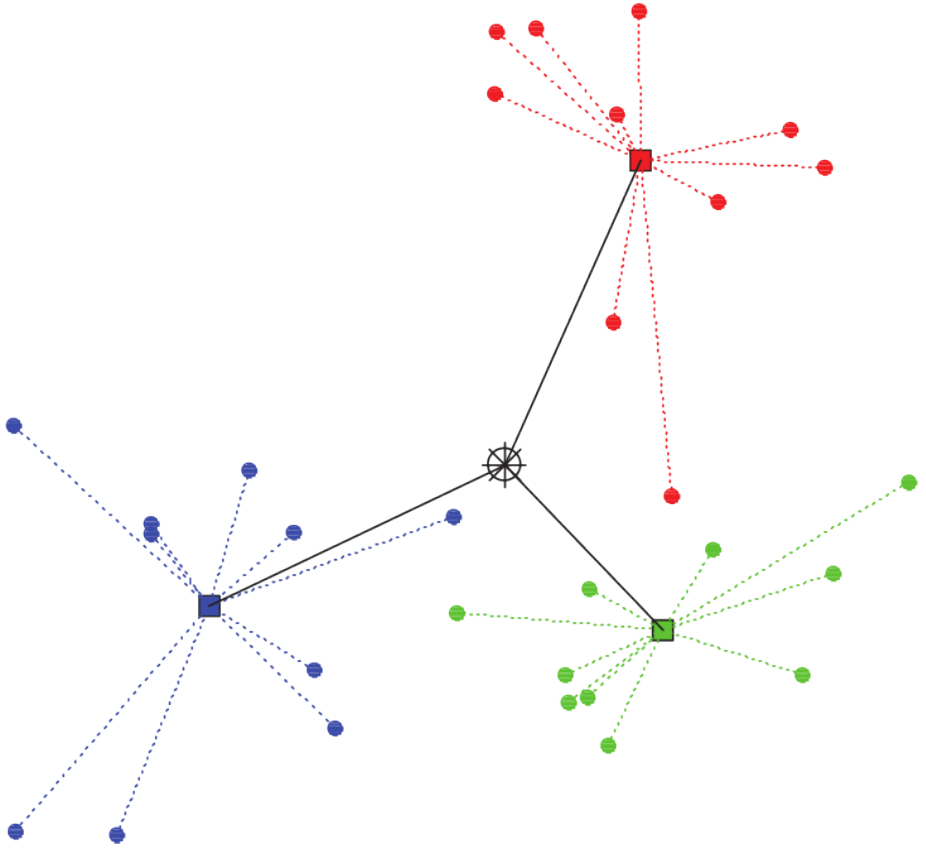
Introduction to statistical modeling of community data



Introduction to statistical modeling of community data

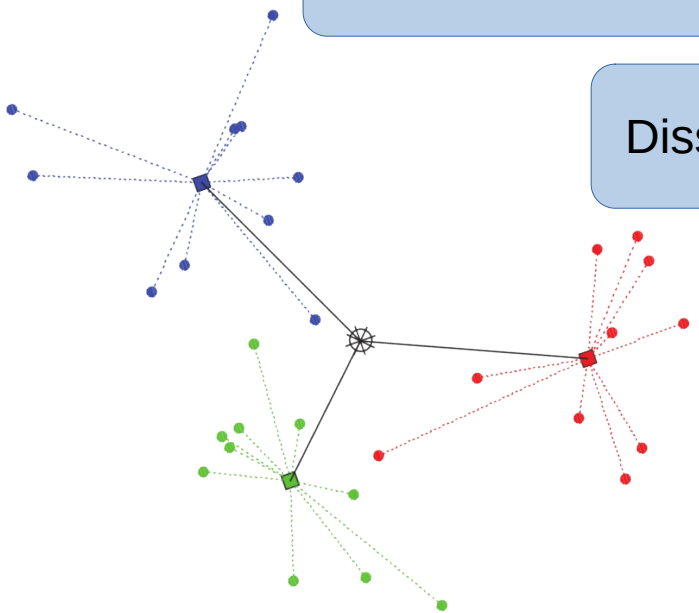
Alpha diversity

Beta diversity

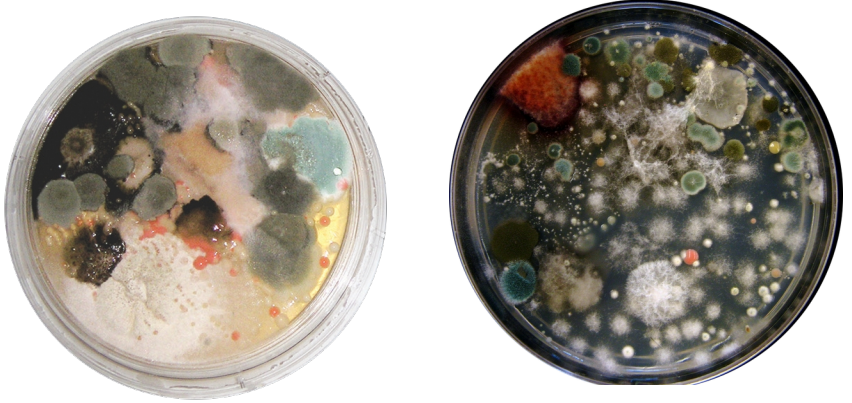
Dissimilarity coefficients

Ordinations

Statistical models

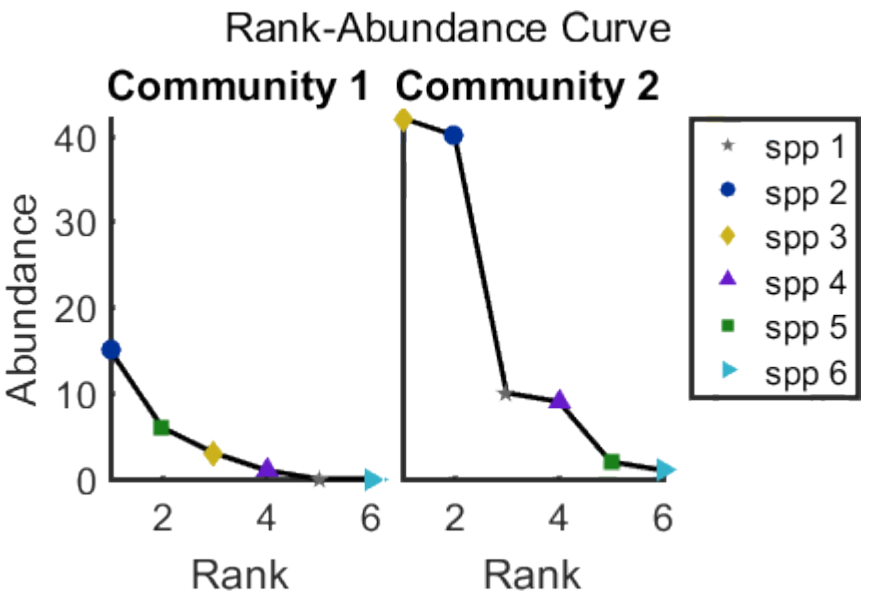


Alpha diversity

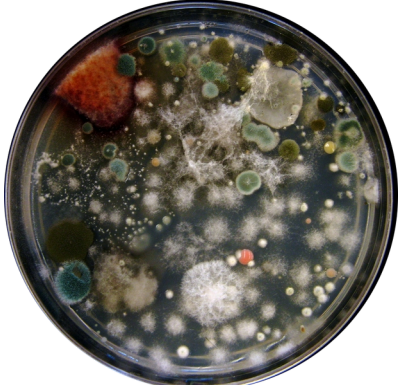


Diversity metrics are simple, **one-sample** indicators of species richness.

They are sometimes also about the shape of the species abundance distributions, called **dominance** or **species evenness**.



Alpha diversity

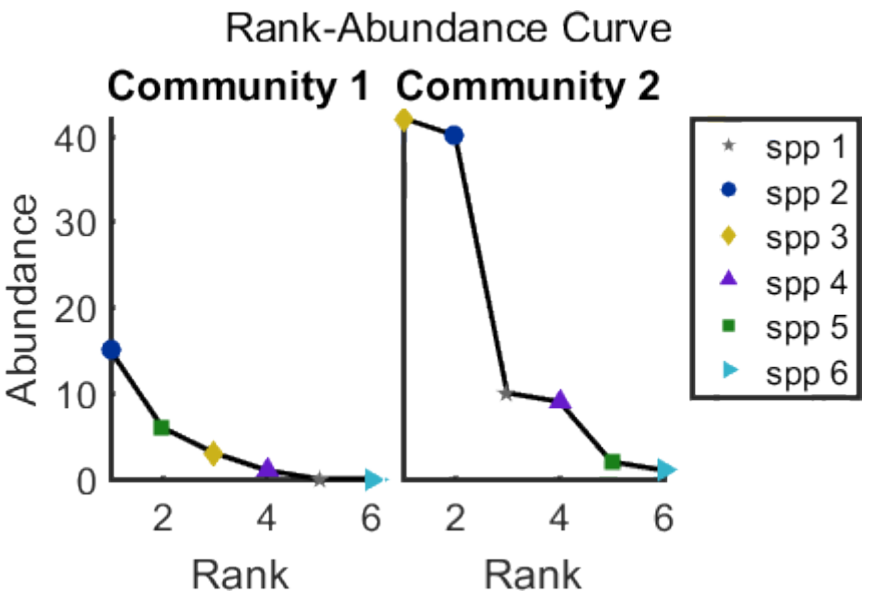


Shannon diversity index

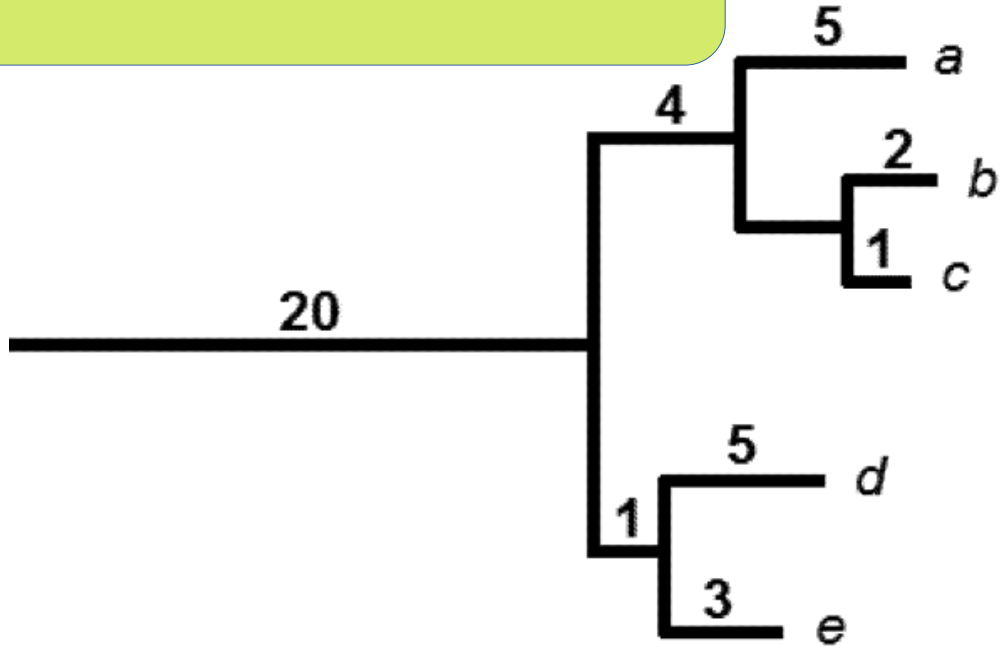
$$H' = -\Sigma\, pi \ln pi$$

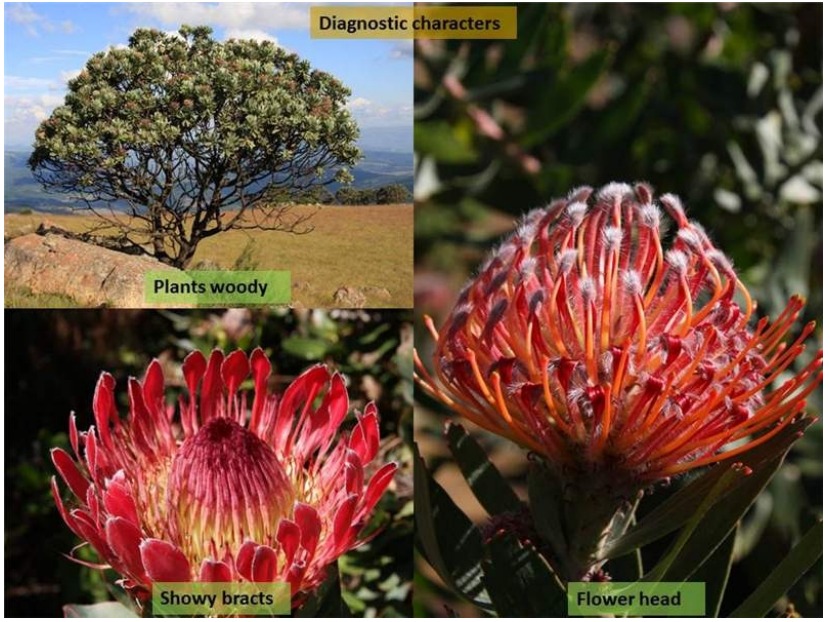
Simpsons reciprocal diversity index

$$D = \frac{N(N-1)}{\sum n(n-1)}$$



Alpha diversity





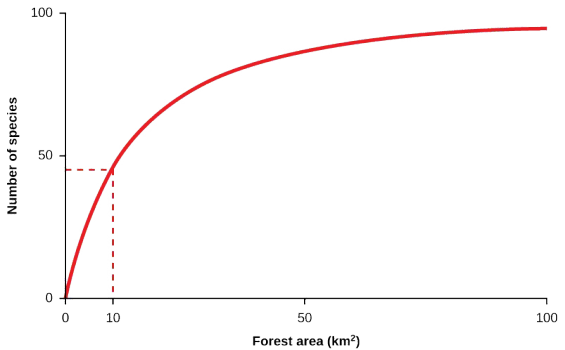
Phylogenetic alpha diversity measures exist (e.g. Faith’s PD).

Alpha diversity

Alpha diversity is interesting, but common problems and limits emerge

Alpha diversity

Comparability problems

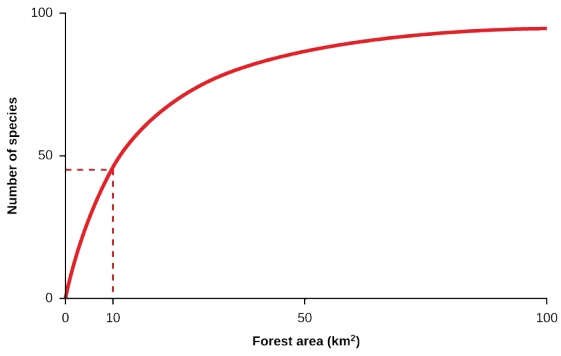


Alpha diversity is interesting, but common problems and limits emerge

Alpha diversity

Loss of information, such as endemics or unique ecological processes

Comparability problems

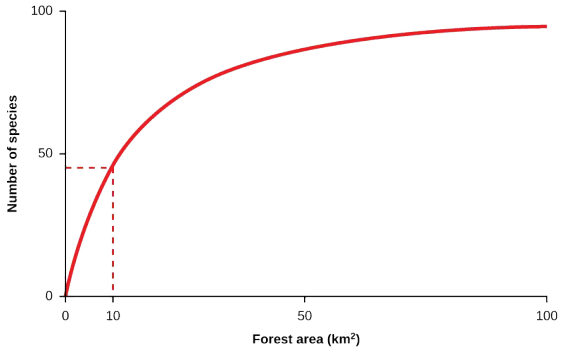


Alpha diversity is interesting, but common problems and limits emerge

Alpha diversity

Loss of information, such as endemics or unique ecological processes

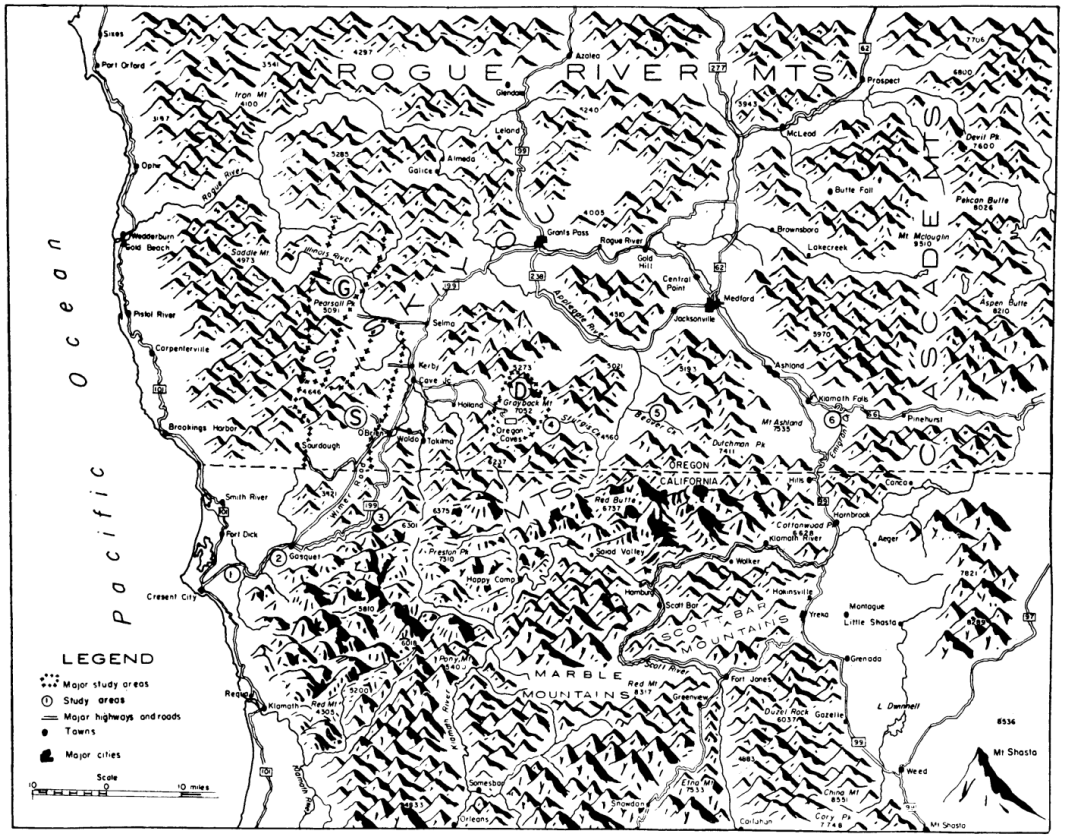
Comparability problems



Over-simplifications:
My forest is more diverse than yours!!!!
Deforestation doesn't affect microbes!!!

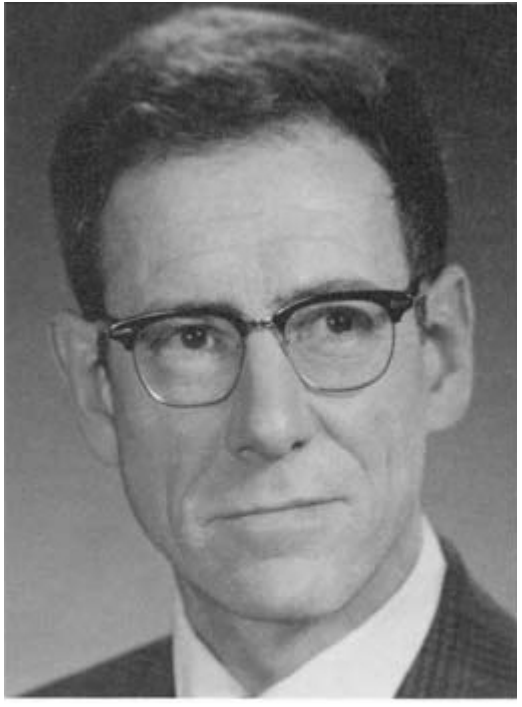
Alpha diversity is interesting, but common problems and limits emerge

Beta diversity

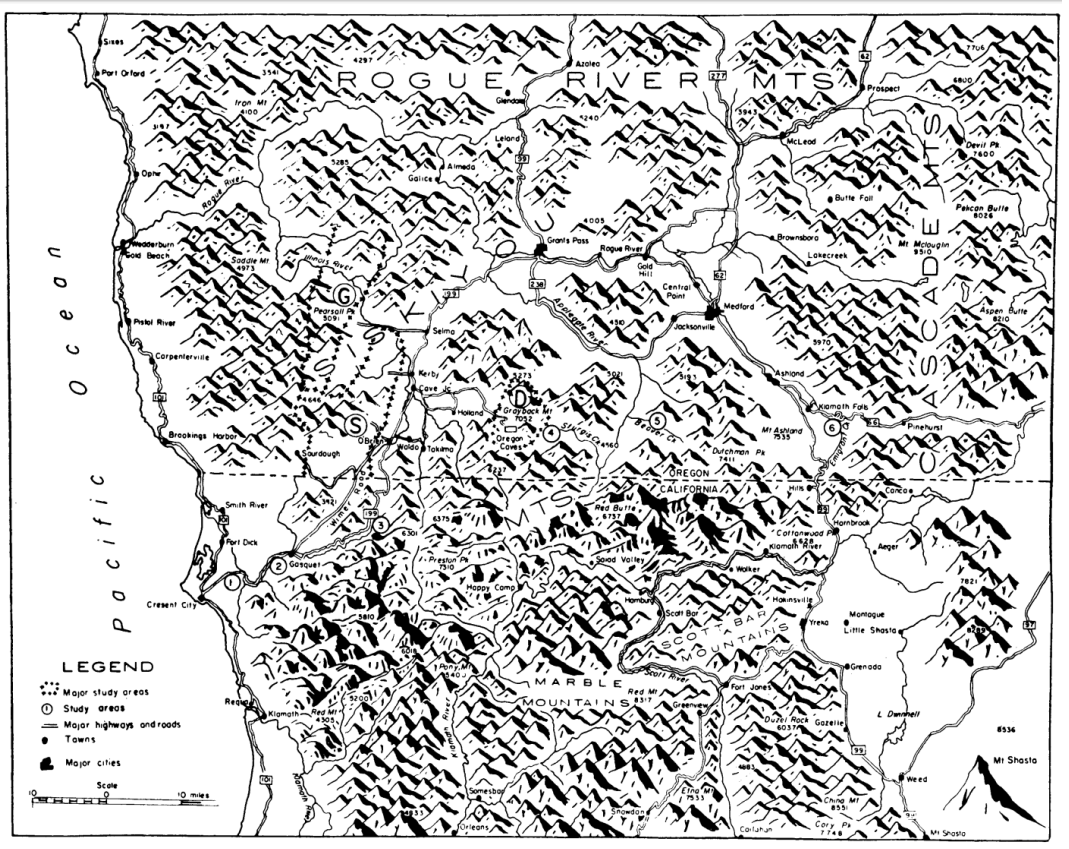


Beta diversity is the comparison of two or more sites.

Beta diversity

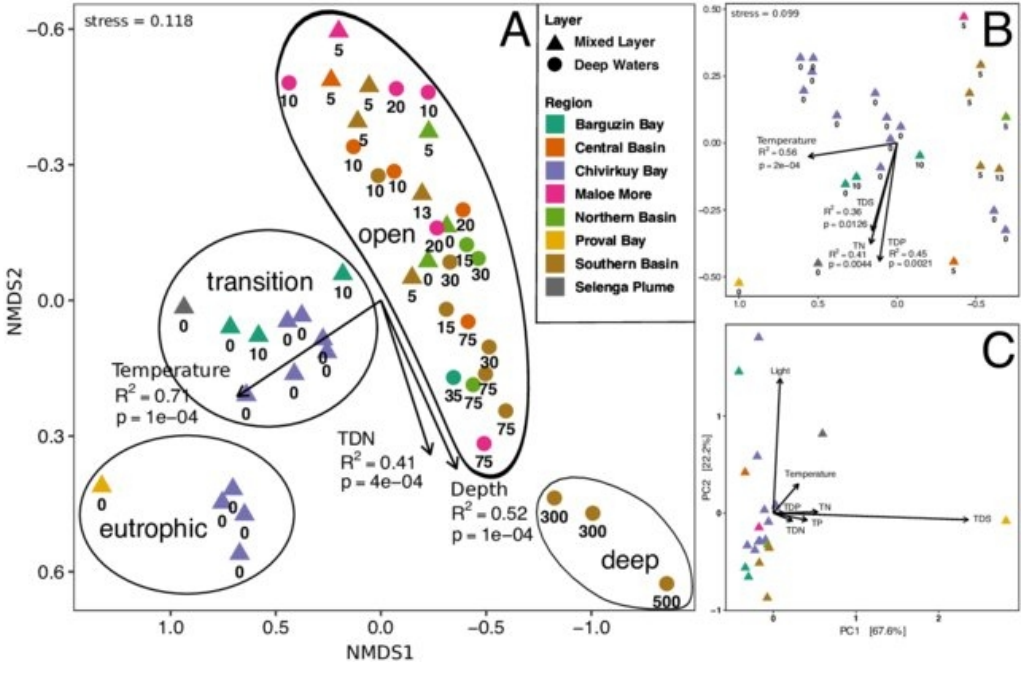


Whitaker (1960)



Beta diversity is the comparison of two or more sites.

Beta diversity



Wilburn, Paul, et al. (2020)

How do we quantify the communities of two sites or samples as similar/different?

Dissimilarity coefficients

$$1 - \left[\frac{\sum_{j=1}^{p} A_j}{\sum_{j=1}^{p} (A_j + B_j + C_j)}\right]$$

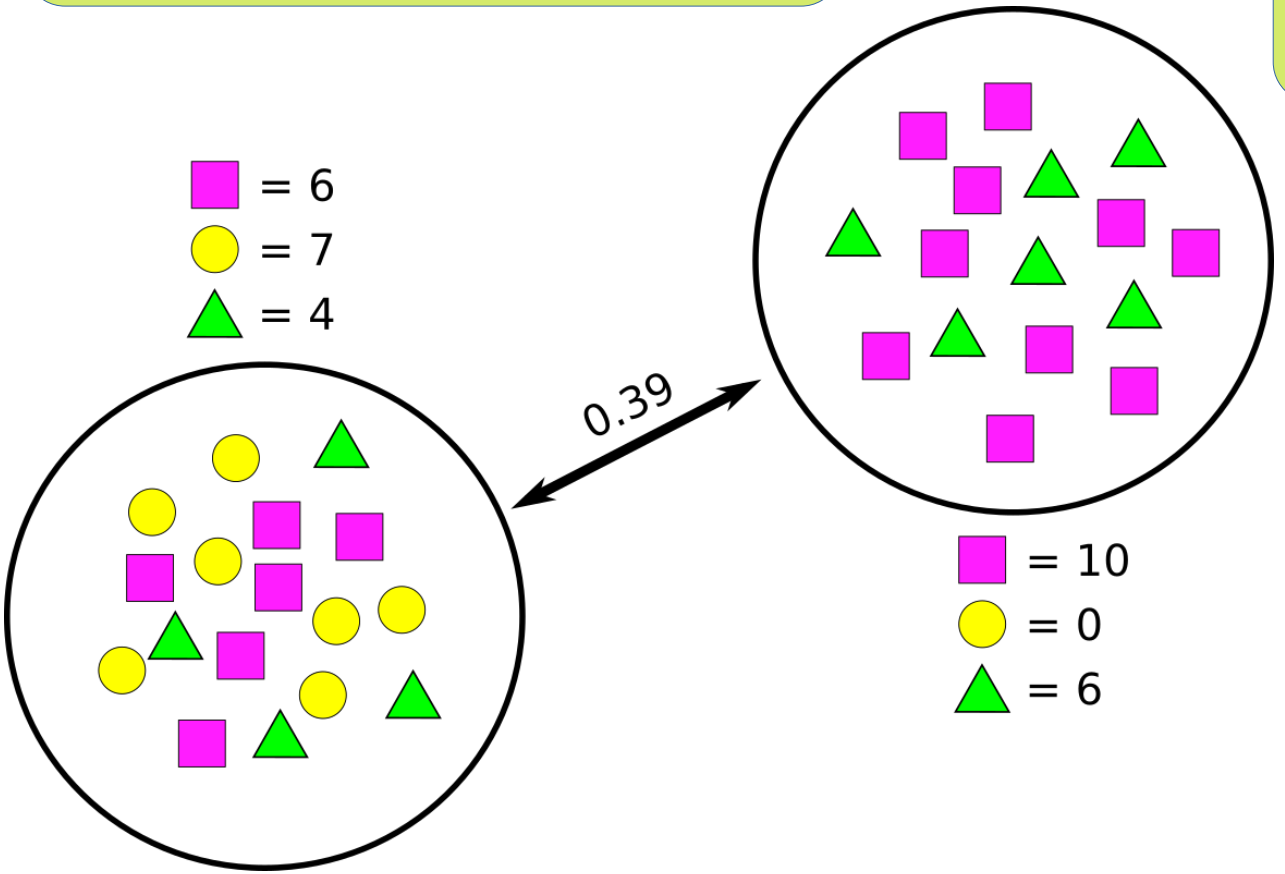
$$\sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2} \qquad 1 - \frac{1}{2}\left[\frac{\sum_{j=1}^{p} A_j}{\sum_{j=1}^{p} (A_j + B_j)} + \frac{\sum_{j=1}^{p} A_j}{\sum_{j=1}^{p} (A_j + C_j)}\right] \qquad \sqrt{\sum_{j=1}^{p} \left(\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}}\right)^2}$$

$$\frac{1}{pp} \sum_{j=1}^{p} \left[\frac{B_j + C_j}{2A_j + B_j + C_j}\right]$$

$$\sum_{j=1}^{p} (B_j + C_j)^{\dagger}$$

$$1 - \left[\frac{\sum_{j=1}^{p} 2A_j}{\sum_{j=1}^{p} (2A_j + B_j + C_j)}\right] \qquad \frac{\sum_{j=1}^{p} |y_{1j} - y_{2j}|}{\sum_{j=1}^{p} (y_{1j} + y_{2j})} \qquad \sqrt{\sum_{j=1}^{p} \left(\frac{y_{1j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2}} - \frac{y_{2j}}{\sum_{j=1}^{p} y_{2j}^2}\right)^2}$$

Dissimilarity coefficients



Bray-Curtis dissimilarity

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

$C_{ij} = 6 + 4 = 10$

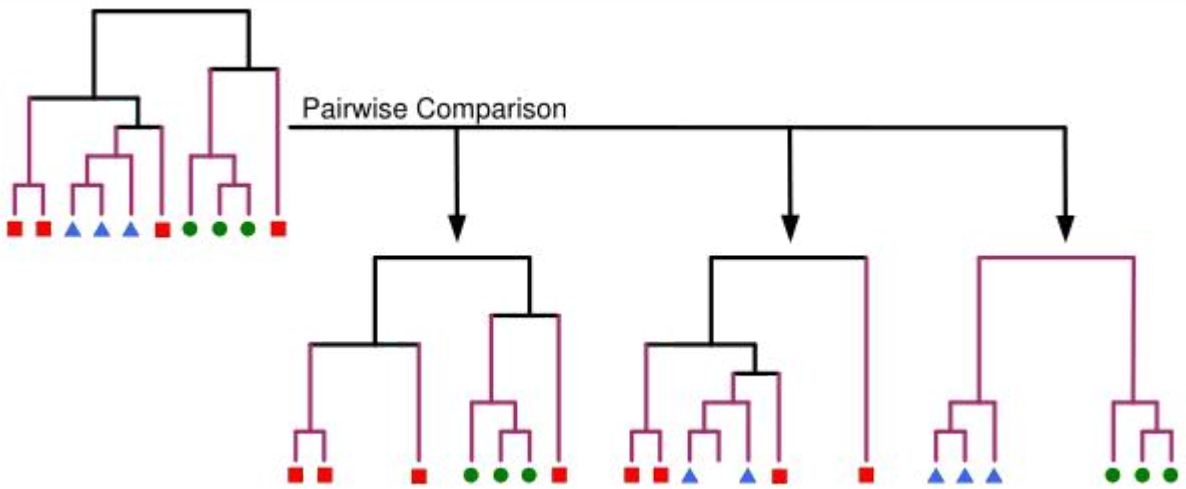
$S_i = 6 + 7 + 4 = 17$

$S_j = 10 + 6 = 16$

$BC_{ij} = 1 - (2 \times 10)/(17 + 16) = 0.39$

Dissimilarity coefficients

Unifrac Metric



	■	●	▲
■	0	0.6	0.5
●	0.6	0	1.0
▲	0.5	1.0	0

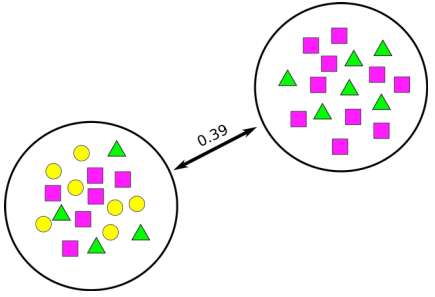
Distance Matrix

“Unique Fraction” = Fraction of unshared branch lengths

Dissimilarity coefficients

We can - and usually do - make many comparisons in a single study.

To record this, we create a **dissimilarity matrix** (or “distance matrix”)

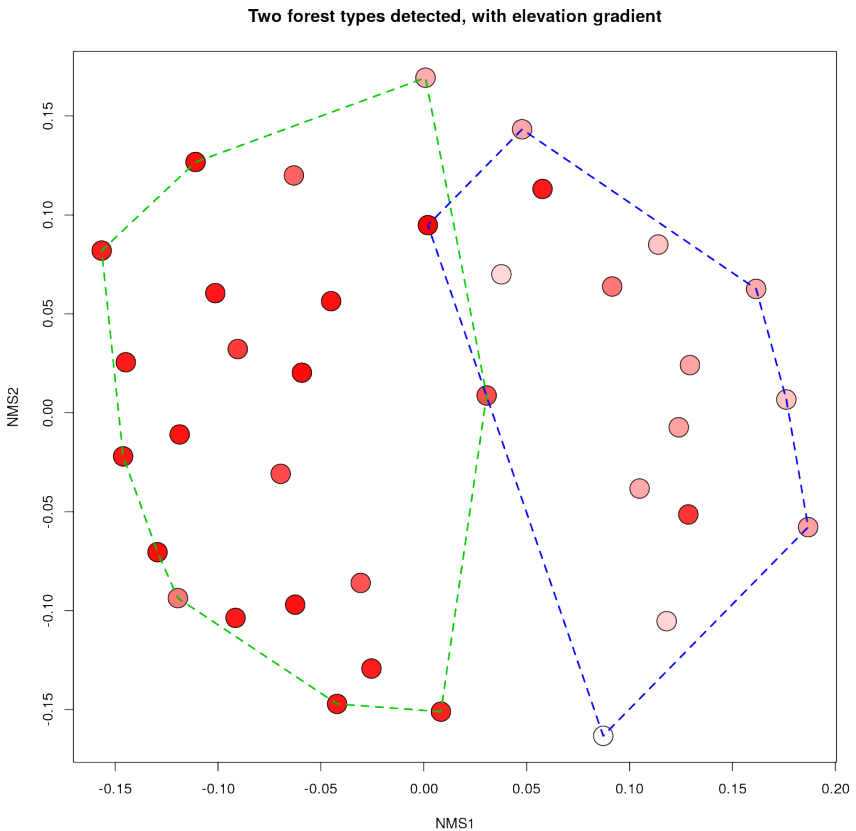


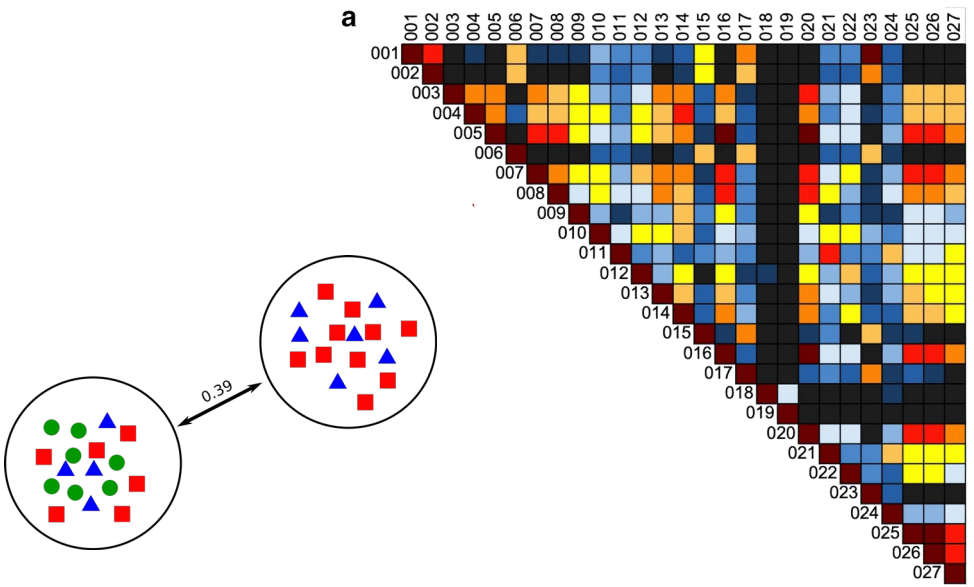
Samples

Samples	S1	S2	S3	S4
S1	0	...	...	...
S2	0.47	0	...	...
S3	0.84	0.64	0	...
S4	0.96	1	1	0

Dissimilarity coefficients

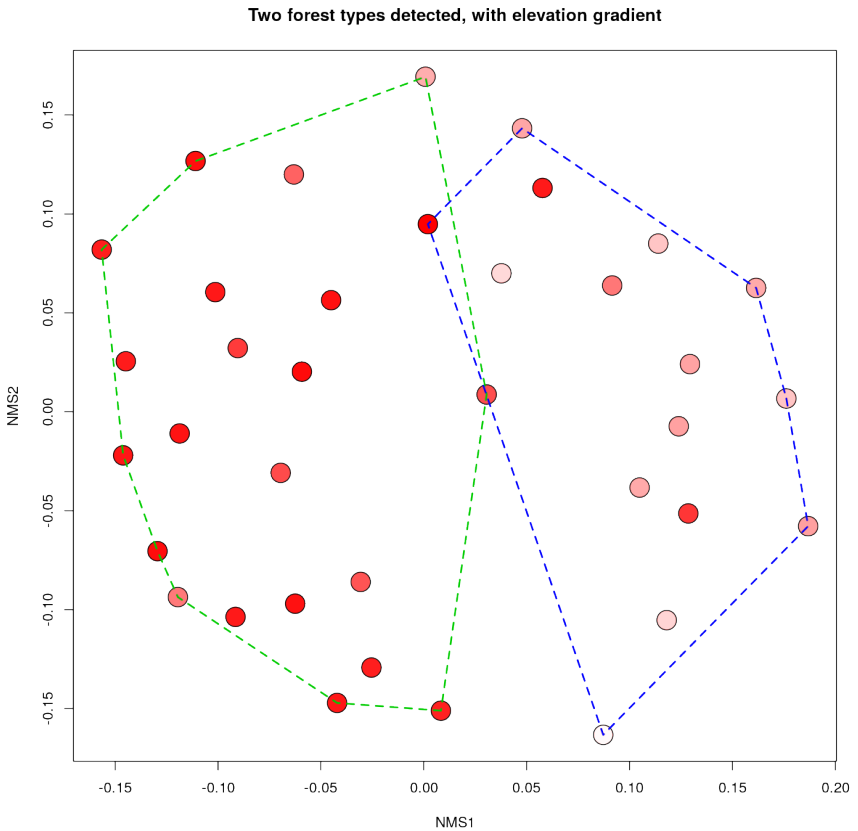
When we attempt to fairly represent the ecological distances among all of our samples in a single 2-dimensional graphic, this is called **ordination**.





Ordinations

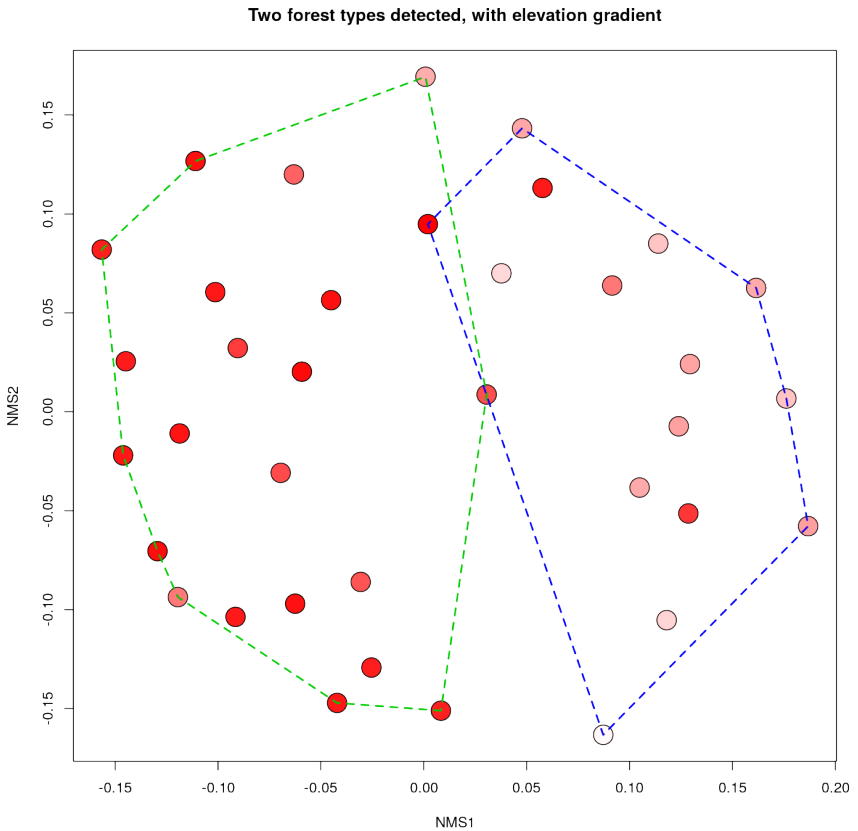
When viewing an ordination, always ask four questions:



1. What is the distance metric used?

Ordinations

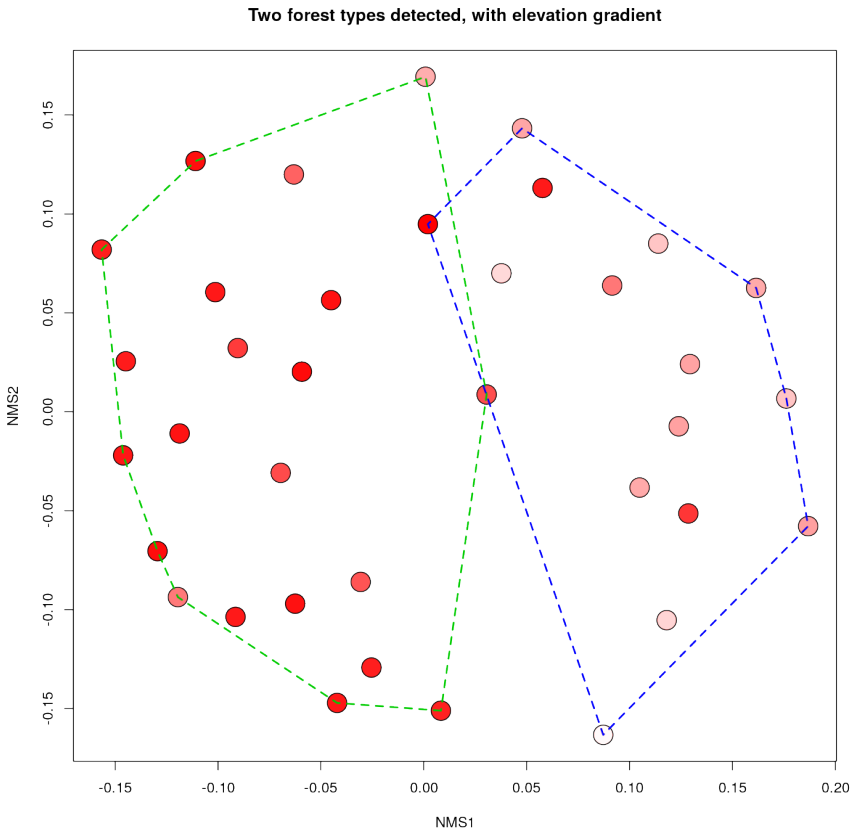
When viewing an ordination, always ask four questions:



1. What is the distance metric used?

2. What is the ordination algorithm used?

Ordinations



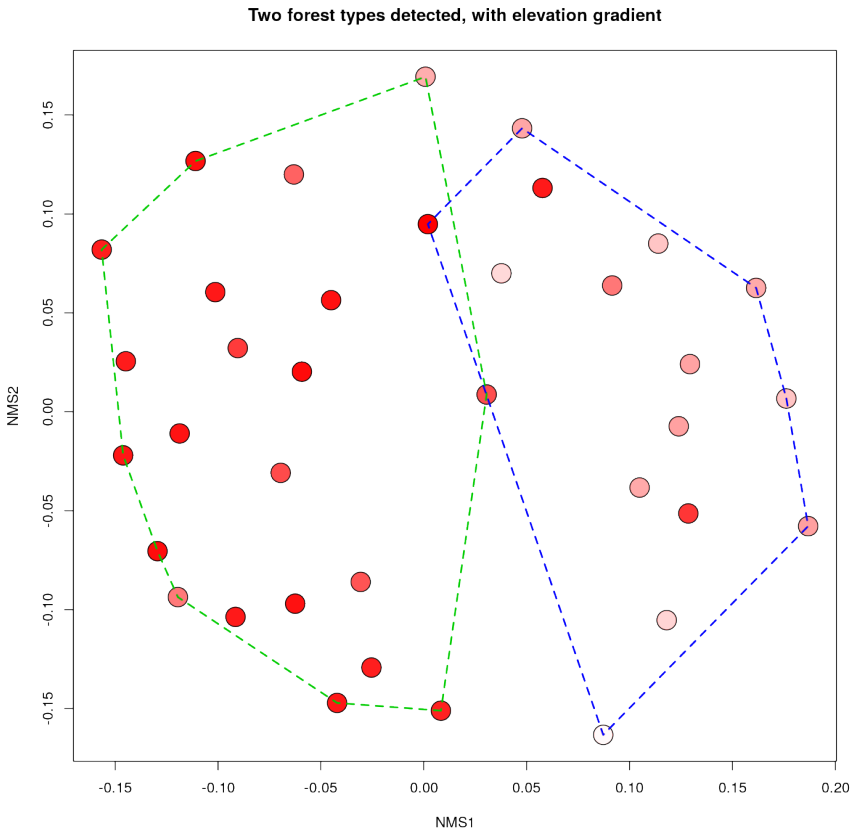
When viewing an ordination, always ask four questions:

1. What is the distance metric used?

2. What is the ordination algorithm used?

3. Is it a good fit? (Or is it “stressed” out?)

Ordinations



When viewing an ordination, always ask four questions:

1. What is the distance metric used?

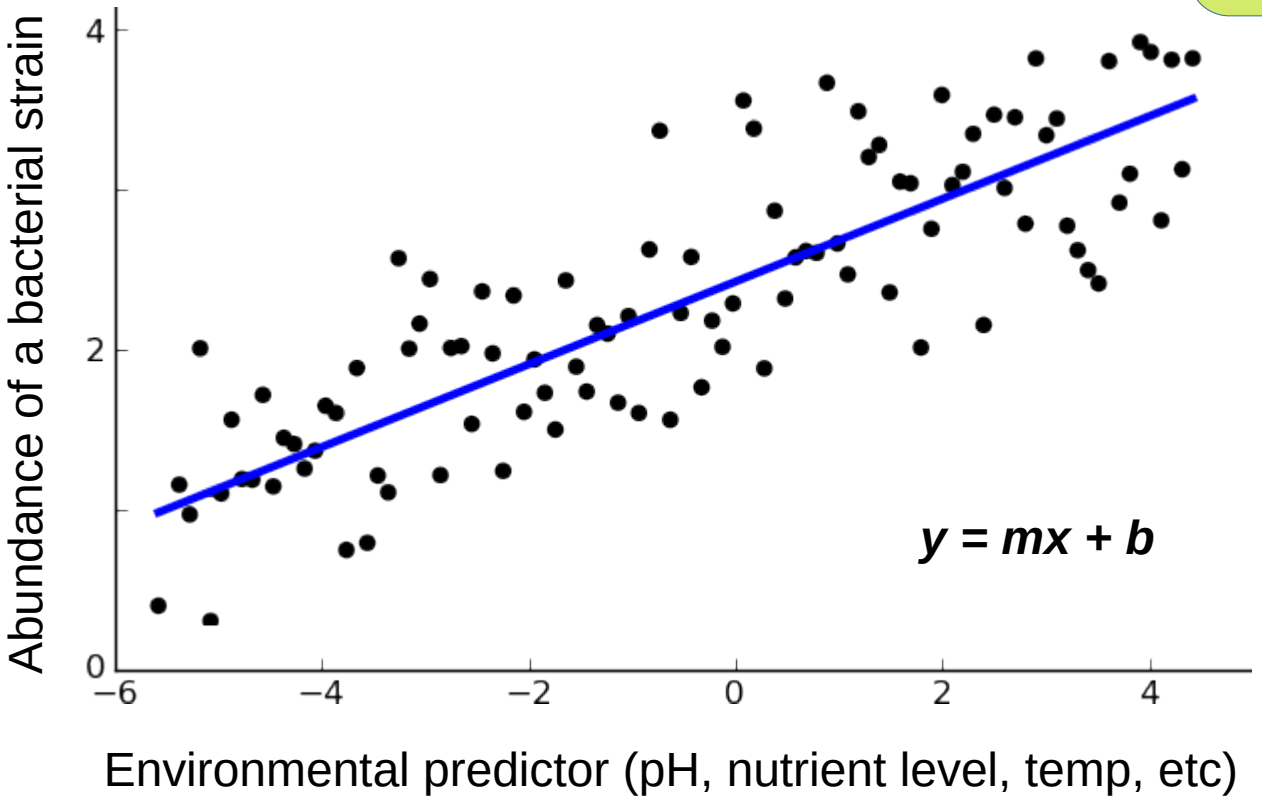
2. What is the ordination algorithm used?

3. Is it a good fit? (Or is it “stressed” out?)

4. Is there a statistical model behind it?

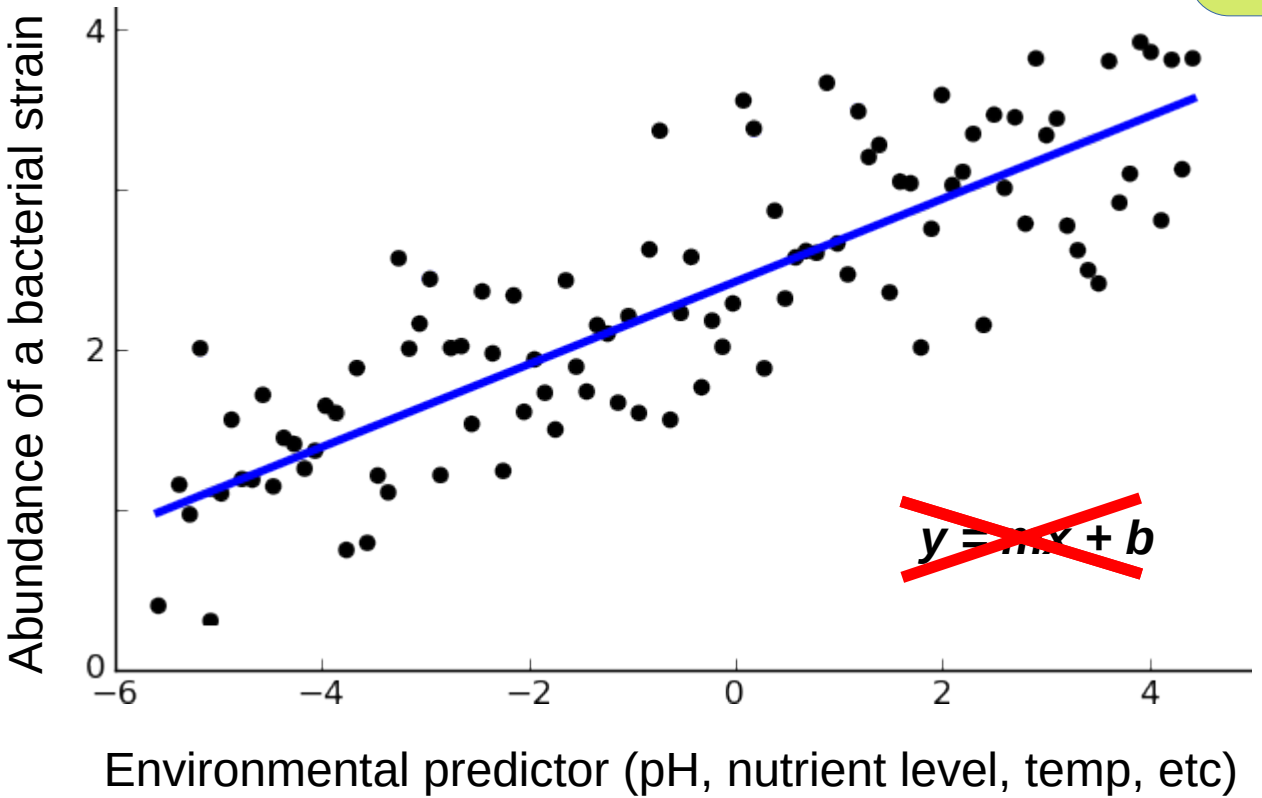
Statistical models

Linear models often don't work well for microbial ecologists….



Statistical models

Linear models often don't work well for microbial ecologists….



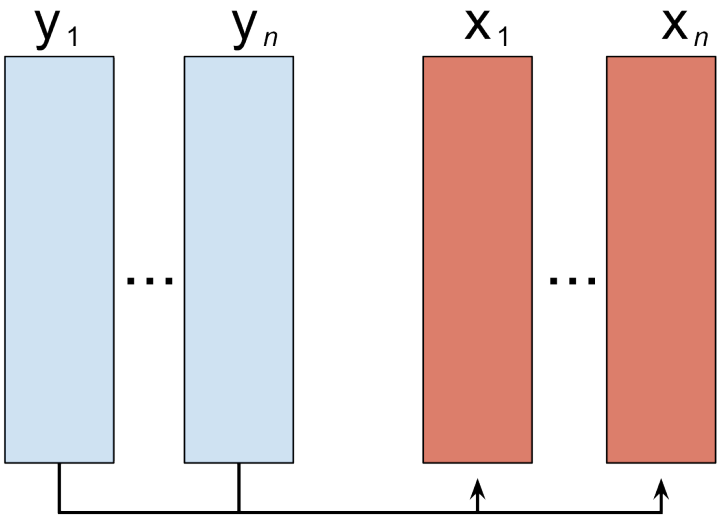
Why not?

Statistical models

Species

Sites	X1	X2	X3	X4
S1	14	2	14	14
S2	10	14	0	8
S3	0	5	0	2
S4	0	0	1	0

Abundance of species X3 at site S4



Environmental parameters

Sites	E1	E2	E3	E4
S1	0.4	45	2.5	33
S2	0.1	2	1.2	17
S3	0.2	24	4	43
S4	0	1	0.5	9

Value of parameter E3 at site S4

Huge response variable matrices (community matrix)!

Statistical models

We can’t use the most basic statistical tool (linear models)….

$$y = mx + b$$

Statistical models

We can’t use the most basic statistical tool (linear models)….

$y = ax + b$

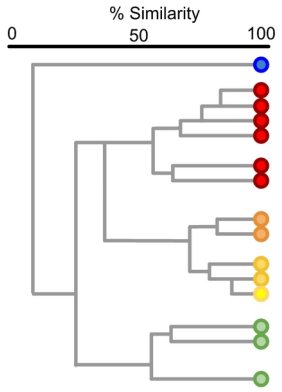
Statistical models

We can’t use the most basic statistical tool (linear models)….

$y = ax + b$

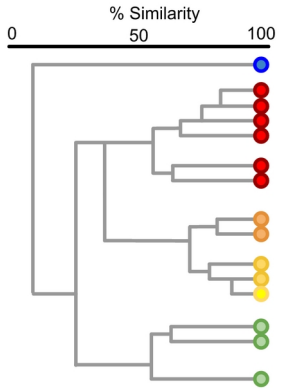
...so what can we do?

Statistical models

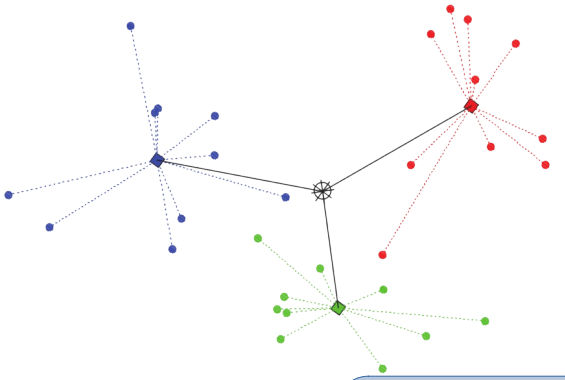


Clustering of samples

Statistical models

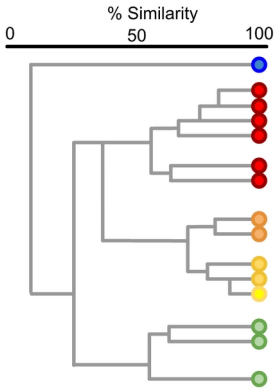


Clustering of samples

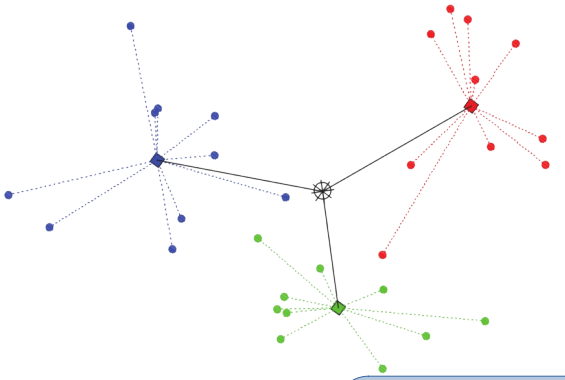


Resemblence-based permutation tests

Statistical models

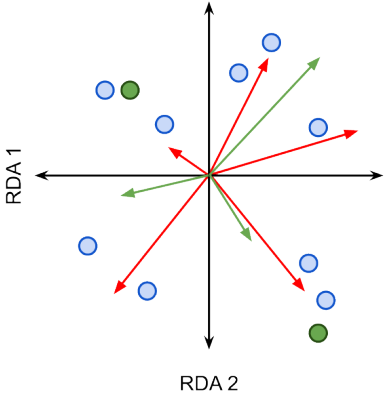


Clustering of samples



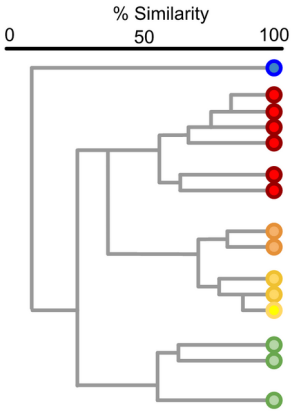
Resemblence-based permutation tests

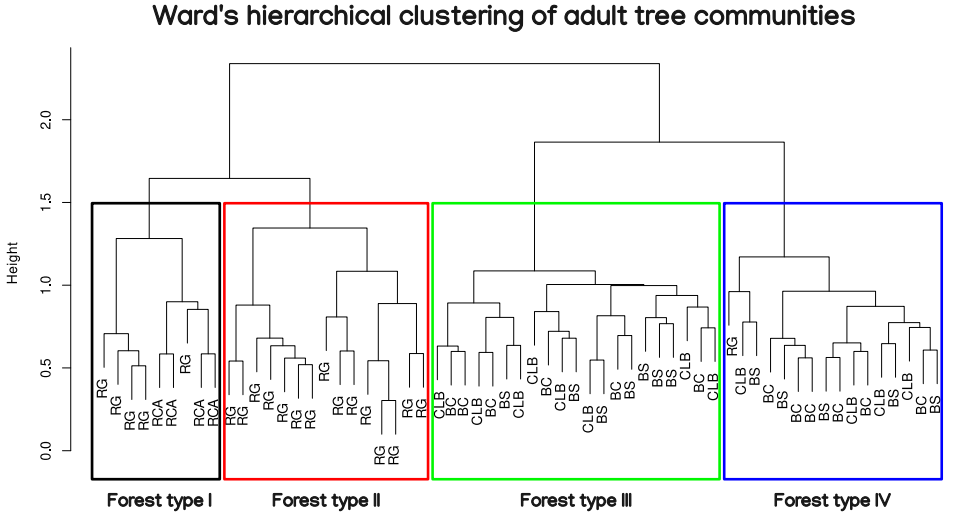
Dimensionality reduction



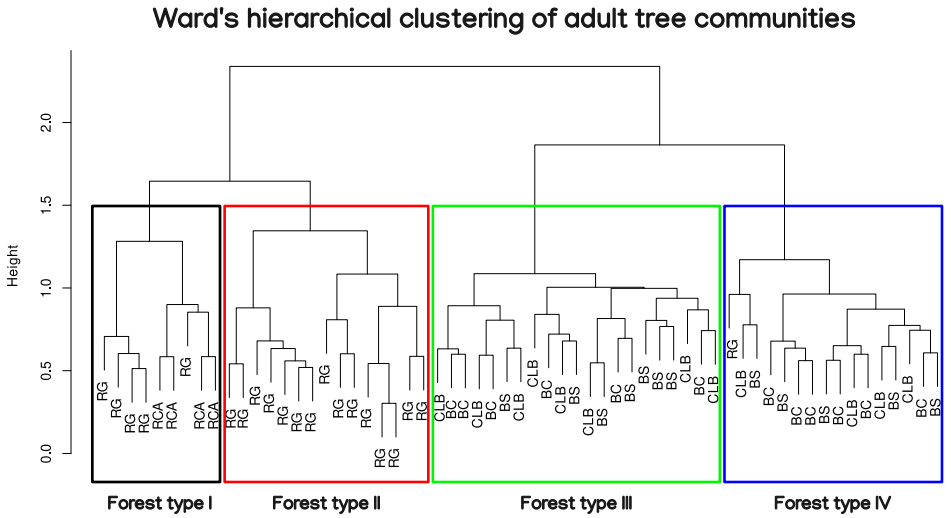
Clustering of samples

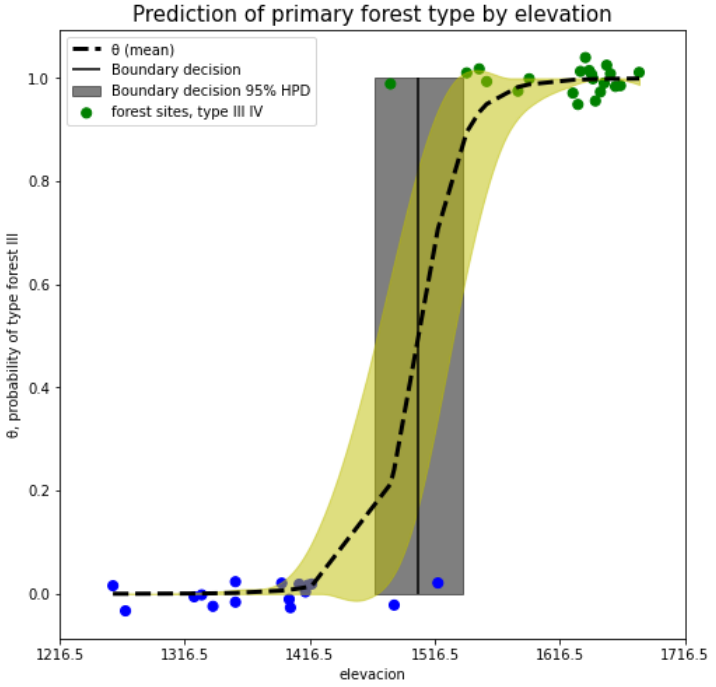
Numerous clustering algorithms exist. They using work either from an agglomerative (“building-up”) or divisive (“tearing apart”) methods.





Clustering of samples

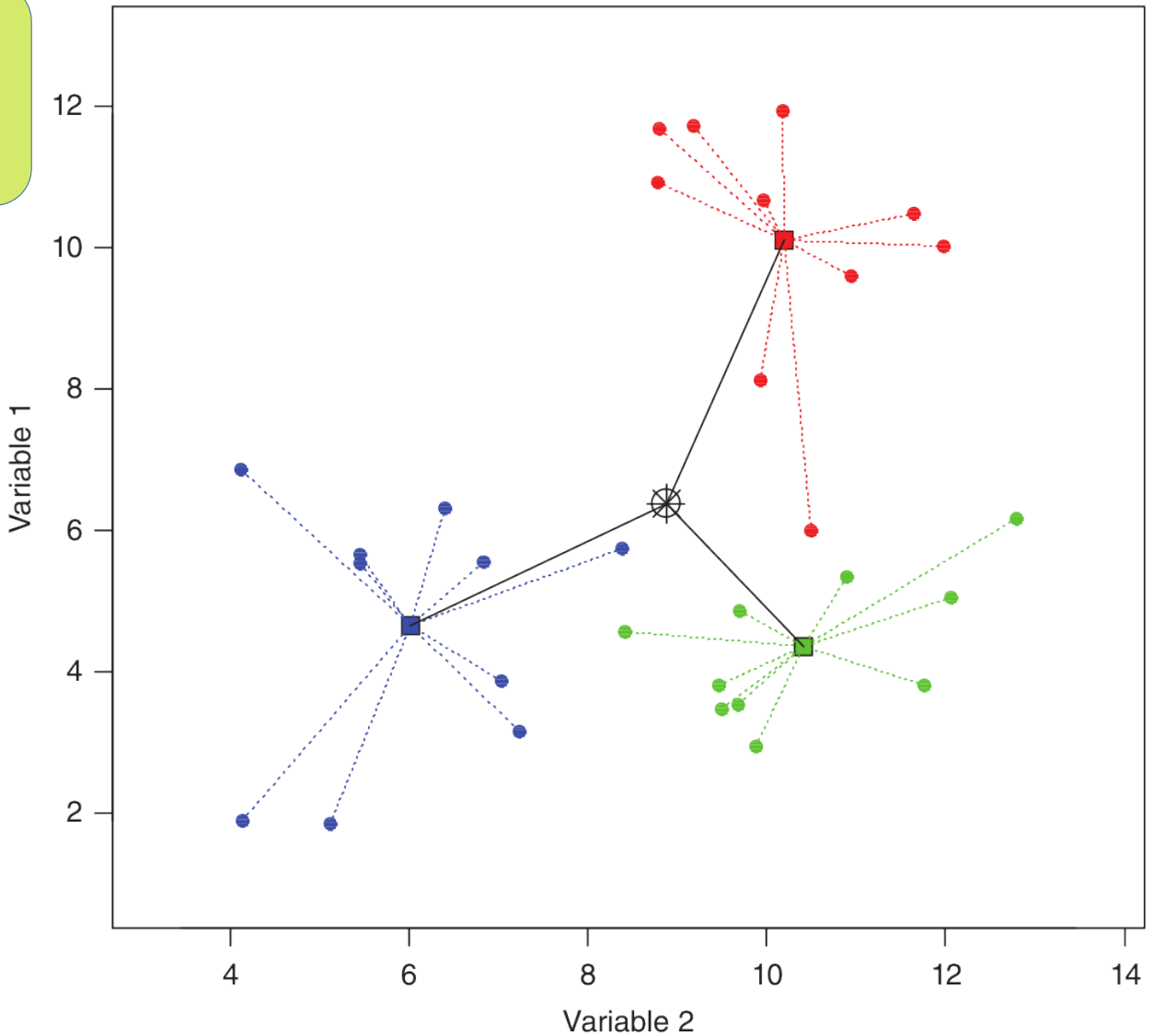




This creates a really simply categorical variable out of our complex community (one column instead of 1800!), so can then use a linear models like we want to.

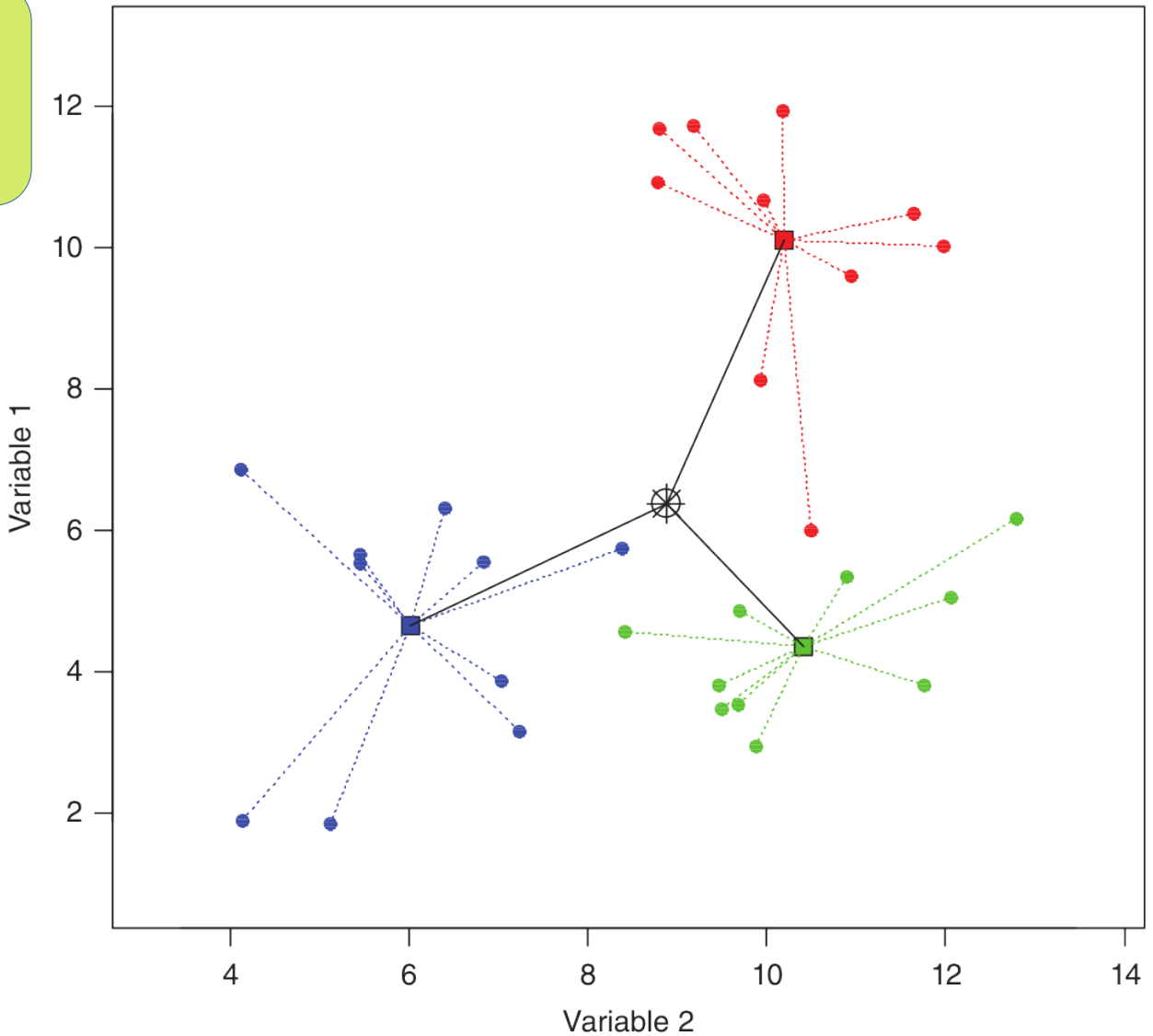
Resemblence-based permutation tests

A different approach is to see if the predicted groupings are any more tightly clustered than would occur by chance alone. PERMANOVA and other “pseudo-linear models”.



Resemblence-based permutation tests

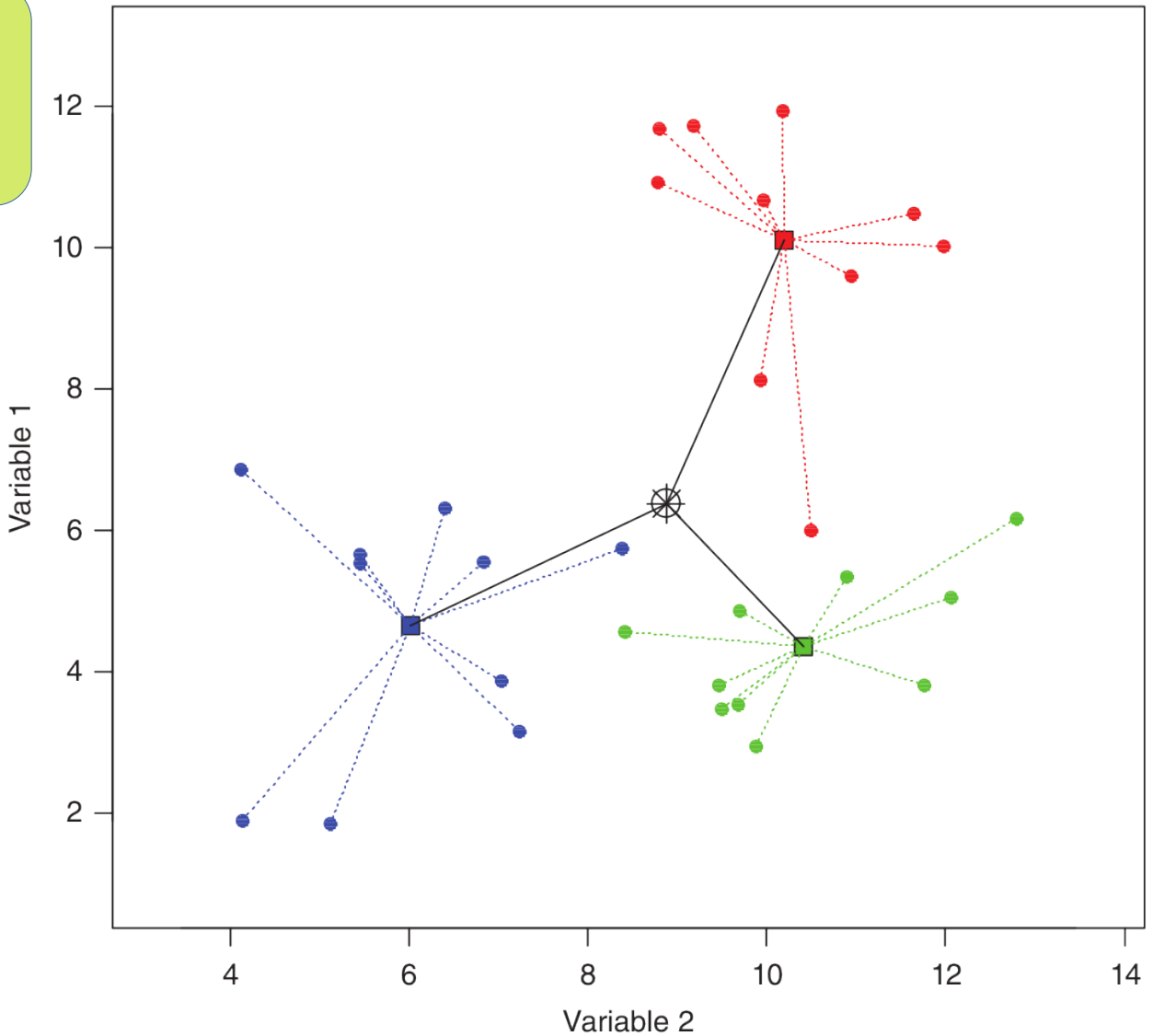
PERMANOVA and its colleagues are very robust tests, and you can do hypothesis testing.



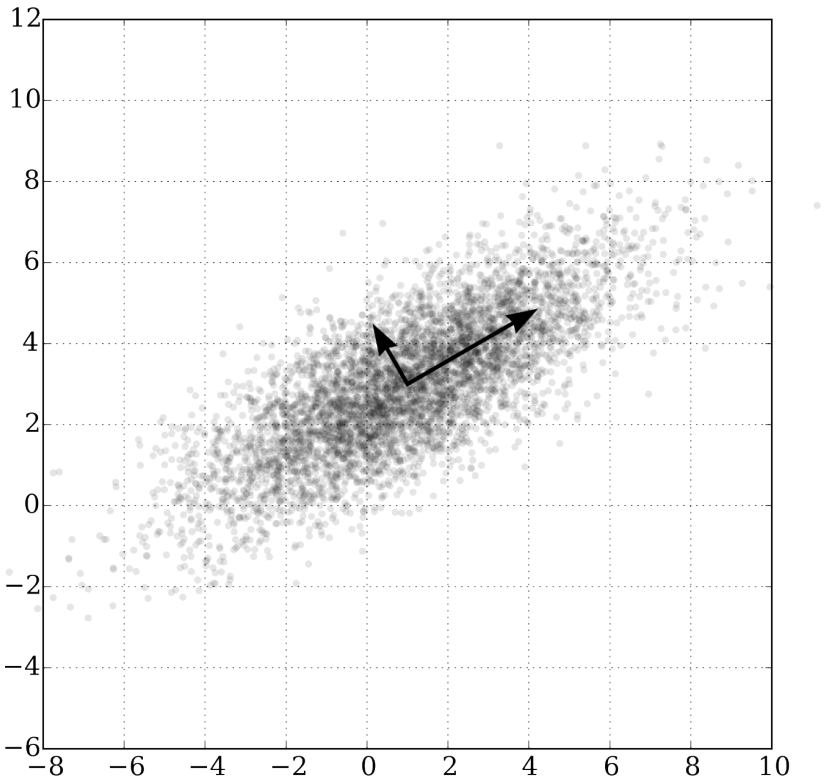
Resemblence-based permutation tests

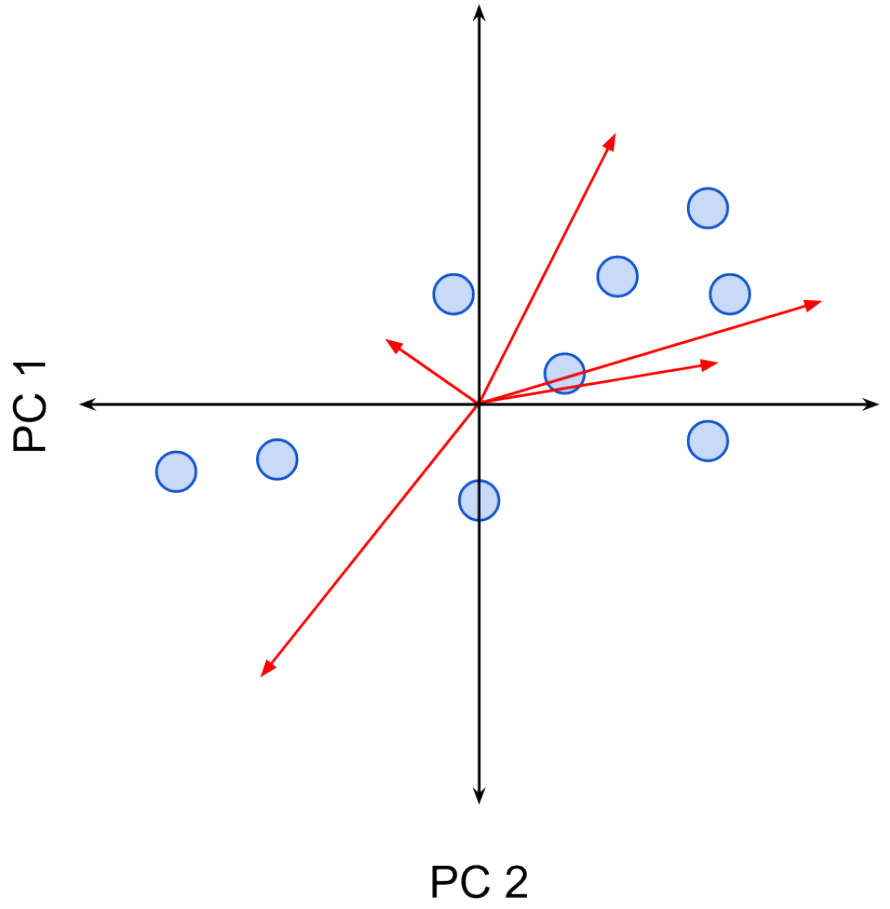
PERMANOVA and its colleagues are very robust tests, and you can do hypothesis testing.

But a lot of information is lost. (Where are my old predictor variables??)



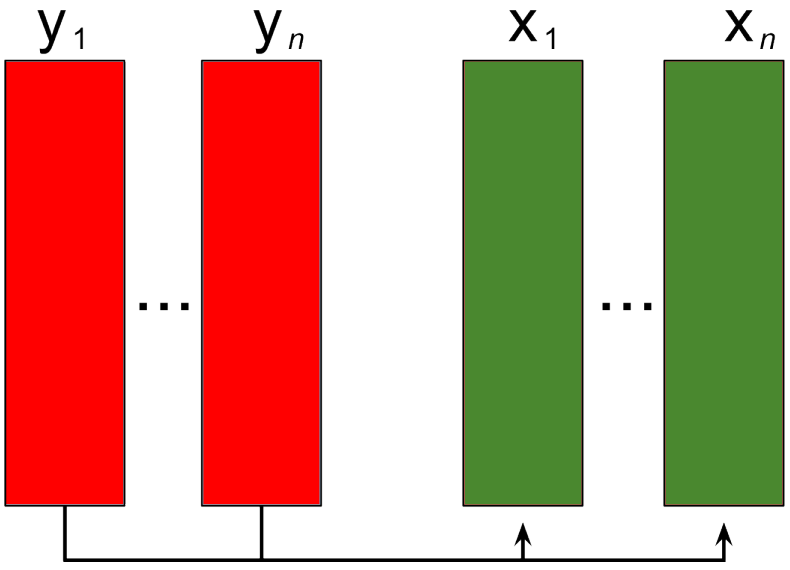
Dimensionality reduction

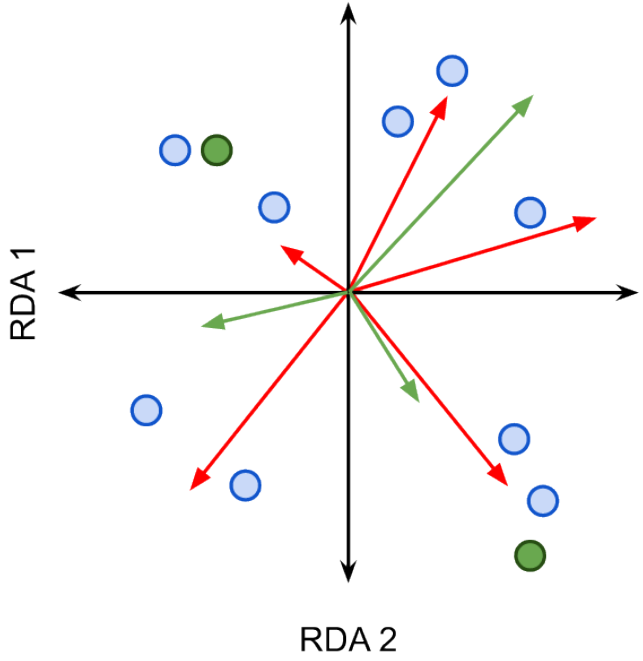




Dimensionality reduction is based on eigen-decomposition to create new axes that explain variance more succinctly.

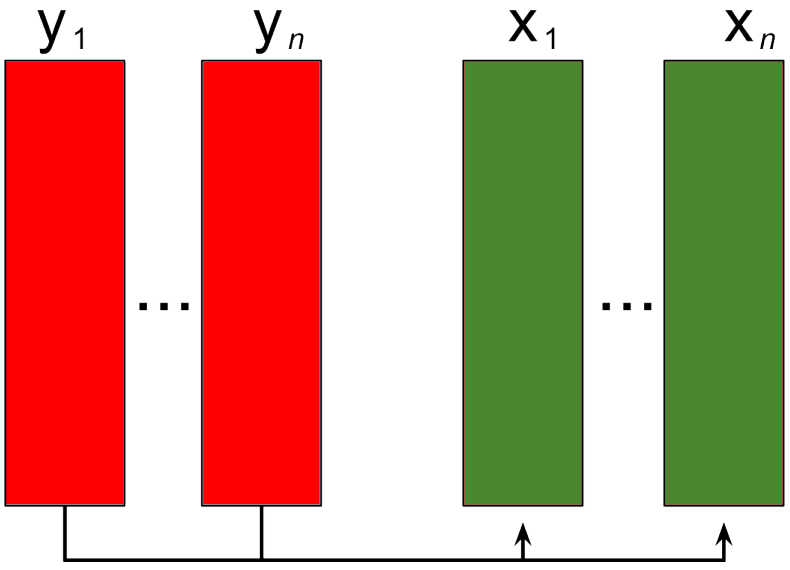
Dimensionality reduction

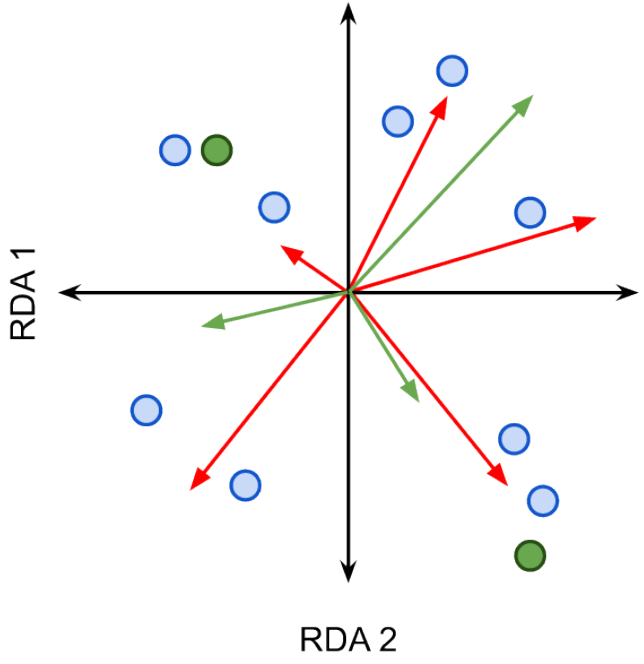




We can create new PCA axes from both the predictors (e.g. environmental variables) and response variable (e.g. bacterial 16s read abundance matrix) to tell us which of our environmental data are most important for predicting changes in our microbial data.

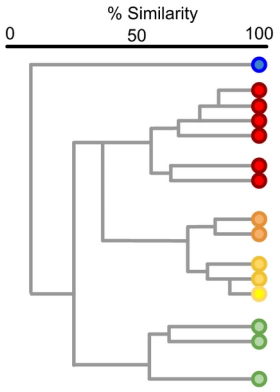
Dimensionality reduction



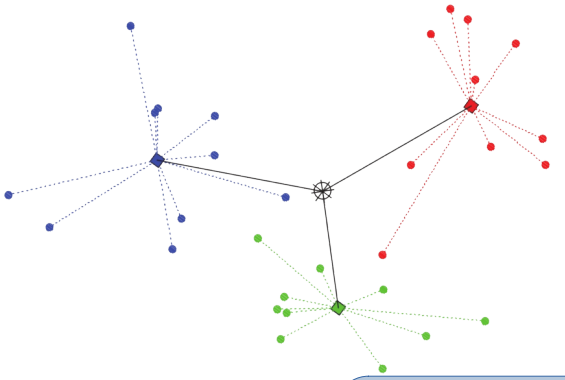


This is as close as we can get to a “pure” linear model, without algorithmic changing of the data. The microbial communities are plotted onto a few, new axes, and we can recover the effect of our original environmental or experimental data from these.

All of these methods can be combined!



Clustering of samples



Resemblence-based permutation tests

Dimensionality reduction