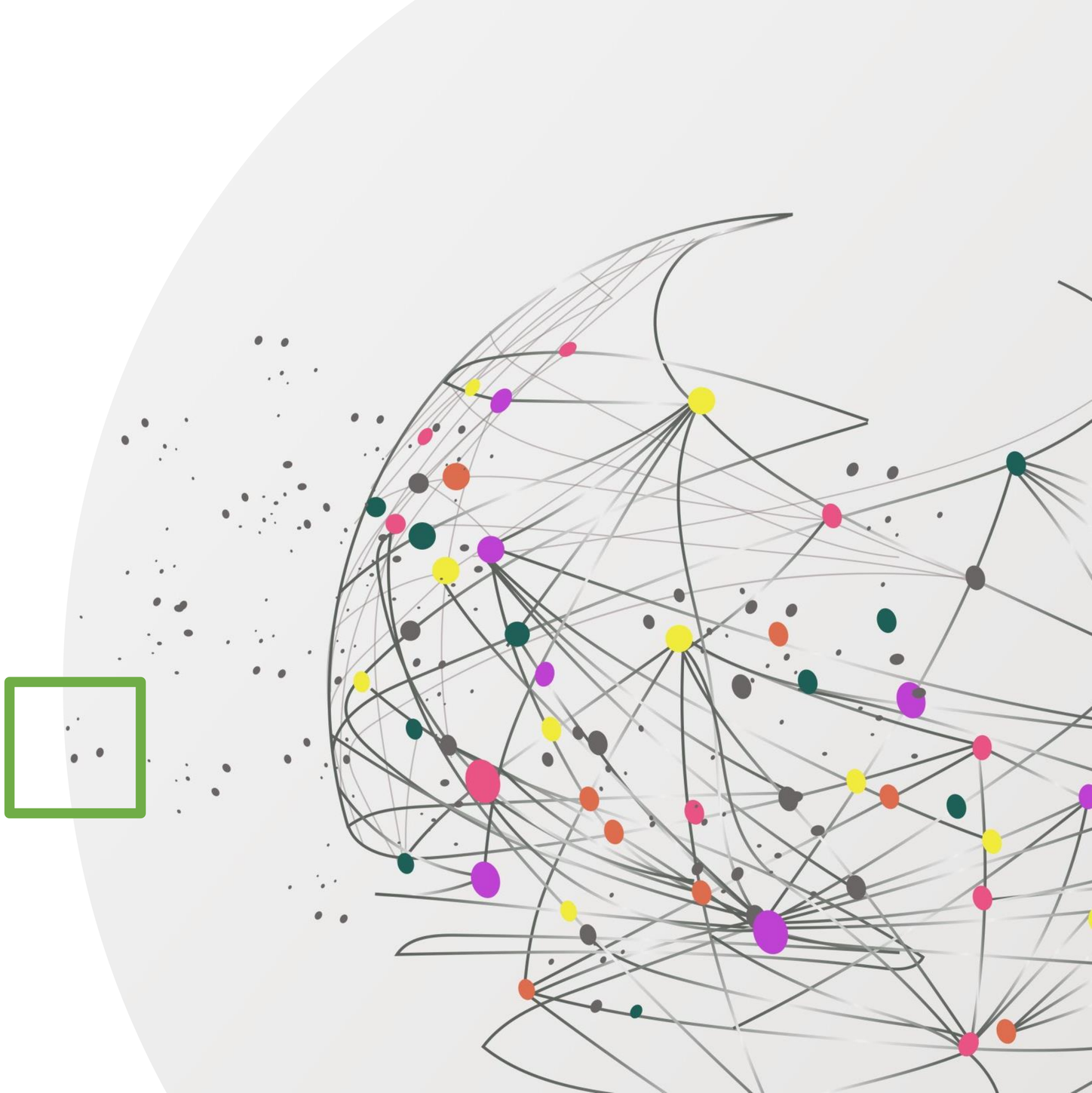
Community ecology and multivariate analyses

Anders K. Krabberød (UiO)
a.k.krabberod@ibv.uio.no
&
Based on code by
Ramiro Logares (ICM-CSIC)
ramiro.logares@icm.csic.es

AB332 - 2024



Purpose

- Learn methods for doing community ecology analysis of high throughput metabarcoding data in R
- Look at some common diversity indexes
- Learn ordination methods
- All analyses are done in R
  - Scripts and lectures can be found on github
  - https://github.com/krabberod/UNIS_AB332_2024

Computer Lab part I : Diversity

Keyterms: Species richness, abundance, evenness, and diversity

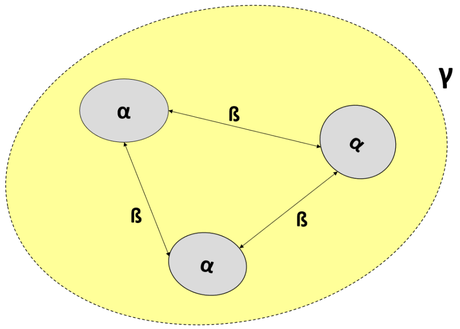
Each forest has the same four tree species (same species richness), but they differ in species evenness (relative abundance of each species).

Community 1

A: 25% B: 25% C: 25% D: 25%

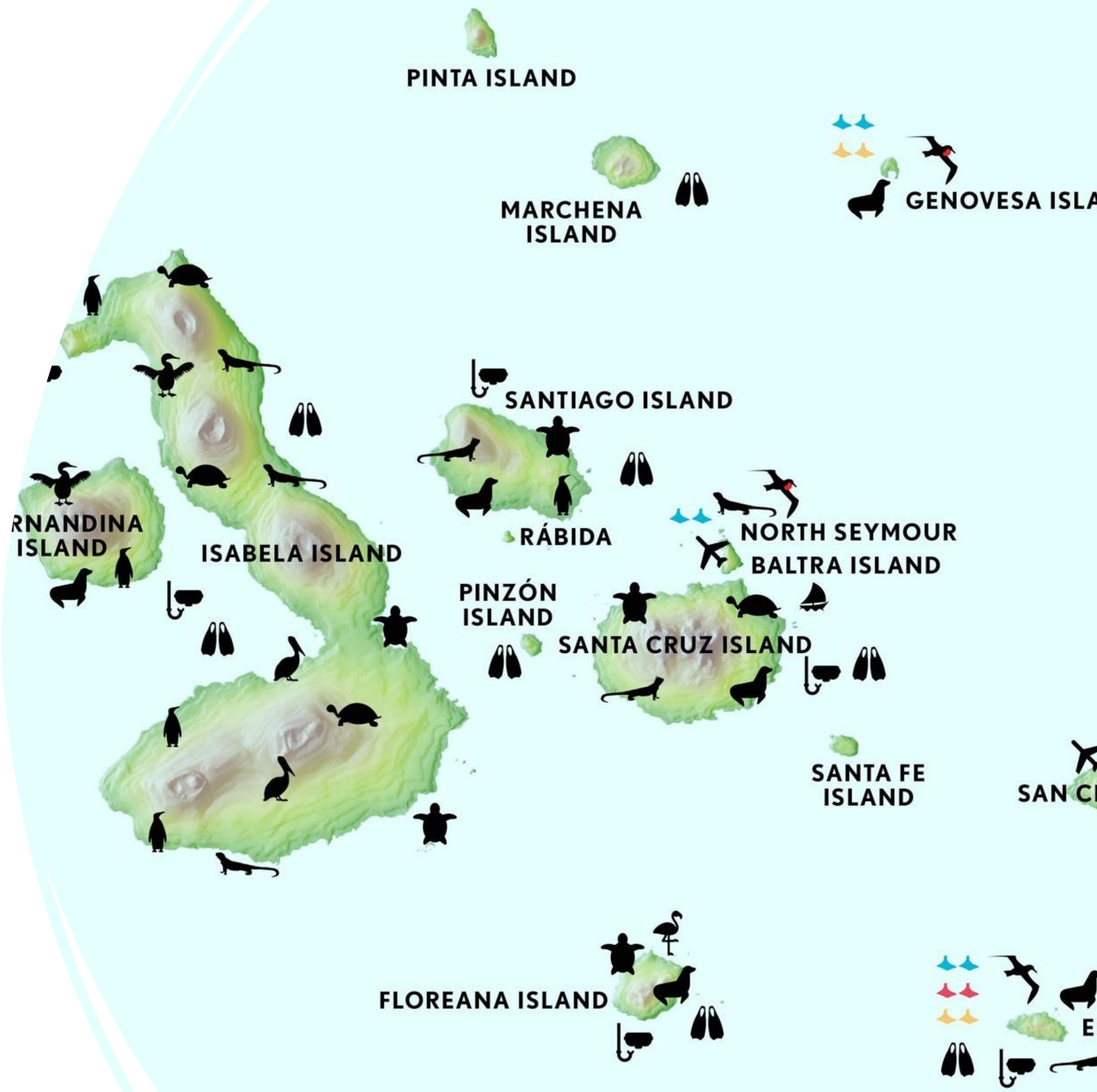
Community 2

A: 80% B: 5% C: 5% D: 10%



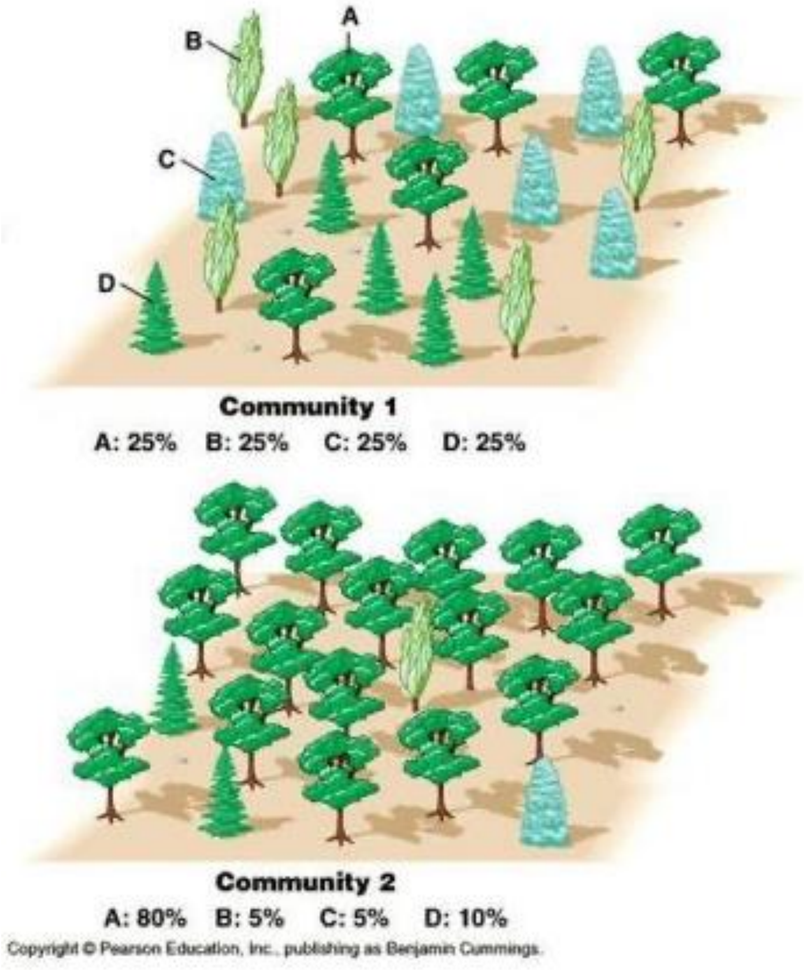
Diversity

- **Alpha diversity**: number of species on each island
- **Beta diversity**: species change between islands
- **Gamma diversity**: species on all islands



Some important terms

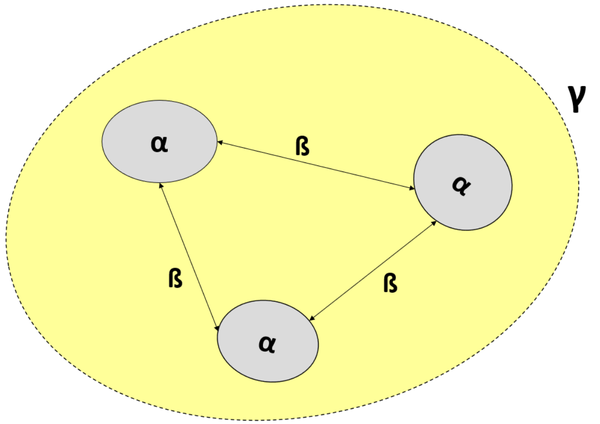
- **Species richness**
  - the *number of species* in an area
- **Species abundance**
  - the *number of individuals* of each species in an area
- **Species diversity**
  - is a term used for *the different number of species* in an area, their *abundance,* and *the distribution of these species* in that particular ecosystem



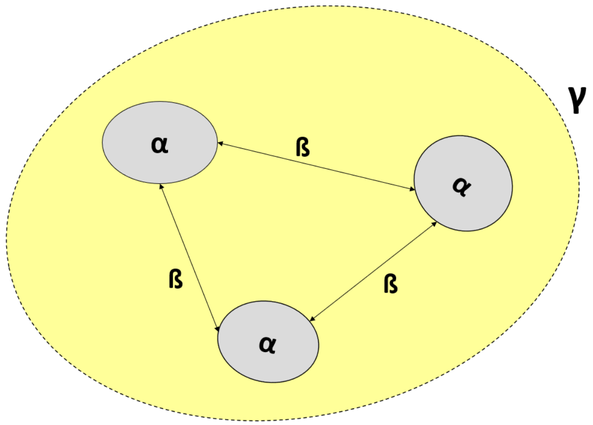
Diversity

- **Alpha**
  - Richness: number of species in a location/sample
  - Evenness: relative species abundance in a location/sample
- **Beta**
  - Species turnover across locations/time points/samples
- **Gamma**
  - Species in all analysed locations/samples

Robert Whittaker



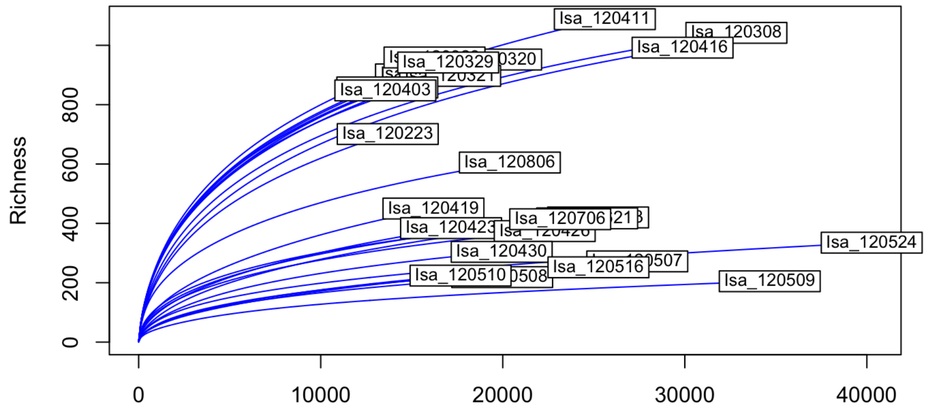
Alpha diversity



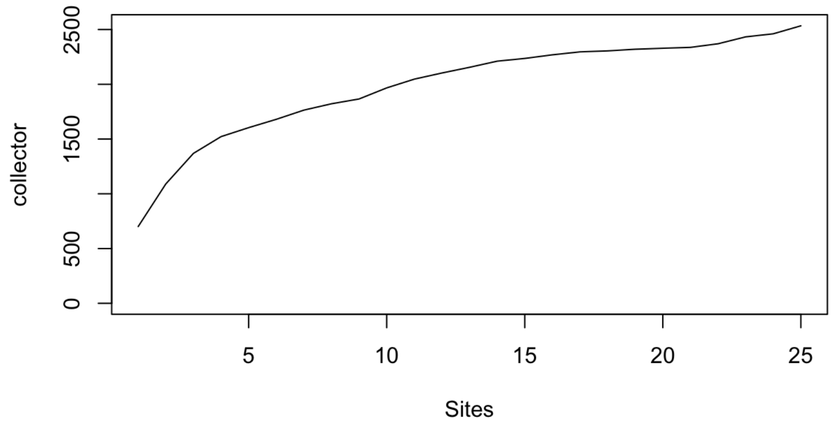
- Number of species in specific samples/location
- Alpha diversity describes the species diversity within a community at a small scale or local scale

Per sample - Alpha

- Rarefaction curves to see if the sequencing effort is sufficient
    - How many new OTUs appear in each sample as we add more reads?

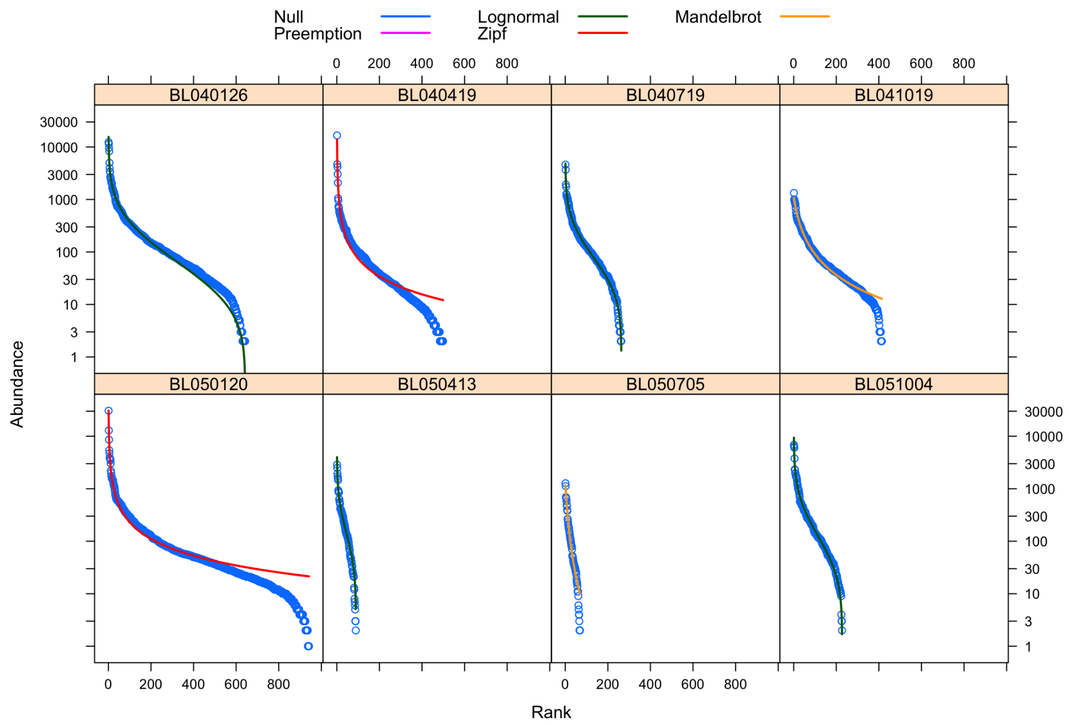


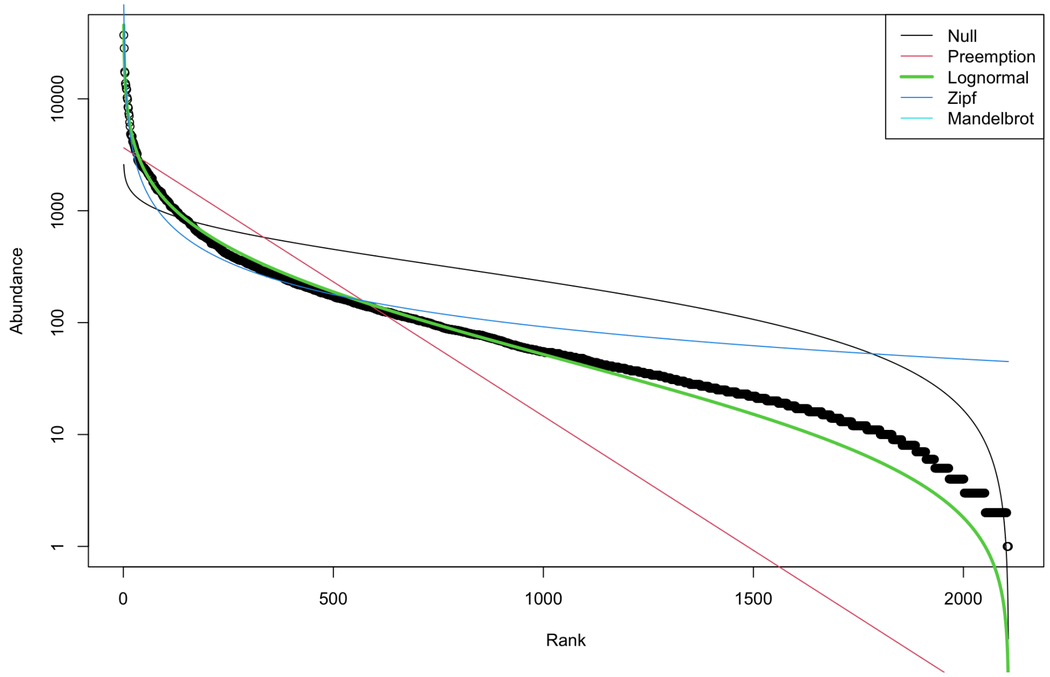
- Richness accumulation curves
    - How much is the richness increasing as we add more samples? Increase in OTUs.



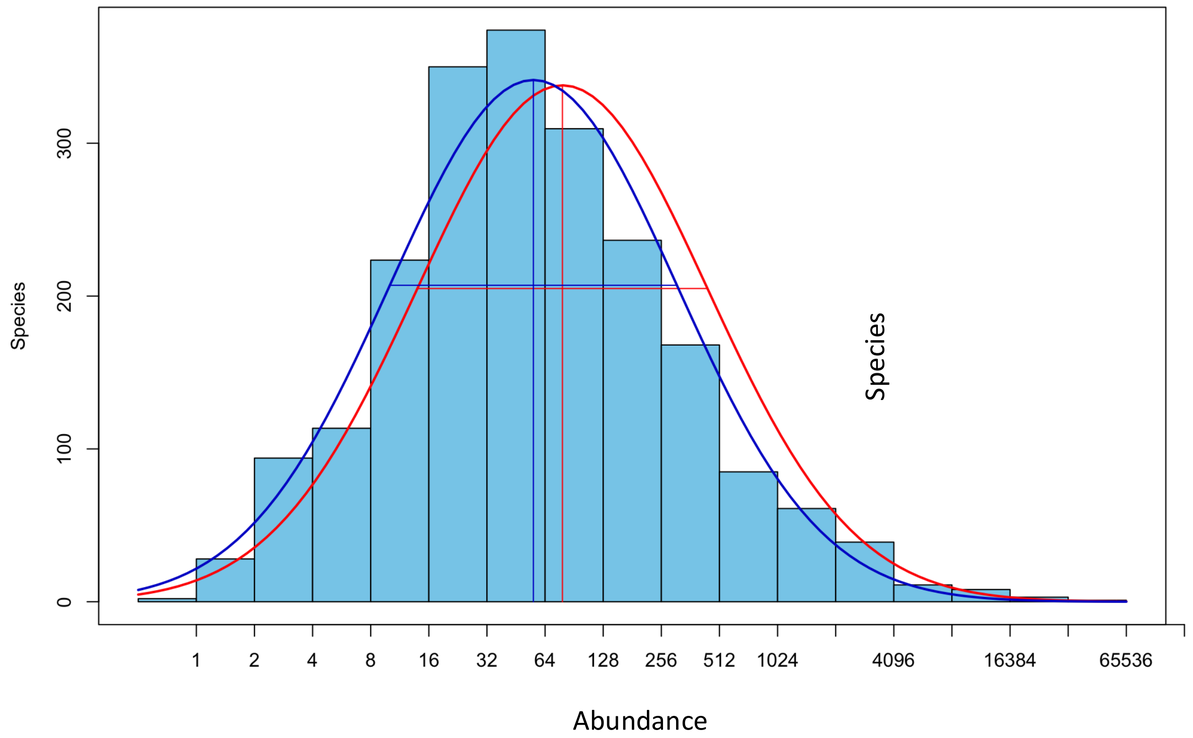
Fitting rank-abundance distributions

- How are the species (OTUs) distributed in each sample/timepoint (alpha) or across all samples (gamma).





- Frank W. Preston (1948) proposed that species abundances (when binned logarithmically) follow a normal distribution.
- This leads to a lognormal abundance distribution.
- <u>Before using ASVs, these plots were rarely observed for microbial data</u>



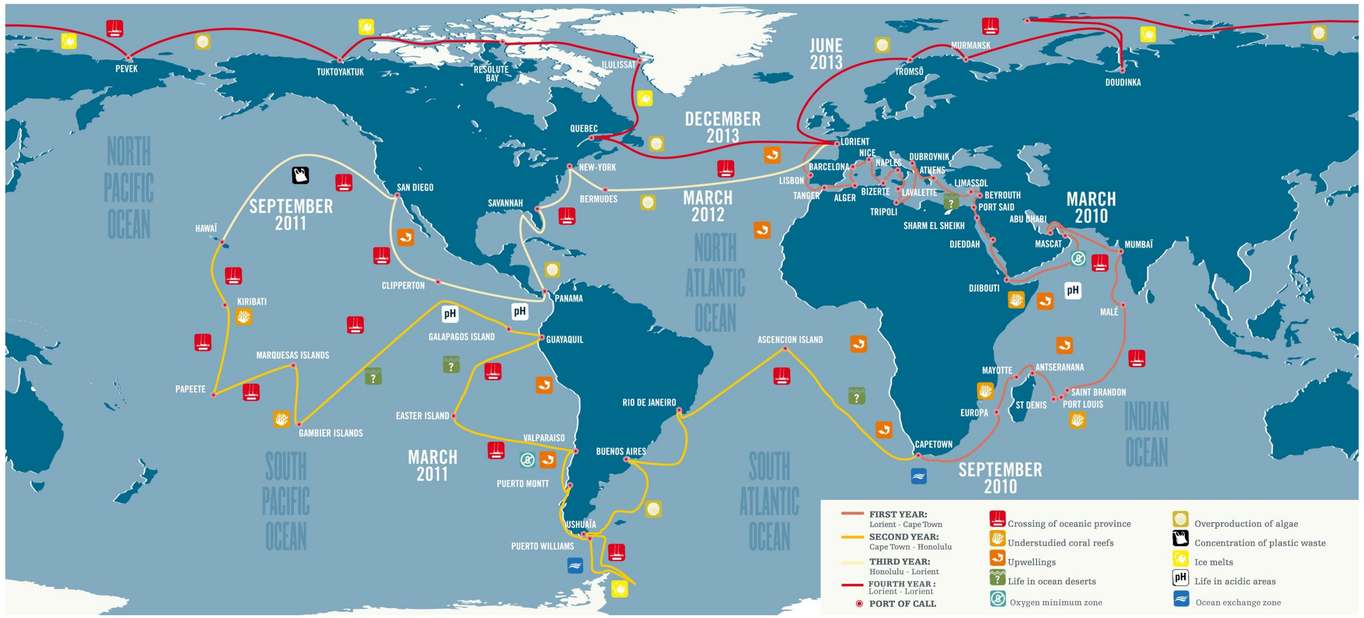
RESEARCH ARTICLE

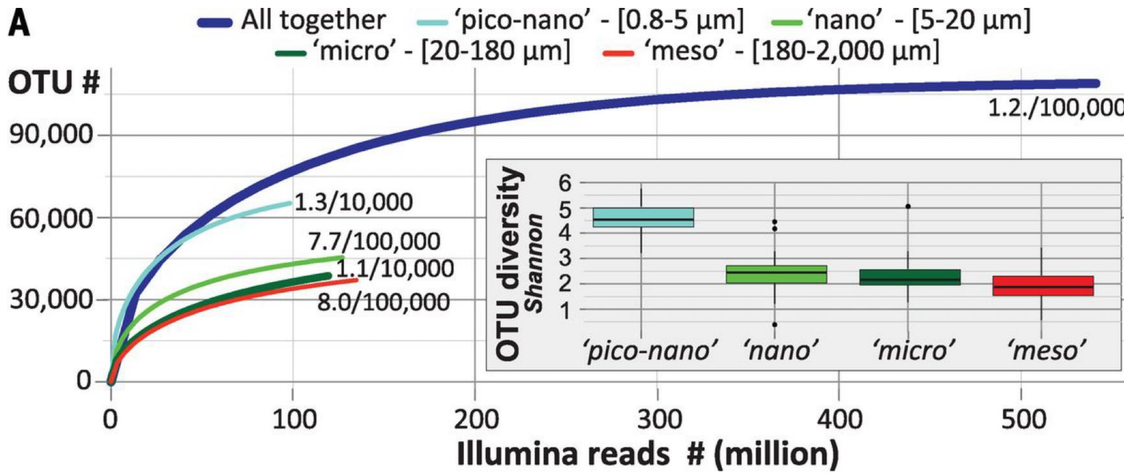
Eukaryotic plankton diversity in the sunlit ocean

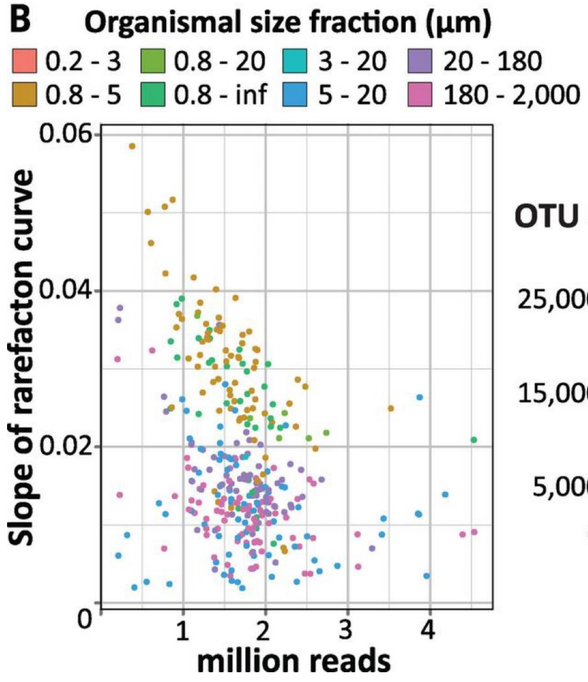
Colomban de Vargas[1,2,*,†], Stéphane Audic[1,2,†], Nicolas Henry[1,2,†], Johan Decelle[1,2,†], Frédéric Mahé[3,1,2,†], Ramiro Logares...

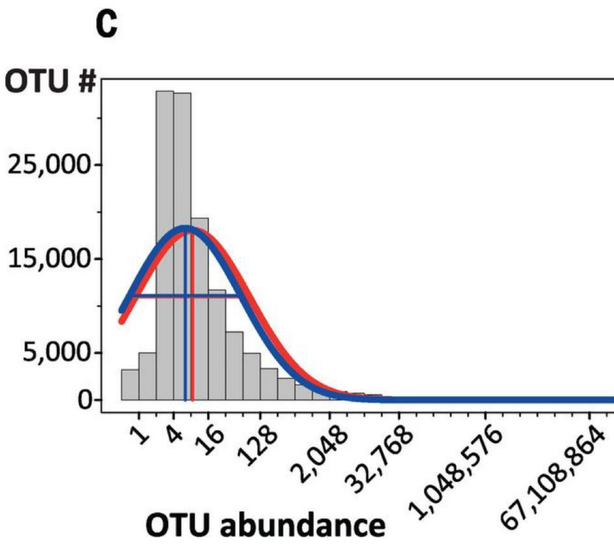
+ See all authors and affiliations

Science 22 May 2015:
Vol. 348, Issue 6237, 1261605
DOI: 10.1126/science.1261605









18S V9 (Swarm)

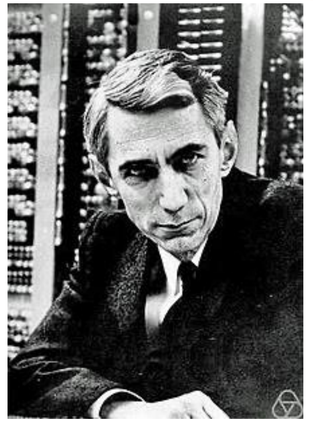
Diversity Indices

Shannon index H (entropy).

A diversity index considers the number species (richness) as well as the number of individuals per species (evenness). Varies from

- 0 for communities with only a single taxon to
- high values for communities with many taxa, each with few individuals

$$H' = -\sum_{i=1}^{R} p_i \ln p_i$$



Claude Shannon

-***Pielou's index of evenness***

A calculated value of Pielou's evenness ranges from

- 0 (no evenness) to
- 1 (complete evenness).

Derived from Shannon (H)

$$J' = \frac{H'}{H'_{\max}}$$

E.C. Pielou

Inverse Simpson's D index

The value of Simpson's D ranges from 0 to 1, with 0 representing infinite diversity and 1 representing no diversity, so the larger the value of D, the lower the diversity.

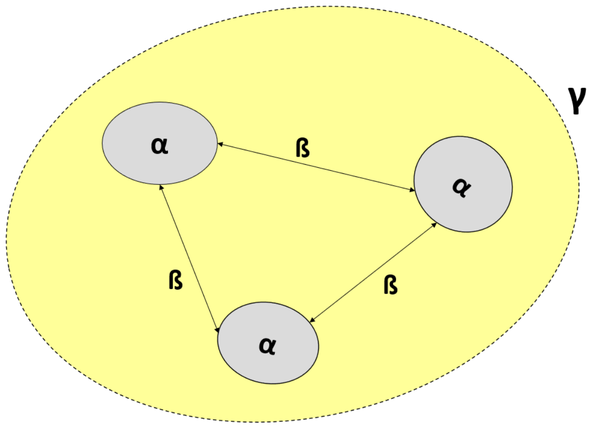
For this reason, Simpson's index is usually expressed as its inverse.

- High value = higher diversity

$$\frac{1}{\lambda} = \frac{1}{\sum_{i=1}^{R} p_i^2} = {}^2D$$

Edward Simpson

Beta diversity



- Species turnover between samples/location
- Beta diversity analyses measure how communities change between different samples (time points, locations, etc.)
- Analyses can be biased if different samples have different sequencing efforts (or sequencing depth)
- High throughput sequencing (HTS/NGS) data are compositional

HTS data are compositional

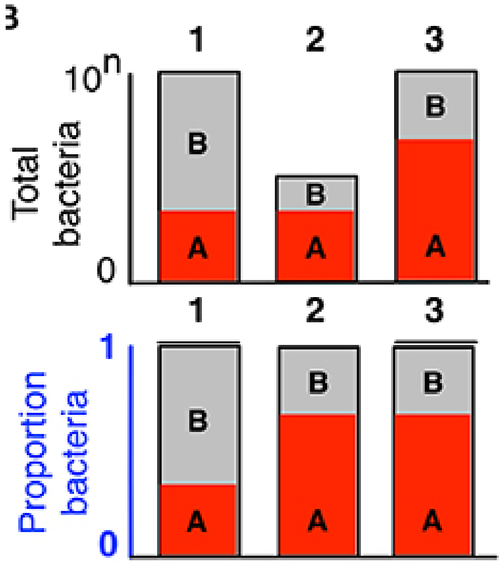
- HTS datasets are compositional, due to the total limits imposed by sequencers. Then, the increase of one OTU means the decrease of another OTU in the HTS dataset.
- The total read count in an HTS run is a fixed-size, random sample of the relative abundance of the molecules in the underlying ecosystem.
- The count cannot be related to the absolute number of molecules in the input sample.

HTS data are compositional

- This is implicitly acknowledged when microbiome datasets are converted to relative abundance values, or normalized counts, or are rarefied
- Data described as proportions or probabilities, or with a constant or irrelevant sum, are referred to as compositional data
- Compositional data is all about the relationships between the parts (can't inform on absolute abundances of molecules)
- The abundance of one OTU is only interpretable relative to another

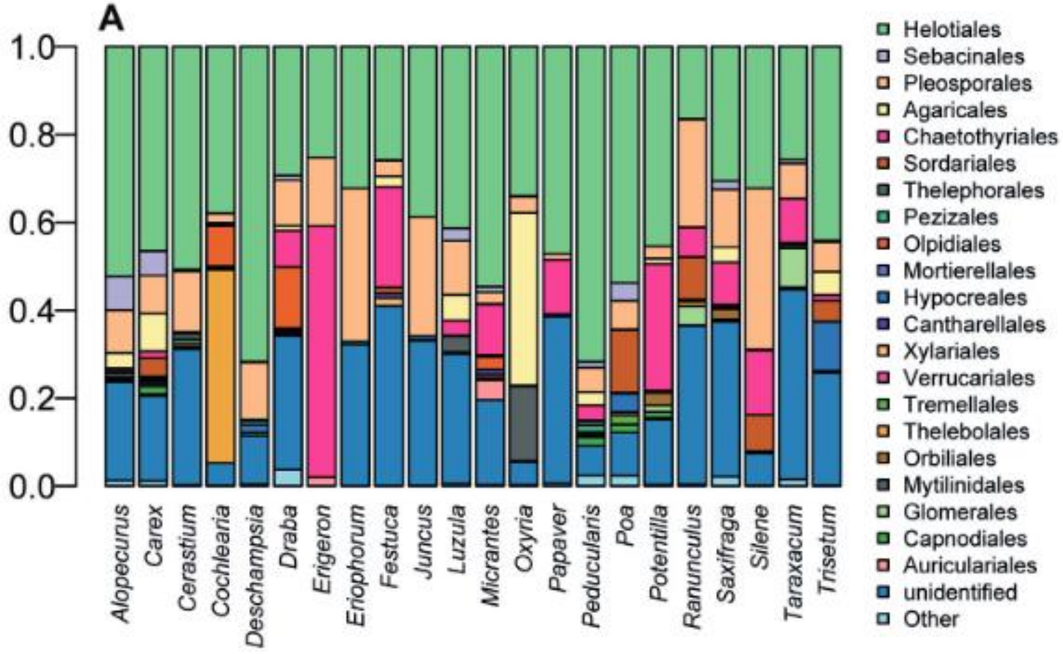
HTS data are compositional

- After samples are sequenced we lose the absolute count information and only have relative abundances, proportions, or “normalised counts”



Absolute abundance before sequencing

Relative abundance after sequencing



Modified from: Gloor et al. 2017, Microbiome Datasets Are Compositional: And This Is Not Optional

Botnen et al. FEMS Microbiology Ecology, 96, 2020, fiaa185

HTS data have different sequencing depth

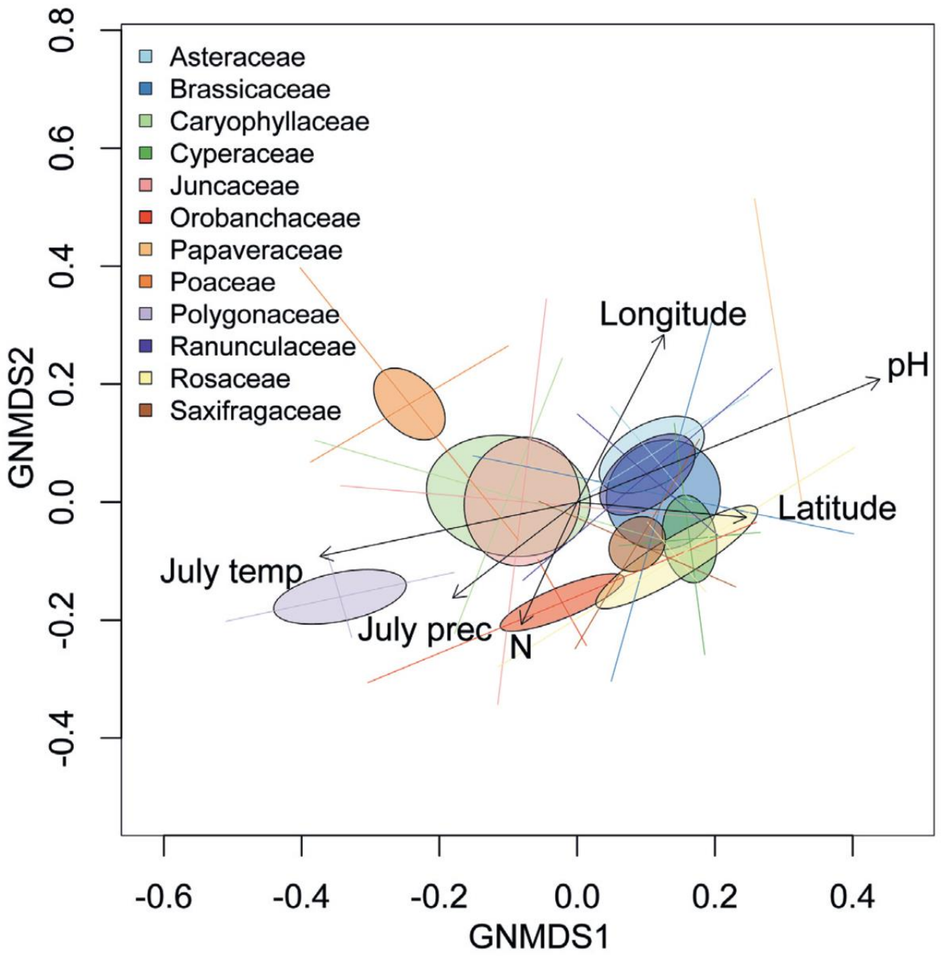
- Different sequencing depths may bias the calculation of distances for multivariate analyses.
- One way to mitigate this is to subsample or “rarefy” samples to the same sequencing depth.
    - But, it has been criticized due to loss of information and precision
- Function rrarefy() in vegan generates a randomly rarefied (without replacement) community data frame or vector of a given sample size.
- Typically rarefied to the size of the sample with the fewest reads.

Other ways of transforming

- The clr transformation is becoming more popular
  - centered log-ratio
- Ratio transformation of the data: captures relationships between features (e.g. OTUs)
- Taking the logarithm of the ratios (log-ratios) makes data symmetric and linearly related
- In a composition, all components (OTUs) are mutually dependent, and can not be understood in isolation
- Analyses of individual components (OTUs) is done with respect to a reference
- The clr transformation uses the geometric mean as a reference

Computer Lab II: Distance measures Ordination, and clustering

- **Keyterms: Bray-curtis, PCA, NMDS, UPGMA**

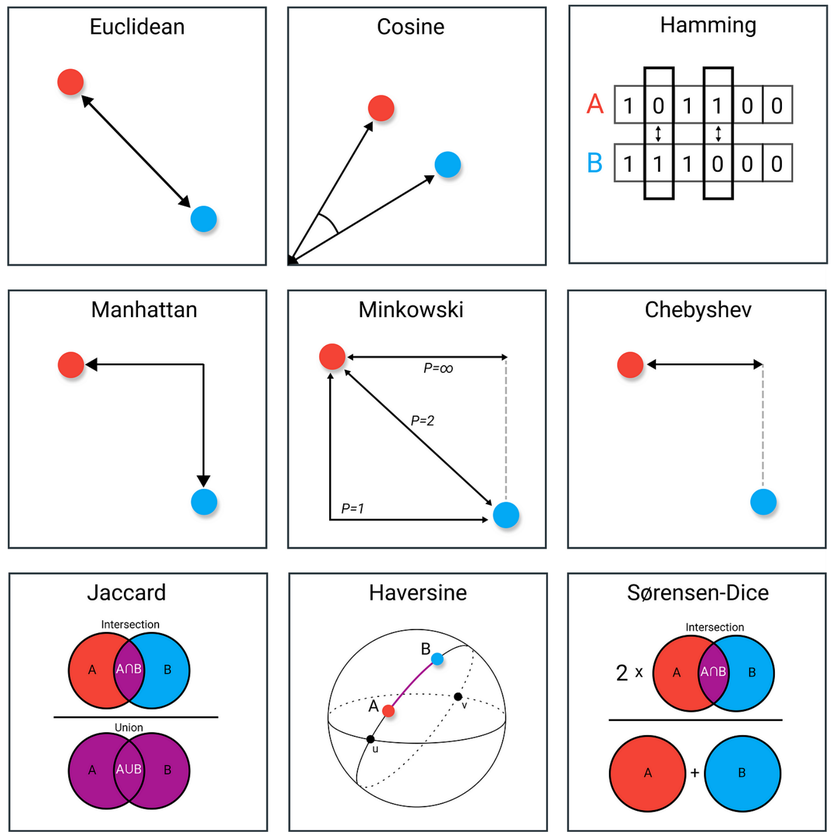


Distance metrics

- Statistical distance: distance between variables
- Distance metrics in ecology: allow measuring the dissimilarity between communities composed by several species (OTUs)
- Several distance metrics available in R
- Often used: Euclidean, Jaccard, Sorensen, Simpson, Bray Curtis
- BC between two sites j and k is

$$BC_{ij} = \frac{\sum |x_{ik} - x_{jk}|}{\sum (x_{ik} + x_{jk})}$$

- Where x_{ik} and x_{jk} are the abundances of species *k* in sample *i* and *j*



https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

Distance metrics

- Bray Curtis example

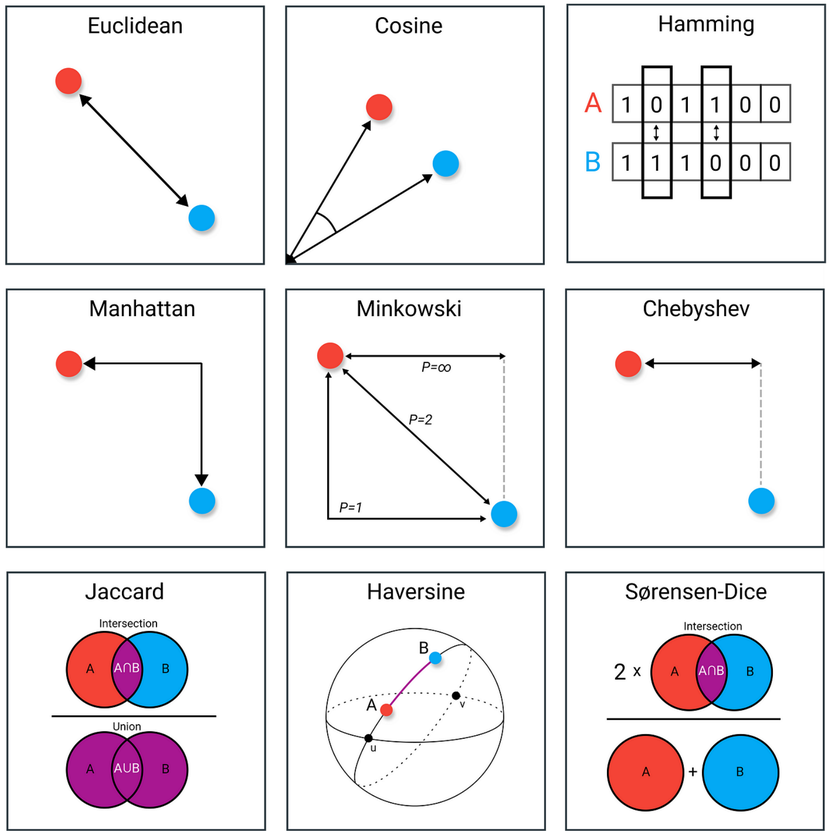
1. Compute the absolute differences between the counts for each species:
   - For Species A: |5−2|=3
   - For Species B: |3−6|=3
   - For Species C: |7−7|=0
2. Sum the absolute differences: 3+3+0=6
3. Compute the total counts (sums) for both samples:
   - For Species A: 5+2=7
   - For Species B: 3+6=9
   - For Species C: 7+7=14
4. Sum the total counts: 7+9+14=30
5. Finally, apply the Bray-Curtis formula: BC=6/30=0.2

Species	Sample 1	Sample 2
A	5	2
B	3	6
C	7	7

$$BC_{ij} = \frac{\sum |x_{ik} - x_{jk}|}{\sum (x_{ik} + x_{jk})}$$

Distance metrics

- It is important to think what is the most appropriate distance metric for the data
- Different distance metrics can have different ranges
- Bray-Curtis: influenced by abundant taxa (but normally used)
  - Ranges between 0-1 (1 most dissimilar)
- Euclidean and Sorensen: influenced by large differences in species abundances, data sparsity (lots of zeros) and many observations
- Euclidean: no upper limit of values



https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

Ordination

- Is a collective term for multivariate techniques which summarise a multidimensional dataset in such a way that when it is projected onto a two dimensional space any intrinsic pattern the data may possess becomes apparent upon visual inspection (Pielou, 1984)
- In ecological terms: ordination serves to summarise community data (such as species abundance data) by producing a *low-dimensional ordination space in which similar species and samples are plotted close together, and dissimilar species and samples are placed far apart.*

Ordination

- Ordination is used in ecology to investigate relationships between species composition patterns and environmental variability.
- Often, these techniques are used to address the question: what environmental variables shape communities?
- The relative importance of environmental gradients in shaping communities can be estimated

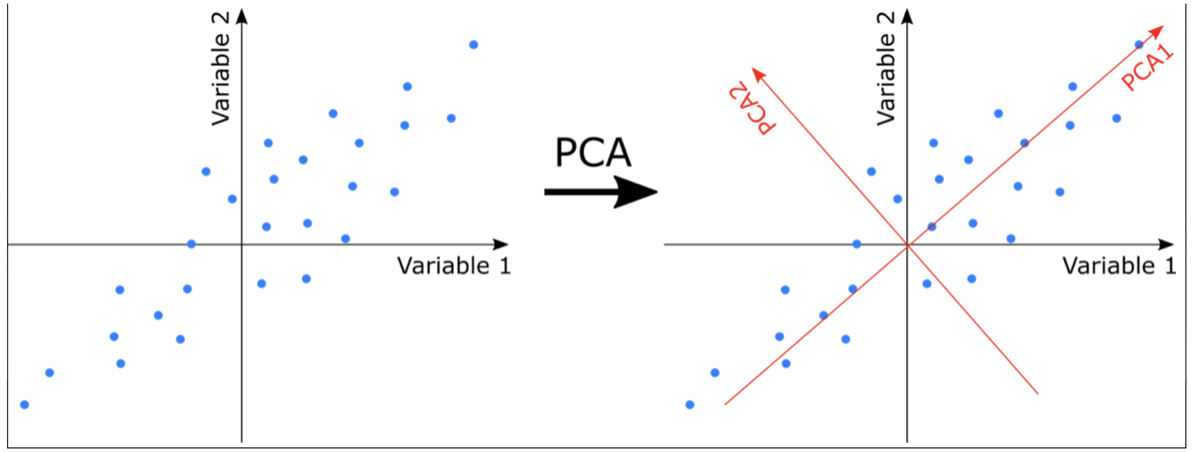
Ordination techniques

Two commonly used unconstrained techniques

- Principal Component Analysis (PCA)
  - Assumes linear relationship between features, variances, and correlations (i.e. euclidean space)
  - Deterministic, always the same outcome for the same dataset
  - Orthogonal linear transformation
- Non-metric Multidimensional Scaling (NMDS)
  - Does not assume linear relationship between features, which allows the use of any (suitable) distance measure (for instance bray-curtis)
  - NMDS relies on rank orders (distances) for ordination
  - Iterative and non-parametric process
  - Outcome will vary for each run

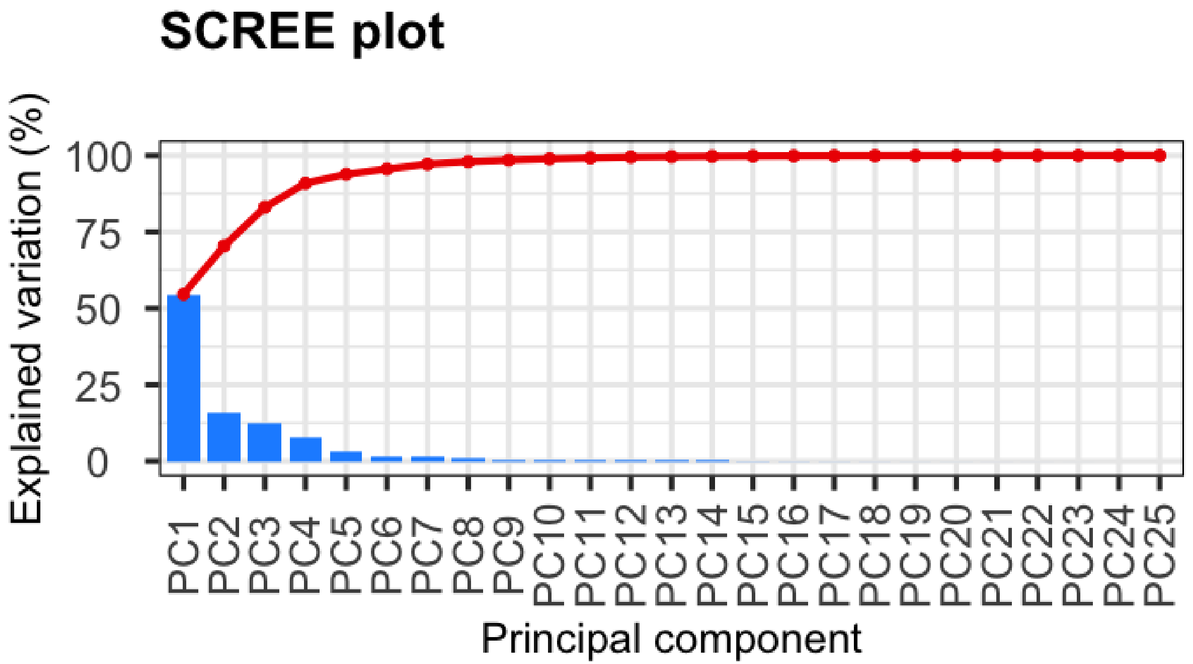
PCA

- A rotation of the coordinate axes, chosen such that each successful axis captures as much variance as possible
- The first principal component (PC) will be placed in the direction of the maximum variability and subsequent axes (PCs) will be generated in the same manner



PCA

- Typically the first 1-4 axes represents the majority of the variation in the data set. The rest is often considered noisy



Percentage of variance explain by each PC

Non-metric Multidimensional Scaling (NMDS)

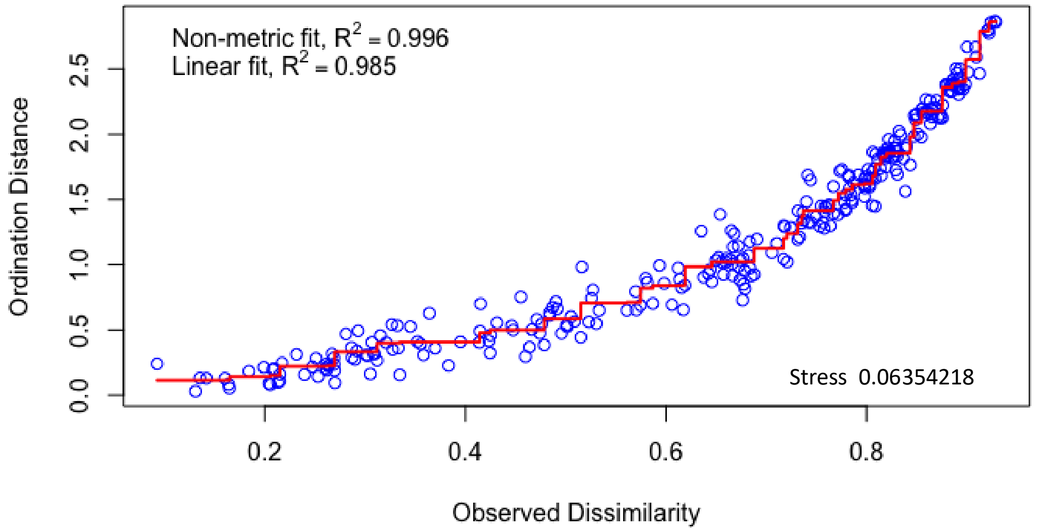
- NMDS is more robust than PCA (e.g. is not affected by the arch effect)
- NMDS attempts to represent the pairwise dissimilarity between objects in a low-dimensional space
- Any distance metric can be used to build the distance matrix
- NMDS is a rank approach, meaning that distances are replaced by ranks
- The stress value indicates how well the ordination summarises the observed distances among the samples

Non-metric Multidimensional Scaling (NMDS)

- NMDS differs from PCA in that:
- There is not a unique ordination result (thus, algorithms run NMDS multiple times)
- The axes of the ordination are not ordered according to the variance they explain (but metaMDS() in Vegan rotates final results to make Axis 1 correspond to the greatest variance among samples)
- The number of dimensions of the low-dimensional space must be specified before running NMDS
- Plotting stress (goodness of fit) vs. dimensionality can be used to assess the choice of dimensions. Stress values should be <0.2. We choose the minimum number of dimensions

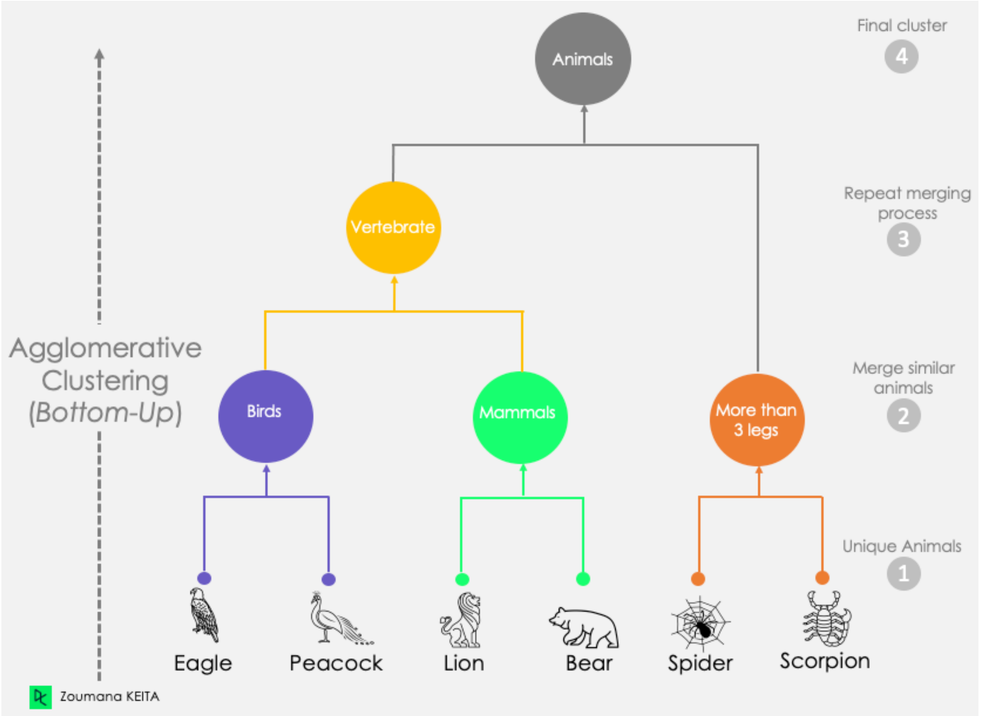
Non-metric Multidimensional Scaling (NMDS)

- Running NMDS:
  - Step 1: run NMDS with e.g. 2 to 10 dimensions
  - Step 2: Check stress vs. dimension plot
  - Step 3: Choose optimal number of dimensions (typically k=2)
  - Step 4: Check for convergent solution and final stress



Hierarchical Clustering

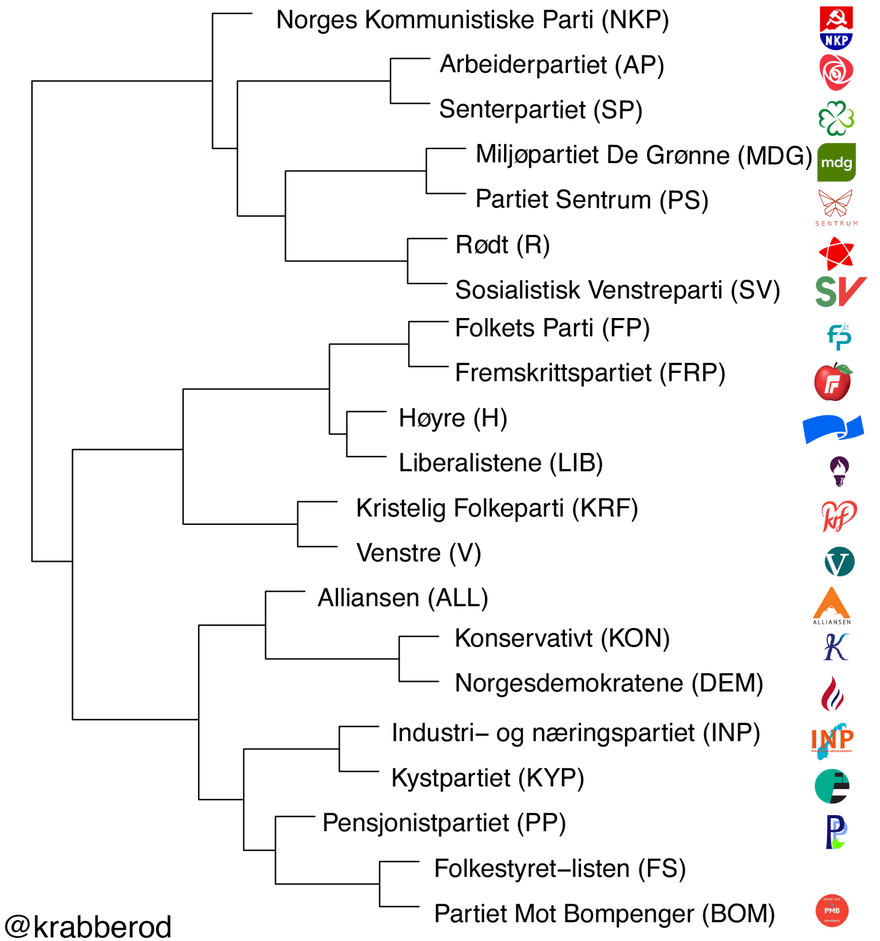
- Hierarchical clustering: groups are organised in ranks according to their similarity
- UPGMA (unweighted pair group method with arithmetic mean): agglomerative (bottom-up) hierarchical clustering
- Based on a dissimilarity matrix



https://www.datacamp.com/tutorial/introduction-hierarchical-clustering-python

Hierarchical Clustering

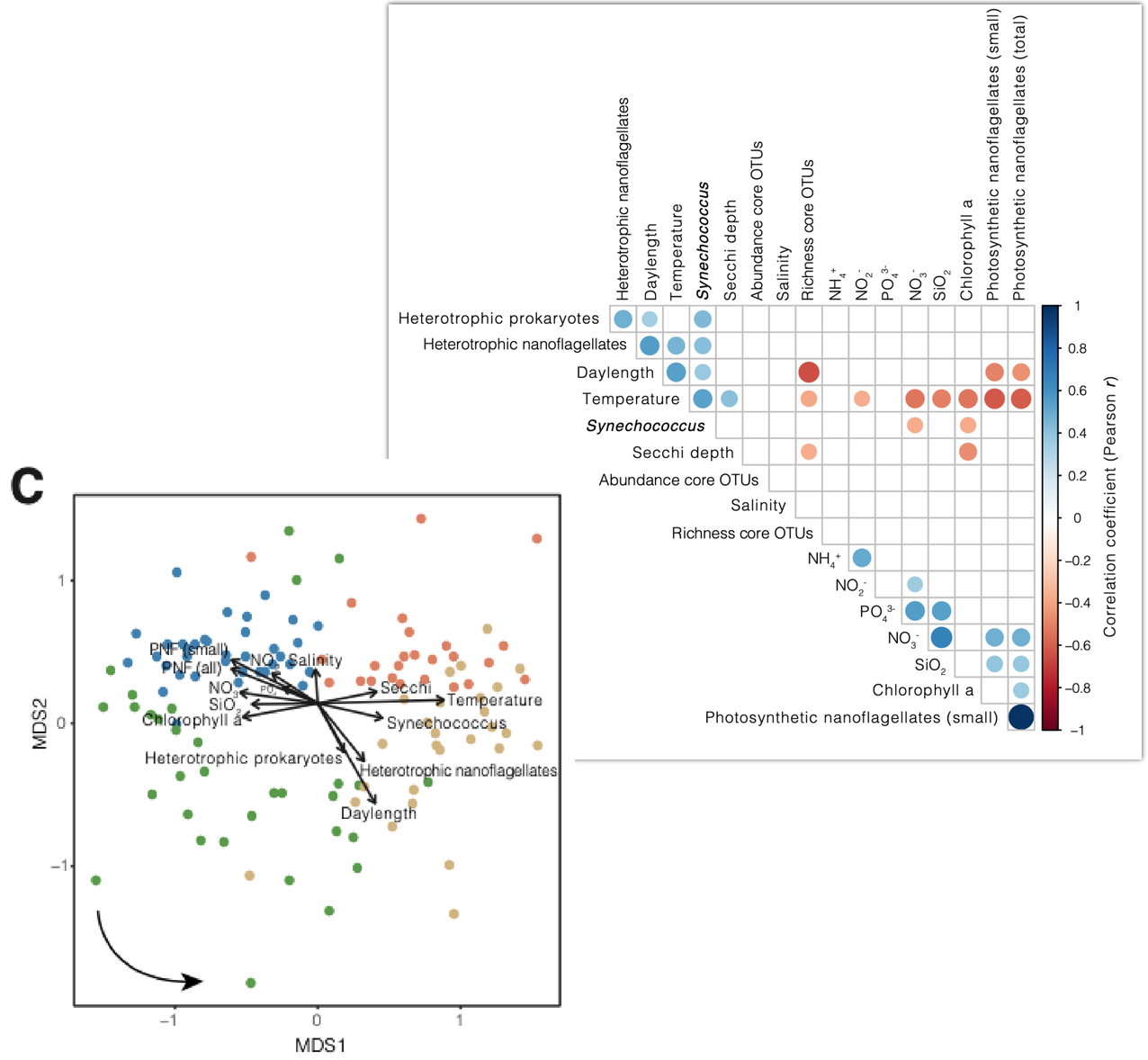
- Hierarchical clustering: groups are organised in ranks according to their similarity
- UPGMA (unweighted pair group method with arithmetic mean): agglomerative (bottom-up) hierarchical clustering
- Based on a dissimilarity matrix



Hierarkisk klyngeanalyse av partiene som stiller til valg i Oslo, basert på NRK sin valgomat med 31 påstander. For hver påstand har NRK plassert partiene i en av fire kategorier avhengig av hvordan partiet stiller seg til påstanden. Kategoriene har i denne klyngeanalysen fått poengene: «helt uenig» (-2), «litt uenig» (-1), «litt enig» (+1) og «helt enig» (+2). Forholdene mellom partiene har så blitt regnet ut med metoden «fullstendig lenking».

Computer Lab III: Incorporating environmental data

- **Keyterms:**
  Unconstrained and
  Constrained ordination



Krabberød et al. 2022. doi.org/10.1186/s40793-022-00417-1

Environmental data

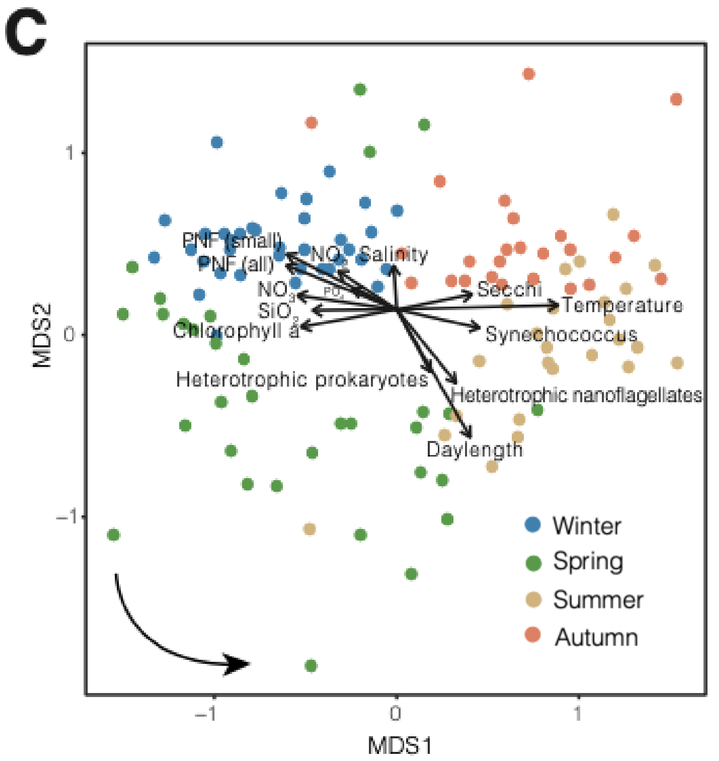
- Our goal is to determine whether environmental variability can explain the observed variance in community composition.
- Environmental variables are standardized to ensure they have comparable ranges, allowing for meaningful comparison and analysis
- Typically centered and scaled:
  - Center: subtracting the mean of each column from the respective column values.
  - Scale: dividing each column by its standard deviation.
- Z-scores represent how many standard deviations a value is from the mean. Standardizing in this way centers the data around 0 (mean) and scales it to have a standard deviation of 1.

Unconstrained vs. Constrained ordination

- In unconstrained ordination, the goal is to identify the main patterns of compositional variation in the data, which can later be linked to environmental factors (e.g., using envfit).
- In contrast, constrained ordination focuses specifically on the variation that can be explained by a predefined set of environmental variables, rather than capturing the full spectrum of compositional variation.
- The constrained ordination is non-symmetric: we have independent variables or constraints (environmental data), and we have dependent variables or the community (otu matrix).
- This is called non-symmetric because the model assumes that the independent variables drive or influence the variation in the dependent variables, not the other way around.

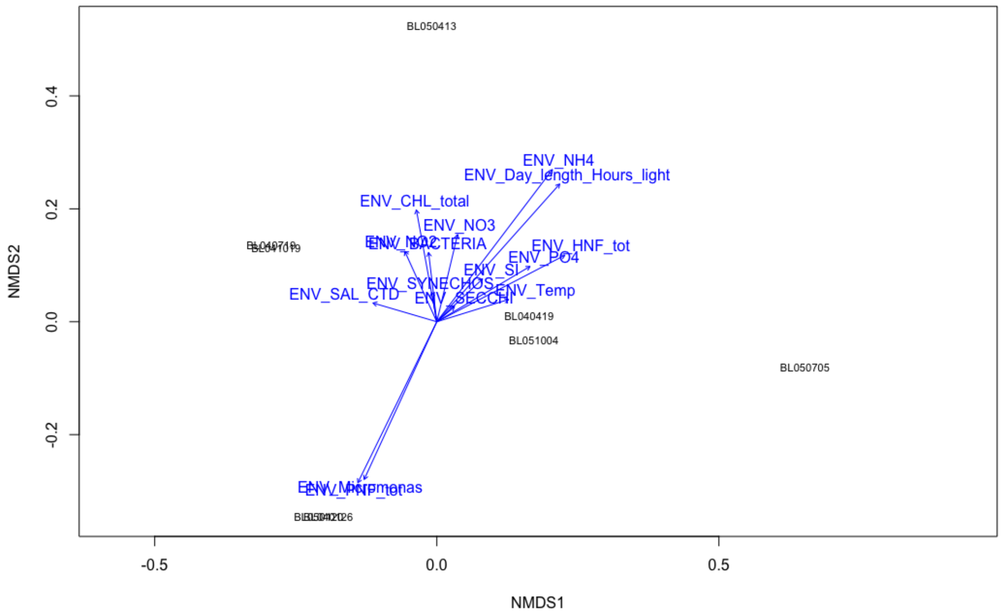
Fitting environmental data to ordination

- Environmental variables are correlated with the ordination axes to understand how they influence community composition.
- The arrow points in the direction of the steepest change in the environmental variable, often referred to as the “gradient direction.”
- The length of the arrow reflects the strength of the correlation between the ordination axes and the environmental variable, indicating the intensity of the gradient. Longer arrows suggest stronger relationships.

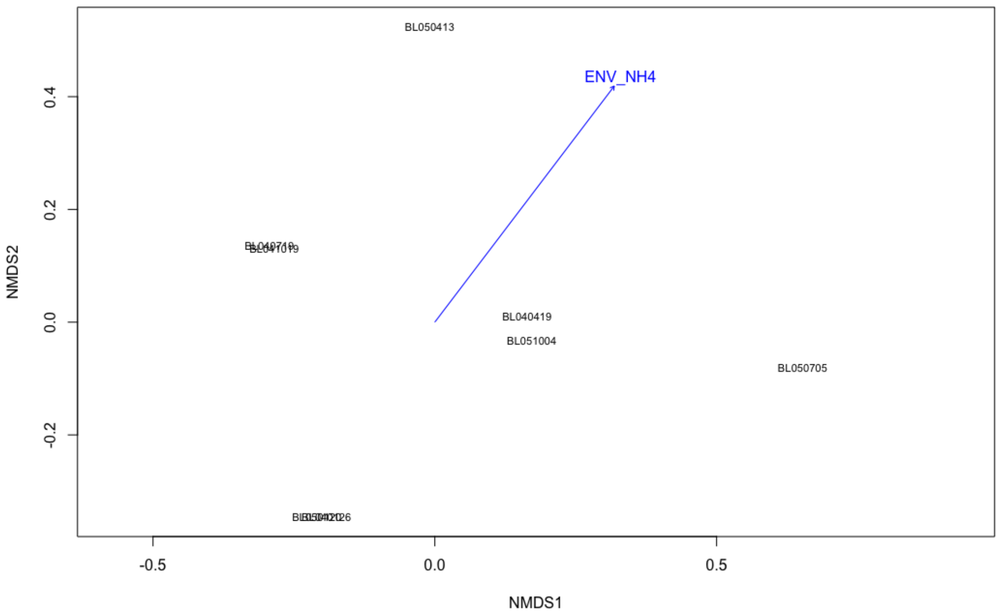


Krabberød et al 2023
Environmental Microbiome doi: 10.1186/s40793-022-00417-1

All



P < 0.05



vegan (version 2.4-2)

envfit: Fits an Environmental Vector or Factor onto an Ordination

Description

The function fits environmental vectors or factors onto an ordination. The projections of points onto vectors have maximum correlation with corresponding environmental variables, and the factors show the averages of factor levels.

Usage

```
"envfit"(ord, env, permutations = 999, strata = NULL,  choices=c(1,2),  display = "sites", w  = weights(ord), na.rm = FALSE, ...)
"envfit"(formula, data, ...)
"plot"(x, choices = c(1,2), labels, arrow.mul, at = c(0,0),  axis = FALSE, p.max = NULL, col = "blue", bg, add = TRUE, ...)
"scores"(x, display, choices, ...)
vectorfit(X, P, permutations = 0, strata = NULL, w, ...)
factorfit(X, P, permutations = 0, strata = NULL, w, ...)
```

Model selection

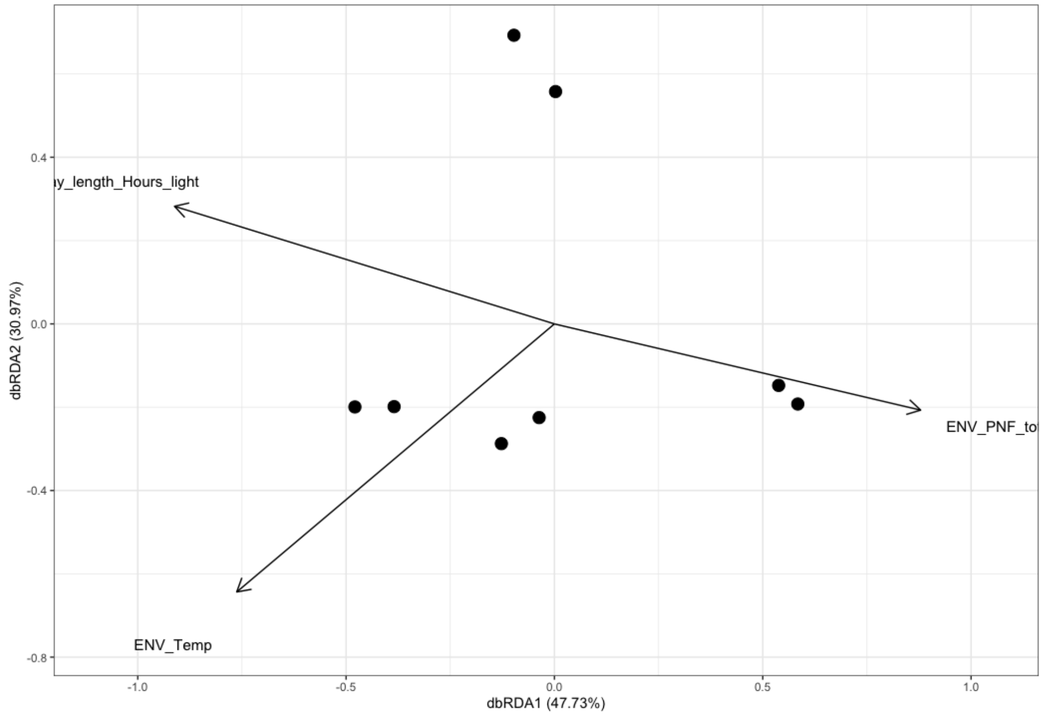
- Which environmental variables should be included in a model?
- To select environmental variables that explain most of the community variance:
  - **Forward selection**: Start with an empty model and add variables one by one. At each step, add the variable that gives the best improvement to the model.
  - **Backward elimination**: Start with a full model including all variables. Remove variables with low explanatory power, one at a time.

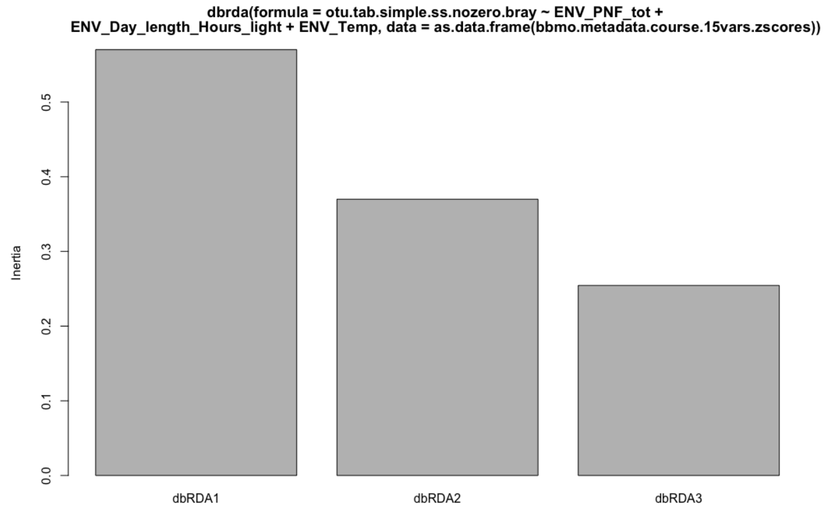
Example of a model:

Community ~ silicate + temperature + salinity + phosphate

Model selection

- The function Ordistep () in Vegan performs a step-wise selection of environmental variables based on two criteria:
  - The inclusion of a variable significantly increases the explained variance (tested using permutation tests)
  - The Akaike Information Criterion (AIC) decreases when the variable is added (in models like redundancy analysis or canonical correspondence analysis).
    - AIC: estimates the quality of models relative to other models (model selection). It is an estimator of prediction error, taking into account the number of parameters.
- We are typically not interested in variables that do not explain the variation in community composition.


0.4
0.0
-0.4
-0.8
dbRDA2 (30.97%)
y_length_Hours_light
ENV_PNF_to
ENV_Temp
-1.0
-0.5
0.0
0.5
1.0
dbRDA1 (47.73%)


dbrda(formula = otu.tab.simple.ss.nozero.bray ~ ENV_PNF_tot +
ENV_Day_length_Hours_light + ENV_Temp, data = as.data.frame(bbmo.metadata.course.15vars.zscores))
Inertia
0.0
0.1
0.2
0.3
0.4
0.5
dbRDA1
dbRDA2
dbRDA3