

Finding Groups

➤ *Objectives:*

Discuss the nature of community composition data

Introduce community analysis with Multi-Variate Statistics

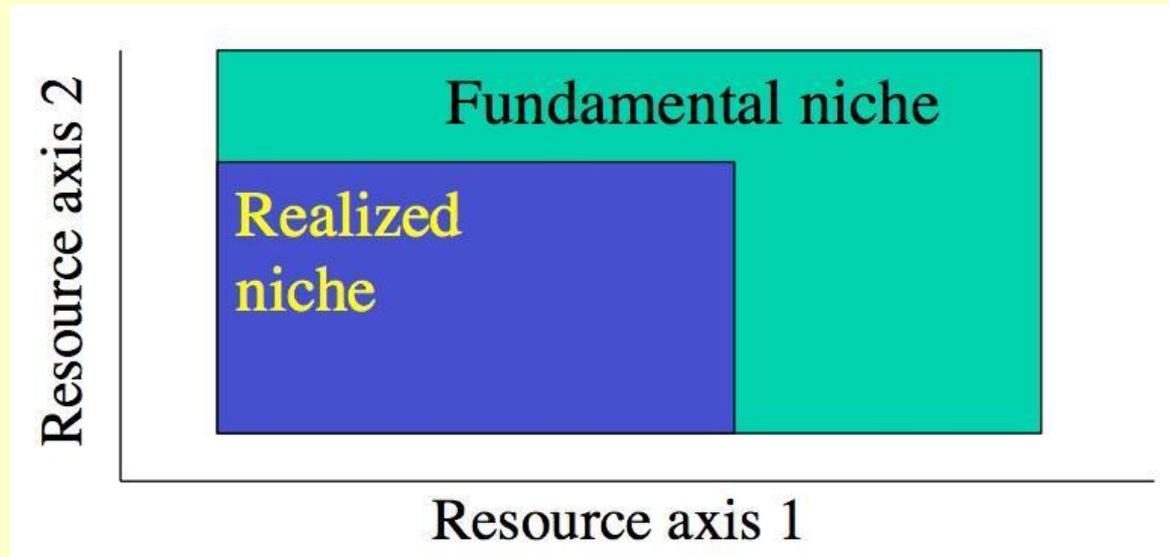
Community Analysis - Introduction

- Community data (species counts in time / space) are multi-variate because each sample unit characterized by:
 - the occurrence (presence / absence) or abundance (counts) of number of co-occurring species
 - a set of (cross-correlated) environmental factors affecting species distributions
 - a set of temporal (e.g., absolute time, relative time) and spatial attributes (e.g., lat / long, habitat type)

(James & McCulloch 1990)

Community Analysis - Introduction

- Community ecologists analyze effects of multiple environmental factors on large numbers of co-occurring species and deal with statistical errors (measurement / structural)
- Why are species data not independent ?

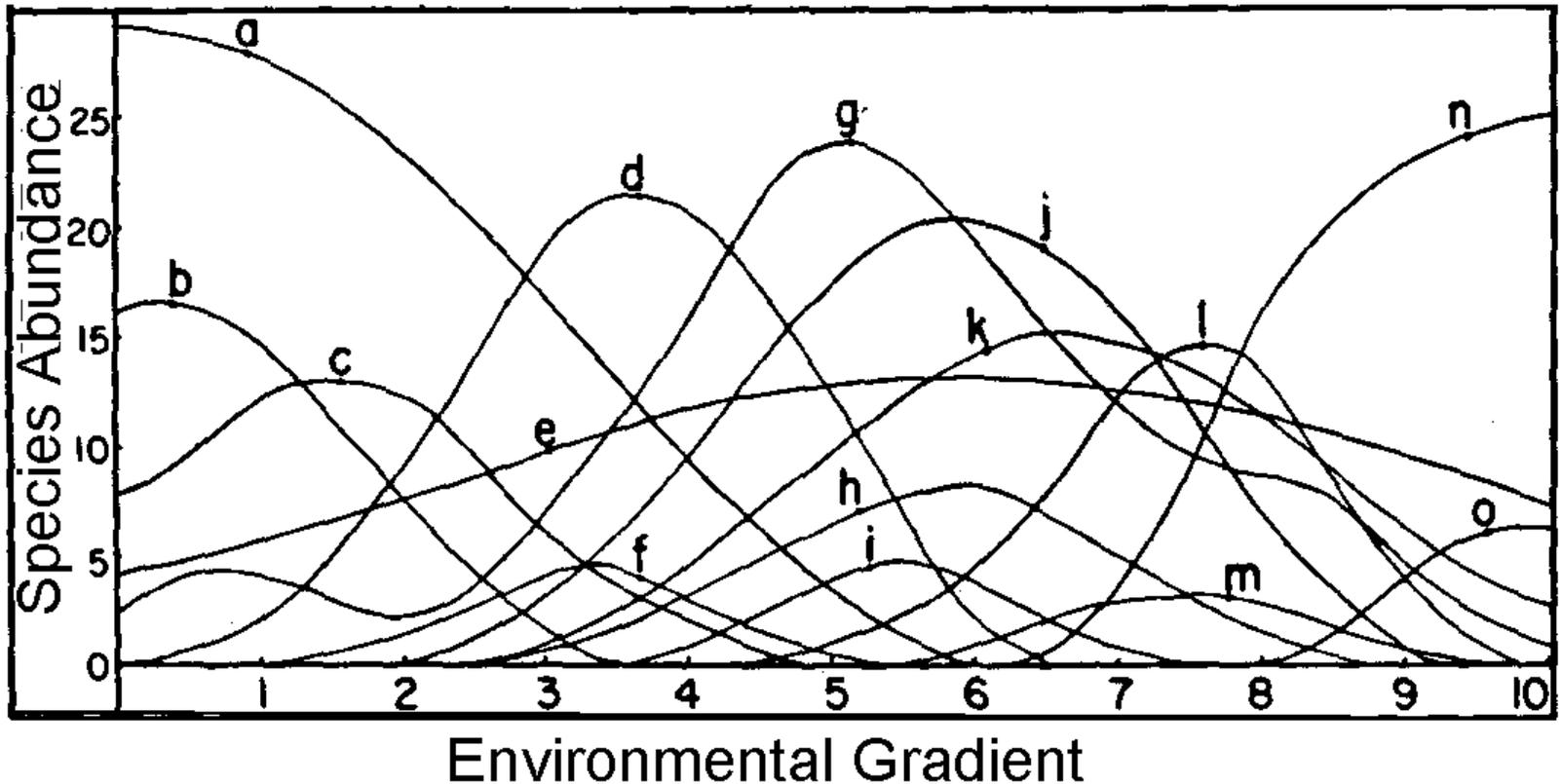


Species on Environmental Gradients - Ideal

Robert H. Whittaker (Ed), *Classification of Plant Communities*,
1978 (*Handbook of Vegetation Science*), Kluwer
Academic Publishers

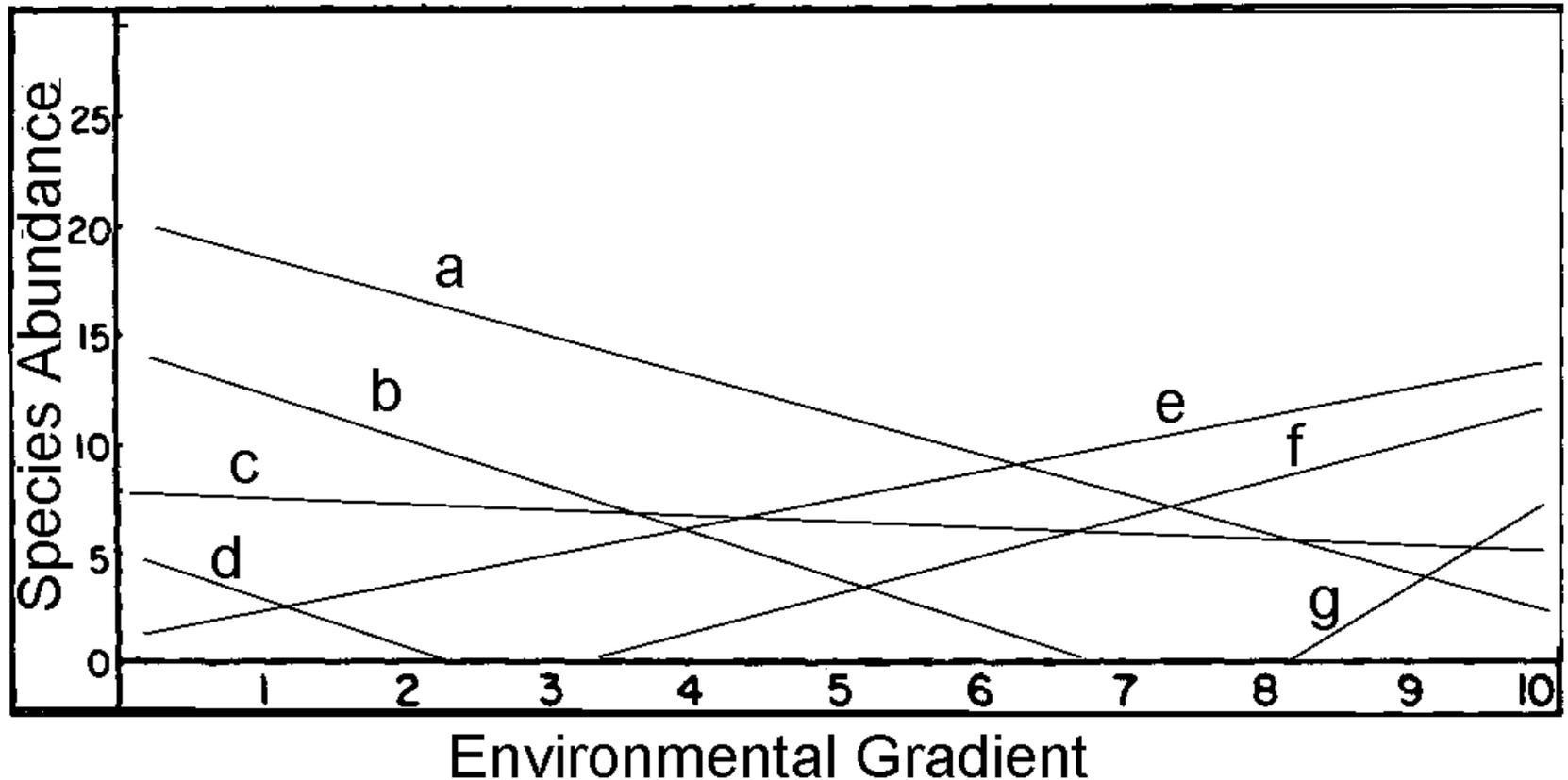
- Ideal species distributions across environmental gradients:
 - Gaussian Response: Smooth normal curves
Characterized by: mean \pm SD, mode
 - Linear Response: Smooth straight lines
Characterized by: range, slope

Gaussian (Normal) Distribution



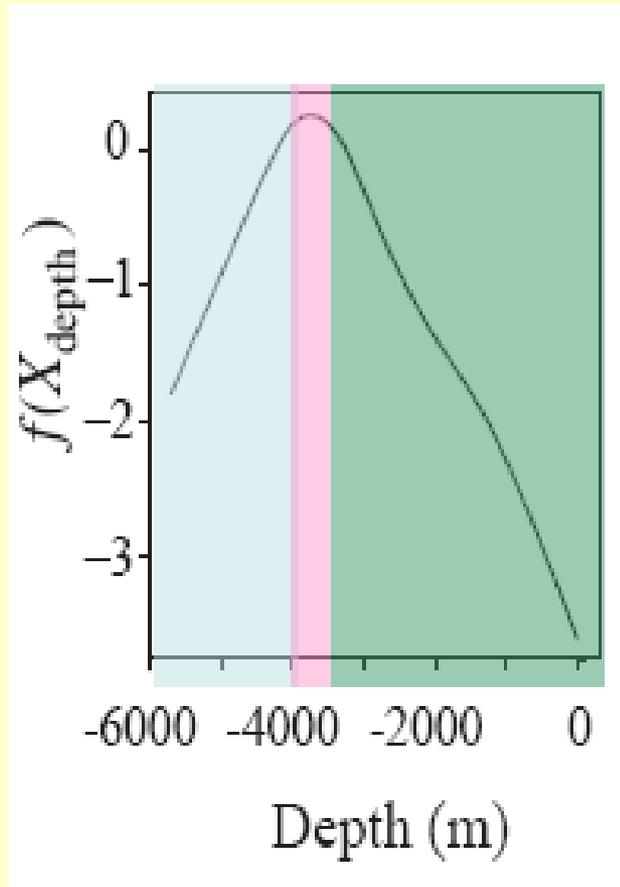
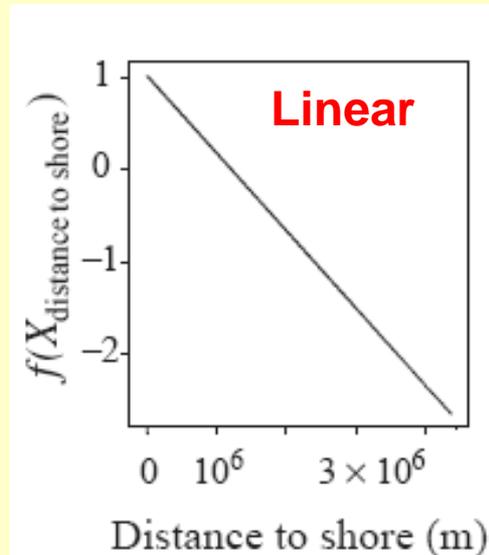
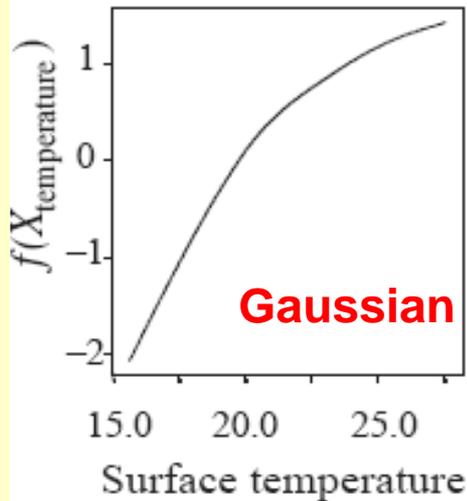
Hypothetical species abundance in response to an environmental gradient. Lettered curves represent different species. From Whittaker (1954).

Linear Distribution



Hypothetical linear responses of species abundance to an environmental gradient. Lettered lines represent different species. From Whittaker (1954).

Species on Environmental Gradients - Real



Study 1:
(4 - 6 km depth):
Species prefers
shallow habitat

Study 3:
(3.5 - 4 km depth):
Species shows
no preference

Study 2:
(0 - 3.5 km depth):
Species prefers
deep habitat

Some times the answer
is scale-dependent

Species on Environmental Gradients - Real

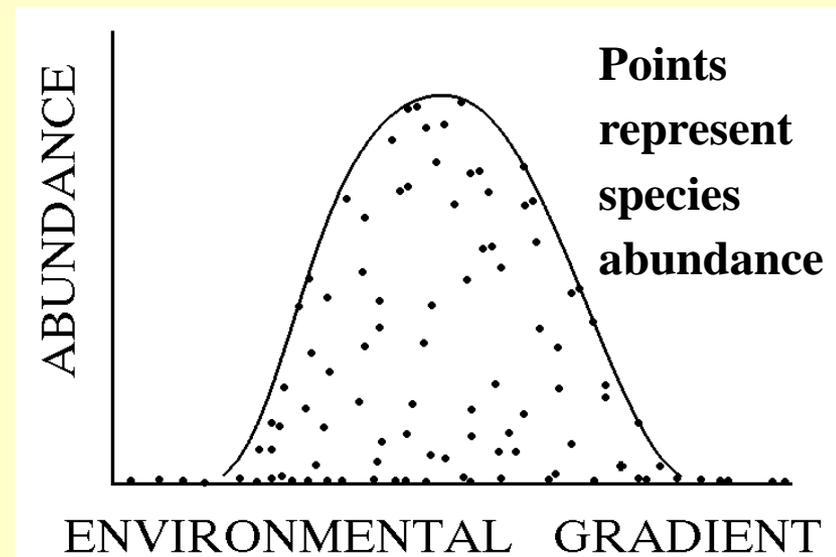
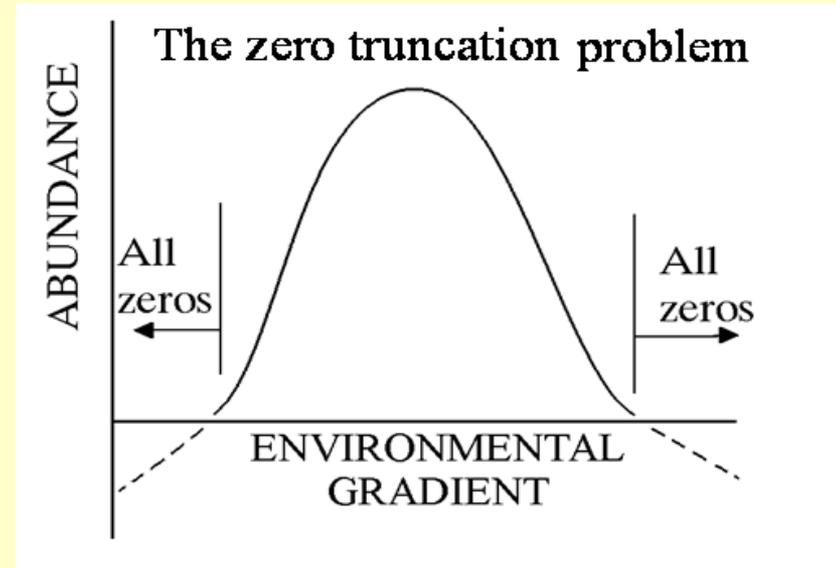
➤ 3 important issues to consider:

- Zero-truncation Problem:
- Solid Curves:
- Complex Curves:

Observation: Species are often below their “optimal abundance”, given the driving environmental factor. Why?

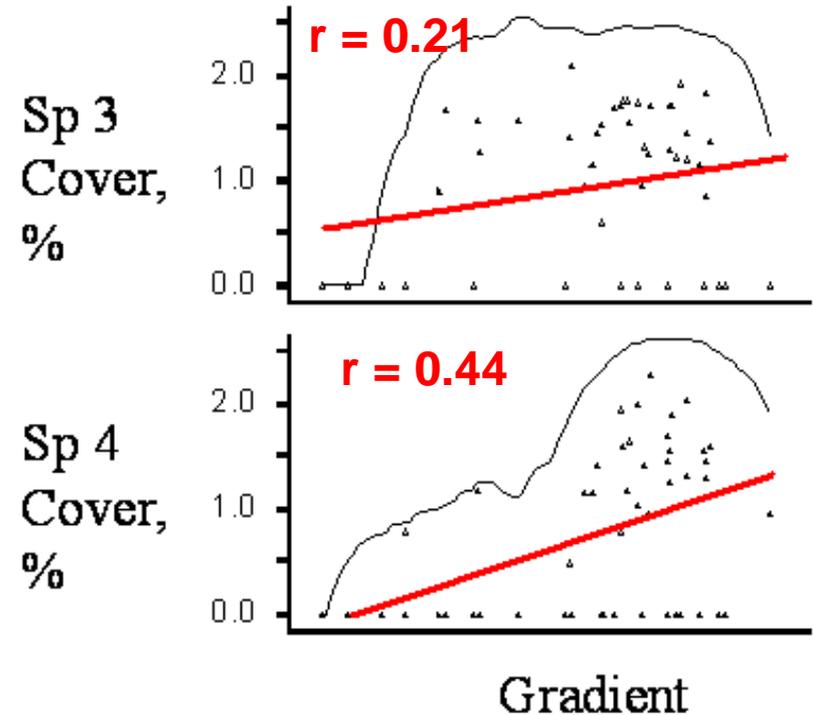
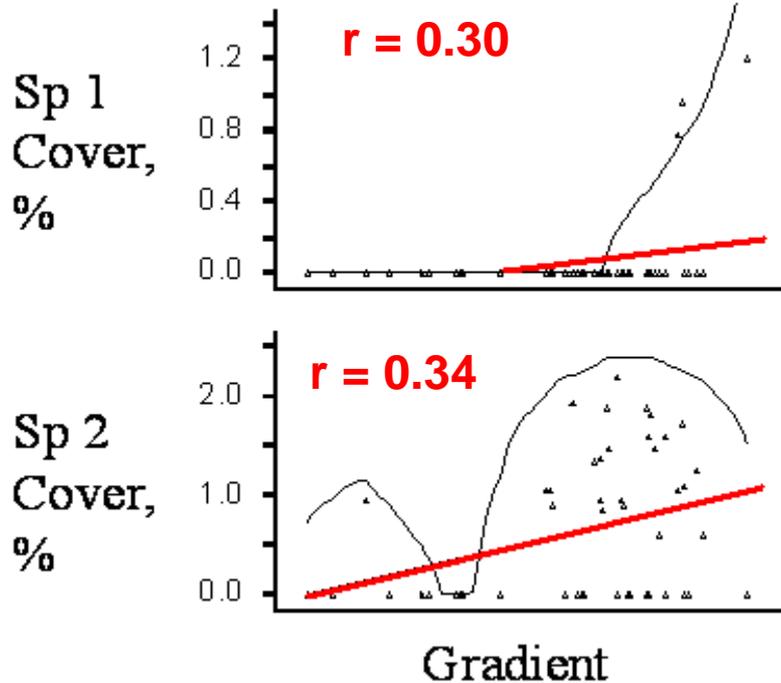
Other limiting factors

(e.g., Other environmental factors, other species, life-history, chance)



Species on Environmental Gradients - Real

— Linear Regression — Fitted Envelope



**Abrupt Ranges
(Boundaries)**

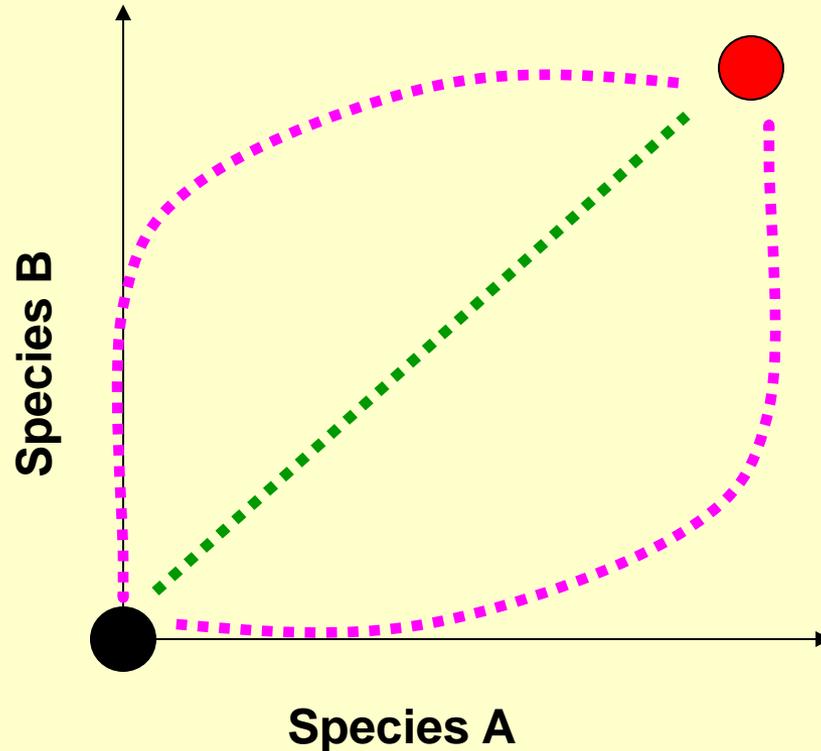
Multiple Modes

Linear Responses

Peaks (Optima)

Community Analysis – Bivariate Plots

More fruitful to explore how pairs of species abundances are related with “bivariate plots”.



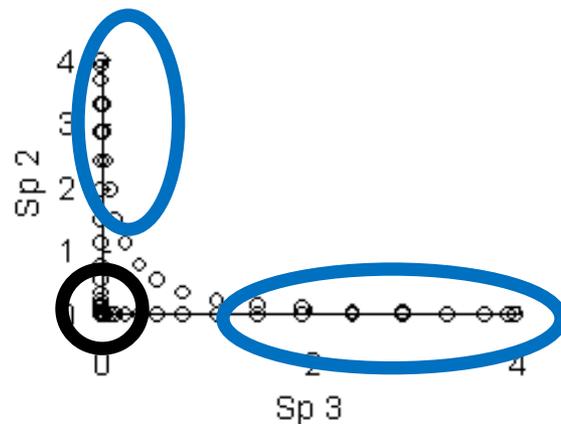
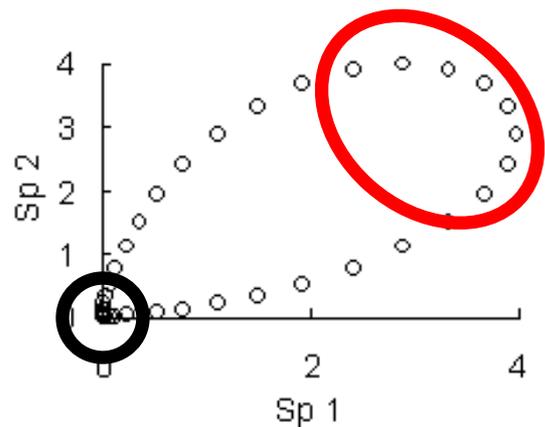
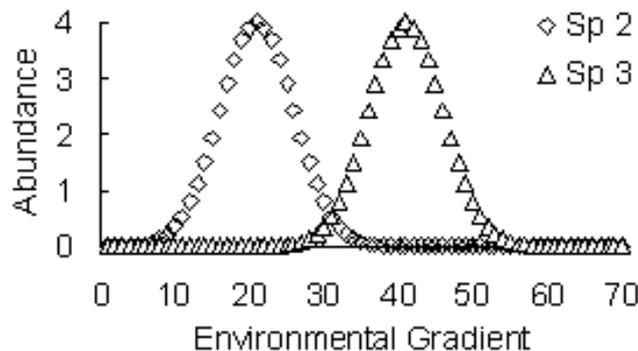
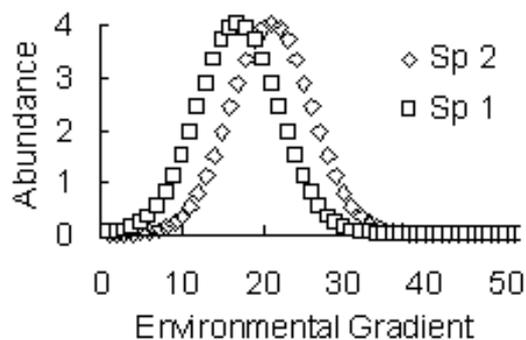
- Joint Absences
(0,0)
- Joint Occurrences
(lots, lots)
- Perfect correlation
- Weak correlation

Community Analysis – Bivariate Plots

Bivariate plots from pairs of species responses to the same environmental gradients

positively associated

negatively associated



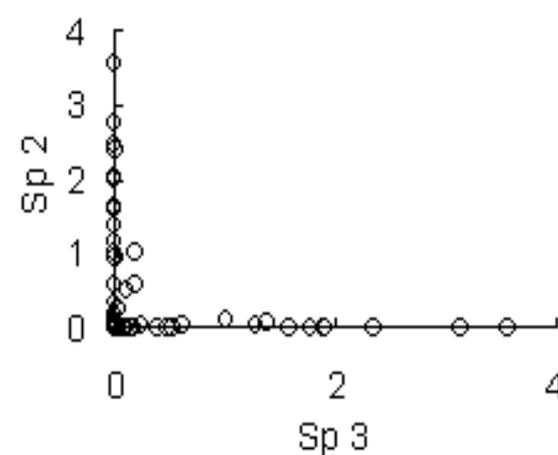
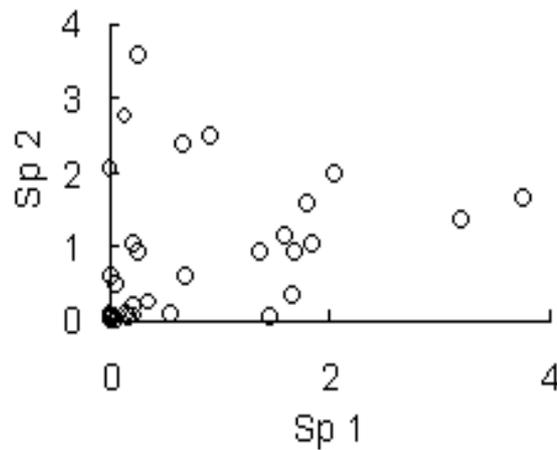
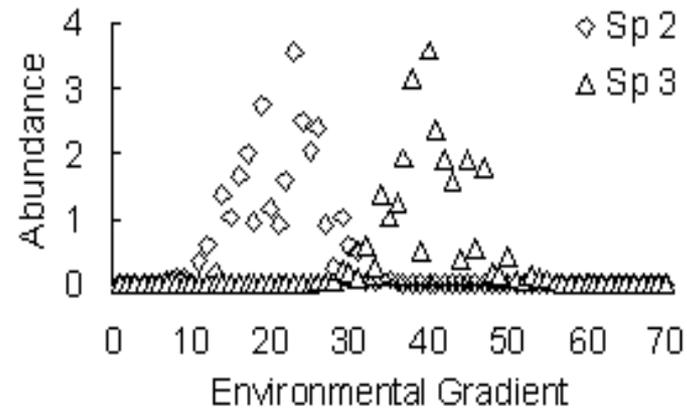
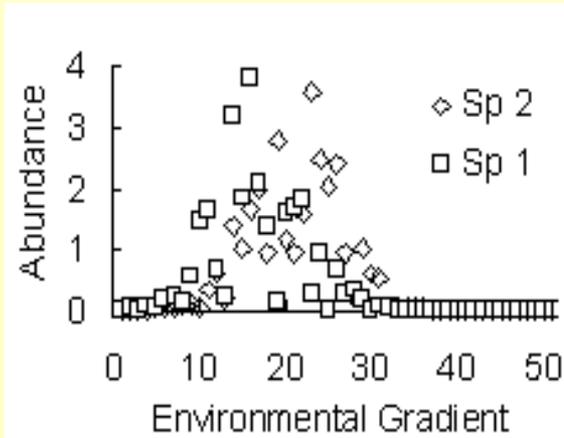
➤ Both have joint absence (0, 0) data

➤ Joint occurrence versus single occurrence

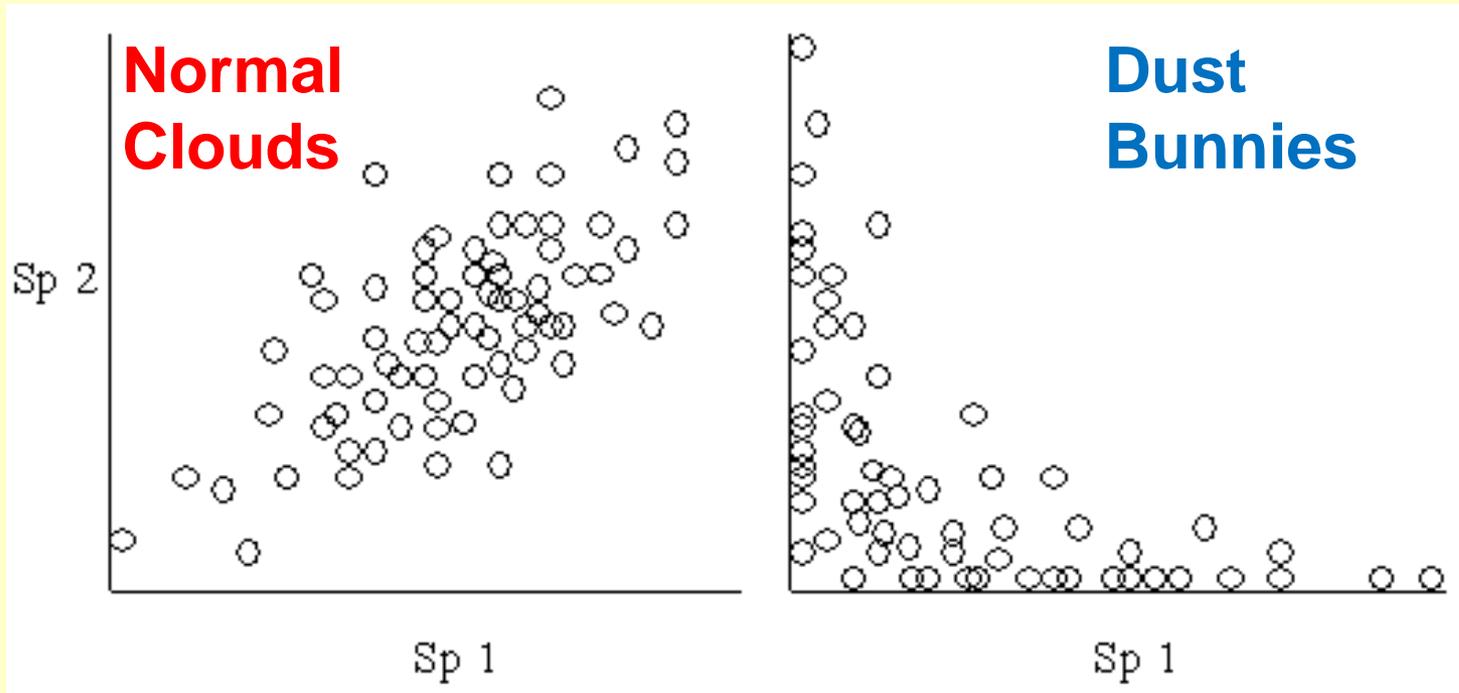
Community Analysis – Bivariate Plots

positively associated

negatively associated



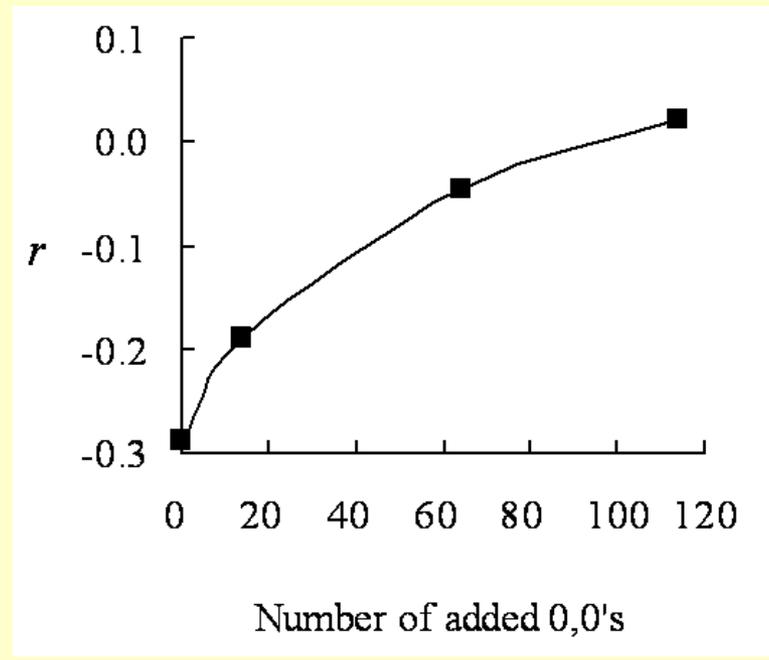
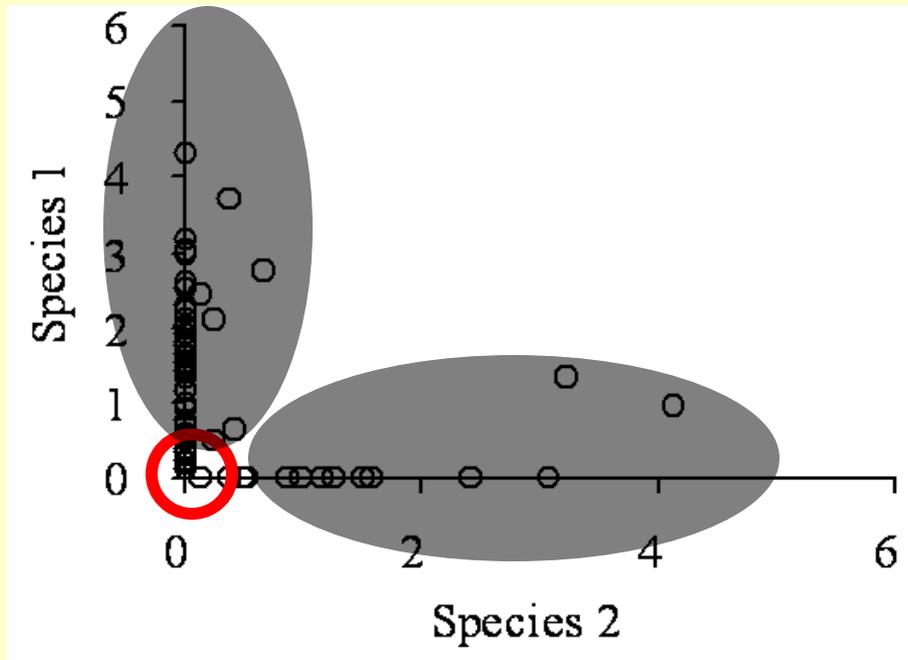
Community Analysis – Correlations



r is positive

r is negative

Community Analysis – Correlations



Consider two species with different habitat responses (a negative association)

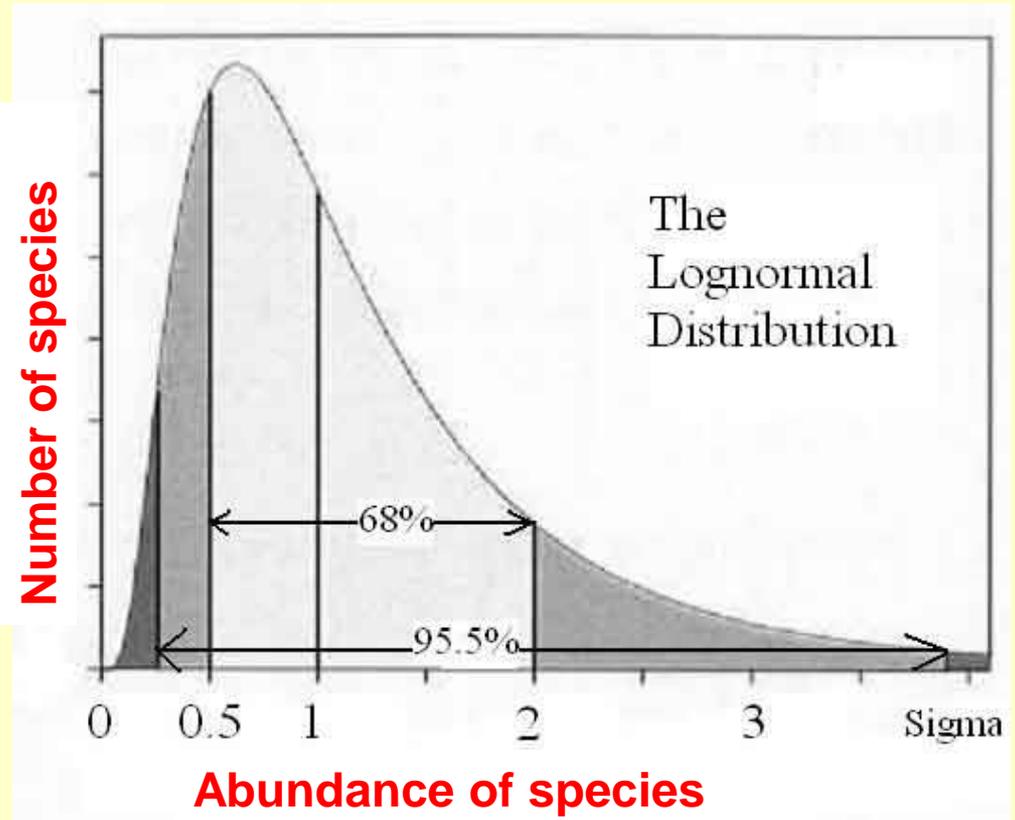
Beware:

As we sample beyond their habitats, we record more and more joint absences...

Summary: Sampling Communities

Species Abundance Data

- Large proportion of species absent
- Some species numerically dominant
- Most species infrequent



So What ?

Carefully consider sampling design (how big a sample ?) and the analysis, starting from choice of similarity measure

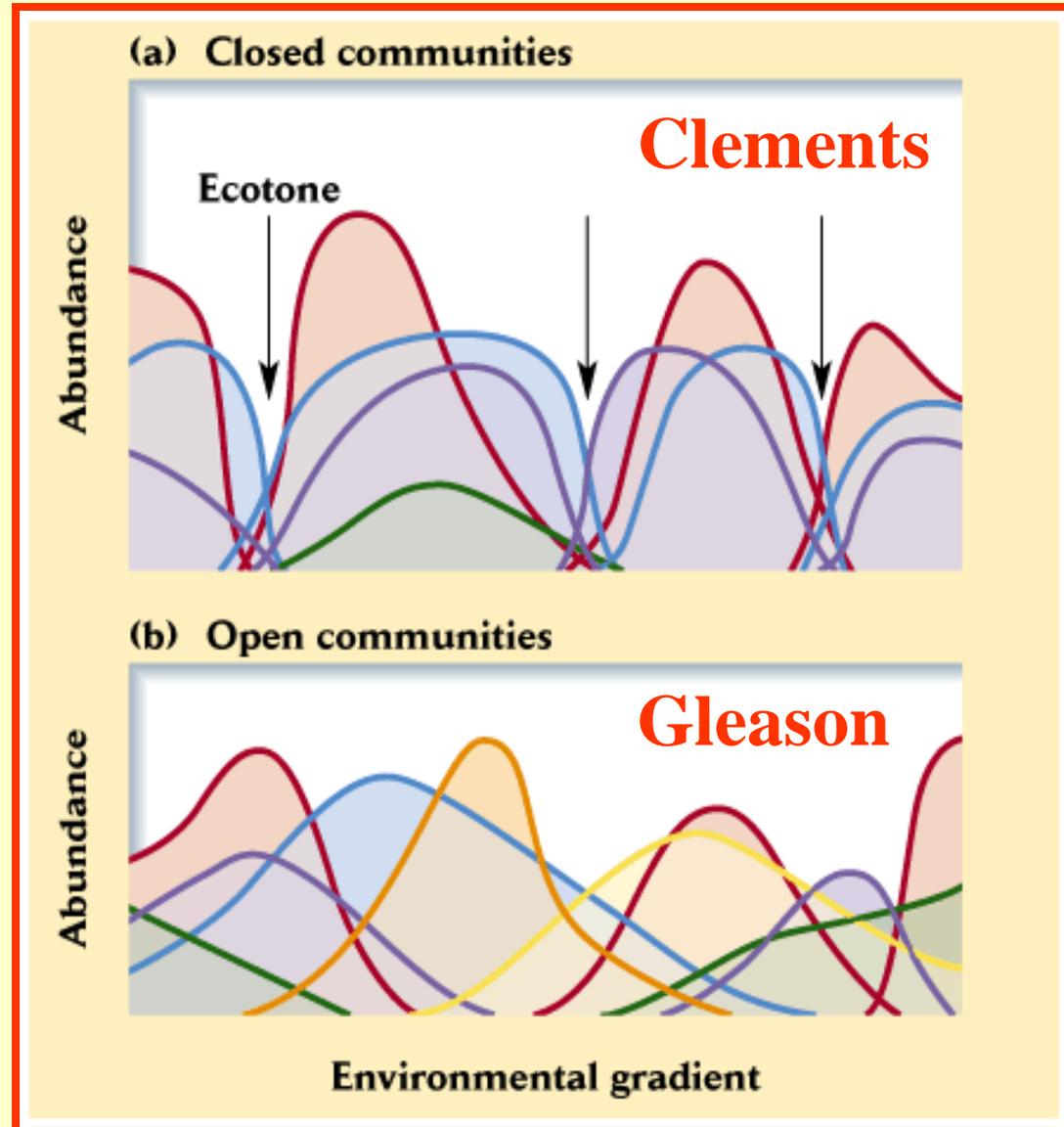
Summary: Sampling Communities

- Plots of species responses (abundance) to single environmental variables are informative:
 - unimodal – multimodal
 - linear / normal
 - peak(s) reveals optimums
- Bivariate plots (sp1 vs sp2) more informative: compare responses of pair of species to all environmental variables (at once)
- Typical responses:
“normal cloud” and “dust bunnies”

What is an Ecological Community ?

➤ Two views have dominated the debate over the nature of ecological communities since the 1920's:

- Clements' discrete unit
- Gleason's loose assemblage of species



What is an Ecological Community ?

➤ Clements' Perspective:

- Discrete entities with recognizable boundaries
- The community fully integrated functionally
- Species have coevolved, enhancing their interdependence

➤ Gleason's Perspective:

- Community is a chance association of species with similar adaptations and ecological requirements
- No distinct boundaries where communities meet

Community Analysis - Introduction

Community analysis techniques fall into two groups:

classification and **ordination**

- **Classification** is the placement of species and / or sample units into (discrete) groups
- **Ordination** is the arrangement or 'ordering' of species and / or sample units along gradients

Classification - Objectives

➤ Objectives and Limitations (James & McCulloch 1990)

Objectives:

1. To classify groups of objects judged to be similar according to distance or similarity measure
2. To reduce consideration of n objects to g (g less than n) group of objects

Limitations:

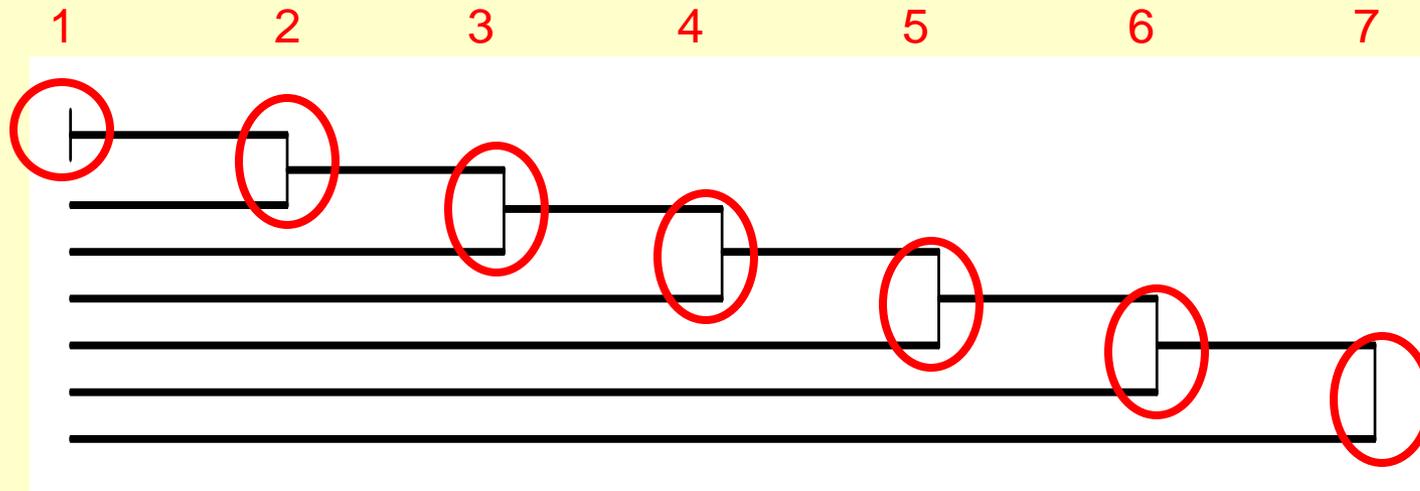
1. Results depend on the distance measure chosen.
2. Results depend on the algorithm chosen for forming clusters

Clustering - Introduction

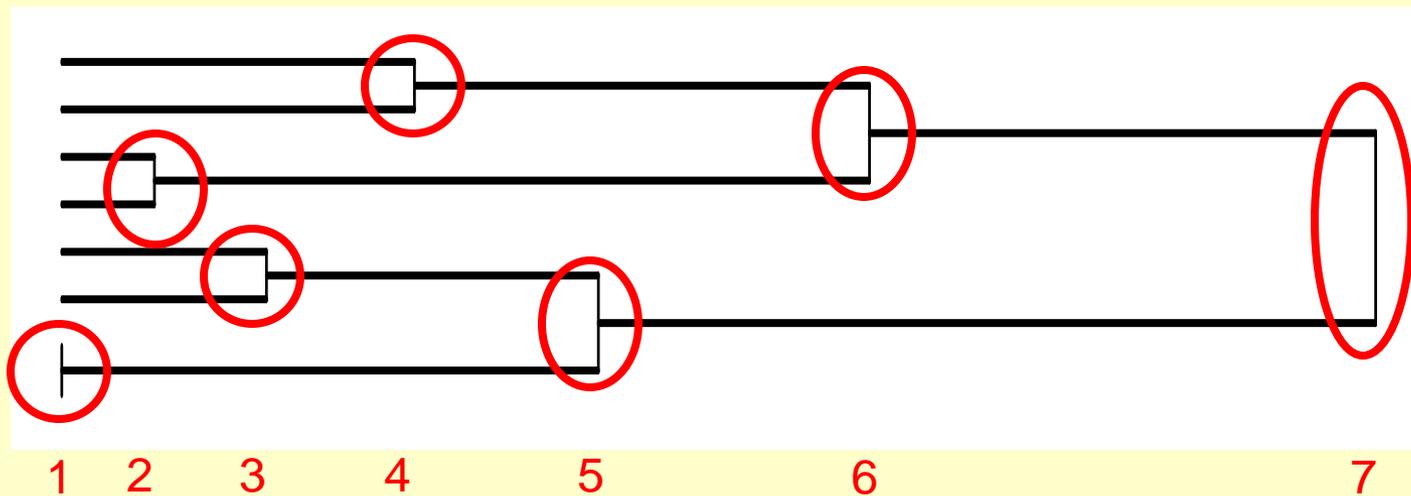
1. Start with pairwise similarity matrix among objects (individuals, sites, populations, taxa).
2. Two most similar objects are joined into a group, and the similarities of this group to all other units are calculated.
3. Repeatedly the two closest groups are combined until only a single group remains.
4. Results usually expressed in the form of a dendrogram, a two-dimensional hierarchical tree diagram representing the complex multi-variate relationships among the objects.

Clustering - Introduction

Two ways to sort eight samples (multiple species) into groups

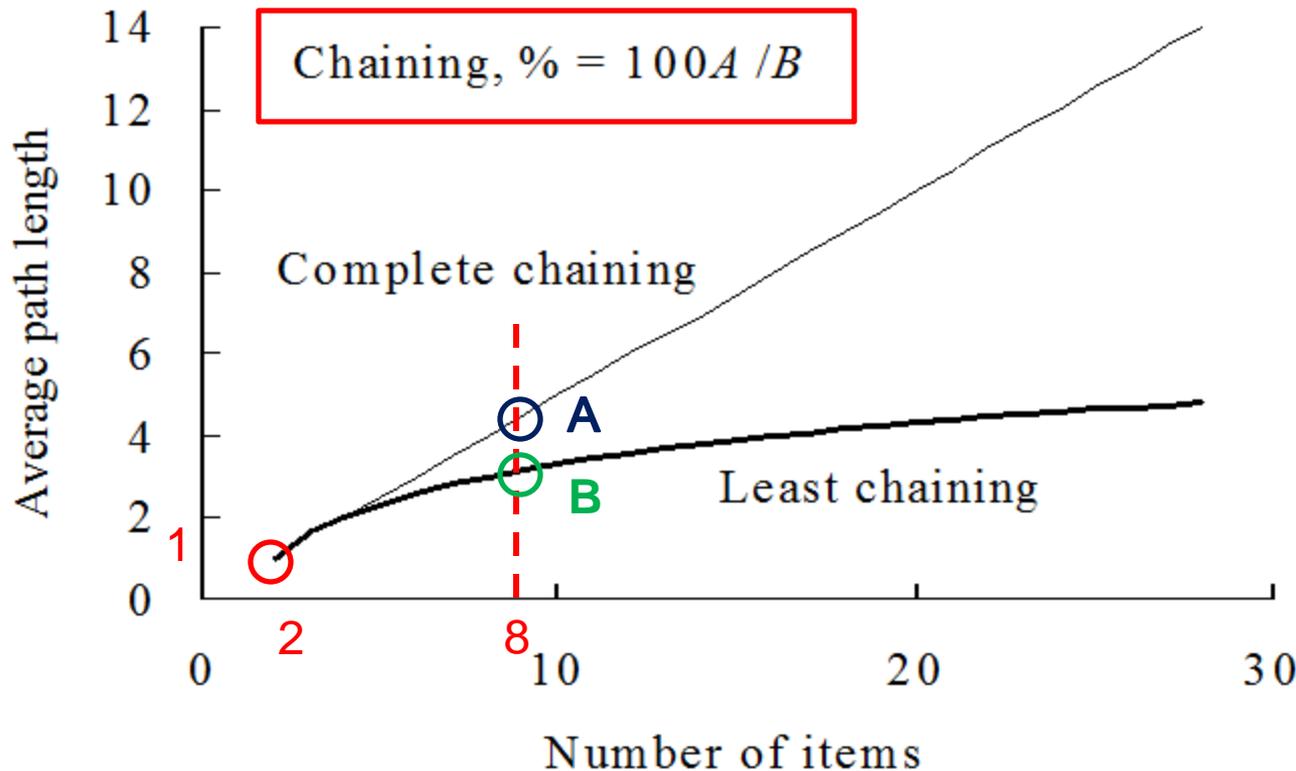


Complete chaining
(1 group)



No chaining
(2 groups)

Clustering - Introduction



Example with 2 items?

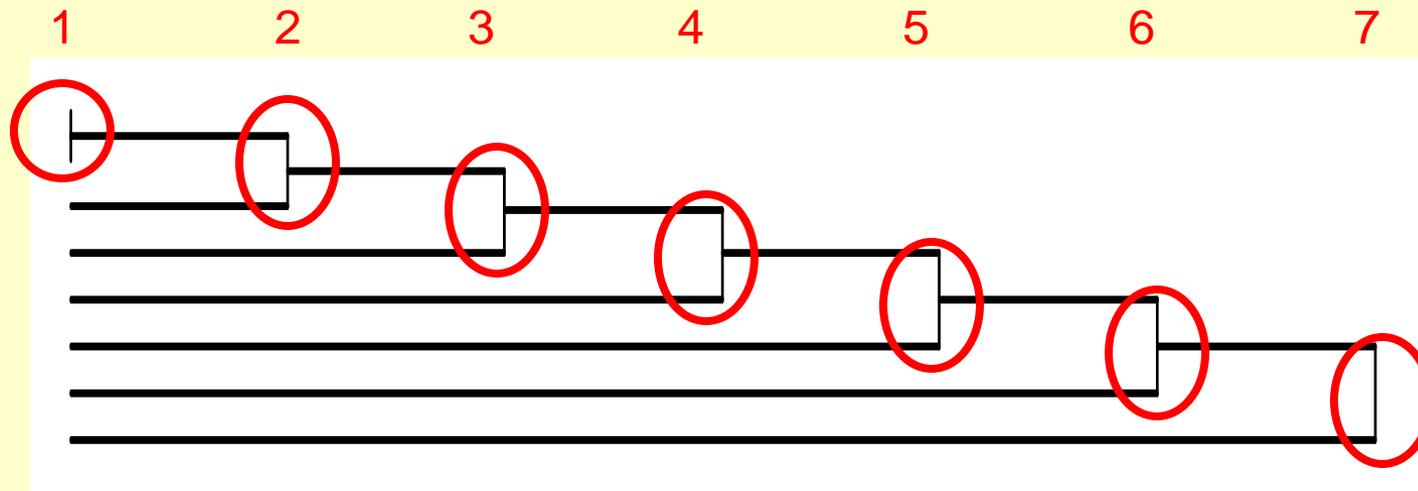
- Two paths
- One node in each
- Avg. Path Length = 1

Average path length used to measure percent chaining in cluster analysis. Path length is the number of nodes between tip of a branch and trunk.

Clustering - Introduction

Two ways to sort eight samples (multiple species) into groups

A) Complete chaining:



Number of paths = 8

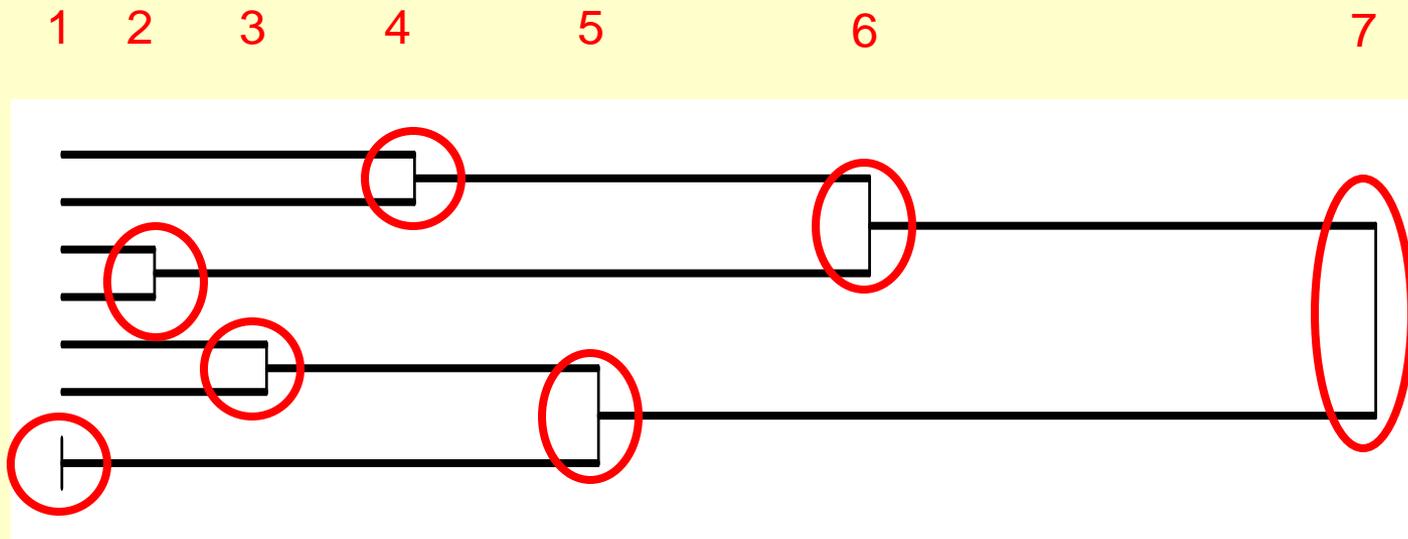
Sum of nodes = $7 + 7 + 6 + 5 + 4 + 3 + 2 + 1 = 35$

Avg. path length = $35 / 8 = 4.375$

Clustering - Introduction

Two ways to sort eight samples (multiple species) into groups

B) No chaining:

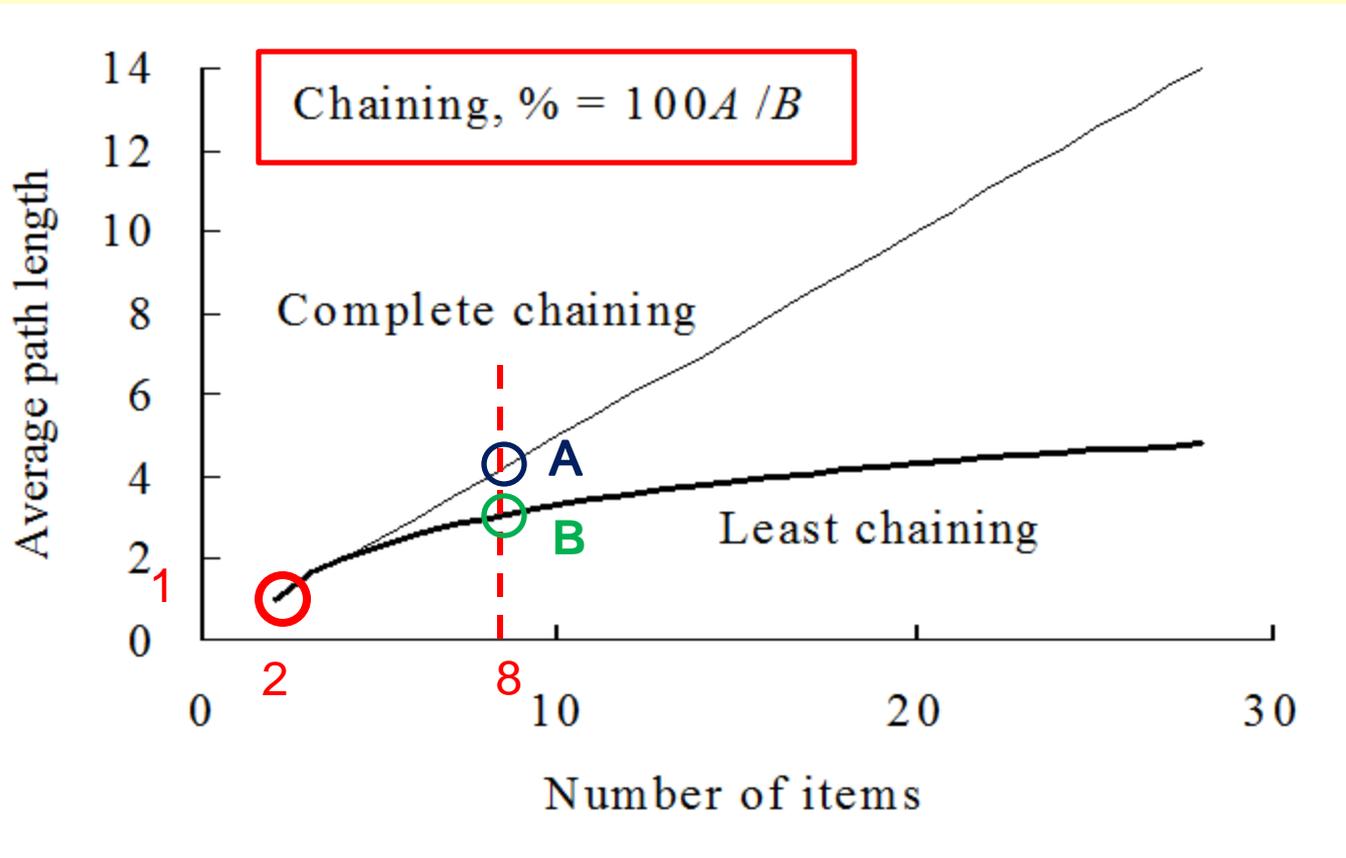


Number of paths = 8

Sum of nodes = 3 3 3 3 3 3 3 3 = 24

Avg. path length = $24 / 8 = 3.00$

Clustering - Introduction



NOTE: Chaining can be calculated for any given clustering pattern

Chaining (B) = $100 * (A / \text{complete-chain}) = 100 * (3 / 4.375) = 68.57 \%$

Chaining (A) = $100 * (A / \text{complete-chain}) = 100 * (4.375 / 4.375) = 100\%$

Clustering – How it Works

A dissimilarity matrix of order $n \times n$ ($n =$ number of entities) is calculated and each of the elements is squared. The algorithm then performs $n-1$ loops (clustering cycles) in which the following steps are done:

1. The smallest element (d_{pq}^2) in dissimilarity matrix sought (groups associated with this element are S_p and S_q).
2. The objective function E_n (the amount of information lost by linking to cycle n) is incremented according to the rule.
3. Group S_p is replaced by $S_p \cup S_q$ by recalculating the dissimilarity between the new group and all other groups (this requires calculating new dissimilarities).
4. Group S_q is inactive and its elements are assigned to new group $S_p \cup S_q$. After joining all items, procedure is complete.

Clustering – How it Works

The **objective function (E)** is the sum of the error sum of squares from each centroid to the items in that group.

Where :

t indexes the T clusters

E_t is the error sum of squares for cluster t

$$E = \sum_{t=1}^T E_t$$

And each E_t is found by:

$$E_t = \sum_{i=1}^{k_t} \sum_{j=1}^p (x_{ijt} - \bar{x}_{jt})^2$$

x_{ijt} is the value of the:

jth variable for the

ith point of cluster t

(which contains k_t points)

\bar{x} is the mean of the jth variable for cluster t.

Clustering – How it Works

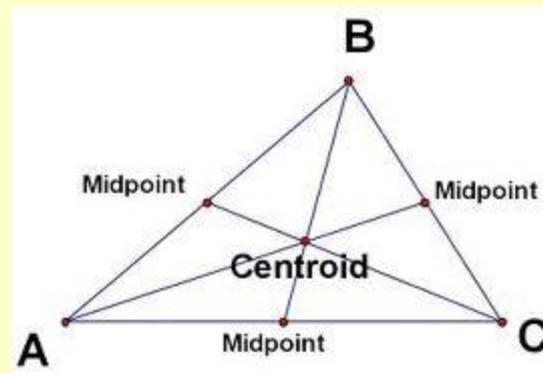
Calculate E for each cluster, separately and sum up
(Note: there are T clusters):

$$E = \sum_{t=1}^T E_t$$

Calculate E by summing the deviations between all points and centroid, for all variables:

$$E_t = \sum_{i=1}^{k_t} \sum_{j=1}^p (x_{ijt} - \bar{x}_{jt})^2$$

What is \bar{x} ?



A cluster of 3 points, plotted in 2 dimensions

Clustering – How it Works

We need a rule to progressively combine the elements, as we go through the cycles and the groups become larger.

The basic combinatorial equation is:

$$d_{ir}^2 = \alpha_p d_{ip}^2 + \alpha_q d_{iq}^2 + \beta d_{pq}^2 + \gamma |d_{ip}^2 - d_{iq}^2|$$

where values of α_p , α_q , β , and γ determine the type of sorting strategy (See Table below).

Why Are there Different Linkage Methods?

Use different coefficients in the basic combinatorial equation.

Linkage method	Coefficient			
	α_p	α_q	β	γ
Nearest neighbor	0.5	0.5	0	-0.5
Farthest neighbor	0.5	0.5	0	0.5
Median	0.5	0.5	-0.25	0
Group average	n_p / n_r	n_q / n_r	0	0
Centroid	n_p / n_r	n_q / n_r	$-\alpha_p \alpha_p$	0
Ward's method	$\frac{n_i + n_p}{n_i + n_r}$	$\frac{n_i + n_q}{n_i + n_r}$	$\frac{-n_i}{n_i + n_r}$	0
Flexible beta	$(1 - \beta)/2$	$(1 - \beta)/2$	β	0
McQuitty's method	0.5	0.5	0	0

n_p = number of elements in S_p

n_q = number of elements in S_q

n_r = number of elements in $S_r = S_p \cup S_q$

n_i = number of elements in S_i $i = 1, n$ except $i \neq p$ and $i \neq q$

Defining Groups (clusters)

The properties of linkage methods (“sorting strategies”) depend on type of dissimilarity measure used.

We consider two generic classes of dissimilarity measures:

Euclidean (absolute, relative)

Proportional (Sorensen, Jaccard)

We consider eight linkage methods:

Nearest neighbor Farthest neighbor

Median Group average

Centroid Ward's method

Flexible beta McQuitty's method

Homework #2 – Due Feb 22

➤ **The objectives of this homework are:**

- A) To review and practice data transformations.
- B) To calculate dissimilarities for a data matrix.
- C) To perform a clustering analysis.
- D) To practice reporting the results of clustering analyses.

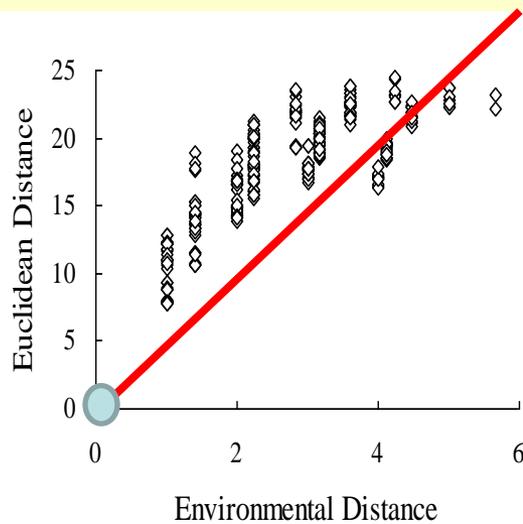
To complete this homework, you will need:

Instruction file: “BIOL6090_hw2.doc” – edit and turn in

“desserts colors.xls” data file: (open with excel) – do not turn in

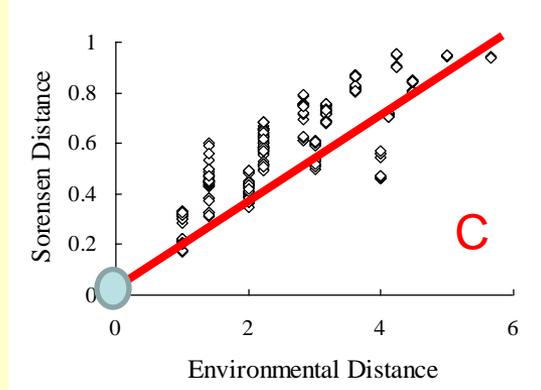
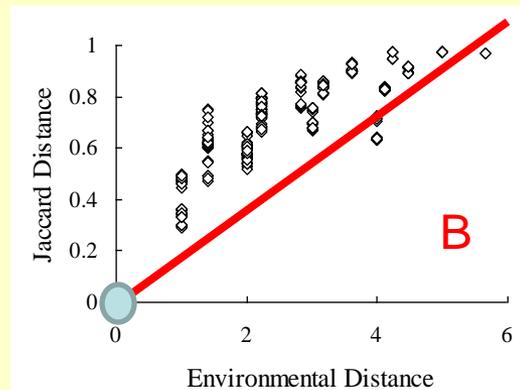
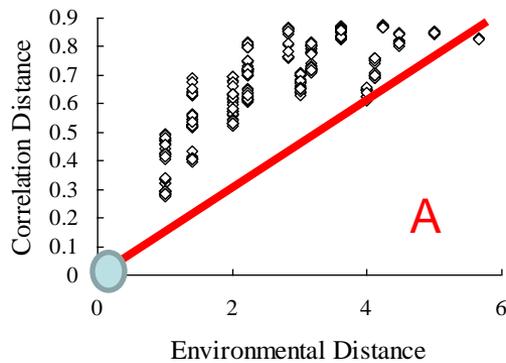
“aMoss1M.WK1” data file: (open with PC-ORD) – do not turn in

Distance Measures – Recommendations



Is the distance metric bounded ?

Is the distance metric monotonic ?



NOTE: use linear regression “anchored” at origin (0,0)

Homework #2

– Readings

- Critically reading
of journal articles



Fisheries Research 31 (1997) 147–158



Cluster analysis of longline sets and fishing strategies within the Hawaii-based fishery

Xi He ^{a,*}, Keith A. Bigelow ^a, Christofer H. Boggs ^b

^a Pelagic Fisheries Research Program, Joint Institute for Marine and Atmospheric Research, School of Ocean and Earth Science and Technology, University of Hawaii, 2570 Dole Street, Honolulu, HI 96822, USA

^b Honolulu Laboratory, Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, 2570 Dole Street, Honolulu, HI 96822, USA

540

JOURNAL OF CLIMATE AND APPLIED METEOROLOGY

VOLUME 26

The Southern Oscillation in Surface Circulation and Climate over the Tropical Atlantic, Eastern Pacific, and Indian Oceans as Captured by Cluster Analysis

KLAUS WOLTER

Department of Meteorology, University of Wisconsin, Madison, WI, 53706

(Manuscript received 30 June 1986, in final form 14 November 1986)

J Chron Dis Vol. 35, pp. 623 to 633, 1982
Printed in Great Britain. All rights reserved

0021-9681/82/080623-07\$03.00/0
Copyright © 1982 Pergamon Press Ltd

CLUSTER ANALYSIS TO DETERMINE HEADACHE TYPES*

PAULA DIEHR, GEORGE DIEHR, THOMAS KOEPEL, ROBERT WOOD,
KIRK BEACH, BARRY WOLCOTT and RICHARD K. TOMPKINS