ECOLOGICAL DATA SOLUTIONS

Data Visualization & Dimensionality Reduction Techniques

Presented by Dr. Ariol RAMA, Biologist

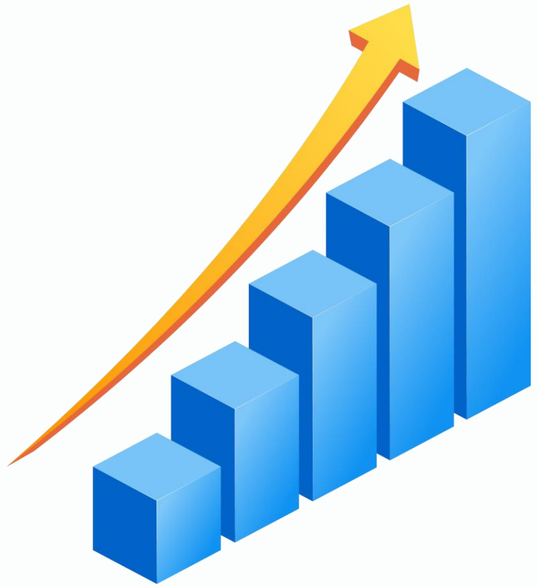
UNIVERSITETI I TIRANËS

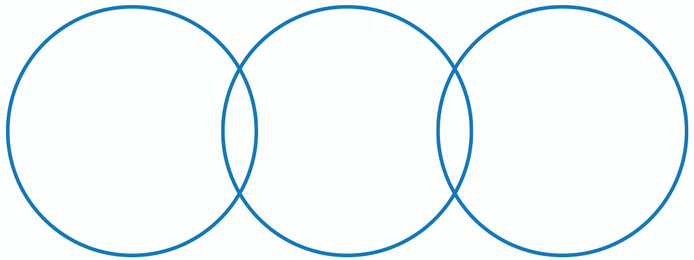
“We are blind to the obvious, and we are also blind to our blindness.”

Daniel Kahneman

Objectives

Key goals for understanding data visualization techniques

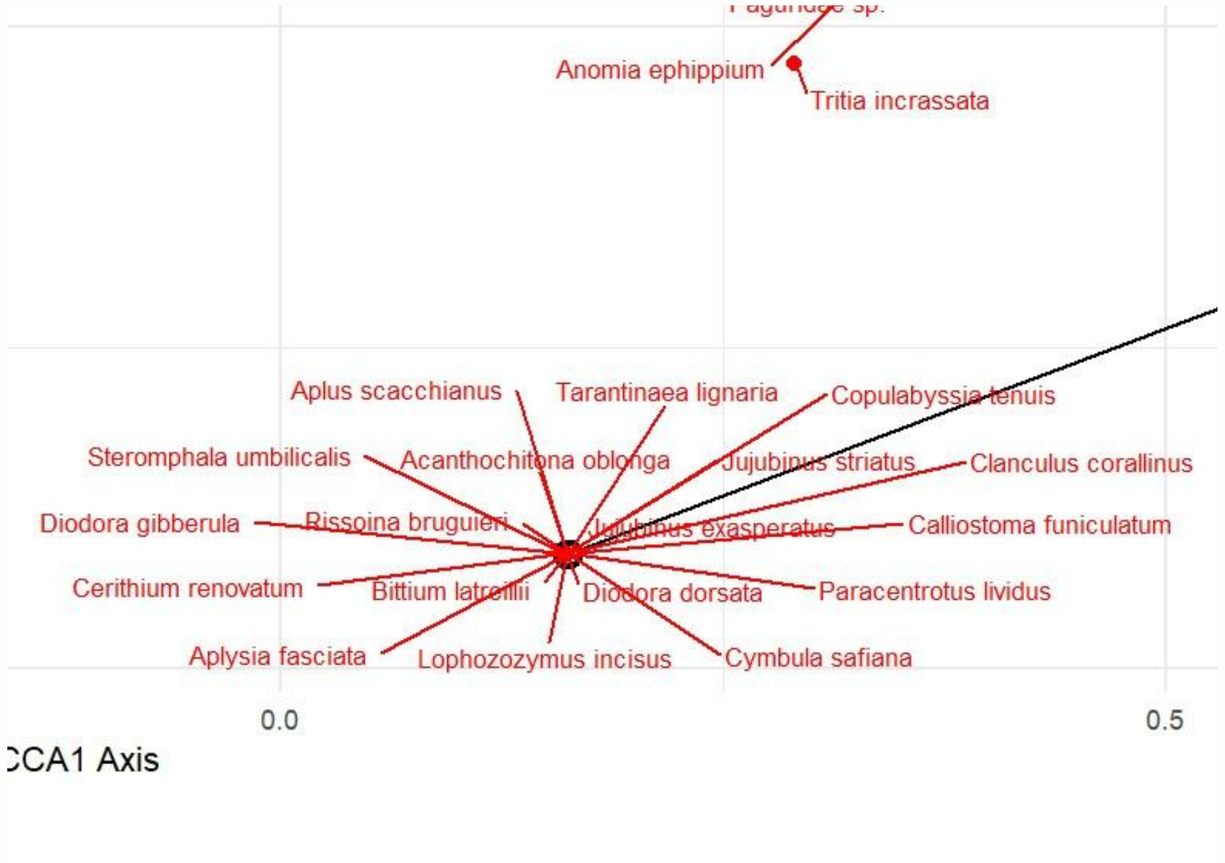




- What dimensionality reduction solves
- Why visualization is essential in modern data analysis
- Key techniques: NMDS, CCA, PCA
- Interpretation and scientific reasoning
- Limitations of classical (univariate) statistics

The Problem: Ecology = High-Dimensional Data

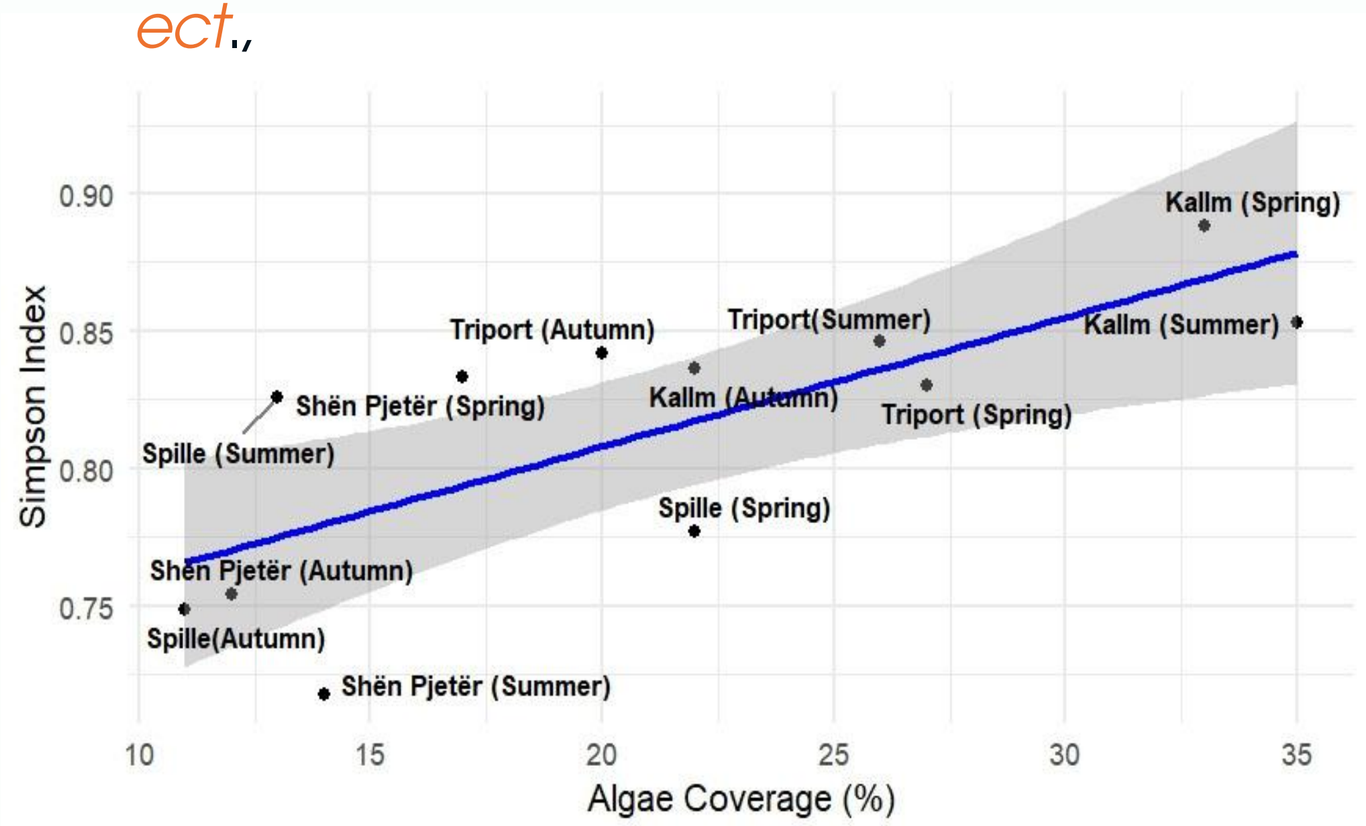
- Non-linear relationships
- Multiple environmental variables
- Dozens to hundreds of species
- Classical statistics become blind to structure
- Zero-inflated and overdispersed data



Why “Classical” Statistical Tests Alone Are Not Enough?

Classical approaches (t-tests, ANOVA, correlations, LM ect.,

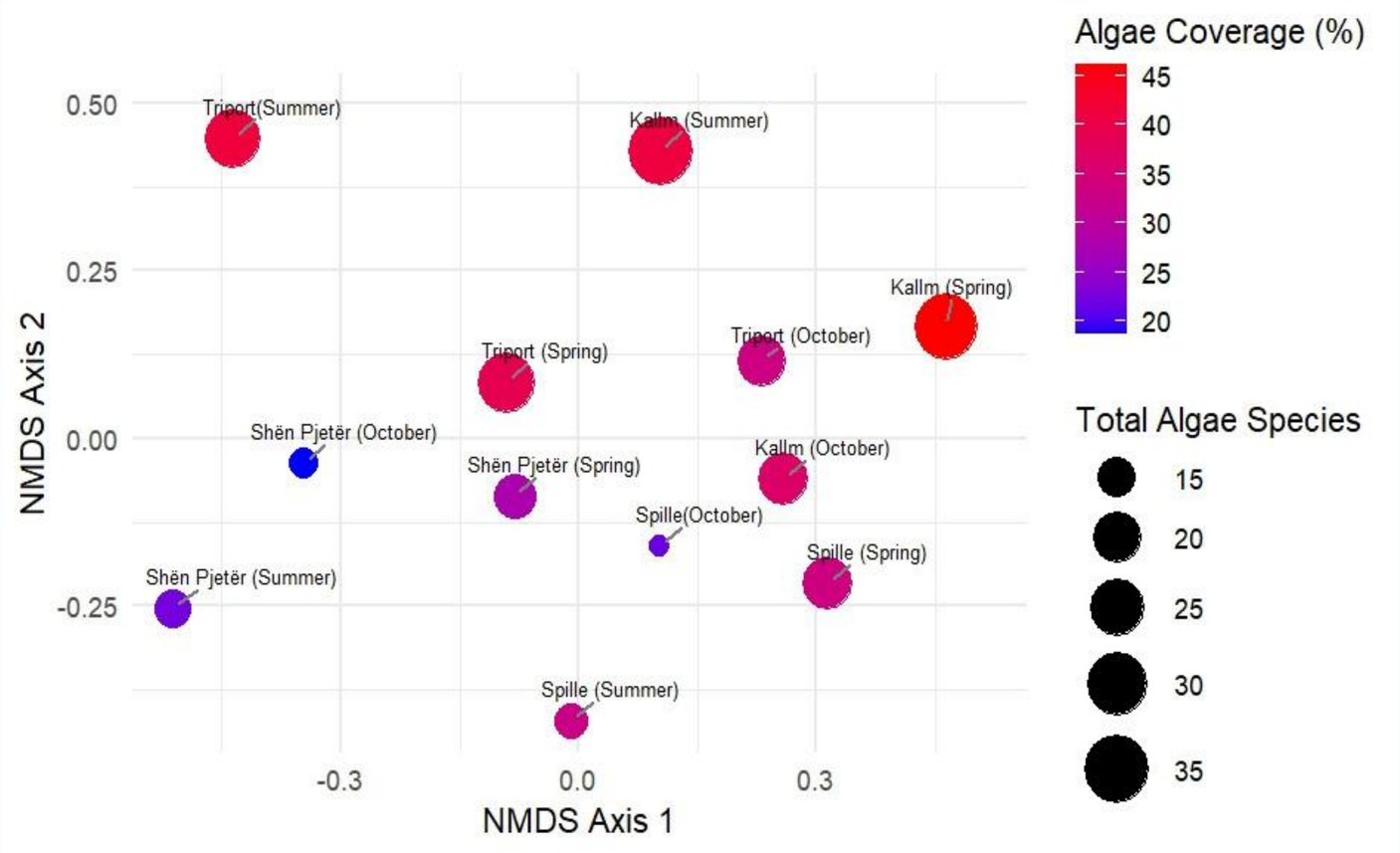
- Usually only test single variables at a time
- Cannot reveal multivariate patterns
- Assume linearity & normality
- Lose ecological meaning when species interact
- Fail with Bray–Curtis distance, presence/absence,



WHY WE NEED DIMENSIONALITY REDUCTION?

Understanding the weaknesses of traditional statistical methods

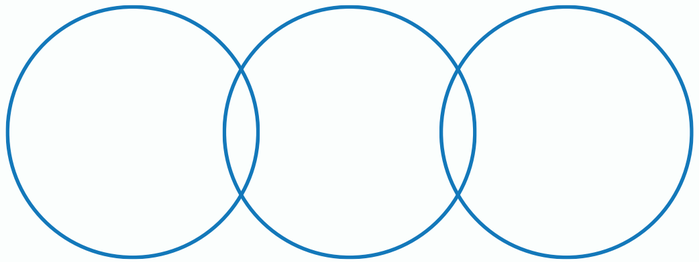
- Mathematical reduction of complex multivariate datasets
- From 100+ species → 2 gradient axes
- Allows visual interpretation of patterns
- Preserves relationships & ecological structure
- Converts dissimilarities into spatial distances

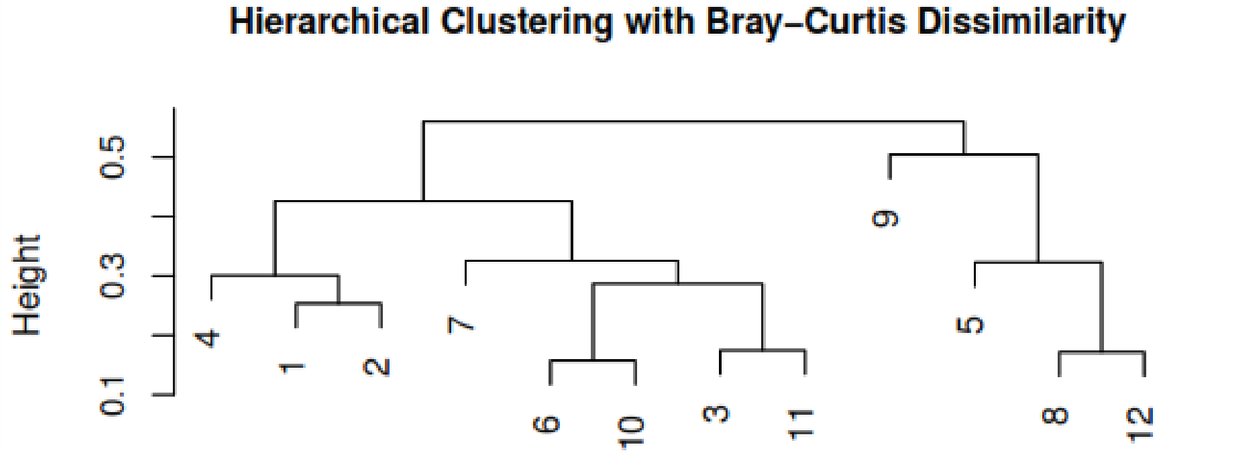


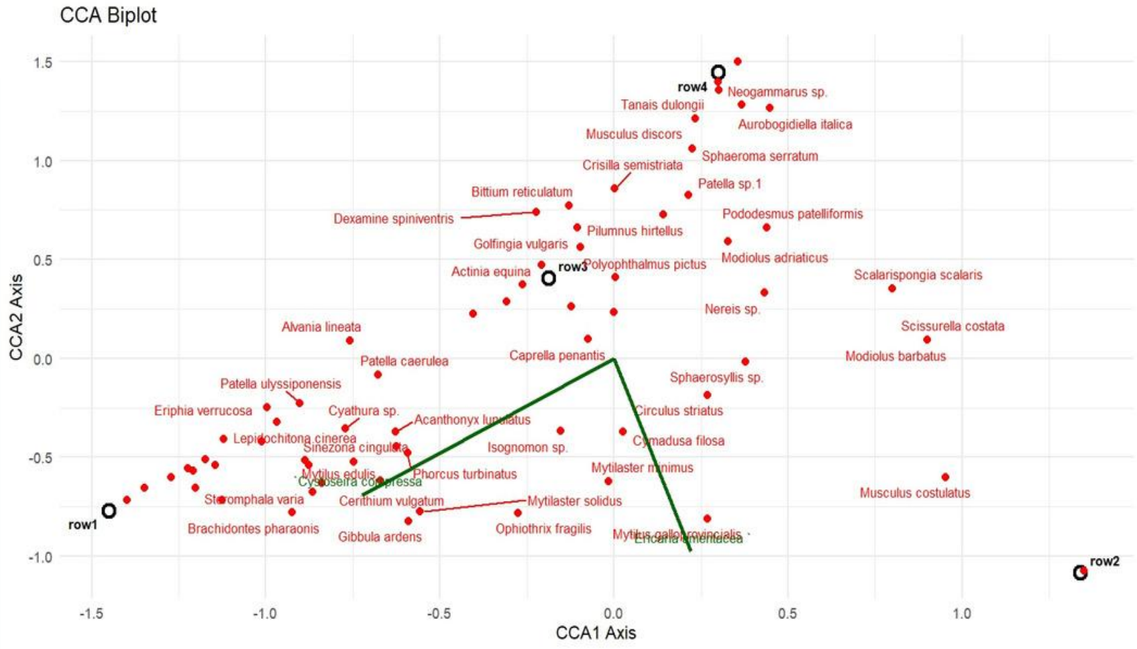
Why Ecologists Depend on It ?

Understanding its role in ecological data analysis: Because such datasets have

- Have high species richness
- Are compositional (relative abundances)
- Contain nonlinear gradients
- Need distance-based approaches (Bray-Curtis, Jaccard)



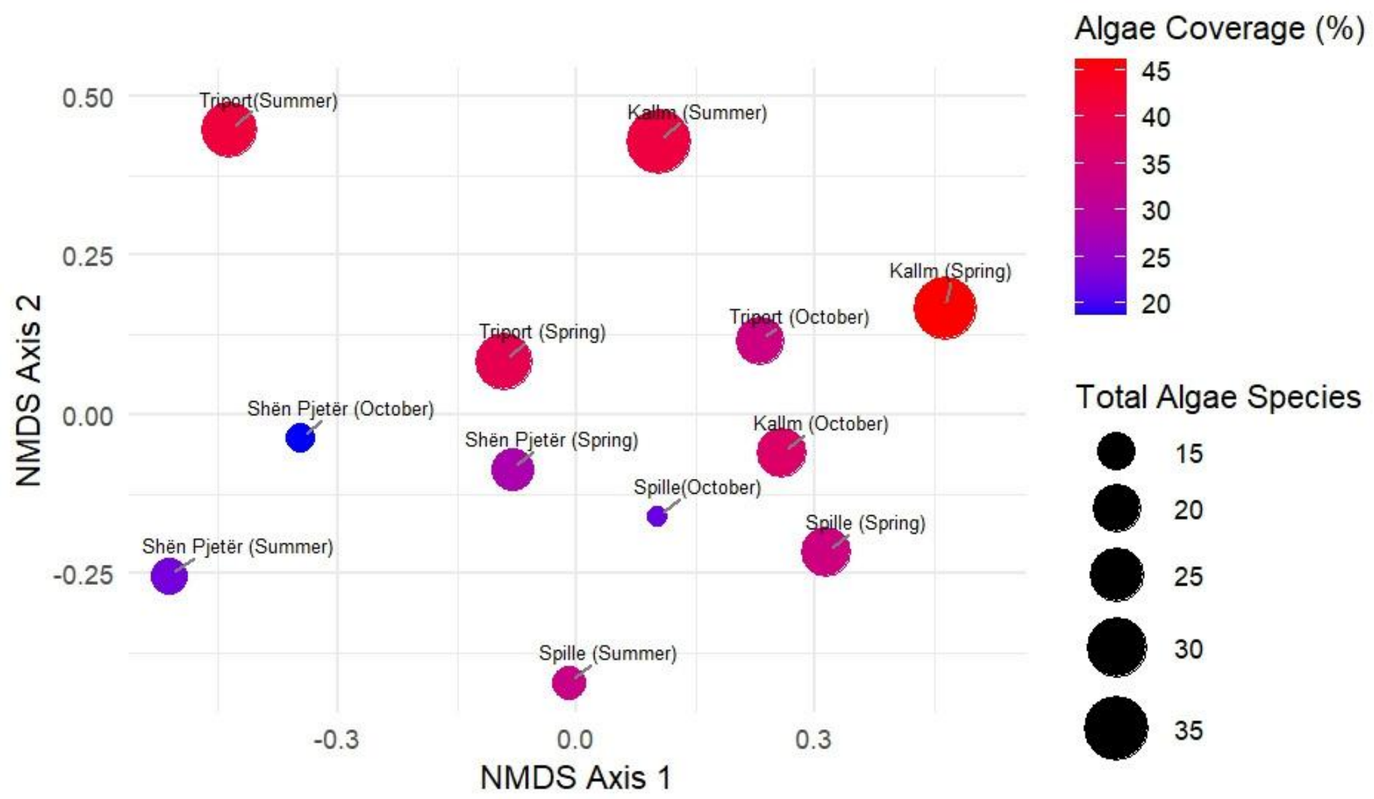




What Dimensionality Reduction Reveals?

Essential methods for effective data representation in ecology

- Similarity among sites
- Gradients in species composition
- Clusters (seasonal, spatial)
- Influence of environmental variables
- Outliers and rare species effects



Classical Statistics vs Multivariate Ordination

Understanding Non-metric Multidimensional Scaling in Ecology

Classical Stats	Dimensionality Reduction
One variable at a time	All variables simultaneously
Assume normality	Non-parametric or distance-based
Weak at detecting structure	Highlights gradients, clusters
Linear thinking	Captures nonlinear relationships
Cannot show spatial/ecological similarity	Shows similarity visually

Example

Understanding the limits

	Stacionet/stinet					
Familjet	Shën Pjetër (Spring)	Kallm (Spring)	Spille (Spring)	Triport (Spring)	Shën Pjetër (Summer)	Kallm (Summer)
Ianthellidae	0	0	0	0	0	0
Actiniidae	220	237	11	59	60	105
Hormathiidae	0	237	0	59	0	0
Enoplidae	1	237	0	59	1	0
Echiuridae	0	237	0	59	0	0
Chitonidae	4	237	59	59	3	1
Acanthochitonidae	1	237	0	59	0	1
Leptochitonidae	0	237		0	0	0

Example Showing Why LM Fails

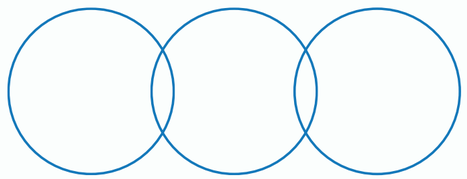
Terminology

Shannon Diversity Index (H'), is a quantitative measure of community diversity that incorporates both species richness and species evenness

An LM model inrefers to a Linear Model, the simplest and most classical regression method.
It assumes a linear relationship between one response variable and one or more predictor variables.

The task

The linear regression model is used to explore the relationship between the Shannon Diversity Index (ShannonDiversity) and two predictors: **Algae Coverage** and **Total Algae Species**



LM: ShannonDiversity ~ AlgaeCoverage + TotalAlgaeSpecies

Example Showing Why LM Fails

LM: ShannonDiversity ~ AlgaeCoverage + TotalAlgaeSpecies

Coefficients:			
Estimate Std. Error t value Pr(>|t|)			
(Intercept) 1.526557 0.152578 10.005 ***3.56e-06 ****			
AlgaeCoverage 0.015659 0.009269 1.689 ***0.125***			
TotalAlgaeSpecies 0.007394 0.009931 0.745 ***0.476***			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			

- **Model significant, predictors non-significant**
- **Interpretation ambiguous**
- **Reason: Shannon responds to multivariate forces, not single variables.**

When GLM Works but Still Falls Short

Terminology

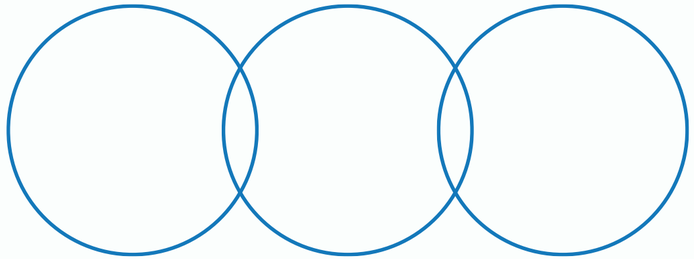
Akaike Information Criterion is used to compare models, with lower values indicating a better model fit, relative to the complexity of the model.

Generalized Linear Models (GLM): extends standard linear regression to handle response variables that aren't normally distributed, unifying models like logistic and Poisson regression

The task

GLM with a Poisson distribution is used to show the relationship between ***species abundance*** and ***Algae Coverage***.

```
glm(formula = SpeciesAbundance ~ AlgaeCoverage, family = "poisson",
data = data)
```



When GLM Works but Still Falls Short

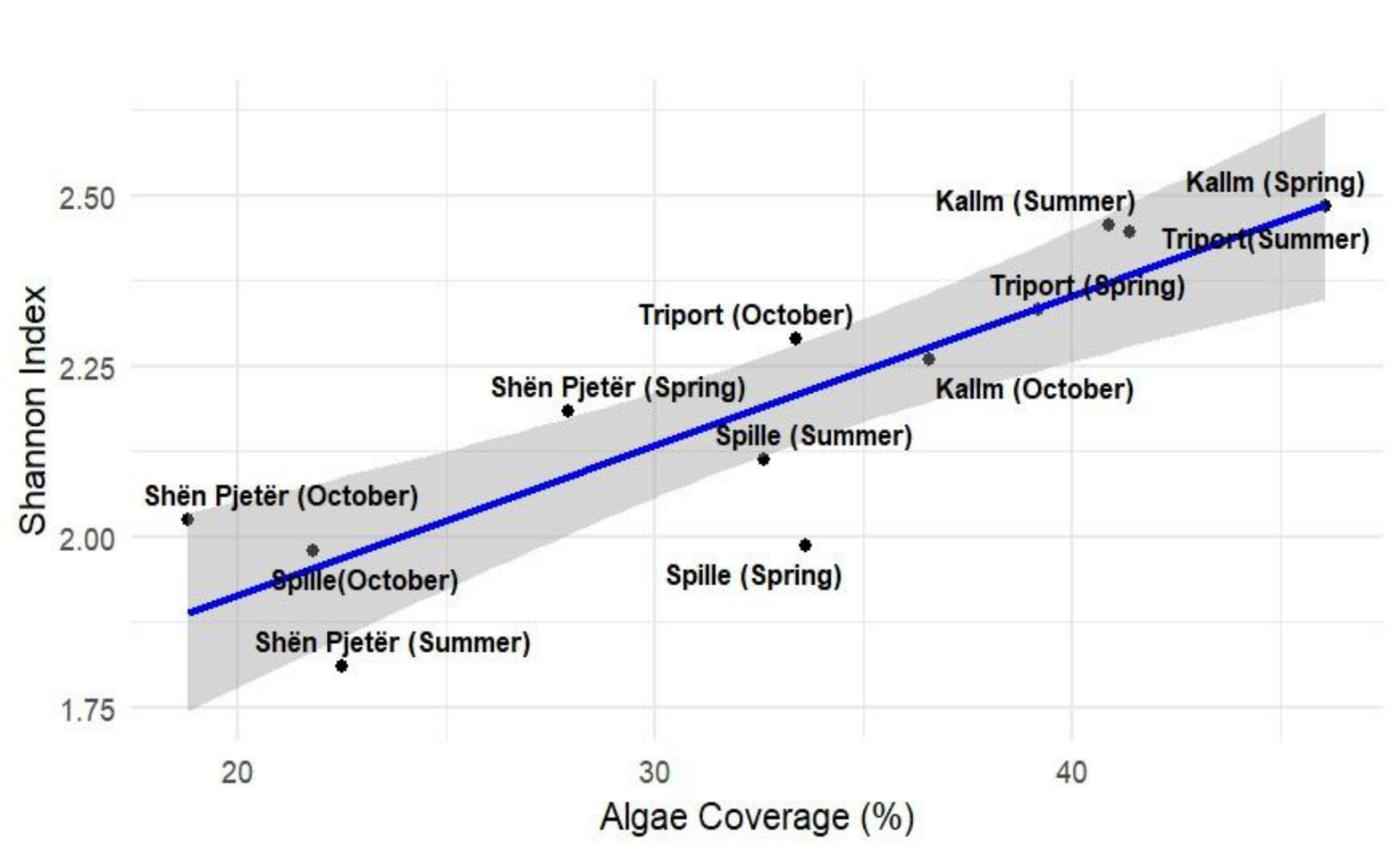
glm(formula = SpeciesAbundance ~ AlgaeCoverage, family = "poisson", data = data)

```
glm(formula = SpeciesAbundance ~ AlgaeCoverage, family = "poisson",
data = data)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 6.929152 0.035171 197.014 < 2e-16 ***
AlgaeCoverage 0.003887 0.001029 3.776 0.000159 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 529.12 on 11 degrees of freedom
Residual deviance: 514.81 on 10 degrees of freedom
AIC: 625.34- inf
```

- **Indicates positive effect of algae coverage**
- **Yet AIC = Inf, signs of overdispersion**
- **Evidence: Even GLM cannot capture community structure → need ordination.**

DATA VISUALIZATION BASICS

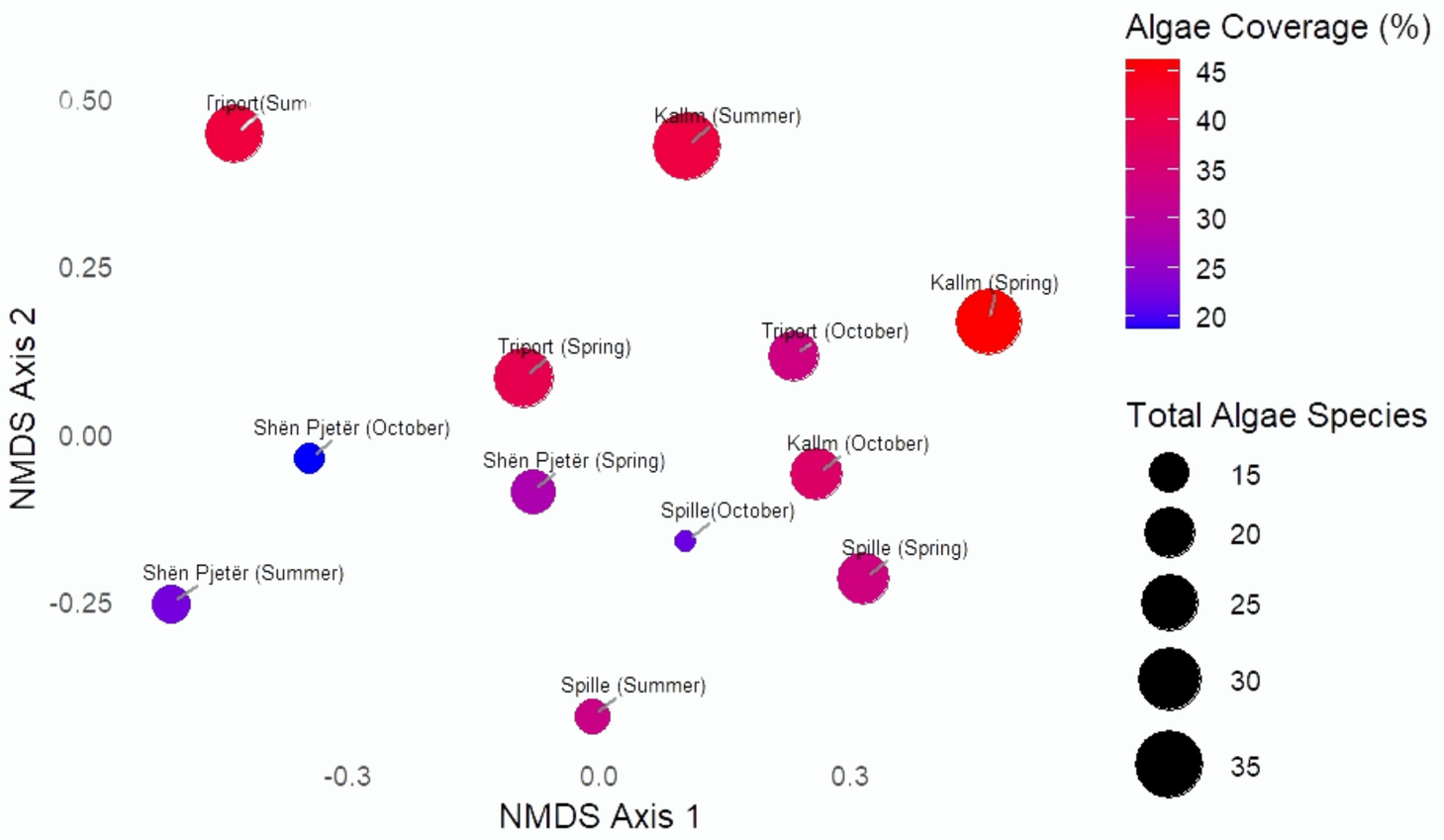
Example: Shannon vs Algae Coverage



- **No strong linear relationship found…"**
- **But ordination** (high-dimensional data (like species counts across many sites) into a lower-dimensional spac) **later reveals deeper patterns.**

NMDS

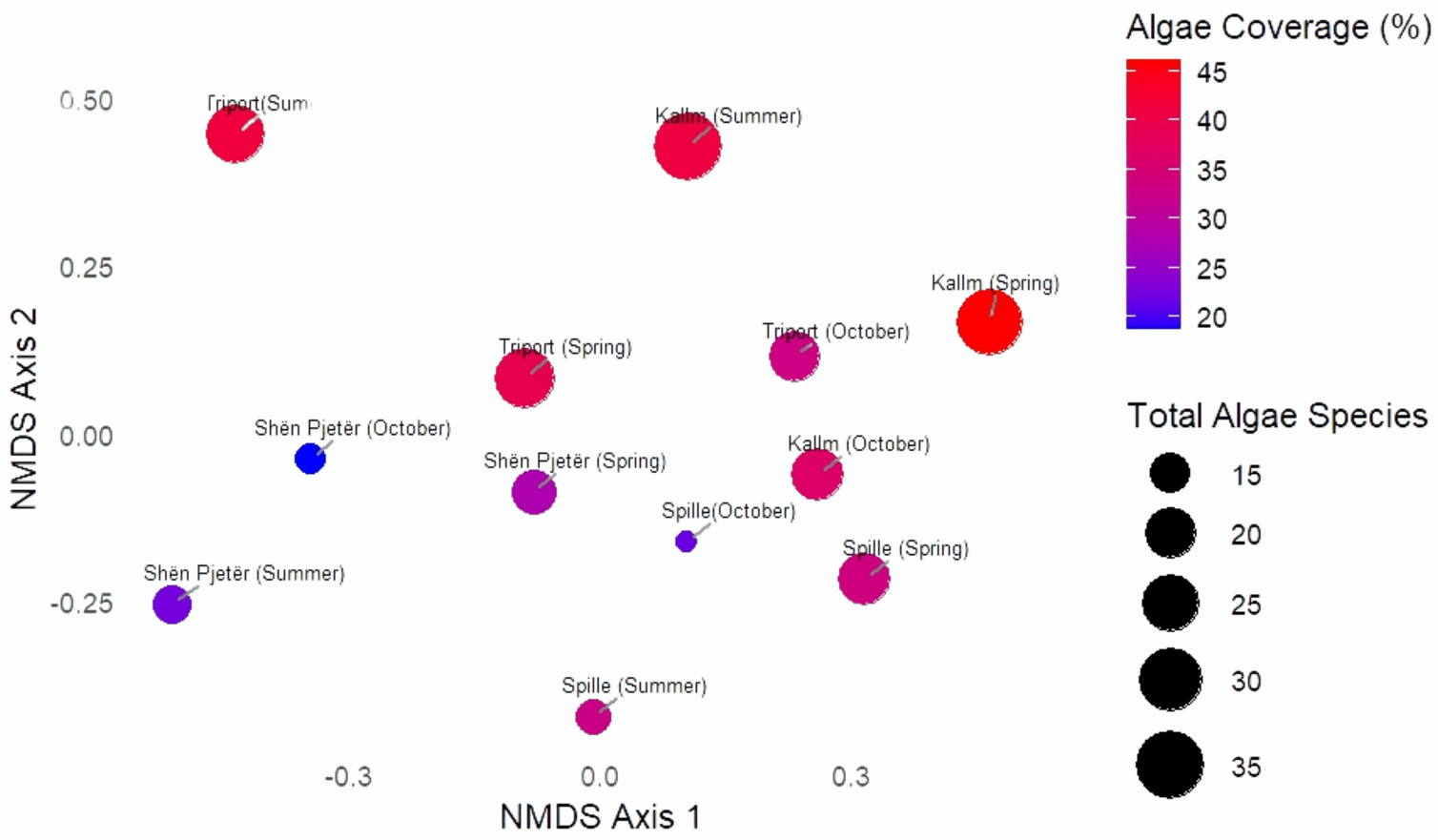
Non-metric Multidimensional Scaling



- **Reduces data using ranked ecological distances**
- **Preserves structure even with many zeroes**
- **Ideal for species abundance data**

How to Read NMDS Axes

Non-metric Multidimensional Scaling



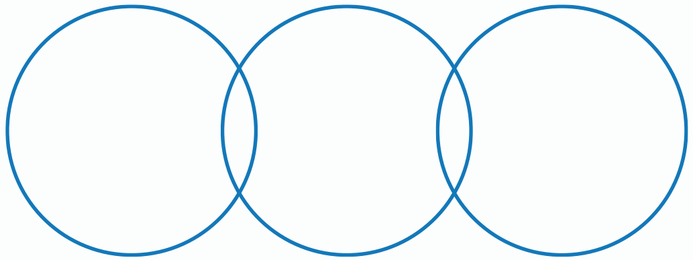
- **"NMDS axes don't have direct biological meaning… only represent maximum variation while maintaining rank order…"**
- **NMDS1 = strongest gradient**
- **NMDS2 = secondary gradient**
- **Closer points = similar communities.**
- **Spring sites cluster**
- **October samples differ**

Why NMDS Is Indispensable

Non-metric Multidimensional Scaling

RECALL!

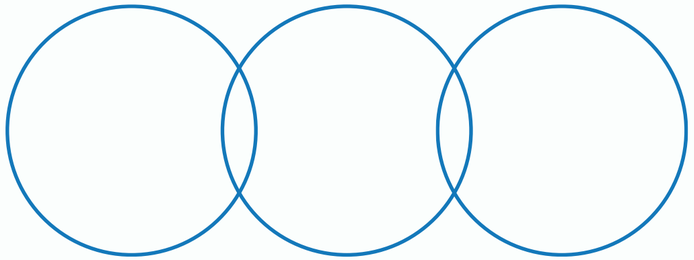
- **Works with non-normal data;**
- **Works with zero-inflated datasets**
- **Works with nonlinear ecological gradients**
- **Reveals patterns classical stats cannot detect**



What can't NMDS Tell Us??

- NMDS is an unconstrained ordination method used to visualize patterns in a single dataset (e.g., species community data) without making assumptions about data distribution (it is non-parametric). It arranges samples so that their distances in the ordination plot best match their rank dissimilarities in the original data.

  - HENCE..... If we needed to know how specific environmental variables influence community composition WE need to move on to CCA analysis

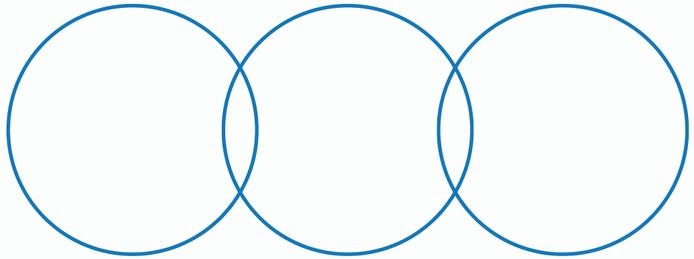


What Is CCA

Canonical Correspondence Analysis

CCA is a *constrained* ordination method that explicitly incorporates a second matrix of environmental variables to model species distribution directly. It aims to extract synthetic environmental gradients that best explain the variation in species composition data.

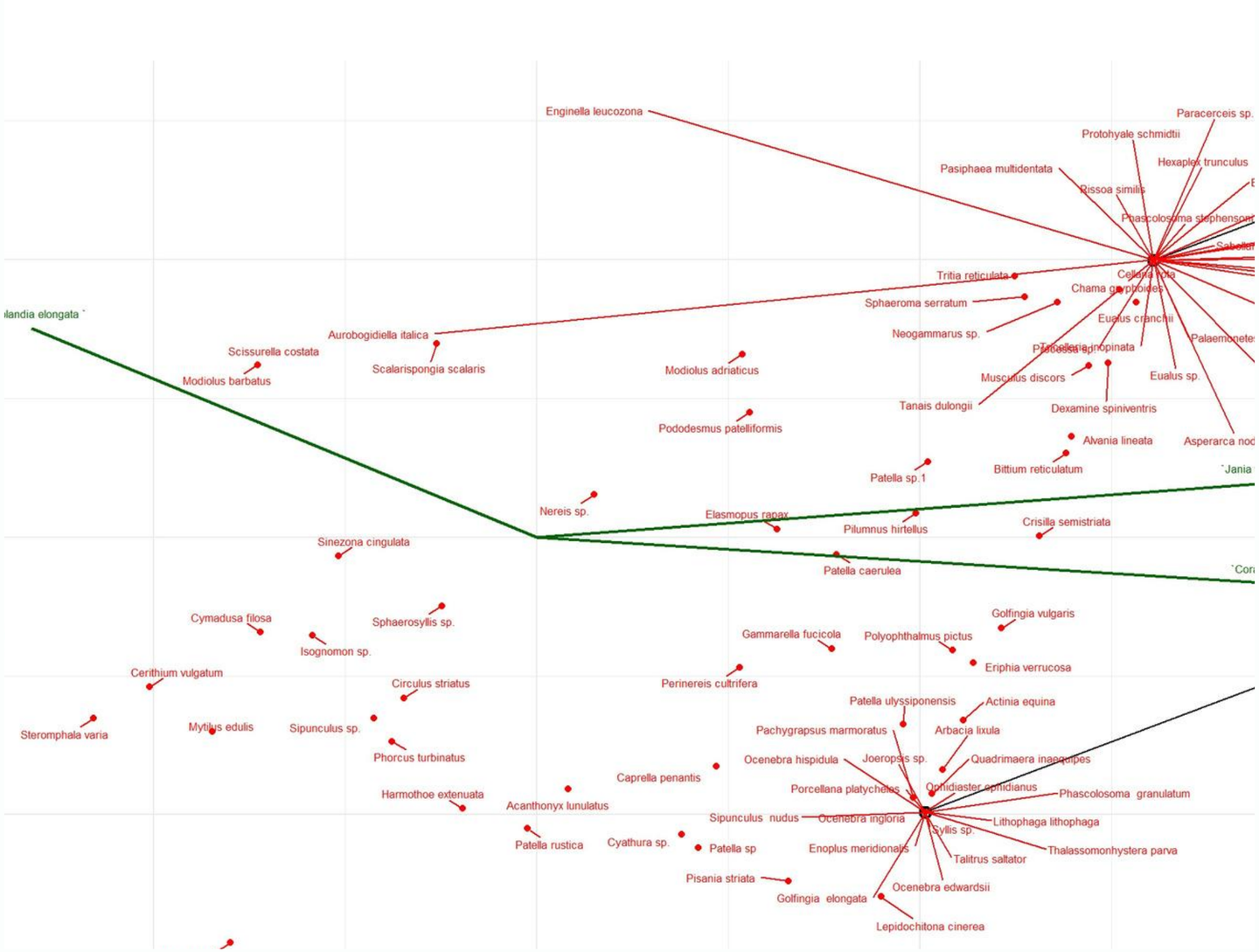
For example: To quantify the influence of specific environmental factors (e.g., pH, nitrogen levels, temperature) on community structure



How to Interpret CCA Vectors

Canonical Correspondence Analysis

- **Arrows show direction & strength of environmental variables**
- **Species near arrow are strongly influenced**
- **Sites align along environmental gradients**



CCA vs NMDS

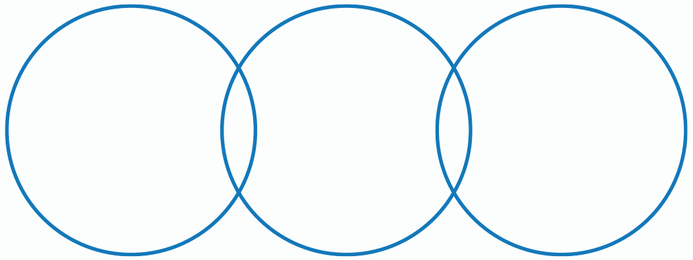
NMDS	CCA
Unconstrained	Constrained by environment
Patterns only	Patterns + environmental drivers
Exploratory	Hypothesis-driven
Good first step	Good explanatory step

Additional Resources

Explore valuable references for improved understanding and application

The Vegan package Vegan: an introduction to ordination by Jari Oksanen provides comprehensive guidelines for ecological analysis and visualization techniques in R programming.

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology, 18(1), 117–143. Key literature on ordination methods enhances understanding of ecological data analysis and interpretation in various contexts.



Do you have questions?

Feel free to ask anything about the presentation