Keep in touch

@SEEC_UCT

SEEC.UCT

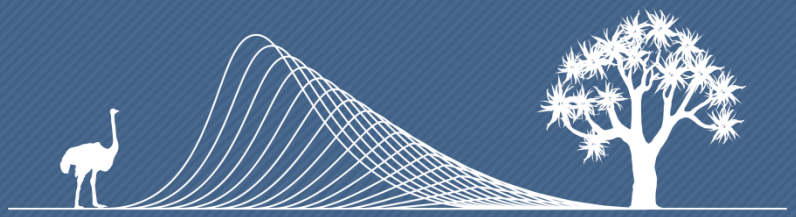
ask to be added to mailing list
(seecuct@gmail.com)

www seec.uct.ac.za

SEEC - Statistics in Ecology, Environment and Conservation

Intro to Multivariate Analyses

Natasha Karenyi



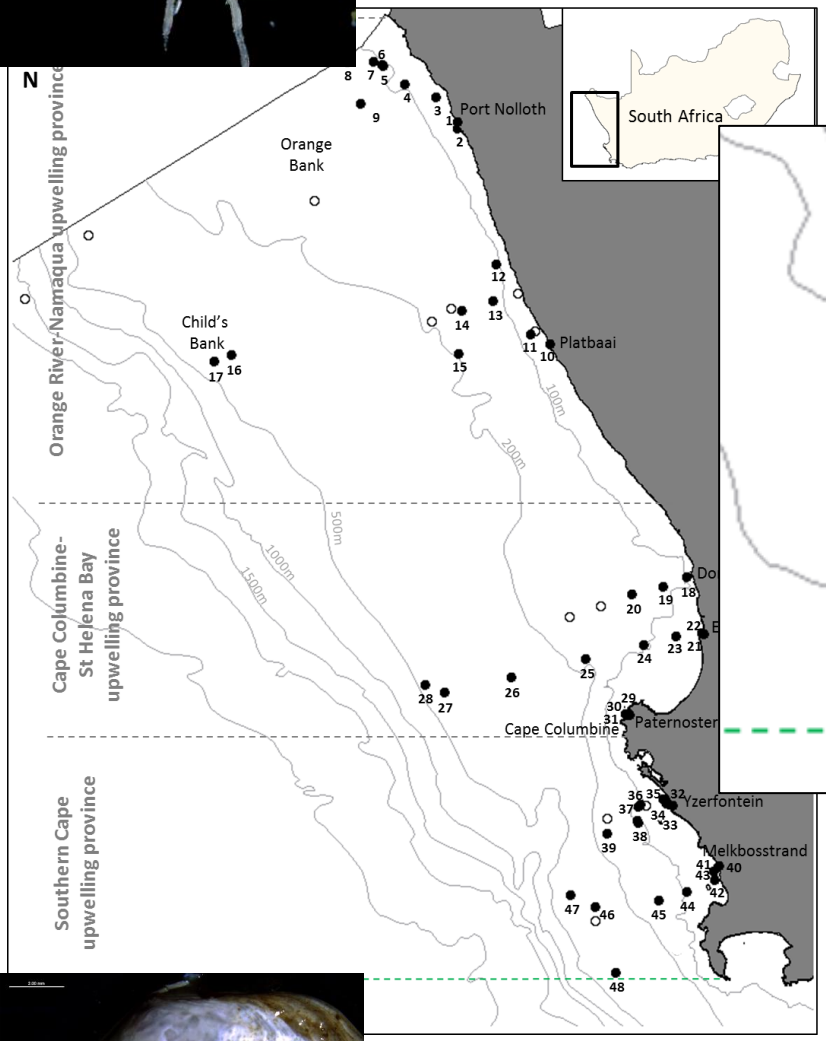
Introduction

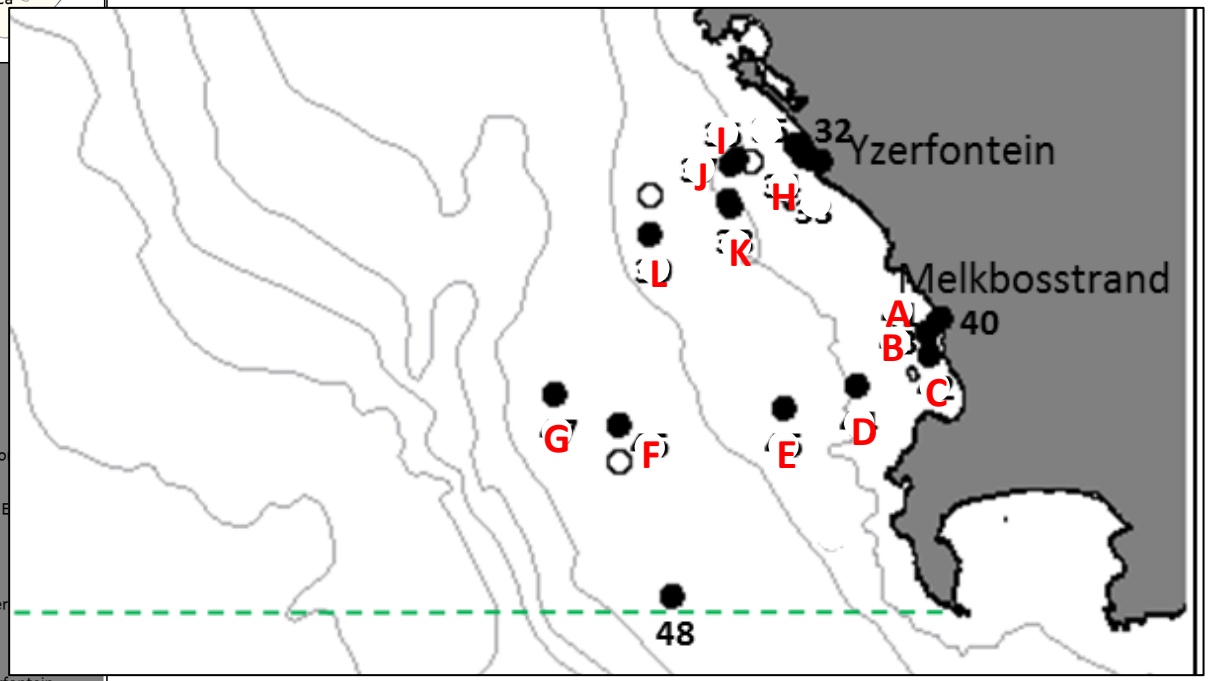
- What is a multivariate analysis?
  - It is a statistical process with multiple dependent and independent variables
- Broad types of analyses: association-based and model-based
- Association-based are most common in ecology
- Only recently model-based analyses have become more accessible

Data types

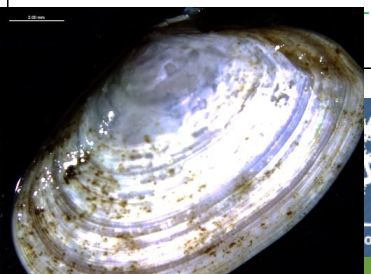
- Species
  - Presence/absence, ordinal, count, biomass, percentage cover
- Environmental
  - Geological, oceanographic, climate
- Morphological/Traits
  - Size or shape measurements, life history traits, sex, etc.
- Survey/Questionaire
  - Nominal, ordinal, measurements

Example

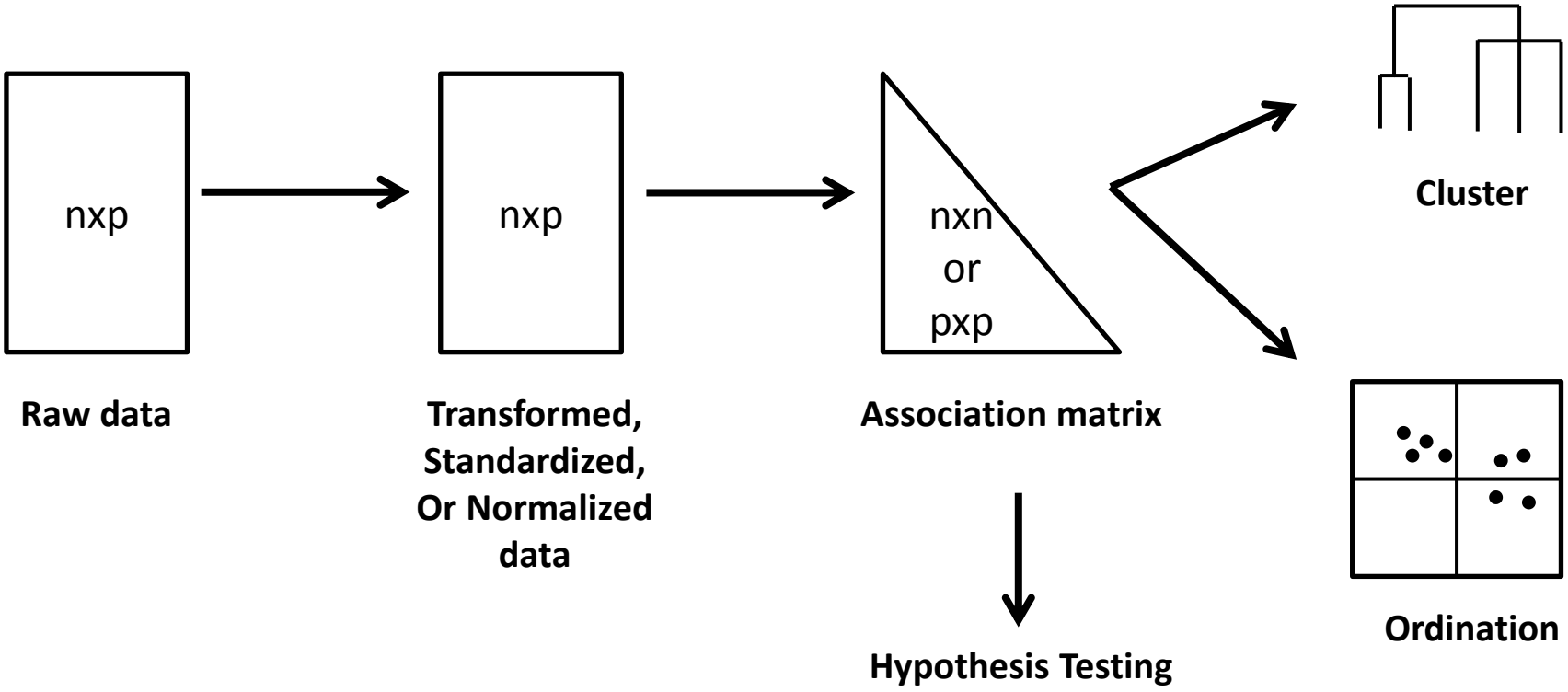




Macrofauna data – species counts
Environmental data e.g. sediment type, depth, oxygen, etc



Common Association-based multivariate analyses



Association Matrices

- Distance matrices
  - Metric distance between objects
  - 0 < D < n: identical < D < different
  - Does not deal well with double zeros
  - E.g. Euclidean distance – quantitative data (env data or morphological measurements)
- Similarity matrices
  - Measure the association between objects
  - 0 < S < 1: completely dissimilar < S < identical
  - E.g. Bray-Curtis similarity – Ordinal, count, biomass, percentage cover, presence-absence (species data)
  - E.g. Gower coefficient – mixed data types (survey data)

Common Association-based Multivariate Analyses

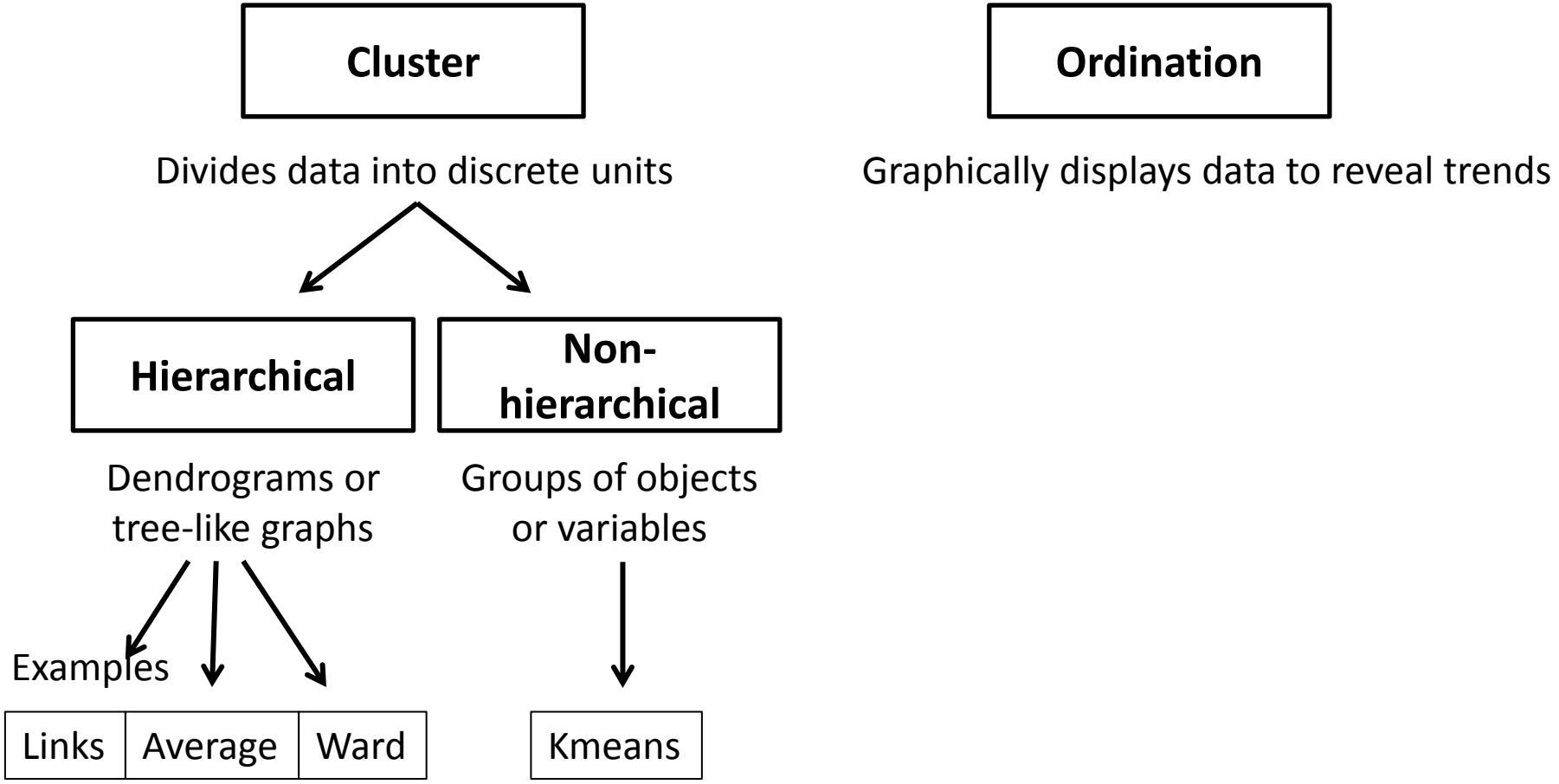
Cluster

Divides data into discrete units

Ordination

Graphically displays data to reveal trends

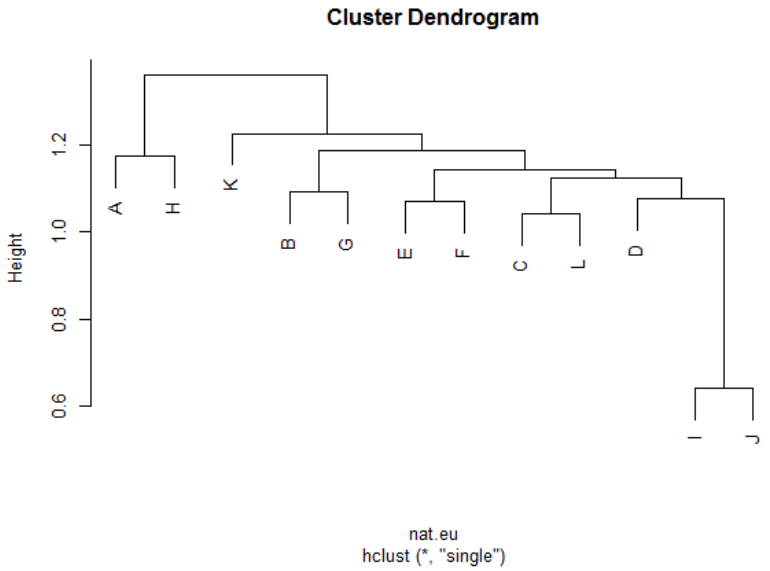
Common Association-based Multivariate Analyses

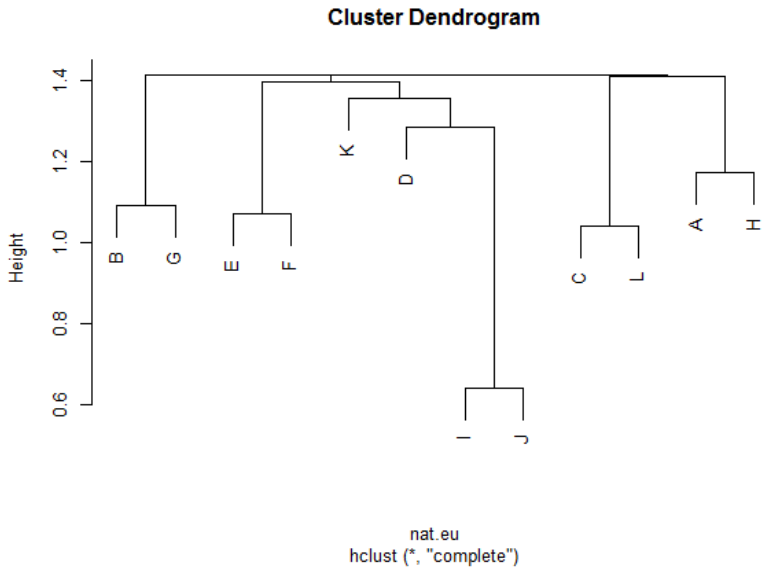


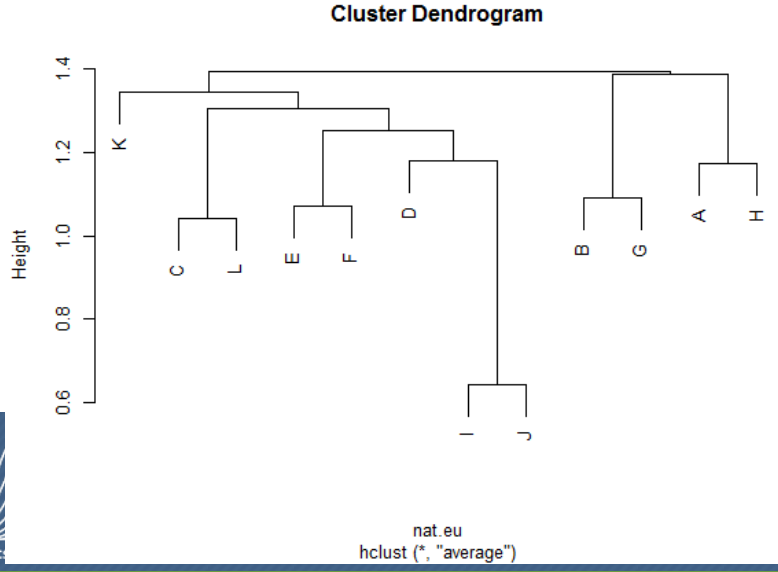
Hierarchical Clustering

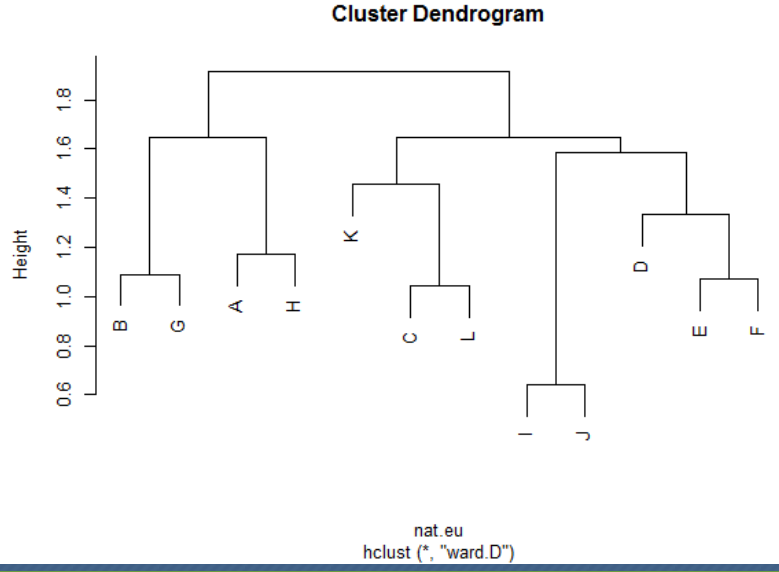
- hclust() in stats package
- Links
  - Groups agglomerate based on nearest (single linkage) or furthest (complete linkage) neighbour sorting
- Average Agglomerative
  - Average similarity of objects among clusters
- Ward's
  - Based on the linear model criterion of least squares. Groups defined so that the sum of squares within a group is minimized.

Hierarchical Clustering e.g.

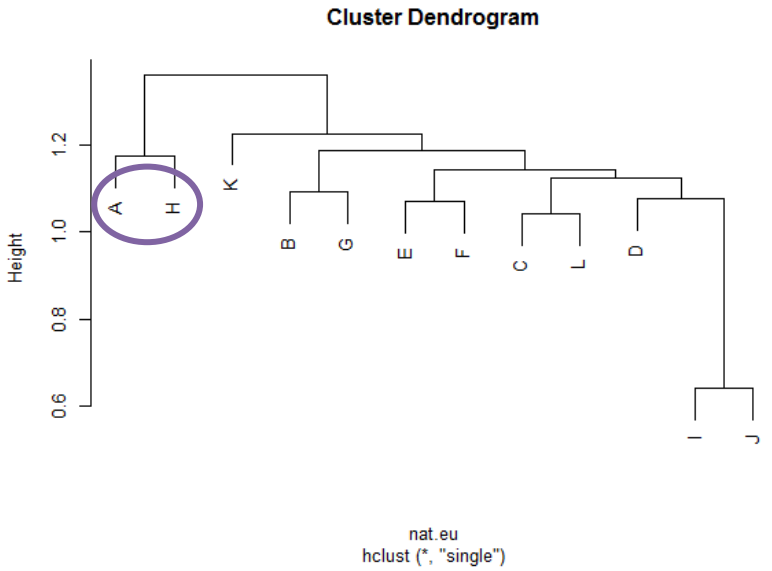


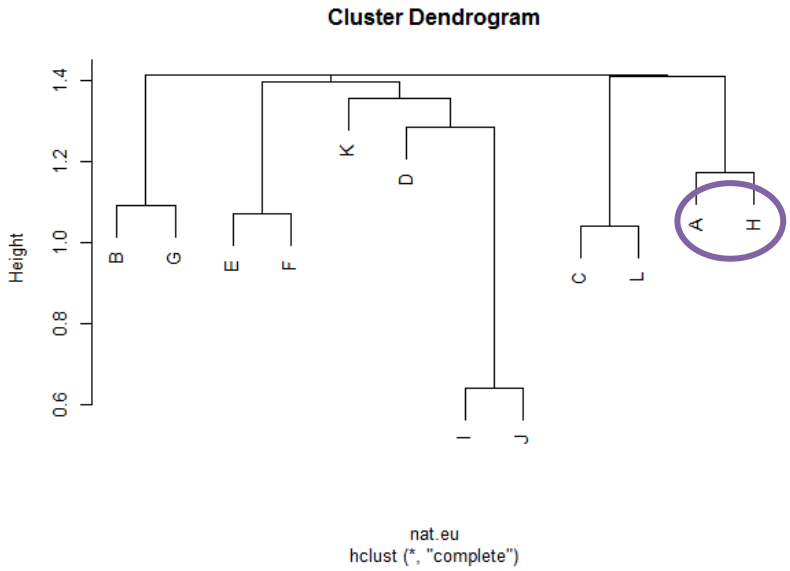


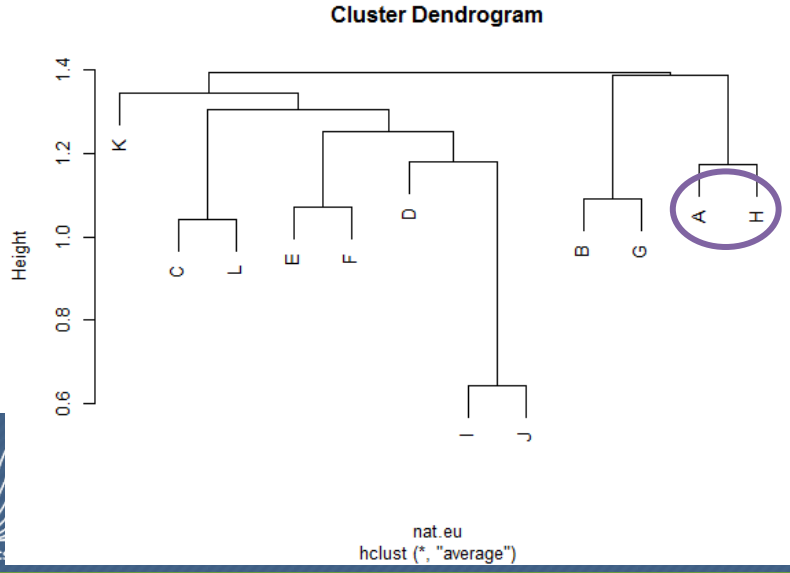


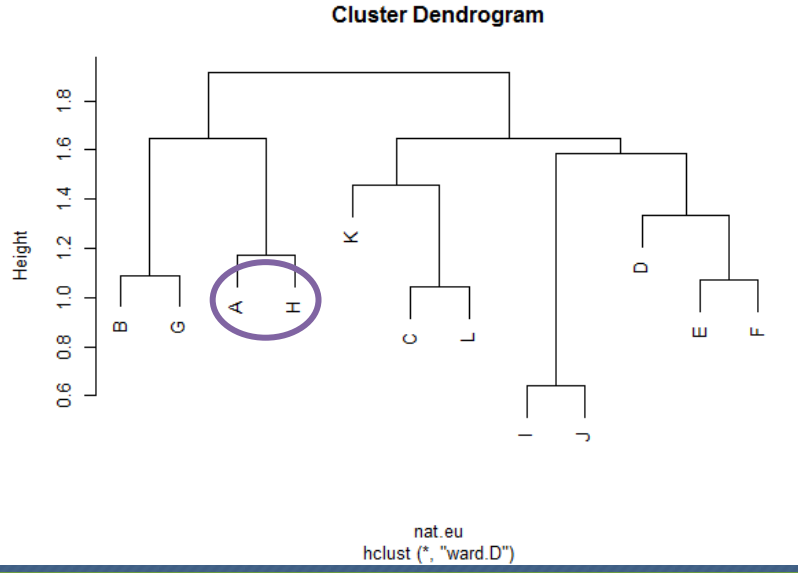


Hierarchical Clustering e.g.





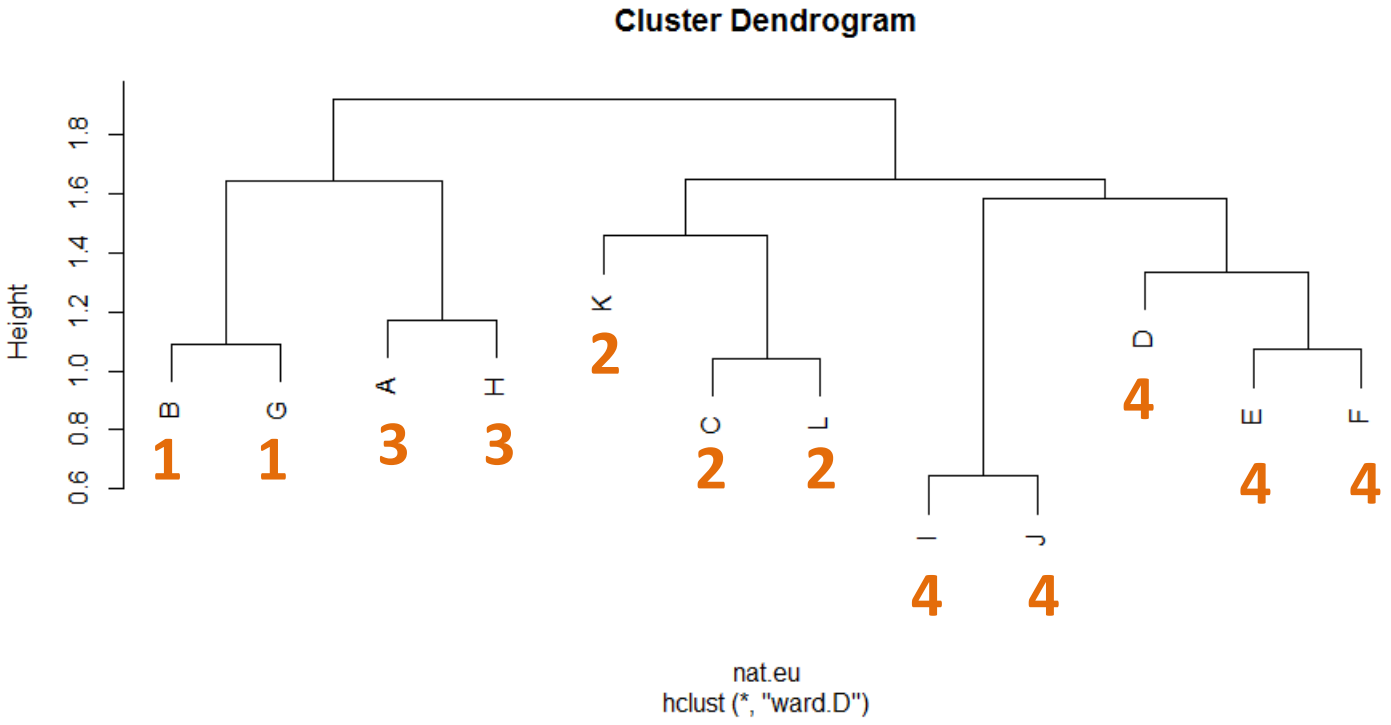




Non-Hierarchical Clustering

- kmeans() in stats package
- Kmeans
  - divisive clustering method
  - n objects in p-dimensional space should be divided into k groups so that the objects within the cluster are more similar to each other than the other groups.
  - Works on Euclidean distance, so other similarity matrices (e.g. Bray-Curtis) will have to be converted to rectangular matrices before applying kmeans command

Example of kmeans

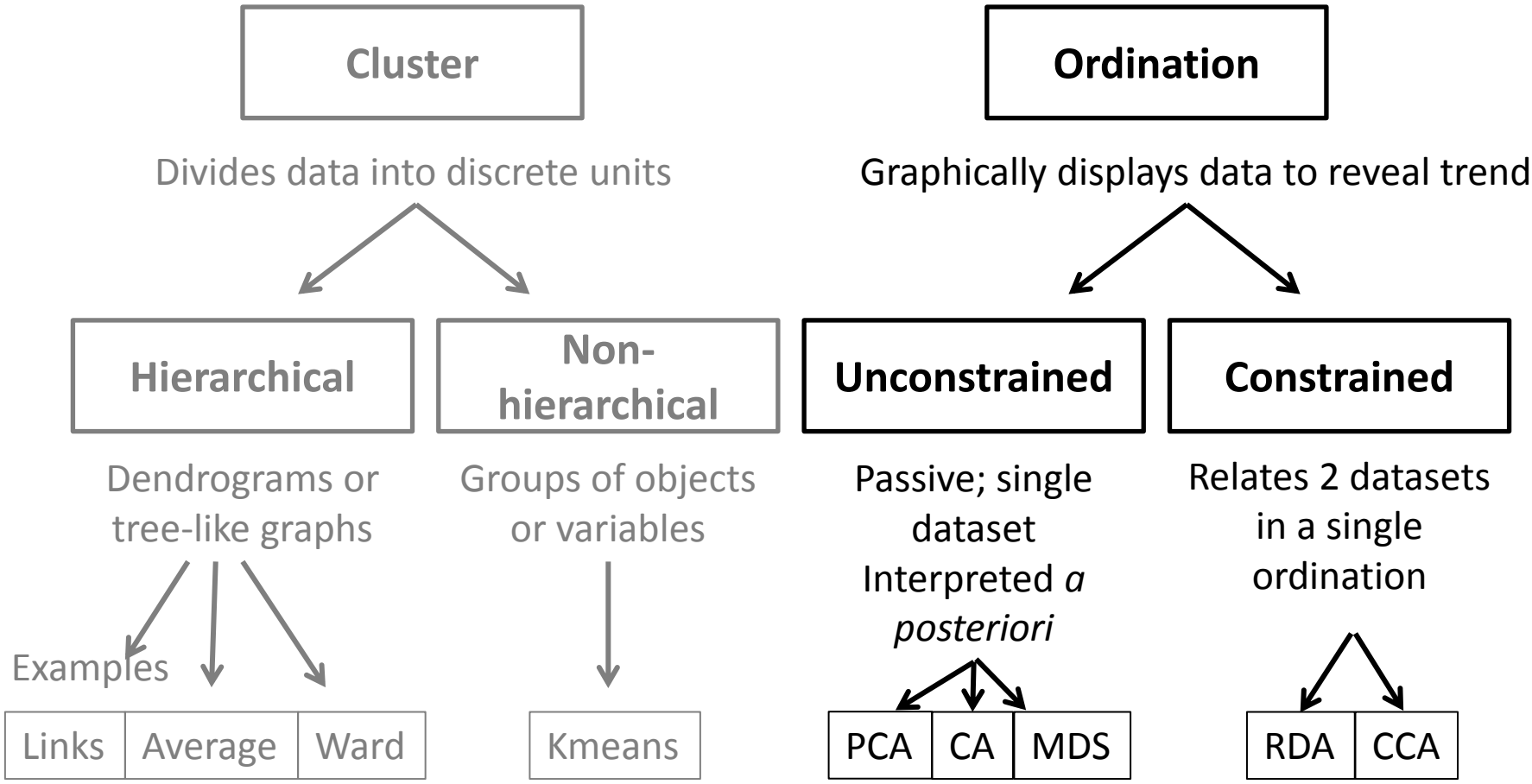


Kmeans groups

Clustering

- No “best” clustering method
- Method and groups are subjective
- Can be tested e.g. cophenetic correlation, bootstrap
- Clustering forces groups, therefore could miss gradients
- Always use clustering with ordination

Common Association-based Multivariate Analyses



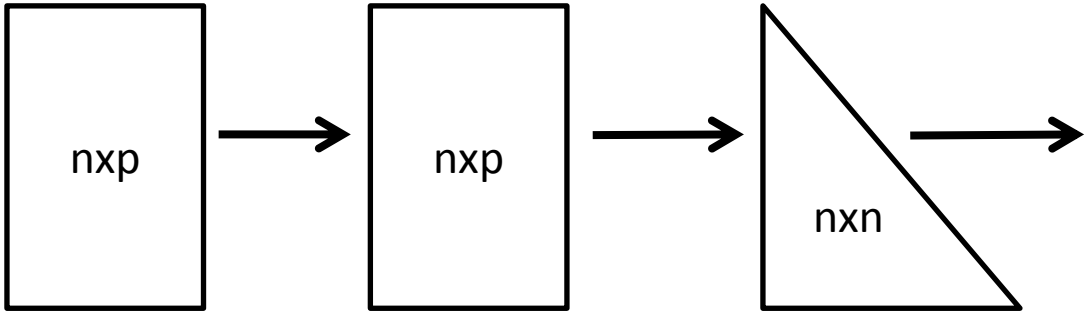
Unconstrained Ordination

- Principal component analysis (PCA)
  - Reduces number of variables to indices (biplots)
- Correspondence analysis (CA)
  - Compares distributions of rows and columns (biplots)
- Non-metric Multidimensional scaling (NMDS)
  - map of how individuals or sites are related based on distance matrices

Principal Component Analysis (PCA)

- Data types:
  - Quantitative variables
  - environmental or morphological measurements
  - Species data (with prior transformations)
  - not too many zeros or too many variables
- Packages in R
  - rda() in vegan
  - dudi.pca() in ade4
  - prcomp() in stats

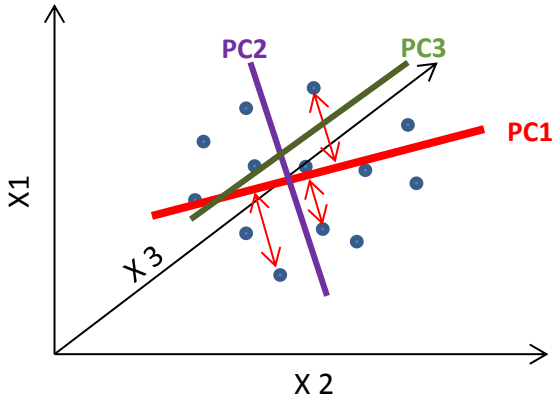
PCA



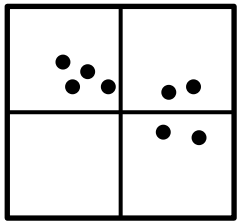
Raw data

Normalized data
i.e. standardize variables to zero means and unit std dev; equal weighting

Covariance matrix = Correlation matrix



Find eigenvalues[1] and eigenvectors[2] for each PC; indices e.g.
$PC1 = a_1X_1 + a_2X_2 + a_3X_3$

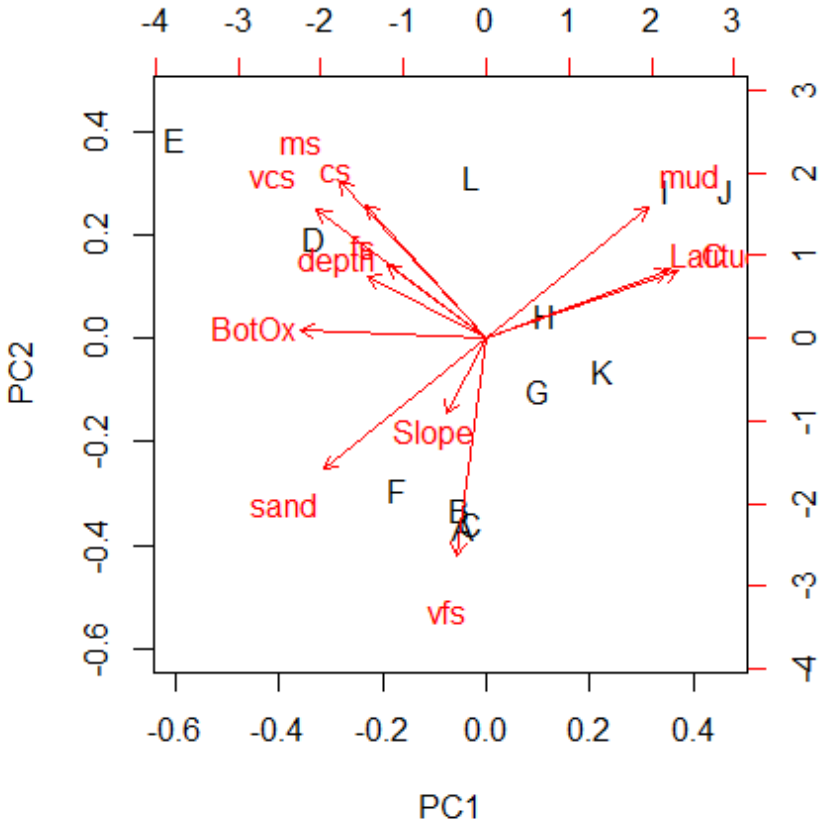


Ordination Biplot
Euclidean distance preserved;
Linear relationships detected
PCs>70% variation

Eigenvalues = variation accounted for by each PC
Eigenvectors = coefficients of each variable in a PC (i.e. a_i)

(Manly 1986)

PCA example



Scaling 1: distances between objects/sites are preserved, general direction of variables show importance for objects, but no interpretation for relationships between variables

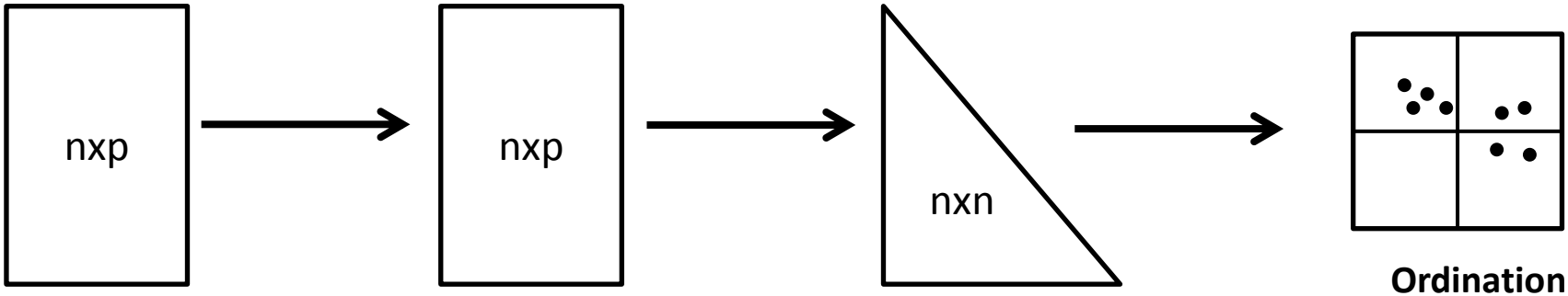
Correspondence analysis

- Data types:
  - species abundance, biomass or percentage cover
  - Quantitative, ordinal or nominal (different ca's)
  - Dimensionally homogenous (i.e. same units)
  - No negative values
- Packages in R
  - cca() in vegan
  - ca() in ca
- Notes
  - Compares distributions of rows and columns
  - Based on proportions of total sums in rows and columns
  - χ^2 distance preserved

Non-metric Multi-Dimensional Scaling

- Data types:
  - species abundance, biomass, percentage cover, presence-absence
  - Quantitative, ordinal or nominal
- Packages in R
  - metaMDS() in vegan
  - isoMDS() in MASS (if some values missing)

NMDS



Raw data

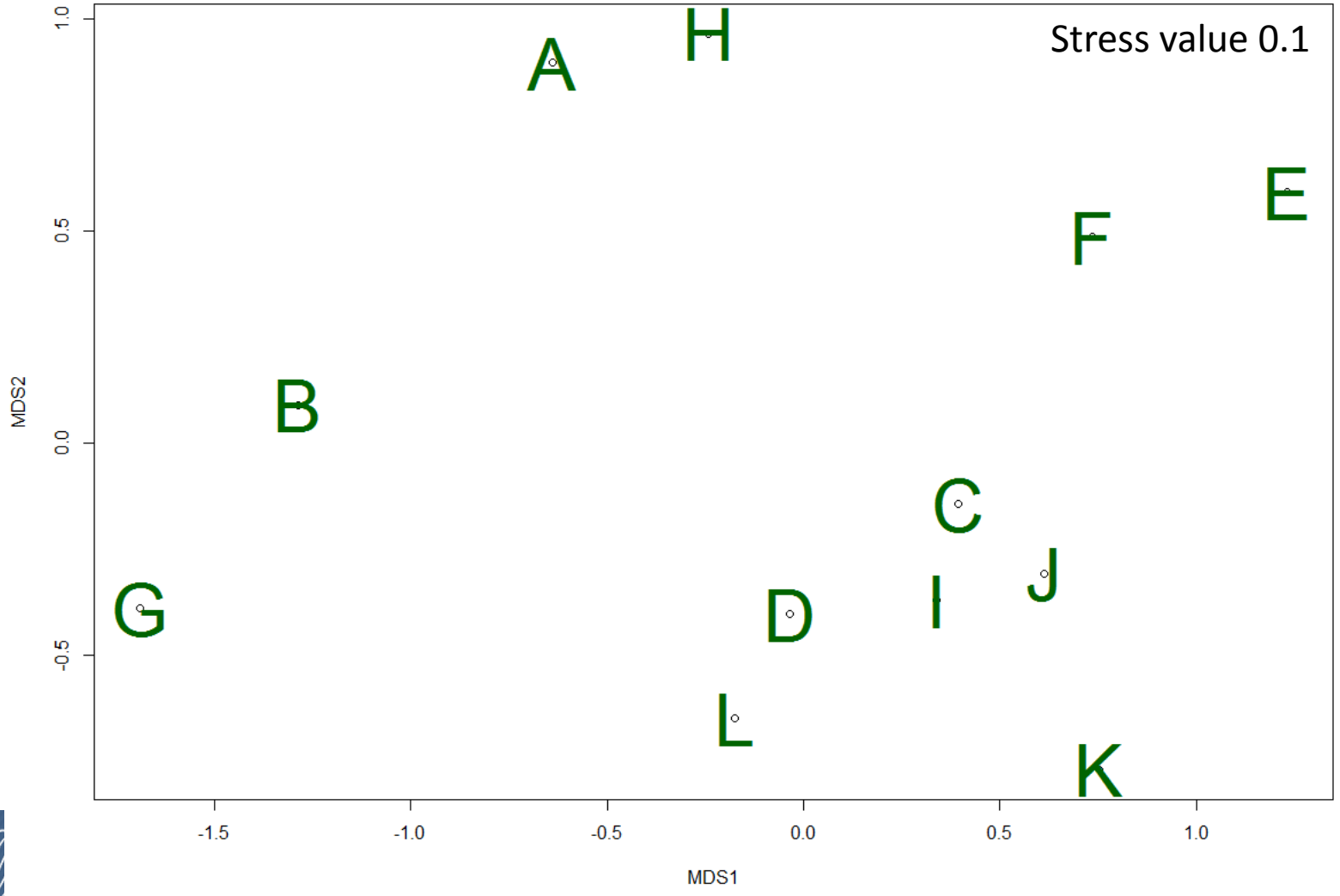
Transformed Data
For more equal weighting
Square root, fourth root, log(x+1), etc

Similarity matrix
Bray- Curtis (species counts), Gower (questionaires), Euclidean distance (env)

Hypothesis Testing
ANOSIM/PERMANOVA

Ordination
Represent ordering relationships between objects; iterative procedure; stress-value

NMDS example



Unconstrained Ordination

- Not hypothesis testing
- Exploratory or descriptive approach
- Overlay cluster groups on ordinations
- A posteriori comparison with environmental variables
  - Overlay environmental variables
  - Can identify possible relationships between env variables and species data

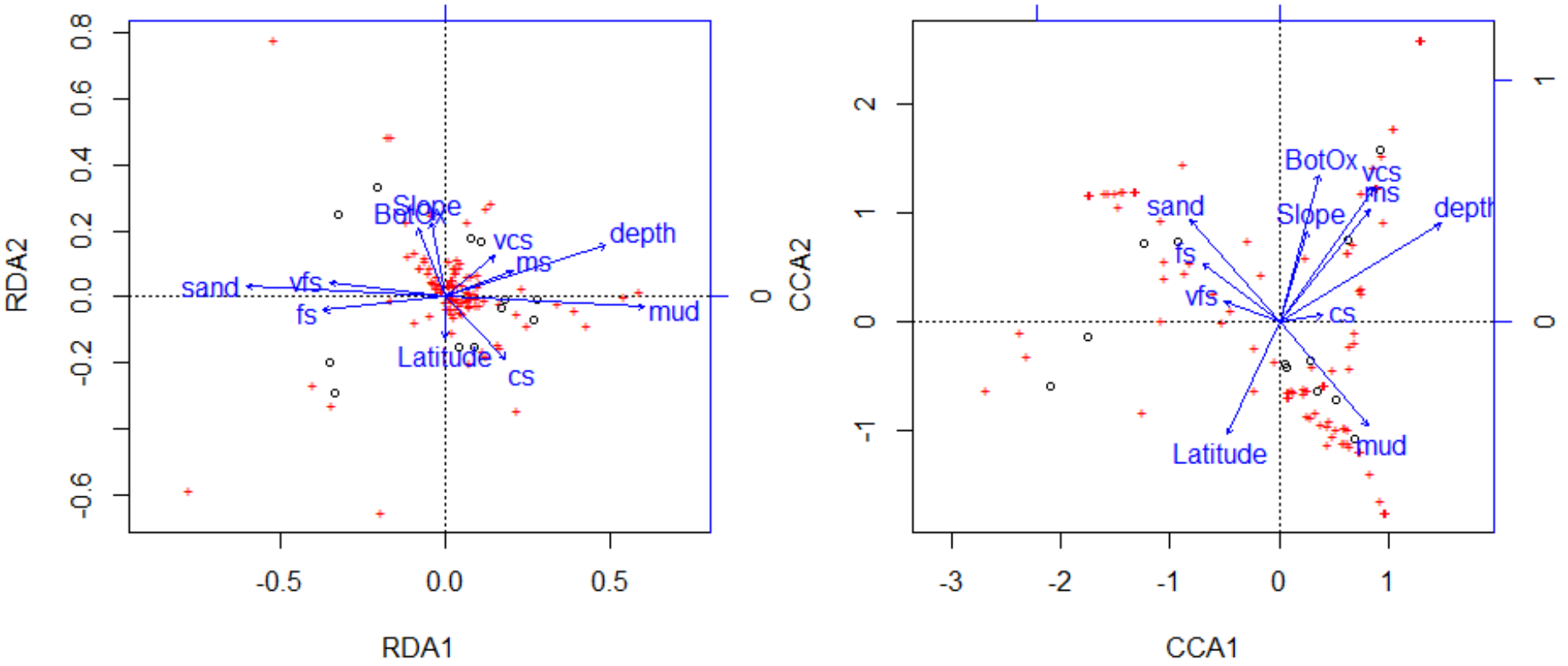
Constrained/Canonical Ordination

- Associates two or more datasets in the ordination process
- Can formally test hypotheses about relationships between the data sets

- Redundancy Analysis (RDA)
  - multivariate multiple linear regression followed by PCA of fitted values table

- Canonical Correspondence Analysis (CCA)
  - Combines multiple regression with CA

RDA and CCA examples



Triplot including sites, species and environmental variables

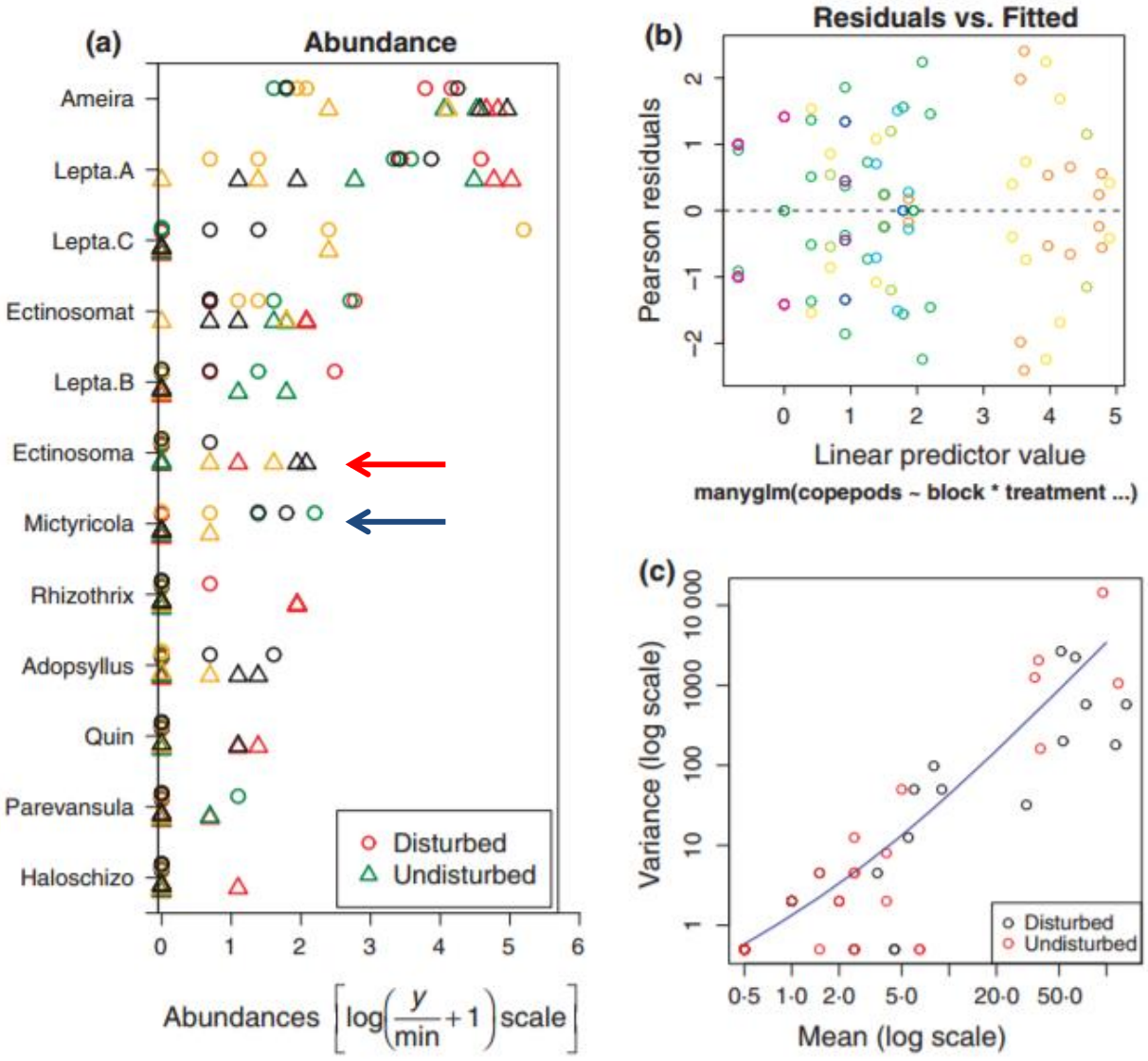
Model-based multivariate analyses

- Don’t reduce data to correlation or similarity matrix
- Can calculate AIC, plot residuals, etc
- Directly relates species and covariates in single analysis

Mvabund package

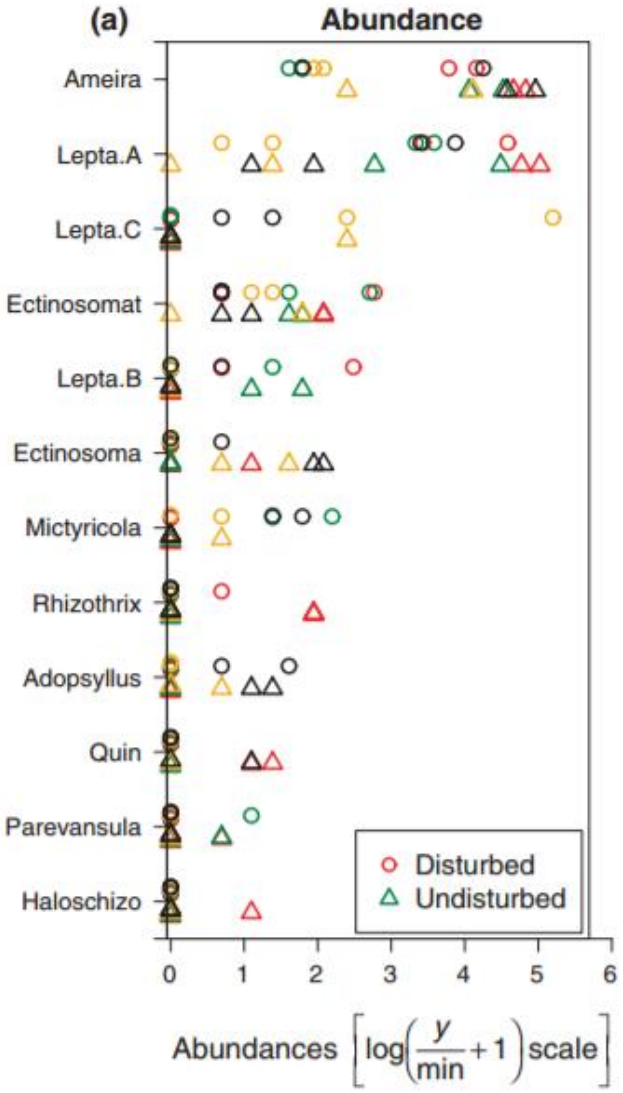
- Multivariate Linear and Generalized Linear Models
- Models can include both species and environmental data in single analysis
- Able to do hypothesis testing
- Data types
  - Species: Presence/absence, count, ordinal, biomass, percentage cover
  - environmental data

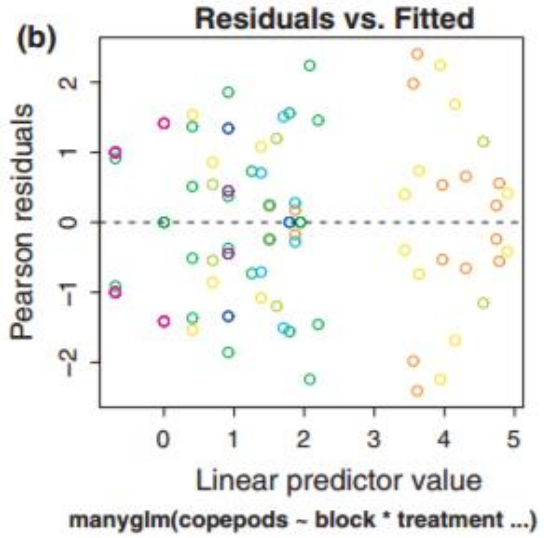
Wang et al. 2012. mvabund – an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution* 3: 471-474

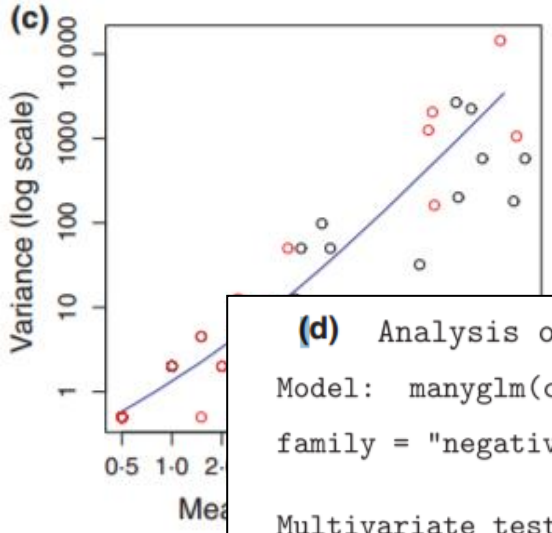


Example: copepods in disturbed and undisturbed sites

Wang et al. 2012. mvabund – an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution* 3:471–474







Example: copepods in disturbed and undisturbed sites

(d) Analysis of Deviance Table

Model: manyglm(copepods ~ block * treatment, family = "negative binomial")

Multivariate test:

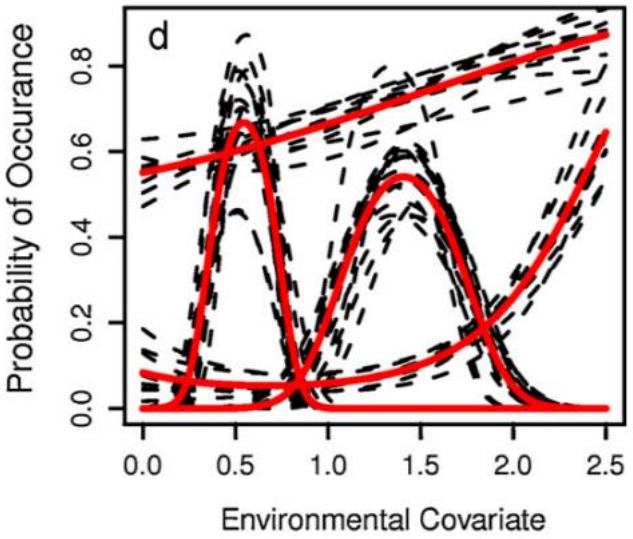
	Res.Df	Df.diff	Dev	Pr(>Dev)	
blocks	12	3	326.1	0.001	***
treatment	11	1	106.5	0.008	**
blocks:treatment	8	3	48.5	0.063	.

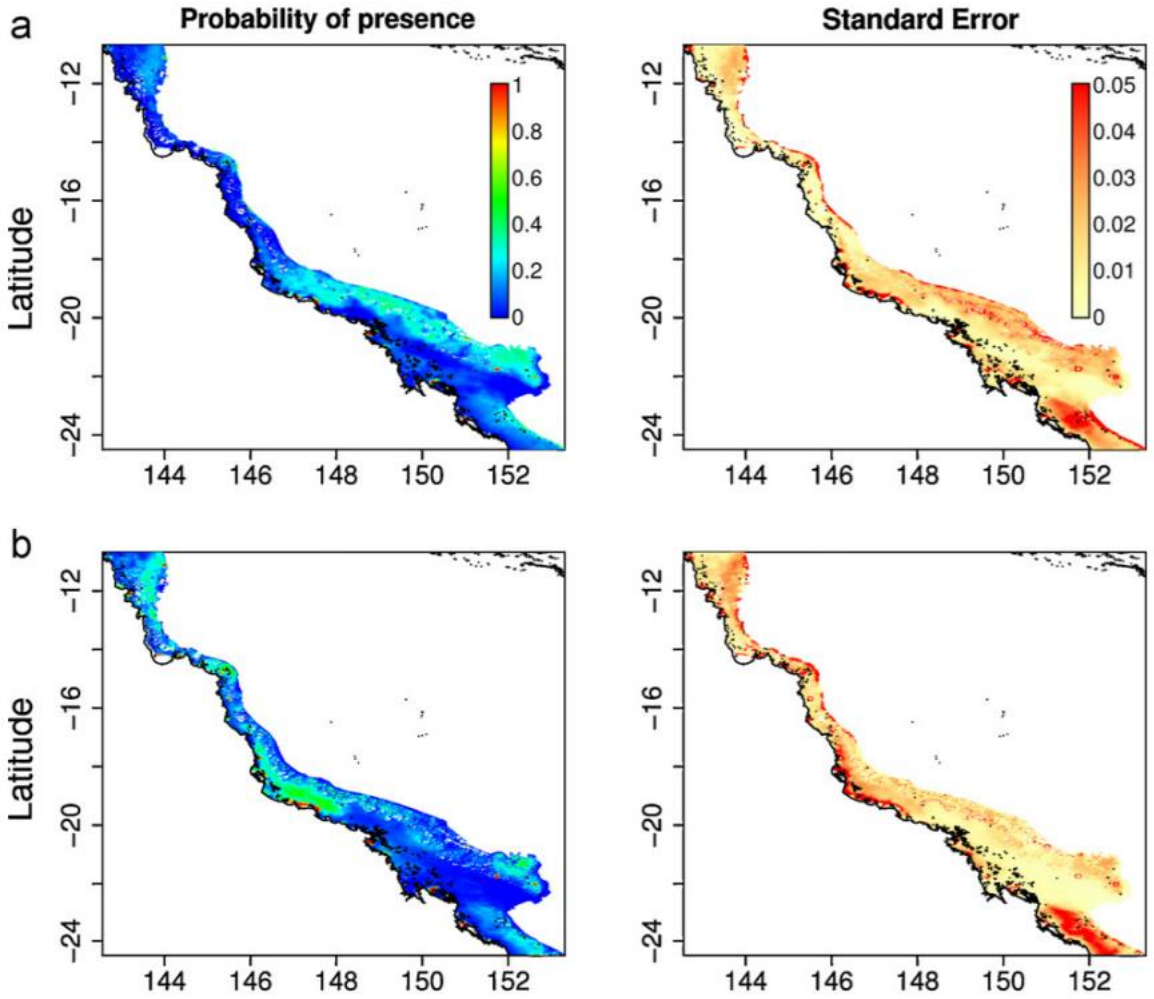
Wang et al. 2012. mvabund – an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution* 3:471–474

Speciesmix package

- Similar to a species distribution model for species archetypes
  - Groups species that have a similar occurrence across an environmental covariate into species archetypes with minimal info loss
  - Predicts species archetype occurrence based on GLMs
- Data types
  - Species: Presence/absence, count
  - environmental data

Dunstan et al. 2011. Model based grouping of species across environmental gradients. *Ecological Modelling* 222:955–963





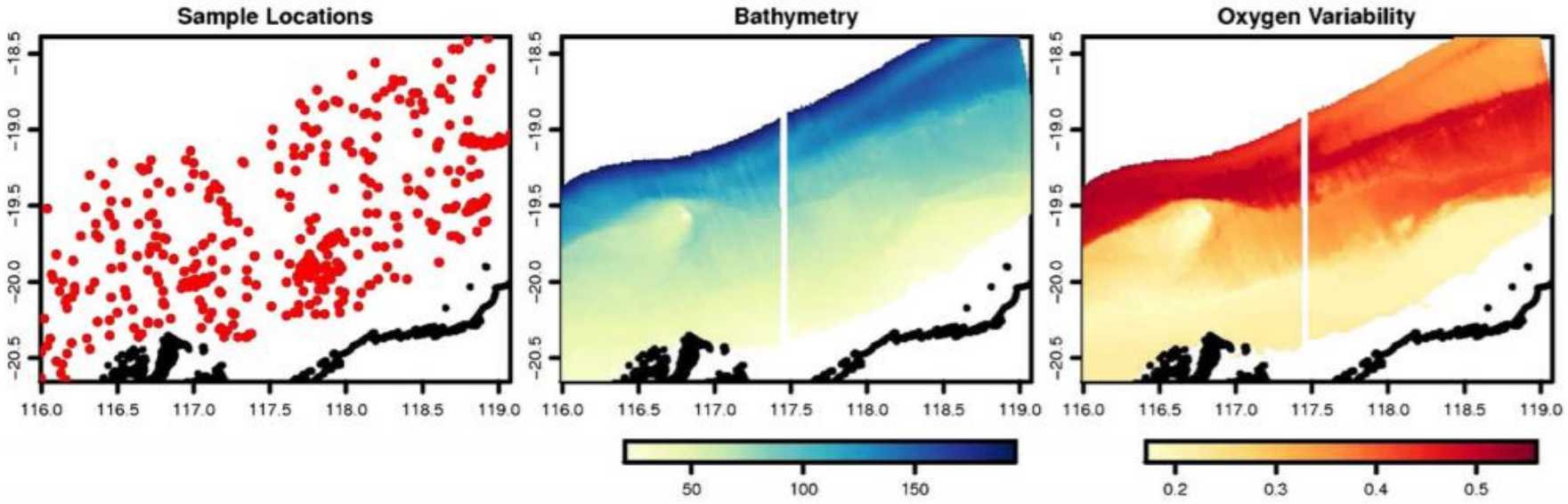
Dunstan et al. 2011. Model based grouping of species across environmental gradients. *Ecological Modelling* 222:955–963

RCPmod package

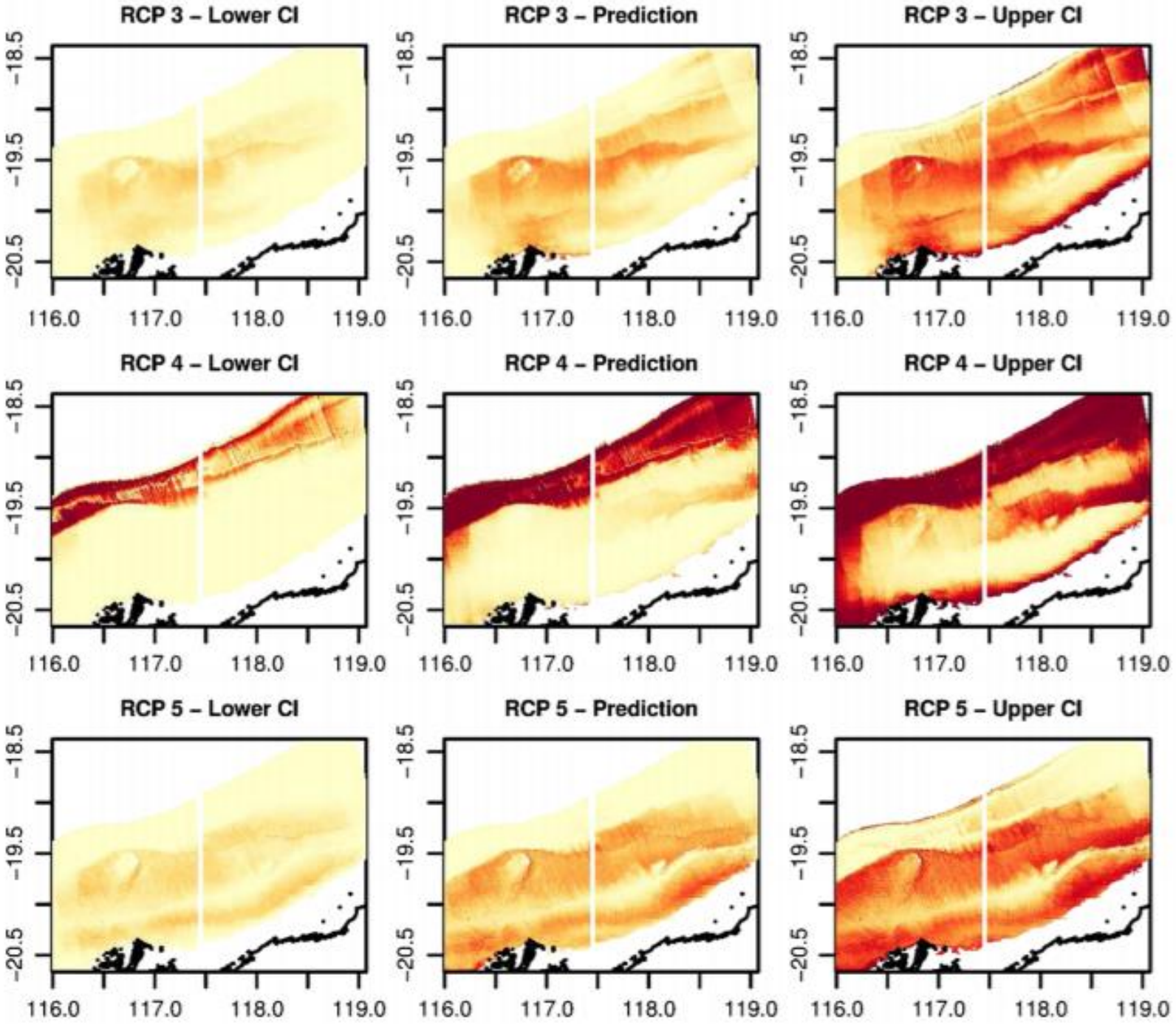
- Regions of common species catch probability profile; not occurrence
- Takes into account where species are caught and not caught
- RCP= single catch prob profile; discrete
- Data types:
  - Presence/absence, count, can include detection if replicates available

Foster et al. 2013. Modelling biological regions from multi-species and environmental data. *Environmetrics* 24: 489–499

E.g. Fish species of North West Shelf off western Australia

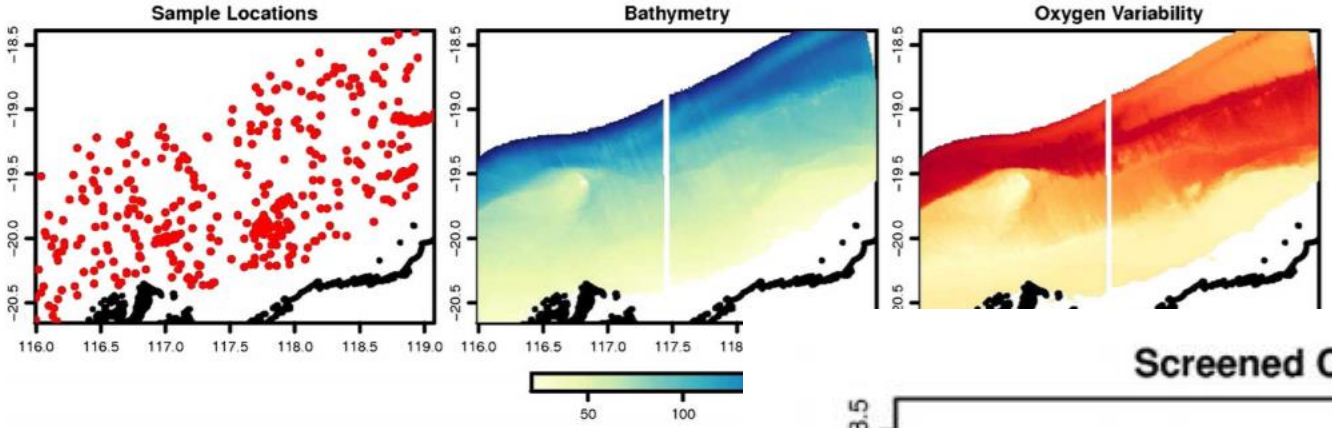


Foster et al. 2013. Modelling biological regions from multi-species and environmental data. *Environmetrics* 24: 489–499

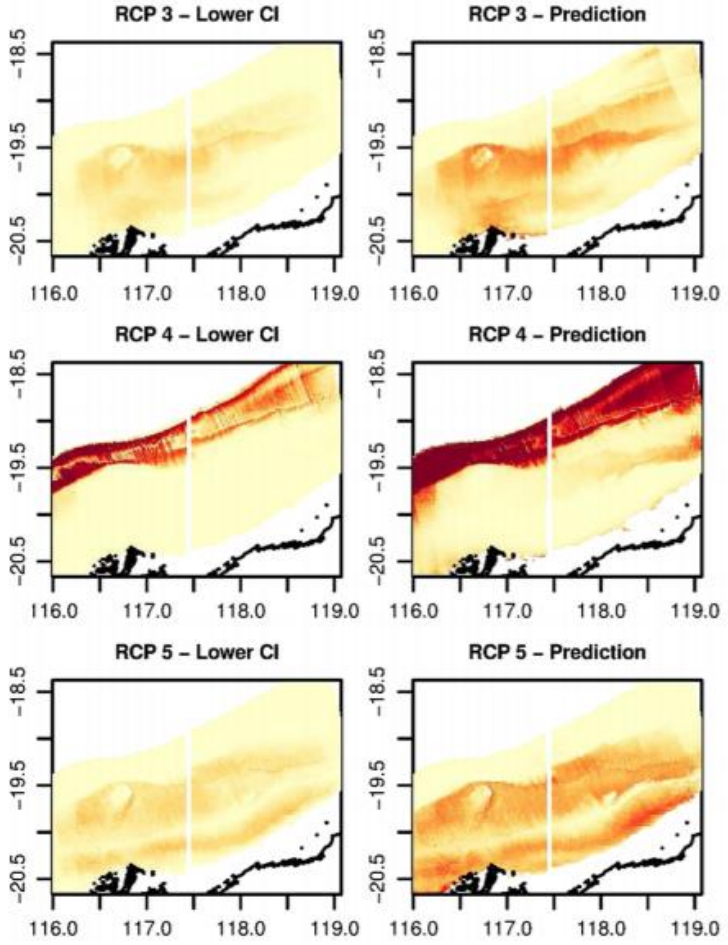


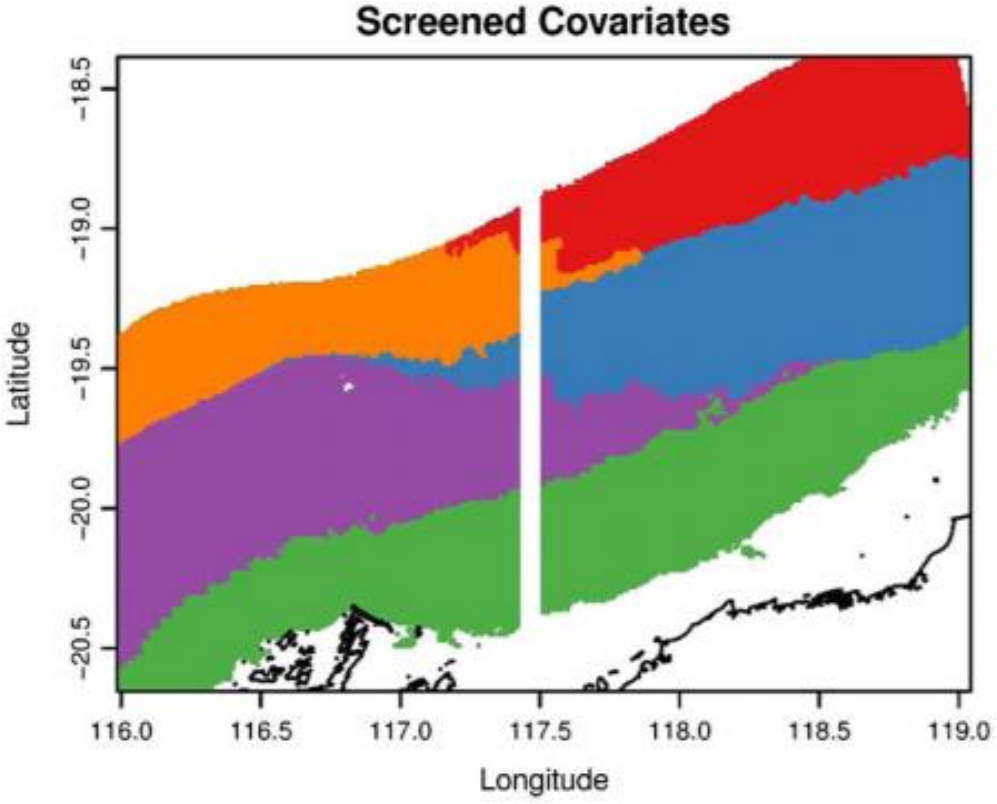
ies of North
f western

delling biological
-species and
Environmetrics
499



E.g. Fish species of North West Shelf off western Australia

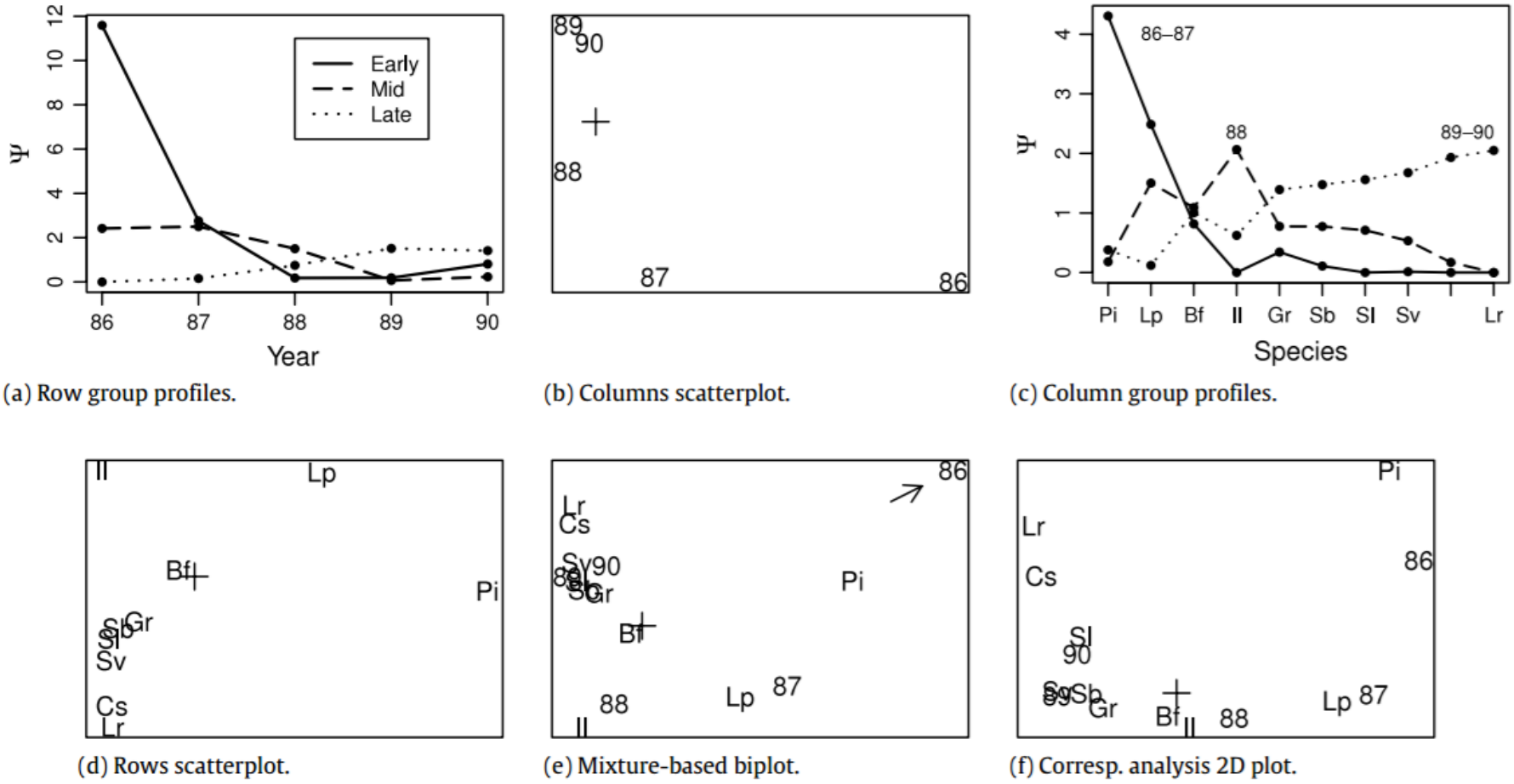




Foster et al. 2013. Modelling biological regions from multi-species and environmental data. *Environmetrics* 24: 489–499

ClustGLM package

- Clustered GLMs
  - Determines the best model-based clustering of data by sites or species or both
  - Can include site and species covariates e.g. altitude level or phylum
- Count data
- Current limitation: mostly common species
- Can produce ordinations
- http://homepages.ecs.vuw.ac.nz/~shirley/



(a) Row group profiles. (b) Columns scatterplot. (c) Column group profiles.

(d) Rows scatterplot. (e) Mixture-based biplot. (f) Corresp. analysis 2D plot.

Fig. 5. Plots from Liphook Forest funghi data. Plots (a) and (b) are from model $\{rR3,\ cp,\ C,\ PD\}$, with species groups Δ = early, $*$ = mid and $\bullet$ = late. Plots (c) and (d) are from model $\{rn,\ cC3,\ C,\ PD\}$, with year groups $\Delta = \{86, 87\}$, $* = \{88\}$ and $\bullet = \{89, 90\}$. The mixture-based biplot is in (e) and (f) is the traditional correspondence analysis 2D plot for comparison. Plots (b), (d) and (e) do not show the triangle vertices, which are outside the plotted region. The centroids are marked +, and the arrow in the biplot indicates that Year 86 is an outlier with true position twice the plotted distance from the centroid.

Example: Toadstool species in a New Zealand Forest

Pledger and Arnold 2014. Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection . *Computational Statistics and Data Analysis* 71:241–261

Resources

- Borcard, D., Gillet, F. and Legendre, P. 2011. Numerical Ecology with R. Springer Science And Business, LLC, New York, pp 306
- Manly, B.F.J. 1986. Multivariate Statistical Methods: A Primer. Chapman and Hall Ltd, London, pp 159
- Legendre, P. and Legendre, L. 2012. Numerical Ecology (Third ed). Elsevier, Amsterdam, pp 990
- http://www.multivariatestatistics.org/ (correspondence analyses)
- http://environmentalcomputing.net/introduction-to-mvabund/
- https://cran.r-project.org/web/packages/mvabund/mvabund.pdf
- https://cran.r-project.org/web/packages/RCPmod/RCPmod.pdf
- https://cran.r-project.org/web/packages/SpeciesMix/SpeciesMix.pdf
- http://homepages.ecs.vuw.ac.nz/~shirley/
- Greenacre, M. 1984. Theory and Applications of Correspondence Analysis. Academic Press Limited, London, pp364 (http://www.carme-n.org/?sec=books5)

SEEC Stats Toolbox

Want to broaden your stats knowledge? Unsure of what you can do with your data? Still developing your proposal?

Join us for the monthly **SEEC Stats Toolbox** seminars where we introduce you to statistical methods that are useful for ecologists, environmental and conservation scientists.

Our next seminar:

Topic: **Spatial Capture-Recapture Models**
Who: Dr Greg Distiller
When: **Thursday 25 May 2017 (1-2pm)**
Where: PD Hahn Lecture Theatre 3,
PD Hahn Building Level 5, UCT

More details: www.seec.uct.ca.za

SEEC - Statistics in Ecology, Environment and Conservation