028
3.1	79	-0.662	0.292
5.6	73	0.506	-0.092
4.3	59	-0.101	-0.988
1.9	54	-1.222	-1.309
2.4	95	-0.989	1.317
4.3	64	-0.101	-0.668
2.0	97	-1.175	1.445
2.5	78	-0.942	0.228
2.1	85	-1.129	0.677
3.4	92	-0.522	1.125
6.0	51	0.693	-1.501
1.9	99	-1.222	1.573

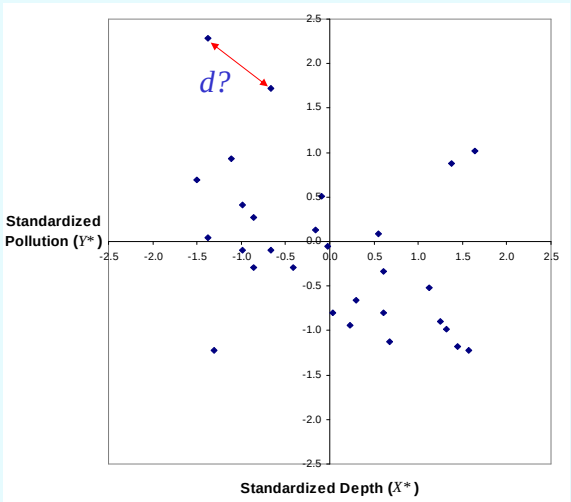
$$[y - \text{mean}(y)]/\text{sd}(y) = [y - 4.52]/2.14 = y^*$$
$$[x - \text{mean}(x)]/\text{sd}(x) = [x - 74.4]/15.6 = x^*$$

Standardized plot



$$x^* = [x - \text{mean}(x)]/\text{sd}(x) = [x - 6.59]/2.56$$
$$y^* = [y - \text{mean}(y)]/\text{sd}(y) = [y - 74.4]/15.6$$

Standardized distance



$$x^* = [x - \text{mean}(x)]/\text{sd}(x) = [x - 6.59]/2.56$$
$$y^* = [y - \text{mean}(y)]/\text{sd}(y) = [y - 74.4]/15.6$$

(Weighted) Euclidean distance

distance between rows (sites) i and i'

$$d(i,i') = \sqrt{(x_i^* - x_{i'}^*)^2 + (y_i^* - y_{i'}^*)^2}$$

$$= \sqrt{\left(\frac{x_i - \bar{x}}{s_x} - \frac{x_{i'} - \bar{x}}{s_x}\right)^2 + \left(\frac{y_i - \bar{y}}{s_y} - \frac{y_{i'} - \bar{y}}{s_y}\right)^2}$$

$$= \sqrt{\left(\frac{x_i}{s_x} - \frac{x_{i'}}{s_x}\right)^2 + \left(\frac{y_i}{s_y} - \frac{y_{i'}}{s_y}\right)^2}$$

$$= \sqrt{\frac{1}{s_x^2}(x_i - x_{i'})^2 + \frac{1}{s_y^2}(y_i - y_{i'})^2}$$

"weights", inverses of variances

For more than two variables, just add similar terms for the other variables

Inter-site distances

for the three continuous variables pollution, depth and temperature

$$d(i,i') = \sqrt{\frac{1}{s_x^2}\left(x_i - x_{i'}\right)^2 + \frac{1}{s_y^2}\left(y_i - y_{i'}\right)^2 + \frac{1}{s_z^2}\left(z_i - z_{i'}\right)^2}$$

	s1	s2	s3	s4	s5	s6	· · ·	s24	s25	s26	s27	s28	s29
s2	3.681												
s3	2.977	1.741											
s4	2.708	2.980	1.523										
s5	1.642	2.371	1.591	2.139									
s6	1.744	3.759	3.850	3.884	2.619								
s7	2.458	2.171	0.890	1.823	0.935	3.510							
.	.	.	.	.	.	.	. .						
.	.	.	.	.	.	.	. .						
s25	2.727	2.299	3.095	3.602	2.496	1.810	· · ·	2.371					
s26	1.195	3.209	3.084	3.324	1.658	1.086	· · ·	1.880	1.886				
s27	2.333	1.918	2.507	3.170	1.788	1.884	· · ·	1.692	0.770	1.503			
s28	1.604	3.059	3.145	3.204	2.122	0.813	· · ·	2.128	1.291	1.052	1.307		
s29	2.299	2.785	1.216	1.369	1.224	3.642	· · ·	1.150	3.488	2.772	2.839	3.083	
s30	3.062	2.136	3.121	3.699	2.702	2.182	· · ·	2.531	0.381	2.247	0.969	1.648	3.639

Distance between rows of matrix $\mathbf{X} = [x_{ij}]$

$$d(i,i') = \sqrt{\sum_{j=1}^{m} \frac{1}{s_j^2}\left(x_{ij} - x_{i'j}\right)^2}$$

Bray-Curtis index of dissimilarity

SITE NO.	SPECIES COUNTS a	b	c	d	e
s1	0	2	9	14	2
s2	26	4	13	11	0
s3	0	10	9	8	0
s4	0	0	15	3	0
s5	13	5	3	10	7
s6	31	21	13	16	5
s7	9	6	0	11	2
s8	2	0	0	0	1
s9	17	7	10	14	6
s10	0	5	26	9	0
s11	0	8	8	6	7
s12	14	11	13	15	0
s13	0	0	19	0	6
s14	13	0	0	9	0
s15	4	0	10	12	0
s16	42	20	0	3	6
s17	4	0	0	0	0
s18	21	15	33	20	0
s19	2	5	12	16	3
s20	0	10	14	9	0
s21	8	0	0	4	6
s22	35	10	0	9	17
s23	6	7	1	17	10
s24	18	12	20	7	0
s25	32	26	0	23	0
s26	32	21	0	10	2
s27	24	17	0	25	6
s28	16	3	12	20	2
s29	11	0	7	8	0
s30	24	37	5	18	1

All the species variables are on the same scale (counts, or frequencies) but there is still a problem of standardization which we shall temporarily ignore...

The Bray-Curtis index is one of the most popular measures of intersample dissimilarity used by ecologists. It inherently assumes that the samples are from the same physically sized sample (area of plot, volume of grab...)

For example, there are

- 0 +2+ 9 +14+2 = 27 species observed in site 1
- 26+4+13+11+0 = 54 species observed in site 2
- 26+2+ 4 + 3 +2 = 37 absolute differences between the two sites
- Bray-Curtis index between sites 1 and 2 = 37/(27+54)
  = 0.4568

usually expressed as a percentage, that is 45.7%

Bray-Curtis index of dissimilarity

SITE NO.	SPECIES COUNTS a	b	c	d	e
s1	0	2	9	14	2
s2	26	4	13	11	0
s3	0	10	9	8	0
s4	0	0	15	3	0
s5	13	5	3	10	7
s6	31	21	13	16	5
s7	9	6	0	11	2
s8	2	0	0	0	1
s9	17	7	10	14	6
s10	0	5	26	9	0
s11	0	8	8	6	7
s12	14	11	13	15	0
s13	0	0	19	0	6
s14	13	0	0	9	0
s15	4	0	10	12	0
s16	42	20	0	3	6
s17	4	0	0	0	0
s18	21	15	33	20	0
s19	2	5	12	16	3
s20	0	10	14	9	0
s21	8	0	0	4	6
s22	35	10	0	9	17
s23	6	7	1	17	10
s24	18	12	20	7	0
s25	32	26	0	23	0
s26	32	21	0	10	2
s27	24	17	0	25	6
s28	16	3	12	20	2
s29	11	0	7	8	0
s30	24	37	5	18	1

Suppose data matrix is **N** with general element n_{ij}.

A dot (·) in the subindex indicates summation over that index, so $n_{i\cdot}$ is the sum of the i-th row.

Then the Bray-Curtis dissimilarity is defined as:

$$d(i,i') = \frac{\sum_j |n_{ij} - n_{i'j}|}{n_{i\cdot} + n_{i'\cdot}}$$

It is clear that species e, for example, which is rare, will not contribute much to the Bray-Curtis index, since the differences between these low values are also going to be low, whereas the differences between frequent species will be generally high.

Often this problem is alleviated by taking square roots or even fourth roots of the frequencies: $\tilde{n}_{ij} = n_{ij}^{0.25}$ and then calculating the Bray-Curtis on the transformed data

Bray-Curtis dissimilarities

	s1	s2	s3	s4	s5	s6	· · ·	s24	s25	s26	s27	s28	s29
s2	45.7												
s3	29.6	48.1											
s4	46.7	55.6	46.7										
s5	47.7	34.8	50.8	78.6									
s6	52.2	22.9	52.2	69.2	41.9								
s7	45.5	41.5	49.1	87.0	21.2	50.9							
.	.	.	.	.	.	.	. .						
.	.	.	.	.	.	.	. .						
.	.	.	.	.	.	.	. .						
s25	70.4	39.3	66.7	93.9	52.9	18.6	· · ·	46.4					
s26	69.6	32.8	60.9	92.8	41.7	15.2	· · ·	39.3	13.7				
s27	63.6	38.1	63.6	93.3	38.2	21.5	· · ·	42.6	16.3	22.6			
s28	32.5	21.5	50.0	57.7	31.9	29.5	· · ·	30.9	41.8	47.5	34.4		
s29	43.4	35.0	43.4	54.5	31.2	53.6	· · ·	39.8	64.5	58.2	61.2	34.2	
s30	60.7	36.7	58.9	84.5	48.0	21.6	· · ·	40.8	18.1	25.3	23.6	37.7	56.8

Chi-square distance

SITE	a	b	c	d	e	sum
s1	0	2	9	14	2	27
s2	26	4	13	11	0	54
s3	0	10	9	8	0	27
s4	0	0	15	3	0	18
s5	13	5	3	10	7	38
s6	31	21	13	16	5	86
s7	9	6	0	11	2	28
s8	2	0	0	0	1	3
s9	17	7	10	14	6	54
s10	0	5	26	9	0	40
s11	0	8	8	6	7	29
s12	14	11	13	15	0	53
s13	0	0	19	0	6	25
s14	13	0	0	9	0	22
s15	4	0	10	12	0	26
s16	42	20	0	3	6	71
s17	4	0	0	0	0	4
s18	21	15	33	20	0	89
s19	2	5	12	16	3	38
s20	0	10	14	9	0	33
s21	8	0	0	4	6	18
s22	35	10	0	9	17	71
s23	6	7	1	17	10	41
s24	18	12	20	7	0	57
s25	32	26	0	23	0	81
s26	32	21	0	10	2	65
s27	24	17	0	25	6	72
s28	16	3	12	20	2	53
s29	11	0	7	8	0	26
s30	24	37	5	18	1	85
sum	404	262	252	327	89	1334

SITE	a	b	c	d	e
s1	0.000	0.074	0.333	0.519	0.074
s2	0.481	0.074	0.241	0.204	0.000
s3	0.000	0.370	0.333	0.296	0.000
s4	0.000	0.000	0.833	0.167	0.000
s5	0.342	0.132	0.079	0.263	0.184
s6	0.360	0.244	0.151	0.186	0.058
s7	0.321	0.214	0.000	0.393	0.071
s8	0.667	0.000	0.000	0.000	0.333
s9	0.315	0.130	0.185	0.259	0.111
s10	0.000	0.125	0.650	0.225	0.000
s11	0.000	0.276	0.276	0.207	0.241
s12	0.264	0.208	0.245	0.283	0.000
s13	0.000	0.000	0.760	0.000	0.240
s14	0.591	0.000	0.000	0.409	0.000
s15	0.154	0.000	0.385	0.462	0.000
s16	0.592	0.282	0.000	0.042	0.085
s17	1.000	0.000	0.000	0.000	0.000
s18	0.236	0.169	0.371	0.225	0.000
s19	0.053	0.132	0.316	0.421	0.079
s20	0.000	0.303	0.424	0.273	0.000
s21	0.444	0.000	0.000	0.222	0.333
s22	0.493	0.141	0.000	0.127	0.239
s23	0.146	0.171	0.024	0.415	0.244
s24	0.316	0.211	0.351	0.123	0.000
s25	0.395	0.321	0.000	0.284	0.000
s26	0.492	0.323	0.000	0.154	0.031
s27	0.333	0.236	0.000	0.347	0.083
s28	0.302	0.057	0.226	0.377	0.038
s29	0.423	0.000	0.269	0.308	0.000
s30	0.282	0.435	0.059	0.212	0.012
ave.	0.303	0.196	0.189	0.245	0.067

The chi-square distance is an alternative measure of dissimilarity that is applicable to frequencies and which is the basis of correspondence analysis (CA).

First convert the rows of frequencies into ***profiles***, that is the frequencies relative to their totals. The row of column sums is also expressed as proportions, called the ***average profile***.

Chi-square distance

SITE	a	b	c	d	e
s1	0.000	0.074	0.333	0.519	0.074
s2	0.481	0.074	0.241	0.204	0.000
s3	0.000	0.370	0.333	0.296	0.000
s4	0.000	0.000	0.833	0.167	0.000
s5	0.342	0.132	0.079	0.263	0.184
s6	0.360	0.244	0.151	0.186	0.058
s7	0.321	0.214	0.000	0.393	0.071
s8	0.667	0.000	0.000	0.000	0.333
s9	0.315	0.130	0.185	0.259	0.111
s10	0.000	0.125	0.650	0.225	0.000
s11	0.000	0.276	0.276	0.207	0.241
s12	0.264	0.208	0.245	0.283	0.000
s13	0.000	0.000	0.760	0.000	0.240
s14	0.591	0.000	0.000	0.409	0.000
s15	0.154	0.000	0.385	0.462	0.000
s16	0.592	0.282	0.000	0.042	0.085
s17	1.000	0.000	0.000	0.000	0.000
s18	0.236	0.169	0.371	0.225	0.000
s19	0.053	0.132	0.316	0.421	0.079
s20	0.000	0.303	0.424	0.273	0.000
s21	0.444	0.000	0.000	0.222	0.333
s22	0.493	0.141	0.000	0.127	0.239
s23	0.146	0.171	0.024	0.415	0.244
s24	0.316	0.211	0.351	0.123	0.000
s25	0.395	0.321	0.000	0.284	0.000
s26	0.492	0.323	0.000	0.154	0.031
s27	0.333	0.236	0.000	0.347	0.083
s28	0.302	0.057	0.226	0.377	0.038
s29	0.423	0.000	0.269	0.308	0.000
s30	0.282	0.435	0.059	0.212	0.012
ave.	0.303	0.196	0.189	0.245	0.067

Calculate the chi-square distance between the first two sites as follows:

$$d(1,2)^2 = \frac{(0-0.481)^2}{0.303} + \frac{(0.074-0.074)^2}{0.196} + \frac{(0.333-0.241)^2}{0.189} + \frac{(0.519-0.204)^2}{0.245} + \frac{(0.074-0)^2}{0.067}$$

$$= 1.297$$

Therefore

$$d(1,2) = 1.139$$

In general, for profiles x_{ij} (sample i, species j) with average profile $x_{.j}$ (for species j), the chi-square distance between samples is:

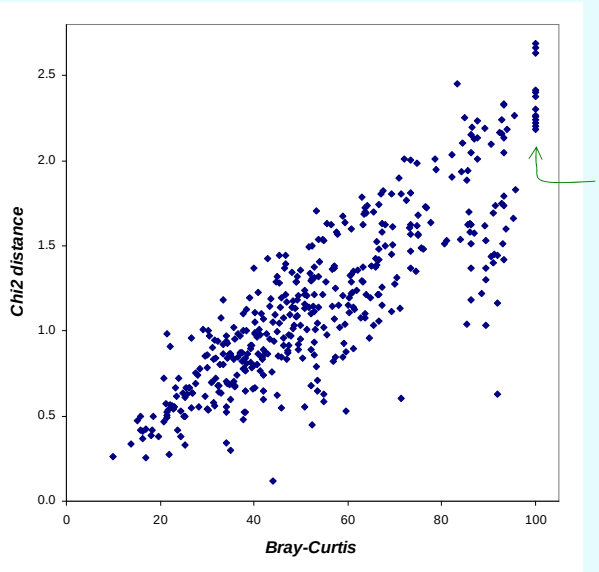
$$d(i,i') = \sqrt{\sum_j \frac{(x_{ij} - x_{i'j})^2}{x_{.j}}}$$

standardization using average

Chi-square distance

	s1	s2	s3	s4	s5	s6	...	s24	s25	s26	s27	s28	s29
s2	1.139												
s3	0.855	1.137											
s4	1.392	1.630	1.446										
s5	1.093	0.862	1.238	2.008									
s6	1.099	0.539	0.887	1.802	0.597								
s7	1.046	0.845	1.081	2.130	0.573	0.555							
...													
s25	1.312	0.817	1.057	2.185	0.858	0.495	...	0.917					
s26	1.508	0.805	1.224	2.241	0.834	0.475	...	0.915	0.338				
s27	1.100	0.837	1.078	2.136	0.520	0.489	...	0.983	0.412	0.562			
s28	0.681	0.504	0.954	1.572	0.724	0.613	...	0.699	0.844	0.978	0.688		
s29	0.951	0.296	1.145	1.535	0.905	0.708	...	0.662	0.956	1.021	0.897	0.340	
s30	1.330	0.986	0.846	2.101	0.970	0.535	...	0.864	0.388	0.497	0.617	1.001	1.142

Chi-square distance vs. Bray-Curtis



Notice the Bray-Curtis values of 100 (maximum dissimilarity, no overlap at all), but the chi-square distance spreads them out

Presence/absence data

SITES	SPECIES S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
A	1	1	1	0	1	0	0	1	1	1
B	1	1	0	1	1	0	0	0	0	1
C	0	1	1	0	1	0	0	1	0	0
D	0	0	0	1	0	1	0	0	0	0
E	1	1	1	0	1	0	1	1	1	0
F	0	1	0	1	1	0	0	0	0	1
G	0	1	1	0	1	1	0	1	1	0

A \ B	1	0	
1	4	3	7
0	1	2	3
	5	5	10

Similarity $s_{12} = 6/10 = 0.6$

Dissimilarity $d_{12} = 4/10 = 0.4 = 1 - s_{12}$

In general there are n sites, p species, and we can count combinations for each pair of sites i and i' as:

i \ i'	1	0	
1	a	b	$a+b$
0	c	d	$c+d$
	$a+c$	$b+d$	p

Similarity $s_{ii'} = (a+d)/p$

Dissimilarity $d_{ii'} = 1 - s_{ii'} = (b+c)/p$

Presence/absence data

SITES	SPECIES S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
A	1	1	1	0	1	0	0	1	1	1
B	1	1	0	1	1	0	0	0	0	1
C	0	1	1	0	1	0	0	1	0	0
D	0	0	0	1	0	1	0	0	0	0
E	1	1	1	0	1	0	1	1	1	0
F	0	1	0	1	1	0	0	0	0	1
G	0	1	1	0	1	1	0	1	1	0

A \ B	1	0	
1	4	3	7
0	1	2	3
	5	5	10

Similarity $s_{AB} = 6/10 = 0.6$

Dissimilarity $d_{AB} = 4/10 = 0.4 = 1 - s_{AB}$

In general there are n sites, p species, and we can count combinations for each pair of sites i and i' as:

i \ i'	1	0	
1	a	b	$a+b$
0	c	d	$c+d$
	$a+c$	$b+d$	p

Similarity $s_{ii'} = (a+d)/p$

Dissimilarity $d_{ii'} = 1 - s_{ii'} = (b+c)/p$

Presence/absence data

SITES	SPECIES S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
A	1	1	1	0	1	0	0	1	1	1
B	1	1	0	1	1	0	0	0	0	1
C	0	1	1	0	1	0	0	1	0	0
D	0	0	0	1	0	1	0	0	0	0
E	1	1	1	0	1	0	1	1	1	0
F	0	1	0	1	1	0	0	0	0	1
G	0	1	1	0	1	1	0	1	1	0

i \ i'	1	0	
1	a	b	$a+b$
0	c	d	$c+d$
	$a+c$	$b+d$	p

Similarity $s_{ii'} = (a+d)/p$ (*simple matching coefficient)*

Dissimilarity $d_{ii'} = 1 - s_{ii'} = (b+c)/p$ (identical to the squared Euclidean distance between rows)

When common absences are not important, similarity is measured by the *Jaccard index* $c_{ii'} = a/(p-d)$; that is, drop the species absent in both sites and recompute the matching coefficient.

Similarity (Jaccard index) $c_{AB} = 4/8 = 0.5$

Dissimilarity (Jaccard) $d_{AB} = 4/8 = 0.45 = 1 - c_{AB}$

Presence/absence data

Jaccard dissimilarities

samples	A	B	C	D	E	F	G
A	0	0.5000	0.4286	1.0000	0.2500	0.6250	0.3750
B	0.5000	0	0.7143	0.8333	0.6667	0.2000	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.6667	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8000	0.8571
E	0.2500	0.6667	0.4286	1.0000	0	0.7778	0.3750
F	0.6250	0.2000	0.6667	0.8000	0.7778	0	0.7500
G	0.3750	0.7778	0.3333	0.8571	0.3750	0.7500	0

Dealing with heterogeneous data

	Continuous variables			*Discrete variables*	
Station	Depth	Temperature	Salinity	Region	Substrate
s3	30	3.15	33.52	Ta	Si/St
s8	29	3.15	33.52	Ta	Cl/Gr
s25	30	3.00	33.45	Sk	Cl/Sa
⋮	⋮	⋮	⋮	⋮	⋮
s84	66	3.22	33.48	St	Cl

Coding of categorical data as dummy variables

	Continuous variables			*Sampled region*				*Substrate character*				
Station	Depth	Temperature	Salinity	Tarehola	Skognes	Njosken	Storura	Clay	Silt	Sand	Gravel	Stone
s3	30	3.15	33.52	1	0	0	0	0	1	0	0	1
s8	29	3.15	33.52	1	0	0	0	1	0	0	1	0
s25	30	3.00	33.45	0	1	0	0	1	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
s84	66	3.22	33.48	0	0	0	1	1	0	0	0	0
mean	58.15	3.086	33.50	0.242	0.273	0.242	0.242	0.606	0.152	0.364	0.182	0.061
s.d.	32.45	0.100	0.076	0.435	0.452	0.435	0.435	0.496	0.364	0.489	0.392	0.242

Dealing with heterogeneous data

	Continuous variables			*Sampled region*				*Substrate character*				
Station	Depth	Temperature	Salinity	Tarehola	Skognes	Njosken	Storura	Clay	Silt	Sand	Gravel	Stone
s3	30	3.15	33.52	1	0	0	0	0	1	0	0	1
s8	29	3.15	33.52	1	0	0	0	1	0	0	1	0
s25	30	3.00	33.45	0	1	0	0	1	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
s84	66	3.22	33.48	0	0	0	1	1	0	0	0	0
mean	58.15	3.086	33.50	0.242	0.273	0.242	0.242	0.606	0.152	0.364	0.182	0.061
s.d.	32.45	0.100	0.076	0.435	0.452	0.435	0.435	0.496	0.364	0.489	0.392	0.242

Gower's general coefficient of (dis)similarity: standardize each variable and multiply all the columns corresponding to dummy variables by 1/√2 = 0.7071, a factor which compensates for their 0/1 coding:

	Continuous variables			*Sampled region*				*Substrate character*				
Station	Depth	Temperature	Salinity	Tarehola	Skognes	Njosken	Storura	Clay	Silt	Sand	Gravel	Stone
s3	-0.868	0.615	0.260	1.231	-0.426	-0.394	-0.394	-0.864	1.648	-0.526	-0.328	2.741
s8	-0.898	0.615	0.260	1.231	-0.426	-0.394	-0.394	0.561	-0.294	-0.526	1.477	-0.177
s25	-0.868	-0.854	-0.676	-0.394	1.137	-0.394	-0.394	0.561	-0.294	0.921	-0.328	-0.177
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
s84	0.242	1.294	-0.294	-0.394	-0.426	-0.394	1.231	0.561	-0.294	-0.526	-0.328	-0.177

compute Euclidean distance between stations

Dealing with heterogeneous data

Alternative approach:

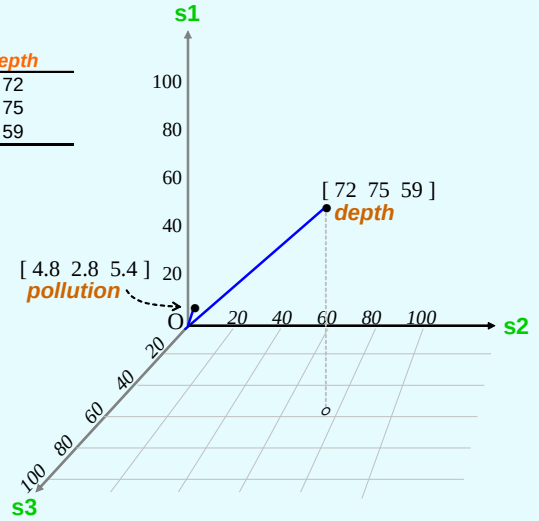
Compute distance matrices $\mathbf{D}_1$, $\mathbf{D}_2$, … for the different blocks of variables (homogeneous types within each block), and then combine the distances (or the squared distances) linearly:

$$\mathbf{D} = \sum_v w_v \mathbf{D}_v$$

where w_v is a set of positive weights that adds up to 1.

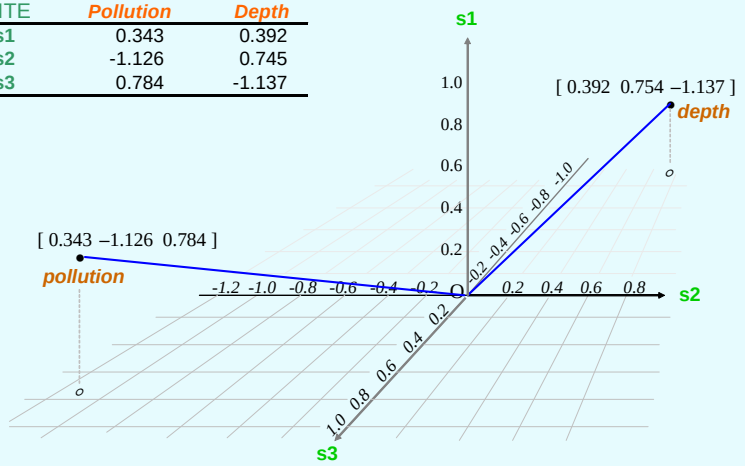
Correlations between variables

SITE	Pollution	Depth
s1	4.8	72
s2	2.8	75
s3	5.4	59

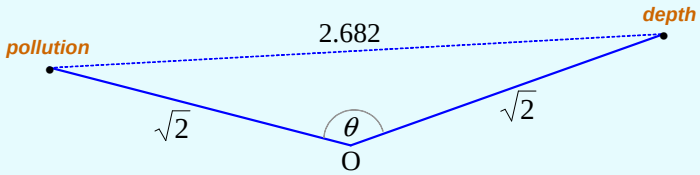


Correlations between variables

SITE	Pollution	Depth
s1	0.343	0.392
s2	-1.126	0.745
s3	0.784	-1.137



Correlations between variables



$c^2 = a^2 + b^2 - 2ab\cos(\theta)$ (cosine rule)

i.e., $2.682^2 = 2 + 2 - 2\sqrt{2}\sqrt{2}\cos(\theta) = 4 - 4\cos(\theta)$

hence, $\cos(\theta) = 1 - \frac{1}{4} \times 7.190 = -0.7975$

and the angle is $\theta = 2.494$ radians, or 142.9 degrees.

That's the correlation coefficient between pollution and depth!

Distance between variables

The relationship between correlation and interpoint distance becomes even simpler if the variables have unit length.

If c is the interpoint distance in this case, then:

$$r = 1 - \tfrac{1}{2} c^2$$

The inverse relationship is:

$$c = \sqrt{2}\sqrt{1-r}$$

Correlation	Poll.	Depth	Temp.
Pollution	1	-0.3955	-0.0921
Depth	-0.3955	1	-0.0034
Temperature	-0.0921	-0.0034	1

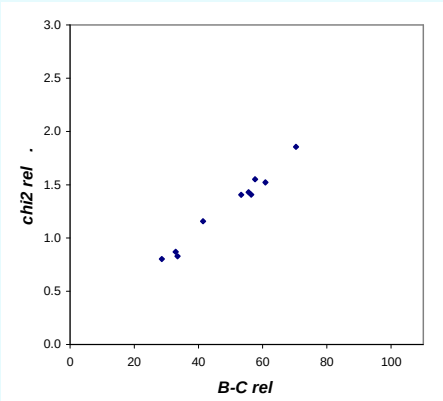
Distance	Poll.	Depth	Temp.
Pollution	0	1.6706	1.4779
Depth	1.6706	0	1.4166
Temperature	1.4779	1.4166	0

Distance between species

Apply chi-square or Bray-Curtis to the column profiles (to remove effect of 'size' of species)

chi2	a	b	c	d
b	0.802			
c	1.522	1.407		
d	0.87	0.828	1.157	
e	1.406	1.55	1.855	1.43

B-C	a	b	c	d
b	28.6			
c	60.9	56.4		
d	32.9	33.5	41.4	
e	53.3	57.6	70.4	55.6



Distance between categorical variables

Cross-tabulate the two variables (here we categorized depth to get another discrete variable):

		Sediment		
		C	S	G
Depth	low	6	5	0
	medium	3	5	1
	high	2	1	7

$n = 30$

χ^2 statistic = 15.88 $\quad \phi^2 = \chi^2/n = 0.519$ ('inertia' in correspondence analysis)

Cramer's $V = \sqrt{\dfrac{\phi^2}{\min\{I-1, J-1\}}} = 0.509$ (can be interpreted as a correlation)

Correspondence Analysis & Related Methods

Michael Greenacre

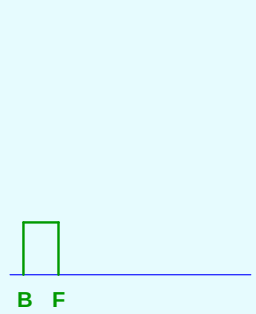
SESSION 2:

HIERARCHICAL CLUSTERING
WARD CLUSTERING
***k*-MEANS CLUSTERING**

Hierarchical Cluster Analysis - 1

- start with symmetric matrix of "Jaccard dissimilarities"

	A	B	C	D	E	F	G
A	0	.50	.43	1	.25	.63	.38
B	.50	0	.71	.83	.67	.20	.78
C	.43	.71	0	1	.43	.67	.33
D	1	.83	1	0	1	.80	.86
E	.25	.67	.43	1	0	.78	.38
F	.63	.20	.67	.80	.78	0	.75
G	.38	.78	.33	.86	.38	.75	0

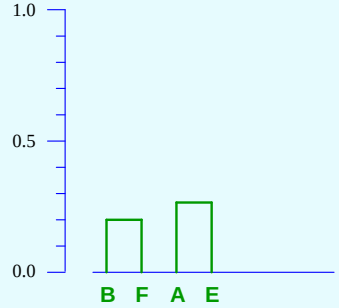


- look for smallest value (closest pair)
- join them together at their level of dissimilarity
- merge in table and recompute distances

Hierarchical Cluster Analysis - 2

- using "maximum" method (complete linkage)

	A	BF	C	D	E	G
A	0	.63	.43	1	.25	.38
BF	.63	0	.71	.83	.78	.78
C	.43	.71	0	1	.43	.33
D	1	.83	1	0	1	.86
E	.25	.78	.43	1	0	.38
G	.38	.78	.33	.86	.38	0

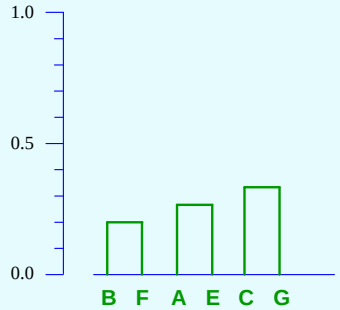


- look for smallest dissimilarity – if there's a tie, then take any one

Hierarchical Cluster Analysis - 3

- using “maximum” method (complete linkage)

	AE	BF	C	D	G
AE	0	.78	.43	1	.38
BF	.78	0	.71	.83	.78
C	.43	.71	0	1	.33
D	1	.83	1	0	.86
G	.38	.78	.33	.86	0

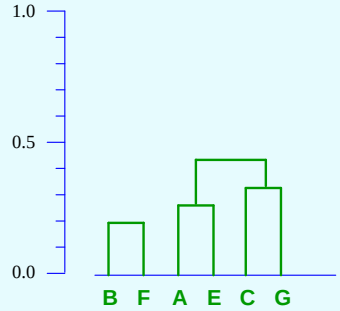


- look for smallest value (closest pair)

Hierarchical Cluster Analysis - 4

- using “maximum” method (complete linkage)

	AE	BF	CG	D
AE	0	.78	.43	1
BF	.78	0	.78	.83
CG	.43	.78	0	1
D	1	.83	1	0

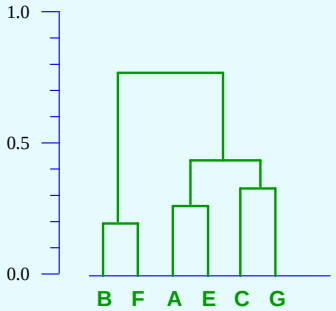


- look for smallest value (closest pair)

Hierarchical Cluster Analysis - 5

- using “maximum” method (complete linkage)

	AECG	BF	D
AECG	0	.78	1
BF	.78	0	.83
D	1	.83	0

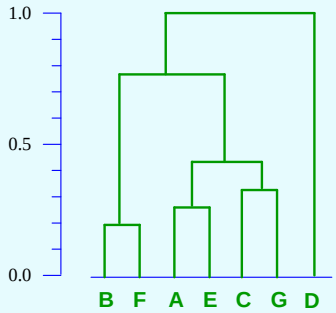


- look for smallest value (closest pair)

Hierarchical Cluster Analysis - 6

- using “maximum” method (complete linkage)

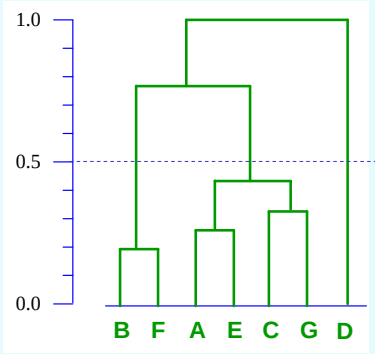
	AECGBF	D
AECGBF	0	1
D	1	0



- look for smallest value (closest pair)

Results: Hierarchical Cluster Analysis

- possible cut of the tree to establish clusters:

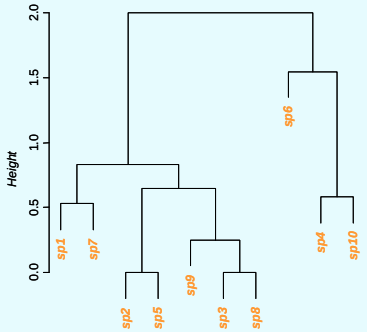


three clusters:
{B,F},
{A,E,C,G}
{D}

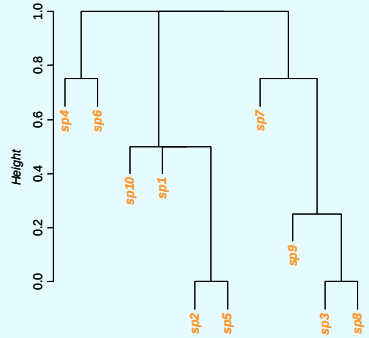
Usually we cut the tree where there is a lot of space between the nodes to cut it. In this case we would also like to cut it at a value below 0.5 because this means that there are more species in common between the samples than not.

Clustering of species

based on correlation
$(1-r)$

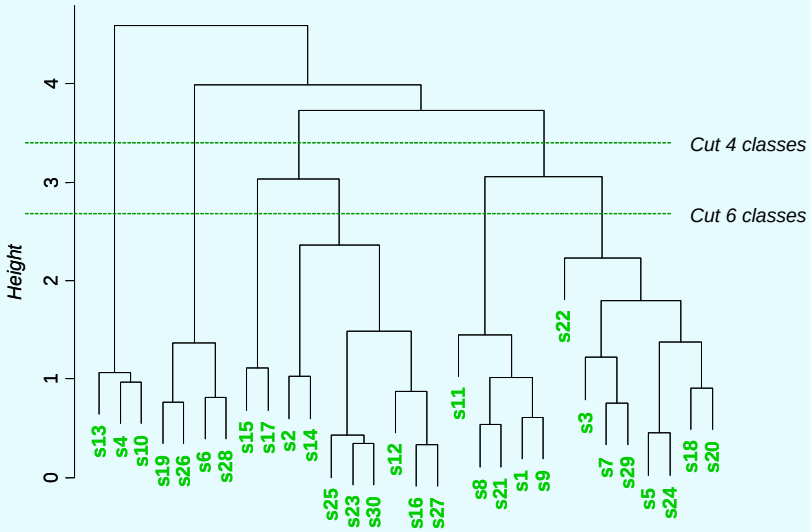


based on Jaccard index
between species



...using R function **hclust**

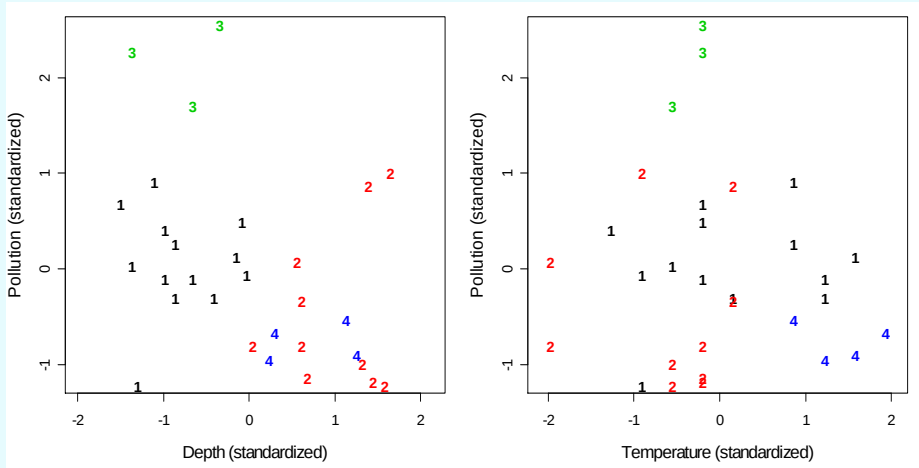
Clustering of sites
based on pollution, depth & temperature



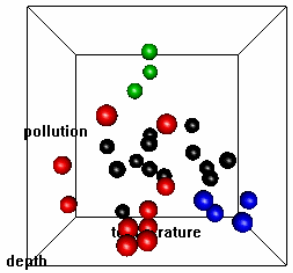
Showing clusters in plot

These data are in three dimensions: pollution, depth and temperature.

Here are two planes of this two-dimensional space, showing pollution versus depth and pollution versus temperature.

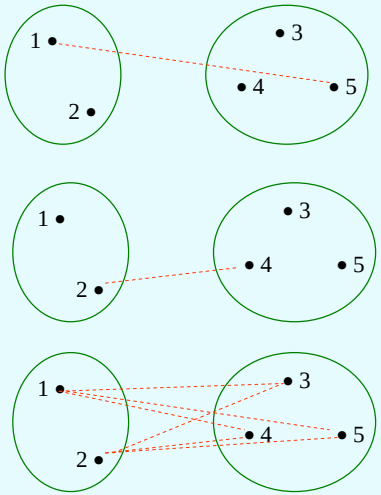


Showing clusters in 3d-plot



Inter-cluster distance measures

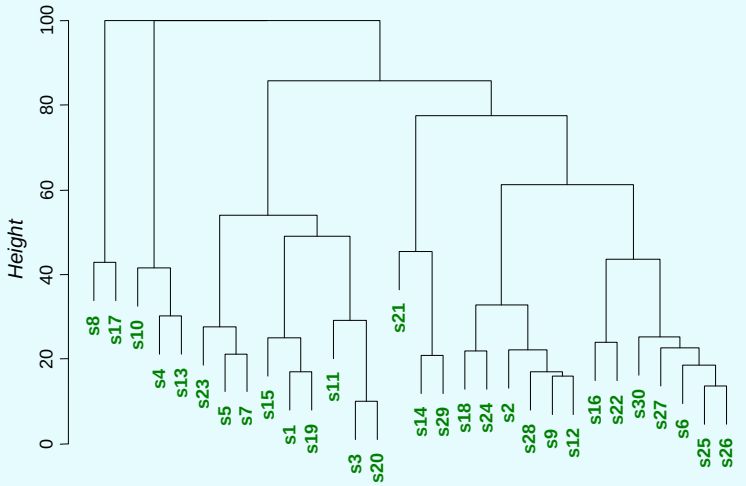
method	cluster distance
'maximum' complete linkage	d_{15}
'minimum' single linkage	d_{24}
'average' average linkage	$\frac{d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{26}}{6}$



Bray-Curtis dissimilarities

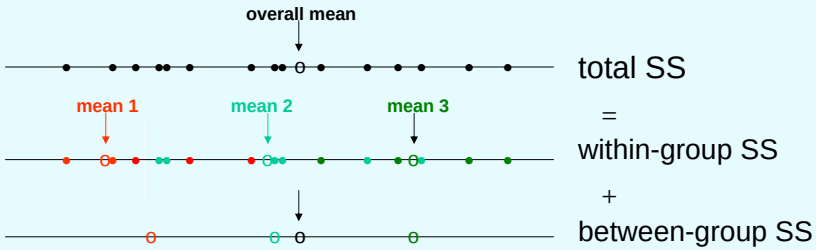
	s1	s2	s3	s4	s5	s6	. . .	s24	s25	s26	s27	s28	s29
s2	45.7												
s3	29.6	48.1											
s4	46.7	55.6	46.7										
s5	47.7	34.8	50.8	78.6									
s6	52.2	22.9	52.2	69.2	41.9								
s7	45.5	41.5	49.1	87.0	21.2	50.9							
.	.	.	.	.	.	.	.						
.	.	.	.	.	.	.	.						
.	.	.	.	.	.	.	.						
s25	70.4	39.3	66.7	93.9	52.9	18.6	. . .	46.4					
s26	69.6	32.8	60.9	92.8	41.7	15.2	. . .	39.3	13.7				
s27	63.6	38.1	63.6	93.3	38.2	21.5	. . .	42.6	16.3	22.6			
s28	32.5	21.5	50.0	57.7	31.9	29.5	. . .	30.9	41.8	47.5	34.4		
s29	43.4	35.0	43.4	54.5	31.2	53.6	. . .	39.8	64.5	58.2	61.2	34.2	
s30	60.7	36.7	58.9	84.5	48.0	21.6	. . .	40.8	18.1	25.3	23.6	37.7	56.8

Clustering Bray-Curtis dissimilarities



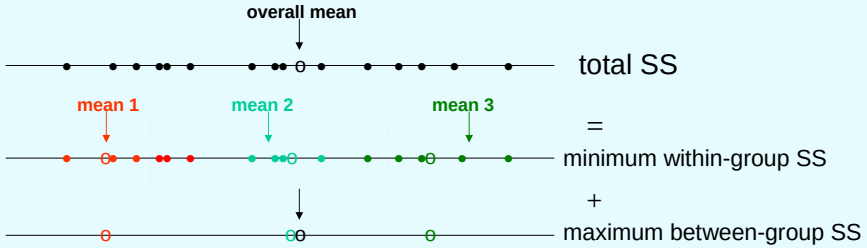
Ward clustering

- of particular interest for all Euclidean and weighted Euclidean distance functions
- decomposes the total variance of the points in space into parts "within-clusters" and "between-clusters"
- the idea can be illustrated by analysis-of-variance (ANOVA) for one variable (SS=sum of squared deviations from the mean):



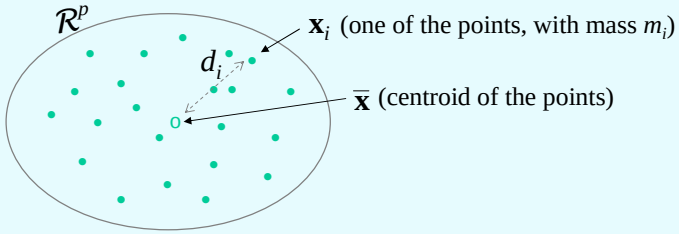
Ward clustering

- how should the points be divided into three groups to minimise the within-group SS, or (equivalently) to maximise the between-group SS?



Ward clustering

- for any set of points in (weighted) Euclidean space:



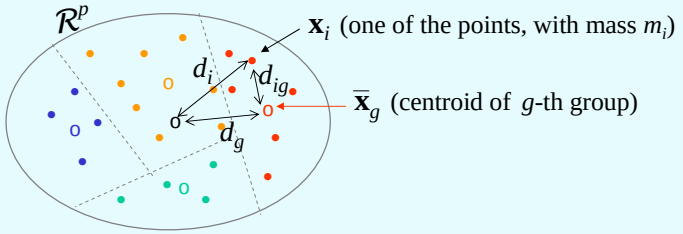
- *total inertia* of the "cloud" of n points:

$$\sum_i m_i d_i^2$$

(this includes regular variance if $m_i = 1/n$)

Ward clustering

- we make any partitioning of the cloud into G groups:



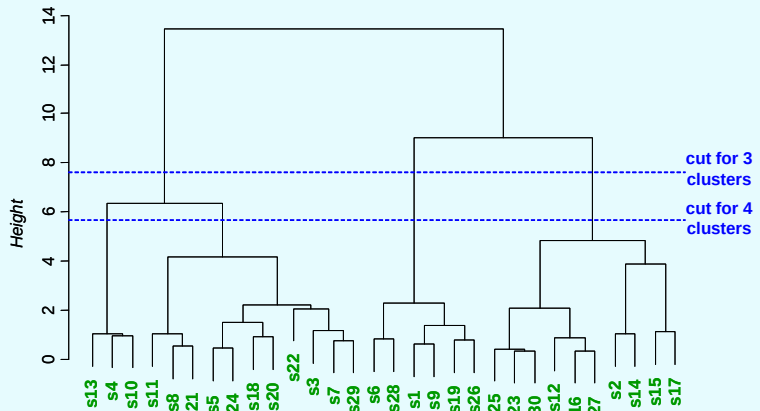
- *total inertia = within-group inertia + between-group inertia*

$$\sum_i m_i d_i^2 = \sum_g \left(\sum_i m_i d_{ig}^2 \right) + \sum_g m_g d_g^2$$

- for successful clustering we want the between-group inertia to be as high as possible, which is equivalent to the within-group inertia being as low as possible

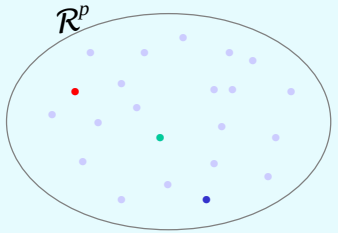
Ward clustering in three dimensions

(pollution, depth & temperature, all standardized)



Nonhierarchical clustering (K-means)

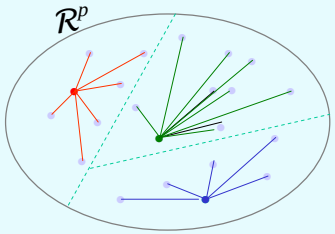
- suppose G groups (e.g. $G = 3$) and select 3 points as "seeds":



- the "seeds" can be chosen
  (1) at random
  (2) by intelligent selection

K-means clustering

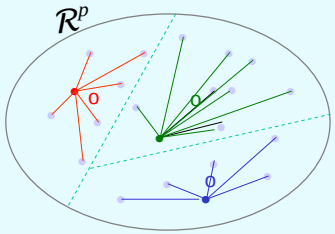
- allocate each object to the closest seed:



- this gives a first classification of the objects...
- ... and this is equivalent to making the above partitioning

K-means clustering

- from the first classification, calculate the centroids of the groups:



- then repeat the exercise, using the centroids as the new seeds, repeating over and over again until stable (convergence)...

Hierarchical vs. nonhierarchical clustering

- Hierarchical:

  - nice binary tree representation (but cannot draw and interpret the tree easily for a large number of objects being clustered);

  - suitable for small to medium size data sets;

  - you don´t have to specify number of clusters in advance.

- Nonhierarchical:

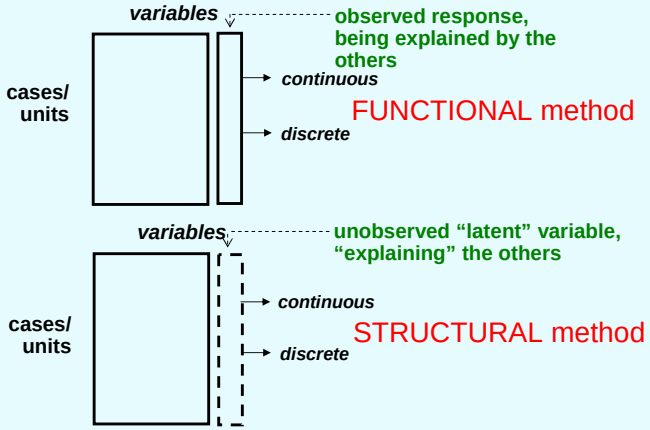
  - no binary tree;

  - suitable for large data sets;

  - you have to specify number of clusters in advance.

A basic scheme of multivariate analysis

All multivariate methods fall basically into two types, depending on the data structure and the question being asked:



Four corners of multivariate analysis