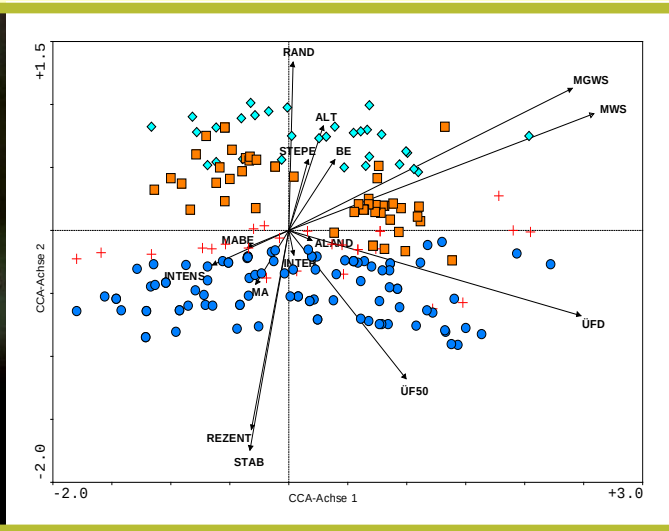
SENCKENBERG
world of biodiversity



Multivariate Methods in Ecology

Karsten Wesche - Botany Görlitz

Henrik von Wehrden, Jan Hanspach –
Leuphana Lüneburg

(Ilona Leyer – Univ. Marburg)

Leibniz Gemeinschaft

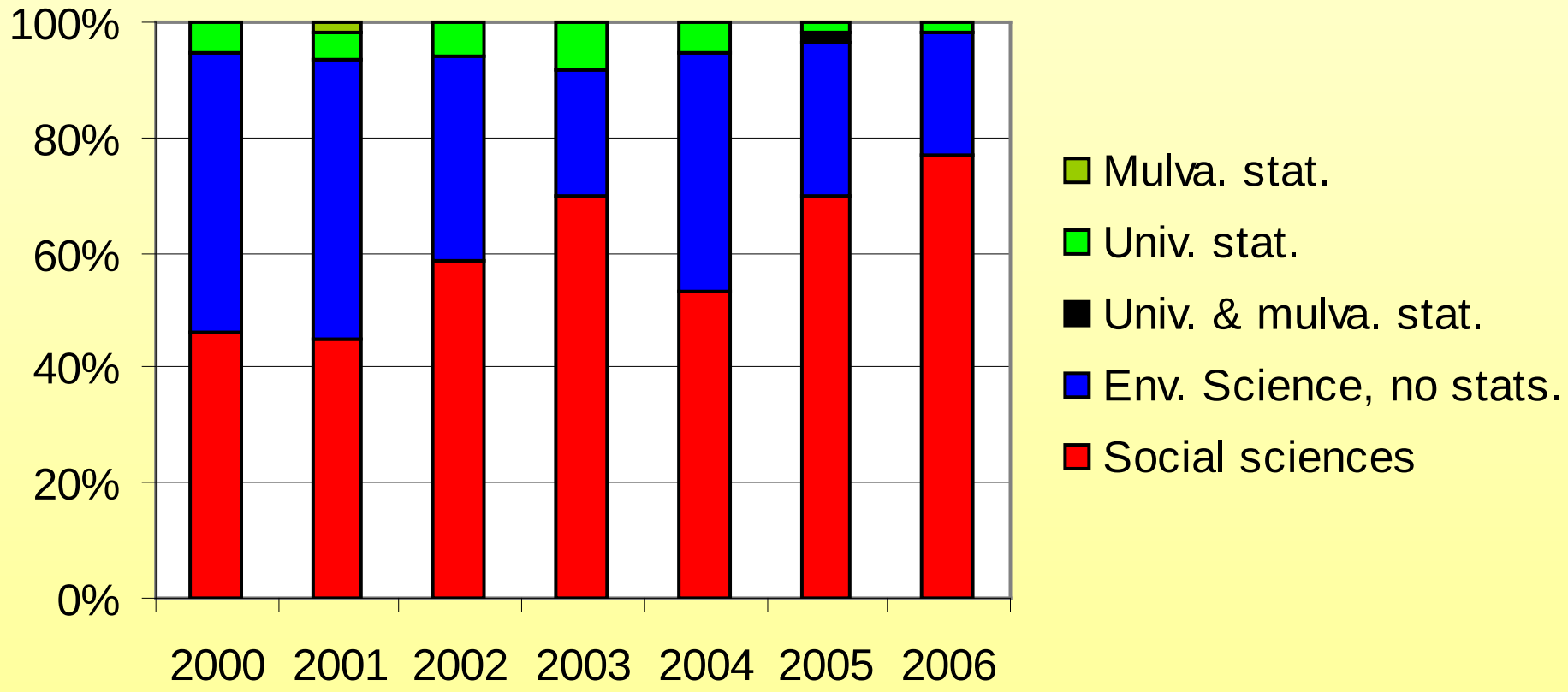
Content

- **Theory of the most widely used ordination and classification methods in ecology**
- **Practicing own multivariate analyses using R (packages vegan, cluster etc.)**
- **Graphing and understanding multivariate analyses including key statistical measures**

Approach

- Why and where?
- Introduction: multivariate analysis, data qualities, transformations, similarities
- (Detrended) Correspondence Analysis – (D) CA
- Principal component analysis – PCA
- *Post hoc* fitting of secondary information & Canonical Correspondence Analysis – CCA (RDA)
- The Fourth Corner Problem: Functional traits analysis (J. Hanspach)
- Non metric multidimensional scaling – NMDS
- Classification

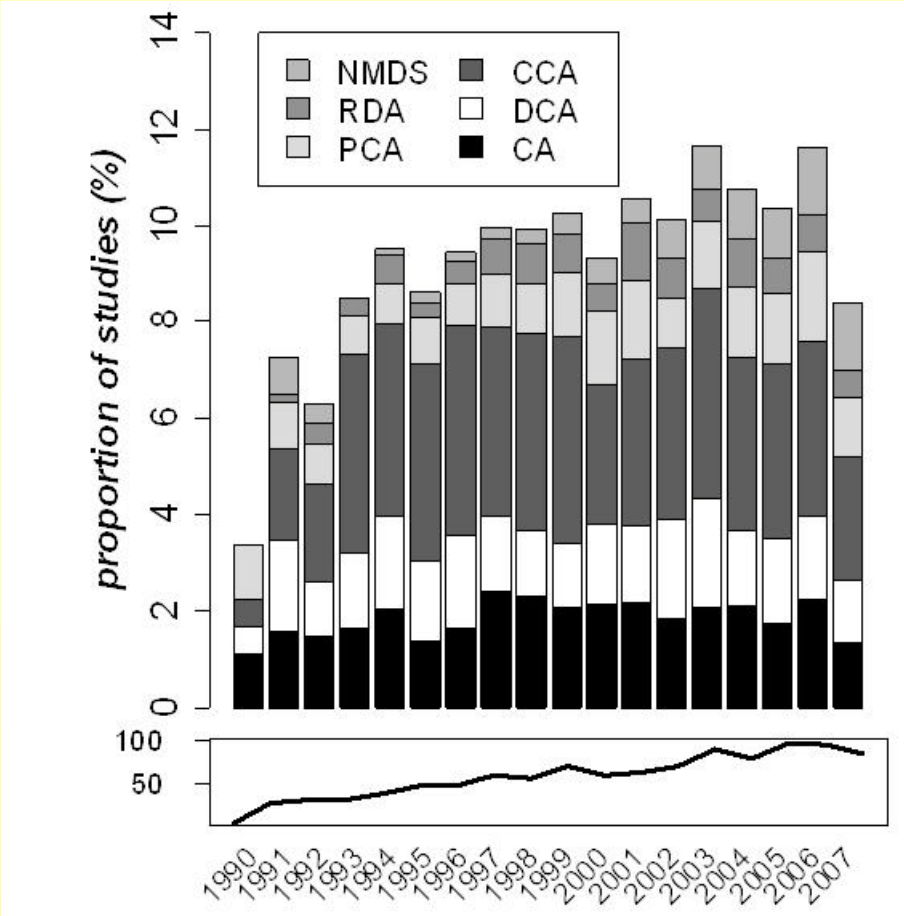
Application of multivariate statistics

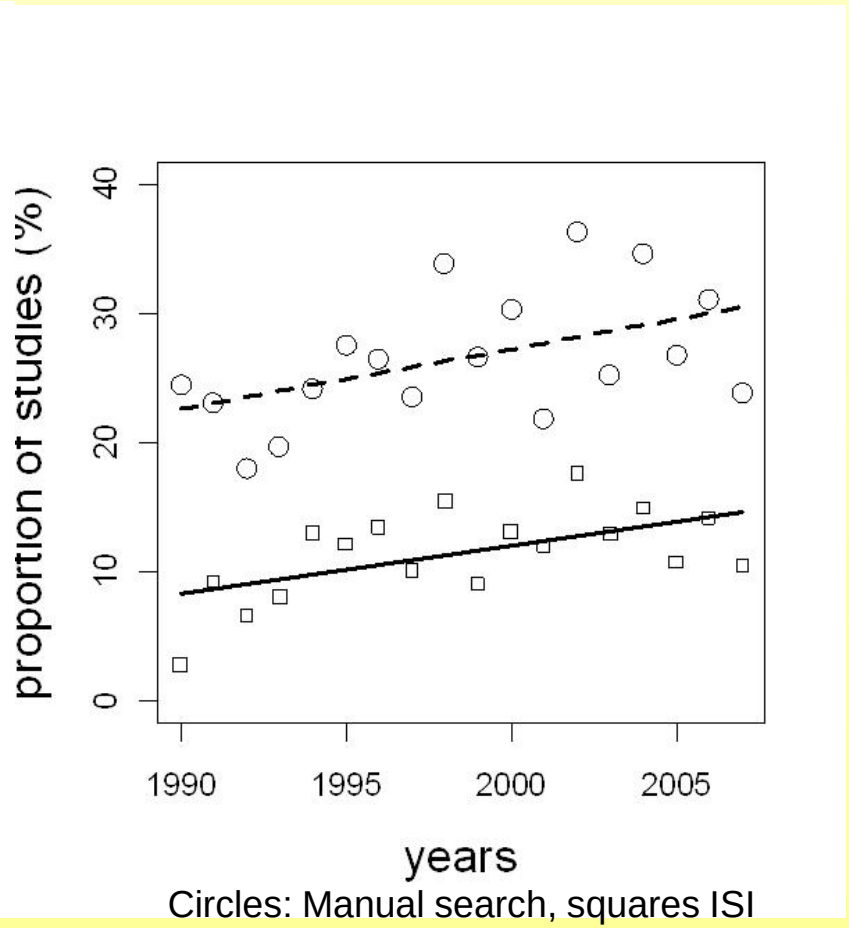


Literature search: Application of multivariate methods in *Natur und Landschaft* (since 1990)

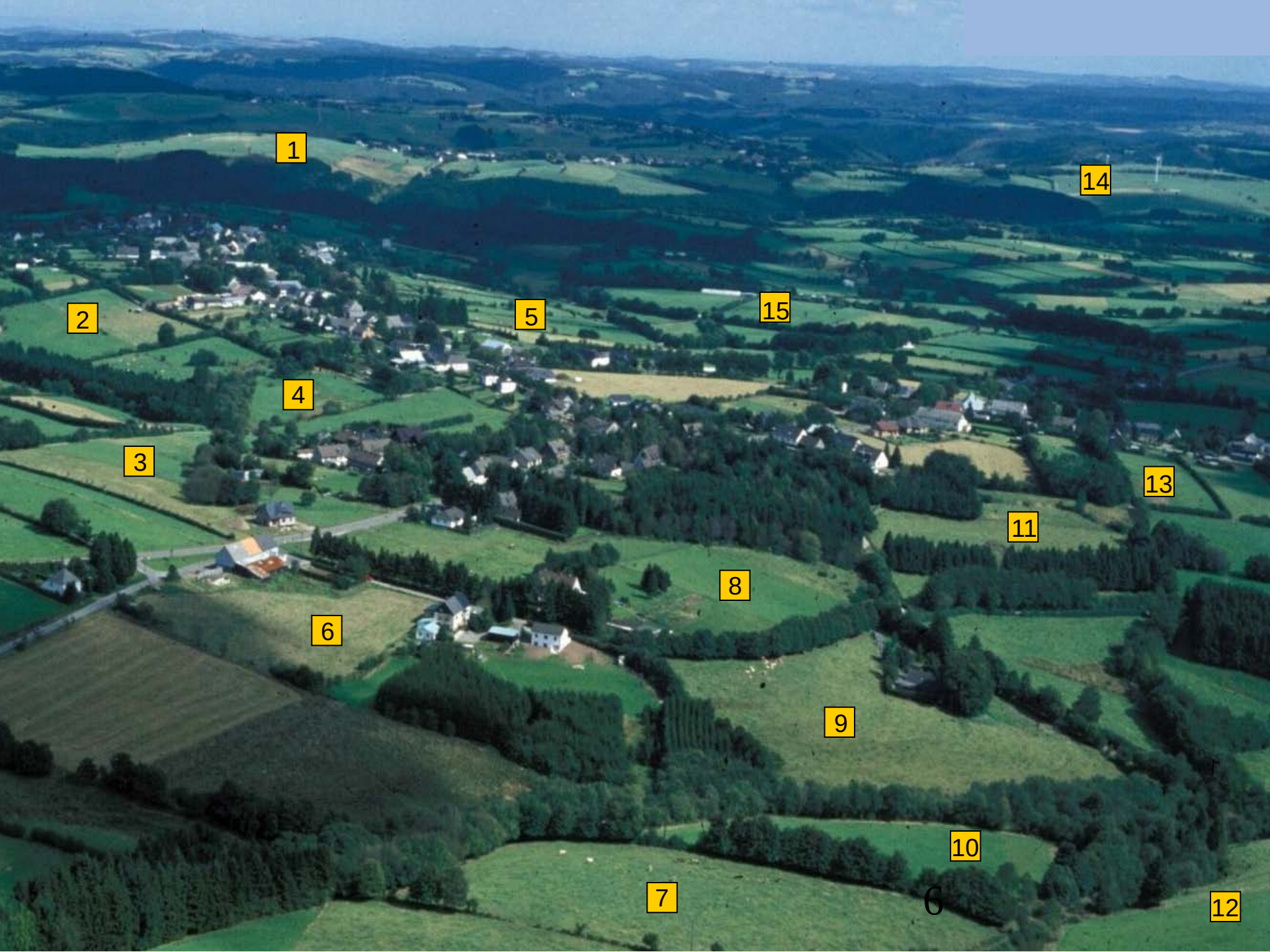
Application of multivariate statistics

Application of ordination methods in major ecological journals, and in those specialising on vegetation ecology (von Wehrden et al. JVS 2011)





Circles: Manual search, squares ISI


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

The term „*sample*“

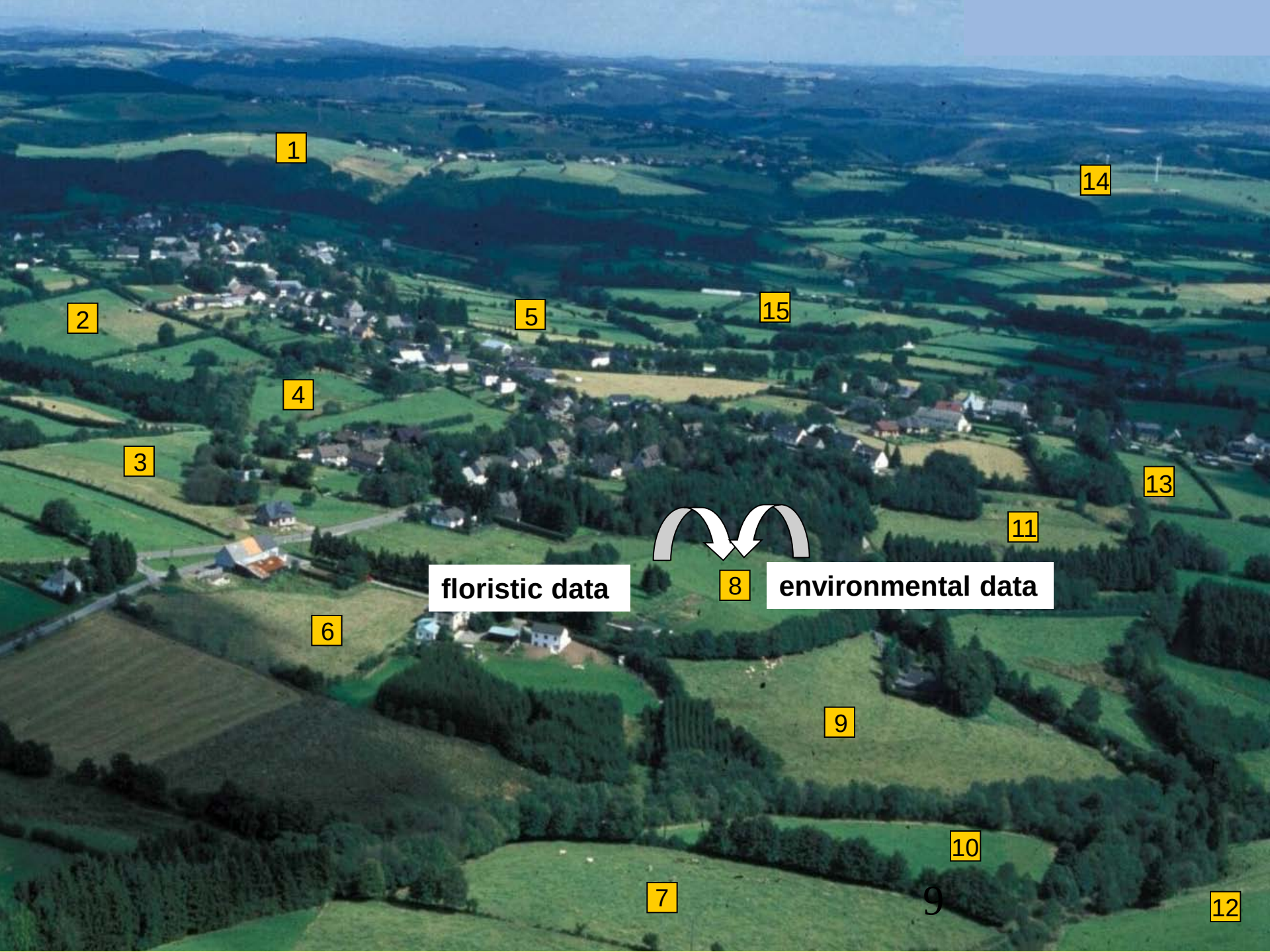
In vegetation ecology: vegetation sample / relevé

In animal ecology: trap unit, net stroke

Biogeography: raster (e.g MTB)
administrative units

Species by sample matrix

Species sample no.	1	2	3	4	5	6	7	8	9	10	11	12
Ranunculus repens	1	0,1	1	2	1	1	2	0,1	1	5	.	0,1
Trifolium repens	.	.	1	1	.	.	1	1	2	.	.	.
Poa palustris	2	2	.	.	1	1	2	1	.	1	1	1
Festuca pratensis	.	.	3	1	0,1	.	.	.	2	.	.	.
Cnidium dubium	1	.	.	.	2	3	2	.	.	.	.	.
Ranunculus auricomus agg.	.	.	1	0,1	.	.	.	.	0,1	.	.	.
Agrostis stolonifera	2	.	1	2	.	.	.	.	.	.	.	.
Veronica arvensis	.	.	.	.	0,1	0,1	.	.	.	.	.	1
Glechoma hederacea	2	1	1	1	.	.	.	.	1	1	0,1	.
Convolvulus arvensis	.	.	.	0,1	.	.	.	.	.	.	.	.
Poa trivialis	3	4	3	2	.	.	.	3	3	2	3	1
Cerastium glutinosum	.	.	.	.	0,1	.	.	.	.	.	.	1
Potentilla reptans	1	r	1	.	1	1	2	1	.	.	0,1	.
Allium vineale	.	.	.	.	.	.	.	.	.	.	1	.
Ornithogalum umbellatum agg.	.	.	.	.	.	.	.	.	1	.	.	.


1
14
2
5
15
4
3
13
11
floristic data
8
environmental data
6
9
10
7
12

Properties ecological data

multivariate data – environmental data / secondary matrix

	Samples															
UP	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	Scale
UP 1	120	330	210	195	125	175	320	275	235	450	95	315	315	210	185	ratio
UP 2	7.5	4.9	4.8	6.2	6.3	7.3	7.1	7.0	7.1	4.5	5.1	5.3	5.9	5.2	5.4	ratio
UP 3	0	1	0	0	1	0	1	0	0	1	1	1	0	1	1	nominal
UP 4	1	0	1	1	0	1	0	1	1	0	0	0	1	0	0	nominal
UP 5	1	3	4	4	4	2	3	3	3	1	1	1	1	2	2	ordinal
UP 6	25	23	24	28	23	25	23	25	28	29	21	20	18	17	16	ratio

Multivariate ecological data

Community data: a data matrix that contains species and samples („species by sample matrix“)

Environmental data (secondary data): a datamatrix, which contains environmental variables and samples

(Trait data: a datamatrix, which contains trait data and species)

Properties of multivariate data

- **elements of the matrix are abundances**
- **community data have many 0-values**
- **most species are not common, only few reach high abundances**
- **the number of potential predictor variables is very large – but usually only few variables are relevant for variation in the species composition**
- **Community data are characterised by high „noise“**
- **Community data are strongly redundant**

multivariate techniques must be capable of handling this

Strategies in multivariate data analysis

Species data

	Samples														
Arten	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Art 1	3	2	.	.	1	2	.	.	3	2	.	.	1	2	.
Art 2	.	.	3	1	.	.	3	1	.	.	3	1	.	.	.
Art 3	.	3	4	.	3	3	1	.	.	3	4	.	.	.	1
Art 4	2	.	3	.	2	.	3	.	2	.	3	.	2	.	3
Art 5	5	8	.	6	.	.	1	1	5	8	.	6	1	.	3
Art 6	.	.	3	.	.	1	.	.	.	.	3	.	.	.	1
Art 7	.	5	.	3	.	5	.	3	.	5	.	.	.	5	.
Art 8	3	.	2	.	3	.	2	.	3	.	2	5	3	1	2
Art 9	.	9	1	.	.	3	1	.	.	9	1	.	.	2	2
Art 10	.	.	7	.	.	.	.	2	.	.	7	.	.	.	.
Art 11	5	2	5	.	2	1	5	.	5	2	5	.	2	1	5
Art 12	3	.	8	7	3	1	8	.	3	.	8	7	3	1	1

Are there major gradients in species composition?

correspondence analysis (principal component analysis)

environmental data

	Samples														
UP	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
UP 1	120	330	210	195	125	175	320	275	235	450	95	315	315	210	185
UP 2	7.5	4.9	4.8	6.2	6.3	7.3	7.1	7.0	7.1	4.5	5.1	5.3	5.9	5.2	5.4
UP 3	0	1	0	0	1	0	1	0	0	1	1	1	0	1	1
UP 4	1	0	1	1	0	1	0	1	1	0	0	0	1	0	0
UP 5	1	3	4	4	4	2	3	3	3	1	1	1	1	2	2
UP 6	25	23	24	28	23	25	23	25	28	29	21	20	18	17	16

Are there relationships between environmental (predictor) variables?

principal component analysis

Strategies in multivariate data analysis

Species data

	Samples														
Arten	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Art 1	3	2	.	.	1	2	.	.	3	2	.	.	1	2	.
Art 2	.	.	3	1	.	.	3	1	.	.	3	1	.	.	.
Art 3	.	3	4	.	3	3	1	.	.	3	4	.	.	.	1
Art 4	2	.	3	.	2	.	3	.	2	.	3	.	2	.	3
Art 5	5	8	.	6	.	1	1	1	5	8	.	6	1	.	3
Art 6	.	.	3	.	.	1	.	.	.	.	3	.	.	.	1
Art 7	.	5	.	3	.	5	.	3	.	5	.	.	.	5	.
Art 8	3	.	2	.	3	.	2	.	3	.	2	5	3	1	2
Art 9	.	9	1	.	.	3	1	.	.	9	1	.	.	2	2
Art 10	.	.	7	.	.	.	.	2	.	.	7	.	.	.	.
Art 11	5	2	5	.	2	1	5	.	5	2	5	.	2	1	5
Art 12	3	.	8	7	3	1	8	.	3	.	8	7	3	1	1

environmental data

	Samples														
UP	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
UP 1	120	330	210	195	125	175	320	275	225	450	95	315	315	210	185
UP 2	7.5	4.9	4.8	6.2	6.3	7.3	7.1	7.0	7.1	4.5	5.1	5.3	5.9	5.2	5.4
UP 3	0	1	0	0	1	0	1	0	0	1	1	1	0	1	1
UP 4	1	0	1	1	0	1	0	1	1	0	0	0	1	0	0
UP 5	1	3	4	4	4	2	3	3	3	1	1	1	1	2	2
UP 6	25	23	24	28	23	25	23	25	28	29	21	20	18	17	16

Is variation in the primary matrix related to (variation in) the secondary matrix?

indirect gradient analysis
canonical techniques

canonical correspondence analysis
(redundancy analysis)

Why multivariate analysis?

It is difficult to assess many dimensions simultaneously. In ecology, however, dozens of dimensions may have influence - ordination aims at reduction of dimensions.

Typically, some few dimensions will represent the main species-gradients (floristic/faunistic) and thus also environmental gradients.

Reduction on the main dimensions results in reduction of noise in the data.

multivariate techniques

ordination

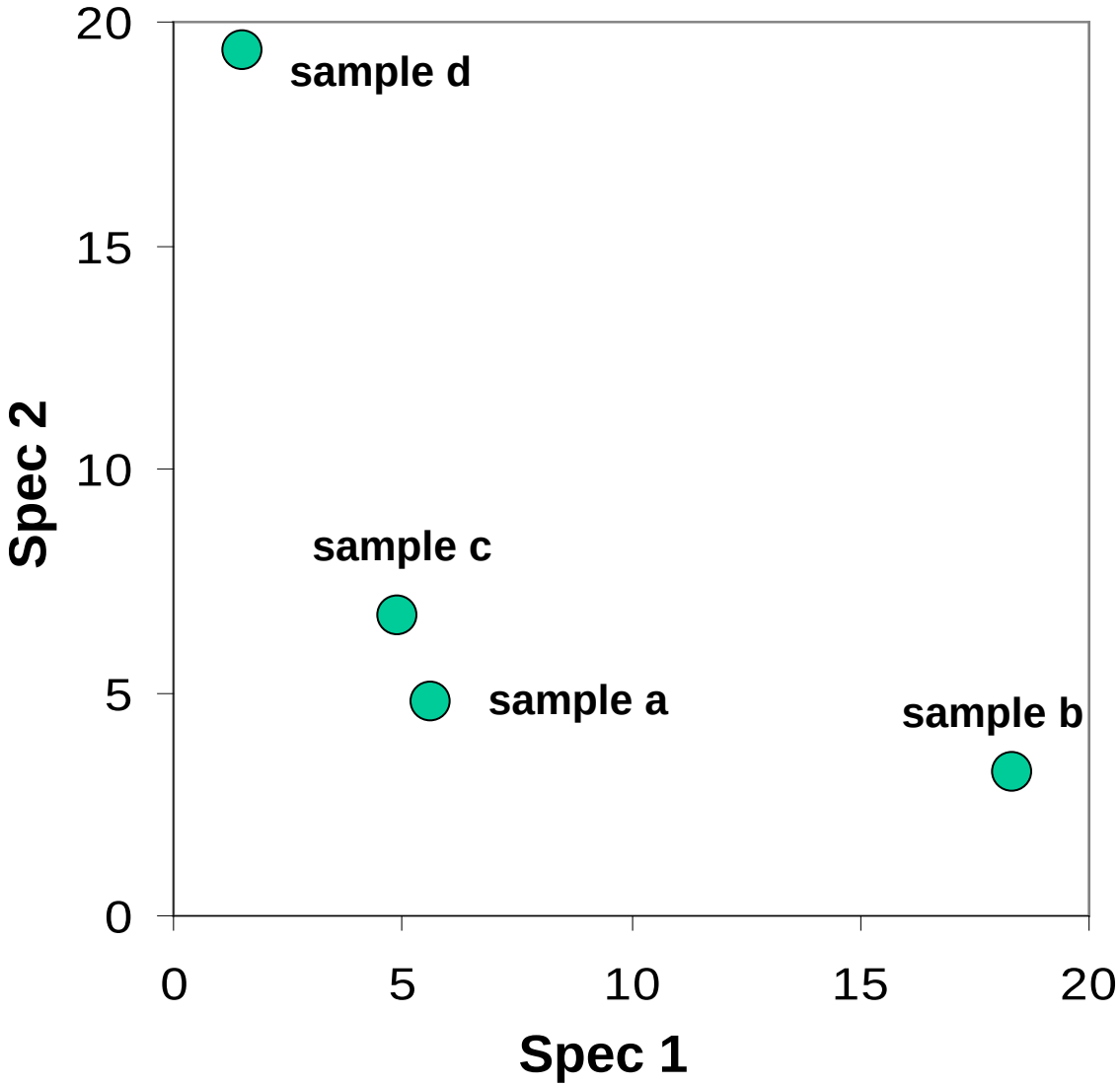
the process of ordering samples and species along abstract axes that (may) represent environmental gradients.
The techniques are also called „gradient analysis“.

classification

samples and species are assigned to classes, with the aim to maximise similarity within classes and to minimise similarity among classes.

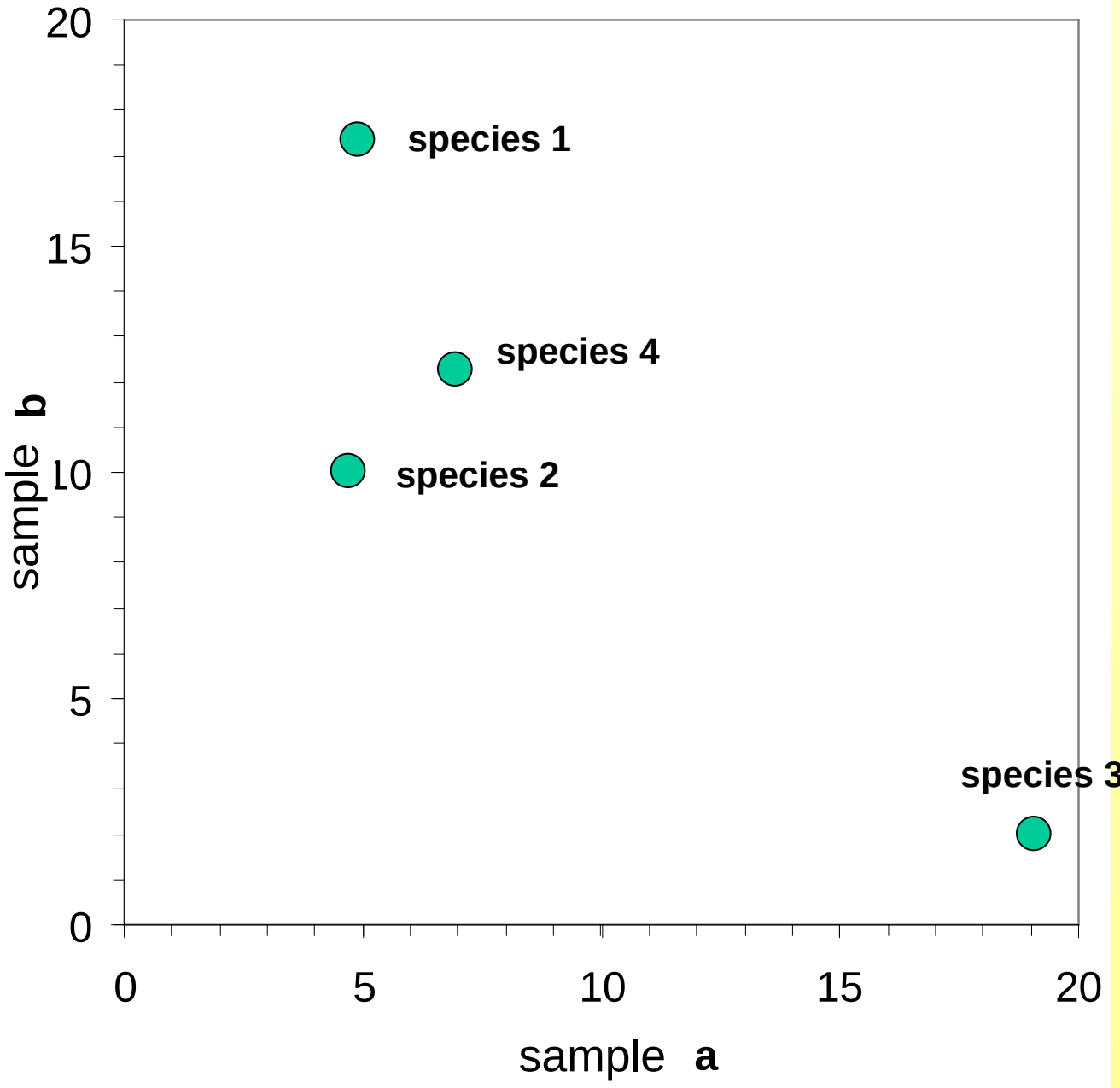
R-analysis

Spec.	Sample a	b	c	d
Spec 1	6	18	5	2
Spec 2	5	3	7	19



Q-analysis

Spec.	Sample a	Sample b
Spec 1	6	17.5
Spec 2	5	10
Spec 3	19	2
Spec 4	7.5	12.5



Data quality

Data can have different scales
Data often have to be **numerical** – recoding may be necessary (e.g Braun-Blanquet-scale).

Nominal - Ordinal - Interval – Ratio

scale		**examples**
Nominal	binary, only two levels	sex, species‘ presence
	several classes	skin colour, geological substrate
Ordinal	ordered /ranked	school grades, Ellenberg indicator values
Interval	intervals defined	temperatures (in C)
Rational	intervals with defined zero	weight

Data quality

Depending on the **scale** a given calculation may be allowed or not; e.g averaging of (ordinal) school grades does not make much sense (from a strictly statistical perspective). Nominal data can be **recoded** to 0/1-scaled **Dummy-variables.**

	a) Habitat	Dummy -“active“	Dummy -“old“	b) Human impact		
					„level 0“	„level 1“
P1	active flood plain	1	0	null	1	0
P2	old plain	0	1	modest	0	1
P3	margin	0	0	strong	0	0

Data quality

In circular variables (e.g exposiure) very high and very low values are similar (1° vs. 359°). For standard statistics, circular variables are recoded in two variables (cosinus & sinus).

example exposiure:

Exposure (°)	0	45	90	135	180	225	270	315	359	
Cosinus	1.00	0.71	0.00	-0.71	-1.00	-0.71	0.00	0.71	1.00	"northness"
Sinus	0.00	0.71	1.00	0.71	0.00	-0.71	-1.00	-0.71	-0.02	"eastness"

Details: MCCUNE B. & DYLAN K. 2002: Equations for potential annual direct incident radiation and heat load. - *J Veg Sci* **13**: 603-606.

Standardising

Transformations to make data comparable

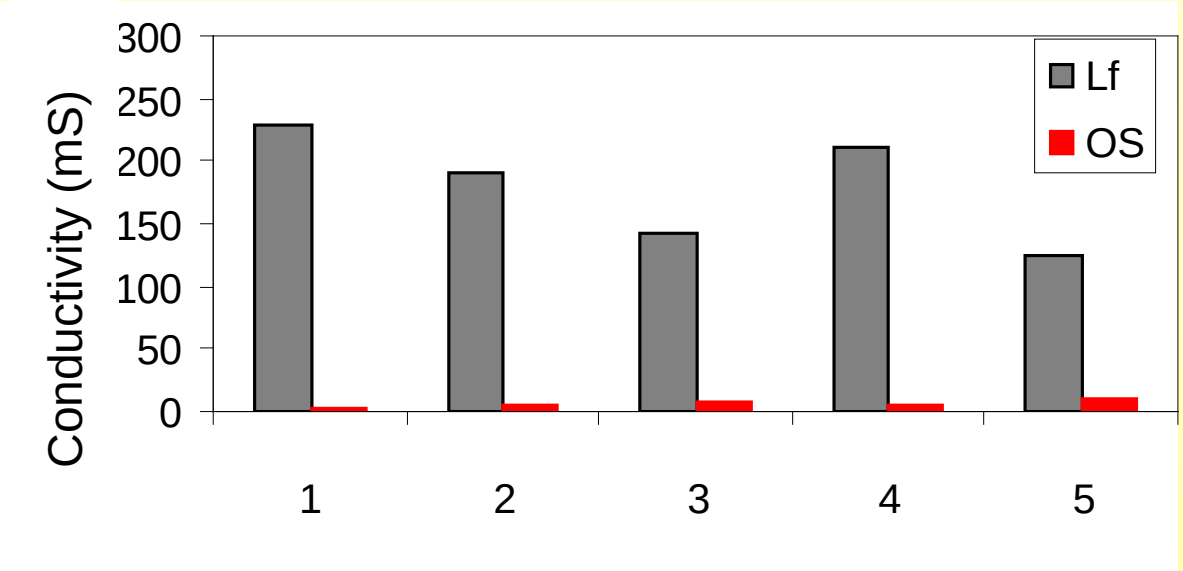
sample	1	2	3	4	5	mean	s.d.
conductivity (mS)	**230**	**190**	**143**	**210**	**125**	179.6	39.73
org. matter(%)	**3.5**	**5.3**	**6.4**	**4.2**	**11.2**	6.12	2.72

standardise by centering:

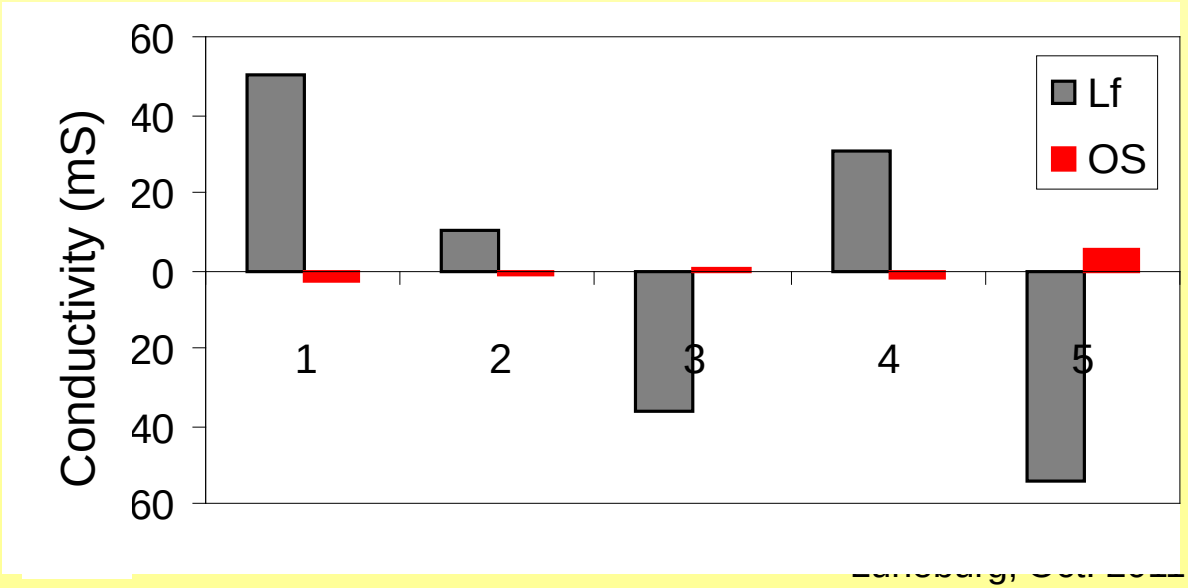
$$x' = x - \bar{x}$$

sample centered	1	2	3	4	5	mean	s.d.
conductivity (mS)	**50.4**	**10.4**	**-36.6**	**30.4**	**-54.6**	0.0	39.73
org. matter(%)	**-2.62**	**-0.82**	**0.28**	**-1.92**	**5.08**	0.0	2.72

Standardising



raw data:
conductivity and
organic matter



centered data

Standardise

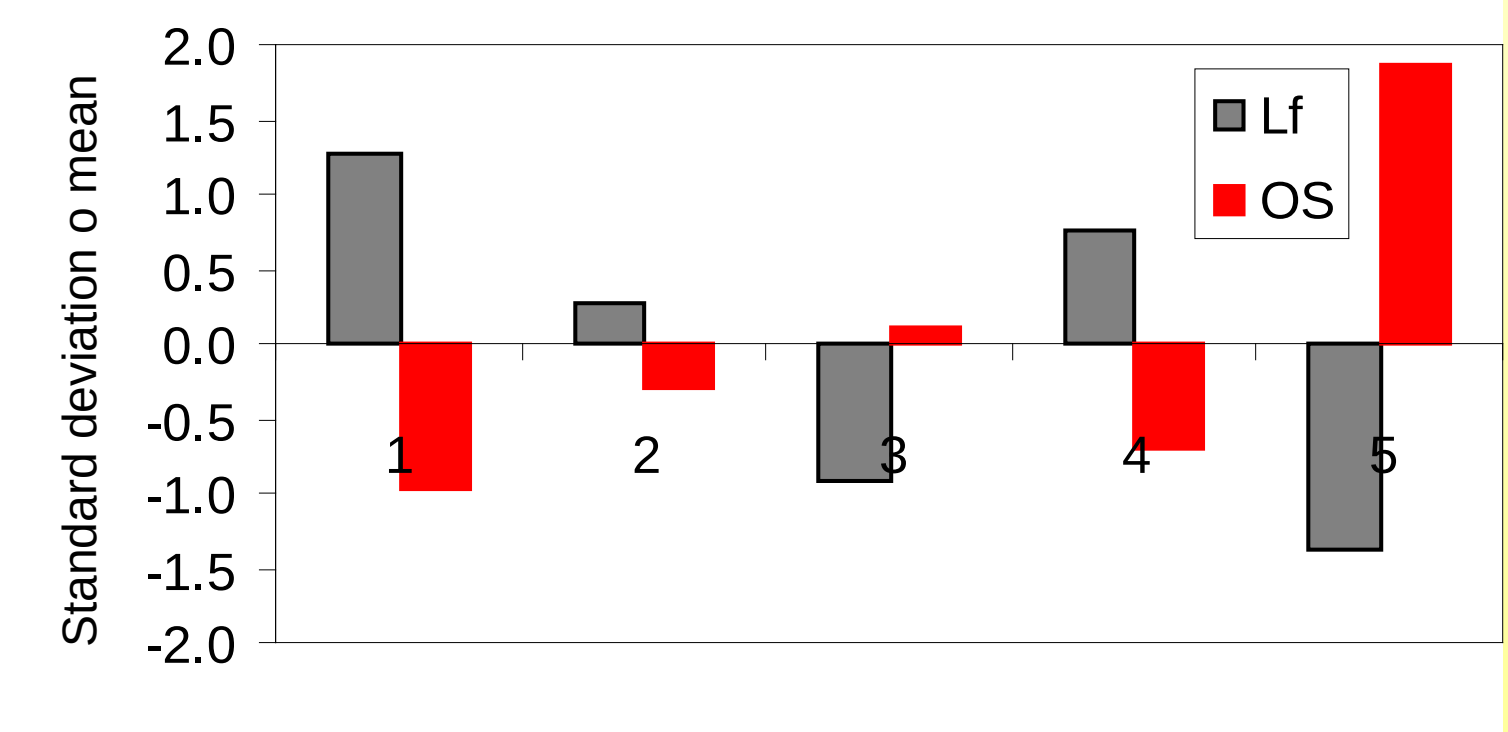
Make variables with **differing units** **comparable**: divide by standard deviation
Standardise to "***zero mean and unit variance***"

$$z' = \frac{x - \bar{x}}{S}$$

sample	1	2	3	4	5	mean	s.d.
standardised							
conductivity (mS)	1.27	0.26	-0.92	0.77	-1.37	0.0	1.0
org. matter(%)	-0.96	-0.3	0.10	-0.70	1.86	0.0	1.0

Standardising

data standardised and centered ("*zero mean, unit variance*")



Transformations

Transformations of data

Down-weighting of dominant species:

e.g **square root-transformation**

log-transformation

Extreme down-weighting dominant species:

0/1 transformation (***presence/absence***,

calculate as $x' = x^0$)

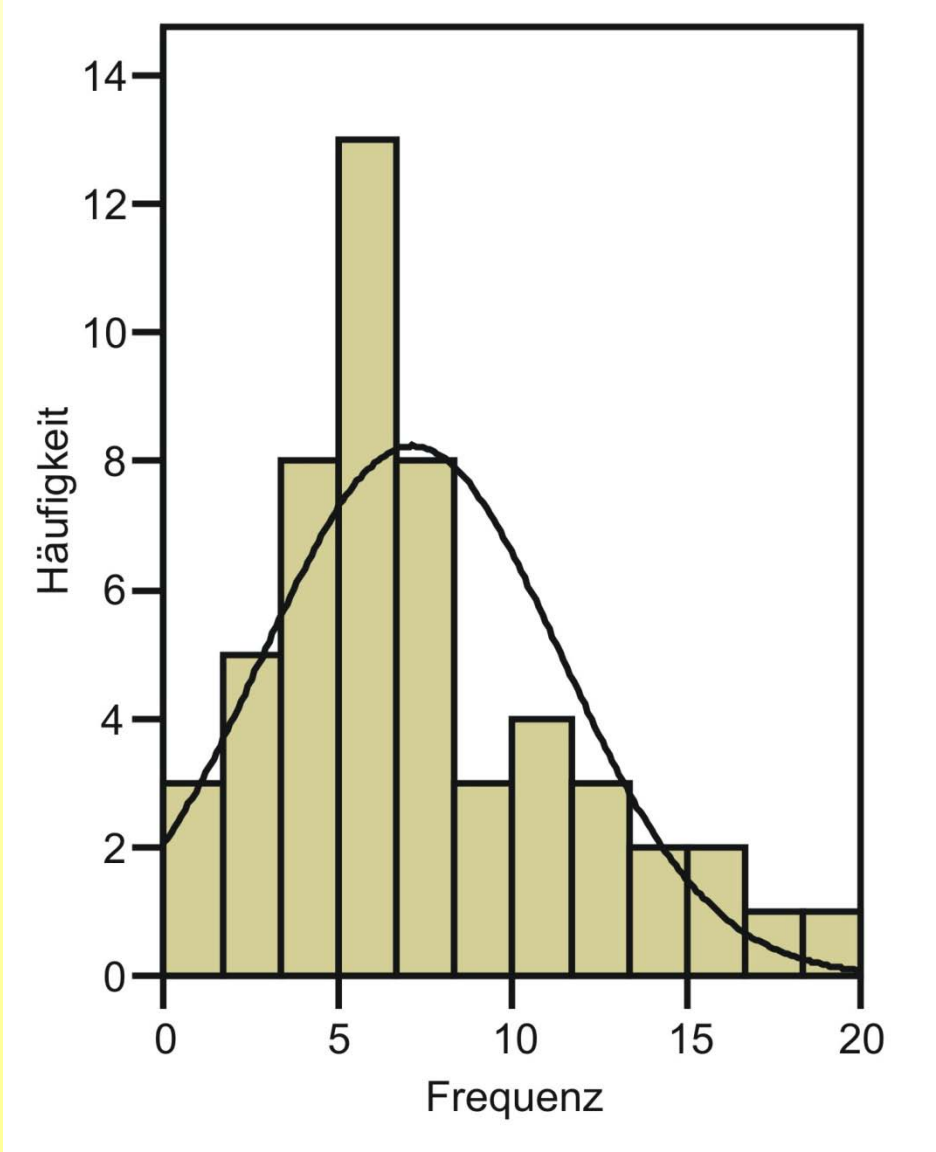
To fit certain statistical distributions or models

e.g **square root-transformation**

log-transformation

transformations

raw data: frequencies of species in samples from the Elbe flood plain

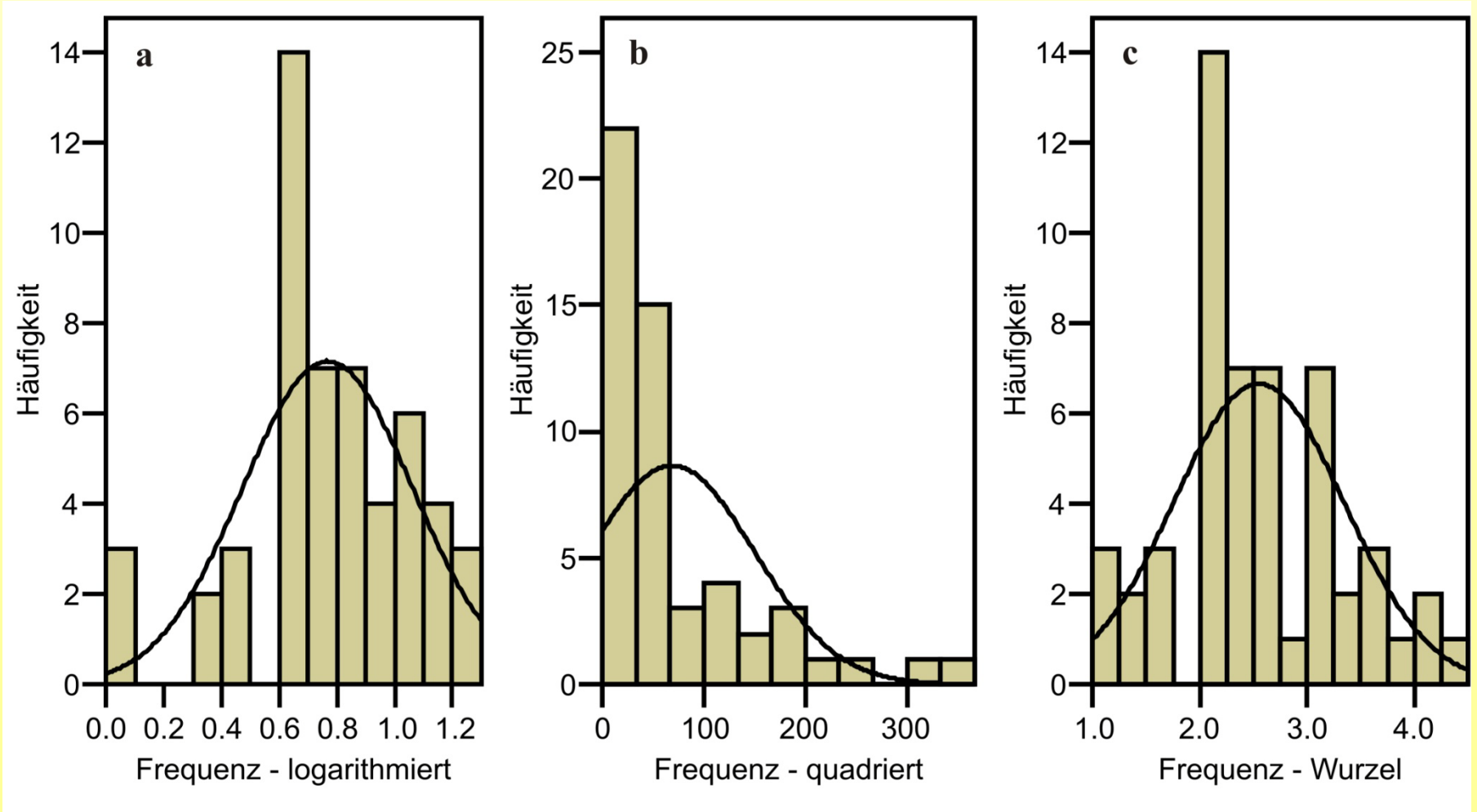


transformations

$x' = log\ (x + a)$

$x' = x^2$

$x' = \sqrt{x}$



Similarity measures

Aim: reduction of multivariate similarity to one or a few values

qualitative similarity

object 2 \ object 1	-present	-absent	
-present	**a**	**b**	a+b
- absent	**c**	**d**	c+d
	a+c	b+d	k

Jaccard index:

Sj = Jaccard similarity

a = shared species of sample 1 and 2

b = number of species in sample 1

c = number of species in sample 2

$$Sj = \frac{a}{a+b+c}$$

d not regarded – asymmetric index

Jaccard-dissimilarity: *Dj = 1.0 – Sj*

(multiplication with 100 yields %-similarity)

Similarity measures

Qualitative similarity

Object 2 \ object 1	-present	-absent	
-present	**a**	**b**	a+b
- absent	**c**	**d**	c+d
	a+c	b+d	k

Sørensen index:

Ss = Sørensen similarity
a = shared species of sample 1 and 2
b = number of species in sample 1
c = number of species in sample 2

$$Ss = \frac{2a}{2a+b+c}$$

Sørensen-dissimilarity: *Dj = 1.0 – Ss*
(multiplication with 100 yields %-similarity)
This dissimilarity is **a semimetric** (Legendre & Legendre 1998)

$D_{1,2} = \sqrt{(1 - S_{1,2})}$ has metric properties

Similarity measures

Quantitative similarity

Bray Curtis index:

$$Sbc = \frac{2w}{A+B}$$

Sbc = Bray Curtis similarity

w = sum of the lower abundance for each pair of shared species of sample 1 and 2

A = sum of the abundance of all species in sample 1

B = sum of the abundance of all species in sample 2

example

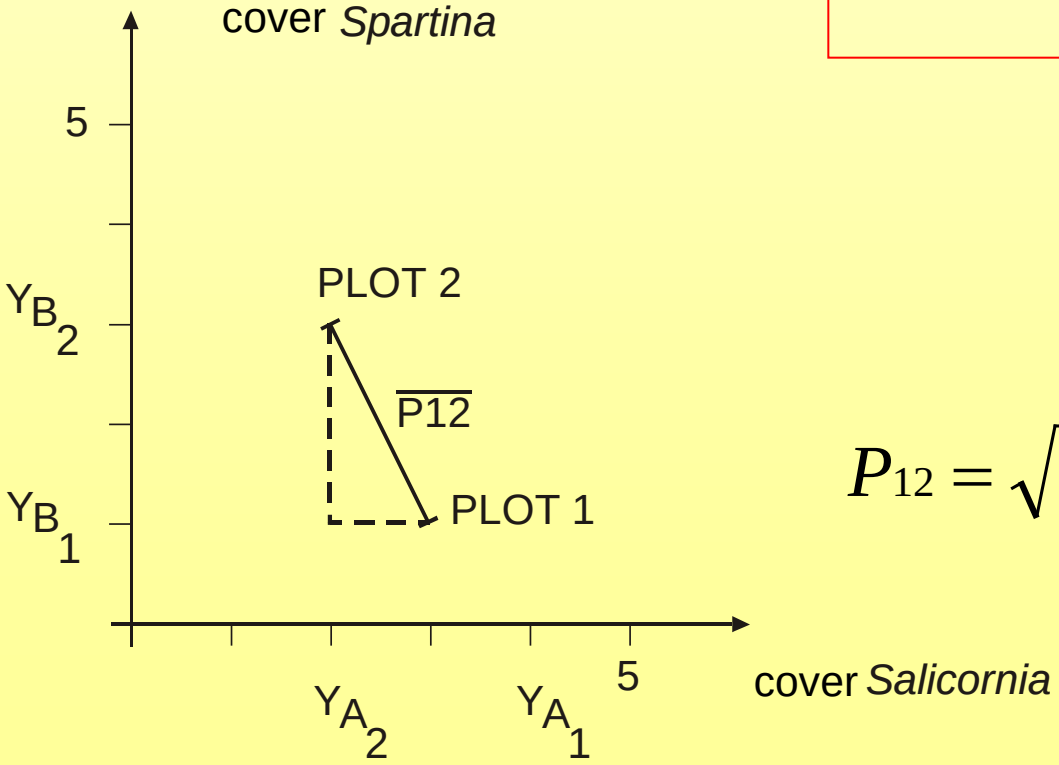
	Spec. 1	Spec. 2	Spec. 3	Spec. 4	Spec. 5	Spec. 6	*B*	*C*	*w*
sample 1	7	3	2	0	4	0	16		
sample 2	4	4	0	0	6	5		19	
minimum	4	3	0	0	4	0			11
							Sbc=2·11/(16+19)=		0.63

Similarity measures

Quantitative similarity (all species, symmetric)

Euclidean distance

$$De_{1,2} = \sqrt{\sum_{k=1}^{m} (y_{1k} - y_{2k})^2}$$



$$P_{12} = \sqrt{(y_{B1} - y_{B2})^2 + (y_{A1} + y_{A2})^2}$$

Similarity measures

Problem of "symmetric" indices in cases with many 0: example 1

	Y1	Y2	Y3
X1	1	3	0
X2	2	2	2
X3	3	1	2

Jaccard-similarity

	X1	X2	X3
X1	1		
X2	0.67	1	
X3	0.67	1	1

Legendre & Legendre 1998

Sørensen-similarity

	X1	X2	X3
x1	1		
x2	0.8	1	
x3	0.8	1	1

Euclidean distance

	X1	X2	X3
x1	0.00		
x2	2.45	0.00	
x3	3.46	1.41	0.00

Similarity measures

Problems with many 0: example 2

	Y1	Y2	Y3
X1	0	1	1
X2	1	0	0
X3	0	4	4

Sørensen-similarity

	x1	x2	x3
X1	1		
x2	0	1	
x3	1	0	1

Euclidean distance

	x1	x2	x3
x1	0.00		
x2	1.73	0.00	
x3	4.24	5.75	0.00

Alternatives: **Nei & Li's *Coefficient*, Gower *General Similarity*, *Manhattan distance***

Overview: Podani 2000, Legendre & Legendre 1998

Similarity measures

An index of indices: Gower‘s similarity

important if variables of different data qualities are to be analysed:

W is weight, *S* respective similarity

$$G_{jk} = \frac{\sum_{i=1}^{n} w_{ijk} s_{ijk}}{\sum_{i=1}^{n} w_{ijk}}$$

S - binary data: 0 or 1

S - continuous variables

$$s_{ijk} = 1 - \frac{|r_{ij} - r_{ik}|}{r_{i.max} - r_{i.min}}$$

S – for ordinal variables (rank-based, *T* no. of objects with ties)

$$s_{ijk} = 1 - \frac{|r_{ij} - r_{ik}| - (T_{ij} - 1)/2 - (T_{ik} - 1)/2}{r_{i.max} - r_{i.min} - (T_{i.max} - 1)/2 - (T_{i.min} - 1)/2}$$

see e.g. package FD (Laliberté and Shipley 2011)

Textbooks

Kent, M. & Coker, P. (1992): vegetation description and analysis - A practical approach. Belhaven Press. London. 363 pp. *Very useful introcution to all methods in vegetation analysis including phytosociology.*

Jongman, R.H. G.; ter Braak, C. J. F. & van Tongeren, O. F. R. (1995): Data analysis in community and landscape ecology. University Press. Cambridge. 299 pp. *Benchmark text: offers help for most questions.*

Gauch, H. G. (1994): multivariate analysis in community ecology. Cambridge University Press. Cambridge. 298 pp. *Old but very readable introduction to all standard methods except CCA/NMDS.*

Lepš J. & Šmilauer, P. (2003): multivariate analysis of ecological data using CANOCO. Cambridge. *Very short introduction to methods im software package Canoco, plus short remarks on classification techniques, useful for advanced questions in Canoco*

(Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2003): Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. Springer, Berlin. *General GERMAN introduction, techniques in ecology not fully covered*)

Textbooks

The previous texts all share a common background / school of thought, the following are more independent:

McCune, B.; Grace, J. B. & Urban, D. L. (2002): Analysis of ecological communities. MjM Software Design. Gleneden Beach. 300 pp. Comprehensive textbook, especially useful in combination with PC-ORD gut beleuchtet.

Legendre, P. & Legendre, L. (1998): Numerical Ecology. Elsevier. Amsterdam. 853 pp. *THE authorative benchmark text, somewhat technical but very very comprehensive.*

Podani, J. (2000): Introduction to the exploration of multivariate biological data. Backhyus Publishers, Leiden. *Extensive treatment from a very independent scientist, very good at classification and similarities*

Borcard, D., Gillet, F. & Legendre, (P. 2011): Numerical ecology with R. Springer, New York, Dordrecht, London, Heidelberg.*Good introduction in multivariate analysis using R*

(Leyer, I. & Wesche, K. (2007): Multivariate Statistik in der Ökologie. - Berlin, Heidelberg, New York. *GERMAN introduction in those techniques widely used in ecology)*


Numerical Ecology
SECOND ENGLISH EDITION
Pierre Legendre
and
Louis Legendre
Developments in
Environmental
Modelling 20
ELSEVIER

Software

Spezialised Software

McCune, B. & Mefford, M. J. PC-ORD. Multivariate Analysis of Ecological Data. Gleneden Beach, Oregon, MJM Software. Version 2006: 5.0

Easy and comprehensive toolbox, full graphical user interface.

ter Braak, C. J. F. & Smilauer, P. CANOCO Reference Manual. Wageningen,Ceske Budejovice: Biometris. Version 2004: 4.5

Widely distributed package for CA and related techniques including advanced statistics, good GUI.

Hammer, Ø. 2011. PAST - PAleontological STatistics. Natural History Museum, University of Oslo, Oslo.

"PAST is a free, easy-to-use data (GUI) analysis package originally aimed at paleontology but now also popular in many other fields"

Oksanen, J., Kindt, R., Legendre, P. & O'Hara, B. 2006. *vegan: Community Ecology Package*, http://cc.oulu.fi/~jarioksa/

example for very useful R package (as always: very comprehensive, and very difficult to use)

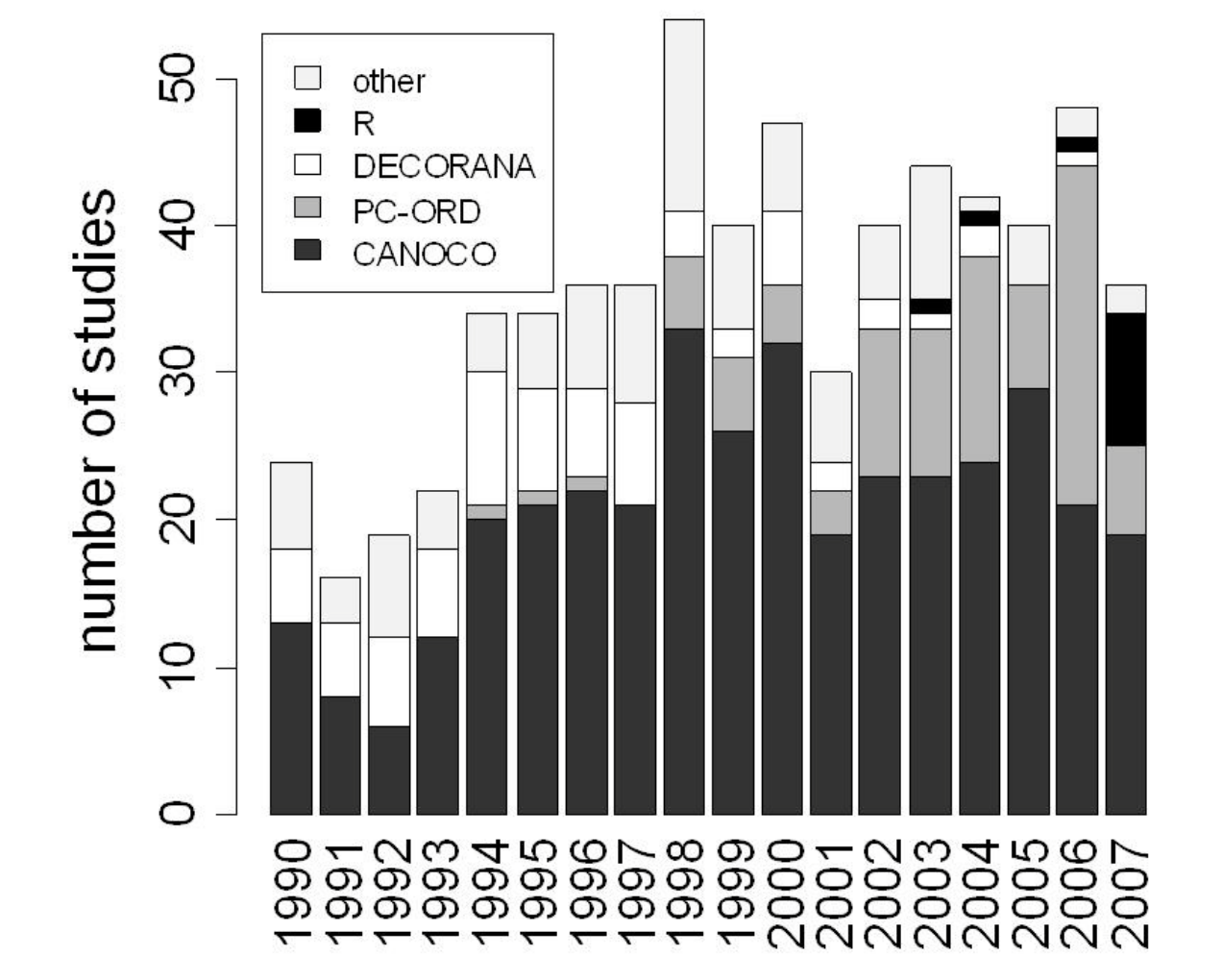
Software

Spezialised Software – overview (Wesche & von Wehrden *JVS* 2009)

Table 1. Current, popularly used software packages and the associated availability of ordination methods; "X" indicates methods available from the early releases, "U" indicates methods added in more recent updates. *ter Braak & Šmilauer (2002). †McCune, B. & Mefford, M.J. (2006). PC-ORD 5.0. MjM Software, Gleneden Beach. ‡Podani (2000). §Kovach (1998). MVSP 3. Kovach Computing Services, Pentraeth. ¶http://cran.r-project.org, Vegan from Oksanen et al. (2006), same as Dixon 2003; due to frequent updates (several times a year) and the open source concept, we only marked the methods currently available. Furthermore, some analyses are available in other packages, e.g. PCA (stats), PCoA (ade4), CCA (ade4). ‖Recently, a routine for NMDS was added to this package (Winkist).

	Decorana/Cornell Ecological Programs	CANOCO*	PC-ORD†	Syntax‡	MVSP§	"R"/Vegan¶
Year of first release	1971/77	1986	1986	1980	1985	1997/2001
(a) *Indirect Ordination*						
Principal Components Analysis (PCA)	X	X	X	X	X	X
Canonical PCA – Redundancy Analysis (RDA)	-	X	-	U	-	X
Correspondence Analysis (CA/RA)	X	X	X	X	X	X
(b) *Direct Ordination*						
Detrended Correspondence Analysis	X	X	X	-	X	X
Canonical Correspondence Analysis (CCA)	-	X	X	U	U	X
Non-Metric Multidimensional Scaling (NMDS)	-	U‖	X	X	-	X
Polar Ordination (Bray-Curtis)	X	-	X	-	-	-
Principal Coordinates Analysis (PCoA)	-	U	-	X	X	X

usage of software



In: *Plant Ecol., AVS, JVS, Folia Geobot., Phytocoenologia*

von Wehrden et al., *Journal of Vegetation Science*, 2009

„In view of the rapidly growing specific functions …R … may soon become the dominating platform for data analysis in vegetation ecology.

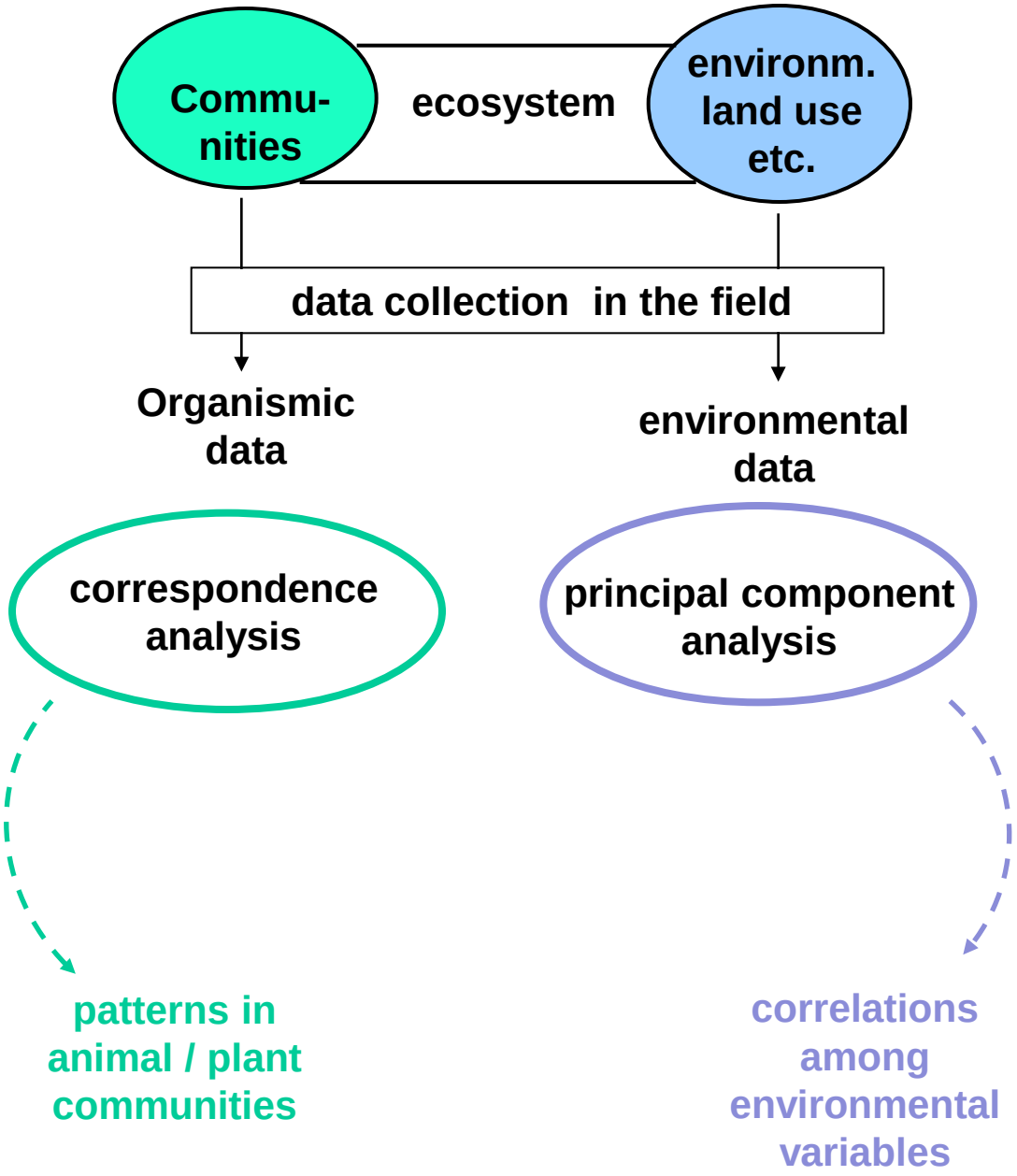
Rumours confirm detailed instructions to be underway“

Wildi, Data analysis… 2010, p. 202

Ordination
Indirect gradient analysis

Environmental gradients are not directly analysed, but instead inferred from patterns in the species composition.

- **Correspondence Analysis (CA, DCA)**
- **Principal Component Analysis (PCA)**
- **Non-metric Multidimensional Scaling (NMDS)**



Analysis of community data

Ordination
Direct gradient analysis

The species composition is analysed in respect to environmental variables

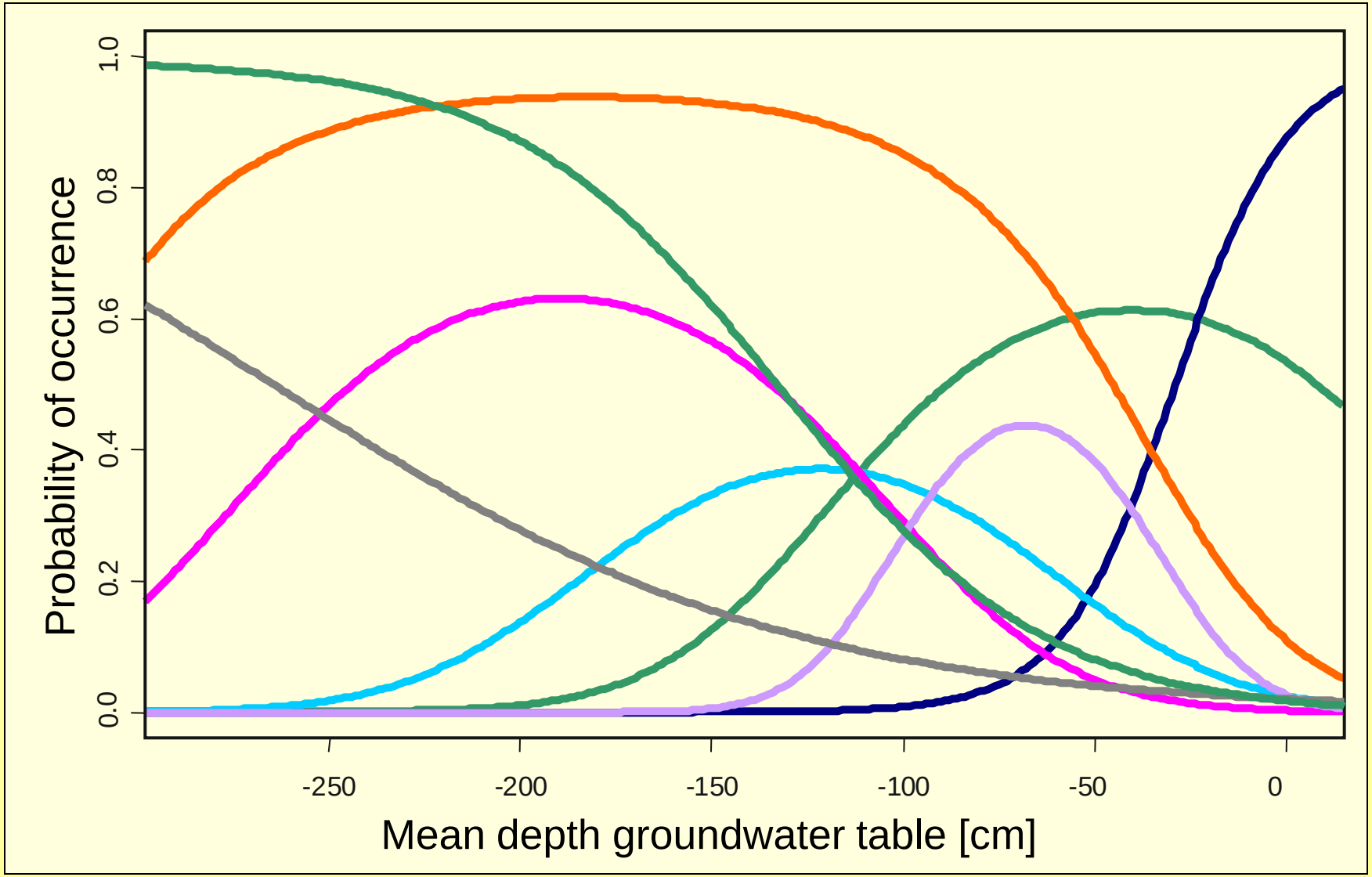
Canonical Correspondence Analysis (CCA)

Redandancy Analysis (RDA)

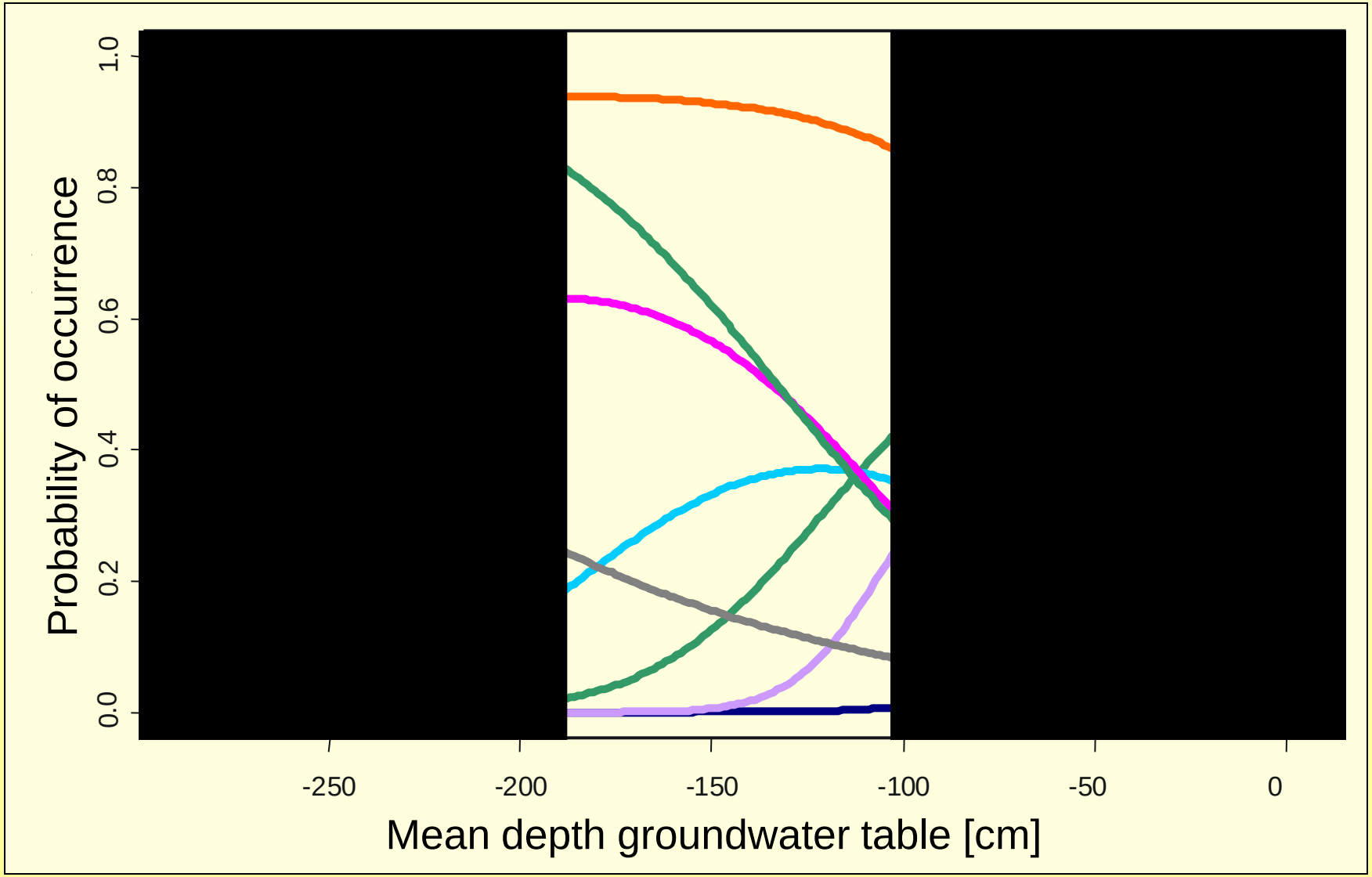
(ecograms / „*Ökogramme*“)

correspondence analysis
(CA)

Species-response-curves

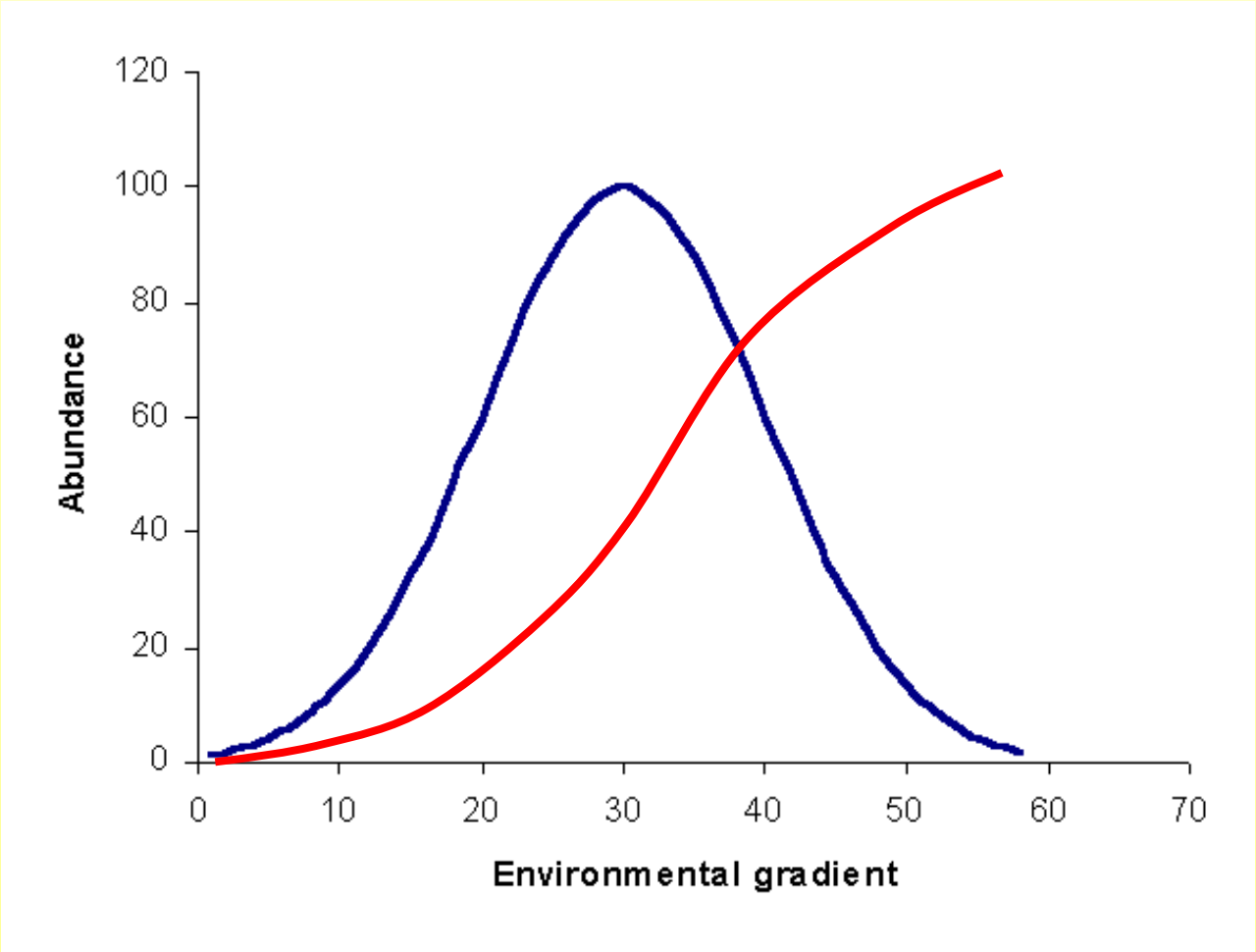


Length of gradient and species-response



Correspondence analysis

Univariate direct gradient analysis



Graph from : webpage of Mike Palmer

Weighted averaging

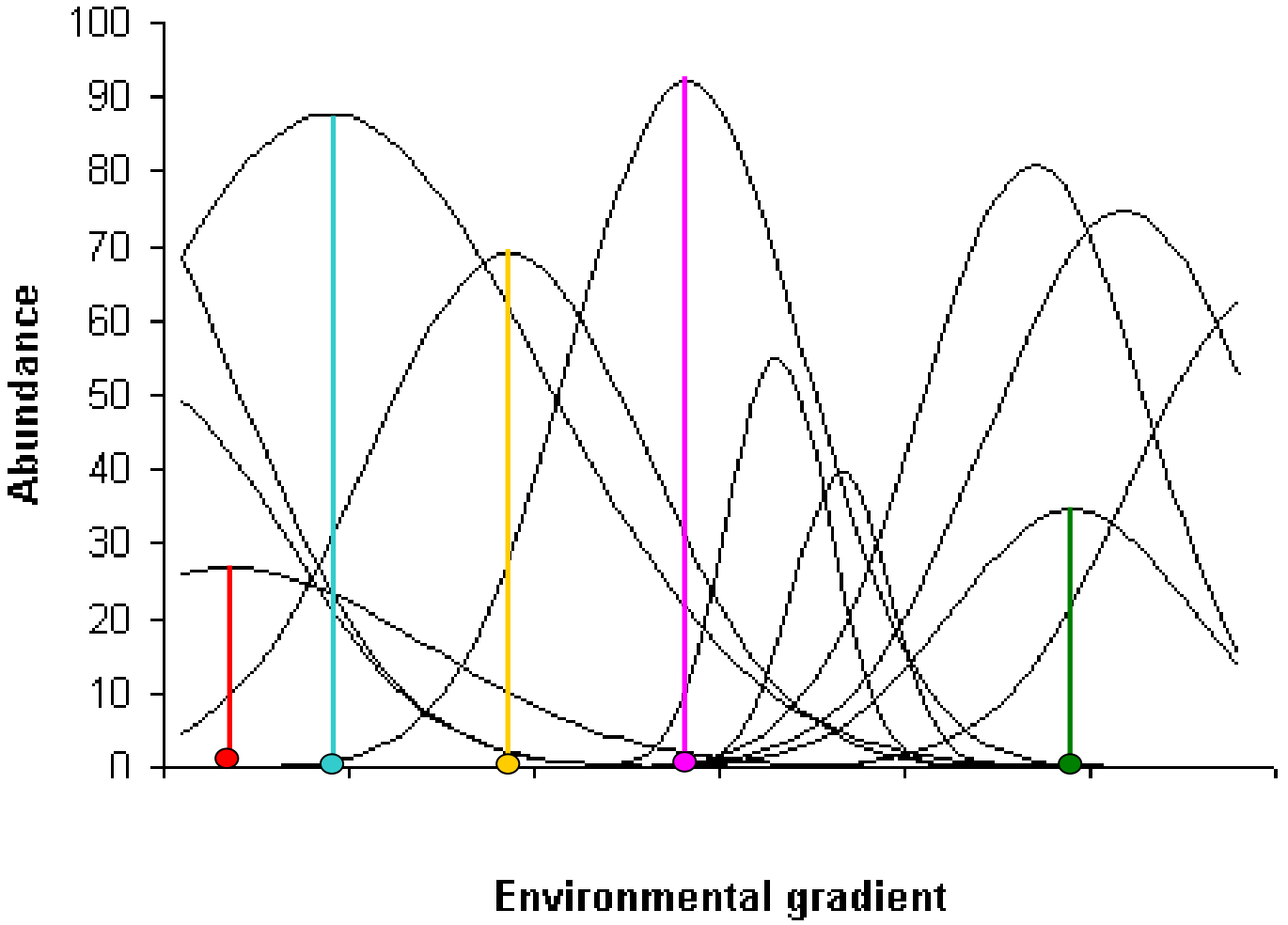
Species that show **unimodal** responses to a given environmental variable are most common near their optimum. The position of the optimum (and thus e.g. the species‘ indicator value) can be found by averaging the value of the environmental variable across all sites where the species occurs.

	x1	x2	x3	x4	x5	x6	x7	x8	x9	M	GM
Y1	0	5	5	9	8	7	2	0	0		
pH	3,5	4,0	4,5	5	5,5	6	6,5	7	7,5	5,25	5,18

***average*: (4,0 + 4,5 + ... + 6,5) / 6 = 5,25**

***weighted average*: (4,0 x 5 + 4,5 x 5 + ...) / 36 = 5,18**

response curves along axis



webpage of Mike Palmer

(CA)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Ach mil	1	3			2	2	2			4							2			
Agr sto			4	8				4	3			4	5	4	4	7				5
Air pra																	2		3	
Alo gen		2	7	2				5	3			8	5			4				
Ant odo					4	3	2			4							4		4	
Bel per		3	2	2	2					2								2		
Bro hor		4		3	2		2			4										
Che alb													1							
Cir arv				2																
Ele pal								4						4	5	8				4
Ely rep	4	4	4	4	4				6											
Emp nig																			2	
Hyp rad											2						2		5	
Jun art								4	4						3	3				4
Jun buf							2		4			4	3							
Leo aut		5	2	2	3	3	3	3	2	3	5	2	2	2	2		2	5	6	2
Lol per	7	5	6	5	2	6	6	4	2	6	7							2		
Pla lan					5	5	5			3	3						2	3		
Poa pra	4	4	5	4	2	3	4	4	4	4	4		2				1	3		
Poa tri	2	7	6	5	6	4	5	4	5	4		4	9			2				
Pot pal														2	2					
Ran fla								2					2	2	2	2				4
Rum ace					5	6	3		2			2								
Sag pro				5				2	2		2	4	2						3	
Sal rep																		3	3	5
Tri pra					2	5	2													
Tri rep		5	2	1	2	5	2	2	3	6	3	3	2	6	1			2	2	
Vic lat										1	2							1		
Bra rut			2	2	2	6	2	2	2	2	4	4			4	4		6	3	4
Cal cus														4		3				3

Dune Meadow – Data (Terschelling)

Jongman et al. (1995): Data analysis in community and landscape ecology

Lüneburg, Oct. 2011

(CA)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Ach mil	1	3			2	2	2			4							2			
Agr sto			4	8				4	3			4	5	4	4	7				5
Air pra																	2		3	
Alo gen		2	7	2				5	3			8	5			4				
Ant odo					4	3	2			4							4		4	
Bel per		3	2	2	2					2								2		
Bro hor		4		3	2		2			4										
Che alb													1							
Cir arv				2																
Ele pal								4						4	5	8				4
Ely rep	4	4	4	4	4				6											
Emp nig																			2	
Hyp rad											2						2		5	
Jun art								4	4						3	3				4
Jun buf							2		4			4	3							
Leo aut		5	2	2	3	3	3	3	2	3	5	2	2	2	2		2	5	6	2
Lol per	7	5	6	5	2	6	6	4	2	6	7							2		
Pla lan					5	5	5			3	3						2	3		
Poa pra	4	4	5	4	2	3	4	4	4	4	4		2				1	3		
Poa tri	2	7	6	5	6	4	5	4	5	4		4	9			2				
Pot pal														2	2					
Ran fla								2					2	2	2	2				4
Rum ace					5	6	3		2			2								
Sag pro				5				2	2		2	4	2						3	
Sal rep																		3	3	5
Tri pra					2	5	2													
Tri rep		5	2	1	2	5	2	2	3	6	3	3	2	6	1			2	2	
Vic lat										1	2							1		
Bra rut			2	2	2	6	2	2	2	2	4	4			4	4		6	3	4
Cal cus														4		3				3
Moisture	**1**	**1**	**2**	**2**	**1**	**1**	**1**	**5**	**4**	**2**	**1**	**4**	**5**	**5**	**5**	**5**	**2**	**1**	**5**	**5**

Dune Meadow – Data (Terschelling) with moisture levels

Jongman et al. (1995): Data analysis in community and landscape ecology

Lüneburg, Oct. 2011

	1	2	5	6	7	11	18	3	4	10	17	9	12	8	13	14	15	16	19	20	
Ach mil	1	3	2	2	2					4	2										1,4
Agr sto								4	8			3	4	4	5	4	4	7		5	4,1
Air pra											2								3		3,8
Alo gen		2						7	2			3	8	5	5			4			3,7
Ant odo			4	3	2					4	4								4		2,1
Bel per		3	2				2	2	2	2											
Bro hor		4	2		2				3	4											
Che alb															1						
Cir arv									2												
Ele pal														4		4	5	8		4	
Ely rep	4	4	4					4	4			6									
Emp nig																			2		
Hyp rad						2					2								5		
Jun art												4		4			3	3		4	
Jun buf					2							4	4		3						
Leo aut		5	3	3	3	5	5	2	2	3	2	2	2	3	2	2	2		6	2	
Lol per	7	5	2	6	6	7	2	6	5	6		2		4							
Pla lan			5	5	5	3	3			3	2										
Poa pra	4	4	2	3	4	4	3	5	4	4	1	4		4	2						
Poa tri	2	7	6	4	5			6	5	4		5	4	4	9			2			
Pot pal																2	2				
Ran fla														2	2	2	2	2		4	
Rum ace			5	6	3							2	2								
Sag pro						2			5			2	4	2	2				3		
Sal rep							3												3	5	
Tri pra			2	5	2																
Tri rep		5	2	5	2	3	2	2	1	6		3	3	2	2	6	1		2		
Vic lat						2	1			1											
Bra rut			2	6	2	4	6	2	2	2		2	4	2			4	4	3	4	
Cal cus																4		3		3	
Moisture	1	1	1	1	1	1	1	2	2	2	2	4	4	5	5	5	5	5	5	5	

(CA)

species scores

Weighted averaging-algorithm

sample scores

Lüneburg, Oct. 2011

(CA)

	1	2	5	6	7	11	18	3	4	10	17	9	12	8	13	14	15	16	19	20	
Tri pra			2	5	2																**1,0**
Pla lan			5	5	5	3	3			3	2										**1,2**
Vic lat						2	1			1											**1,3**
Ach mil	1	3	2	2	2					4	2										**1,4**
Bel per		3	2				2	2	2	2											**1,5**
Bro hor		4	2		2				3	4											**1,5**
Rum ace			5	6	3							2	2								**1,7**
Lol per	7	5	2	6	6	7	2	6	5	6		2		4							**1,7**
Cir arv									2												**2,0**
Ely rep	4	4	4					4	4			6									**2,0**
Poa pra	4	4	2	3	4	4	3	5	4	4	1	4		4	2						**2,0**
Ant odo			4	3	2					4	4								4		**2,1**
Poa tri	2	7	6	4	5			6	5	4		5	4	4	9			2			**2,6**
Leo aut		5	3	3	3	5	5	2	2	3	2	2	2	3	2	2	2		6	2	**2,6**
Tri rep		5	2	5	2	3	2	2	1	6		3	3	2	2	6	1		2		**2,7**
Bra rut			2	6	2	4	6	2	2	2		2	4	2			4	4	3	4	**2,9**
Hyp rad						2					2								5		**3,4**
Sag pro						2			5			2	4	2	2				3		**3,5**
Alo gen		2						7	2			3	8	5	5			4			**3,7**
Jun buf					2							4	4		3						**3,8**
Air pra											2								3		**3,8**
Sal rep							3												3	5	**3,9**
Agr sto								4	8			3	4	4	5	4	4	7		5	**4,1**
Jun art												4		4			3	3		4	**4,8**
Che alb															1						**5,0**
Ele pal														4		4	5	8		4	**5,0**
Emp nig																			2		**5,0**
Pot pal																2	2				**5,0**
Ran fla														2	2	2	2	2		4	**5,0**
Cal cus																4		3		3	**5,0**
Moisture	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**2**	**4**	**4**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5,0**

Arrange samples along diagonal by *weighted averaging*

Jongman et al. (1995): Data analysis in community and landscape ecology

(CA)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Ach mil	1	3			2	2	2			4							2			
Agr sto			4	8				4	3			4	5	4	4	7				5
Air pra																	2		3	
Alo gen		2	7	2				5	3			8	5			4				
Ant odo					4	3	2			4							4		4	
Bel per		3	2	2	2					2								2		
Bro hor		4		3	2		2			4										
Che alb													1							
Cir arv				2																
Ele pal								4						4	5	8				4
Ely rep	4	4	4	4	4				6											
Emp nig																			2	
Hyp rad											2						2		5	
Jun art								4	4						3	3				4
Jun buf							2		4			4	3							
Leo aut		5	2	2	3	3	3	3	2	3	5	2	2	2	2		2	5	6	2
Lol per	7	5	6	5	2	6	6	4	2	6	7							2		
Pla lan					5	5	5			3	3						2	3		
Poa pra	4	4	5	4	2	3	4	4	4	4	4		2				1	3		
Poa tri	2	7	6	5	6	4	5	4	5	4		4	9			2				
Pot pal														2	2					
Ran fla								2					2	2	2	2				4
Rum ace					5	6	3		2			2								
Sag pro				5				2	2		2	4	2						3	
Sal rep																		3	3	5
Tri pra					2	5	2													
Tri rep		5	2	1	2	5	2	2	3	6	3	3	2	6	1			2	2	
Vic lat										1	2							1		
Bra rut			2	2	2	6	2	2	2	2	4	4			4	4		6	3	4
Cal cus														4		3				3
sample scores	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**

First step correspondence analysis: selection of random (but unequal) sample scores

Jongman et al. (1995): Data analysis in community and landscape ecology
Lüneburg, Oct. 2011

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Ach mil	1	3			2	2	2			4							2				7.31
Agr sto			4	8				4	3			4	5	4	4	7				5	11.33
Air pra																	2		3		18.20
Alo gen		2	7	2				5	3			8	5			4					9.03
Ant odo					4	3	2			4							4		4		11.24
Bel per		3	2	2	2					2								2			6.62
Bro hor		4		3	2		2			4											5.60
Che alb													1								13.00
Cir arv				2																	4.00
Ele pal								4						4	5	8				4	14.84
Ely rep	4	4	4	4	4				6												4.38
Emp nig																			2		19.00
Hyp rad											2						2		5		16.78
Jun art								4	4						3	3				4	13.39
Jun buf							2		4			4	3								10.54
Leo aut		5	2	2	3	3	3	3	2	3	5	2	2	2	2		2	5	6	2	10.94
Lol per	7	5	6	5	2	6	6	4	2	6	7							2			6.31
Pla lan					5	5	5			3	3						2	3			9.27
Poa pra	4	4	5	4	2	3	4	4	4	4	4		2				1	3			7.25
Poa tri	2	7	6	5	6	4	5	4	5	4		4	9			2					7.25
Pot pal														2	2						14.50
Ran fla								2					2	2	2	2				4	15.14
Rum ace					5	6	3		2			2									6.89
Sag pro				5				2	2		2	4	2						3		10.35
Sal rep																		3	3	5	19.18
Tri pra					2	5	2														6.00
Tri rep		5	2	1	2	5	2	2	3	6	3	3	2	6	1			2	2		9.47
Vic lat										1	2							1			12.50
Bra rut			2	2	2	6	2	2	2	2	4	4			4	4		6	3	4	12.02
Cal cus														4		3				3	16.40
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	

(CA)

Second step correspondence analysis: calculation of new species scores

Jongman et al. (1995): Data analysis in community and landscape ecology

Lüneburg, Oct. 2011

(CA)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Ach mil	1	3			2	2	2			4							2				7.31
Agr sto			4	8				4	3			4	5	4	4	7				5	11.33
Air pra																	2		3		18.20
Alo gen		2	7	2				5	3			8	5			4					9.03
Ant odo					4	3	2			4							4		4		11.24
Bel per		3	2	2	2					2								2			6.62
Bro hor		4		3	2		2			4											5.60
Che alb													1								13.00
Cir arv				2																	4.00
Ele pal								4						4	5	8				4	14.84
Ely rep	4	4	4	4	4				6												4.38
Emp nig																			2		19.00
Hyp rad											2						2		5		16.78
Jun art								4	4						3	3				4	13.39
Jun buf							2		4			4	3								10.54
Leo aut		5	2	2	3	3	3	3	2	3	5	2	2	2	2		2	5	6	2	10.94
Lol per	7	5	6	5	2	6	6	4	2	6	7							2			6.31
Pla lan					5	5	5			3	3						2	3			9.27
Poa pra	4	4	5	4	2	3	4	4	4	4	4		2				1	3			7.25
Poa tri	2	7	6	5	6	4	5	4	5	4		4	9			2					7.25
Pot pal														2	2						14.50
Ran fla								2					2	2	2	2				4	15.14
Rum ace					5	6	3		2			2									6.89
Sag pro				5				2	2		2	4	2						3		10.35
Sal rep																		3	3	5	19.18
Tri pra					2	5	2														6.00
Tri rep		5	2	1	2	5	2	2	3	6	3	3	2	6	1			2	2		9.47
Vic lat										1	2							1			12.50
Bra rut			2	2	2	6	2	2	2	2	4	4			4	4		6	3	4	12.02
Cal cus														4		3				3	16.40
sample scores	6.25	7.46	8.04	8.18	7.93	8.48	8.18	10.34	8.80	8.37	9.73	9.78	9.64	12.84	12.97	12.62	11.81	10.79	13.89	14.36	

Third step correspondence analysis: calculation new sample scores

Reciprocal averaging

Jongman et al. (1995): Data analysis in community and landscape ecology

Lüneburg, Oct. 2011

(CA)

	1	2	5	3	4	7	10	6	9	13	11	12	8	18	17	16	14	15	19	20	
Cir arv					2																**4.00**
Ely rep	4	4	4	4	4				6												**4.38**
Bro hor		4	2		3	2	4														**5.60**
Tri pra			2			2		5													**6.00**
Lol per	7	5	2	6	5	6	6	6	2		7		4	2							**6.31**
Bel per		3	2	2	2		2							2							**6.62**
Rum ace			5			3		6	2			2									**6.89**
Poa pra	4	4	2	5	4	4	4	3	4	2	4		4	3	1						**7.25**
Poa tri	2	7	6	6	5	5	4	4	5	9		4	4			2					**7.25**
Ach mil	1	3	2			2	4	2							2						**7.31**
Alo gen		2		7	2				3	5		8	5			4					**9.03**
Pla lan			5			5	3	5			3			3	2						**9.27**
Tri rep		5	2	2	1	2	6	5	3	2	3	3	2	2			6	1	2		**9.47**
Sag pro					5				2	2	2	4	2						3		**10.35**
Jun buf						2			4	3		4									**10.54**
Leo aut		5	3	2	2	3	3	3	2	2	5	2	3	5	2		2	2	6	2	**10.94**
Ant odo			4			2	4	3							4				4		**11.24**
Agr sto				4	8				3	5		4	4			7	4	4		5	**11.33**
Bra rut			2	2	2	2	2	6	2		4	4	2	6		4		4	3	4	**12.02**
Vic lat							1				2			1							**12.50**
Che alb										1											**13.00**
Jun art									4				4			3		3		4	**13.39**
Pot pal																	2	2			**14.50**
Ele pal													4			8	4	5		4	**14.84**
Ran fla										2			2			2	2	2		4	**15.14**
Cal cus																3	4			3	**16.40**
Hyp rad											2				2				5		**16.78**
Air pra															2				3		**18.20**
Emp nig																			2		**19.00**
Sal rep														3					3	5	**19.18**
sample scores	6 2 5	7 4 6	7 9 3	8 0 4	8 1 8	8 1 8	8 3 7	8 4 8	8 8 0	9 6 4	9 7 3	9 7 8	10 . 3 4	10 . 7 9	11 . 8 1	12 . 6 2	12 . 8 4	12 . 9 7	13 . 8 9	14 . 3 6	

Arrange after first iteration

Jongman et al. (1995): Data analysis in community and landscape ecology
Lüneburg, Oct. 2011

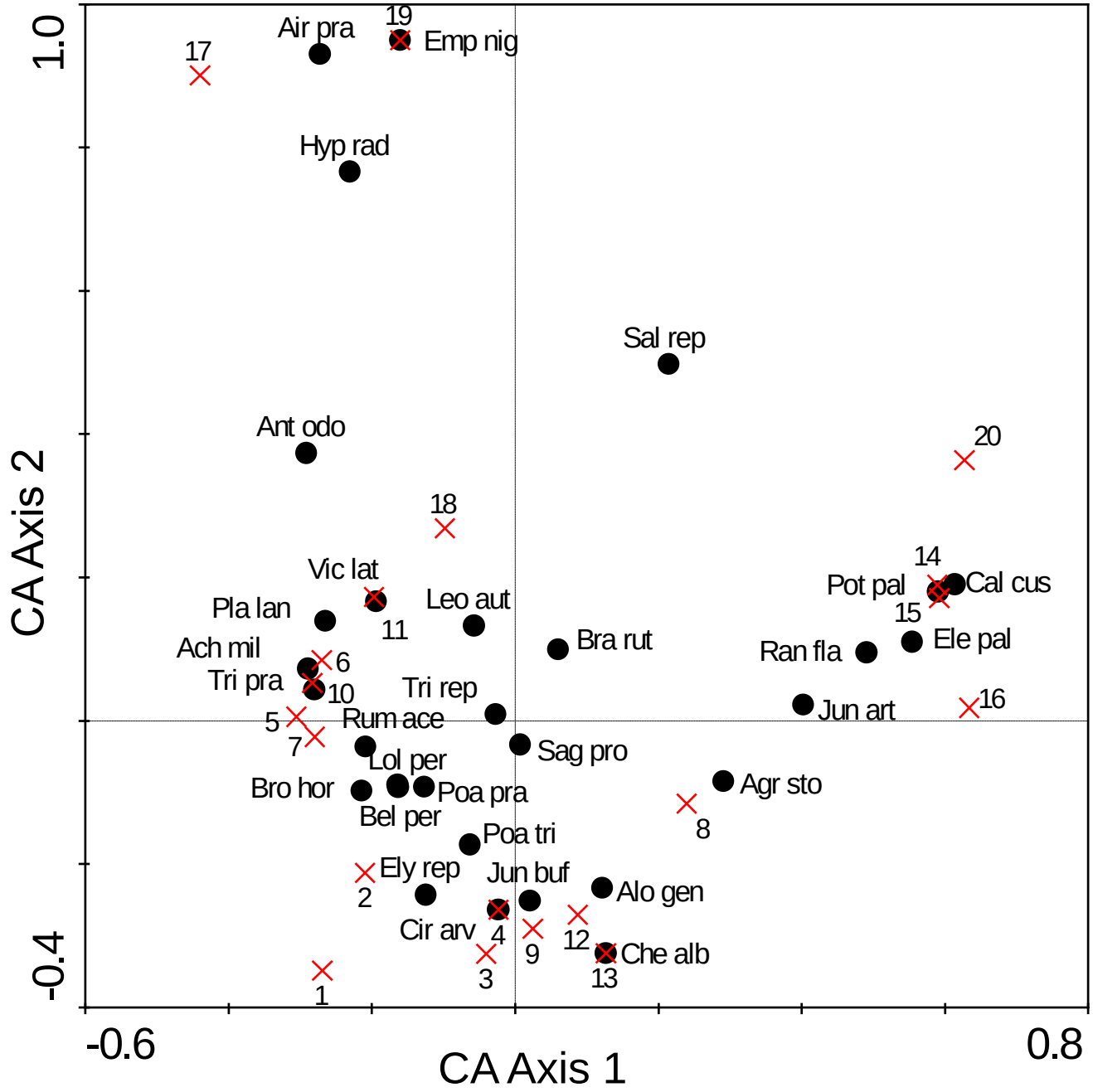
(CA)

	17	5	10	7	6	1	19	11	2	18	3	4	9	12	13	8	15	14	20	16	
Air pra	2						3														-0,99
Ant odo	4	4	4	2	3		4														-0,96
Ach mil	2	2	4	2	2	1			3												-0,91
Tri pra		2		2	5																-0,88
Hyp rad	2						5	2													-0,84
Pla lan	2	5	3	5	5			3		3											-0,84
Emp nig							2														-0,67
Bro hor		2	4	2					4			3									-0,66
Rum ace		5		3	6								2	2							-0,65
Vic lat			1					2		1											-0,62
Bel per		2	2						3	2	2	2									-0,50
Lol per		2	6	6	6	7		7	5	2	6	5	2			4					-0,50
Poa pra	1	2	4	4	3	4		4	4	3	5	4	4		2	4					-0,39
Ely rep		4				4			4		4	4	6								-0,37
Leo aut	2	3	3	3	3		6	5	5	5	2	2	2	2	2	3	2	2	2		-0,19
Poa tri		6	4	5	4	2			7		6	5	5	4	9	4				2	-0,18
Tri rep		2	6	2	5		2	3	5	2	2	1	3	3	2	2	1	6			-0,08
Cir arv												2									-0,06
Sag pro							3	2				5	2	4	2	2					0,00
Jun buf				2									4	4	3						0,08
Bra rut		2	2	2	6		3	4		6	2	2	2	4		2	4		4	4	0,18
Alo gen									2		7	2	3	8	5	5				4	0,40
Che alb															1						0,42
Sal rep							3			3									5		0,62
Agr sto											4	8	3	4	5	4	4	4	5	7	0,93
Jun art													4			4	3		4	3	1,28
Ran fla															2	2	2	2	4	2	1,56
Ele pal																4	5	4	4	8	1,77
Pot pal																	2	2			1,92
Cal cus																		4	3	3	1,96
sample scores	-1,46	-0,95	-0,88	-0,87	-0,86	-0,82	-0,68	-0,64	-0,64	-0,31	-0,11	-0,06	0,09	0,28	0,42	0,76	0,92	0,92	0,95	1,00	

Next steps correspondence analysis: iterative circles of reciprocal averaging until values have stabilised

Jongman et al. (1995): Data analysis in community and landscape ecology

Lüneburg, Oct. 2011



CA Biplot Dune Meadow Data

Jongman et al. (1995): Data analysis in community and landscape ecology

Lüneburg, Oct. 2011

steps two-way weighted averaging-algorithm of the CA

1. axis:

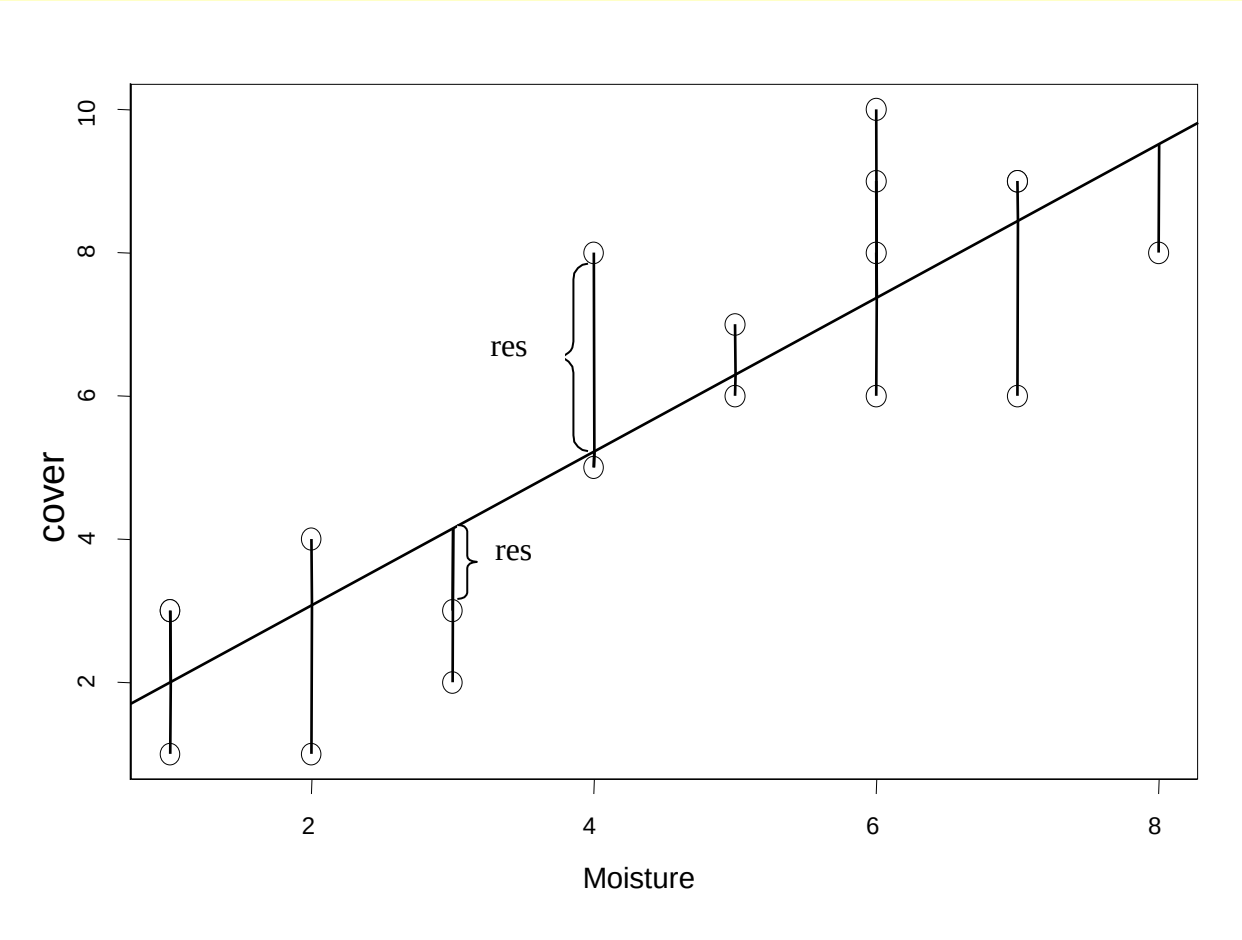
- **random selection of (unequal) sample scores**
- **calculate new species scores by weighted averaging of sample scores**
- **calculate new sample scores by weighted averaging of species scores**
- **standardise sample scores**
- **stop process once new scores are similar to previous scores**

steps two-way weighted averaging-algorithm of the CA

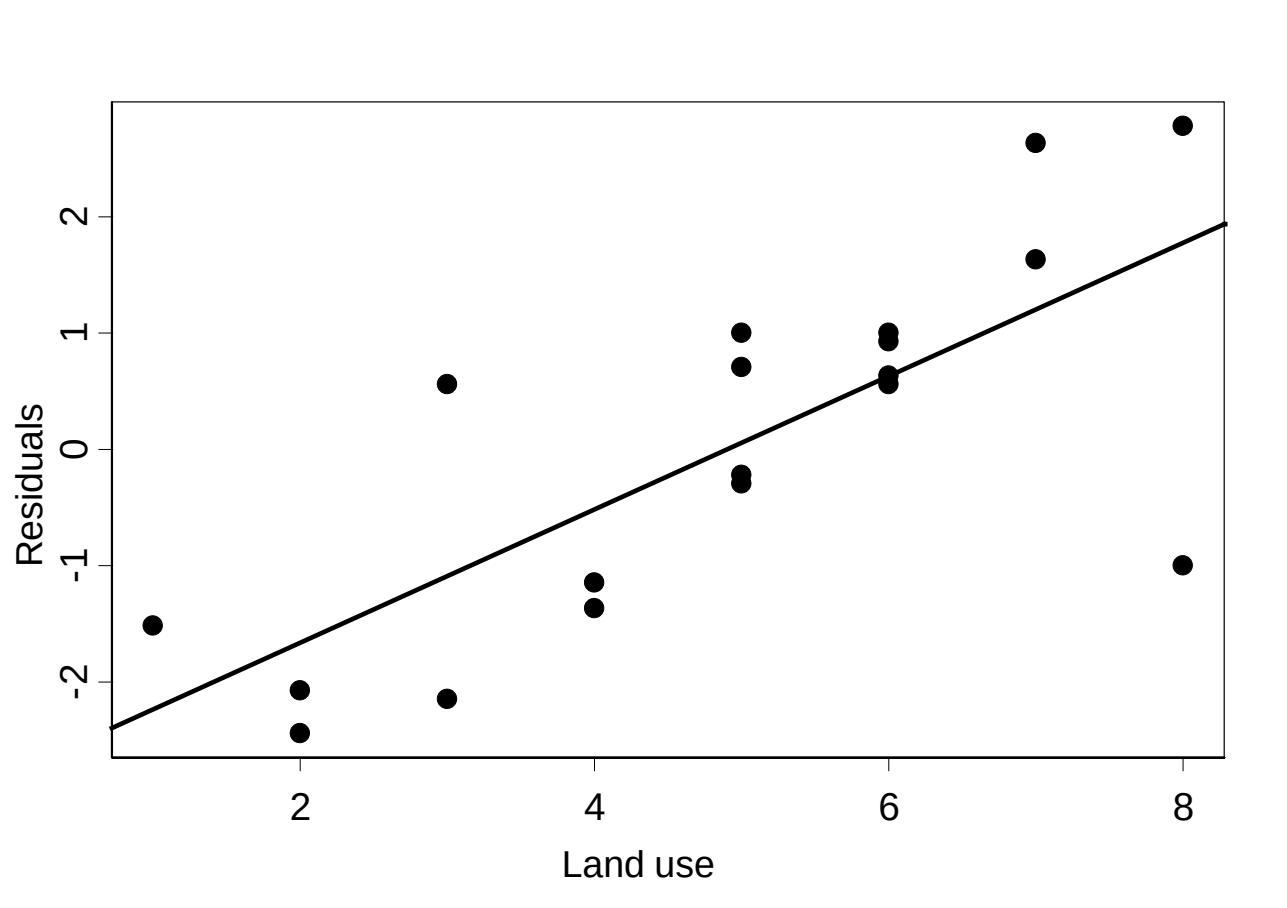
2. axis:

- **random selection of (unequal) sample scores**
- **calculate new species scores by weighted averaging of sample scores**
- **calculate new sample scores by weighted averaging of species scores**
- **make sample scores uncorrelated (orthogonal) to those of previous axis**
- **standardise sample scores**
- **stop process once new scores are similar to previous scores**

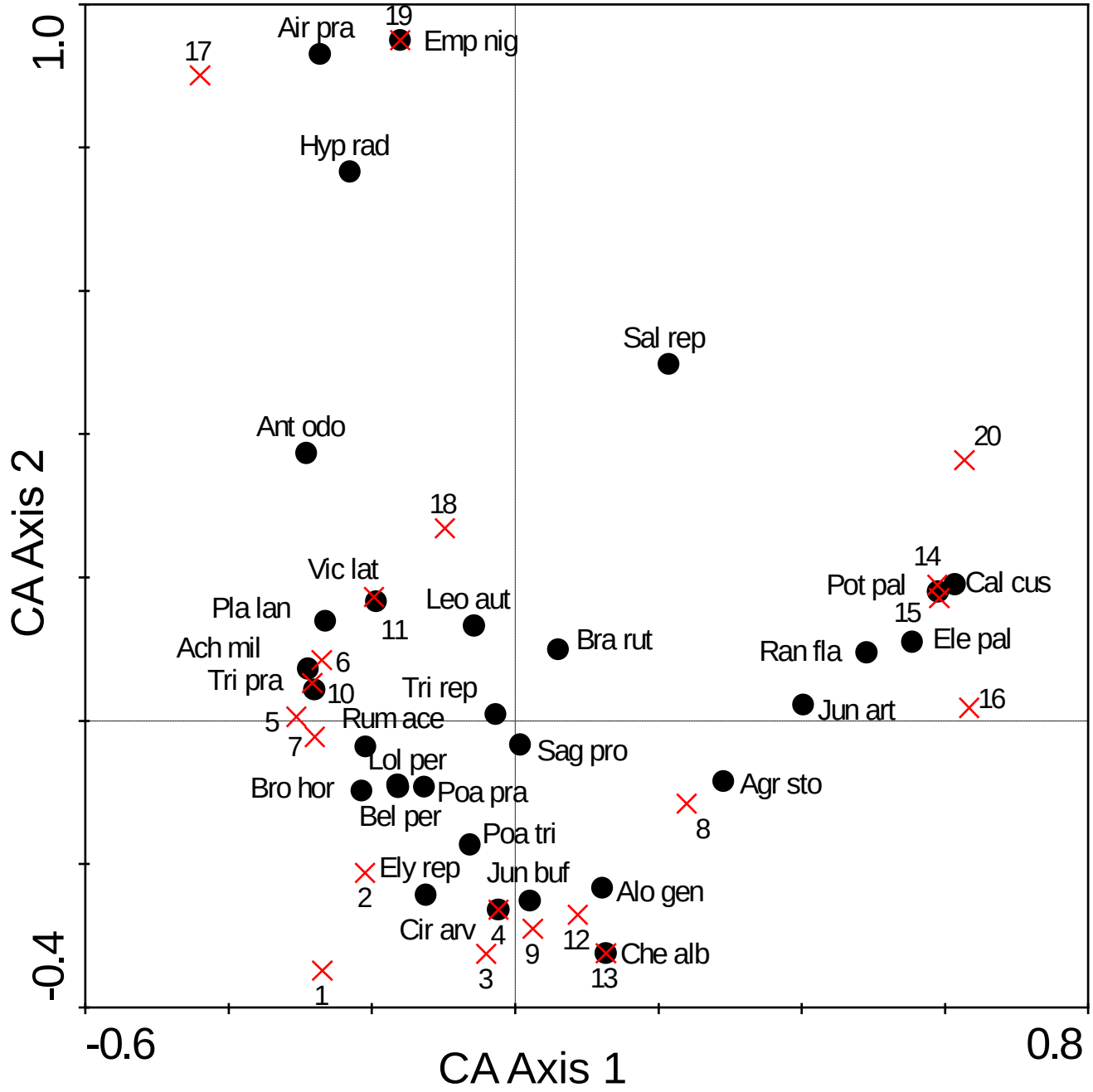
Principle orthogonality: regression species' abundance ~soil moisture



regression residuals ~land use intensity



-> correlation of residuals ~ moisture now zero



CA Biplot Dune Meadow Data

CA is implicitly based on **chi-square similarities** among objects, which results in unwanted upweighting of rare species, can be corrected by **down-weighting** them *a priori*

Jongman et al. (1995): Data analysis in community and landscape ecology

Lüneburg, Oct. 2011

ordination diagnostics

The **Eigenvalue** is a measure for the strength of the ordination axis (**Eigenvector**). It is usually scaled between 0 and 1. The larger the eigenvalue, the better is the spread of species/sample scores along the axis and the more important is the axis for capturing variation in the communities. The first axis has the largest EV, followed by the second etc. (exception NMDS).
The eigenvalue is important to asses the relevance of an axis within a given analysis, not across different data sets.

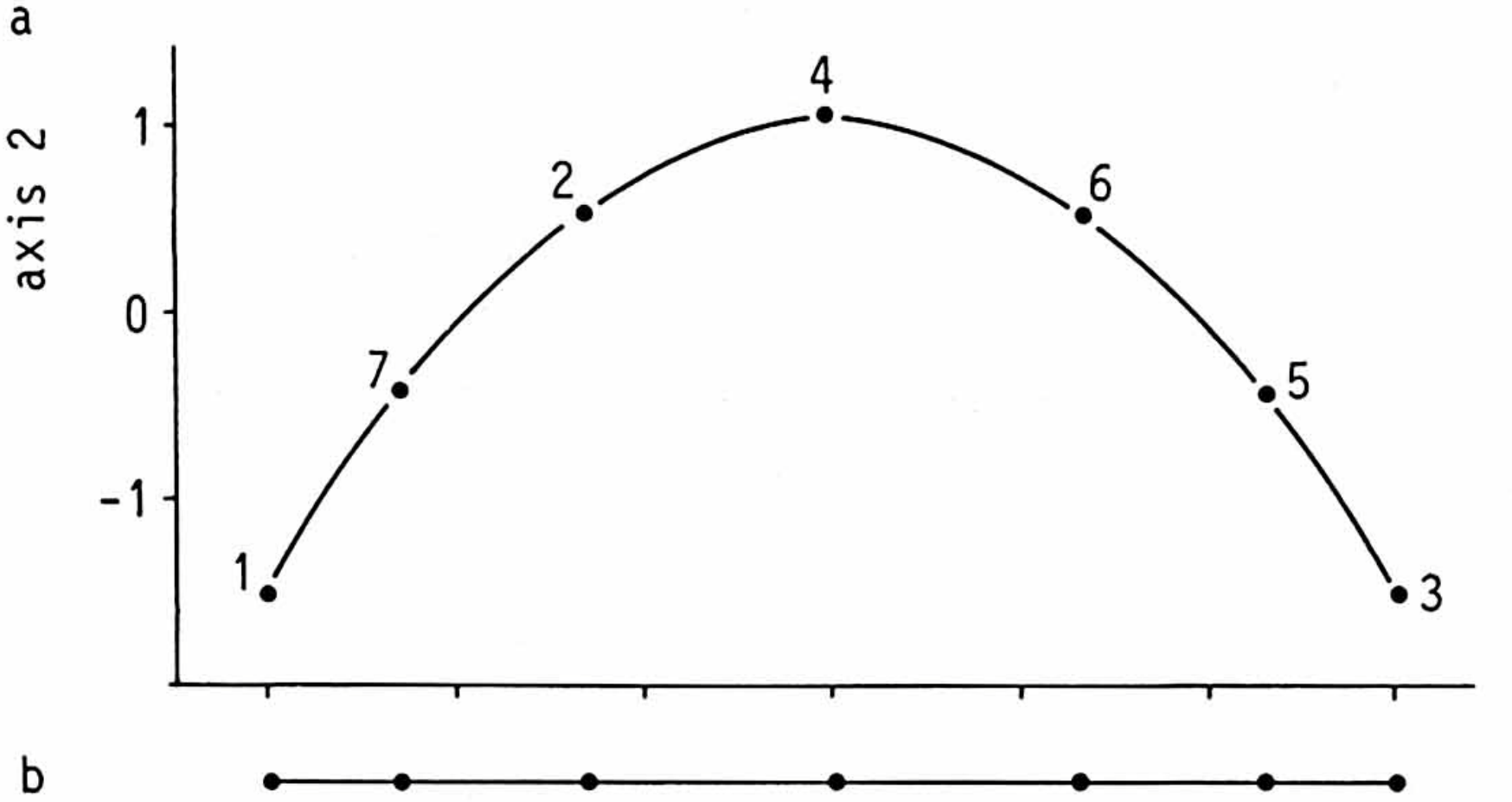
In **unimodal methods**, the **total inertia** is a measure of the total variance in the species data and equals the sum of all Eigenvalues (of e.g. the CA).

Two-way Petrie-matrix

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Art 1	1									
Art2	1	1								
Art 3	1	1	1							
Art 4		1	1	1						
Art 5			1	1	1					
Art 6				1	1	1				
Art 7					1	1	1			
Art 8						1	1	1		
Art 9							1	1	1	
Art 10								1	1	1
Art 11									1	1
Art 12										1

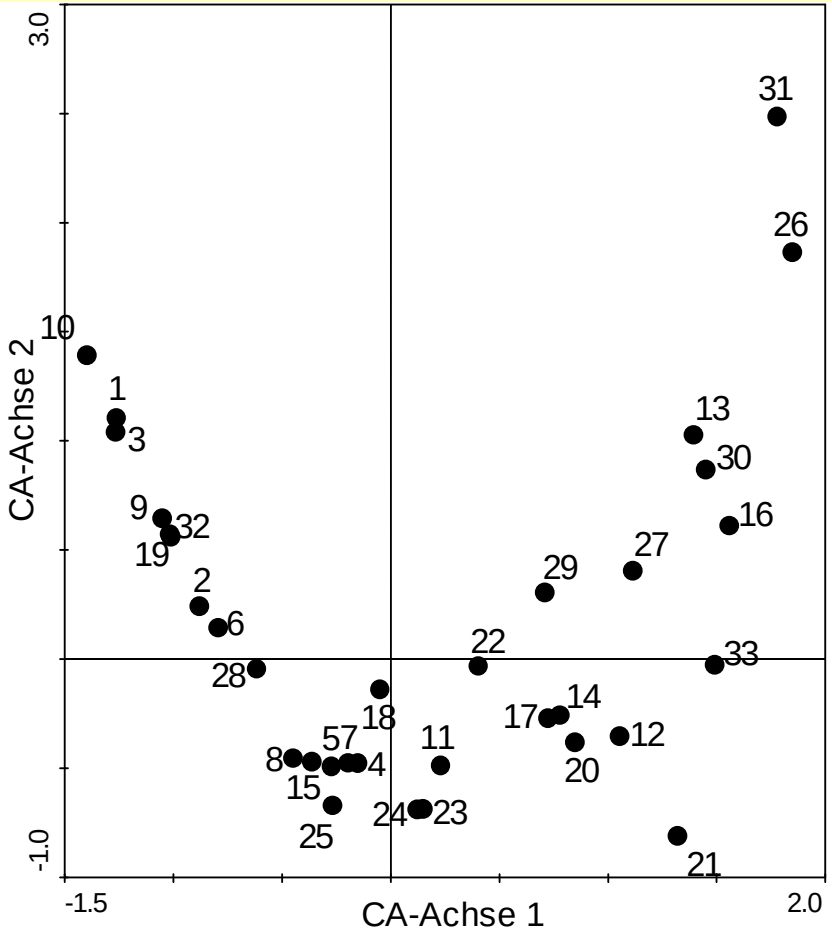
Jongman et al. (1995): Data analysis in community and landscape ecology

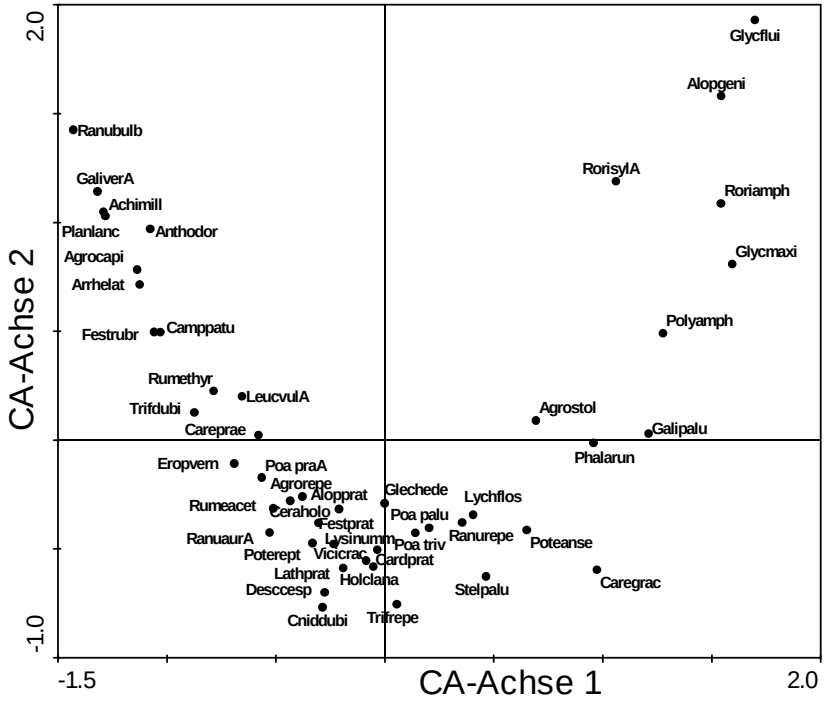
Arch-effect



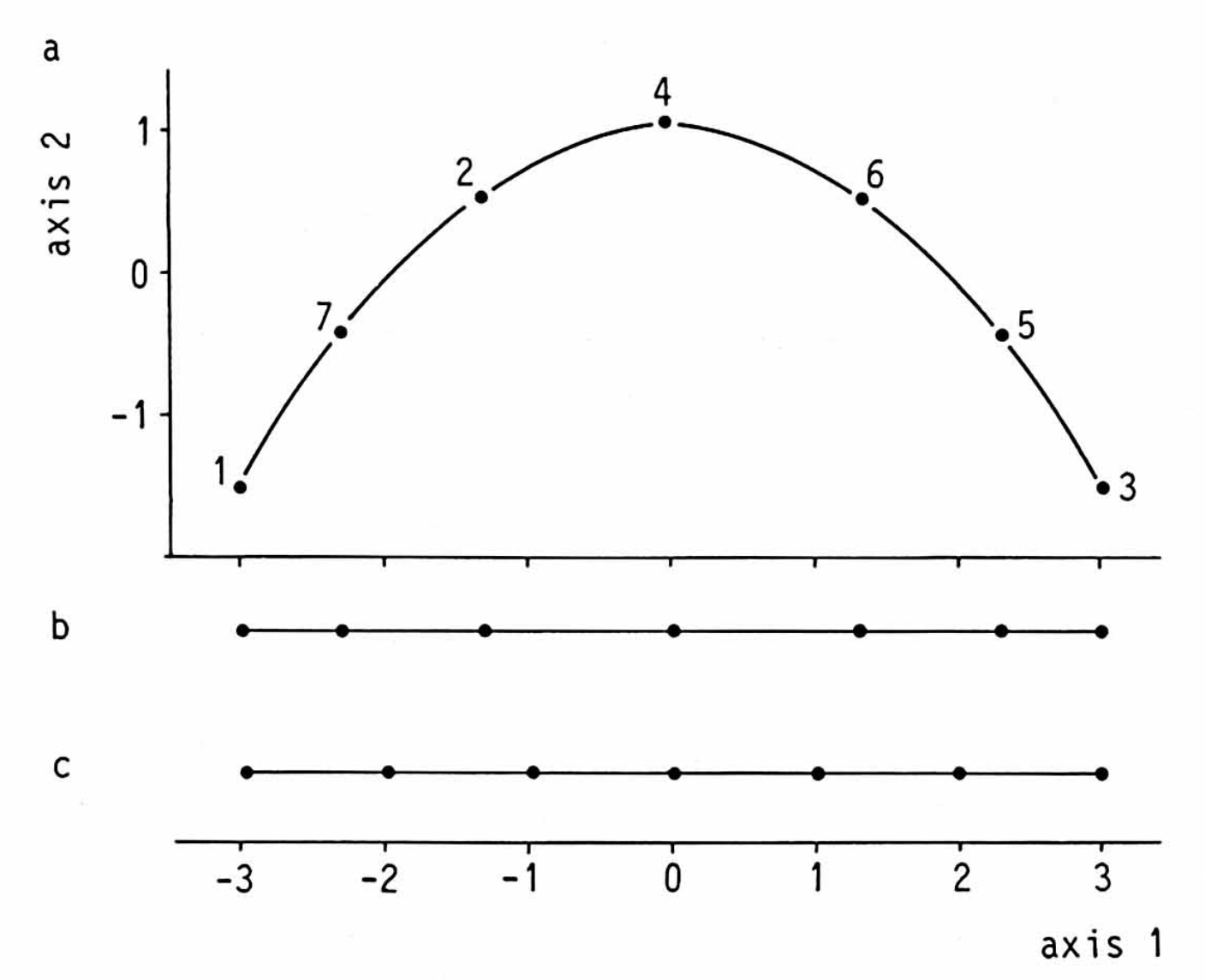
Jongman et al. (1995): Data analysis in community and landscape ecology

Arch-effect for real data



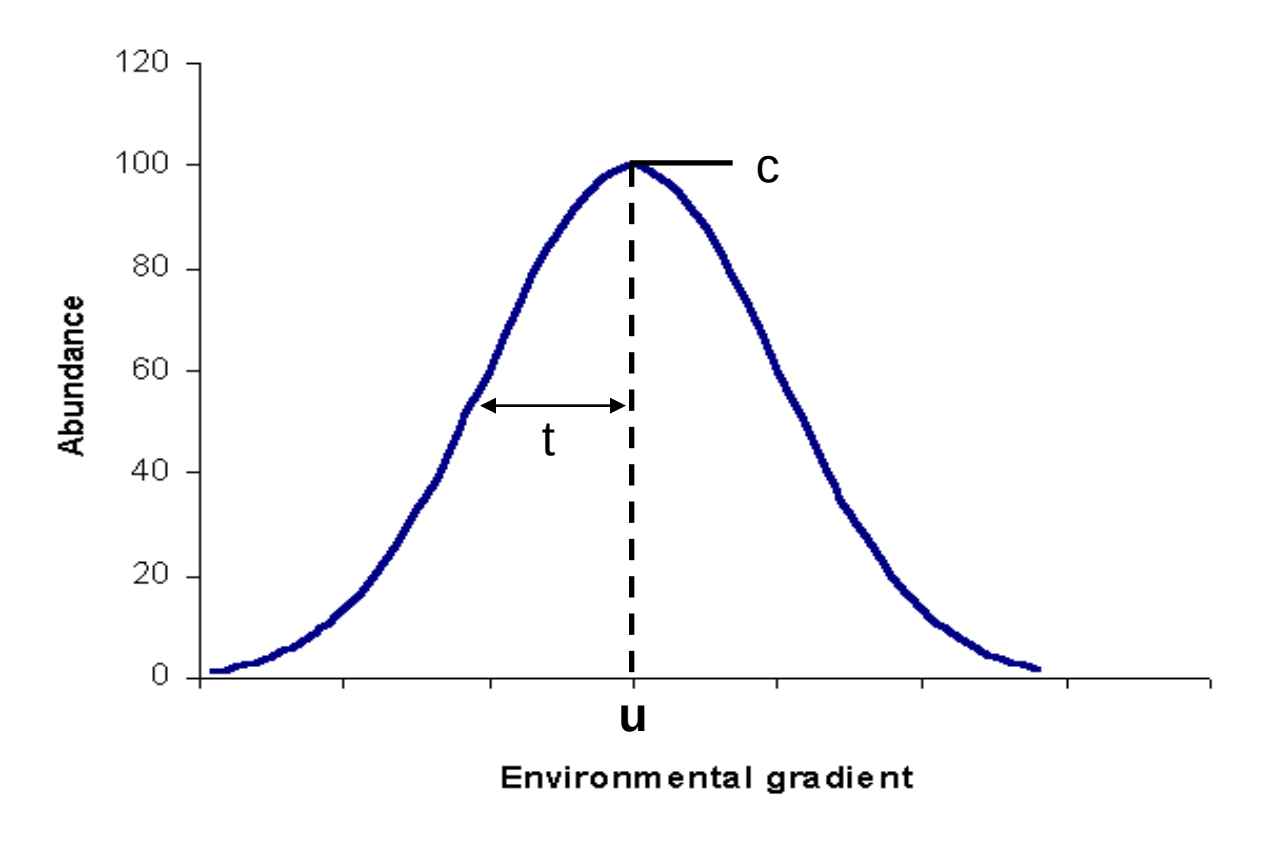


CA –DCA



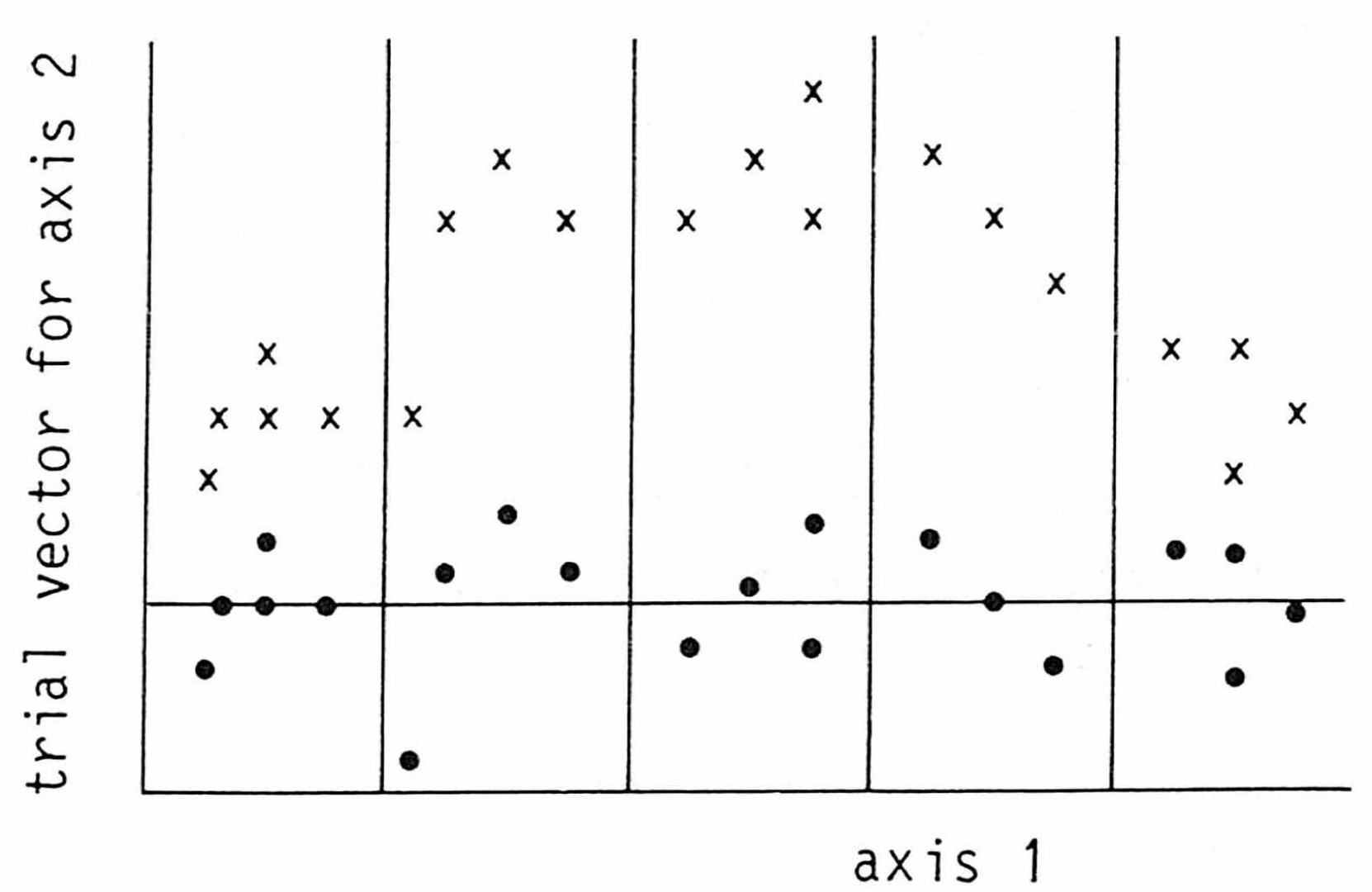
Jongman et al. (1995): Data analysis in community and landscape ecology

Gaussian response curve



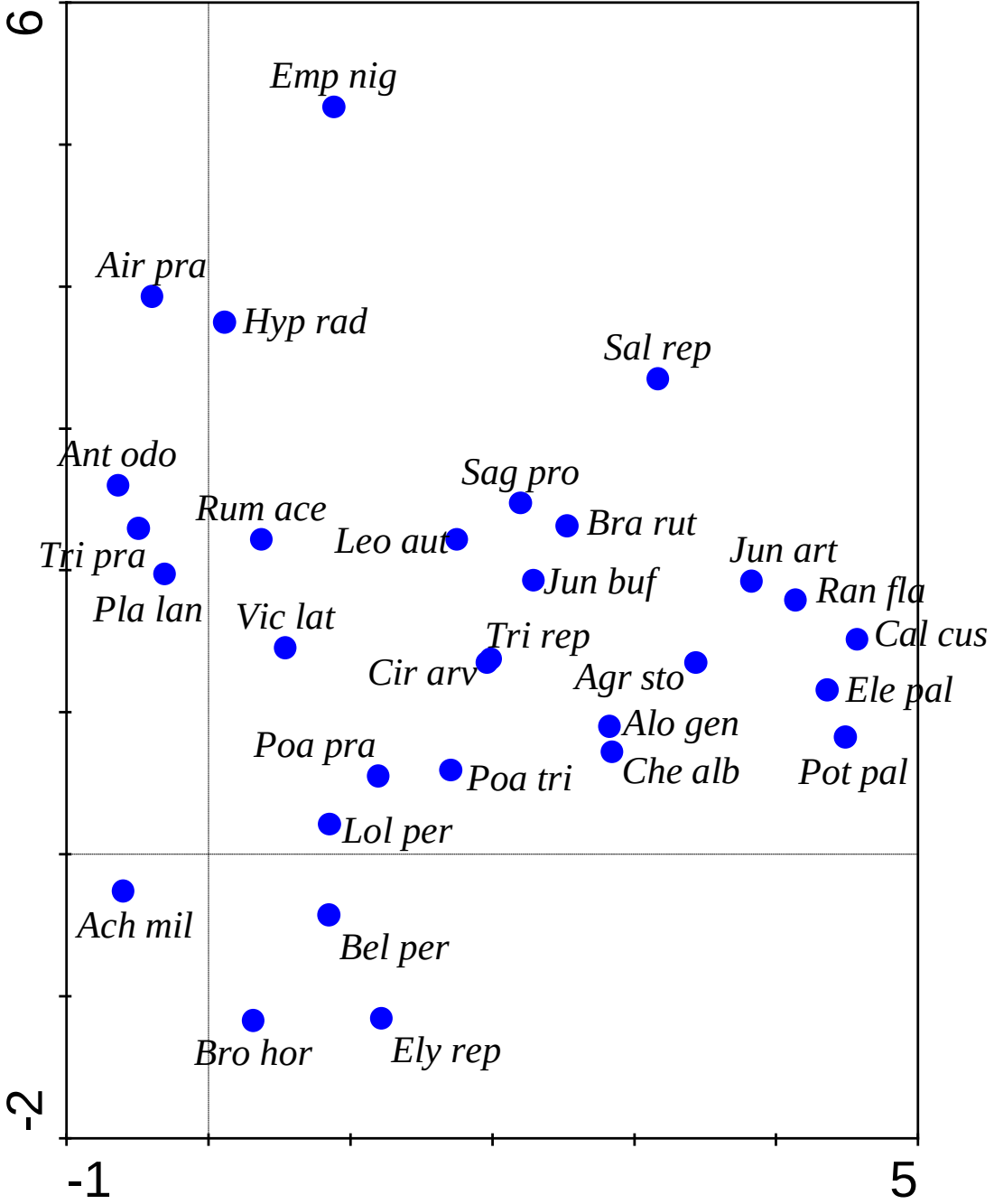
$$z = c \exp[-0.5(x-u)^2/t^2]$$

Detrending by segments



Jongman et al. (1995): Data analysis in community and landscape ecology

(CA)

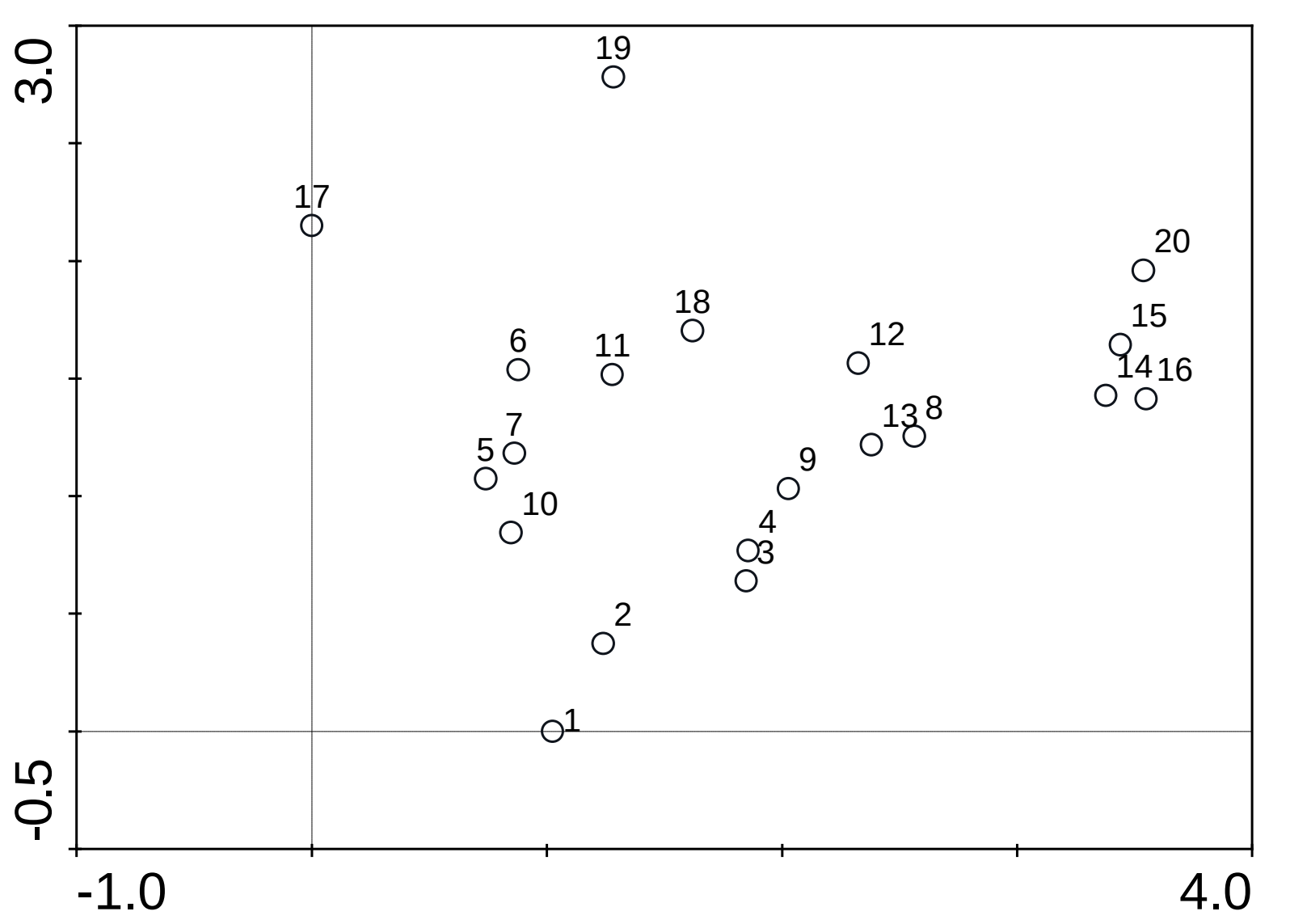


Dune meadow Data
DCA
scatter plot
species scores

Lüneburg, Oct. 2011

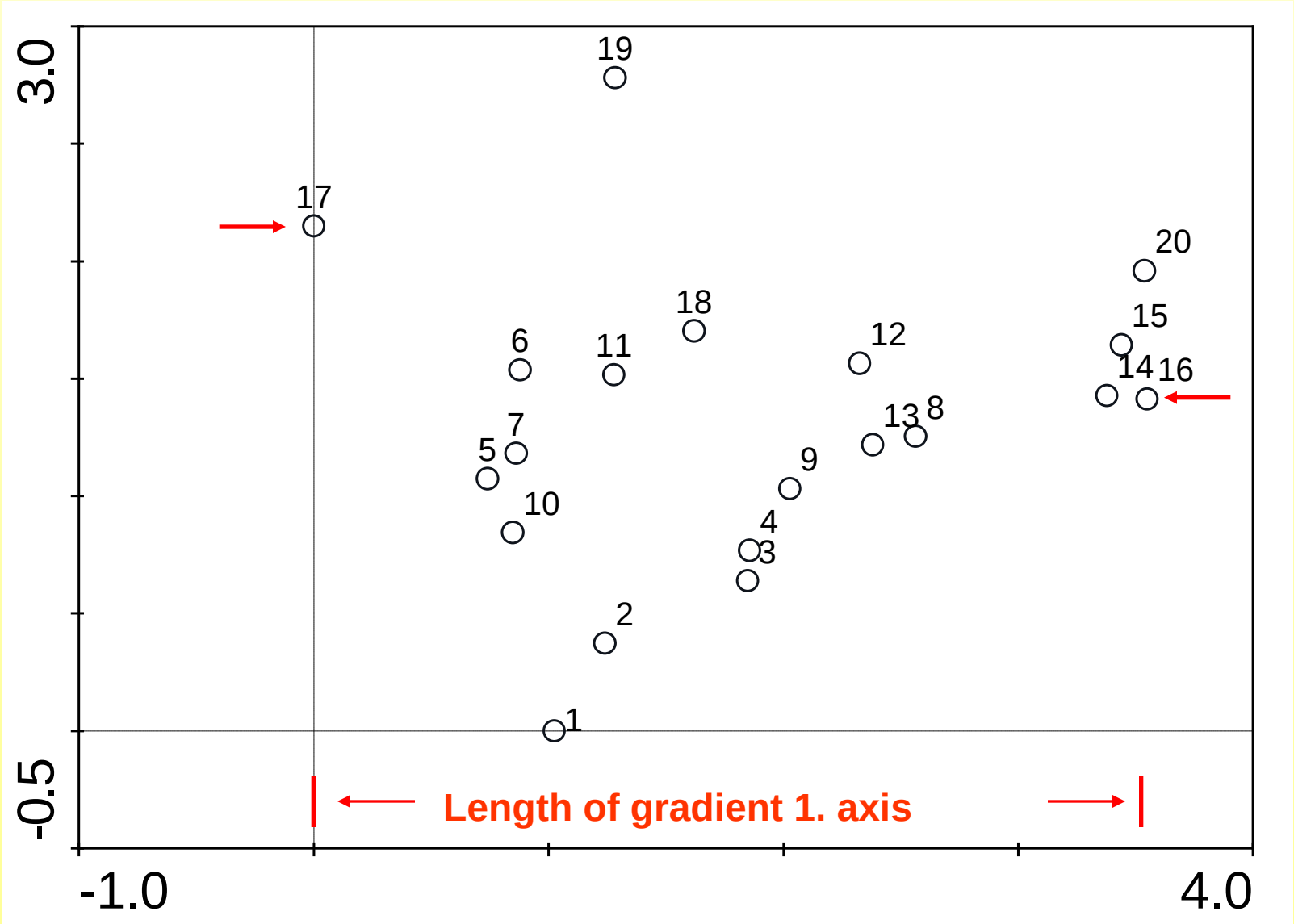
correspondence analysis
(CA)

Dune Meadow Data
DCA scatter plot sample scores



correspondence analysis (CA)

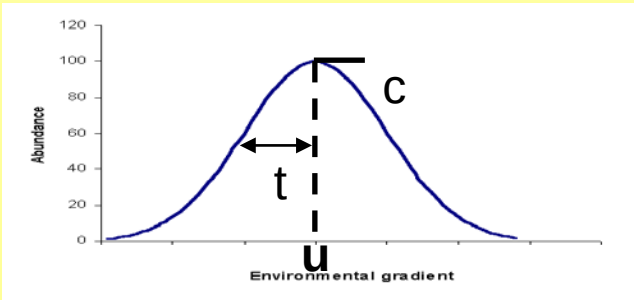
Dune meadow Data
DCA scatter plot *sample scores*



Dune Meadow Data
ordination diagnostics

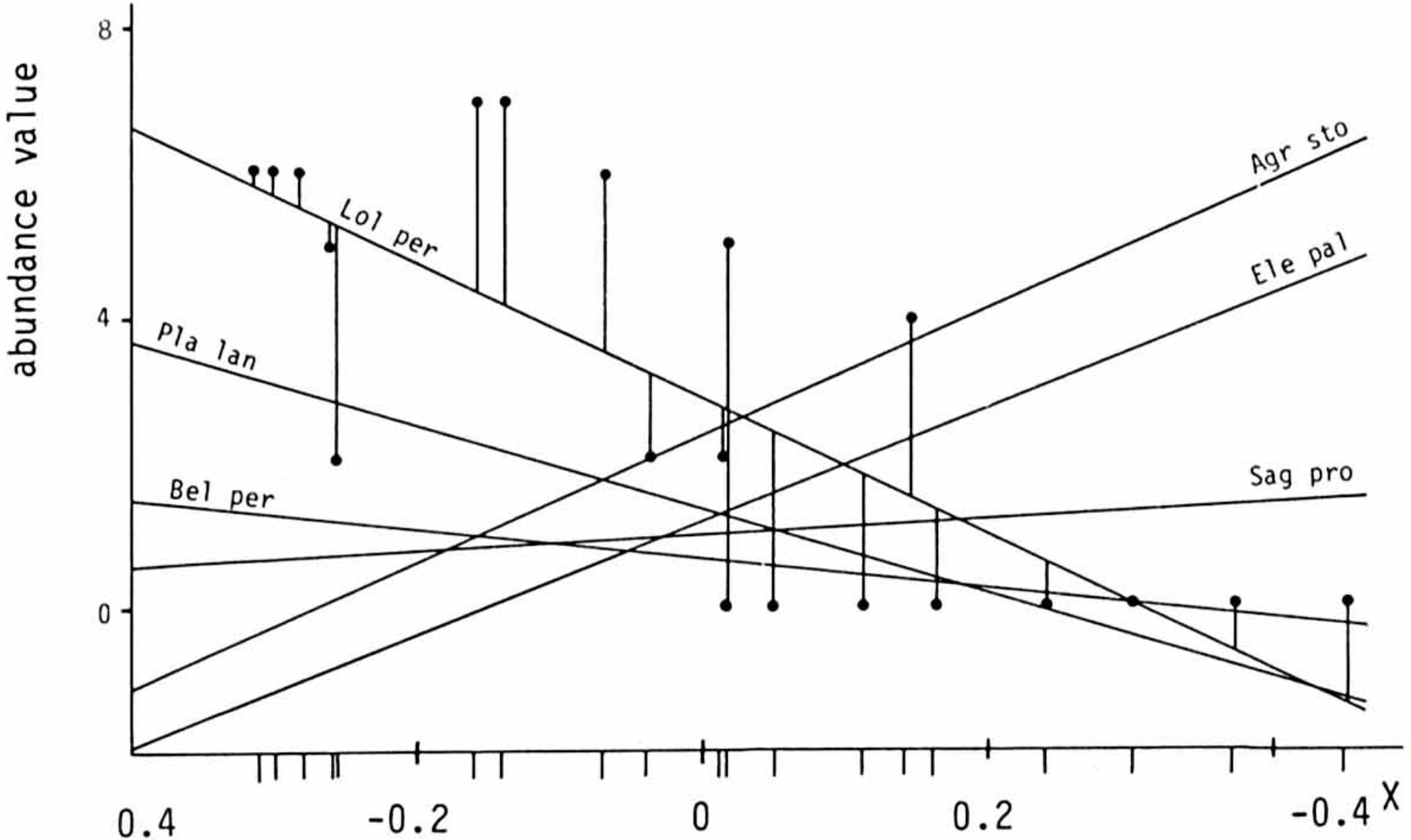
Axis	1	2	3	4
Eigenvalues CA:	0.530	0.360	0.238	0.168
Eigenvalues DCA:	0.530	0.224	0.050	0.031
Length of gradient:	3.548	2.783	1.488	1.246

(the length of gradient indicates multivariate standard deviations: 2 units correspond to 50% species turnover)



Principal Component Analysis (PCA)

regression lines of dune meadow species for main gradient (principal component I)

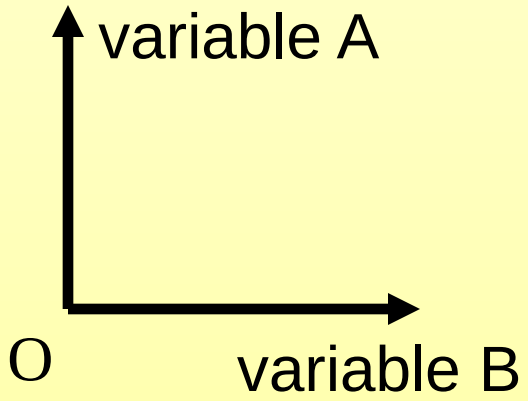


Jongman et al. (1995): Data analysis in community and landscape ecology

Principal Component Analysis (PCA) - Principles

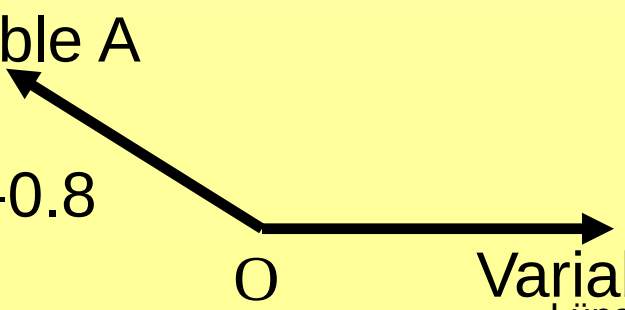
Vector graphic for the correlation between two variables

a) Uncorrelated R = 0 (cos = 0)



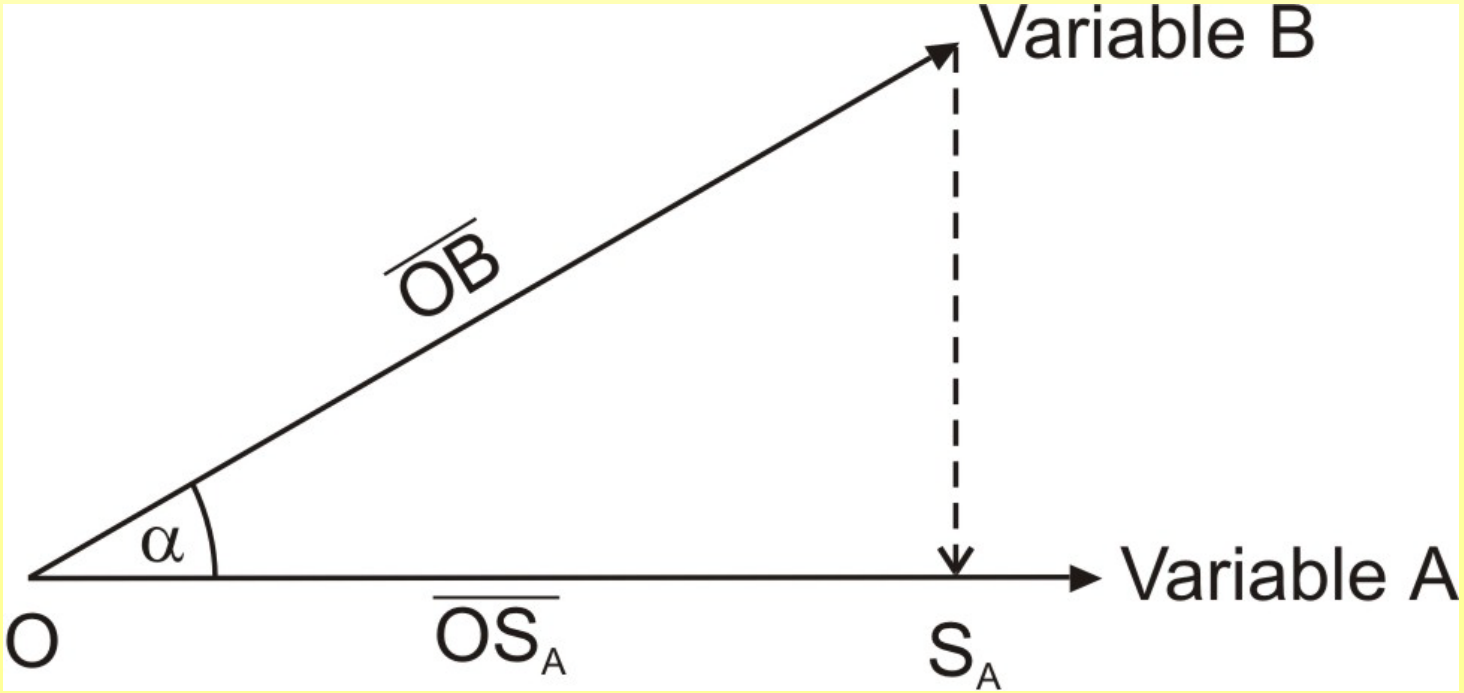
b) 100% correlated r = 1 (cos = 1)

c) Negatively correlated r = cos = -0.8 (α= 143)



Principles

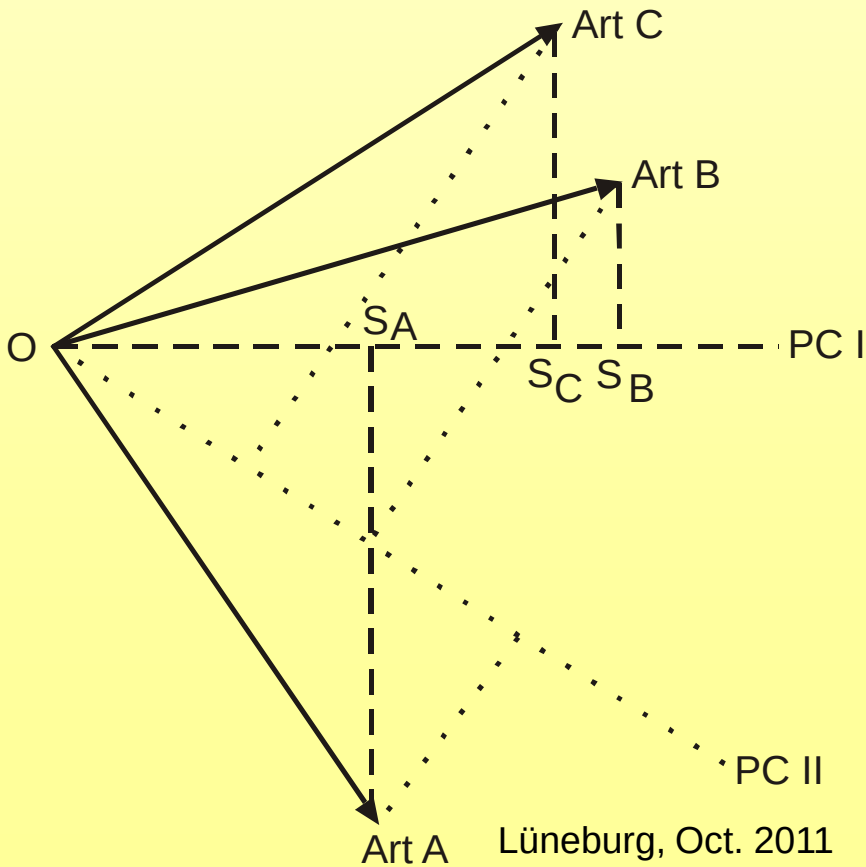
Vector graphic for the correlation between two variables (centered values, unit vectors, Kent & Coker, 1992). Cosinus α varies between –1 and +1 like a correlation coefficient. If length of vector is equal (standardised), line segment $\overline{OS}_A$ becomes a measure for the correlation ($\overline{OS}_A$= cos α as vectors have unit length)



Principles

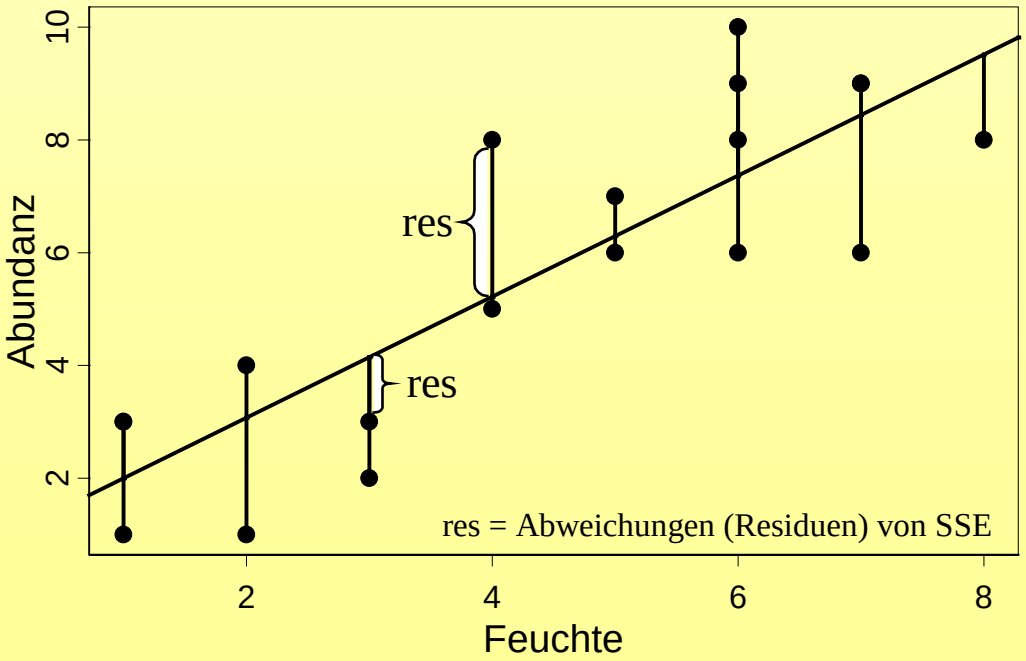
The PCA finds the new (synthetic) axis that correlates most with as many variables as possible. This so called **principle component** is akin to a synthetic species that captures information of many real species/variables.

The principal components are selected to maximise the sum of correlations of the variables with the new axis (the sum is proportional to the sum of all line segments between 0 and the intercepts of the perpendiculars - it‘s called **Eigenvalue**).



General: regression coefficients

The PCA is based on linear regression (Pearson regression). The calculation of the linear models follows the standard method of minimising the residual sum of squares:



$$a = \bar{y} - b\bar{x}$$

$$b = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) / \sum_{i=1}^{n}(x_i - \bar{x})^2$$

General: regression coefficients

For PCA, the regression equation is simplified:

- The response variable (e.g species) is centered: $y_{ik} = y_i - \bar{y}$
- The predictor variable (e.g axes value) is centered and standardised (mean 0, sum of squares 1): $\sum_{i=1}^{n} (x_i - \bar{x})^2 = 1$

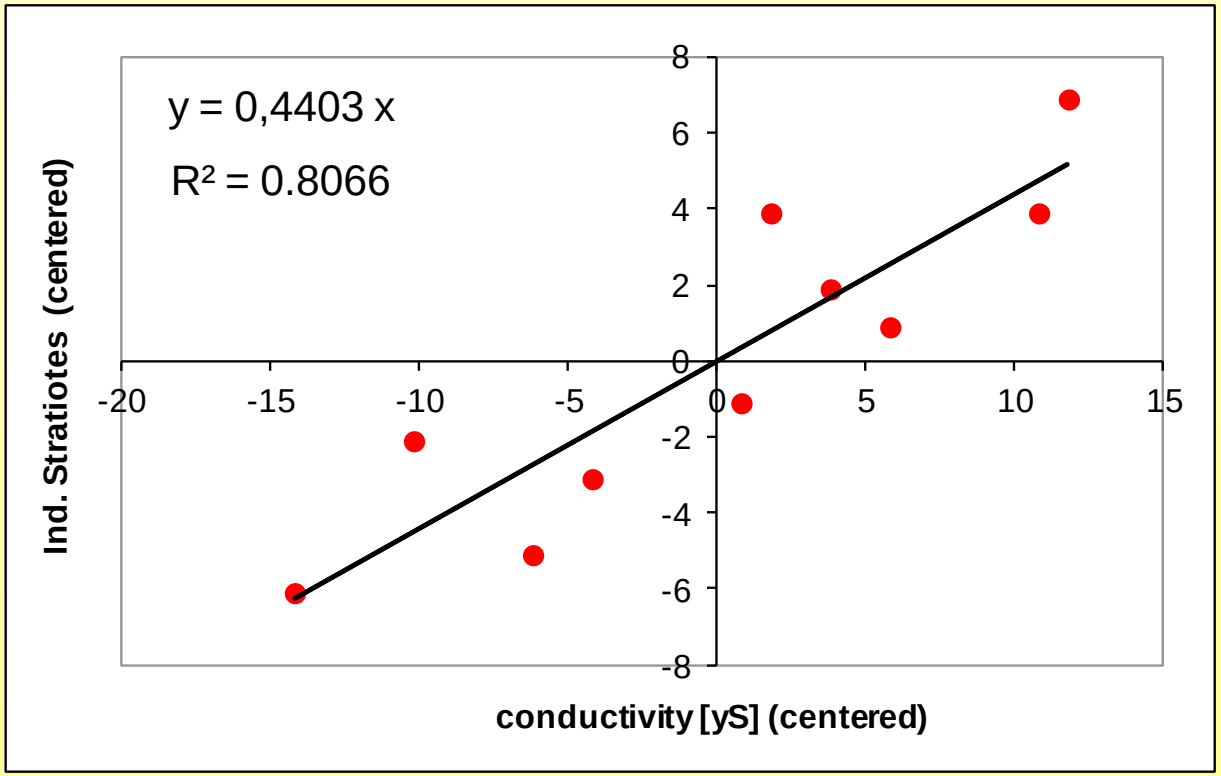
This results in a strong simplification of the regression coefficients:

$a = \bar{y}_{ik} - b\bar{x}_i$ → $a = 0$, $b\bar{x}_i = 0$

Linear model – regression analysis with centered data

Conduct.	*Stratiotes aloe.* (Ind)
-14,2	-6,1
-10,2	-2,1
-6,2	-5,1
-4,2	-3,1
0,8	-1,1
1,8	3,9
3,8	1,9
5,8	0,9
10,8	3,9
11,8	6,9

centered



General: regression coefficients

For PCA, the regression equation is simplified:

- The response variable (e.g species) is centered: $y_{ik} = y_i - \bar{y}$
- The predictor variable (e.g axes value) is centered and standardised (mean 0, sum of squares 1): $\sum_{i=1}^{n} (x_i - \bar{x})^2 = 1$

This results in a strong simplification of the regression coefficients:

$a = \bar{y}_{ik} - b\bar{x}_i$ (0, 0)

$b = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) / \sum_{i=1}^{n} (x_i - \bar{x})^2$ (y_{ik}, 0, 1) $\rightarrow$ $b_k = \sum_{i=1}^{n} y_{ik} x_i$

$r = \sum (x_i - \bar{x})(y_i - \bar{y}) / \sqrt{\sum (x_i - \bar{x})^2 \cdot (y_i - \bar{y})^2}$ (0, y_{ik}, 1, y_{ik}) $\rightarrow$ $r_k = \sum_{i=1}^{n} y_{ik} x_i / \sqrt{\sum y_{ik}}$

Weighted Summation algorithm

1. step: choose standardised sample scores (predictor variable):

2. step: regression of response variables (species)

Plot	Site score	=(xi-MW(x))²	*Achillea*	yik = yi-MW(y)	bk=yik*site score
MW	**0.00**		**0.800**		
Sum		**1.00**		**0.000**	**-1.980**
1	-0.37	0.14	1.000	0.200	-0.074
2	-0.33	0.11	3.000	2.200	-0.726
3	-0.29	0.08		-0.800	0.232
4	-0.25	0.06		-0.800	0.200
5	-0.21	0.04	2.000	1.200	-0.252
6	-0.17	0.03	2.000	1.200	-0.204
7	-0.14	0.02	2.000	1.200	-0.168
8	-0.10	0.01		-0.800	0.080
9	-0.06	0.00		-0.800	0.048
10	-0.02	0.00	4.000	3.200	-0.064
11	0.02	0.00		-0.800	-0.016
12	0.06	0.00		-0.800	-0.048
13	0.10	0.01		-0.800	-0.080
14	0.14	0.02		-0.800	-0.112
15	0.17	0.03		-0.800	-0.136
16	0.21	0.04		-0.800	-0.168
17	0.25	0.06	2.000	1.200	0.300
18	0.29	0.08		-0.800	-0.232
19	0.33	0.11		-0.800	-0.264
20	0.37	0.14		-0.800	-0.296

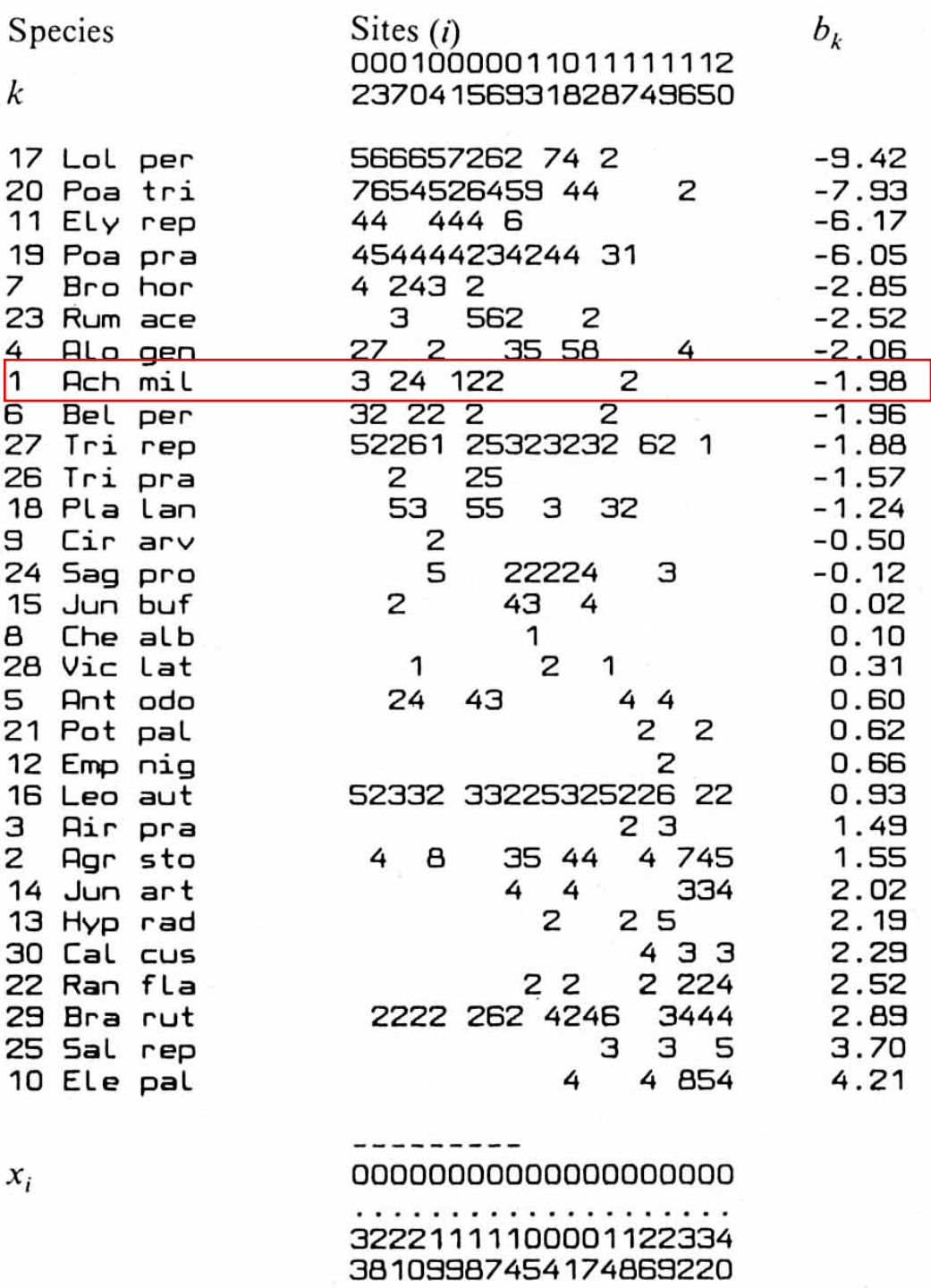
principal component analysis

Weighted Summation algorithm

3. step – calibration: calculate new values for the predictor variable (sample scores) based on regression-coefficients for response variables.

$$x_i = \sum_{k=1}^{m} y_{ki} b_k$$

4. step: standardise predictor variables (sample scores, as in CA)



Species k	Sites (i)	b_k
	00010000011011111112	
	23704156931828749650	
17 Lol per	566657262 74 2	-9.42
20 Poa tri	7654526459 44 2	-7.93
11 Ely rep	44 444 6	-6.17
19 Poa pra	454444234244 31	-6.05
7 Bro hor	4 243 2	-2.85
23 Rum ace	3 562 2	-2.52
4 Alo gen	27 2 35 58 4	-2.06
1 Ach mil	3 24 122 2	-1.98
6 Bel per	32 22 2 2	-1.96
27 Tri rep	52261 25323232 62 1	-1.88
26 Tri pra	2 25	-1.57
18 Pla lan	53 55 3 32	-1.24
9 Cir arv	2	-0.50
24 Sag pro	5 22224 3	-0.12
15 Jun buf	2 43 4	0.02
8 Che alb	1	0.10
28 Vic lat	1 2 1	0.31
5 Ant odo	24 43 4 4	0.60
21 Pot pal	2 2	0.62
12 Emp nig	2	0.66
16 Leo aut	52332 33225325226 22	0.93
3 Air pra	2 3	1.49
2 Agr sto	4 8 35 44 4 745	1.55
14 Jun art	4 4 334	2.02
13 Hyp rad	2 2 5	2.19
30 Cal cus	4 3 3	2.29
22 Ran fla	2 2 2 224	2.52
29 Bra rut	2222 262 4246 3444	2.89
25 Sal rep	3 3 5	3.70
10 Ele pal	4 4 854	4.21
	---------	
x_i	00000000000000000000	
	....	
	32221111100001122334	
	38109987454174869220	

principal component analysis

Weighted Summation **algorithm**

4. and following steps: Repeated calculation of new values for response variable (species scores) and predictor variable (sample scores); values will stabilise after a number of iterations

6. step: Calculate higher axes in a similar way, but always make axes orthogonal

Species		Sites (i)	b_k
		01000100100111011121	
k		60725113894793824506	
17	Lol per	66652776225 4	-9.21
18	Pla lan	535 53 3 2	-5.77
19	Poa pra	344424453441 24	-5.69
20	Poa tri	44576 26 55 944 2	-4.80
1	Ach mil	24232 1 2	-3.81
23	Rum ace	6 3 5 2 2	-3.68
27	Tri rep	562523 2231 222361	-3.67
5	Ant odo	342 4 44	-3.52
7	Bro hor	4242 3	-3.31
16	Leo aut	333535 252226232222	-2.86
11	Ely rep	44 44 64	-2.86
26	Tri pra	5 2 2	-2.63
6	Bel per	2 32 22 2	-2.11
28	Vic lat	1 2 1	-0.67
13	Hyp rad	2 25	-0.08
9	Cir arv	2	0.01
12	Emp nig	2	0.09
8	Che alb	1	0.11
3	Air pra	23	0.15
15	Jun buf	2 4 3 4	0.40
29	Bra rut	622 24 2622 3 24 444	0.94
24	Sag pro	2 25 3224	0.98
21	Pot pal	22	1.07
25	Sal rep	3 3 5	1.86
4	Alo gen	2 7 32 558 4	3.33
30	Cal cus	4 33	3.40
22	Ran fla	22 2242	3.95
14	Jun art	4 4 343	4.29
10	Ele pal	4 4548	8.08
2	Agr sto	4 38 5444457	8.67
		---------	
x_i		00000000000000000000	
		.....	
		33222110000001112334	
		10866647401151464075	

Weighted Summation **algorithm**

7. step – calculate sample coordinates: for each axis the position of each sample is calculated:

For each species, multiply its abundance (in the given sample) with the regression coefficients for the given axis.
The sum of these products gives the position of the sample on these axis:

e.g Plot 6, previous table:
site score: 6*(-9.21) + 5*(-5.77) + 3*(-5.69) + ... + 0*8.08 + 0*8.67

ordination diagnostics

***Loadings of* variables**

Are basically the regression coefficients (the correlation) of the variable with a given principal component (or: the slope of the corresponding regression line after standardising).

***Eigenvalue* of the principal component**

Indicates how well all variables in the data set correlate with the main axis; is largest for first principal component and then declines for higher order components.

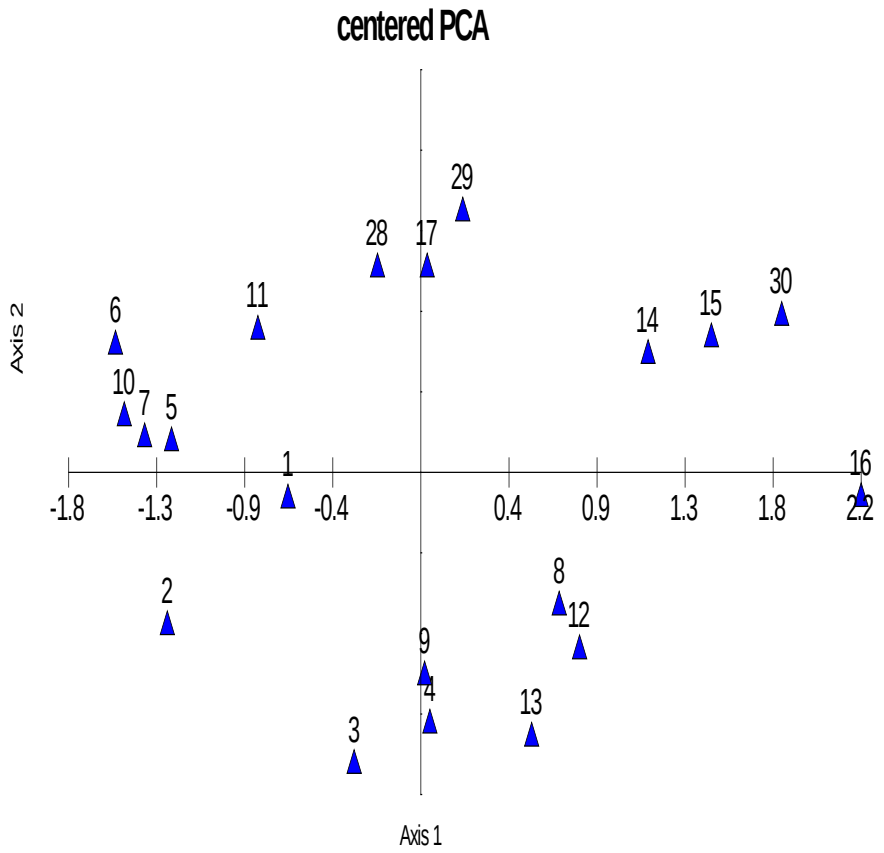
Explained variance of principal component X_i

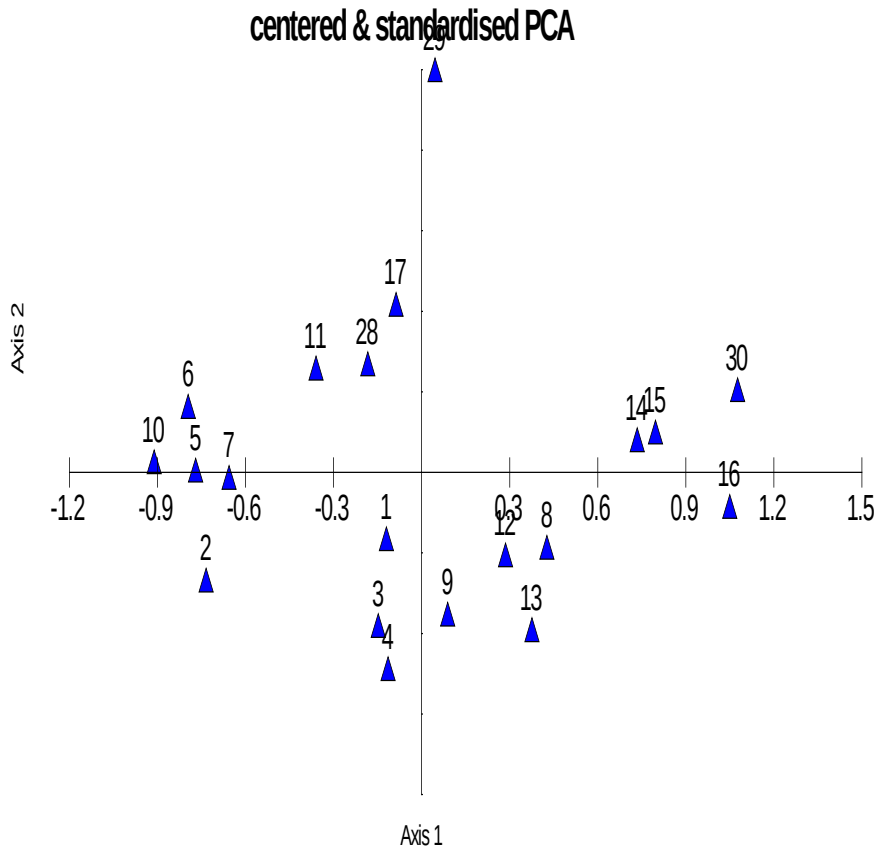
Eigenvalue axis X_i / sum of all Eigenvalues.

-> reported often in a cumulative fashion for the (first 4) axes.

Centered vs. correlations matrix PCA

PCA *Dune Meadow Data*: centered PCA (explained) and PCA of centered and standardised species data (*correlation matrix* PCA; needed if variables have different scales).





Dune Meadow Data - ordination diagnostics

Centred PCA

	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5
Eigenvalues	24.793	18.143	7.604	7.154	5.695
Percentage	29.472	21.567	9.040	8.504	6.770
Cum. Percentage	29.472	51.039	60.078	68.582	75.352

Correlation matrix PCA (centered & standardised)

	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5
Eigenvalues	7.032	4.997	3.555	2.644	2.138
Percentage	23.441	16.657	11.849	8.812	7.128
Cum. Percentage	23.441	40.098	51.947	60.759	67.887

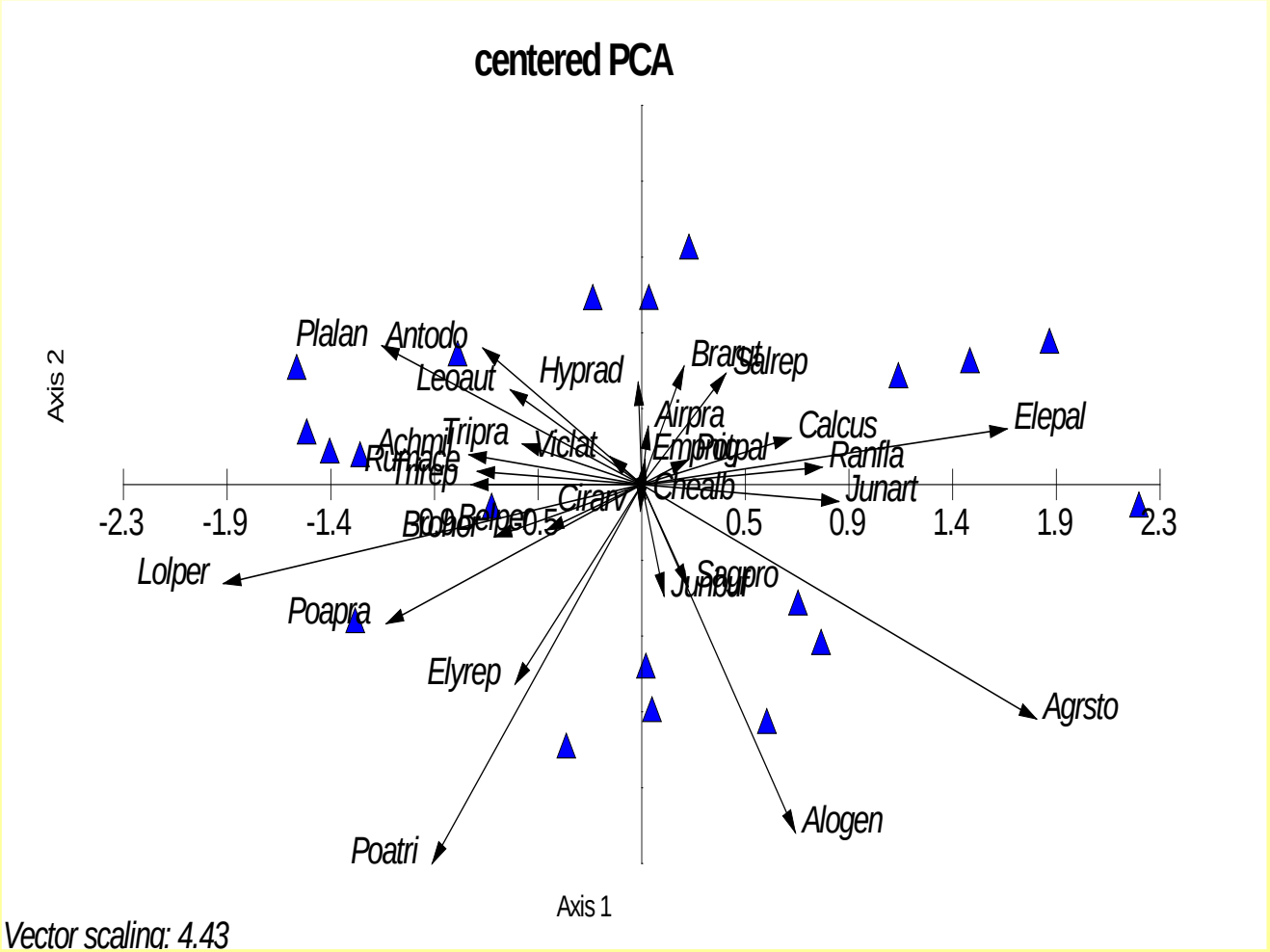
Dune Meadow Data – loadings principal component

***Loadings* of variables:** Loadings of the first 13 species on the first 6 principal components

	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6
Achmil	**-0.176**	0.041	0.009	-0.096	0.235	0.093
Agrsto	**0.402**	**-0.327**	0.111	-0.117	-0.061	0.036
Airpra	0.007	0.081	-0.128	0.147	0.028	-0.044
Alogen	**0.157**	**-0.485**	**-0.253**	-0.019	-0.286	0.202
Antodo	**-0.162**	0.190	**-0.243**	-0.073	0.200	-0.059
Belper	-0.097	-0.063	0.060	0.061	0.065	0.110
Brohor	**-0.150**	-0.073	0.060	0.003	0.232	0.201
Chealb	0.004	-0.020	-0.024	0.005	0.006	0.005
Cirarv	-0.001	-0.036	0.026	0.044	-0.015	-0.032
Elepal	**0.372**	0.077	**0.306**	-0.326	0.116	0.155
Elyrep	**-0.129**	**-0.278**	0.153	0.145	0.190	-0.489
Empnig	0.004	0.034	-0.077	0.068	-0.012	0.004
Hyprad	-0.004	0.142	-0.174	0.261	-0.046	0.000

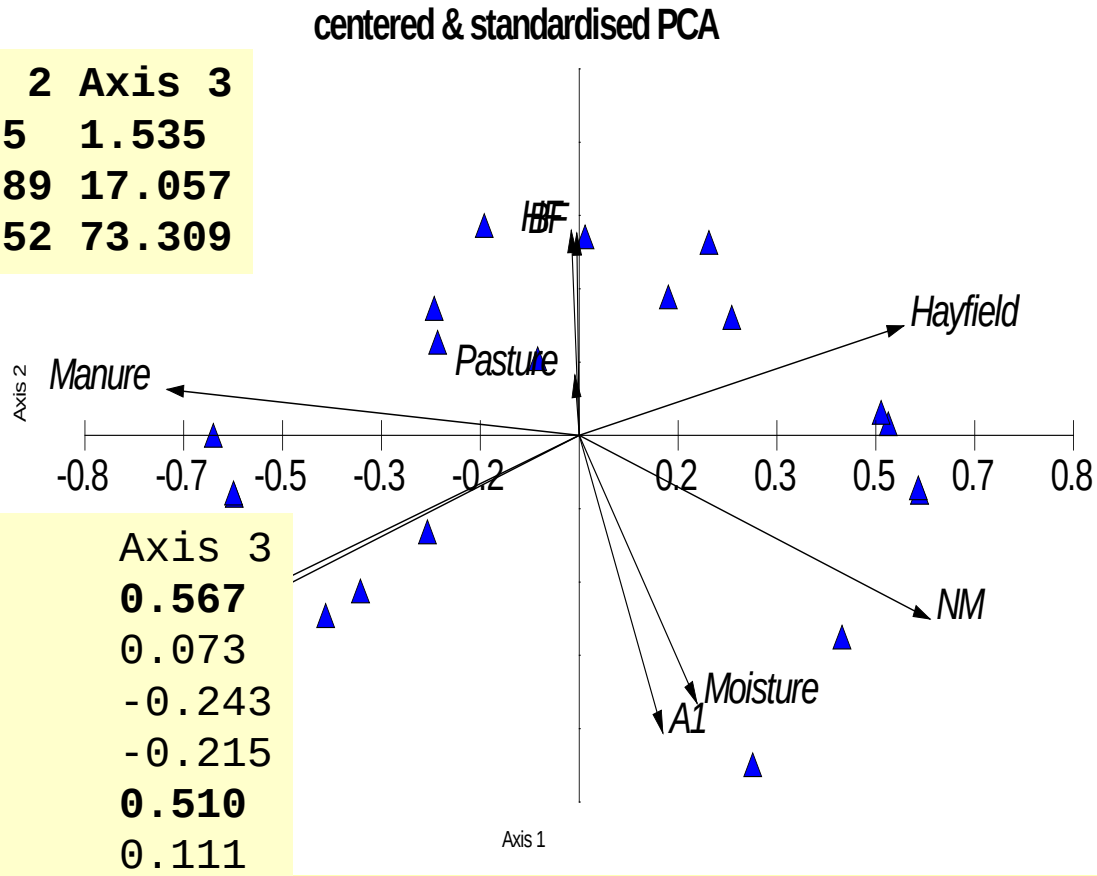
Dune Meadow Data – loadings principal component

Euclidean Biplot of species and samples



Dune Meadow Data – correlation-PCA of environmental variables

	Axis 1	Axis 2	Axis 3
Eigenvalues	**3.228**	**1.835**	**1.535**
Percentage	**35.863**	**20.389**	**17.057**
Cum. Percentage	**35.863**	**56.252**	**73.309**



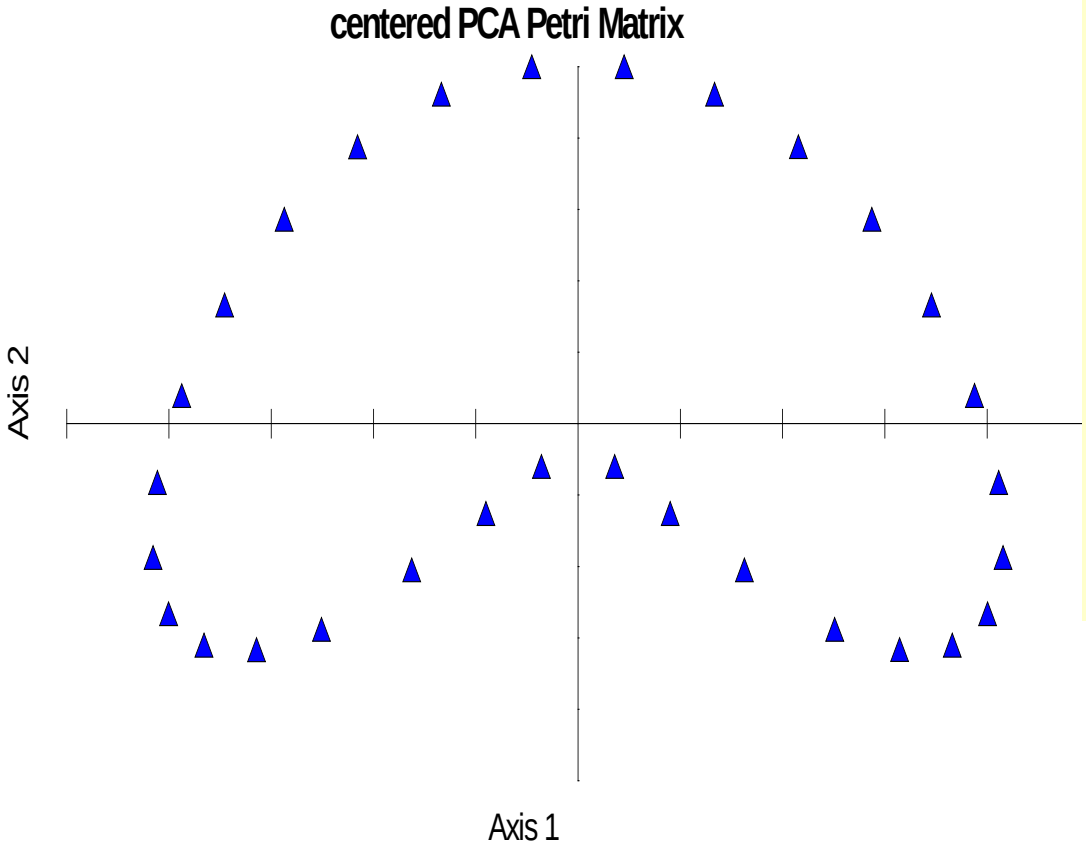
	Axis 1	Axis 2	Axis 3
Moisture	0.119	**-0.326**	**0.567**
Manure	**-0.512**	0.132	0.073
Hayfield	**0.426**	-0.038	-0.243
Haypasture	**-0.409**	**-0.330**	-0.215
Pasture	-0.006	**0.414**	**0.510**
SF	**-0.442**	**-0.325**	0.111
BF	0.018	0.270	**-0.501**
HF	0.005	**0.522**	0.182
NM	**0.424**	**-0.379**	0.107

Two-way Petrie-matrix

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Art 1	1									
Art2	1	1								
Art 3	1	1	1							
Art 4		1	1	1						
Art 5			1	1	1					
Art 6				1	1	1				
Art 7					1	1	1			
Art 8						1	1	1		
Art 9							1	1	1	
Art 10								1	1	1
Art 11									1	1
Art 12										1

Jongman et al. (1995): Data analysis in community and landscape ecology

Horseshoe-Effect PCA



Euclidean distance regards shared absences as similarity. Thus, samples with few shared species are disproportionally similar. They are placed adjacent in the scatter plot; the result is a horse-shoe shaped pattern.

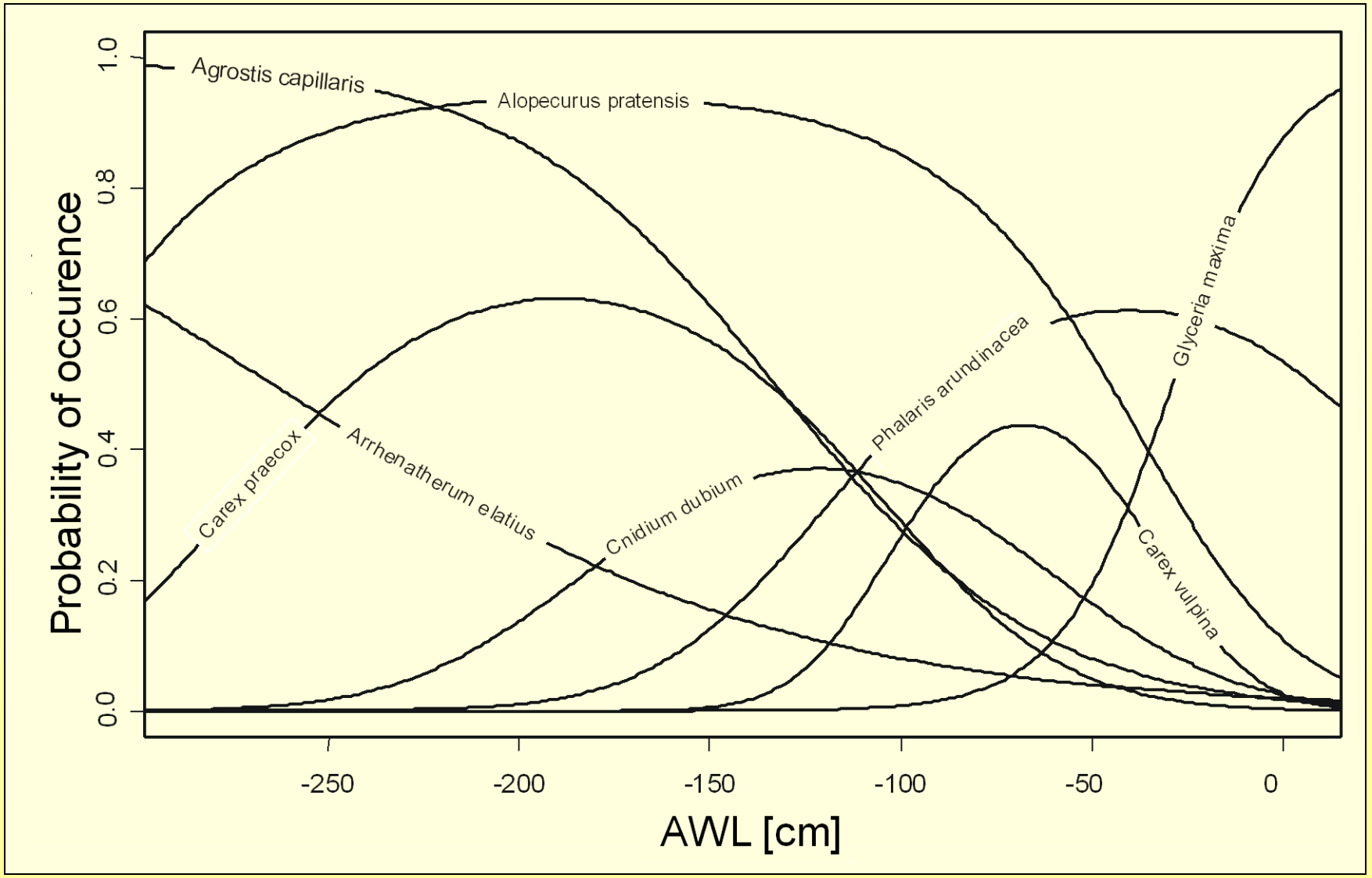
Properties PCA

- Centred PCA/correlation matrix PCA: for data with homogenous scales (e.g. cover values) centered PCA is more suitable, because values respond to transformations; correlation-PCA is e.g. often used for environmental variables that were measured on different scales (Legendre & Legendre 1998, Franklin et al. 1995).
- PCA is either only for for samples ("R-Mode") or only for species ("Q-mode"). Biplots are generated by plotting "site scores" and "loadings" in one diagram with a suitable scale (Details s. Jongman et al. 1995, S.129).
- Data distribution should be (largely) normal, otherwise the first axis will only reflect the gradient to the extreme samples (=> data transformation).
- There should be (many) more species/variables than samples. Potential artefacts are, however, usually small for first axes (solution => exclude species).
- The implicite distance measure in PCA is Euclidean distance. Thus PCA is also sensitive to „the double zero -problem" (s. introduction - similarities). Very unsimilar samples (many shared zero^s) are relatively similar (horseshoe-effect).

- PCA is only suitable for relatively homogeneous data (=> initial DCA, multivariate standard deviations <3); relatively similar ecological samples; real measurements

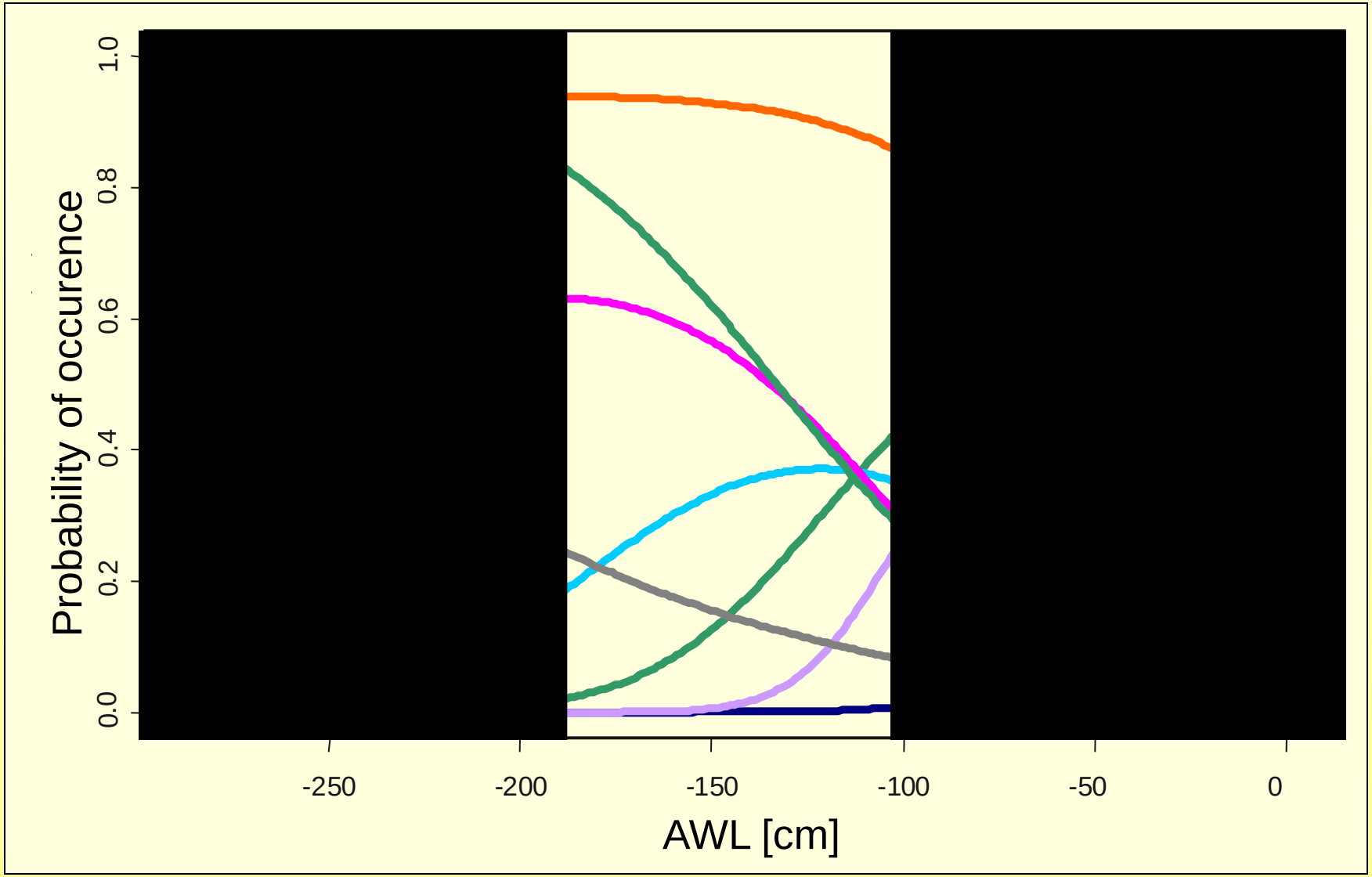
principal component analysis

Species-response-curves

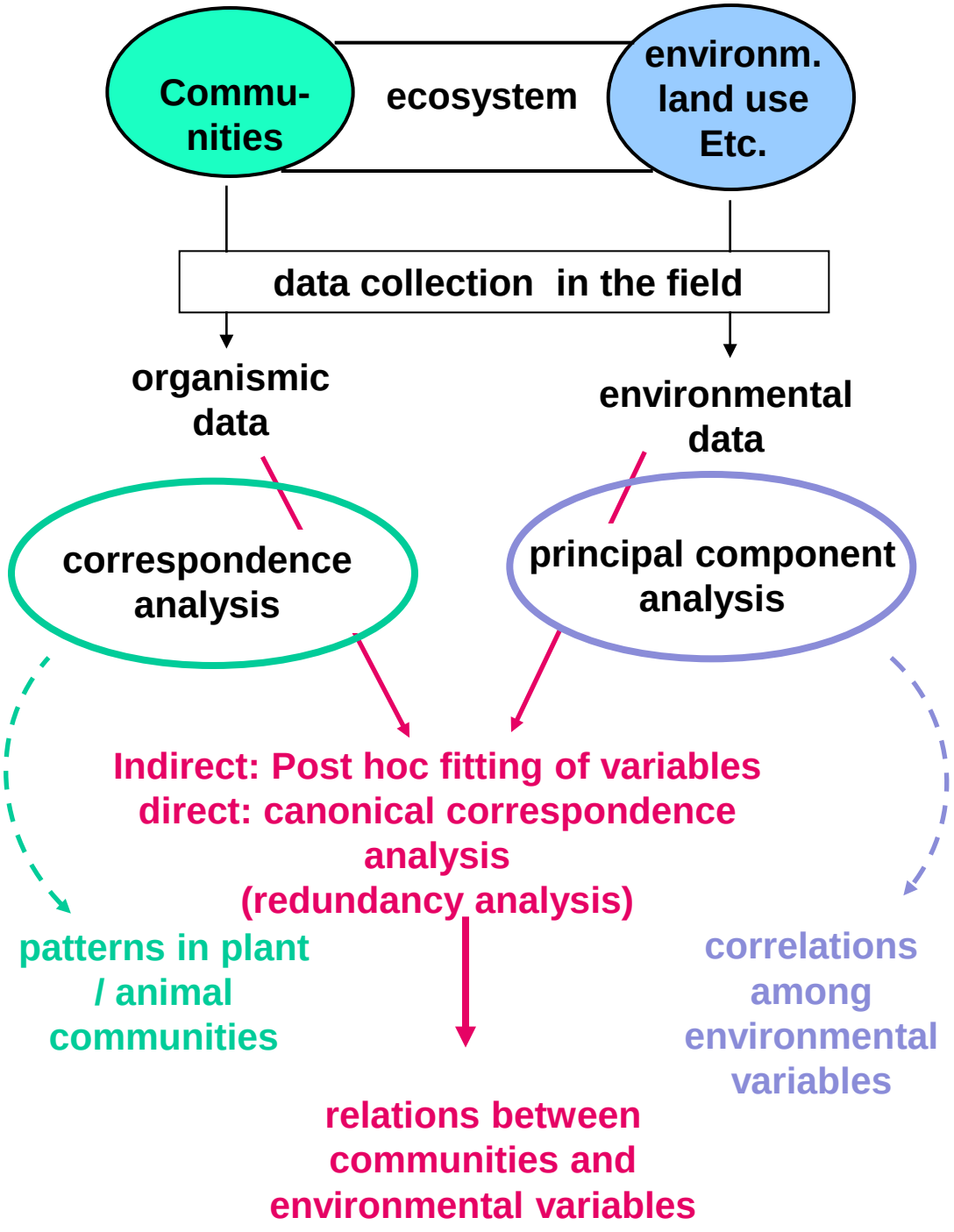


principal component analysis

length of gradient and species-response curves



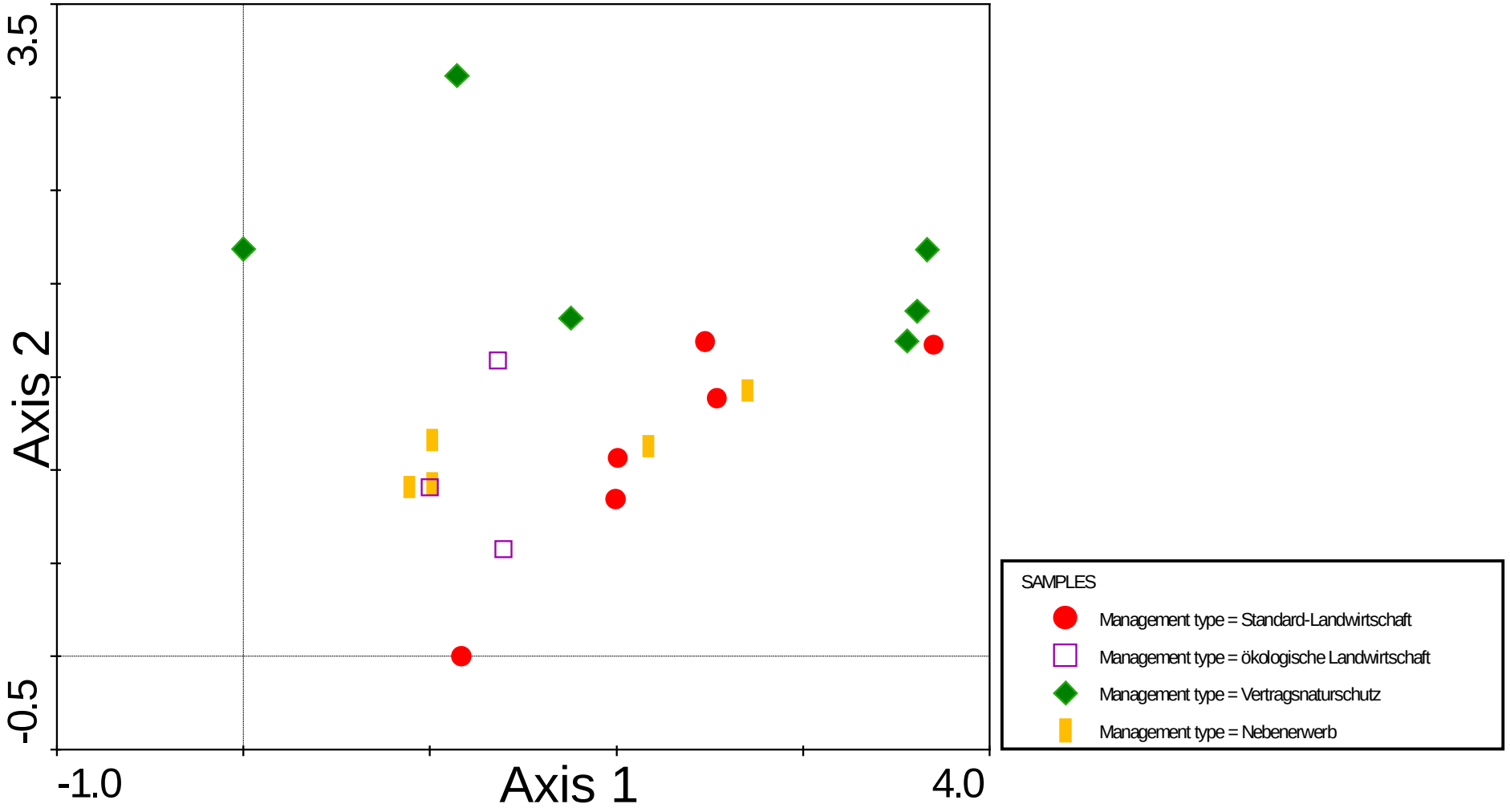
Correspondence analysis and environmental data



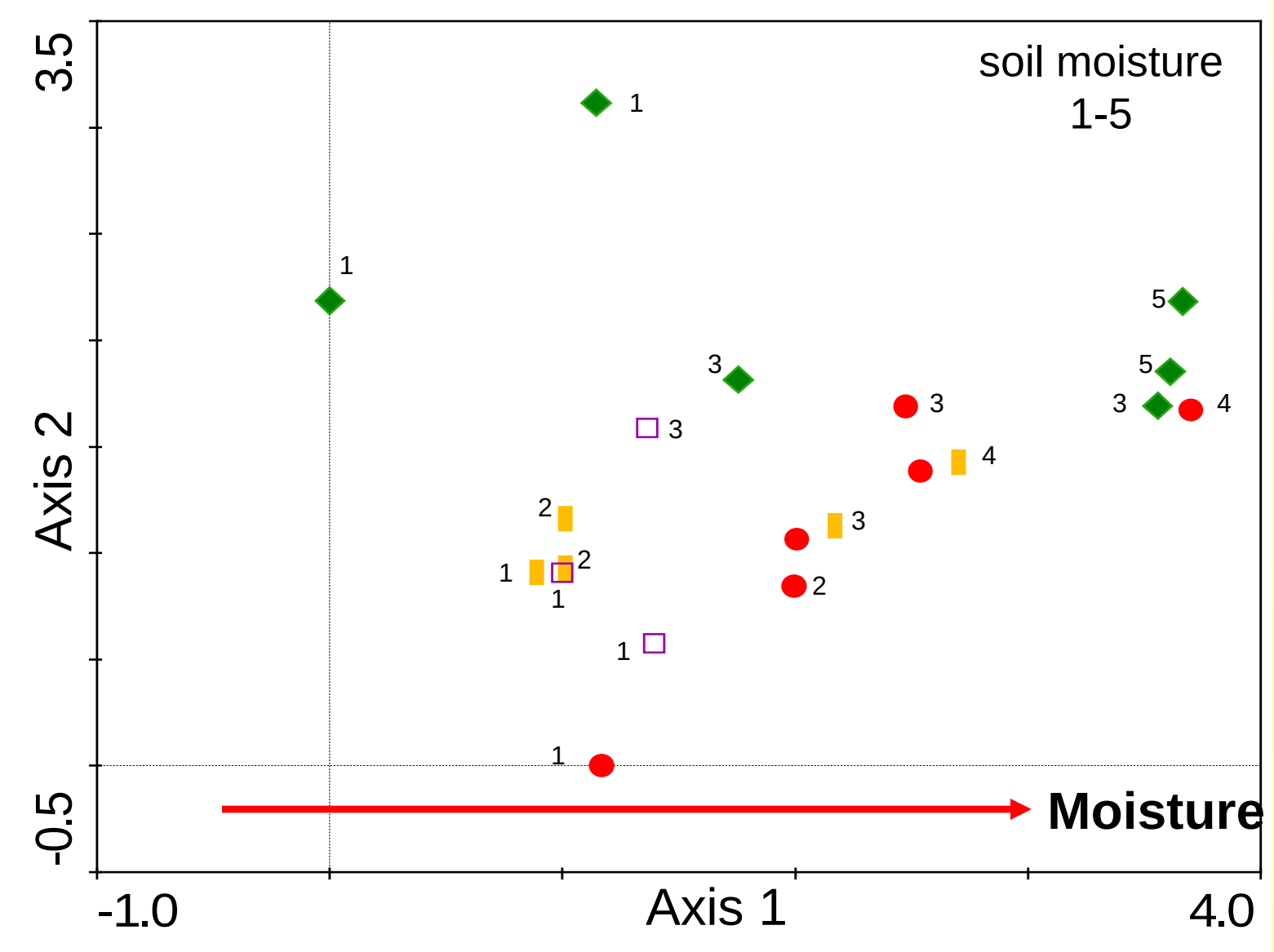
Analysis of community data

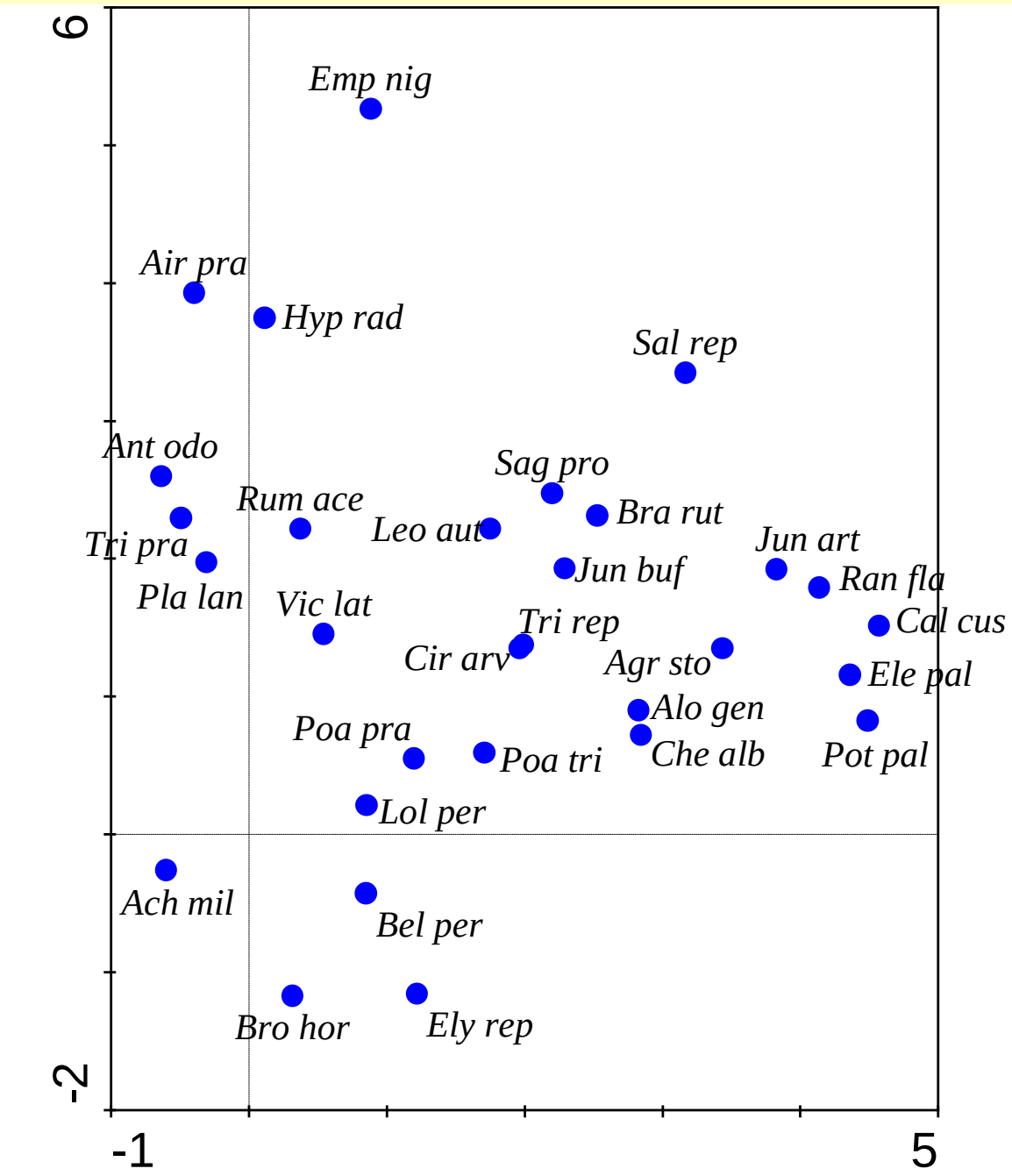
secondary data

Indirect gradient analysis and environmental data - assign samples to classes



indirect gradient analysis and environmental data - assign ratio-scaled data

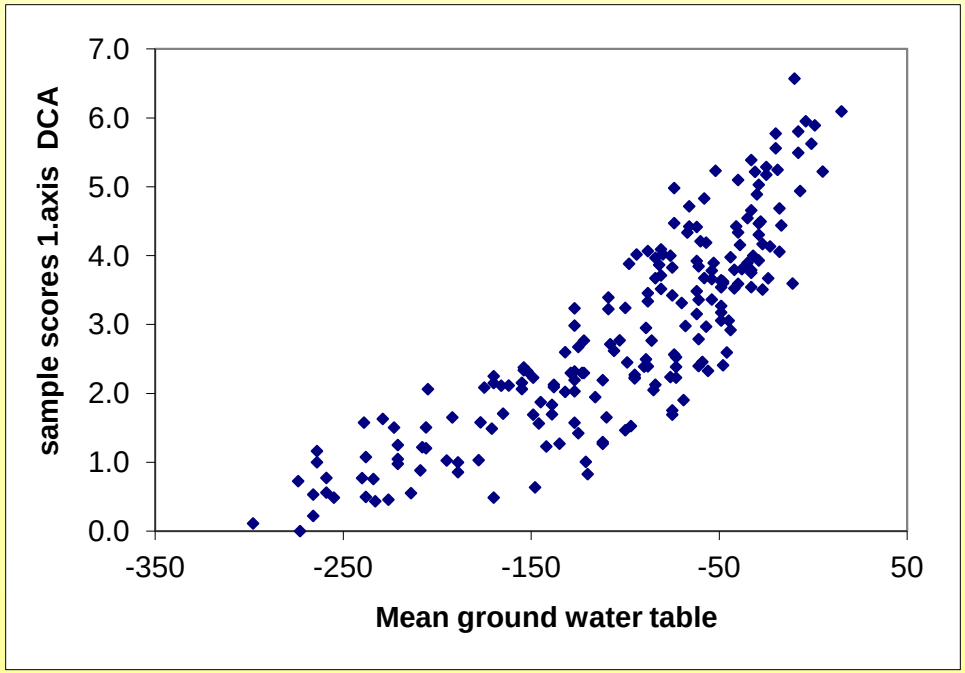


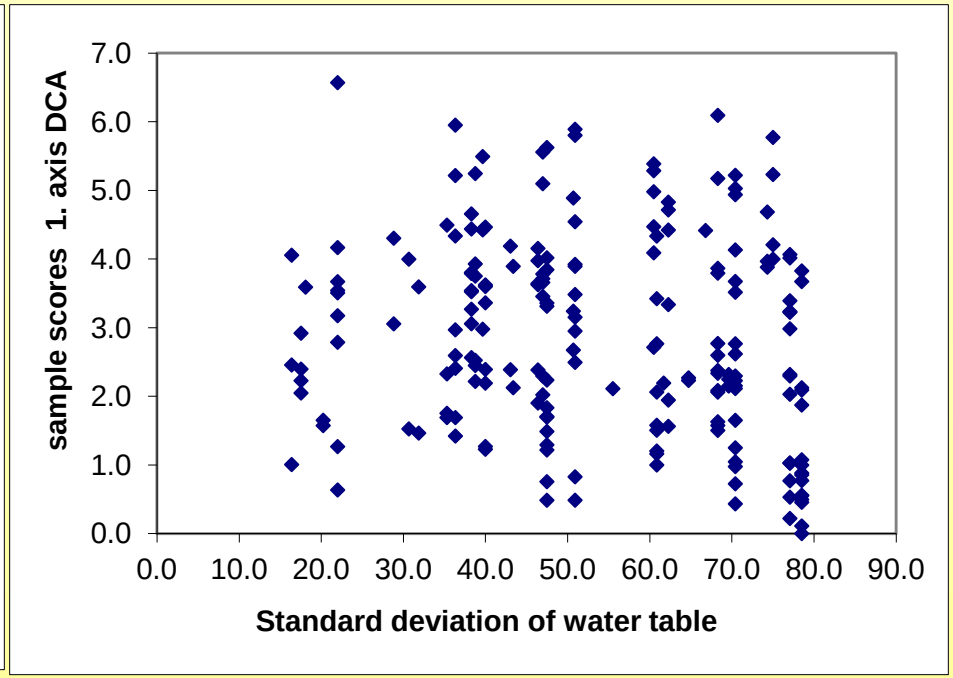


Dune meadow Data
DCA
scatter plot
species scores

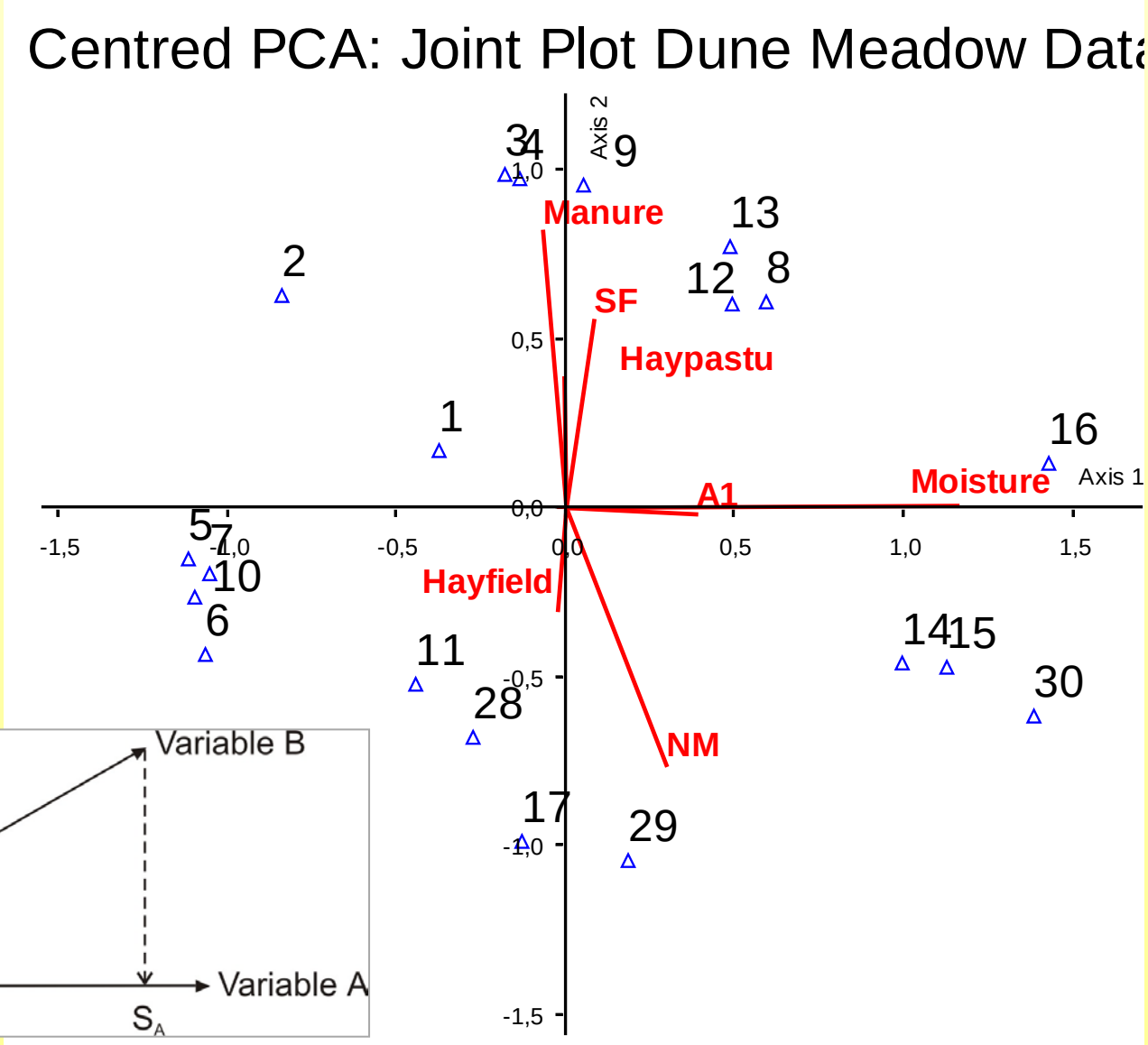
e.g Ellenberg-indicator values, life forms etc.

Correlation of environmental variable with axis in indirect gradient analysis

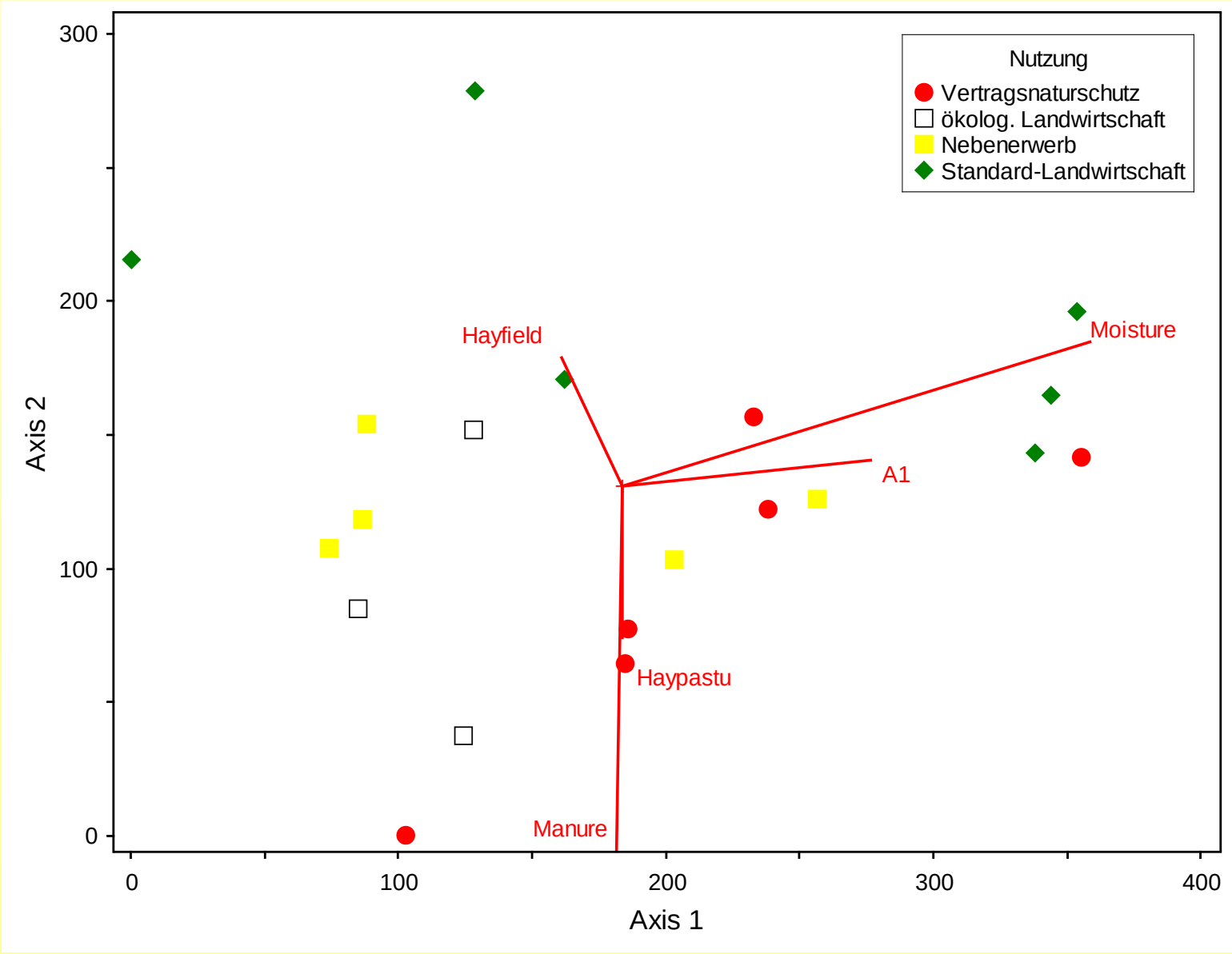




Dune Meadow Data – indirect ordination



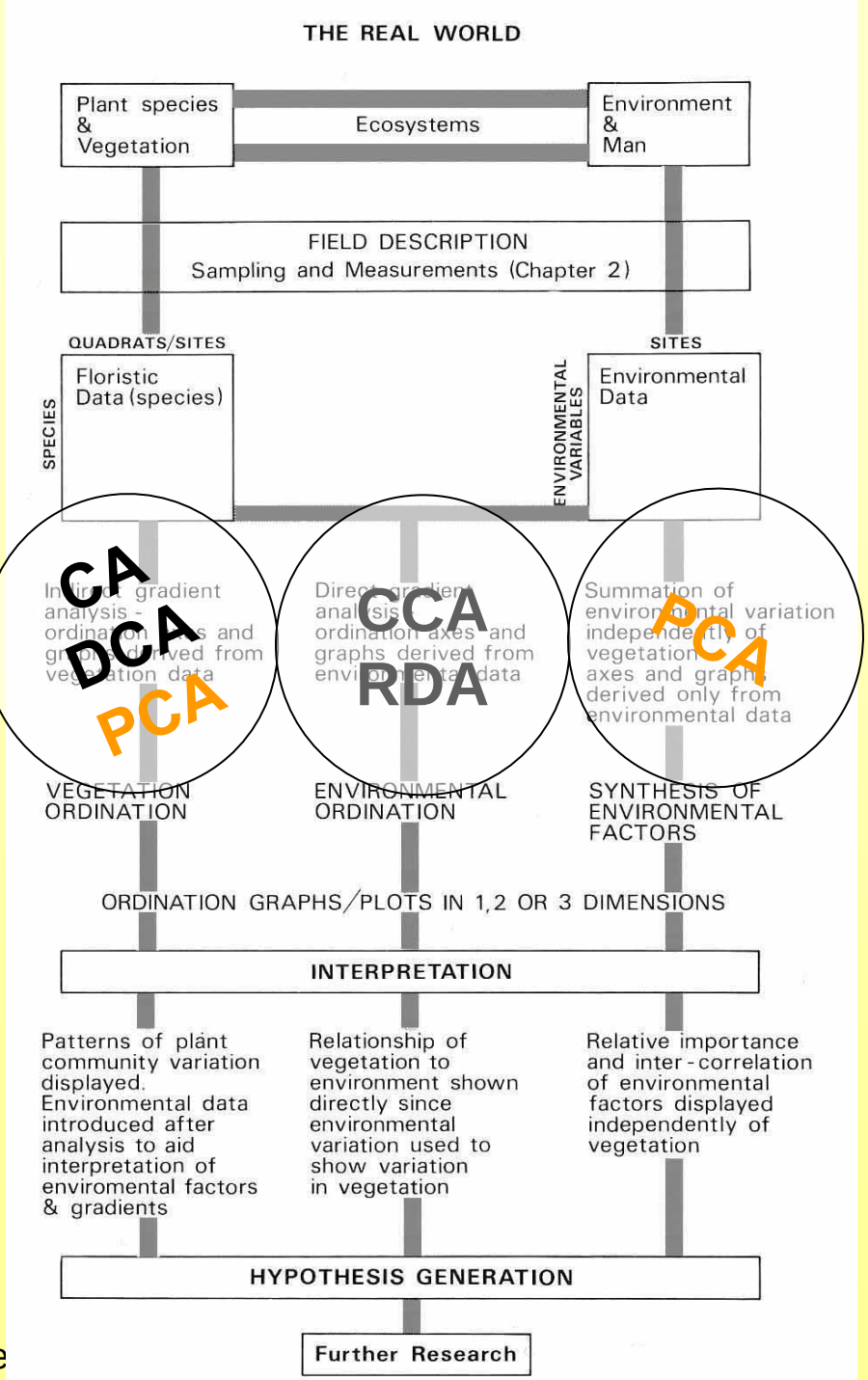
Dune Meadow Data, DCA with environmental variables



correlation (r) sample scores with environmental variables (dune meadow data)

Axis:	1		2		3	
	r	tau	r	tau	r	tau
A1	0.58	0.36	0.19	0.19	0.05	-0.20
Moisture	0.79	0.64	0.44	0.30	-0.18	-0.09
Manure	-0.09	-0.04	-0.70	-0.57	-0.32	-0.24
Hayfield	-0.29	-0.25	0.42	0.28	-0.01	0.01
Haypastu	0.03	0.06	-0.45	-0.30	-0.29	-0.25
Pasture	0.28	0.21	0.05	0.03	0.34	0.28
SF	0.21	0.24	-0.40	-0.32	-0.38	-0.32
BF	-0.29	-0.27	-0.27	-0.23	0.50	0.40
HF	-0.24	-0.21	-0.08	-0.13	-0.33	-0.29
NM	0.24	0.17	0.69	0.62	0.31	0.29

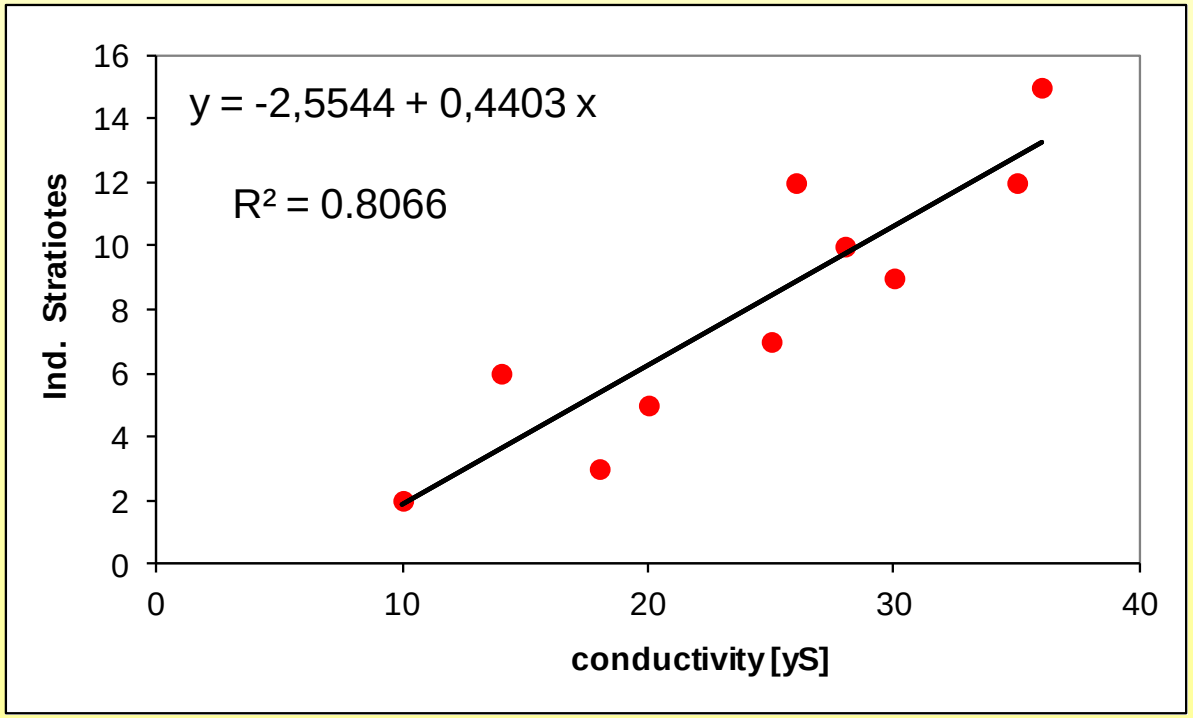
direct ordination



ecological analysis of community-data

linear model – regression analysis

Conduct. [ms]	*Stratiotes aloe.* (Ind)
10	2
14	6
18	3
20	5
25	7
26	12
28	10
30	9
35	12
36	15



-> residual sum of squares is minimised

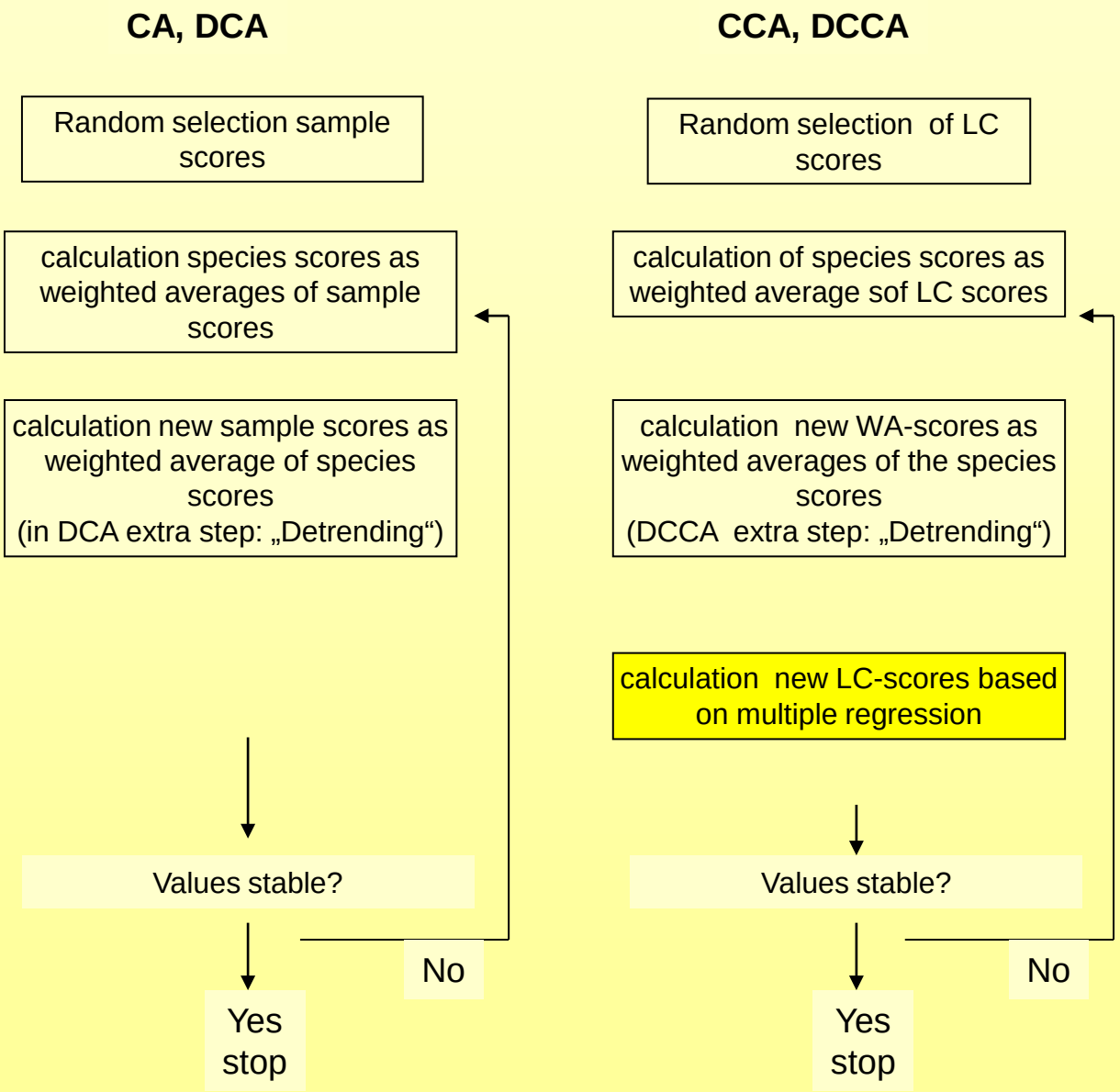
Multiple linear regression

Linear regression can not only describe the response of species along one gradient,
but also along combined gradients.

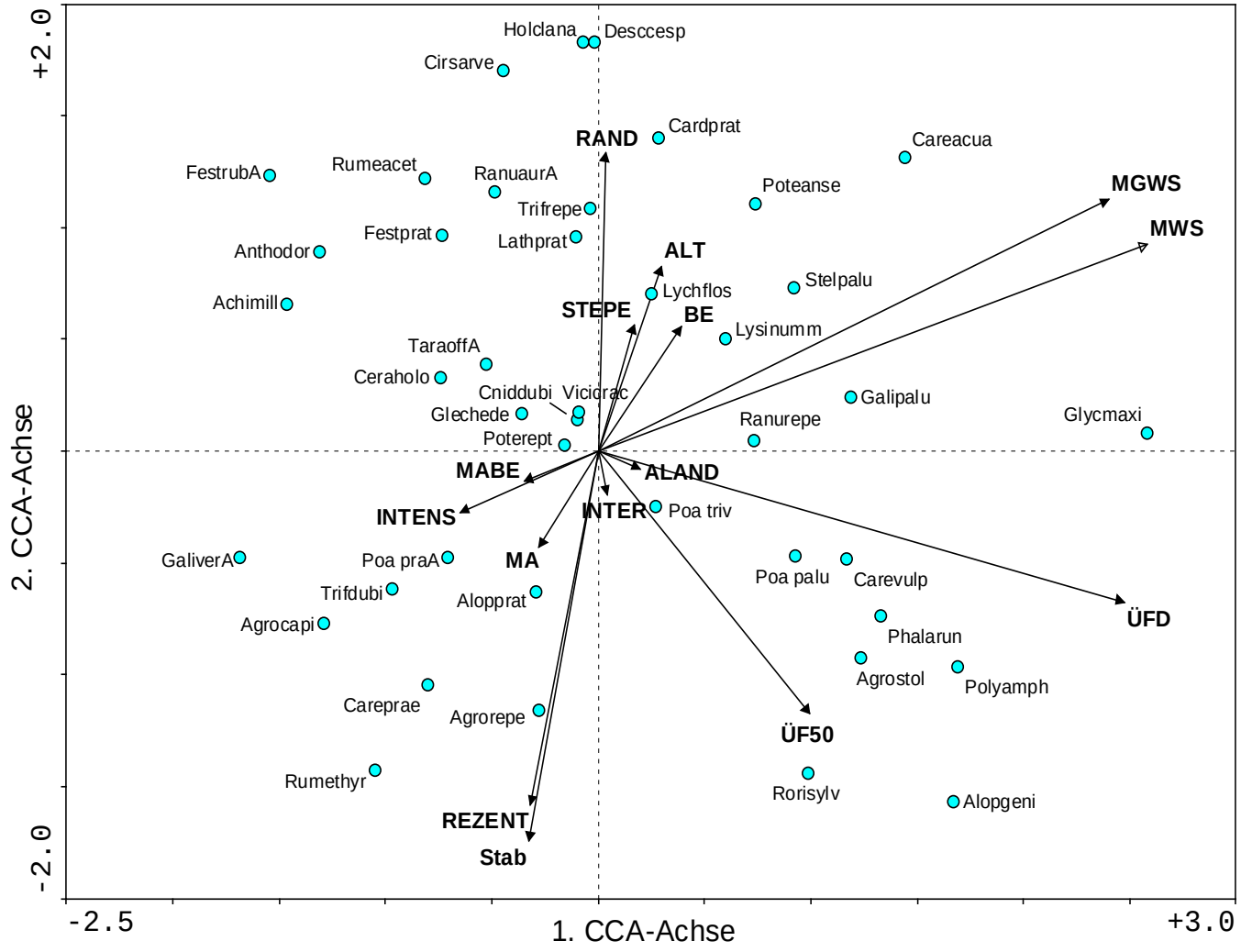
$$y = a + bx$$

$$y = a + b_1x_1 + b_2x_2 + ... + b_qx_q$$

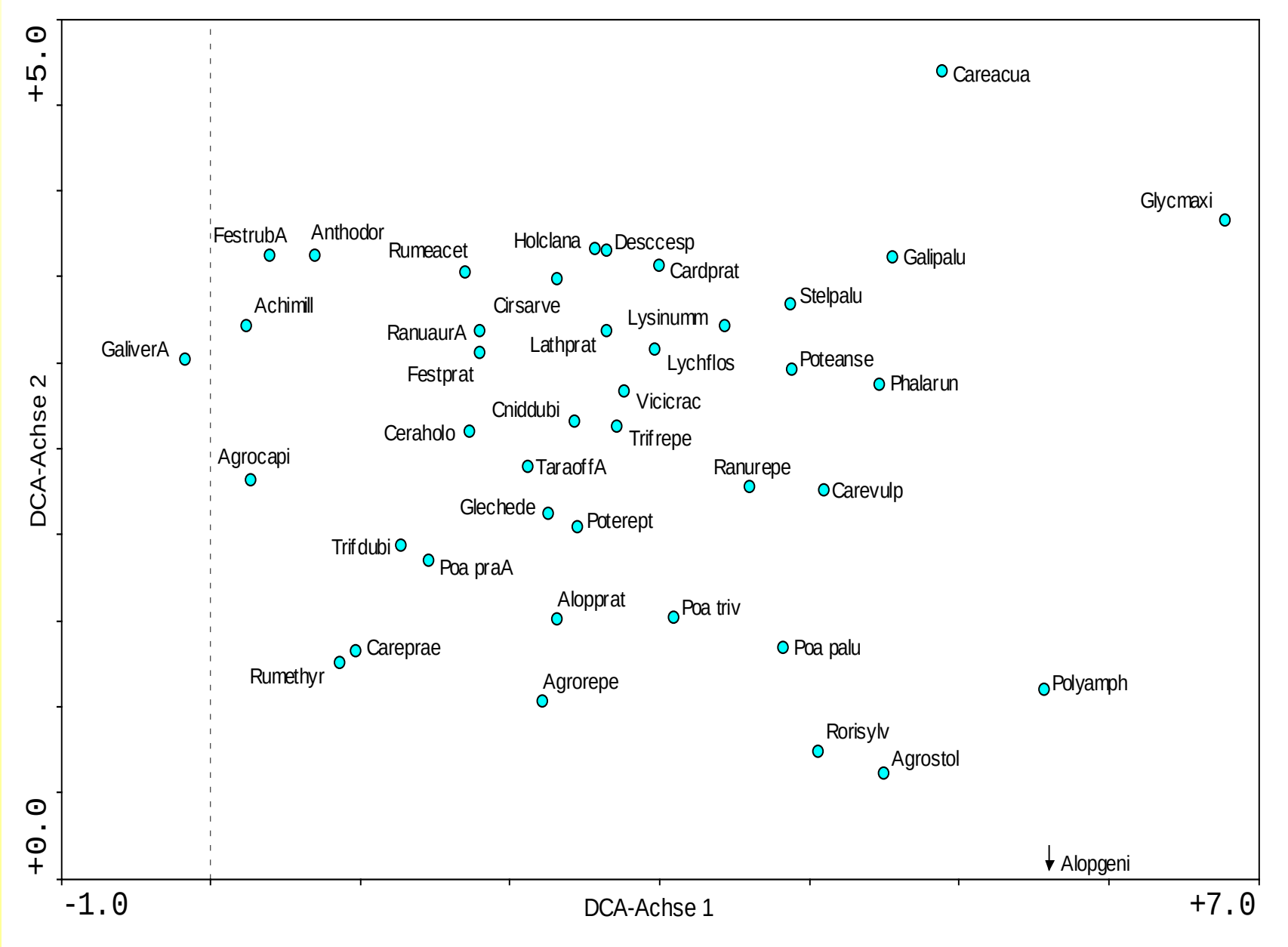
Algorithms for CA, DCA, CCA



CCA - biplot (species/env. variables)

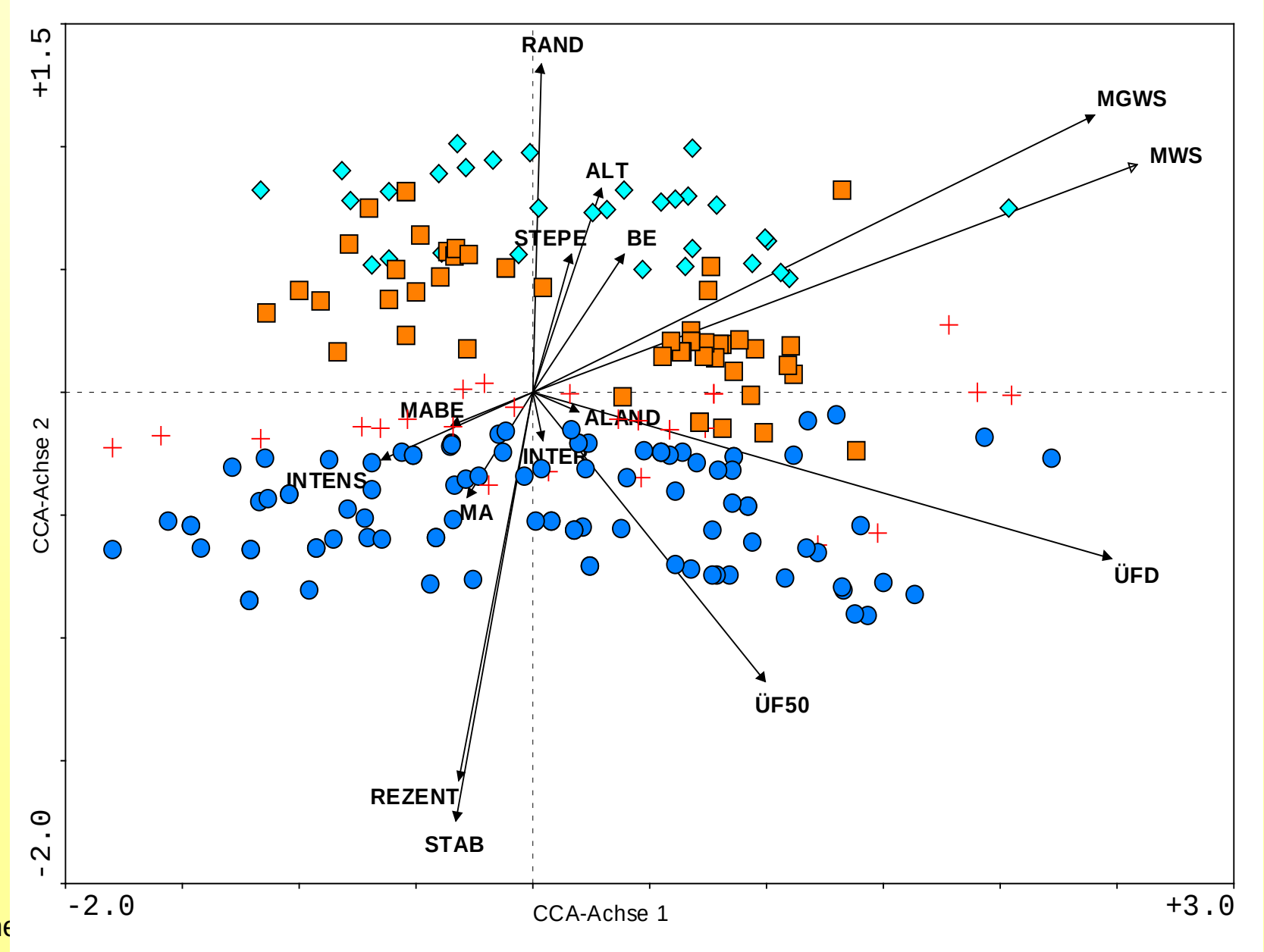


DCA scatter plot species

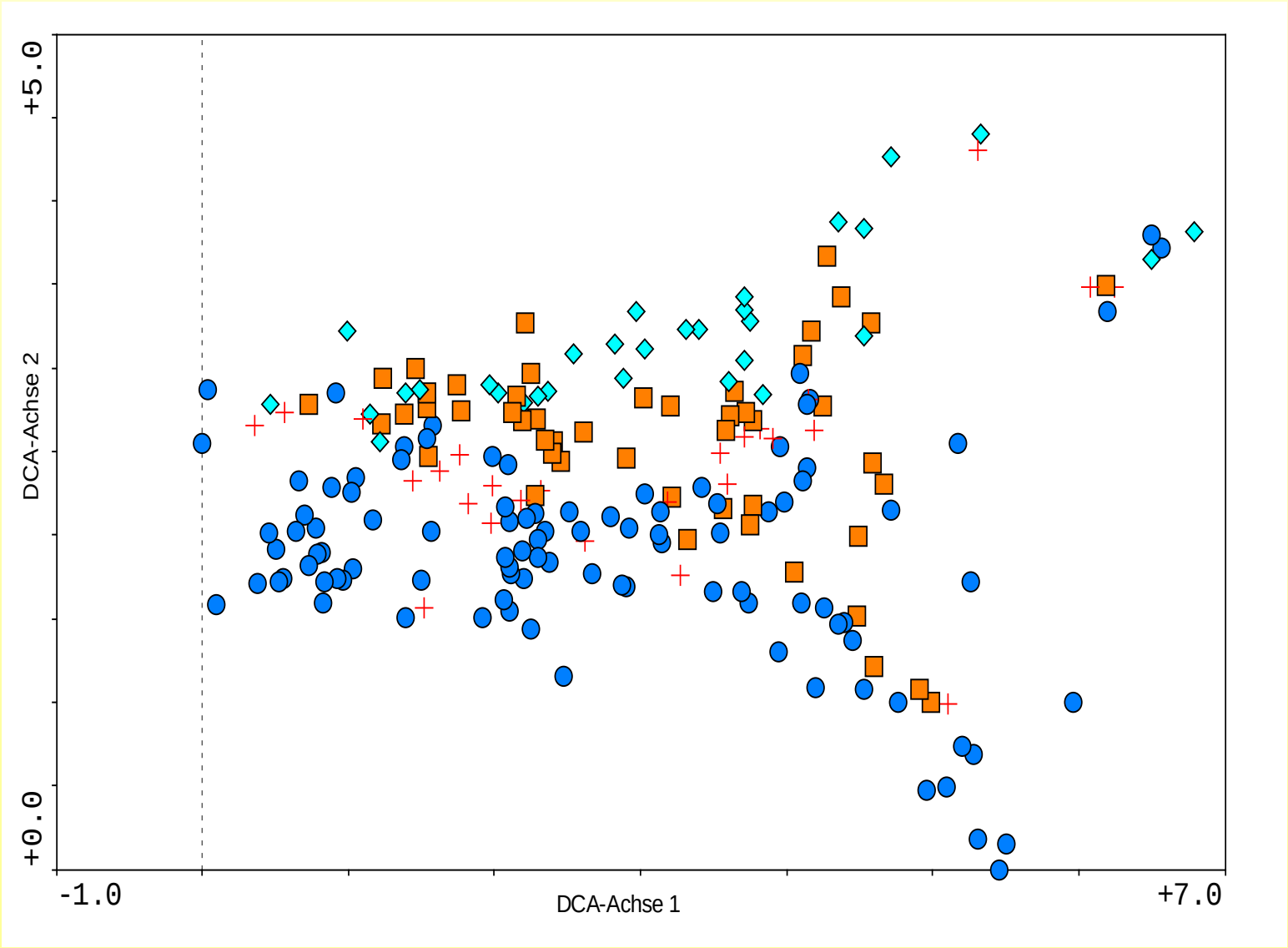


direct ordination

CCA - biplot (samples/env. variables)



DCA scatter plot samples



direct ordination

comparison correlation environmental variables vs. axes of DCA and CCA

	Parameter	DCA Axis 1	DCA Axis 2	CCA Axis 1	CCA Axis 2
Hydrology	AWL	0.82	-0.26	0.80	-0.40
	AGWL	0.76	-0.32	0.74	-0.49
	Flood Duration	0.70	0.22	0.73	0.29
	Flood50	0.26	0.48	0.28	0.49
	SD	-0.15	0.66	-0.13	0.70
Location within the floodplain	Recent	-0.15	0.60	-0.12	0.64
	Inter	0.01	0.03	0.01	0.09
	Older	0.13	-0.25	0.11	-0.36
	Margin	0.03	-0.52	0.02	-0.50
	Stepe	0.04	-0.25	0.05	-0.19
	Aland	0.07	0.06	0.07	0.03
Land use	Hay	-0.10	0.08	-0.09	0.16
	Past	0.15	-0.14	0.14	-0.21
	Mix	-0.11	0.12	-0.12	0.03
	Intens	-0.16	0.27	-0.17	0.03

Summary DCA

**** Summary ****

Axes		1	2	3	4	Total inertia
Eigenvalues	:	0.610	0.252	0.138	0.117	5.022
Lengths of gradient	:	6.570	4.284	3.195	2.251	
Cumulative percentage variance						
of species data	:	12.1	17.2	19.9	22.2	

Sum of all eigenvalues 5.022

Summary CCA

**** Summary ****

Axes	1	2	3	4	Total inertia
Eigenvalues:	0.520	0.206	0.138	0.081	5.022
Species-environment correlations :	0.931	0.868	0.720	0.671	
Cumulative percentage variance of species data:	10.4	14.5	17.2	18.8	
of species-environment relation:	42.9	59.9	71.4	78.0	

Sum of all eigenvalues	5.022
Sum of all canonical eigenvalues	1.212

Environmental variables in constrained ordination

The choice of environmental variables has large impact on ordination outcome!

For explorative analysis:
- All variables known to be important should be used
- It is often desirable to include further (easily measured) variables

=> Unneeded variables can be removed later

For hypothesis-testing:
All environmental variables have to be carefully selected *a priori*, *post-hoc* removal is not permissable

Environmental variables in *constrained ordination*

If number of environmental variables is equal or higher as number of samples, CCA becomes CA
-> 100% of variation in species composition explained (overfitting)

Environmental variables should not be linear combinations of other variables

examples:
Soil types
Dummy variables
Solution: remove redundant variables
(CANOCO discovers linear combinations and removes variables directly)

Environmental variable in constrained ordination

Transformations:
Values for (soil) nutrients, area etc, should be log-transformed.
It may also be sensible to remove impact of outliers (square root, log)

Circular data
Very low and very high values are seemingly similar

-> data have to be transformed
Example:
exposiure
Season-> replace with suitable dummy variable

Randomization-Test

Allow to test significances in correlations between data sets, without constraints of data distribution or many other assumptions.

The Monte-Carlo-permutations-test assesses significance of CCA axis by testing the relation between species and environmental variables**.**

example: Is the number of invertebrate species related to lake area [ha]?

Null-hypothesis: Species number is not correlated to lake area

Lake	Size	Richness
1	0.9	20
2	3.1	40
3	3	55
4	1	36
5	2	41
6	4	62
7	3.5	75
8	3	77

example: species number vs. lake area

Lake	Size	Richness
1	0.9	20
2	3.1	40
3	3	55
4	1	36
5	2	41
6	4	62
7	3.5	75
8	3	77

Question: Which fraction of variance in one variable (e.g. species number) is explained by (related to) variation in the other variable (lake area)?

coefficient of determination $r^2 = 0.623$

The strength of the relation is indicated by one (positive) number.

Which value (of r^2) could be expect edif null-hypothesis is true (i.e. no relation between species number and lake area)?

randomisation-test

Permutation and r^2

Lake	Size	Richness
1	0.9	20
2	3.1	40
3	3	55
4	1	36
5	2	41
6	4	62
7	3.5	75
8	3	77

$r^2 =$ 0.622667309

Lake	Size	Richness
1	0.9	41
2	3.1	75
3	3	62
4	1	77
5	2	36
6	4	20
7	3.5	40
8	3	55

$r^2 =$ 0.101711059

Randomization-Test

Observed („real“) r^2 = 0.622667309

Derived from permutation r^2 =

0.101711059	0.000176755
0.219452888	0.057343488
0.521551724	0.039769863
0.142857143	0.114246593
0.003778495	0.080946695
0.038042288	
0.405050436	
0.002520831	
0.405050436	

Etc.

3 of 100 = 0.03 > r^2_{true}

→ **P < 0.05**

Monte Carlo-Permutation Test

permutation test

1. Choose a test-statistic that is suitable for the data set (e.g. t-Test, F-Test).
2. Define null-hypothesis.
3. Generate new data sets from real data, which are random assemblages → samples of environmental variable dataset are randomly redistributed to samples of species data set.
4. Calculate test-statistic for random assemblage and compare values with original value.
5. Repeat steps 3 and 4 many times (e.g. 9999 permutations).
6. If the original value is larger than 95% of random values, reject null-hypothesis (level $p < 0.05$).

Exkurs: Monte Carlo Tests

The technique can be extended to many other problems of inference statistics / hypothesis testing.

Example: Test similarity between two data matrices; calculate correlation between matrices.
R^2 is than tested by keeping main matrix constant and permute the second matrix. If >95% of r^2 values for rearranged matrices are lower than initially observed, the relationship can be regarded siginificant.

$\Rightarrow$ Mantel-Test

(e.g. non-parametric alternative for ANOVA, s. Sokal and Rohlf 1995, Manly 1998)

Forward Selection

Forward Selection allows testing of each environmental variable for being redundant or not significant with respect to explanation of variance in species composition.

The available environmental variables are tested one by and the overall model is built up stepwise.

Forward Selection

Initially, the explained variance is separately estimated for each environmental variable, i.e. the contribution of each variable to explain the variation in the species composition.

The variable with the hightest variance is chosen and tested for significance with respect to the fraction of variance explained in the species data set (permutation test, or AIC). If the contribution is significant, the variable is built in the model.

There may be (and very often are) correlations between already selected environmental variables and remaining variables. In consequence, building new variables in the model reduces the variance potentially explained by remaing variables to the non-redundant fraction, resulting in changes of explained variance in the remaining variables.

All variables are successively tested and selected until further improvements of the model have no sifnificant effects. The final model is than reported.

Ordination
Direct gradient analysis

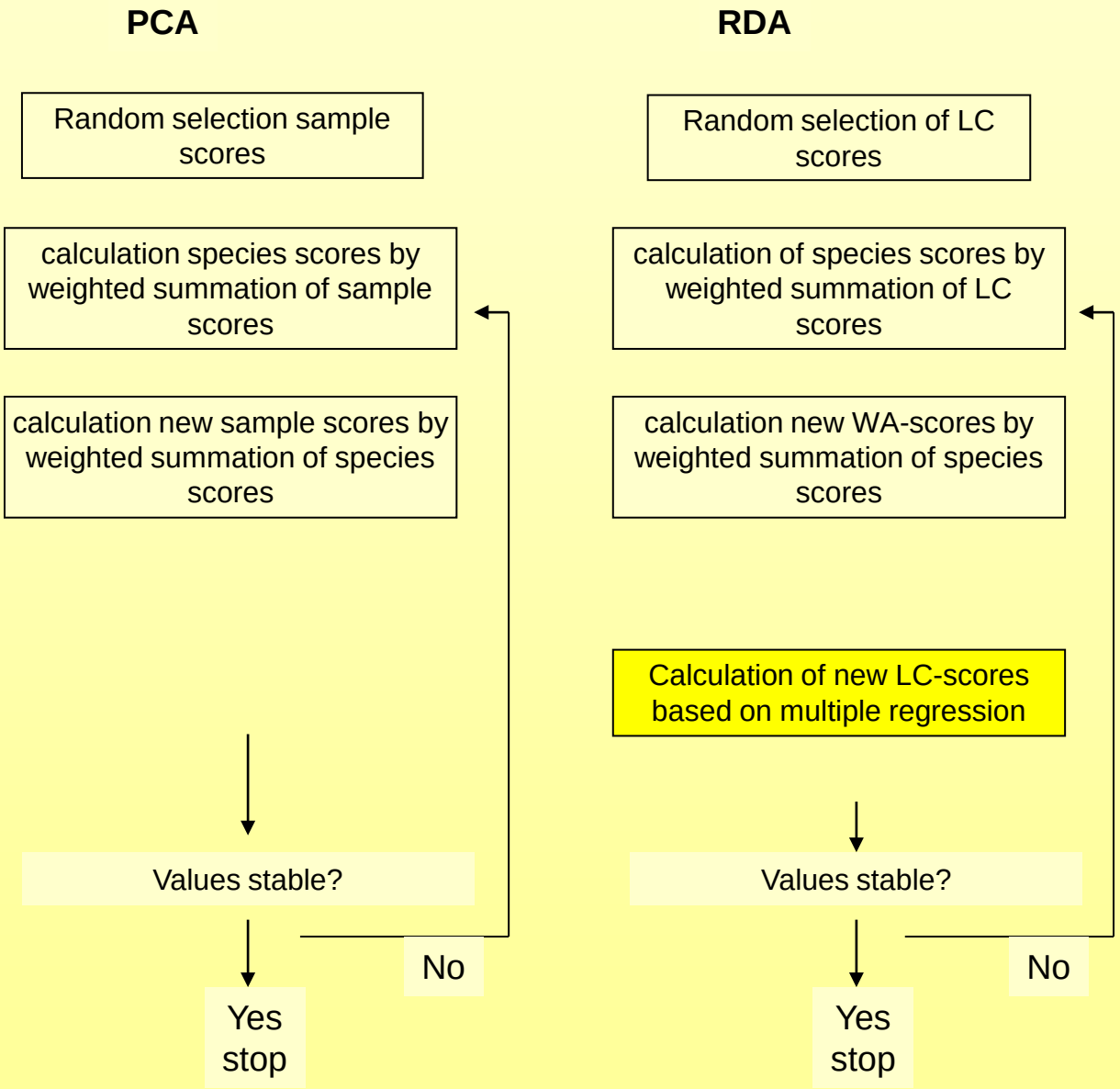
The species composition is directly analysed with respect to measured environmental variables

canonical correspondence analysis (CCA)

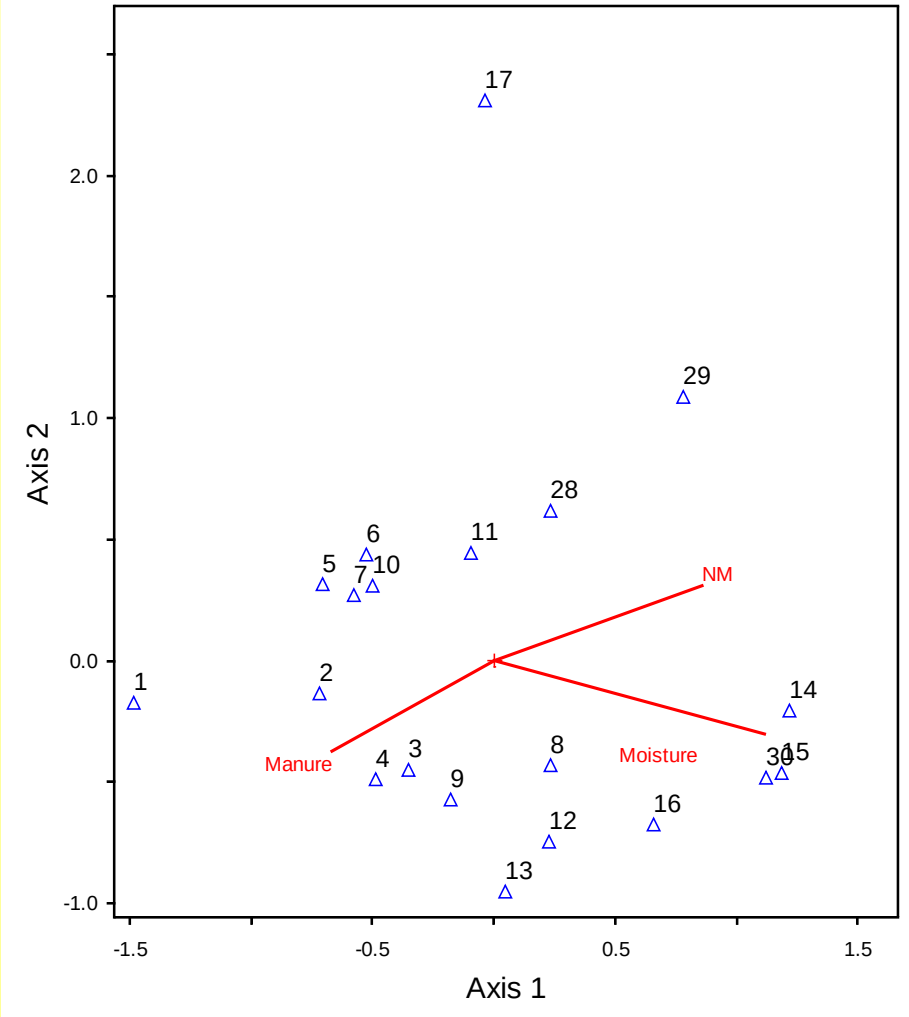
„detrended“ form of CCA (DCCA)

redundancy analysis (RDA)

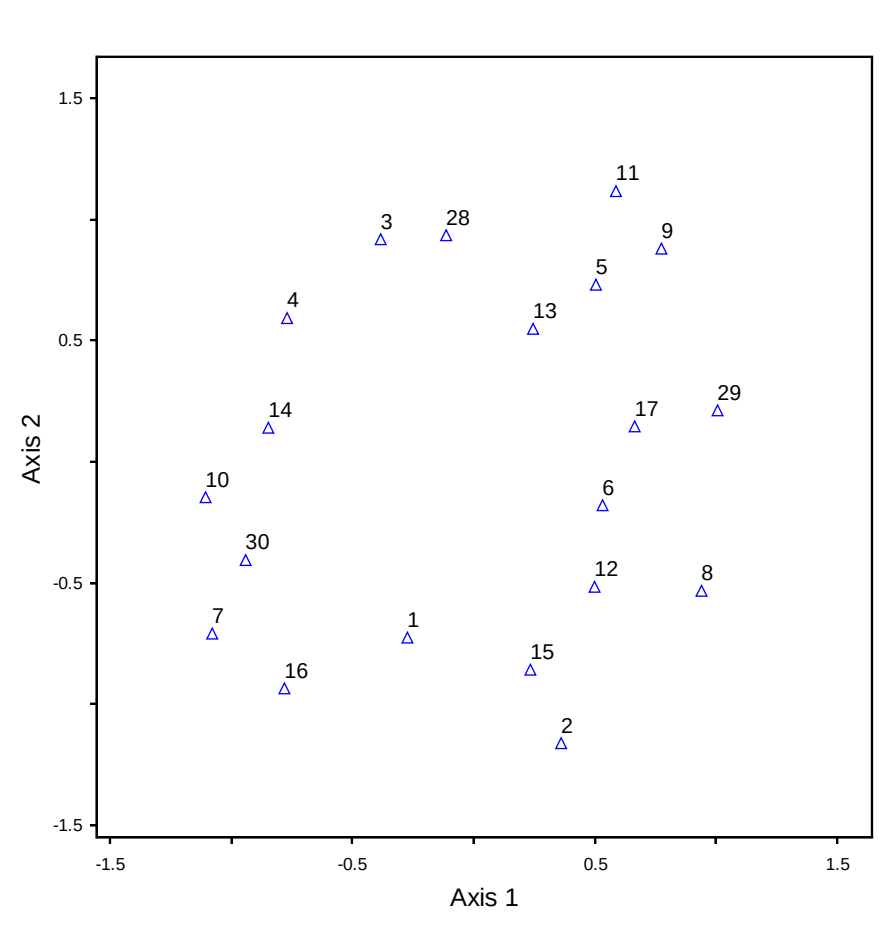
algorithm for PCA, Redundancy Analysis



Non-Metric Multidimensional Scaling (NMDS)



Non-metric multidimensional scaling



Principle:

- calculation of ecological dissimilarity (e.g Bray-Curtis)
- Choose no. of axes (dimensions) to be tested
- Distribution (usually random) of sites in ordination space
- calculation of distance in ordination space (Euclidean distance)
- Move samples until ecological distance and ordinations distance are strongly correlated

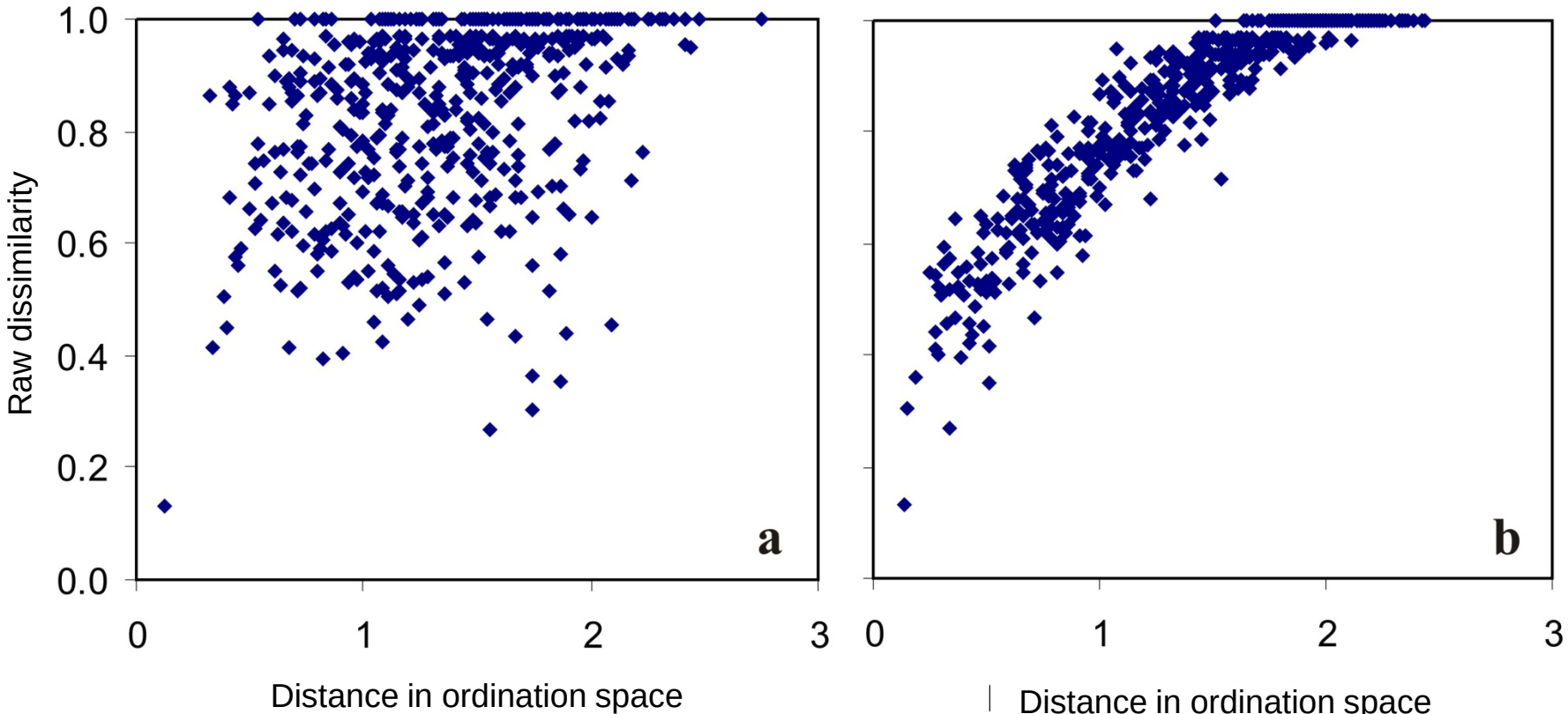
Dune Meadow Data: Random start configuration for NMDS, Bray-Curtis-similarity, stress 44.7

Non-metric multidimensional scaling

Shepard-Diagram for comparison ecological distance – distance in ordination space

a) Random configuration

b) Optimised (500 Iterationen)

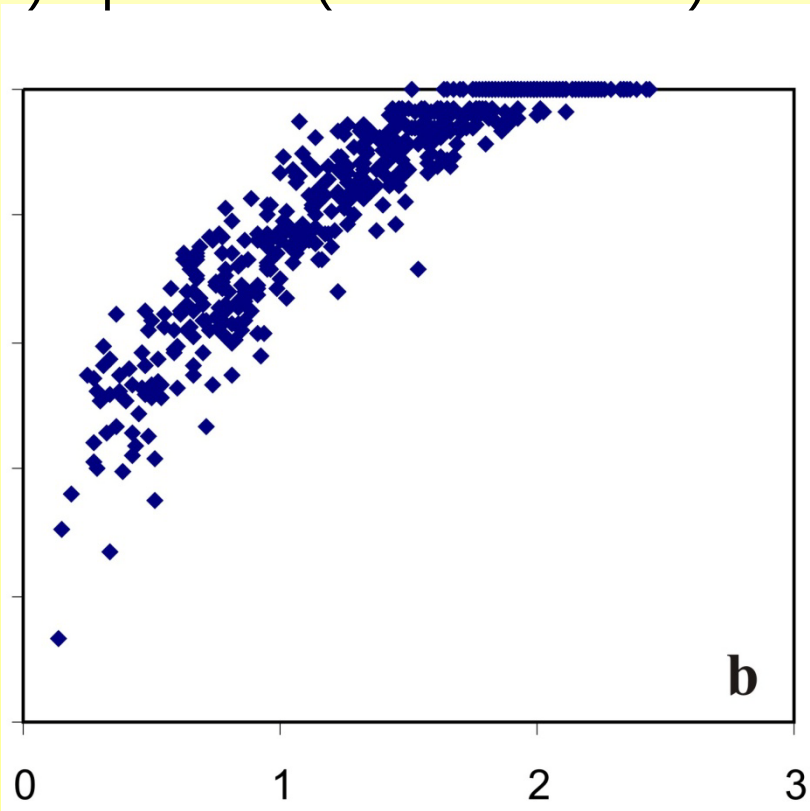


Non-metric multidimensional scaling

Shepard-Diagram for comparison ecological distance – distance in ordination space

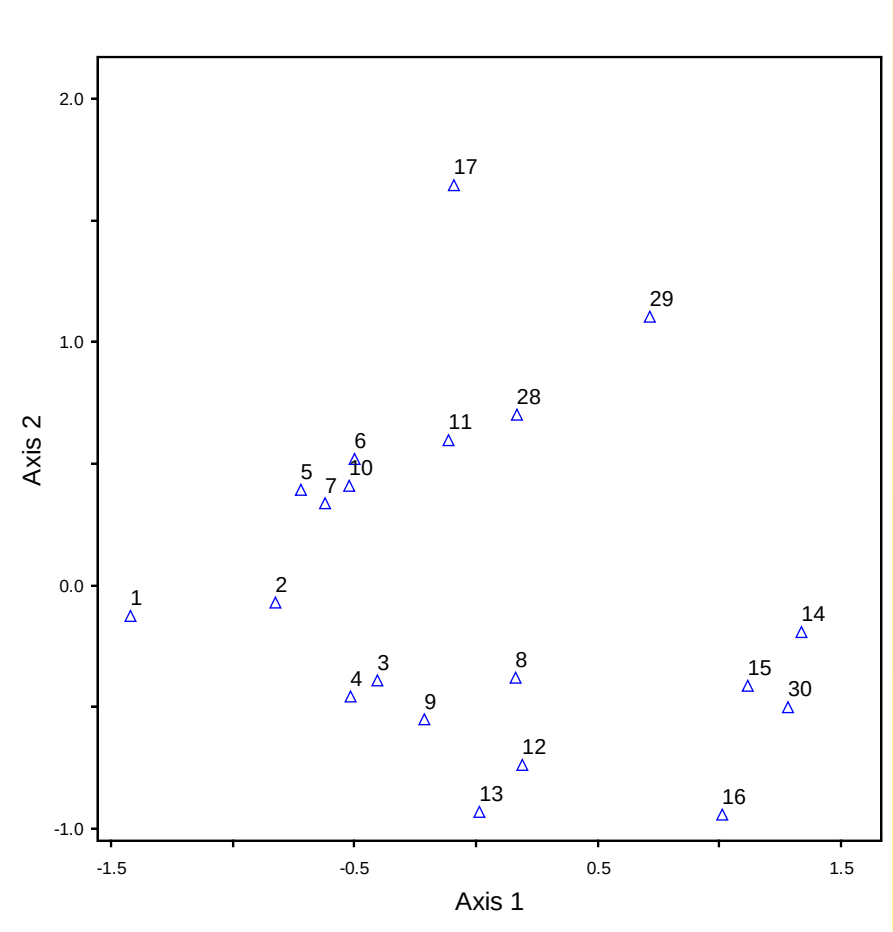
$$Stress_I = \sqrt{\sum_{hi}(d_{hi} - \hat{d}_{hi})^2 / \sum_{hi} d_{hi}^2}$$

b) Optimised (500 Iterationen)

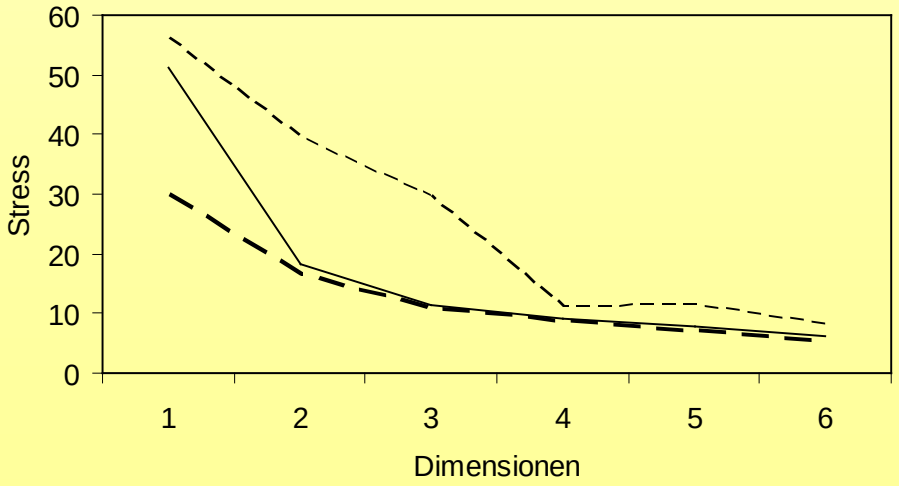


Non-metric multidimensional scaling

Optimised configuration, chosen from trials



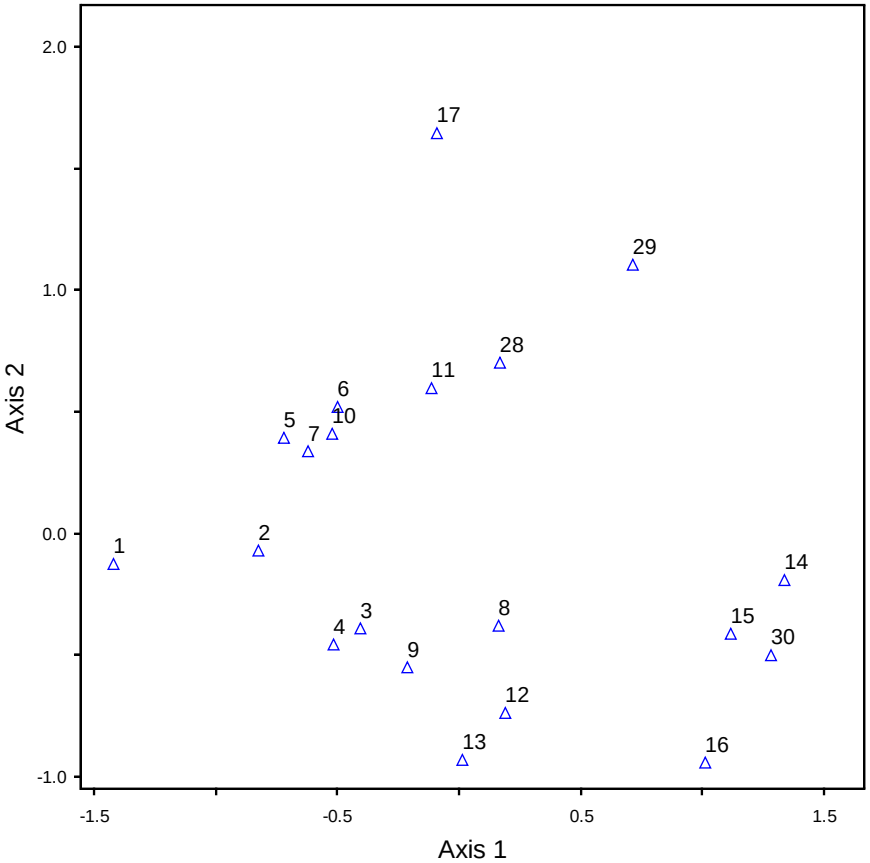
Development of stress across various runs (250, randomised)



Dune Meadow Data: Optimised configuration, 250 runs, Bray-Curtis-similarity, Stress 2D: 12.0

Non-metric multidimensional scaling

Optimised configuration, chosen from trials



Significance tests: Permutation by comparison of „real“ result (stress) with randomised runs (without optimisation)

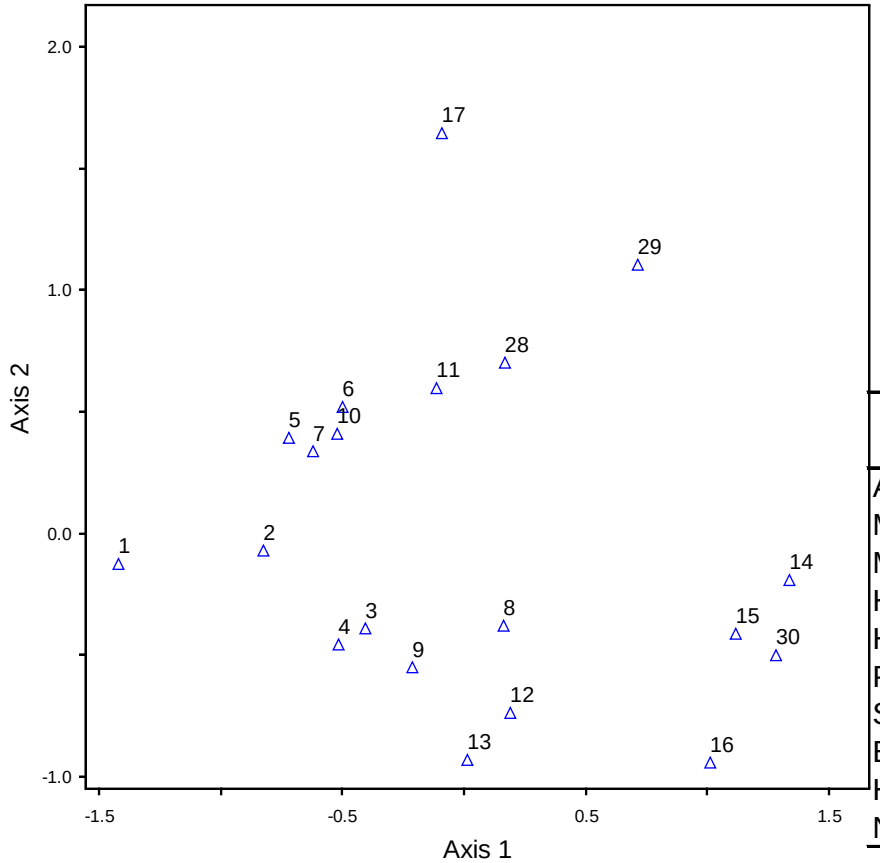
STRESS IN RELATION TO DIMENSIONALITY (Number of Axes)

	Stress in real data 1 run(s)			Stress in randomized data Monte Carlo test, 249 runs			
Axes	Minimum	Mean	Maximum	Min	Mean	Max	p
2	12.022	12.022	12.022	36.721	41.928	47.72	0.004

Dune Meadow Data: Optimised configuration, 250 runs, Bray-Curtis-similarity, Stress 2D: 12.0

Non-metric multidimensional scaling

Optimised configuration, chosen from trials



Fitting environmental variables

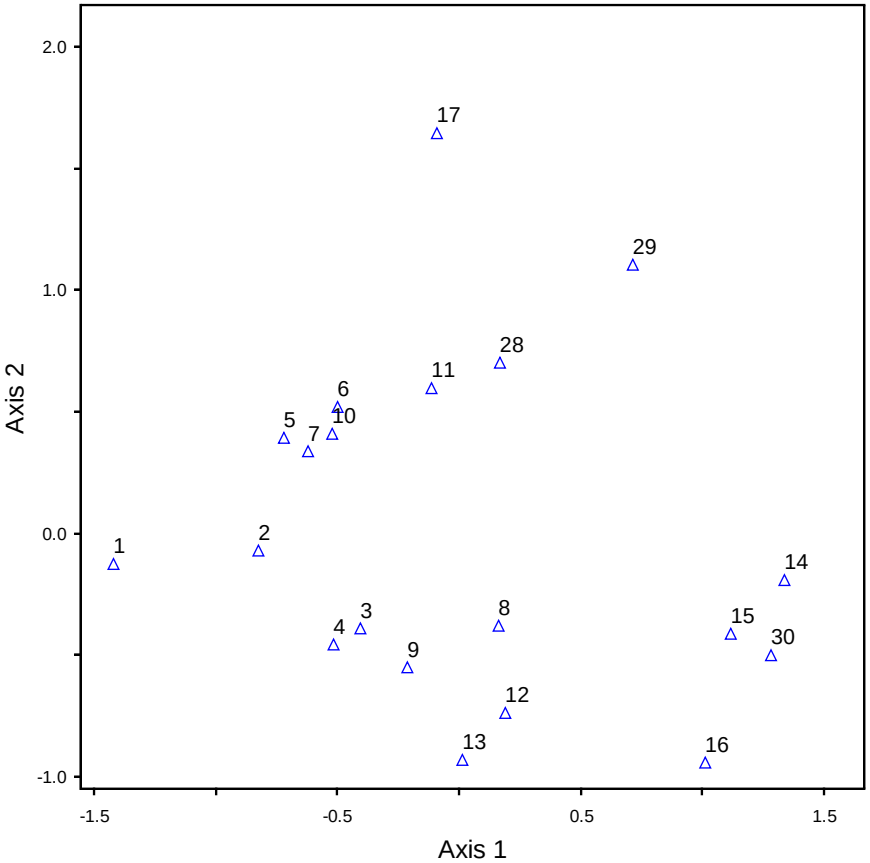
As in other indirect ordination methods: *post hoc* correlation

	Achse 1		Achse 2	
	r	tau	r	tau
A1	0.59	0.34	-0.27	-0.22
Moisture	0.80	0.70	-0.42	-0.41
Manure	-0.62	-0.43	-0.46	-0.34
Hayfield	0.11	0.07	0.51	0.34
Haypastu	-0.31	-0.25	-0.41	-0.33
Pasture	0.23	0.21	-0.09	-0.01
SF	-0.22	-0.10	-0.52	-0.52
BF	-0.26	-0.27	0.12	0.23
HF	-0.29	-0.28	0.00	0.08
NM	0.70	0.57	0.42	0.27

Dune Meadow Data: Optimised configuration, 250 runs, Bray-Curtis-similarity, Stress 2D: 12.0

Non-metric multidimensional scaling

Optimised configuration, chosen from trials

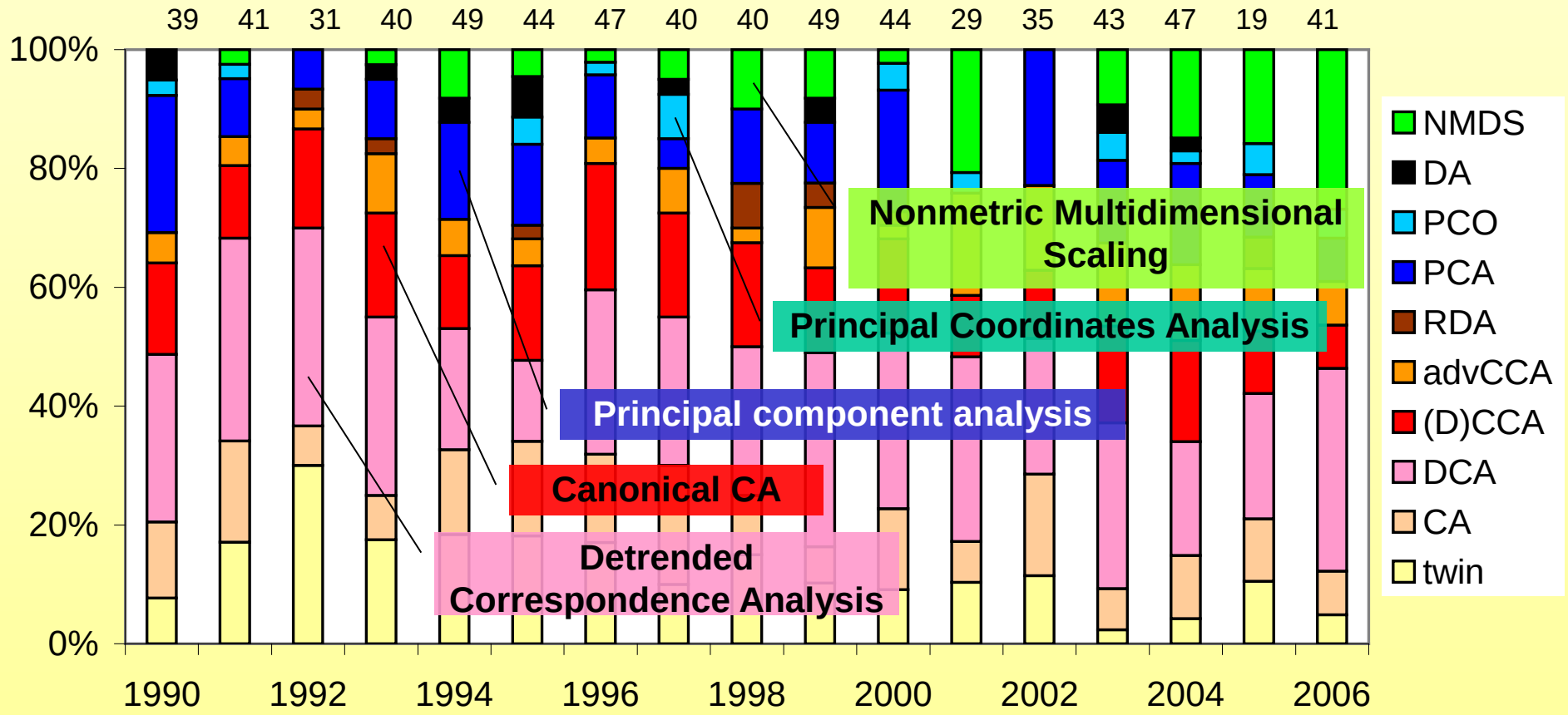


Summary approach

- Choose distance measure (transform data if needed)
- Find best configuration by comparing several NMDS (criterion stress value)
- Test axis for stability by permutation test against non-optimised configuration
- If needed, rotate axes to display the most important axes
- Understand pattern in diagram, *post hoc* fitting if needed

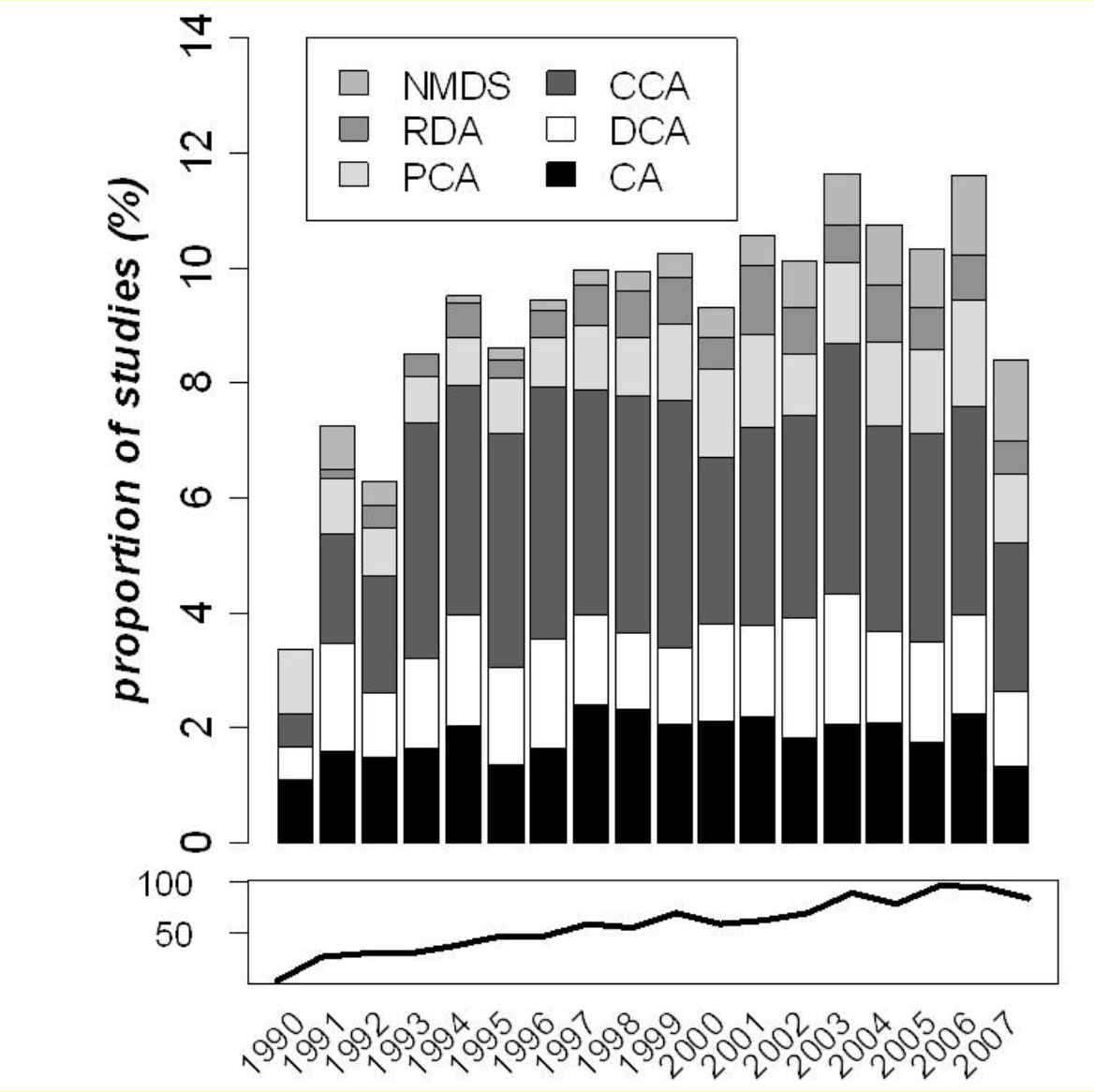
Dune Meadow Data: Optimised configuration, 250 runs, Bray-Curtis-similarity, Stress 2D: 12.0

usage of ordination methods



Literature search: Application of multivariate ordination methods in *Journal of Vegetation Science* (since 1990)

usage of ordination methods



In: all ISI journals ecology

von Wehrden et al., *Journal of vegetation Science*, 2009

Analysis of trait data

The „Fourth Corner Problem“: Relating species (trait) data to environmental information

(D)CA, NMDS, PCA

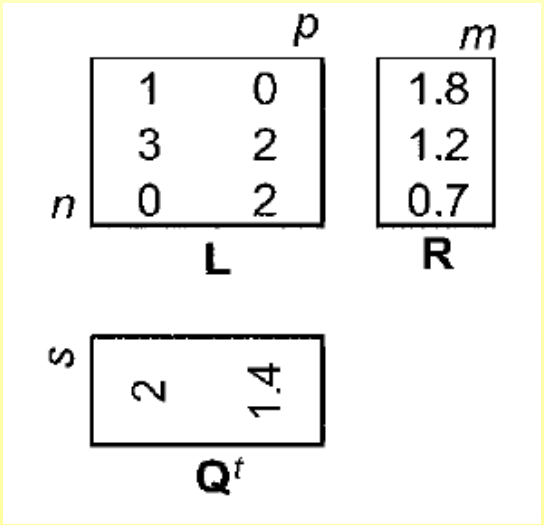
Species – sample matrix

CCA, RDA, indirect gradient analysis

Sample – environ. matrix

PCA

R mode:
CCA
RDA
Indirect gradient analysis



Dray & Legendre *Ecology* 2008

PCA

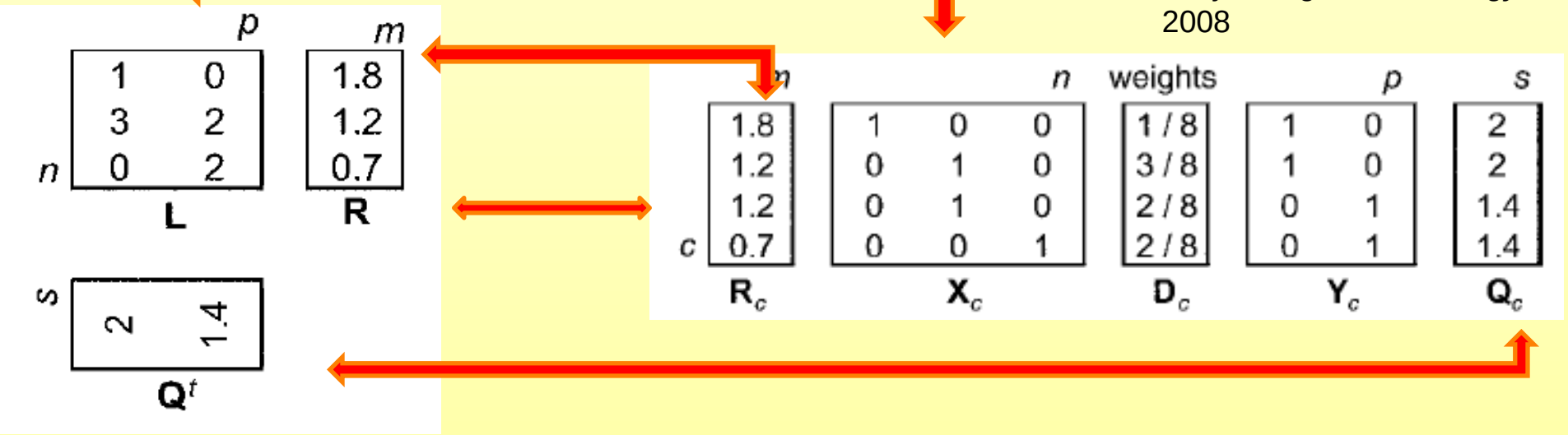
Species – trait matrix

?

Analysis of trait data

Fourth corner analysis: Inflate original tables and then perform separate analysis (qualitative, quantitative possible)

Dray & Legendre *Ecology* 2008



Statistics: Qualitative (presence) vs. qualitative variable (e.g. lifeform): **chi-square**

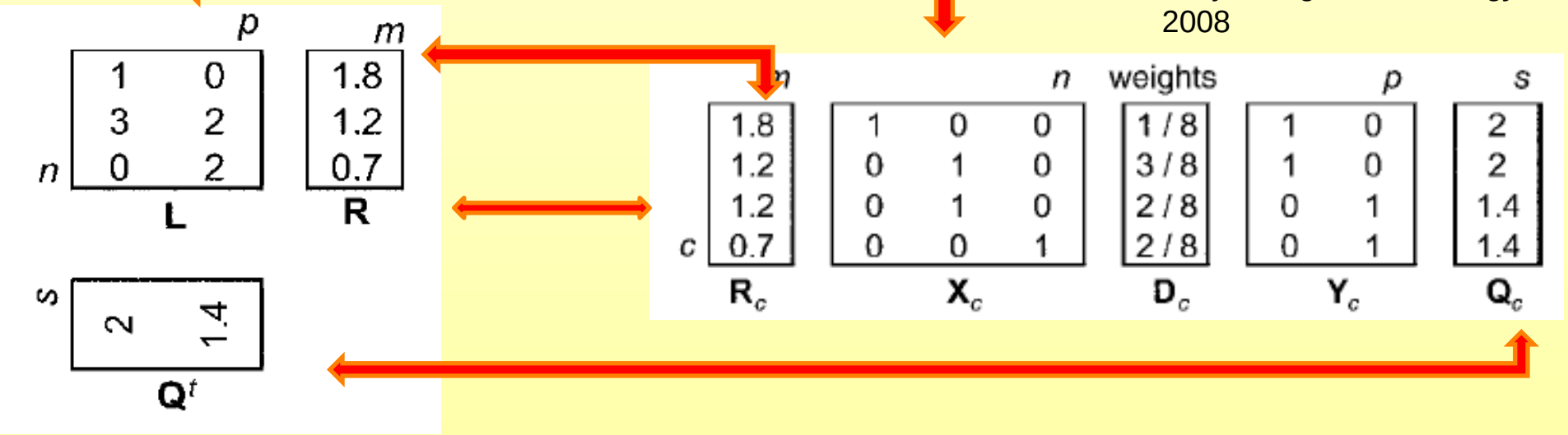
Qualitative (presence) vs. quantitative variable (e.g. height): **F-test** or **correlation ratio**

Quantitative (abundance) vs. quantitative variable (e.g. height): **Pearson r**

Analysis of trait data

Fourth corner analysis: Inflate original tables and then perform separate analysis (qualitative, quantitative possible)

Dray & Legendre *Ecology* 2008



Statistics:

Multivariate summary fourth corner statistic:

Quantitative variables: standardise to zero mean unit variance

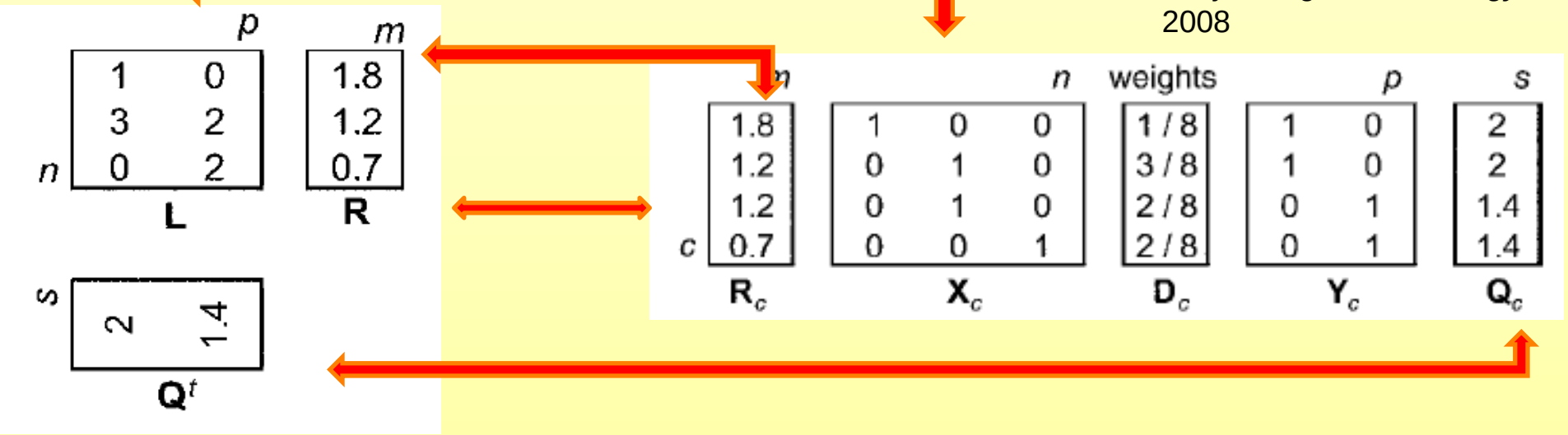
Qualitative variables: replace with dummy variables

➢ **Sum chi-square, r^2, correlation ratio**

Analysis of trait data

Fourth corner analysis: Inflate original tables and then perform separate analysis (qualitative, quantitative possible)

Dray & Legendre *Ecology* 2008



Statistics:

Test statistics by permutation under appropriate null model (for details s. Dray & Legendre 2008)

Analysis of trait data

Fourth corner analysis: Example with real data:

Moist meadows sampled in the 1950/60s and 2008 (606 samples, 281 species, 11 (groups of) traits) (Wesche et al. *Biol. Cons.* in revision)

Decreasing	$trend_{time}$	p_{adj}	$trend_N$	p_{adj}	Increasing	$trend_{time}$	p_{adj}	$trend_N$	p_{adj}
Lifeform	χ^2	***	F	ns	Land use IVs	F		r	
– Hydrophyte	↓		ns		Mowing tolerance (F)	↑	(*)	↑	***
– Hemicryptophyte	↓		ns		Grazing tolerance (F)		ns	↑	***
– Phanerophyte	↓		ns		Fodder value (F)	↑	***	↑	***
– Therophyte	↓		ns						
Strategy	χ^2	***	F	***	Phenological groups	F		r	
– CR strategy	↓		ns		Season	↑	***	↑	***
– CSR strategy	↓		↓						
– R strategy	↓								
– SR strategy	↓								
Clonality	χ^2	***	F	***					
– runner	↓		↓						
– running rhizome	↓		ns						
– bulbillae	↓		ns						
– buds root	↓		↓						
– rhizome	↓		↓						
– pleiocorm	↓		ns						

Analysis of trait data

RLQ-Analysis (Jan Hanspach)

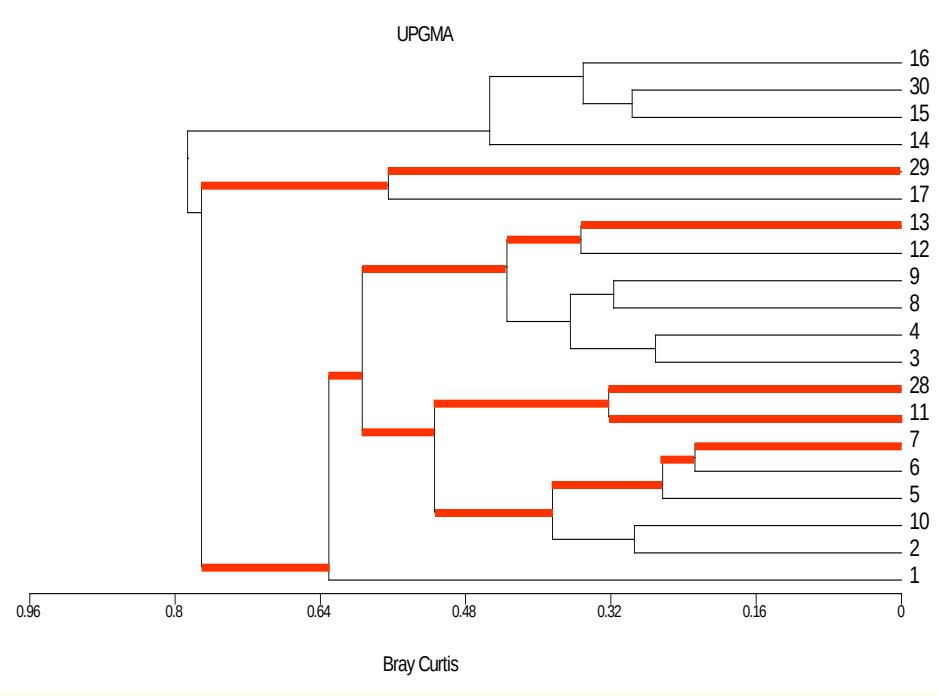
Analysis of trait data

Functional diversity: Measuring diversity of functional traits

FD (sensu Petchey & Gaston *Ecol. Letters* 2002, modified according to Podani & Schmera *Oikos* 2007): quantitative

- Calculate Gower-similarity among species in trait space
- Calculate (UPGMA) dendrogram based on similarity matrix
- For each community, sum length of branches in dendrogram for occurring species

➢ FD

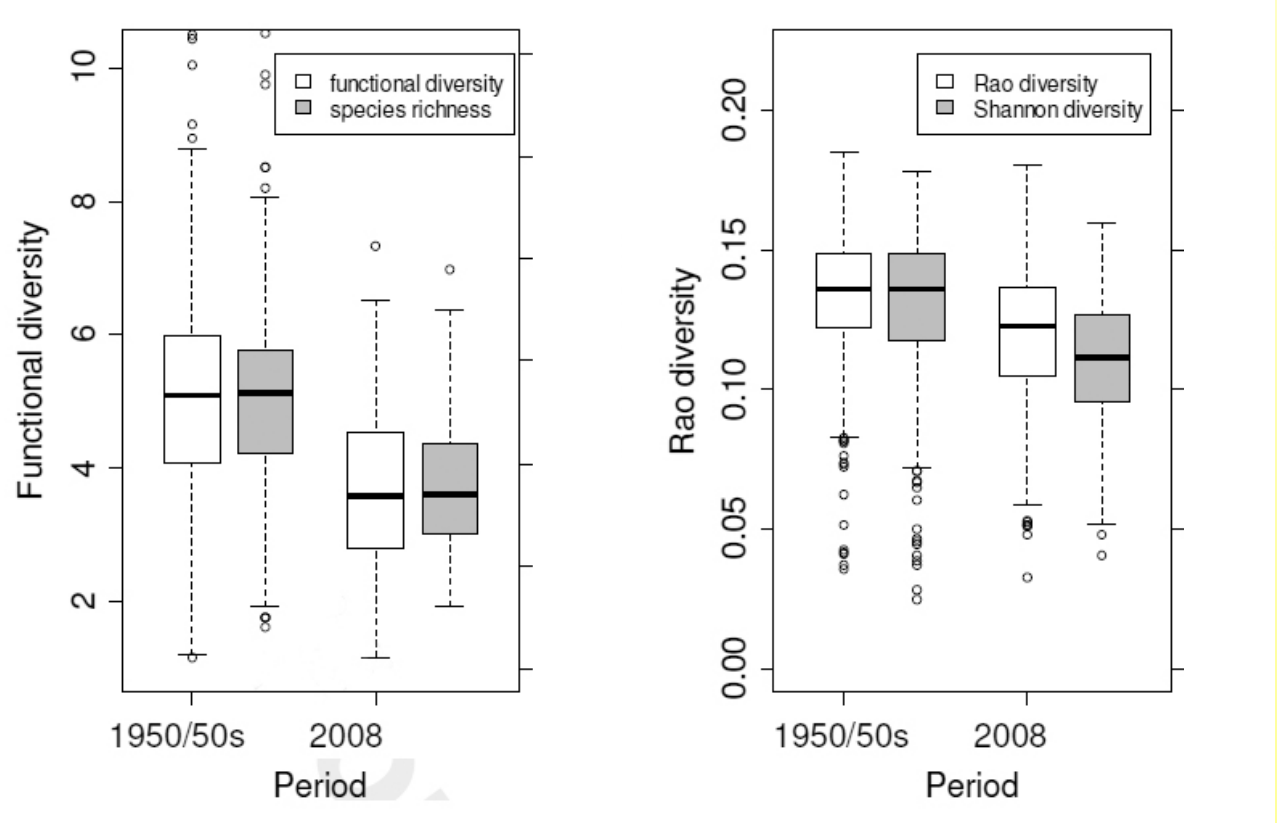


Analysis of trait data

FD: Example moist meadows (1950/60s vs. 2008)

- Correlated with richness

Wesche et al. Biol. Cons. in revision

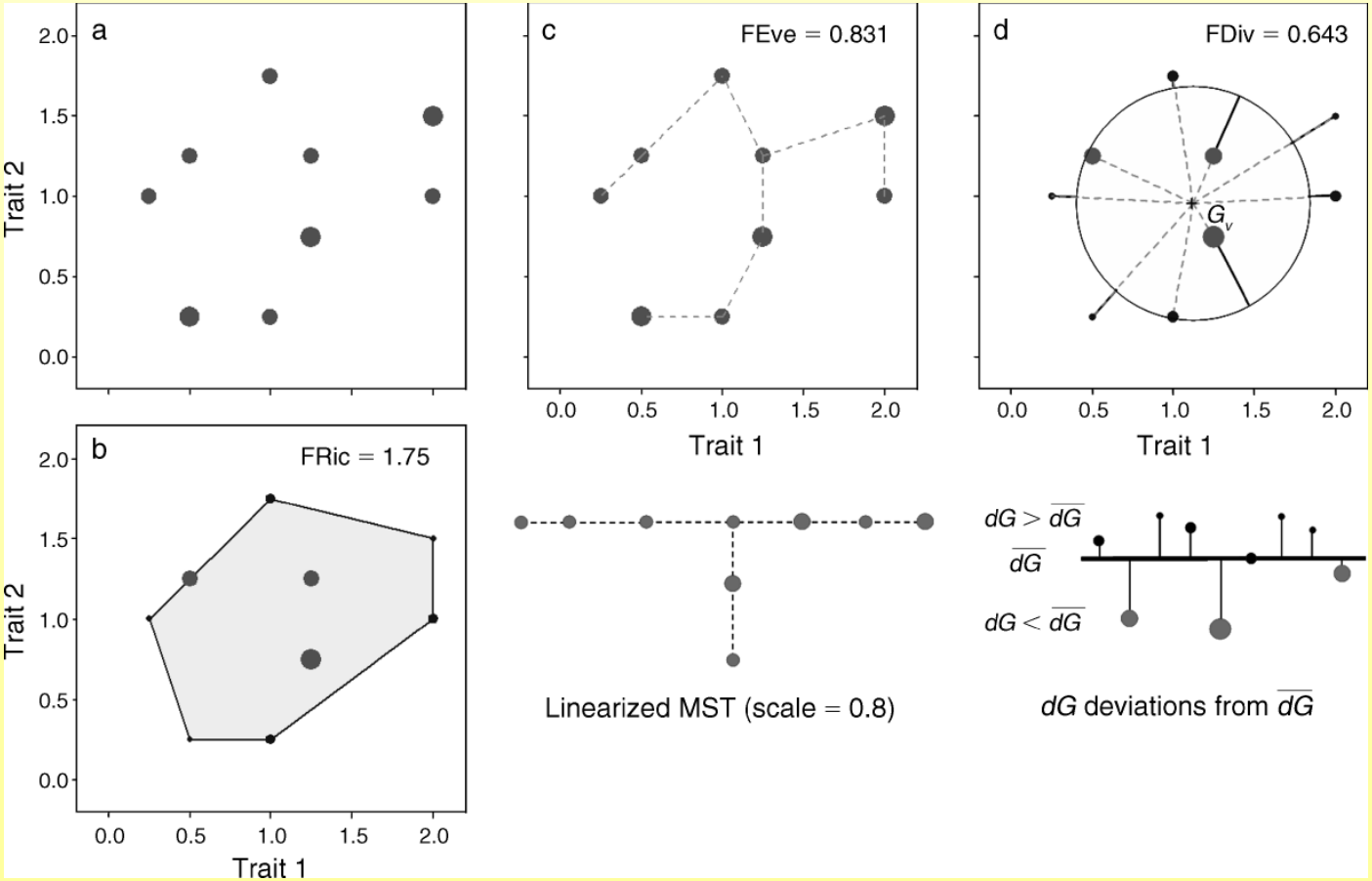


Rao‘s quadratic entropy – considering abundance

- Correlated with diversity

Analysis of trait data

Alternative: View different facets of functional diversity (Villeger et al. *Ecology* 2008)

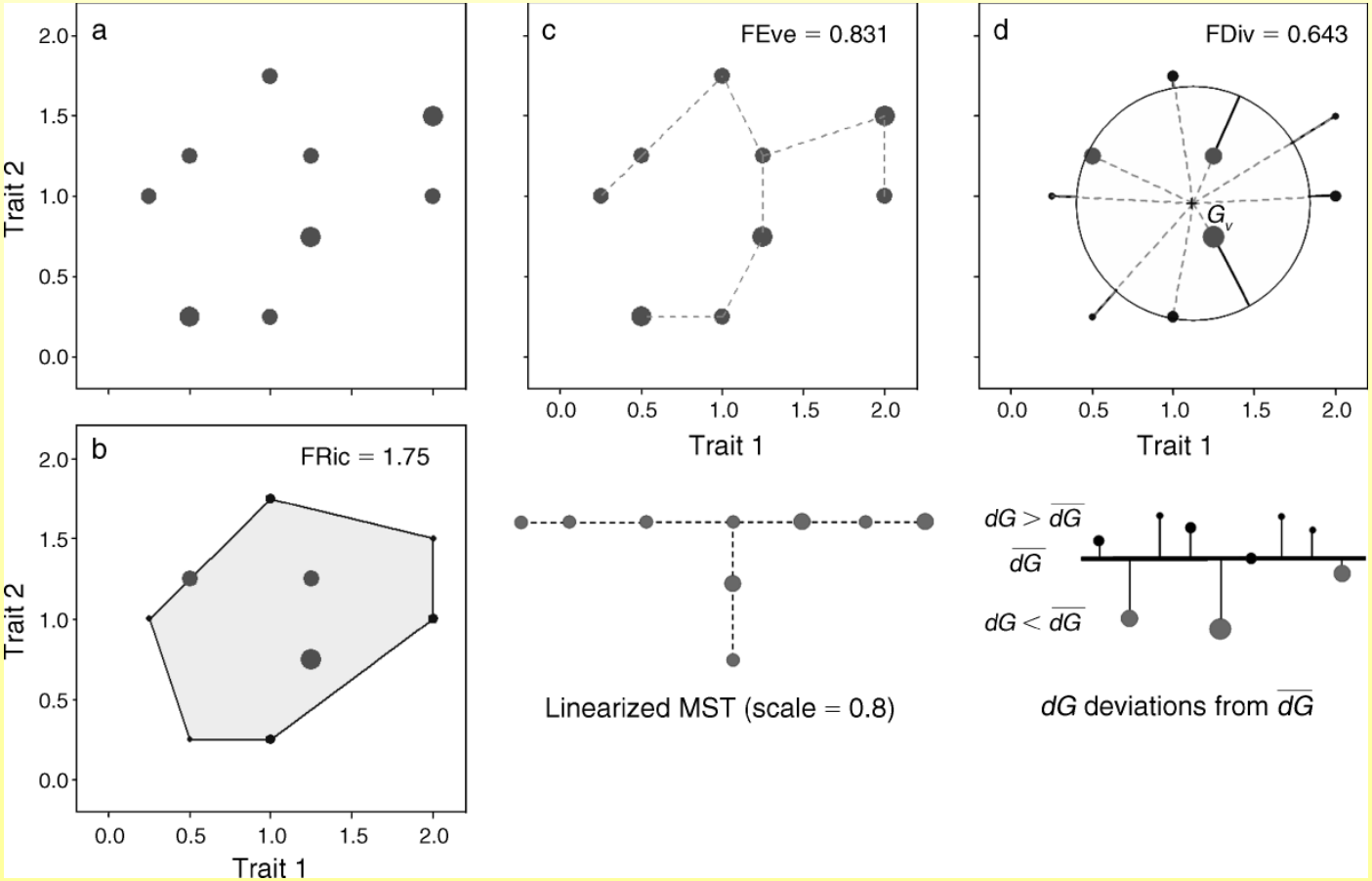


a) 9 species, two traits (principal coordinates analysis on Gower similarities)

b) **Functional richness:** Volume of convex hull around all species

Analysis of trait data

Alternative: View different facets of functional diversity (Villeger et al *Ecology* 2008)



c) **Functional evenness**: Regularity of species along minimum spanning tree connecting all species

d) **Functional divergence**: Mean distance to centroid

Analysis of trait data

Alternative: View different facets of functional diversity (Villeger et al *Ecology* 2008)

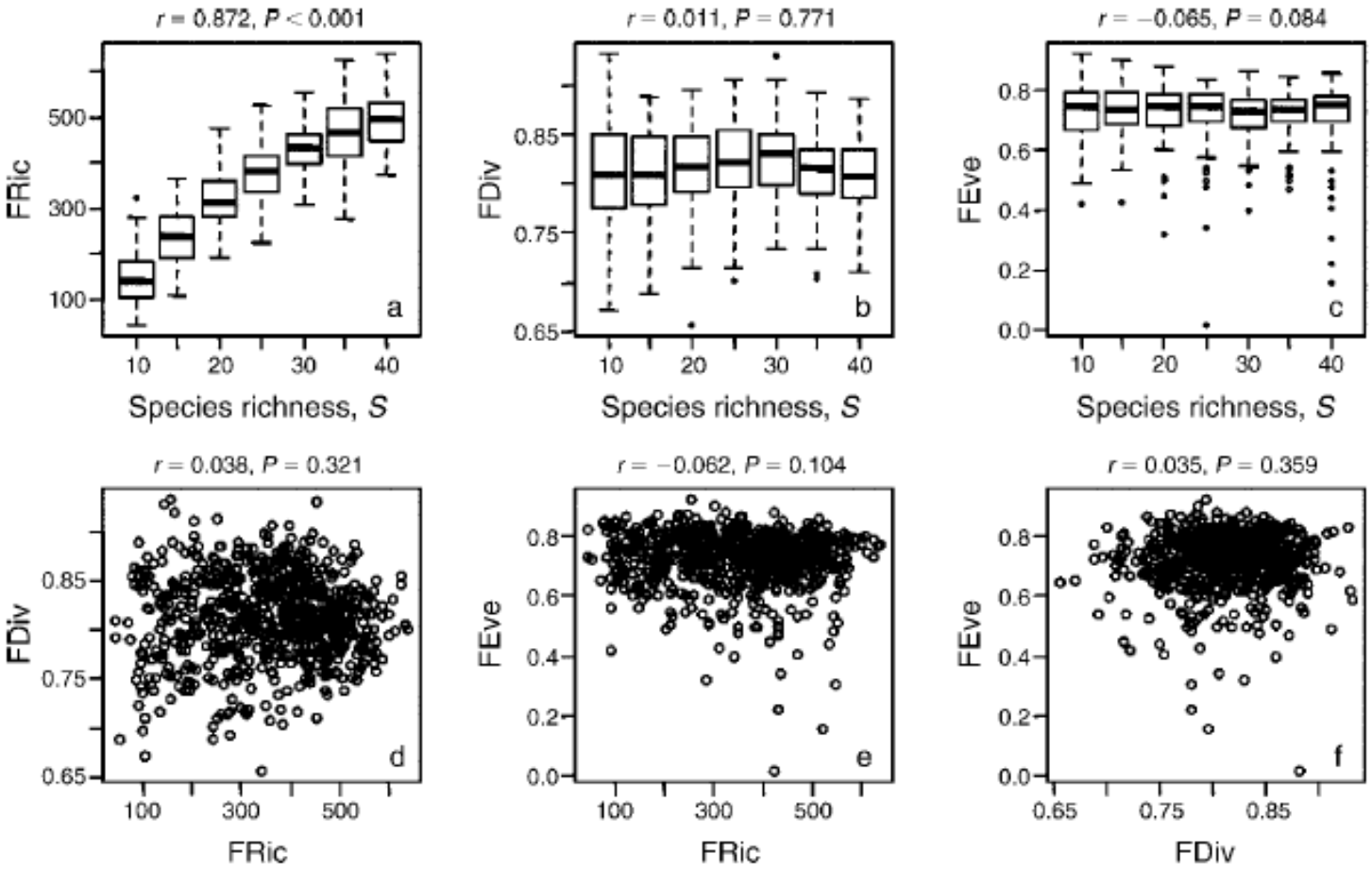
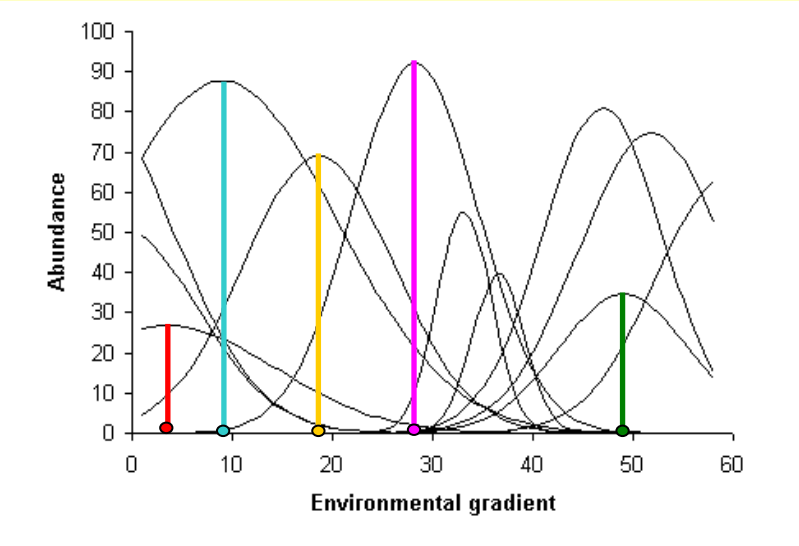


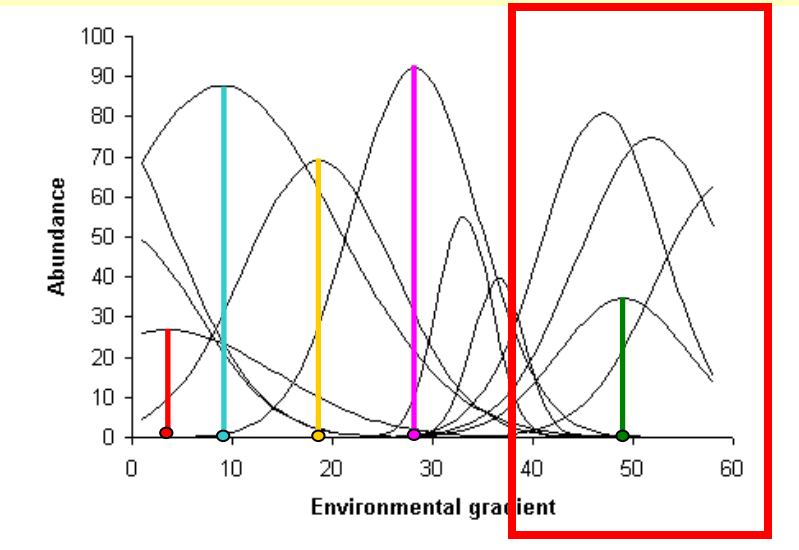
FIG. 4. Properties of the three functional diversity indices for artificial communities. Three traits were considered, and both the coordinates and the abundances of the species were generated under a uniform law (with respective range of 10 and 100). Seven species richness levels (S) were considered. Each species richness level was replicated 100 times. For each community, functional richness (FRic), functional divergence (FDiv), and functional evenness (FEve) were estimated. The first three panels (a, b, c) show the relations between each index and species richness. The three last panels (d, e, f) present the correlations between the three indices. Pearson's coefficients of correlation and levels of significance are given above the panels. FRic is the only index correlated to species richness. The three indices are independent of each other.

Classification

Classification

Ordinations / gradient analyses assume a **continuous** model. The abundance of species changes along environmental gradients.





Classifications search for **discontinuities / breaks** in data. There are no continuous gradients but **abrupt changes** that delimit groups of samples (e.g. communities).
These are groups of similarly responding species if seen from a gradient analysis perspective.

Classification

Phytosociology - most important ecological classification technique

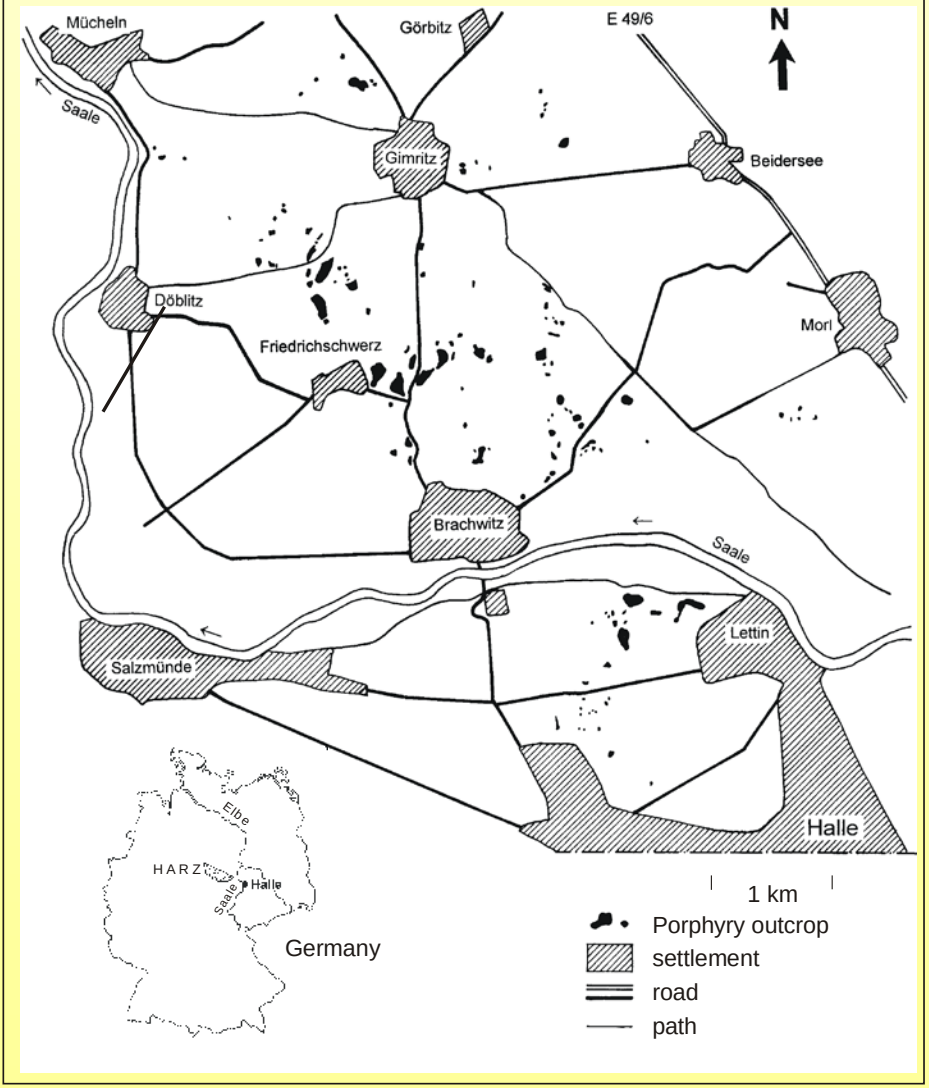
globally many 100 000 samples (= relevés) available.

The classification is (largely) **qualitative**, because the focus is on presence or absence rather than abundance of species in a given sample.

The classification is **agglomerative**, because samples are united into groups.

The classification is (in most cases) polythetic, because several species (and not only one) are used (monothetic: only one diagnostics species).

phytosociology: porphyry-outcrops near Halle



1. Galio-Agrostietum
2. Sisymbrio-Atriplicetum
3. Festuco-Brachypodietum
4. Euphorbio-Callunetum
5. Convolvulo-Agropyretum
6. Cardario-Agropyretum
7. Dauco-Arrhenatheretum
8. Falcario-Agropyretum
9. Filipendulo-Helictotrichetum
10. *Poa angustifolia*-community
11. Festuco-Stipetum
12. *Festuca rupicola*-community
13. Tanaceto-Arrhenatheretum
14. Thymo-Festucetum

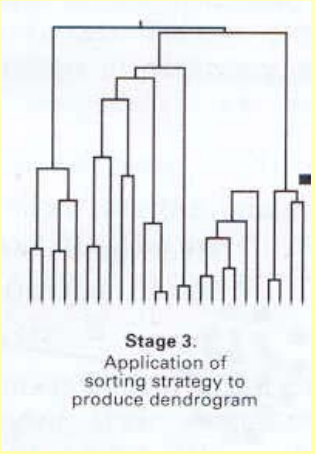
Wesche et al. *Folia Geobotanica* 2005

Classification

Main classes of methods

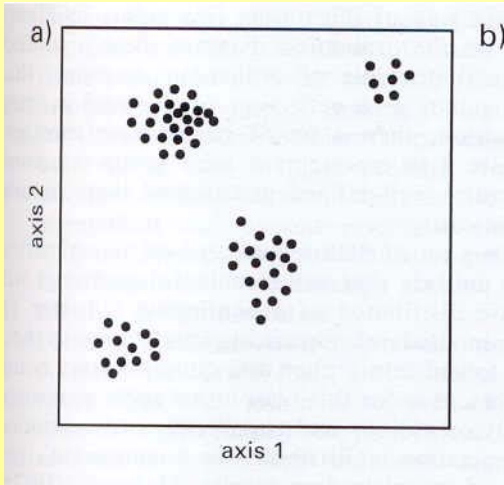
agglomerative methods:

- **Cluster-analysis:** Single-Linkage, UPGMA, Ward's Method
- **phytosociology**



Divisive methods:

- ***Ordination Space Partioning***
- **TWINSPAN:** Two Way Indicator Species Analysis
- (Discriminant analysis)



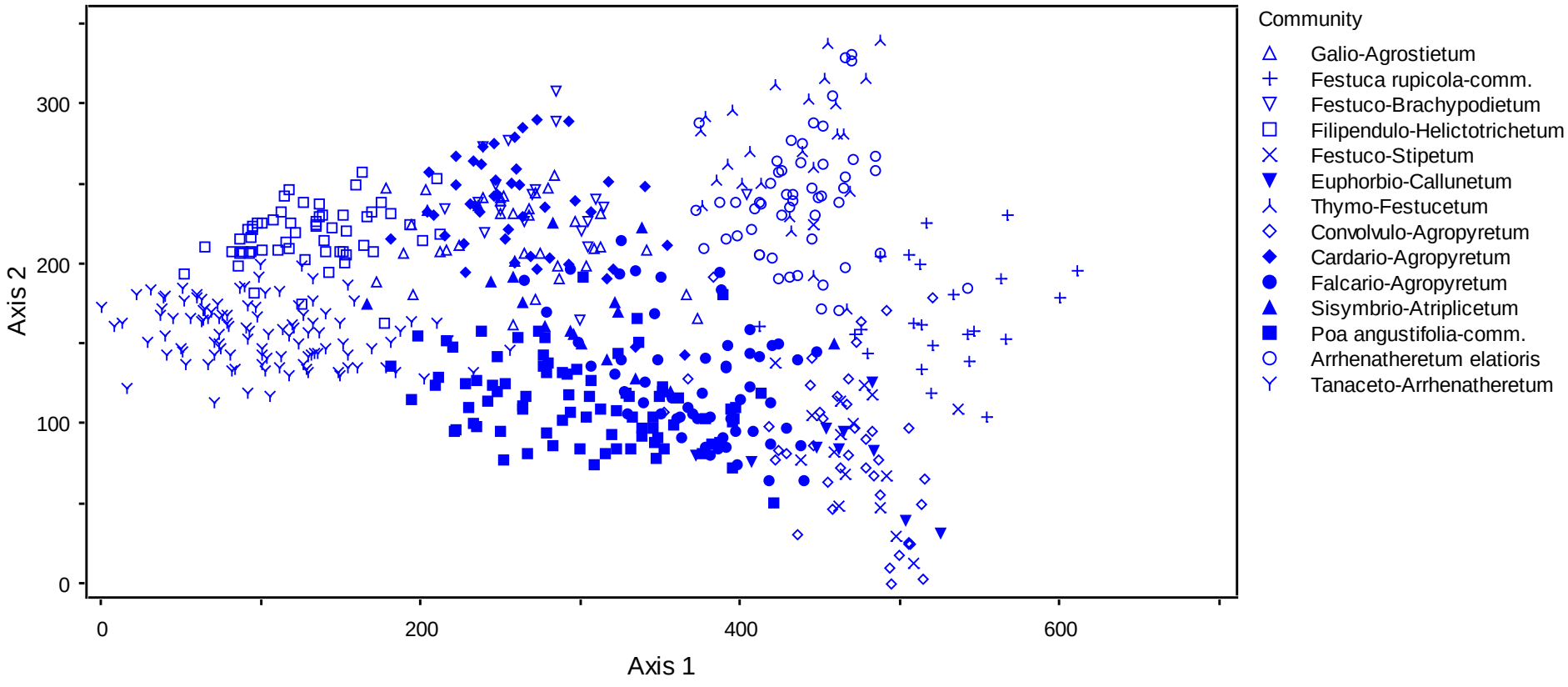
Non-hierarchical methods:

- *K-Means clustering*
- TABORD, COMPCLUS, FLEXCLUS

(relatively uncommon in ecology)

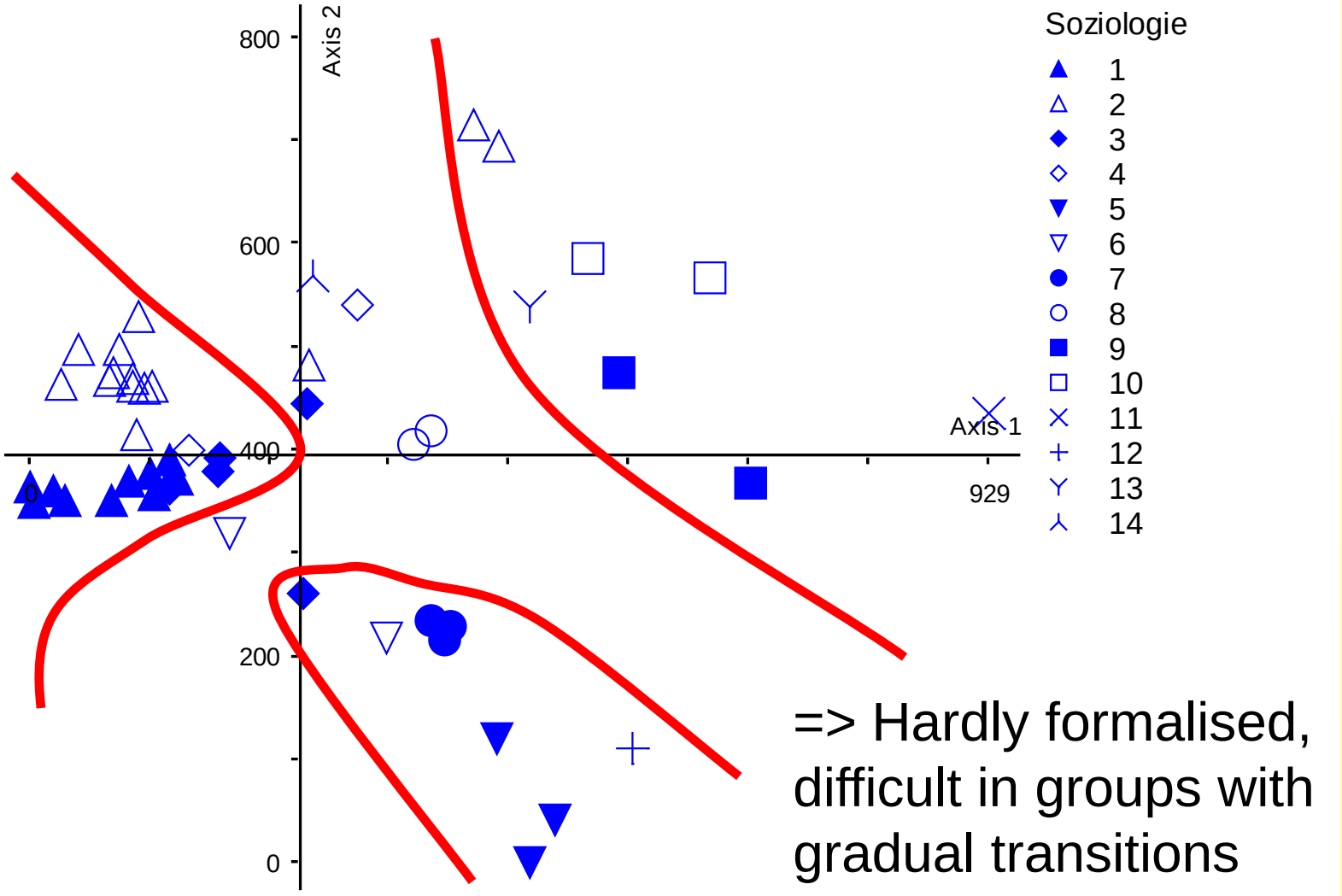
Phytosociology: porphyry-outcrops near Halle

DCA, species frequency >1, square root-transformation, downweighting of rare species, detrending by segments



Ordination space partitioning

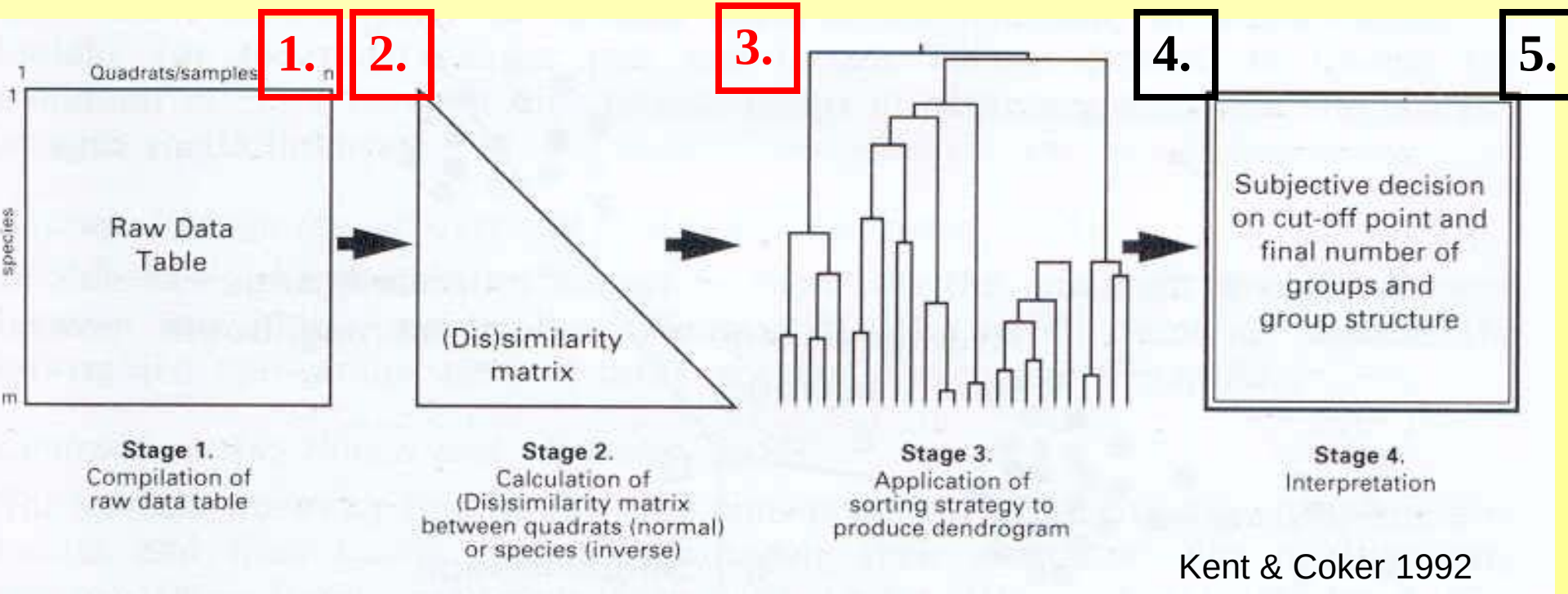
DCA with visual divisoning of ordination space



=> Hardly formalised, difficult in groups with gradual transitions

Cluster-analysis

Cluster-analysis: **Agglomerative** techniques that try to display similarity between samples in a **dendrogram**.
The first step is calculation of a **similarity** or **distance matrix**.
Multivariate similarities are than displayed in a two-dimensionale cluster-dendrogram.



Kent & Coker 1992

Cluster-analysis

1. step: data manipulations

transformations: Depending on aim (s. introduction). The most common applications are downweigthing of dominant species (logarithm-, square-root-transformation). Comparisons of real data have (surprisingly) shown that results are more similar to phytosociology if abundances are considered (rather than using 0/1 transformations).

Standardising: Usually not needed, unless variables were measured on different scales.

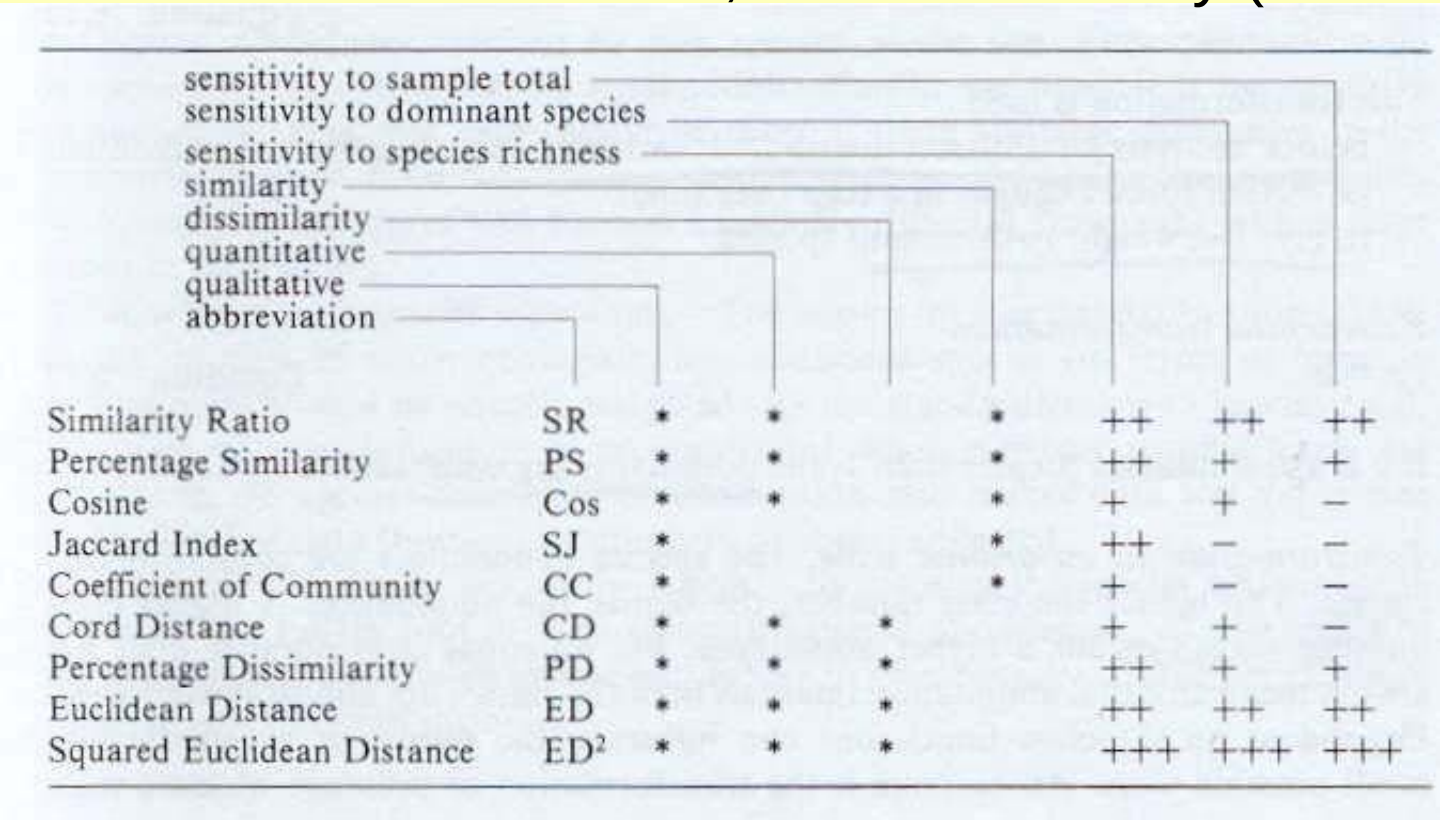
Cluster-analysis

2. step: similarity

There is a huge number of indices, which have own strengths and weaknesses (s.a. introduction). In ecology, **Steinhaus / Bray Curtis** / Sørensen similarity (=percentage similarity) is very important.

In systematics vor allem Nei's & Li's index, Gower Similarity (for a detailed Discusion s. Legendre & Legendre 1998, Seite 254ff).

Jongman et al. 1992



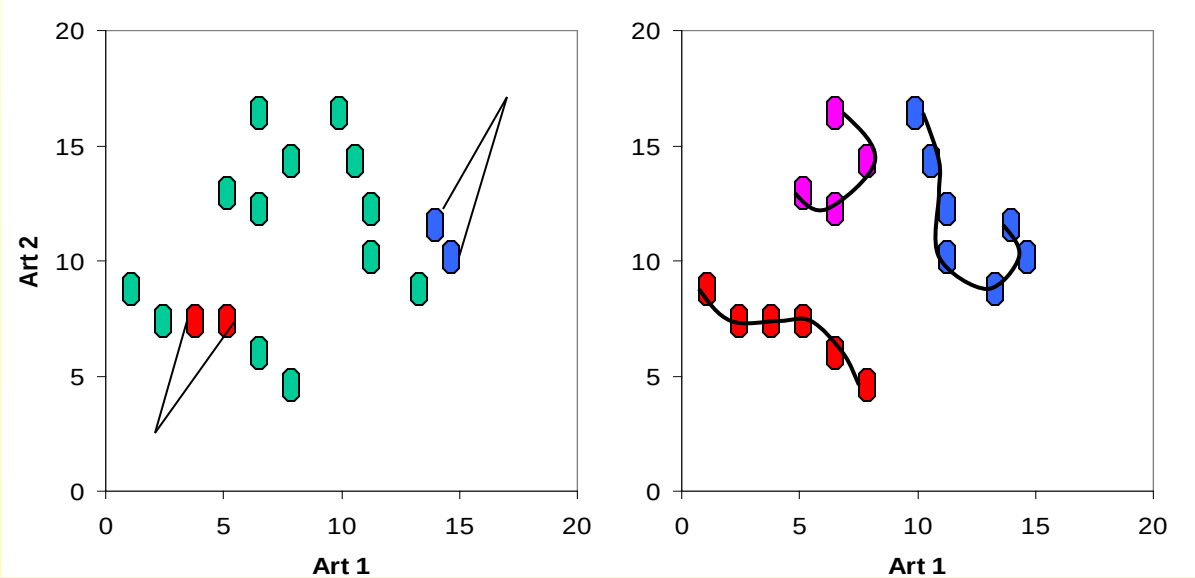
	sensitivity to sample total	sensitivity to dominant species	sensitivity to species richness	similarity	dissimilarity	quantitative	qualitative	abbreviation
Similarity Ratio	++	++	++	*		*	*	SR
Percentage Similarity	++	+	+	*		*	*	PS
Cosine	+	+	–	*		*	*	Cos
Jaccard Index	++	–	–	*		*		SJ
Coefficient of Community	+	–	–	*		*		CC
Cord Distance	+	+	–		*	*	*	CD
Percentage Dissimilarity	++	+	+		*	*	*	PD
Euclidean Distance	++	++	++		*	*	*	ED
Squared Euclidean Distance	+++	+++	+++		*	*	*	ED^2

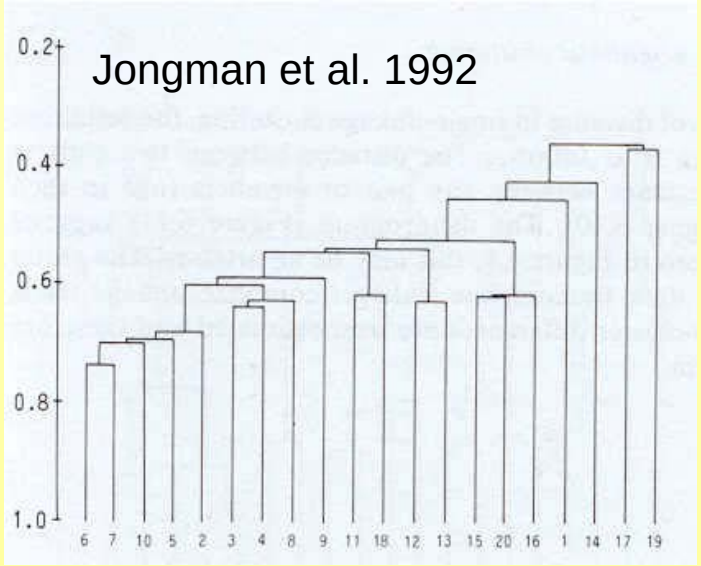
Cluster-analysis

3. step: algorithms

3a) *Single Linkage = Nearest Neighbour Clustering*

The most similar sample pairs are fused, followed by the next similar and so on.





Problem *chaining*: Unwanted chains are formed, which renders the method hardly useful. May be appropriate if large discontinuities are searched at early stage of analysis

Cluster-analysis

3. step: algorithms

3b) *Group Average* = usually calculated as the ***Unweighted Pair Groups Method***

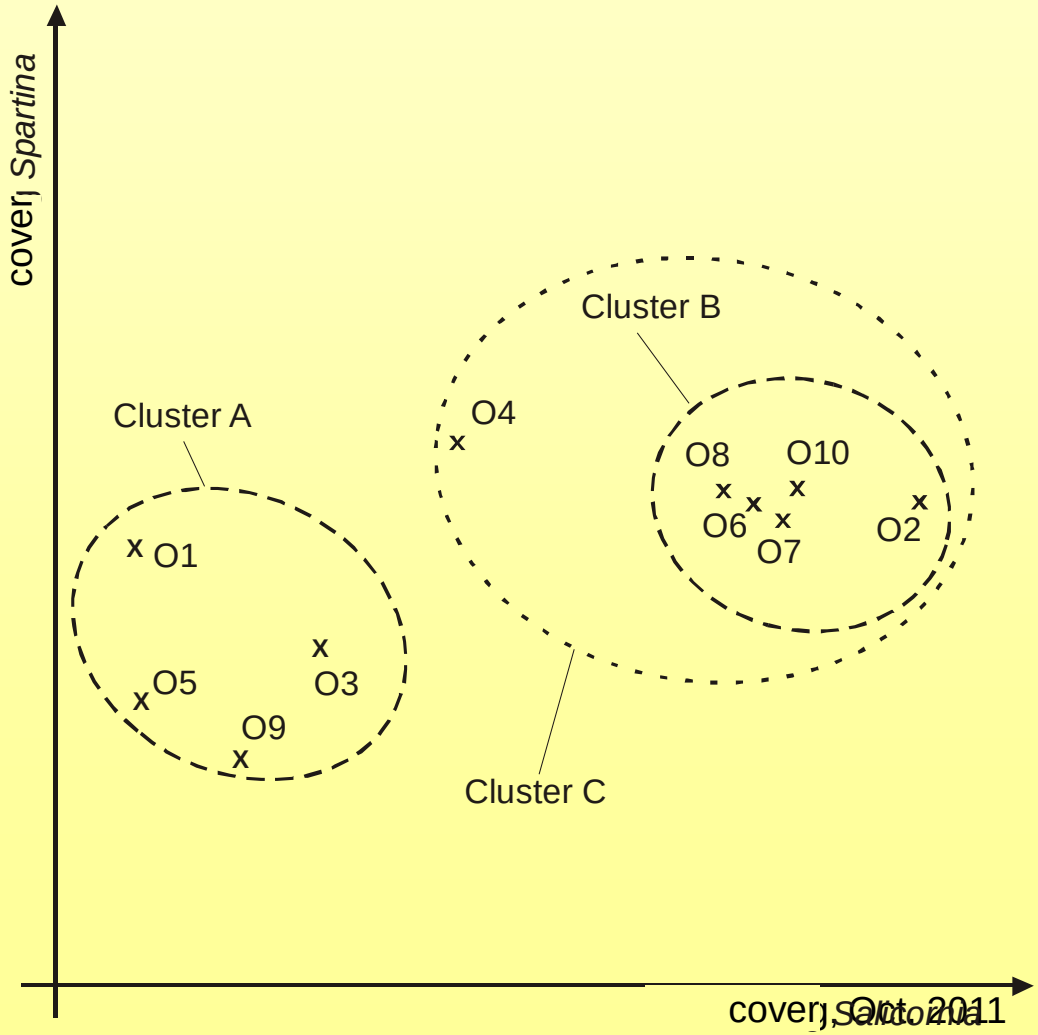
- Cluster are compared with respect to **mean dissimilarity**.
- This is based on dissimilarity between all elements of two clusters.
- Dissimilarity between two cluster is mean dissimilarity in all pair-wise comparison of samples.
- Samples / clusters are fused so that mean dissimilarity remains as low as possible (iterative).
- The dendrogram is scaled to display dissimilarity among groups as length of dendrogram branches.
- GA-Clustering is the method that displays the original dissimilarity best ("space-conserving").

Cluster-analysis

3. step: Algorithmen

3b) *Group Average* = usually calculated with the ***Unweighted Pair Groups Method***

- Clusters are fused according to their mean dissimilarity.
- This is the mean of all pairwise dissimilarities between all elements of the two clusters.



Cluster-analysis

Cluster-analysis 3b: Algorithmen GA / UPGMA

=>The most widely spread and recommended method in ecology (for non-ordinal data)

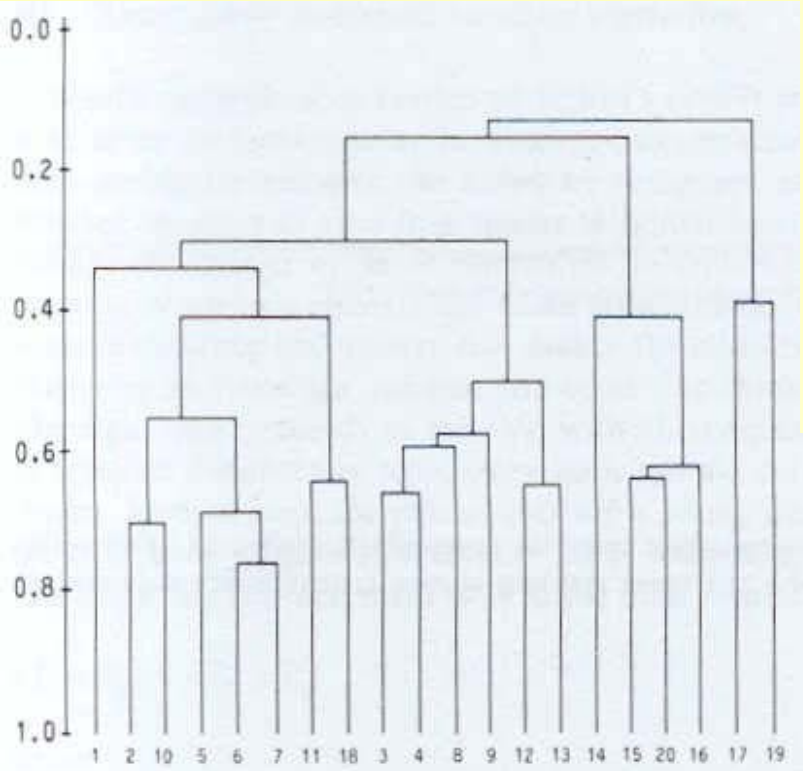


Figure 6.12 Average-linkage dendrogram of the Dune Meadow Data using the similarity ratio.

(There are modifications in the way mean similarity is calculated, UPGMA is the most common)

Jongman et al. 1995

Cluster-analysis

Cluster-analysis 3: algorithms

3c) Ward's Method: Minimum Variance Clustering

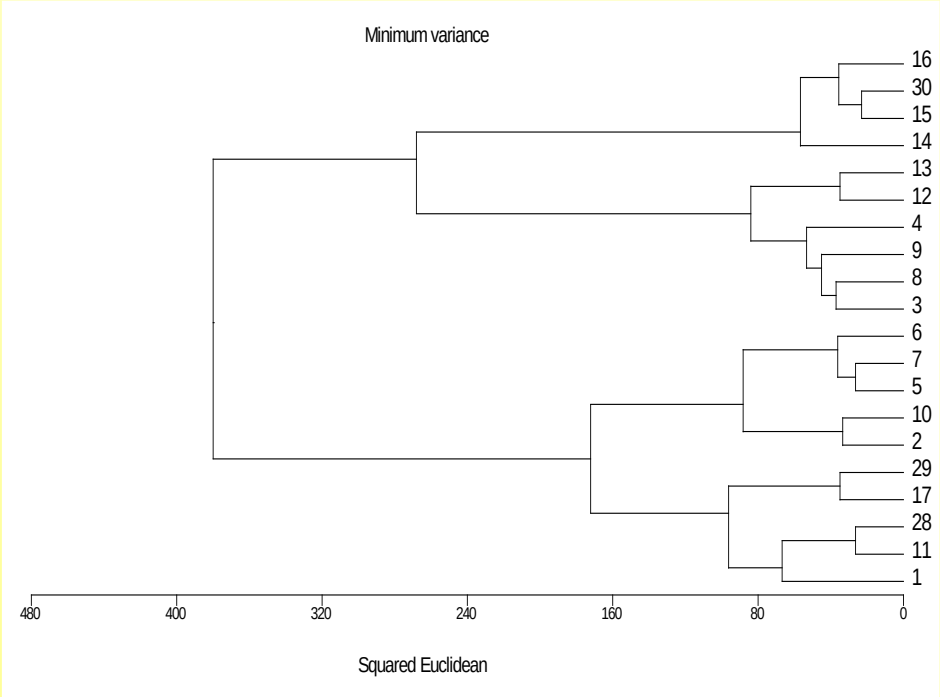
Iterative as Group Average

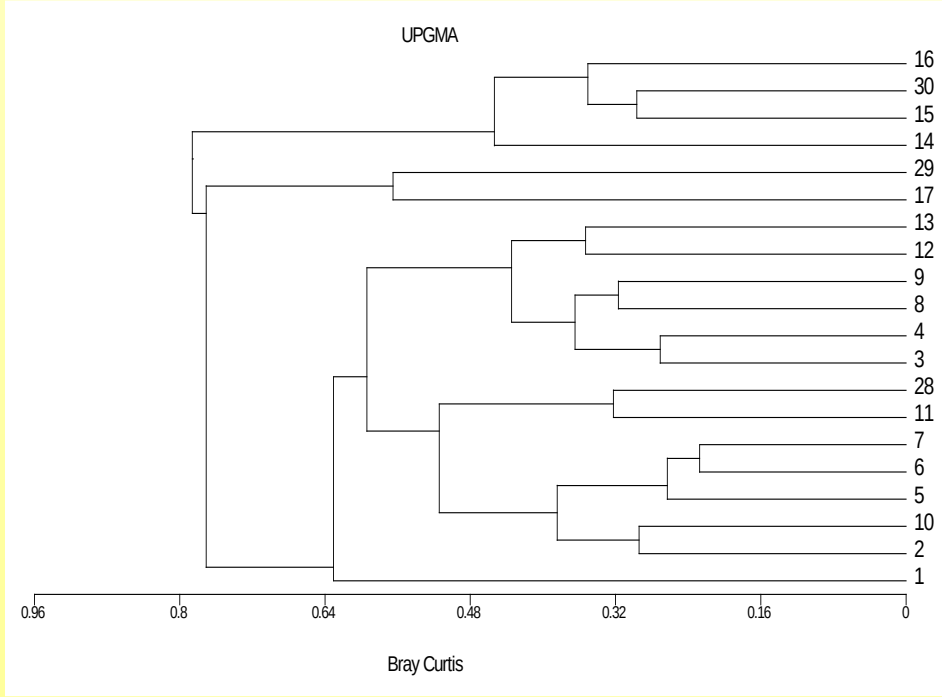
- Criterion is **variance of dissimilarities** instead of mean dissimilarity (variance is the mean squared distance of elements in a cluster to the cluster centroid, often calculated as the mean **squared Euclidean** distance between elements).
- Elements / clusters are fused with the aim to keep increase of variance as small as possible
- The squaring of distances overemphasises larger distances in the dendrogram
- Assumption: multivariate normal distributions, multivariate variance homogeneity, random sampling.

=> **Stritly speaking not suitable for most data sets!**

Cluster-analysis

Cluster-analysis 3: Algorithmen
3c) Wards method: Minimum Variance Clustering vs. UPGMA (example Dune Meadow Data)





Cluster-analysis

comparison of methods:

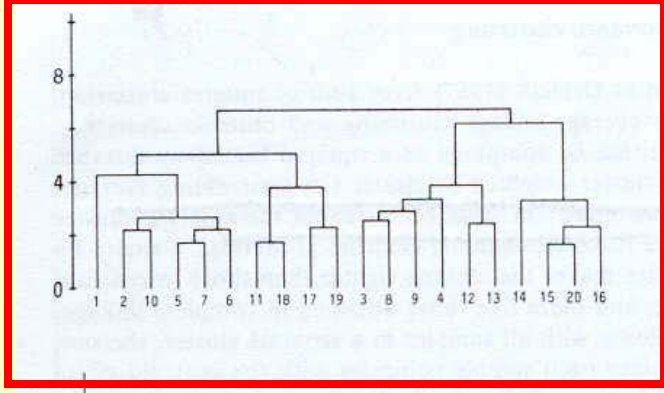
Average Linkage

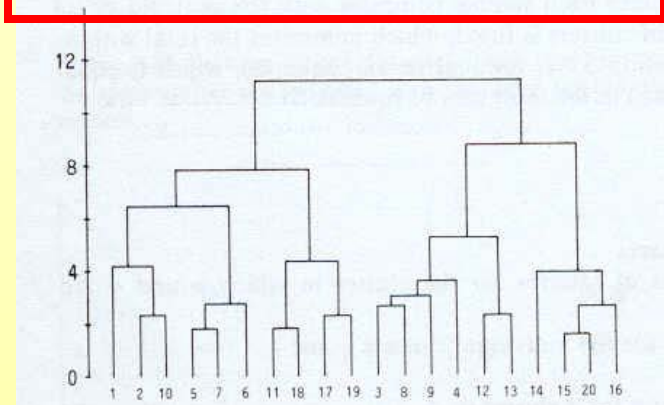
(*Complete Linkage*)

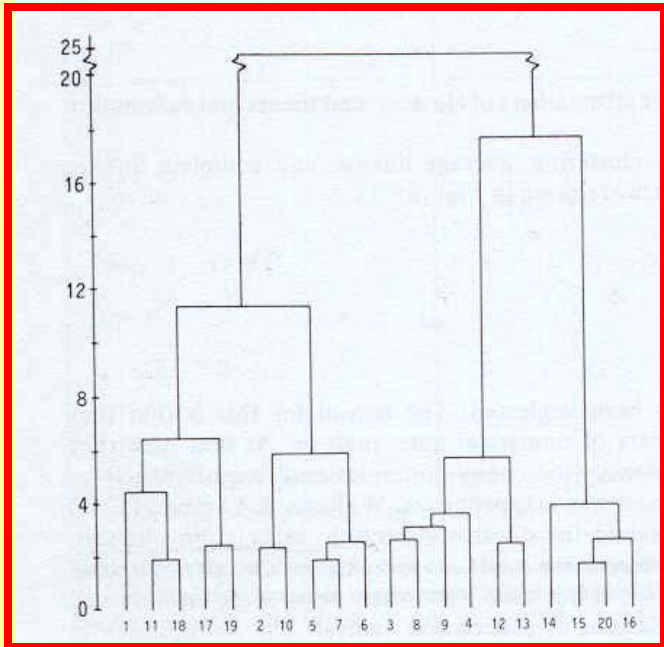
Ward's Method

(Dune Meadow Data, Euclidean distance)

Jongman et al. 1995







Problems:

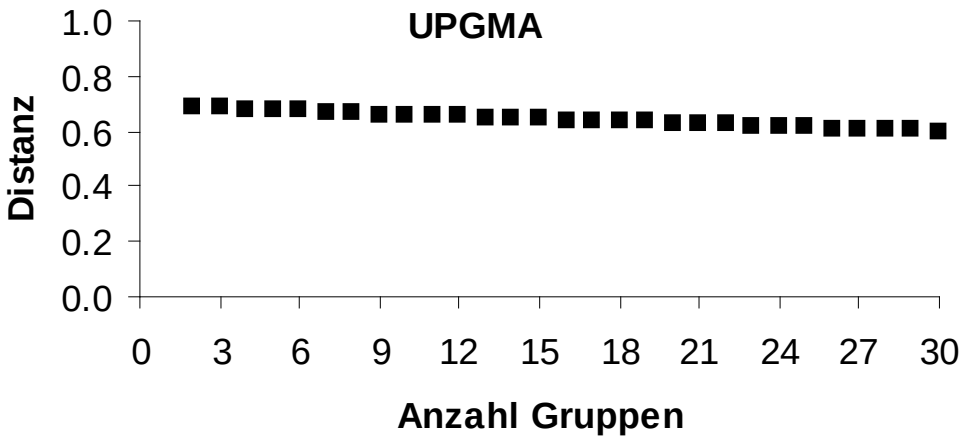
- **modifying** transformations, similarities or fusion-algorithms allows to calculate almost unlimited numbers of distinct cluster dendrograms.
- Choice of **cut levels** and therefore number of clusters is rather subjective.
- **There is no such thing as an objectively correct classification!**

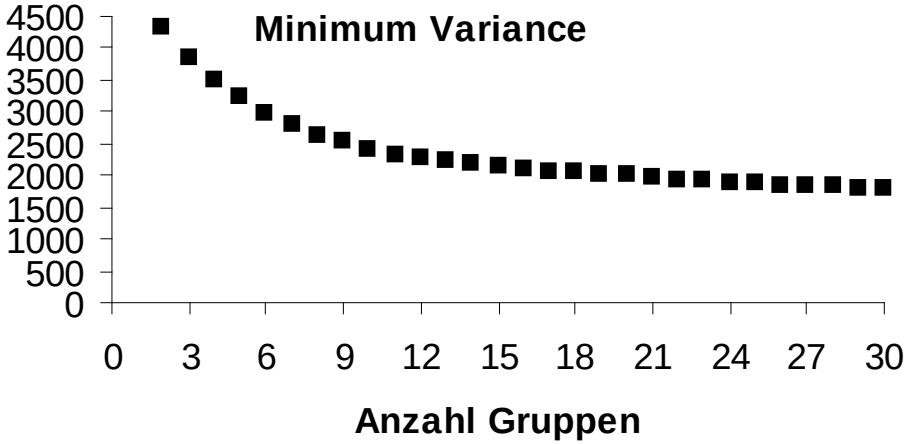
Cluster-analysis

Cluster-analysis 4: pruning dendrograms

Mathematical techniques from univariate statistics: permutation test, chi-square; parametric techniques often not suitable(s. Jongman et al. 1995).

Graphical techniques: Plotting number of clusters against fusion levels. Search graph for breaks.



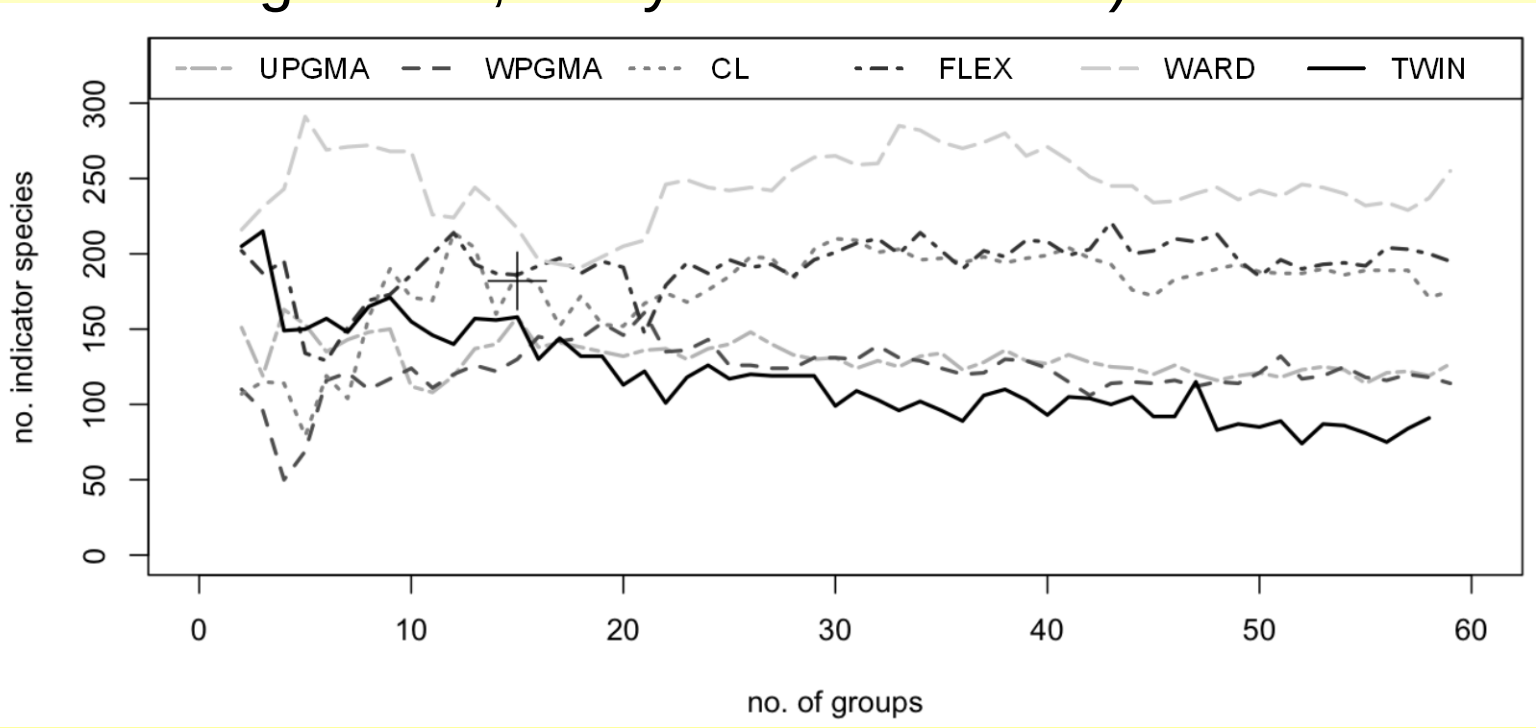


Heuristic techniques (visual inspection)

Cluster-analysis

Cluster-analysis 4: pruning dendrograms

Calculate number of indicator species (Dufrene & Legendre Ecol. Monogr. 1997; Tichy et al JVS 2010)



Number of significant indicator species for dendrograms pruned at various levels. 6 cluster algorithms applied to samples from Mongolian deserts (Wesche & von Wehrden AVS 2011)

Cluster-analysis

Cluster-analysis 5: validation

correlation of distances in dendrogram with original distances (cophenetic correlation).

comparison with external structures: cross-tabulation with phytosociological classification

example Wesche et al. 2005

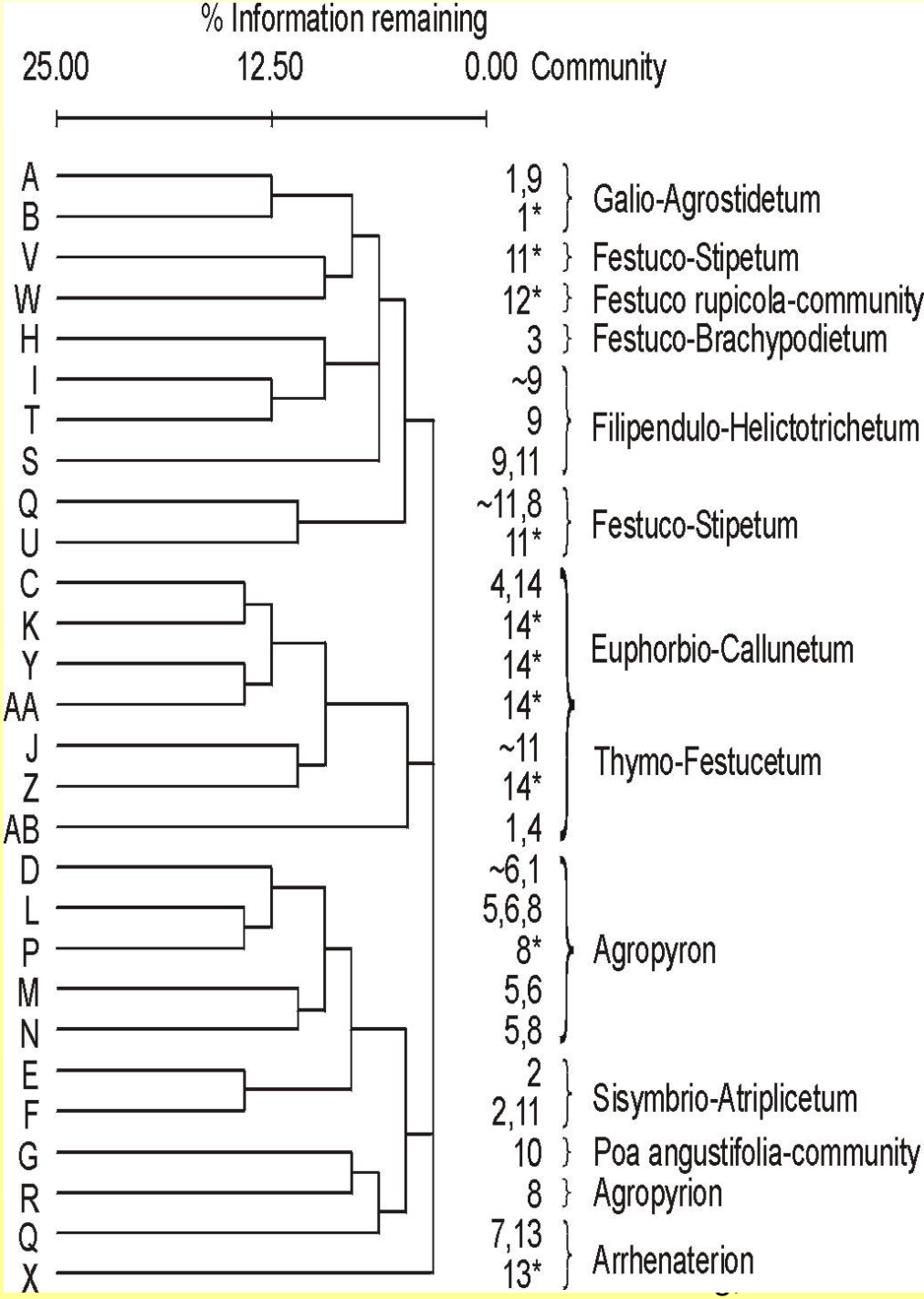
Association \ Cluster ID	1	2	26	29	36	41	47	58	74	81	83	132	133	134	148	193	197	209	230	232	244
1. Galio-Agrostidetum	14	73	2.7	5.4																	
2. Sisymbrio-Atriplicetum					91	4.3	4.3														
3. Festuco-Brachypodietum								89	11												
4. Euphorbio-Callunetum	3.4		88							1.7	1.7										
5. Convolvulo-Agropyretum							5.9					76	5.9	5.9	5.9						
6. Cardario-Agropyretum				10								80	10								
7. Dauco-Arrhenatheretum									4						96						
8. Falcario-Agropyretum	2.4				4.9							78		2.4	2.4	4.9	2.4	2.4			
9. Filipendulo-Helictotrichetu	20	4.5							59										4.5	9.1	2.3
10. *Poa angustifolia*-comm	3.4			1.7			86		6.9						1.7						
11. Festuco-Stipetum						6.3				6.3							6.3		6.3		56
12. *Festuco rupicola*-comm	1.2						1.2	1.2		1.2		3.6						1.2			
13. Tanaceto-Arrhenatheretum												5.7	1.9		91						
14. Thymo-Festucetum	1.8		7.3							0.9	82										
N	22	29	60	4	23	2	53	18	33	4	91	59	3	2	75	2	2	2	3	4	10

Cluster-analysis

Cluster-analysis 5: validation comparison with external

Struktur:

example Porphyry outcrops near Halle



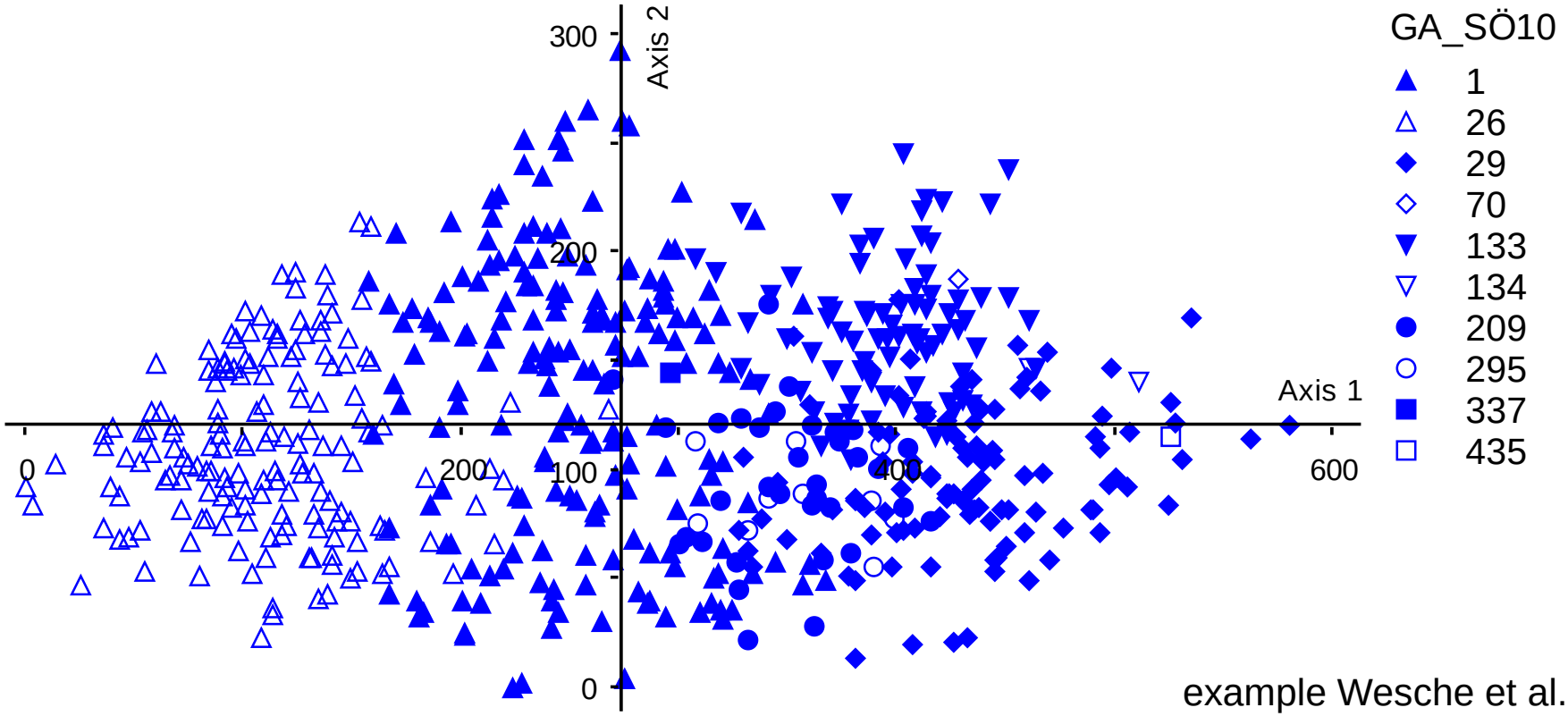
example Wesche et al. 2005

Cluster-analysis

Cluster-analysis 5: validation

Plot on ordination

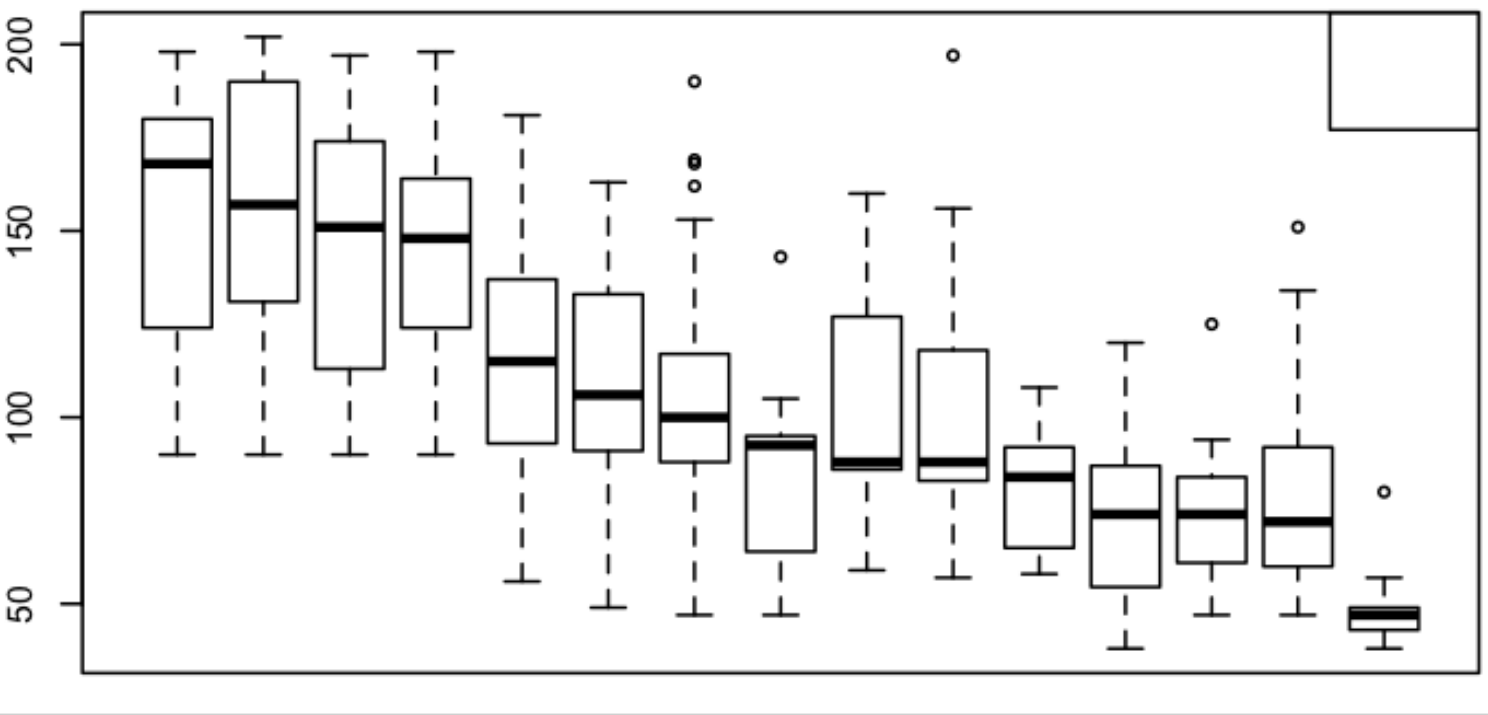
example Porphyry outcrops near Halle



example Wesche et al. 2005

Cluster-analysis

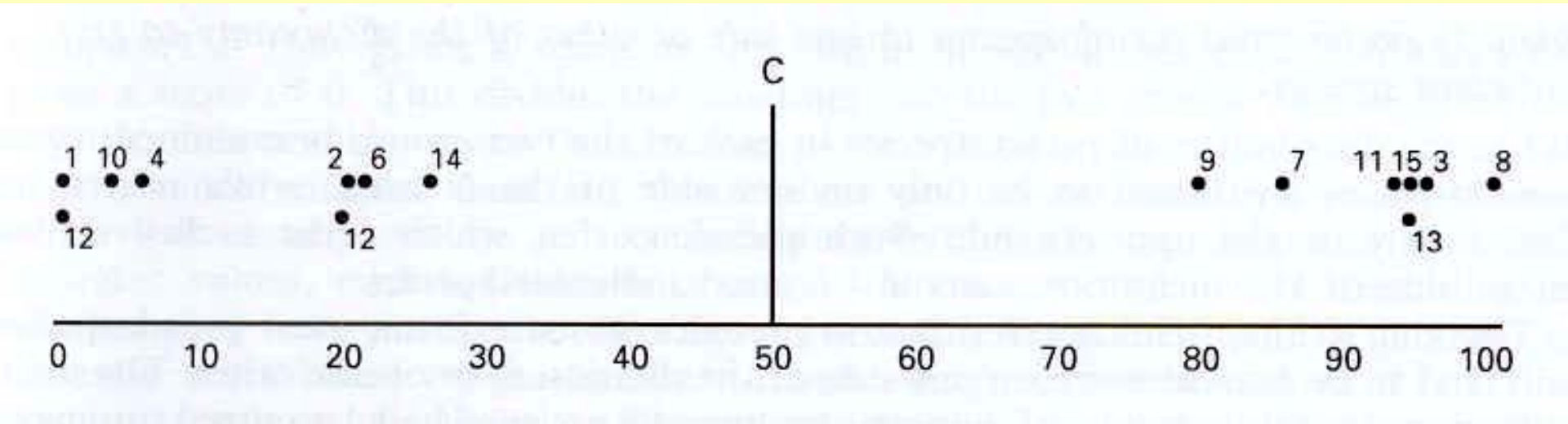
Cluster-analysis 5: validation
Comparison with environmental data



(Modelled) precipitation data for clusters of vegetation samples from Mongolian deserts (Wesche & von Wehrden AVS 2011) example Wesche & von Wehrden 2011

TWINSPAN

***TwoWayINdicatorSPeciesANalysis*:**
Sucessive ordination space partitioning based on coordinates along first axis of a correspondence analysis („primary ordination“).



Terminology: left is **negative group**, right is **positive.**
Species are searched that are characteristic for one of the groups (CA ordinates species and samples!).

Kent & Coker 1992

TWINSPAN

Indicator Values: Indicate whether a species is confined to one of the groups:
(nj = occurrence of species in a given group)
(n = number of all samples in a group)

$$I_j = \frac{n_j^+}{n_+} - \frac{n_j^-}{n_-}$$

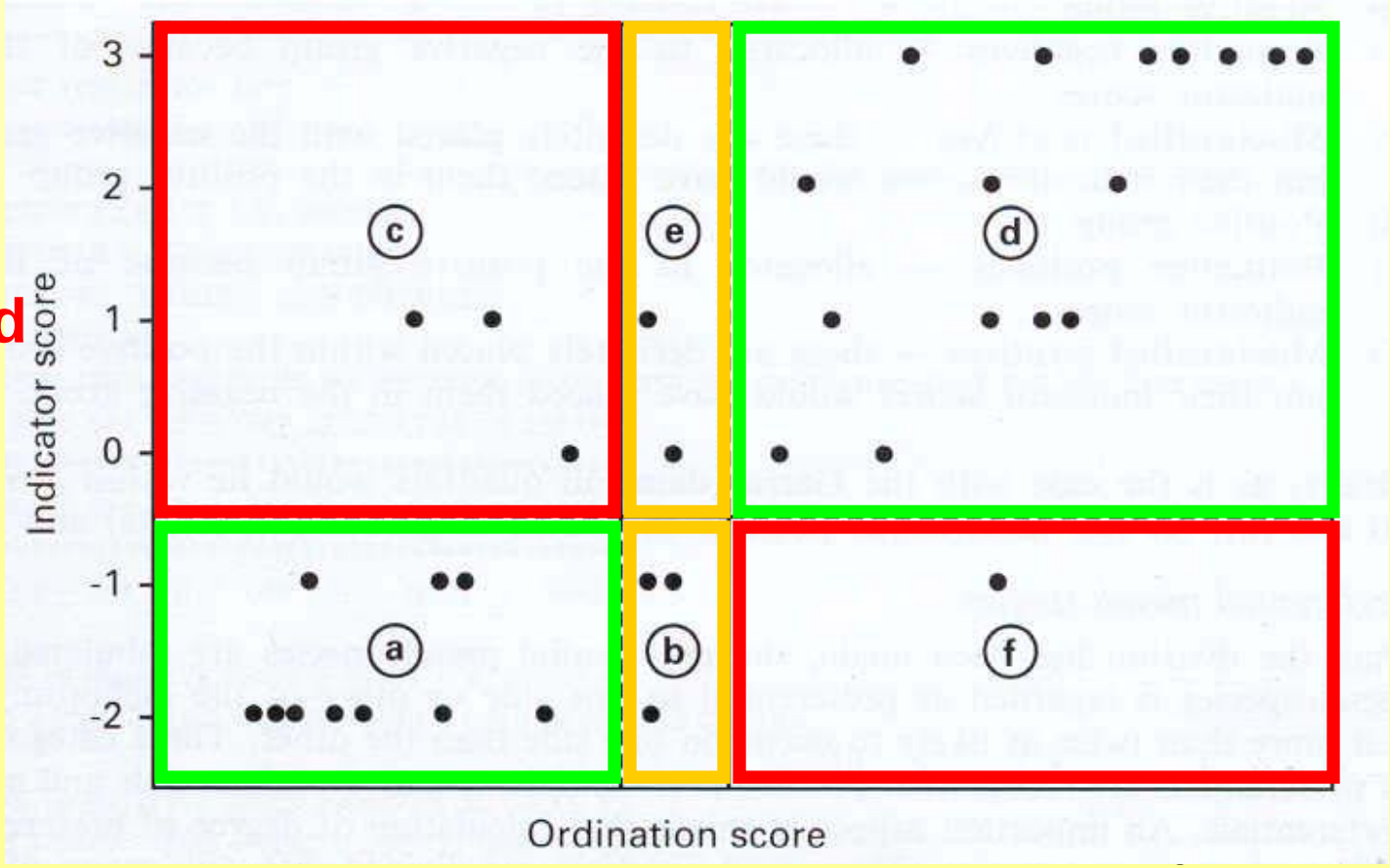
Those species with the most extreme indicator value are selected, and for each sample the sum of indicator values of indicator species is calculated ("indicator score“).
Samples are plotted along new axis based on indicator scores ("refined ordination" – 1-dimensional).
Usually up to 5 indicator species are used to safe calculation time.

New coordinates are compared with original ordination and a threshold for the sum of indicator values is fixed.
Samples above come in positive group, samples below in negative group.

Kent & Coker 1992

TWINSPAN

Comparison CA / *primary ordination* vs. *refined ordination*:
Possibilities: 1) CA values and indicator values give similar picture: **a, d** => reliable classification
2) sample scores are near threshold: **b, e** => "**borderliners**"
3) Values in both ordinations give contradictory patterns: **c, f**: =>"**misclassified**"



Kent & Coker 1992

TWINSPAN

Further iterations: ordinations and divisions are repeated for each respective subgroup until fixed level of resolution has been reached.

Abundance: TWINSPAN uses ordinal abundance values by generating for each species a number of ***pseudospecies*** that corresponds to the number of different abundance values (Dummy-variables). The ordinations are performed with these pseudospecies.

Example: a species with 3 cover values (beyond 0) is replaced by 3 pseudospecies.

	Cover			Pseudo-species					
	Galium verum	*Poa trivialis*		GV1	GV 2	GV 3	GV 4	PT1	PTr 2
Sample 6	4	1		1	1	1	1	1	0
Sample 8	0	2		0	0	0	0	1	1
Sample 25	3	1		1	1	1	0	1	0

TWINSPAN

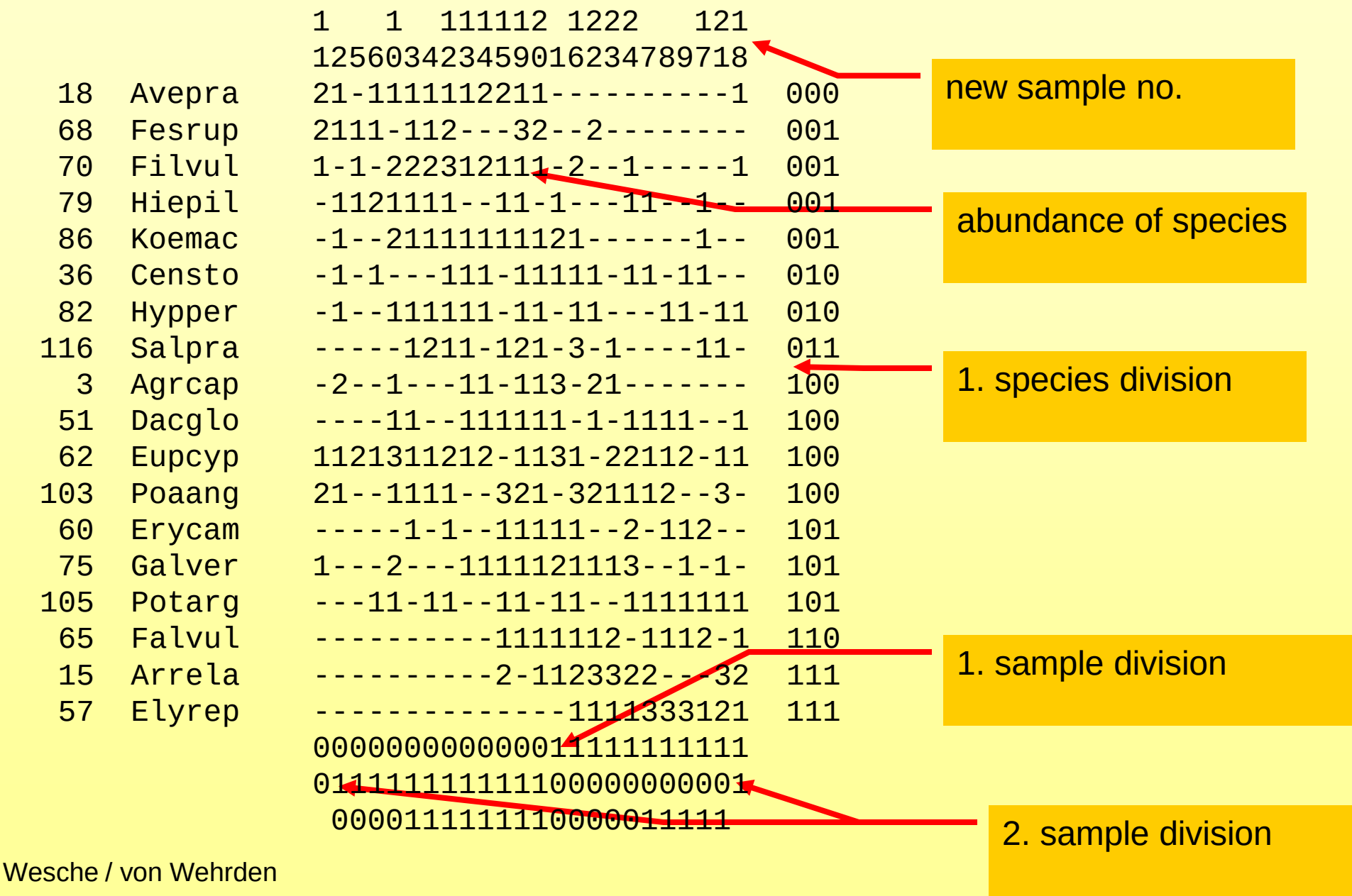
Further iterations: ordinations and divisions are repeated for each respecitve subgroup until fixed level of resolution has been reached

Abundance: TWINSPAN uses ordinal abundance values by generating for each species a number of ***pseudospecies*** that corresponds to the number of different abundance values (Dummy-variables). The ordinations are performed with these pseudospecies.

Example: a species with 3 cover values (beyond 0) is replaced by 3 pseudospecies.

Indicator species is always at most one pseudospecies per taxonomical unit (with highest *indicator value*).

TWINSPAN


1 1 111112 1222 121
125603423459016234789718
18 Avepra 21-1111112211----------1 000
68 Fesrup 2111-112---32--2-------- 001
70 Filvul 1-1-222312111-2--1-----1 001
79 Hiepil -1121111--11-1---11--1-- 001
86 Koemac -1--21111111121------1-- 001
36 Censto -1-1---111-11111-11-11-- 010
82 Hypper -1--111111-11-11---11-11 010
116 Salpra -----1211-121-3-1----11- 011
3 Agrcap -2--1---11-113-21------- 100
51 Dacglo ----11--111111-1-1111--1 100
62 Eupcyp 1121311212-1131-22112-11 100
103 Poaang 21--1111--321-321112--3- 100
60 Erycam -----1-1--11111--2-112-- 101
75 Galver 1---2---1111121113--1-1- 101
105 Potarg ---11-11--11-11--1111111 101
65 Falvul ----------1111112-1112-1 110
15 Arrela ----------2-1123322---32 111
57 Elyrep ------------1111333121 111
0000000000001111111111
0111111111110000000001
0000111111110000011111
new sample no.
abundance of species
1. species division
1. sample division
2. sample division

TWINSPAN

Advantages
The results broadly resemble phytosociological classifications.

TWINSPAN has therefore become very widespread in vegetation science and possible (still) is the most widely used classification technique there.

Kent & Coker 1992

TWINSPAN

TWINSPAN – problems

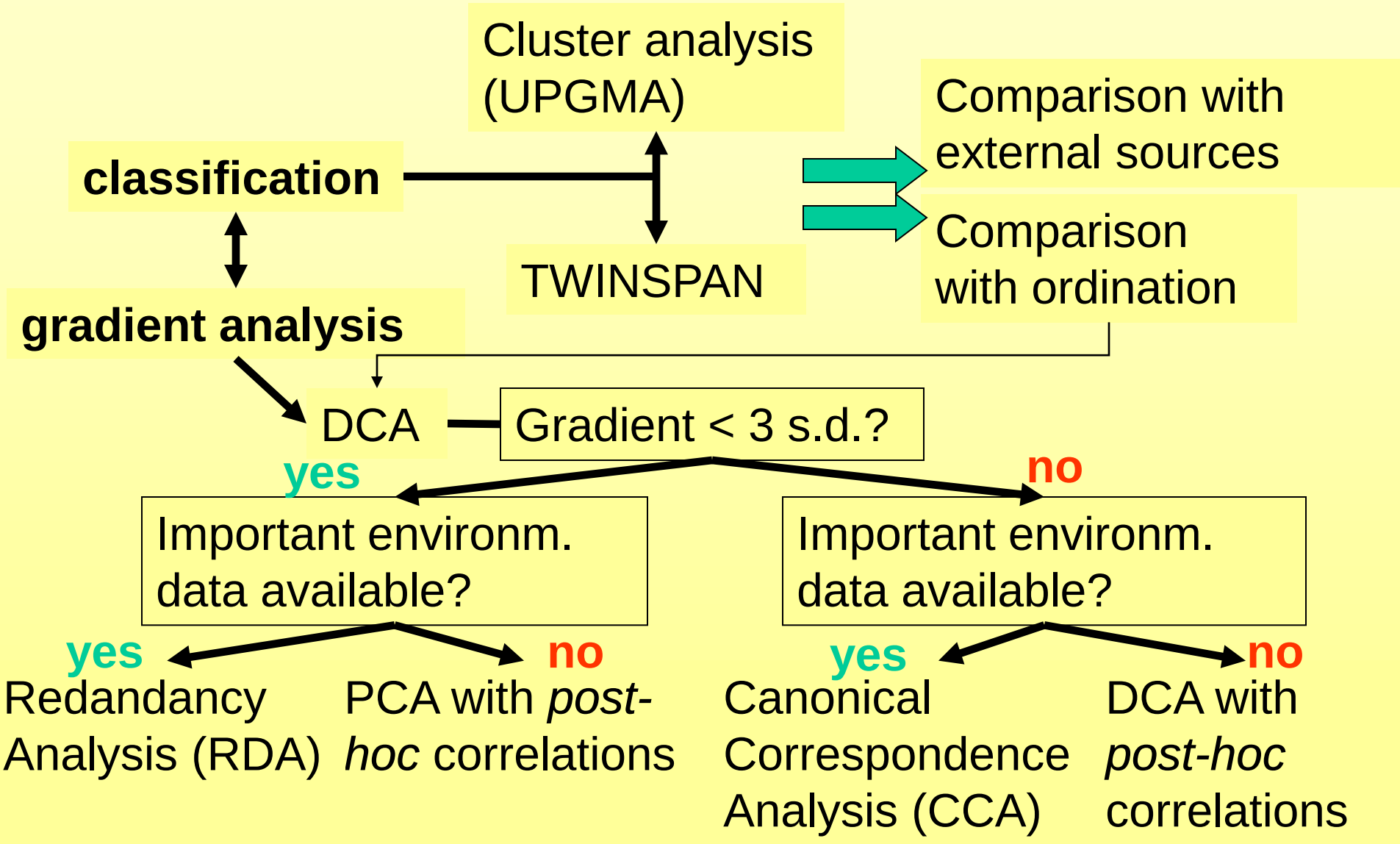
Problems inherited from CA:

- only suitable for data with strong gradients along first axis (and comparatively unimportant second axies).
- Length of gradient decreases with each interation, the unimodal model becomes questionable (see Roleček et al. 2009 for solution).
- Threshold may be unreliable if axis is strongly contorted.
- Rare species are always *downweighted*.

Additionally:

- Complex algorithm, hard to follow in detail.
- Relatively unflexible, hard to assess in details (s.a. van Grounewoud, 1992)
- The finally table is not easily translated into cluster dendrogram (but see JUICE, Tichý 2005).

Analysis of Community data



Synthesis

Complementary analysis - combination of classification and ordination: e.g. superpose classifications on ordination scatter plot

multivariate methods **reduce dimensionality** and facilitate **data exploration**. This should be followed by generation of hypotheses that could be tested.

Tests are often performed with **univariate statistics**, **permutations tests** for multivariate data mediate (Monte Carlo Tests in CCA, Mantel Tests)