

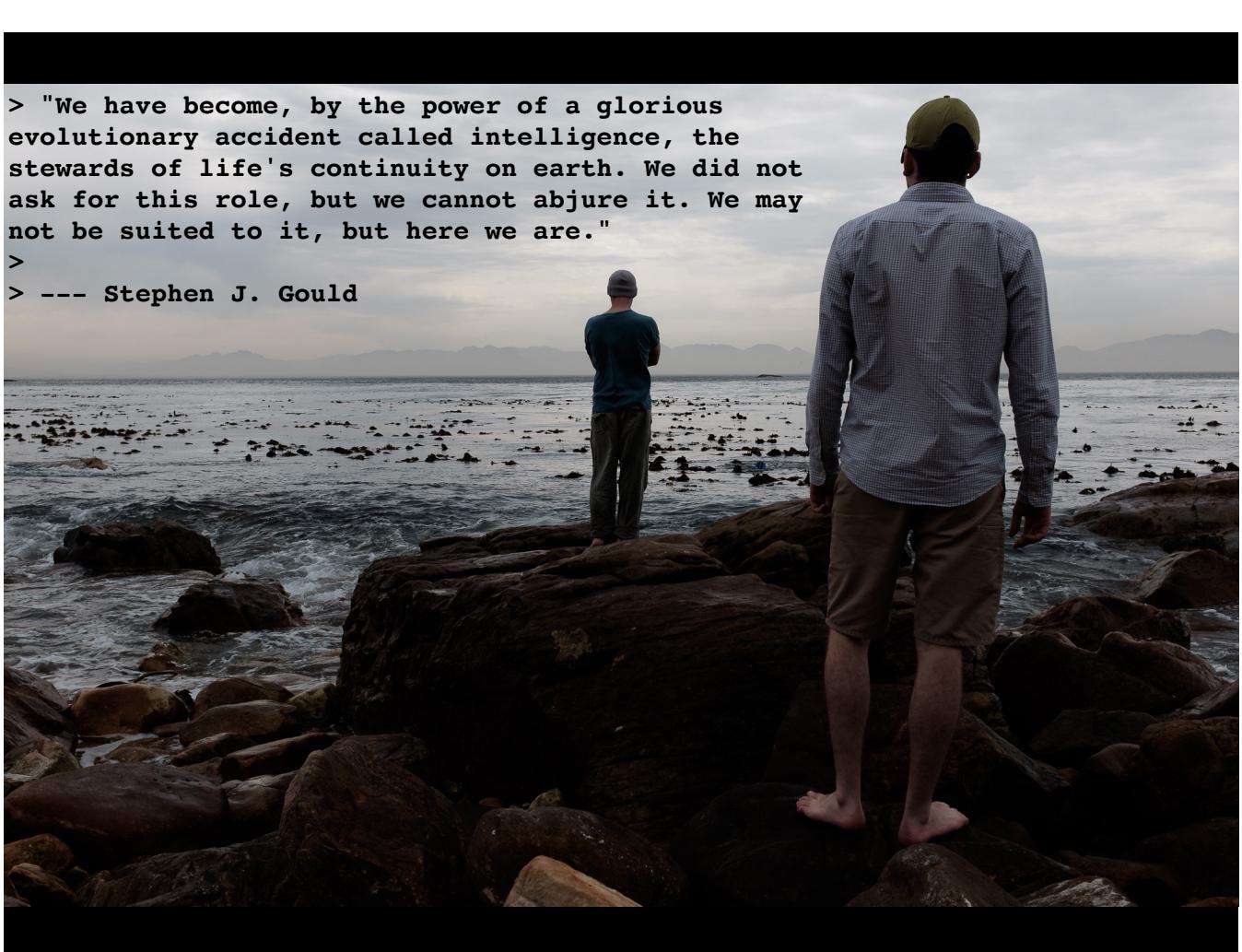
## INTRODUCTION

# QUANTITATIVE ECOLOGY



AJ Smit

© 2012 AJ Smit



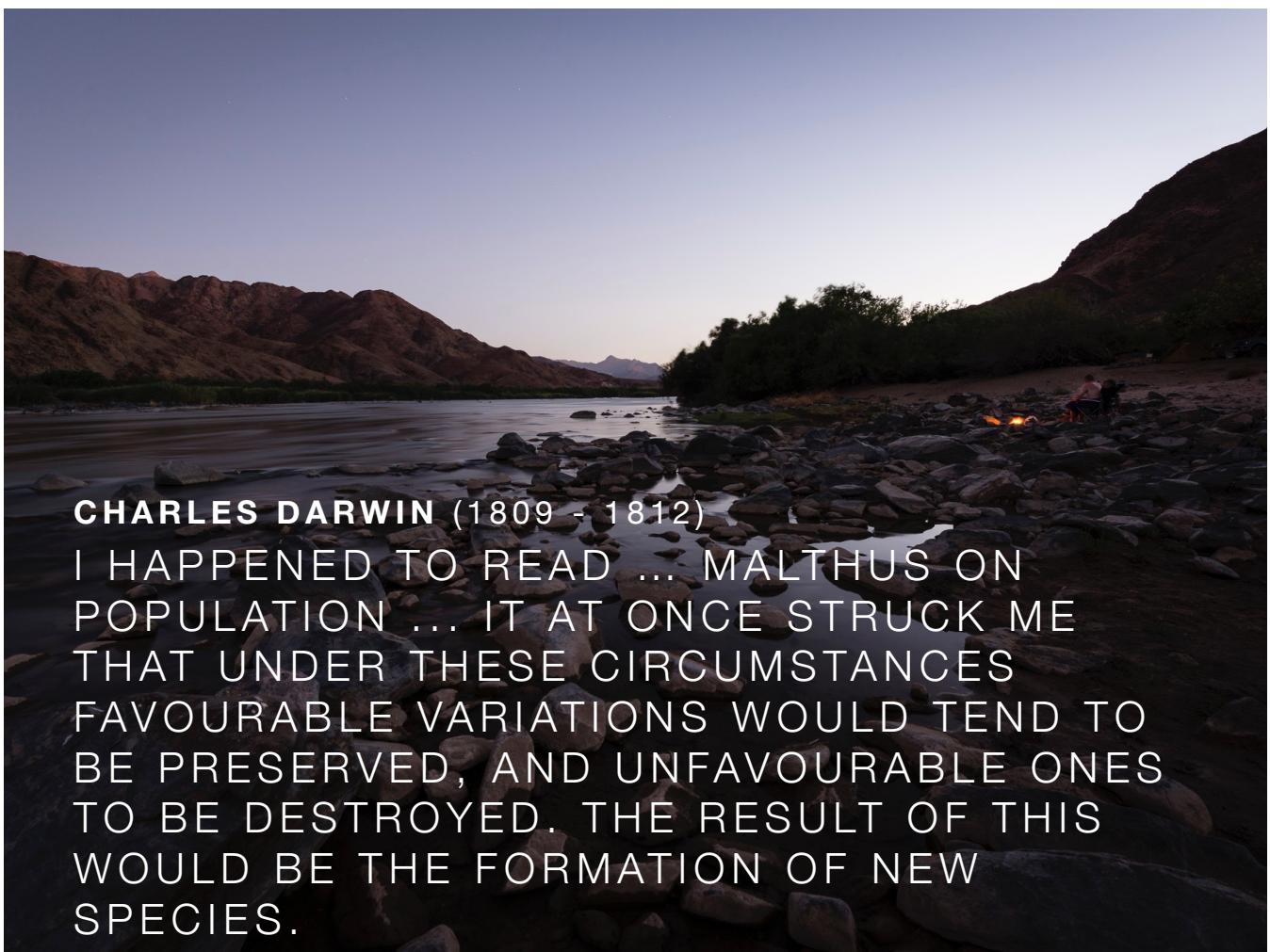
**ALFRED RUSSEL WALLACE (1823 - 1913)**

... IT WAS THE FIRST GREAT WORK I HAD  
YET READ TREATING OF ANY OF THE  
PROBLEMS OF PHILOSOPHICAL BIOLOGY, ...  
TWENTY YEARS LATER GAVE ME THE LONG-  
SOUGHT CLUE TO THE EFFECTIVE AGENT IN  
THE EVOLUTION OF ORGANIC SPECIES



**CHARLES DARWIN (1809 - 1812)**

I HAPPENED TO READ ... MALTHUS ON  
POPULATION ... IT AT ONCE STRUCK ME  
THAT UNDER THESE CIRCUMSTANCES  
FAVOURABLE VARIATIONS WOULD TEND TO  
BE PRESERVED, AND UNFAVOURABLE ONES  
TO BE DESTROYED. THE RESULT OF THIS  
WOULD BE THE FORMATION OF NEW  
SPECIES.



# Intent of Day 1

What is ecology?

Ecological questions:

What kinds of questions have been asked by ecologists?

What kinds of questions can be asked?

What has caused the difference in kinds of questions being asked?

What is QE?

Why not use the biology and statistics knowledge you already have?



© 2014 AJ Smit

## What is ecology?

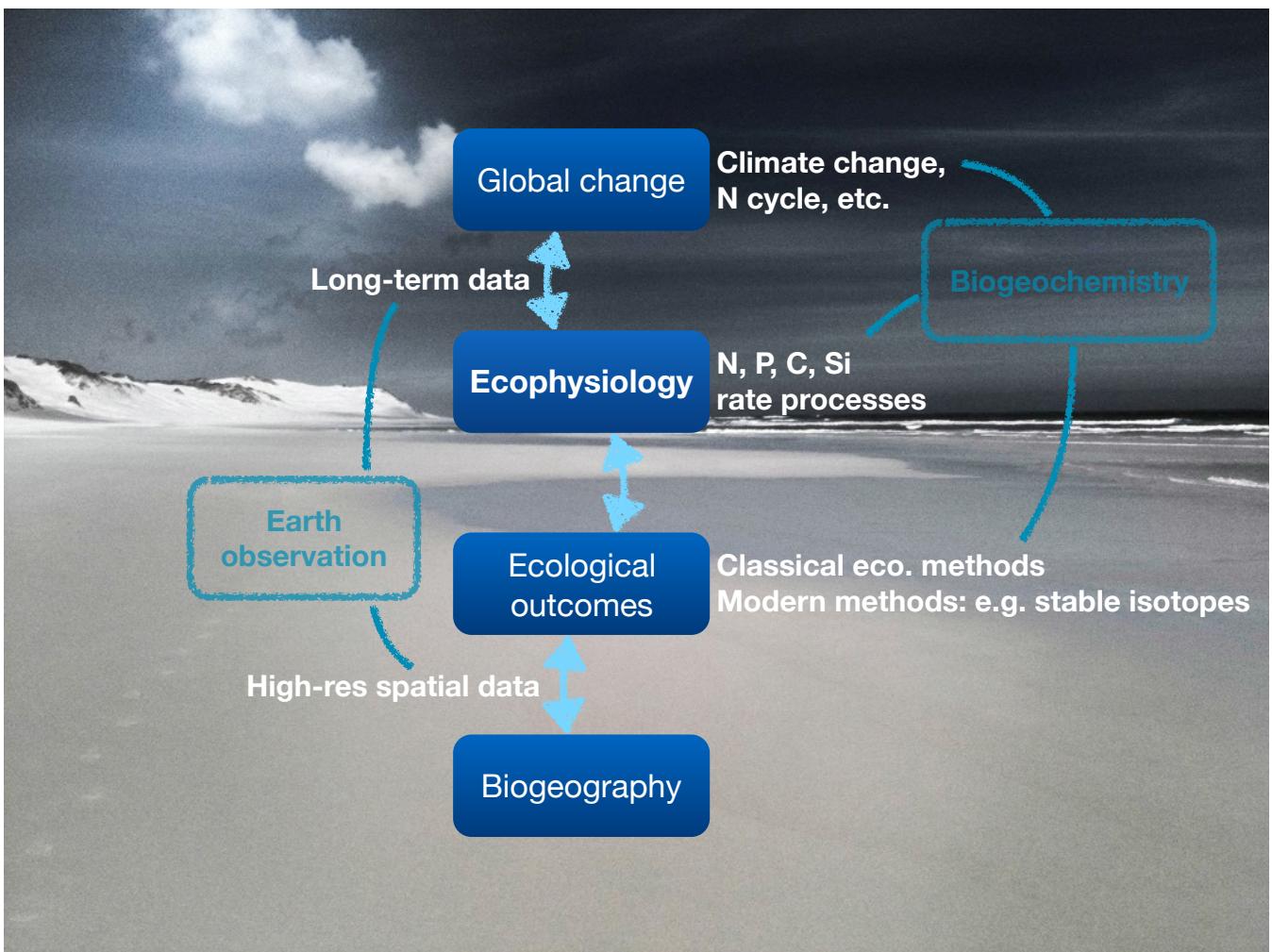
Community ecology vs. population ecology

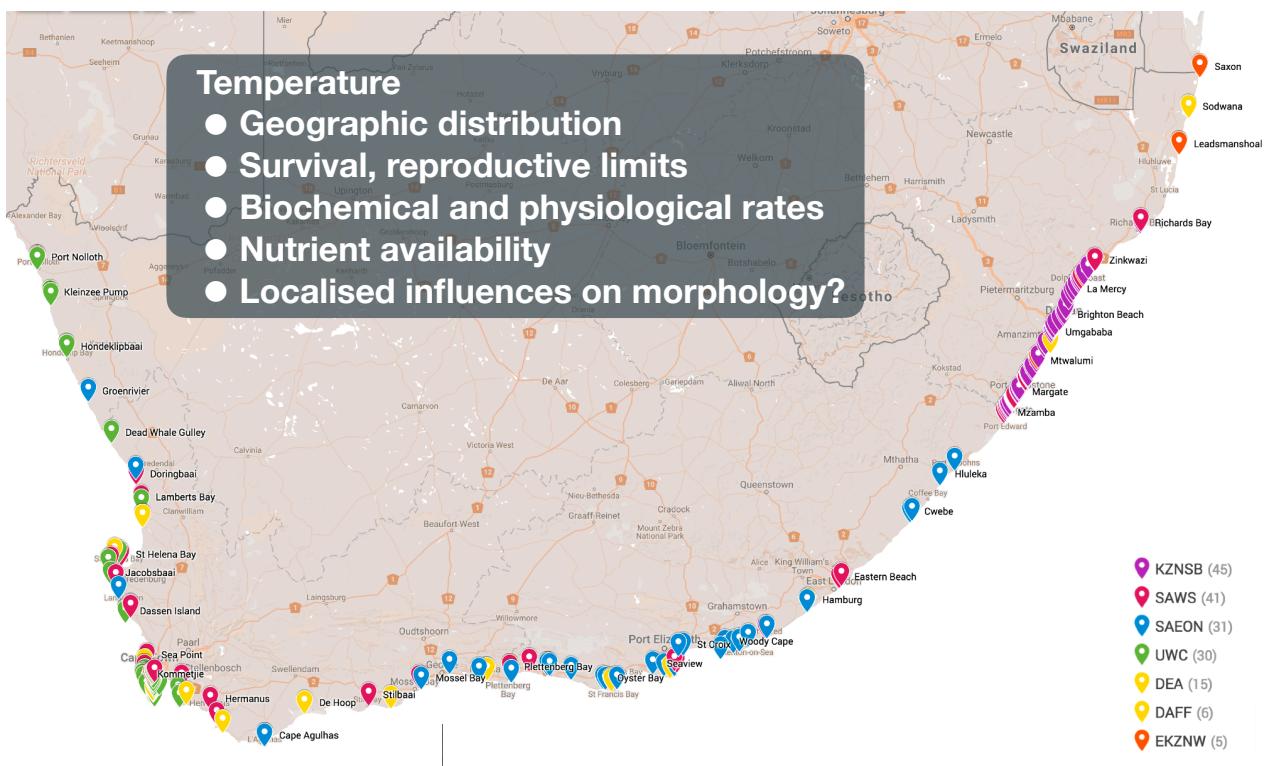
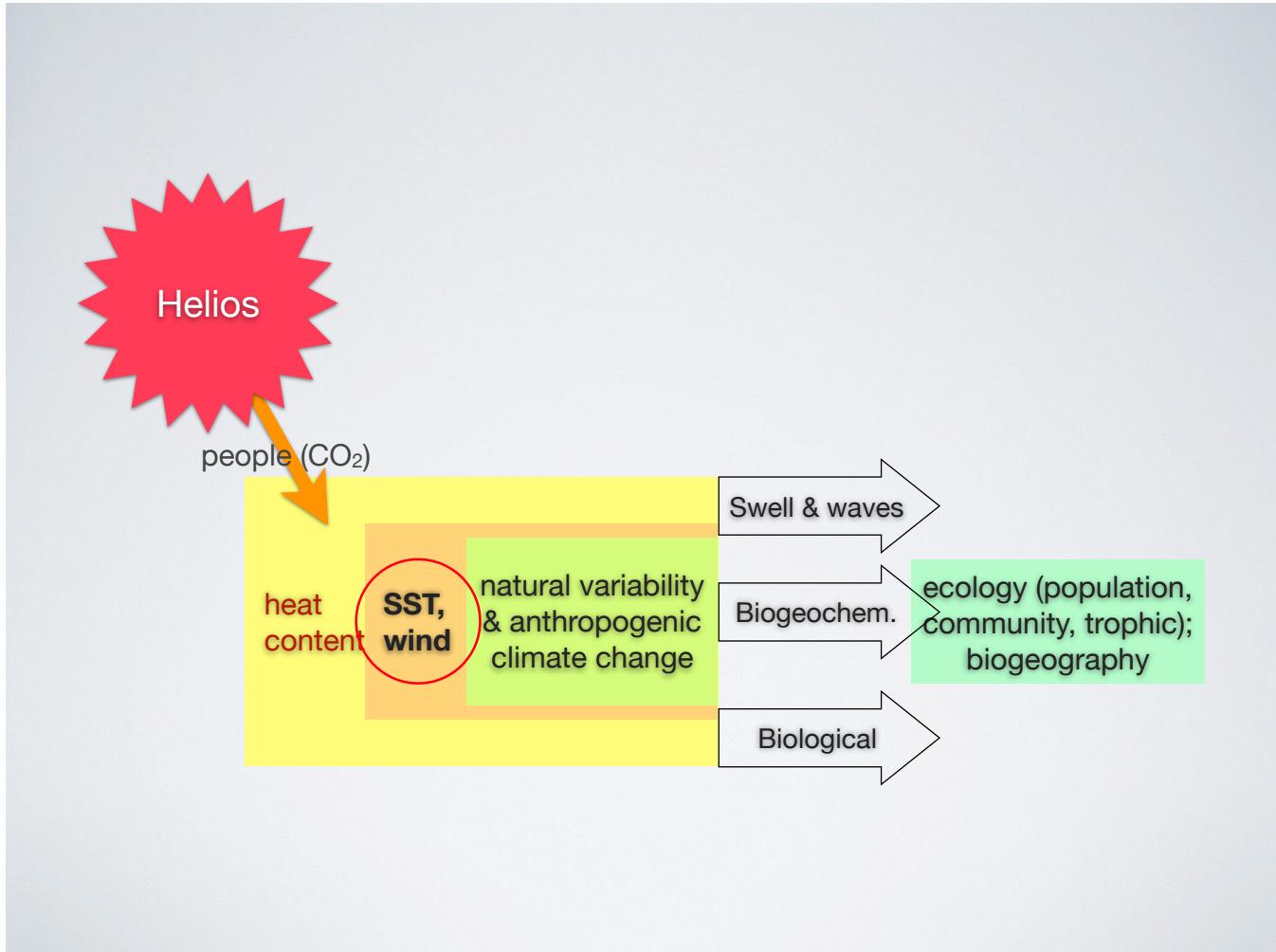


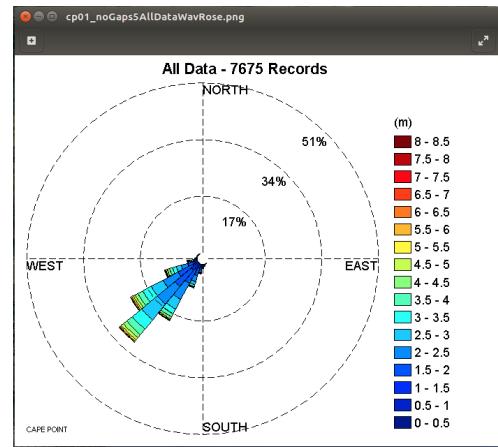
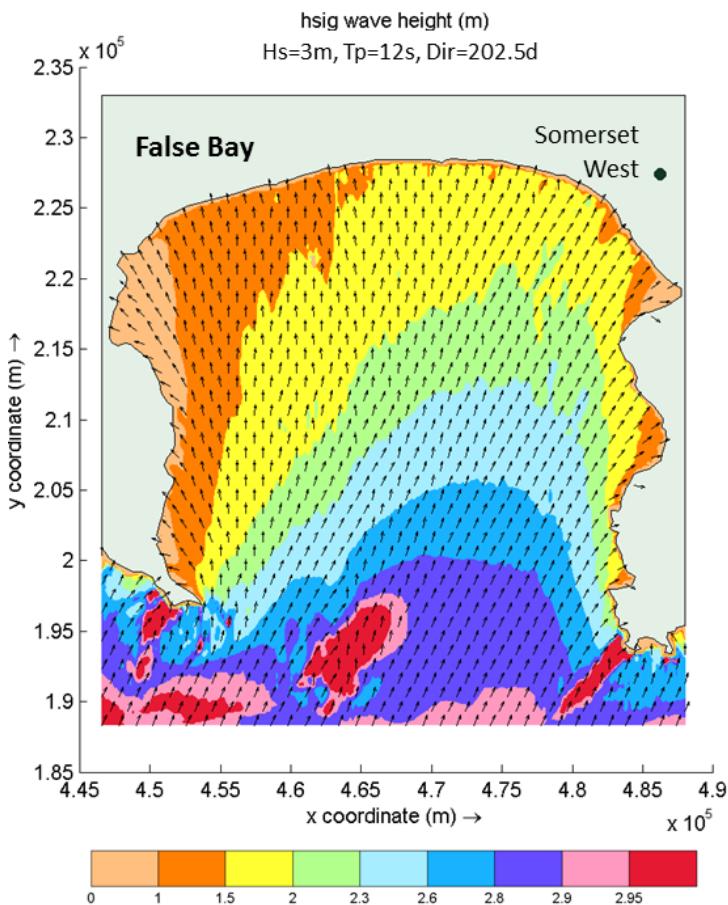
© 2014 AJ Smit

# What kinds of questions?

Community ecology underpins the vast fields of biodiversity and biogeography, and concerns spatial scales from squares of meters to all of Earth. We can look at historical and contemporary processes that have been implicated in shaping the distribution of life on our planet. And the processes that will shape Earth in the Future.







### Numerical wave model

- Wave Watch III forced by NCEP winds at 3hr resolution, hindcast from 1994-2013
- Wave parameters modelled using SWAN
- 200m alongshore resolution

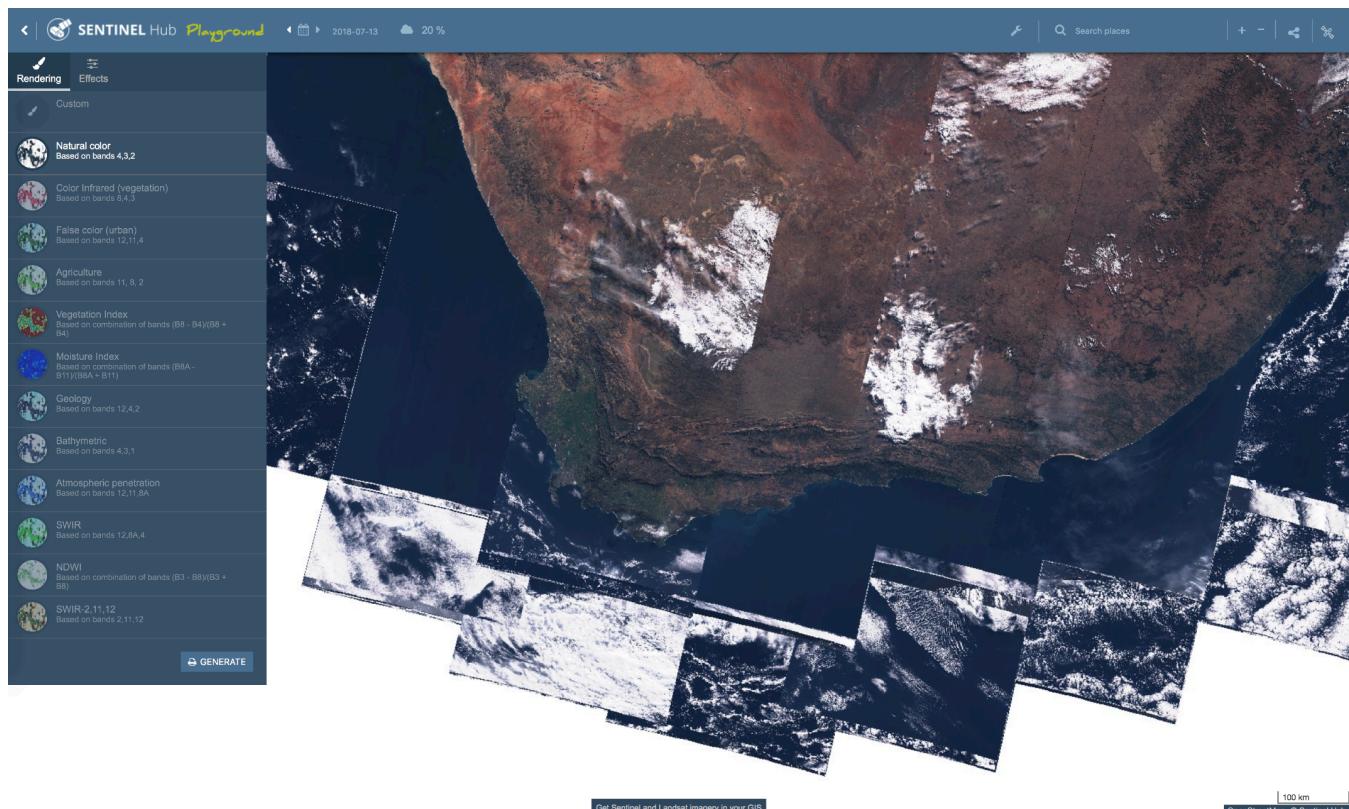
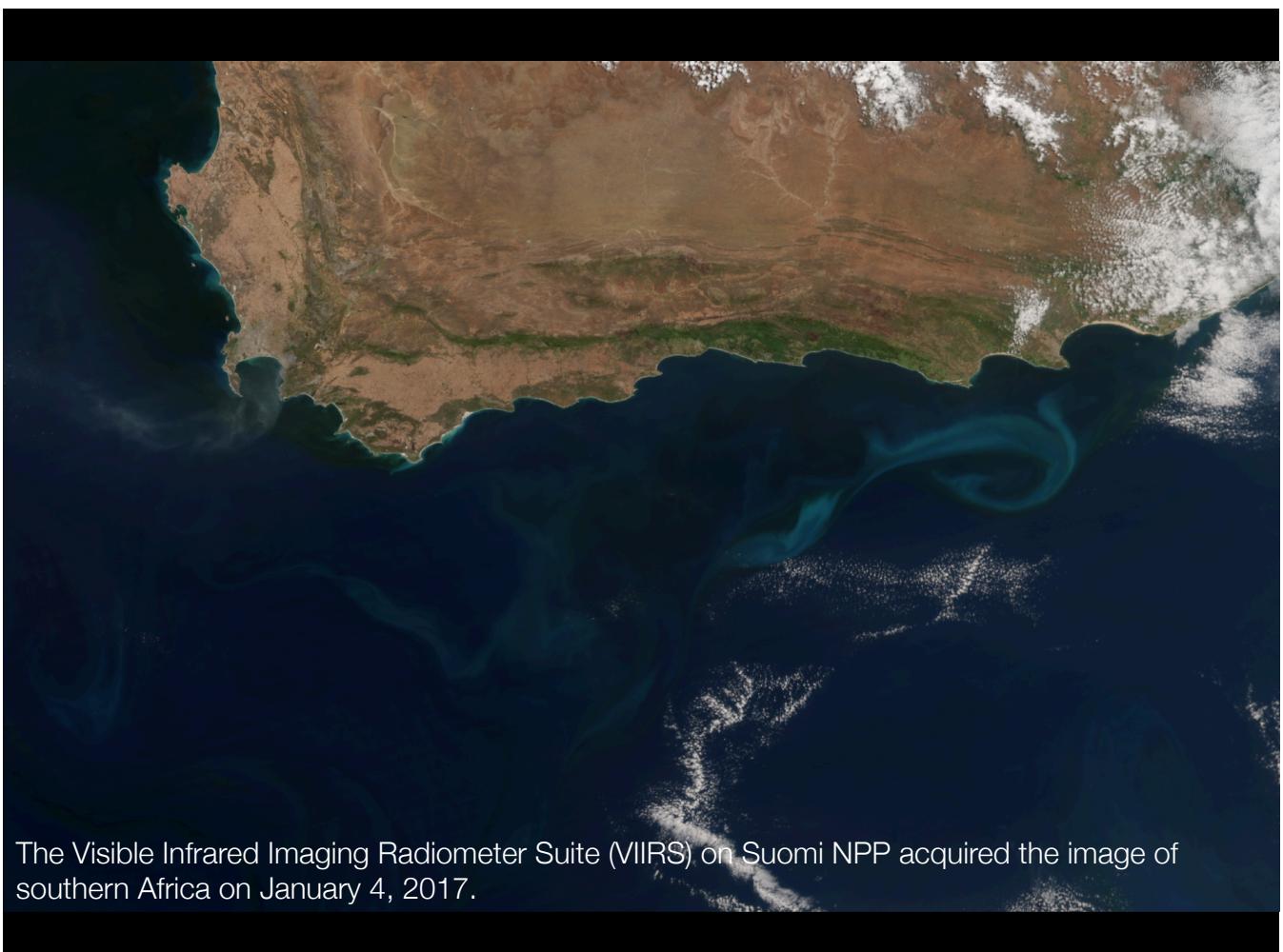




Image of Cape Agulhas was acquired by the Operational Land Imager (OLI) on Landsat 8 on May 25, 2016.



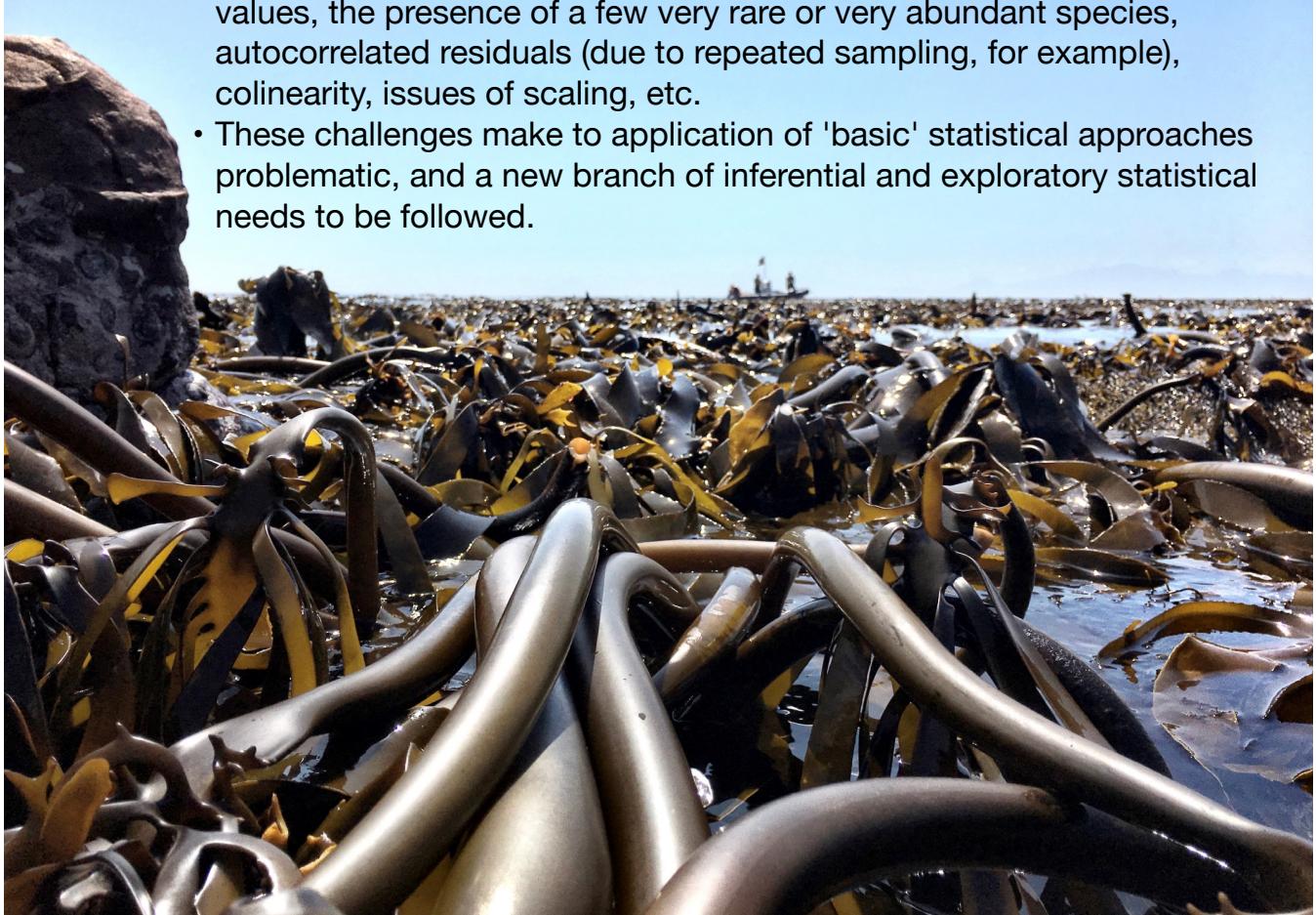
The Visible Infrared Imaging Radiometer Suite (VIIRS) on Suomi NPP acquired the image of southern Africa on January 4, 2017.

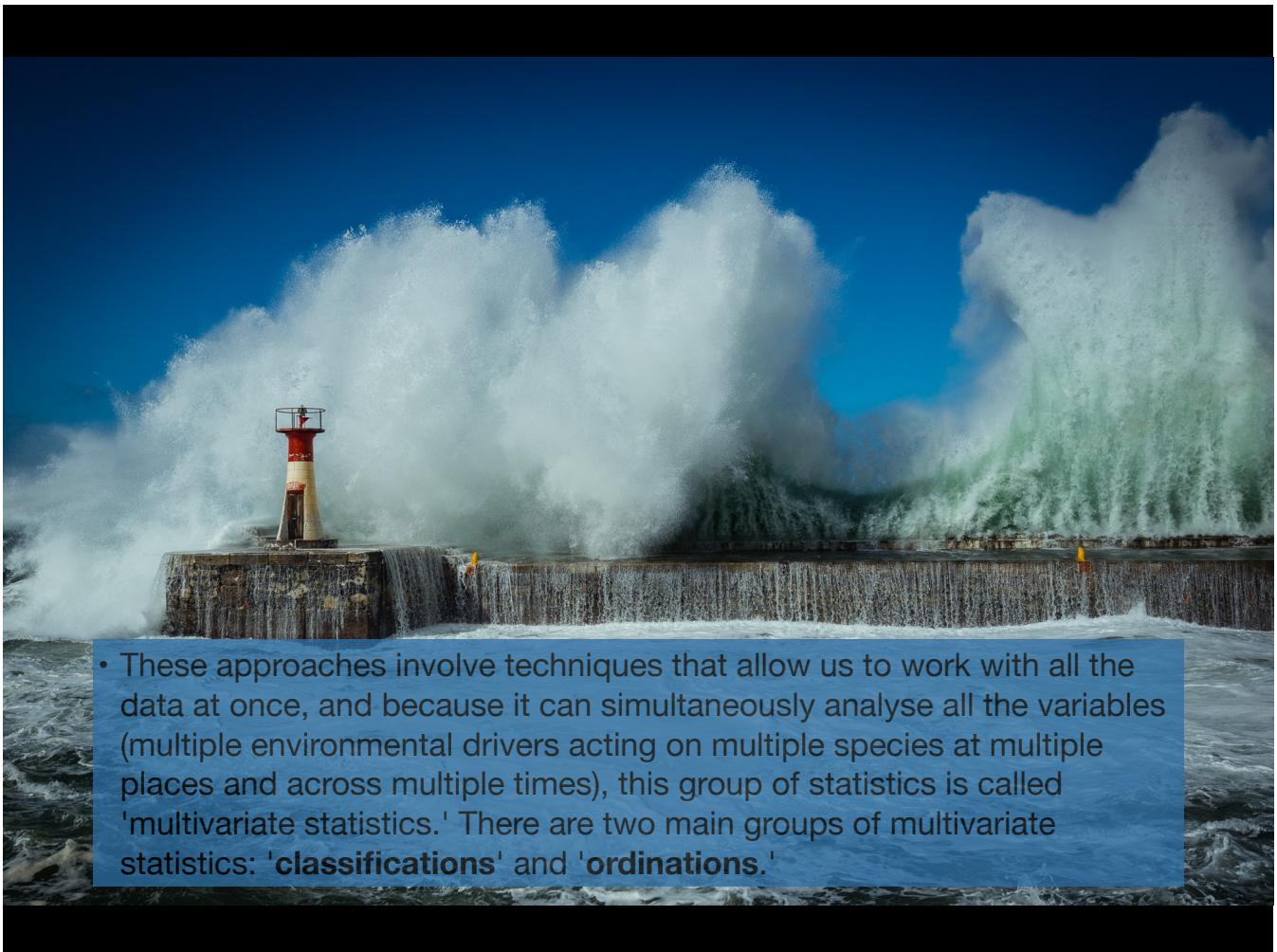
# What is Quantitative Ecology?

- Community ecologists tend to analyse how multiple environmental factors act as drivers that influence the distribution of tens or hundreds of species.
- These data tend to often be messy (not in the sense of untidy data as per the 'tidyverse' definition of tidy data, but it can be that too!) and statistical considerations need to be understood within the context of the data available to us.

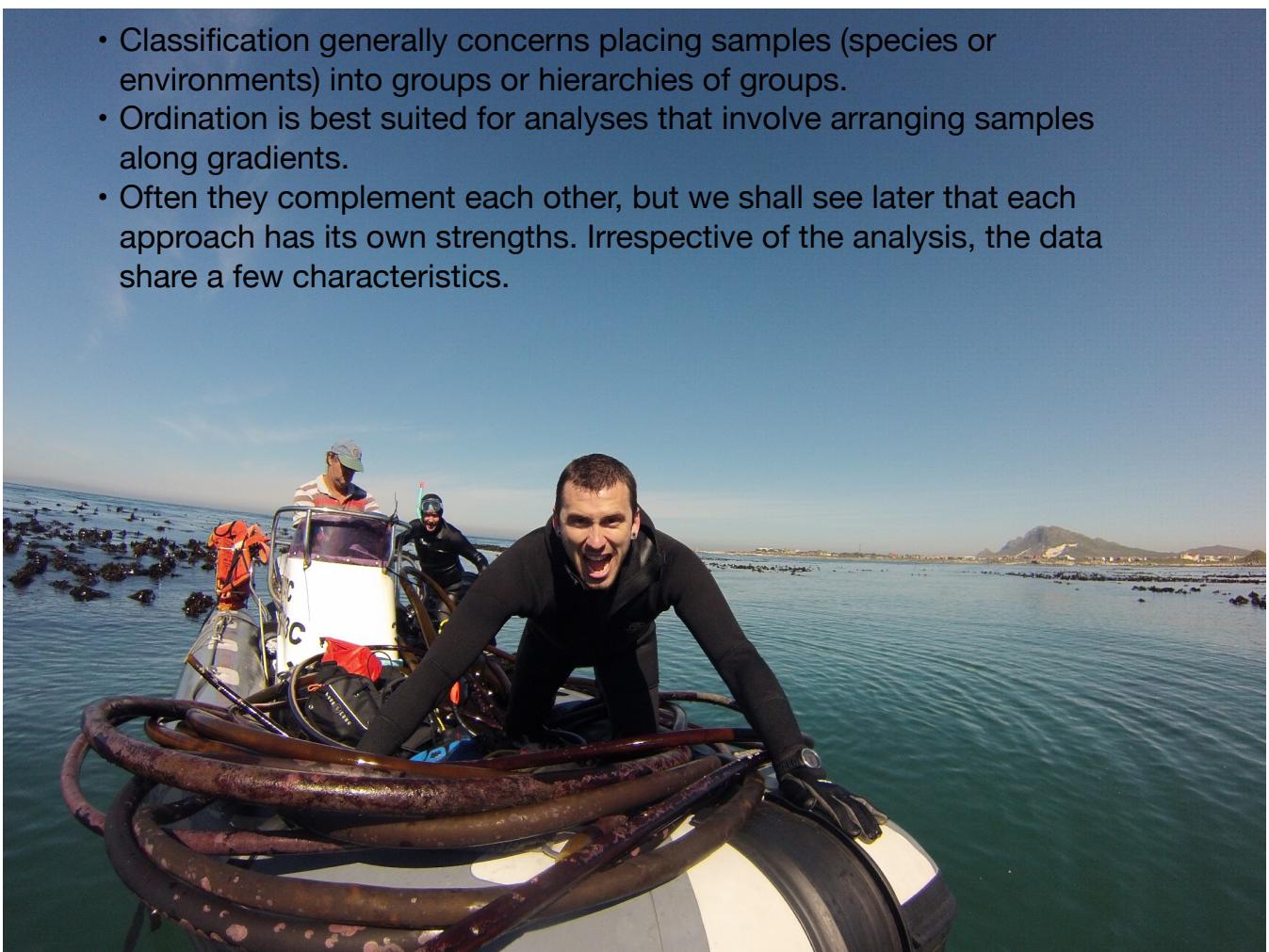


- This translates to errors of measurement and errors due to extreme values, the presence of a few very rare or very abundant species, autocorrelated residuals (due to repeated sampling, for example), collinearity, issues of scaling, etc.
- These challenges make the application of 'basic' statistical approaches problematic, and a new branch of inferential and exploratory statistical needs to be followed.



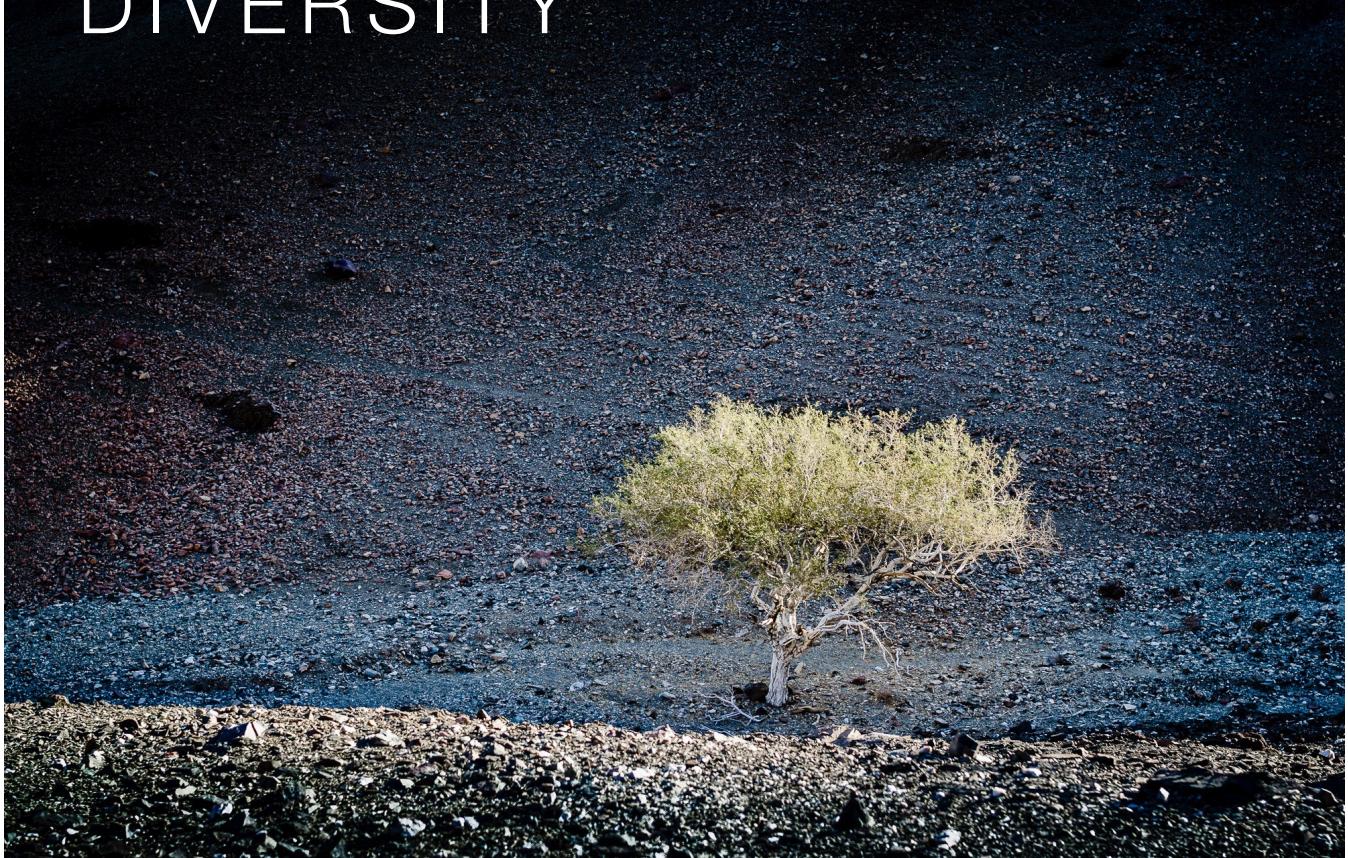


- These approaches involve techniques that allow us to work with all the data at once, and because it can simultaneously analyse all the variables (multiple environmental drivers acting on multiple species at multiple places and across multiple times), this group of statistics is called 'multivariate statistics.' There are two main groups of multivariate statistics: '**classifications**' and '**ordinations**.'



TOPIC 2

# DIVERSITY



Some more important reading...

Ordination Methods for Ecologists

<http://ordination.okstate.edu>

GUide to STatistical Analysis in Microbial Ecology (GUSTA ME)

<https://sites.google.com/site/mb3gustame/>

# Alpha, beta, and gamma diversity

- *alpha ( $\alpha$ ) diversity* is the diversity of a community, and it captures the diversity within a site, plot, transect, or quadrat
- often used to represent the diversity of the smallest sampling unit
- can be seen as representing the local scale
- usually represented as a **synthetic diversity index** (a.k.a. *dissimilarity index*, e.g. Bray-Curtis, Sørensen, Jaccard, etc.) or simply as species richness

# Alpha, beta, and gamma diversity



- *beta ( $\beta$ ) diversity* is also known as ‘**species turnover**’, and represents is the rate of change in species composition from one community to another
- usually applied along **gradients**

# Alpha, beta, and gamma diversity

- *gamma ( $\gamma$ ) diversity* is the **total diversity** of a region or landscape
- can represent the diversity of all samples combined
- same metric as alpha diversity (i.e. one of the dissimilarity indices)

# Alpha, beta, and gamma diversity

- which diversity measure to use depends on the ecological question being asked
- it depends on considerations of the spatial scales of the landscape being studied
- here we shall primarily be concerned with alpha diversity as captured by some of the dissimilarity indices

# Univariate measures of diversity

- Richness ( $R$  or  $S$ , Either local or regional)
- Shannon index ( $H'$ ; Shannon & Weaver 1949): Incorporates richness as well as the relative abundances into a metric
- Simpsons index ( $D$  or  $\lambda$ ; Simpson 1949): Emphasizes evenness

Species	low.light	mid.light	high.light
A	0.75	0.38	0.08
B	0.62	0.15	0.15
C	0.24	0.52	0.18
D	0.33	0.57	0.52
E	0.21	0.28	0.54
F	0.14	0.29	0.56

Metric	low.light	mid.light	high.light
Richness	6	6	6
$H'$	1.63	1.71	1.60
D	0.78	0.81	0.77

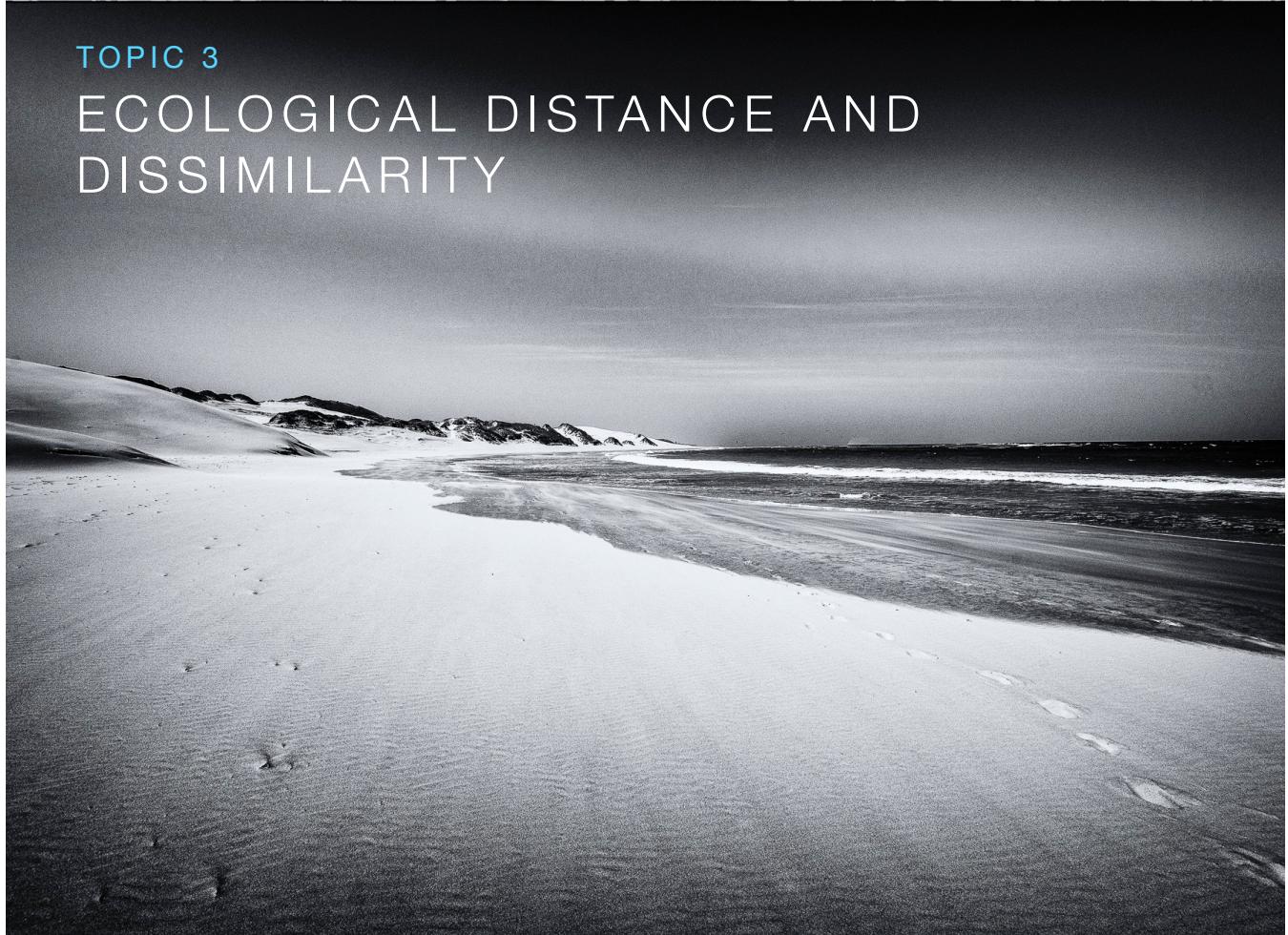
No information about individual species responses

- Species A and B are dominant in low light
- Species E and F are dominant in high light

<https://www.molmed.nl/uploads/abstracts/1542/vegan.pdf>

## TOPIC 3

# ECOLOGICAL DISTANCE AND DISSIMILARITY



# Similarity and dissimilarity

- sites sharing a similar species composition are ecologically similar
  - *i.e.* high similarity / low dissimilarity
- how similar sites are depends on measurable environmental differences that influence species composition, or it can be due to unmeasured influences; it can also simply be noise
- it is the ecologist's role to figure out what influences the similarity/dissimilarity among sites

# Distance matrices

- a **distance matrix** is produced from the data matrix (**species table** or **environment table**) by calculating one of several dissimilarity indices
- also called **association** or **resemblance** matrices
- see `vegdist {vegan}` for a list of dissimilarity indices
- the result is a matrix of **pairwise differences** in community composition (as synthesised by the chosen index) or ecological distance between all sites

# Distance matrices

- the matrices are **square** and **symmetrical**, and they will have as many rows and columns as the number of sites present in the original species or environment table
- the **diagonals are zero** (a site is the same as itself, so it has zero dissimilarity)
- the table can be read directly, and each cell represents the species or ecological difference between a pair of sites
- all information of the species ID (and maybe also abundance) of a site is lost, as this info is condensed into one metric

## Distance matrix for environmental data

- Euclidian distance is “*the ‘ordinary’ straight-line distance between two points in Euclidean space*” (i.e. in its simplest form a planar area, which you know of as a graph with x- and y-axes)
- so, in 2D and 3D, gives distance in cartesian units between points on a plane ( $x, y$ ) or in a volume ( $x, y, z$ )
- conforms to our physical concept of distance
  - e.g. short geographic distances between points on a map
  - (loses accuracy over large distances, as Earth’s surface is not on a plane but on a sphere... correct by using great circle distances, e.g. use the Haversine formula)
- calculated using the Pythagorean theorem
  - differences are squared, so single large differences become very important
  - this is not useful for species data

## Two dimensions [edit]

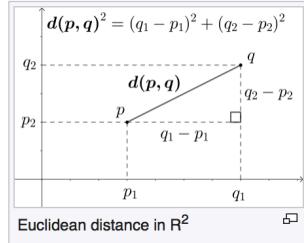
In the Euclidean plane, if  $\mathbf{p} = (p_1, p_2)$  and  $\mathbf{q} = (q_1, q_2)$  then the distance is given by

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}.$$

This is equivalent to the Pythagorean theorem.

Alternatively, it follows from (2) that if the polar coordinates of the point  $\mathbf{p}$  are  $(r_1, \theta_1)$  and those of  $\mathbf{q}$  are  $(r_2, \theta_2)$ , then the distance between the points is

$$\sqrt{r_1^2 + r_2^2 - 2r_1 r_2 \cos(\theta_1 - \theta_2)}.$$



## Three dimensions [edit]

In three-dimensional Euclidean space, the distance is

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}.$$

## n dimensions [edit]

In general, for an  $n$ -dimensional space, the distance is

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

[https://en.wikipedia.org/wiki/Euclidean\\_distance](https://en.wikipedia.org/wiki/Euclidean_distance)

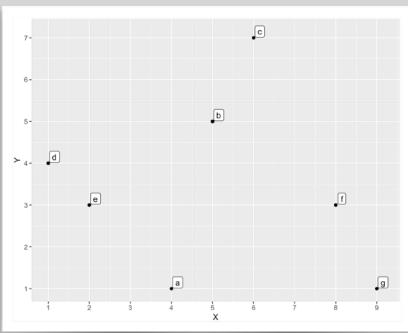
# Euclidian distance

- where  $x_{ij}$   $x_{ik}$  refer to the quantity in (column)  $i$ , at sites (rows)  $j$  and  $k$
- $d_{jk} = \sqrt{\sum(x_{ij} - x_{ik})^2}$

e.g. example with position (such as geographic) coordinates...  
use **vegan**'s `vegdist()` function

### Raw data

	x	y
a	4	1
b	5	5
c	6	7
d	1	4
e	2	3
f	8	3
g	9	1



### Euclidian distances

```
> ex.xy.euc <- vegdist(ex.xy, method = "euclidian")
> ex.xy.euc
```

	a	b	c	d	e	f
b	4.123106					
c	6.324555	2.236068				
d	4.242641	4.123106	5.830952			
e	2.828427	3.605551	5.656854	1.414214		
f	4.472136	3.605551	4.472136	7.071068	6.000000	
g	5.000000	5.656854	6.708204	8.544004	7.280110	2.236068

e.g. example with higher dimension environmental data...

### Raw data

	pH	O2	temp	depth
a	7.1	6.5	12.1	1.1
b	7.5	5.5	12.3	1.3
c	7.6	5.7	11.9	1.5
d	7.0	5.4	11.8	1.6
e	7.1	6.3	12.0	1.8
f	7.2	6.3	12.1	1.9
g	6.9	6.1	12.2	2.2

(transformation)

### Standarised data

```
> ex.env.std <- decostand(xy.env, method = "standardize")
> ex.env.std
```

	pH	O2	temp	depth
a	-0.3872983	1.2156767	0.2494233	-1.41749621
b	1.1618950	-1.0842522	1.4133987	-0.88114629
c	1.5491933	-0.6242664	-0.9145521	-0.34479637
d	-0.7745967	-1.3142450	-1.4965398	-0.07662142
e	-0.3872983	0.7556909	-0.3325644	0.45972850
f	0.0000000	0.7556909	0.2494233	0.72790346
g	-1.1618950	0.2957051	0.8314110	1.53242833

### Euclidian distances

```
> ex.env.euc <- vegdist(xy_dat, method = "euclidian")
> ex.env.euc
```

	a	b	c	d	e	f
b	4.123106					
c	6.324555	2.236068				
d	4.242641	4.123106	5.830952			
e	2.828427	3.605551	5.656854	1.414214		
f	4.472136	3.605551	4.472136	7.071068	6.000000	
g	5.000000	5.656854	6.708204	8.544004	7.280110	2.236068

## Distance matrix for species data

- use Bray-Curtis for the case where data are abundances
- use Jaccard (with `binary = TRUE`) for presence/absence data
- many more in **vegan**; see `?vegdist`

### Bray-Curtis

- again, where  $x_{ij}$   $x_{ik}$  refer to the quantity in (column)  $i$ , at sites (rows)  $j$  and  $k$
- $$d_{jk} = \frac{\sum(|x_{ij} - x_{ik}|)}{\sum(x_{ij} + x_{ik})}$$

# Distance matrix for species data

- using the Doubs species data:
  - ... look at the data
  - ... would we use Bray-Curtis or Jaccard dissimilarities?
  - ... apply the calculation
  - ... explain the meaning of the data in broad terms
  - ... examine it more closely: what general pattern comes out?
  - ... plot this pattern (*hint, it is best seen in the 1st column of the resemblance matrix*)
  - ... what explanation can you offer for this pattern?
  - ... using the decostand() function, create presence/absence data, and apply the appropriate vegdist() function to obtain a suitable dissimilarity matrix

## Distance matrices: some old, some new

- recap
  - **Euclidian distance** for environmental data
  - **Bray-Curtis** for species **abundances** (but also many others in vegdist()) and elsewhere
  - **Jaccard** for **presence-absence** (binary) species data
- ...also **Gower's distance** for **categorical** (factor; must be declared) data, i.e. use daisy() in the **cluster** package or gowdis() in the **FD** package

# Correlations and associations

- e.g. **associations between species** (info on co-varying species) based on dissimilarities, or **correlations between environmental variables**
- associations are usually found on a **transposed** species table

## Associations: species presence-absence

Species table

```
> spe[1:10, 1:10]
# A tibble: 10 x 10
  Cogo Satr Phph Babl Thth Teso Chna Pato Lele Sqce
  <int> <int> <int> <int> <int> <int> <int> <int> <int>
1     0     3     0     0     0     0     0     0     0     0
2     0     5     4     3     0     0     0     0     0     0
3     0     5     5     5     0     0     0     0     0     0
4     0     4     5     5     0     0     0     0     0     1
5     0     2     3     2     0     0     0     0     5     2
6     0     3     4     5     0     0     0     0     1     2
7     0     5     4     5     0     0     0     0     1     1
8     0     0     1     3     0     0     0     0     0     5
9     0     1     4     4     0     0     0     0     2     2
10    1     3     4     1     1     0     0     0     0     1
```

Transposed

```
> spe.t <- t(spe)
> spe.t[1:10, 1:10]
   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
Cogo  0     0     0     0     0     0     0     0     0     1
Satr  3     5     5     4     2     3     5     0     1     3
Phph  0     4     5     5     3     4     4     1     4     4
Babl  0     3     5     5     2     5     5     3     4     1
Thth  0     0     0     0     0     0     0     0     0     1
Teso  0     0     0     0     0     0     0     0     0     0
Chna  0     0     0     0     0     0     0     0     0     0
Pato  0     0     0     0     0     0     0     0     0     0
Lele  0     0     0     0     5     1     1     0     2     0
Sqce  0     0     0     1     2     2     1     5     2     1
```

Jaccard coefficient

```
> spe.t.S7 <- vegdist(spe.t, "jaccard", binary = TRUE)
> round(as.matrix(spe.t.S7)[1:10, 1:10], 2)
   Cogo Satr Phph Babl Thth Teso Chna Pato Lele Sqce
Cogo  0.00 0.53 0.60 0.67 0.22 0.40 0.89 0.81 0.82 0.73
Satr  0.53 0.00 0.24 0.36 0.53 0.61 0.88 0.83 0.65 0.55
Phph  0.60 0.24 0.00 0.17 0.60 0.60 0.77 0.71 0.54 0.39
Babl  0.67 0.36 0.17 0.00 0.67 0.67 0.62 0.60 0.38 0.25
Thth  0.22 0.53 0.60 0.67 0.00 0.40 0.82 0.81 0.82 0.73
Teso  0.40 0.61 0.60 0.67 0.40 0.00 0.75 0.64 0.70 0.73
Chna  0.89 0.88 0.77 0.62 0.82 0.75 0.00 0.23 0.42 0.52
Pato  0.81 0.83 0.71 0.60 0.81 0.64 0.23 0.00 0.39 0.56
Lele  0.82 0.65 0.54 0.38 0.82 0.70 0.42 0.39 0.00 0.28
Sqce  0.73 0.55 0.39 0.25 0.73 0.73 0.52 0.56 0.28 0.00
```

also available are the Sørensen and Ochiai coefficients

# Correlation: environmental variables

## Environment table

```
> env
# A tibble: 30 x 11
  dfs alt slo flo pH har pho nit amm oxy bod
* <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.3 934 48 0.84 7.9 45 0.01 0.2 0 12.2 2.7
2 2.2 932 3 1 8 40 0.02 0.2 0.1 10.3 1.9
3 10.2 914 3.7 1.8 8.3 52 0.05 0.22 0.05 10.5 3.5
4 18.5 854 3.2 2.53 8 72 0.1 0.21 0 11 1.3
5 21.5 849 2.3 2.64 8.1 84 0.38 0.52 0.2 8 6.2
6 32.4 846 3.2 2.86 7.9 60 0.2 0.15 0 10.2 5.3
7 36.8 841 6.6 4 8.1 88 0.07 0.15 0 11.1 2.2
8 49.1 792 2.5 1.3 8.1 94 0.2 0.41 0.12 7 8.1
9 70.5 752 1.2 4.8 8 90 0.3 0.82 0.12 7.2 5.2
10 99 617 9.9 10 7.7 82 0.06 0.75 0.01 10 4.3
# ... with 20 more rows
```



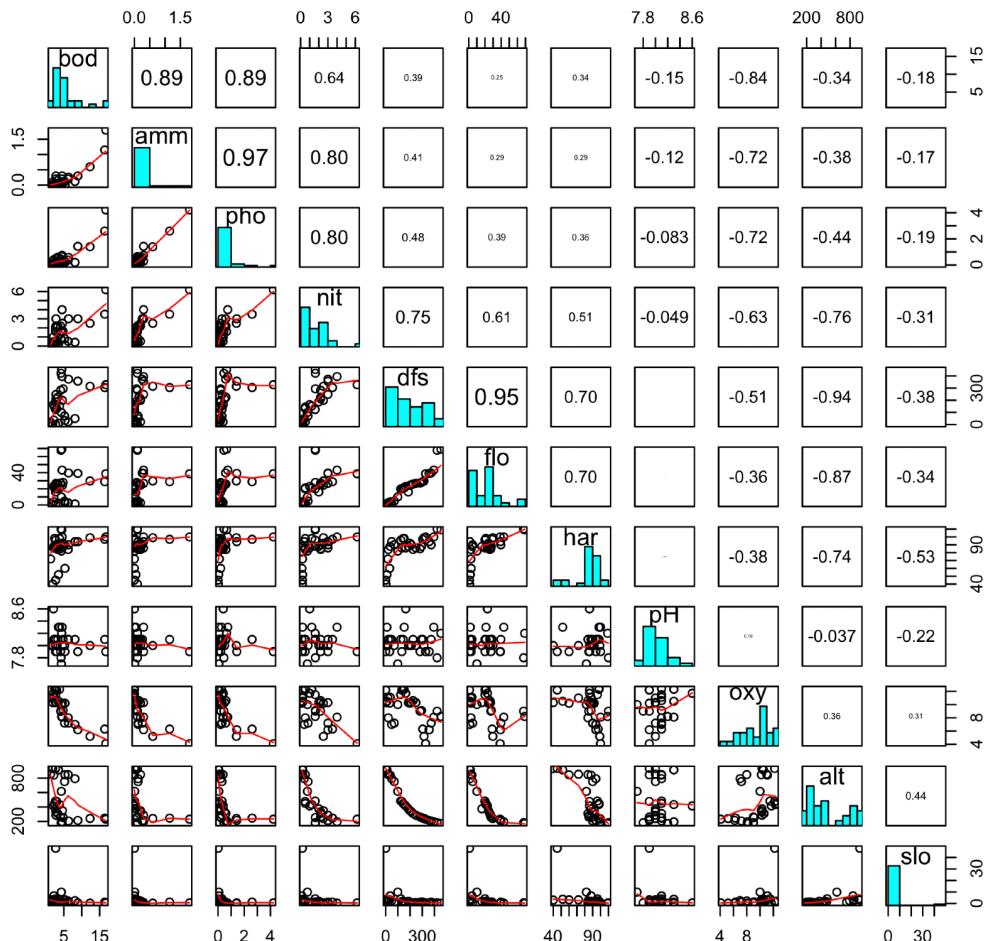
## Pearson correlation coefficients

```
> env.pearson <- cor(env) # default method = "pearson"
> round(env.pearson, 2)
   dfs alt slo flo pH har pho nit amm oxy bod
dfs 1.00 -0.94 -0.39 0.95 0.02 0.73 0.47 0.74 0.41 -0.57 0.43
alt -0.94 1.00 0.46 -0.86 -0.05 -0.79 -0.44 -0.75 -0.38 0.42 -0.38
slo -0.39 0.46 1.00 -0.36 -0.22 -0.53 -0.20 -0.31 -0.17 0.31 -0.17
flo 0.95 -0.86 -0.36 1.00 0.03 0.74 0.38 0.59 0.29 -0.42 0.30
pH 0.02 -0.05 -0.22 0.03 1.00 0.08 -0.08 -0.04 -0.12 0.19 -0.16
har 0.73 -0.79 -0.53 0.74 0.08 1.00 0.37 0.53 0.30 -0.37 0.34
pho 0.47 -0.44 -0.20 0.38 -0.08 0.37 1.00 0.80 0.97 -0.76 0.91
nit 0.74 -0.75 -0.31 0.59 -0.04 0.53 0.80 1.00 0.80 -0.69 0.68
amm 0.41 -0.38 -0.17 0.29 -0.12 0.30 0.97 0.80 1.00 -0.75 0.90
oxy -0.57 0.42 0.31 -0.42 0.19 -0.37 -0.76 -0.69 -0.75 1.00 -0.84
bod 0.43 -0.38 -0.17 0.30 -0.16 0.34 0.91 0.68 0.90 -0.84 1.00
```

no need to transpose or standardise as the `cor()` function does this internally

"comparison among ordinal variables, or among quantitative variables that may be monotonically but not linearly related, can be achieved using rank correlation coefficients like Spearman's  $\rho$  (rho) or Kendall's  $\tau$  (tau)"

## Pearson Correlation Matrix



# Associations: species tables

- we do not derive Euclidian distances from species data, nor do we determine pairwise covariances or correlations (this is fine with the environmental data)
- it may be necessary to apply various transformations to the species data, e.g. when there are a few rare species
- transformation options are provided by `decostand()`; see section 3.5 in *Numerical Ecology with R*

**Table 3.1** Commonly-used distance and dissimilarity functions in Q mode available in R packages. The symbol  $\Rightarrow$  means that applying the function designed for quantitative data to presence-absence data produces the same result as computing the corresponding function designed for presence-absence data

Quantitative data	Presence-absence data
<i>Community composition data</i>	
Ruzicka dissimilarity <code>vegdist(., "jac")</code>	$\Rightarrow$ Jaccard dissimilarity <code>vegdist(., "jac", binary=TRUE)</code> <code>dist.binary(., method=1)</code>
Hellinger distance <code>decostand(., "hel")</code> followed by <code>vegdist(., "euc")</code>	$\Rightarrow$ Ochiai dissimilarity <code>dist.binary(., method=7)</code>
Chord distance <code>decostand(., "norm")</code> followed by <code>vegdist(., "euc")</code>	$\Rightarrow$ Ochiai dissimilarity <code>dist.binary(., method=7)</code>
Bray-Curtis dissimilarity <code>vegdist(., "bray")</code>	$\Rightarrow$ Sørensen dissimilarity <code>dist.binary(., method=5)</code>
Chi-square distance <code>decostand(., "chi.square")</code>	Chi-square distance (idem)
Canberra distance <code>vegdist(., "canberra")</code>	
<i>Other variables, mixed physical units</i>	
Standardized variables: Euclidean distance <code>vegdist(., "euc")</code>	Standardized variables: Simple matching coefficient <code>dist.binary(., method=2)</code>
Non-standardized variables: Gower distance <code>daisy(., "gower")</code>	

# Correlations and associations

- the square association and correlation matrices are generally only used as intermediate steps in our workflow, and are not usually scrutinised directly
- however, meaningful information is already present in these matrices, and it is beneficial to be able to read them
- it is definitely necessary to understand how they are calculated
- the next step in the workflow takes the association and/or correlation matrices and applies the multivariate analyses on them

TOPIC 4  
CLUSTERING



## TOPIC 5

# UNCONSTRAINED ORDINATION



## Questions

**Question 1:**

- One of the graphs are in a circle and the other does not. Why is this? What does the circle represent?
- What does the circle of equilibrium mean?

**Question 2:**

- Will the PCA plots always only have the two axis?
- To calculate the proportion of variance, do you only use the values of the which ever axis are shown (eg. PC1 or PC2)?

**Question 3:**

- How do we compare variables between different sites exactly?

**Question 4:**

- What does the length of the arrows mean, why are some of them short and others longer?

**Question 5:**

- Why do some of the principal components produce negative values in their eigenvalues?

**Question 6:**

- Why should you standardise variables first before running PCA?

**Question 7:**

- Should you leave some variables unstandardised if you want certain variables to have more weighting than others?

**Question 8:**

- Can PCAs be used for things like GIS or molecular work involving large Datasets? In order to summarise climate data for ecological niche modeling.?

**Question 9:**

- I do not understand why and how PCAs are relevant for certain datasets and not for others.

**Question 10:**

- Explain what the distance between objects mean?

**Question 11:**

- How do you know if the variables are positively or negatively correlated with one another by looking at the graph?

# Ordinations

- clustering find discontinuities, but **ordinations highlight gradients**
- gradients are particularly present in ecological communities
- suited to multivariate data, which represent...
  - ...a **space** (e.g. the landscape) comprised of many **sites** (i.e. the samples forming the rows), each one occupied by many **variables** (i.e. species or environmental variables in the columns)
  - there are as many dimensions as variables
  - it would be a silly and not very revealing to analyse it all using a series of univariate or bivariate analyses;
  - $(27 \times 26)/2 = 351$  in the Doubs data set to be precise

# Ordinations

- ordinations represent the data along a reduced number of **orthogonal axes** (linearly independent, uncorrelated), constructed in such a way that they represent, in decreasing order, the main trends of the data
- interpretation is aided by visualisations, regressions, and clustering
- **unconstrained** ordinations are not statistical (no inference testing); they are purely descriptive
  - *also called indirect gradient analysis*
  - based only on the species  $\times$  sites matrix or the environment  $\times$  sites matrix, each seen in isolation
- **constrained** ordination adds a level of statistical testing (next topic)
  - *also called direct gradient analysis*
  - typically uses explanatory variables (in the env. matrix) to explain the patterns seen in the spp. matrix

# Ordinations

- basically, ordinations geometrically arrange sites or species so that distances between them in the 2D or 3D graph represent their ecological distances
- the further the distances between sites or species are on the graph, the larger the ecological differences between them
- they are therefore dimension reduction methods
  - take high-dimensional data
  - reduces the complexity to low-dimensional space

According to Gauch (1982): "**Ordination primarily endeavors to represent sample and species relationships as faithfully as possible in a low-dimensional space**". But why is this objective desirable? There are a number of answers, but most are derived from the 'properties of community' data as described above:

- It is **impossible to visualize multiple dimensions simultaneously** [...] ecologists typically grapple with dozens of dimensions (species and/or samples).
- A single **multivariate analysis saves time**, in contrast to a separate univariate analysis for each species.
- Ideally and typically, **dimensions of this 'low dimensional space' will represent important and interpretable environmental gradients**.
- If statistical tests are desired, **problems of multiple comparisons are diminished** when species composition is studied in its entirety.
- **Statistical power is enhanced** when species are considered in aggregate, because of redundancy.
- **By focusing on 'important dimensions', we avoid interpreting (and misinterpreting) noise**. Thus, ordination is a 'noise reduction technique' (Gauch 1982).
- We can **determine the relative importance of different gradients**; this is virtually impossible with univariate techniques.
- **Community patterns may differ from population patterns**.
- Some techniques provide a measure of beta diversity.
- The **graphical results from most techniques often lead to ready and intuitive interpretations** of species-environment relationships.

# Ordinations

- **Principal component analysis (PCA):** the main eigenvector-based method. Works on raw, quantitative data. Preserves the Euclidean distance among sites.
- **Correspondence analysis (CA):** works on data that must be frequencies or frequency-like, dimensionally homogeneous, and non-negative. Preserves the  $\chi^2$  distance among rows or columns. Mainly used in ecology to analyse species data tables.
- **Principal coordinate analysis (PCoA):** devoted to the ordination of distance matrices, most often in the Q mode, instead of site-by-variables tables. Hence, great flexibility in the choice of association measures.
- **Nonmetric multidimensional scaling (NMDS):** unlike the three others, this is not an eigenvector-based method. NMDS tries to represent the set of objects along a predetermined number of axes while preserving the ordering relationships among them.
- PCoA and NMDS can produce ordinations from any square distance matrix.

Taken verbatim from 5.2.2 in Numerical Ecology with R

## Interpretation of ordination plots

- The **direction of the axes** (e.g. left vs. right; up vs. down) is arbitrary and should not affect the interpretation.
- The **numeric scale on the axis is not very useful for the interpretation** (an exception for this is DCA, in which the scales are in units of beta diversity).
- In most techniques (but not NMDS), **the order of the axes is important**. Thus, axis 1 is more important than axis 2, etc. The meaning of ‘importance’ depends on the technique employed, but ideally **related to the relative influence of environmental gradients**.
- **Third and higher axes can be constructed.** The choice of ‘when to stop’ interpreting new axes is largely a matter of taste, the quantity and quality of the data, and the ability to interpret the results. Fortunately, most of the techniques presented later provide supplemental statistics that can assist in the task.
- It is desirable that axes not be correlated, because you would like them to represent different gradients. **Most techniques automatically result in uncorrelated** (or orthogonal) axes.
- **A biologist's insight, experience, and knowledge of the literature are the most important tools for interpreting indirect gradient analysis.**

Taken verbatim from <http://ordination.okstate.edu/overview.htm>

# Unconstrained ordinations

- recap:
  - unconstrained ordinations are not statistical (no inference testing); they are purely descriptive
    - i.e. no inferences of causation
  - also called indirect gradient analysis
  - based only on the **species × sites** matrix or the **environment × sites** matrix, each seen in isolation
- ...apply to one multivariate (>>2) matrix (spp. or env.)
  - ✓ Principal Component Analysis (**PCA**); **eigenvector** | **Euclidian distance**
  - ✓ Correspondence Analysis (**CA**); **eigenvector** |  $\chi^2$ -distance
  - Detrended Correspondence Analysis (**DCA**)
  - ✓ Principal Coordinates Analysis (**PCoA**); **eigenvector** | **dissimilarities**
  - ✓ Non-Metric Multidimensional Scaling (**NMDS**); **non-metric** | **dissimilarities**
- (Cluster analysis also applies to one matrix (spp. or env.) with >>2 variables)

## Multivariate Techniques

Obs	Group	X-set	Y-set
1	A	$a_{11} \ a_{12} \ a_{13} \ \dots \ a_{1p}$	$b_{11} \ b_{12} \ b_{13} \ \dots \ b_{1m}$
2	A	$a_{21} \ a_{22} \ a_{23} \ \dots \ a_{2p}$	$b_{21} \ b_{22} \ b_{23} \ \dots \ b_{2m}$
3	A	$a_{31} \ a_{32} \ a_{33} \ \dots \ a_{3p}$	$b_{31} \ b_{32} \ b_{33} \ \dots \ b_{3m}$
.	.	.	.
.	.	.	.
n	A	$a_{n1} \ a_{n2} \ a_{n3} \ \dots \ a_{np}$	$b_{n1} \ b_{n2} \ b_{n3} \ \dots \ b_{nm}$
$n+1$	C	$c_{11} \ c_{12} \ c_{13} \ \dots \ c_{1p}$	<b>Unconstrained ordination</b> Principal Component Analysis ( <b>PCA</b> ) Correspondence Analysis ( <b>CA</b> ) Detrended Correspondence Analysis ( <b>DCA</b> ) Principal Coordinates Analysis ( <b>PCoA</b> ) Non-Metric Multidimensional Scaling ( <b>NMDS</b> )
$n+2$	C	$c_{21} \ c_{22} \ c_{23} \ \dots \ c_{2p}$	
$n+3$	C	$c_{31} \ c_{32} \ c_{33} \ \dots \ c_{3p}$	
.	.	.	
.	.	.	
N	C	$c_{n1} \ c_{n2} \ c_{n3} \ \dots \ c_{np}$	

also...

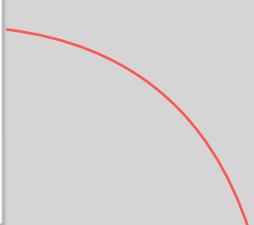
Cluster Analysis

# Principal Component Analysis (PCA)

- in ecology, PCA is done by the eigen-decomposition of an association matrix
- this yields a rigid rotation of axes in that the positions of points relative to one another (euclidean distances) are maintained during rotation
- the rotated axes are referred to as principle components
- the higher the degree of associations between the variables (species) is, the more strongly ‘focused’ the data cloud is, and therefore the more ‘successful’ the PCA is in terms of being able to represent many variables by only a few new variables (*i.e.* a few PC axes)

# Principal Component Analysis (PCA)

```
> env
# A tibble: 29 x 11
  dfs   alt   slo   flo   pH   har   pho   nit   amm   oxy   bod
  <dbl> <int> <dbl> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <dbl> <dbl>
1  0.3   934   48    0.84   7.9   45    0.01   0.2    0     12.2   2.7
2  2.2   932   3     1      8     40    0.02   0.2    0.1   18.3   1.9
3 10.2   914   3.7   1.8    8.3   52    0.05   0.22   0.05   10.5   3.5
4 18.5   854   3.2   2.53   8     72    0.1    0.21   0     11     1.3
5 21.5   849   2.3   2.64   8.1   84    0.38   0.52   0.2    8     6.2
6 32.4   846   3.2   2.86   7.9   60    0.2    0.15   0     10.2   5.3
7 36.8   841   6.6   4     8.1   88    0.07   0.15   0     11.1   2.2
8 70.5   752   1.2   4.8    8     90    0.3    0.82   0.12   7.2    5.2
9 99    617   9.9   10     7.7   82    0.06   0.75   0.01   10     4.3
10 123.   483   4.1   19.9   8.1   96    0.3    1.6    0     11.5   2.7
# ... with 19 more rows
```



```
# PCA based on a correlation matrix
# Argument scale = TRUE calls for a standardization of the variables and it
# creates a correlation matrix; this is necessary because the variables each
# have a different measurement scale
env.pca <- rda(env, scale = TRUE)
```

...intermediate correlation matrix produced by `scale = TRUE`...

```
> cor(env)
  dfs   alt   slo   flo   pH   har   pho   nit   amm   oxy   bod
  dfs 1.0000000 -0.93837904 -0.3947527 0.94742121 0.01604625 0.73277932 0.4729141 0.73809569 0.4976923 -0.5790685 0.43446159
  alt -0.93837894 1.00000000 0.4571298 -0.86289088 -0.85835674 -0.78551547 -0.4371891 -0.75268515 -0.3811344 0.4248239 -0.3825222
  slo -0.39475266 0.45712978 1.0000000 -0.35751430 -0.22222941 -0.52669175 -0.1953638 -0.31433998 -0.1746442 0.3976363 -0.1738556
  flo 0.94742121 -0.86289088 -0.3576143 1.00000000 0.03312637 0.73662526 0.3786878 0.59315883 0.2925252 -0.4219945 0.2951891
  pH  0.01604625 -0.05835674 -0.2222294 0.83312637 1.00000000 0.08451511 -0.0794825 -0.04946292 -0.1228818 0.1923988 -0.1617856
  har 0.73277932 -0.78551547 -0.5266918 0.73662526 0.08451511 1.00000000 0.3731861 0.53495392 0.2961368 -0.3736374 0.3367747
  pho 0.47291414 -0.43718911 -0.1953638 0.37868776 -0.07948250 0.37318607 1.0000000 0.80893149 0.9699346 -0.7575015 0.9991698
  nit 0.73809569 -0.75268515 -0.3143399 0.59315083 -0.04946292 0.53495392 0.8089315 1.00000000 0.8822323 -0.6867146 0.6832380
  amm 0.49769227 -0.38113442 -0.1746442 0.29252515 -0.12200180 0.29163608 0.9699345 0.80822323 1.0000000 -0.7462700 0.9028247
  oxy -0.57906855 0.42482297 0.3676363 -0.4219947 0.19239884 -0.37363740 -0.7576015 -0.68671462 -0.7462700 1.0000000 -0.8398165
  bod 0.43461586 -0.38252216 -0.17388556 0.29518914 -0.16170564 0.33697473 0.9891698 0.68322998 0.9828247 -0.8398165 1.0000000
```

# Principal Component Analysis (PCA)

```
> env.pca
Call: rda(X = env, scale = TRUE)

          Inertia Rank
Total           11
Unconstrained   11   11
Inertia is correlations

Eigenvalues for unconstrained axes:
  PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9   PC10  PC11
  6.098 2.167 1.038 0.704 0.352 0.319 0.165 0.112 0.023 0.017 0.006
```

continue...

```
> # Since the association matrix is a correlation matrix, the sum of the
> # eigenvalues along the diagonal equals the number of 'species'
> # (here env. vars.)
> sum(env.pca$CA$eig)
[1] 11
```

# Principal Component Analysis (PCA)

```
> summary(env.pca) # Default scaling 2

Call:
rda(X = env, scale = TRUE)

Partitioning of correlations:
          Inertia Proportion
Total           11       1
Unconstrained   11       1

Eigenvalues, and their contribution to the correlations

Importance of components:
          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9     PC10    PC11
Eigenvalue  6.0979 2.1672 1.03761 0.70353 0.35174 0.31912 0.16453 0.11173 0.023134 0.01737 0.0060521
Proportion Explained 0.5544 0.1970 0.09433 0.06396 0.03198 0.02901 0.01496 0.01016 0.002103 0.00158 0.0005502
Cumulative Proportion 0.5544 0.7514 0.84571 0.90967 0.94164 0.97065 0.98561 0.99577 0.997870 0.99945 1.0000000

Scaling 2 for species and site scores
* Species are scaled proportional to eigenvalues
* Sites are unscaled: weighted dispersion equal on all dimensions
* General scaling constant of scores: 4.189264
```

- continue...
- eigenvalues are the amount of variation (inertia) explained by the new variables (PC axes), i.e.  $(6.0979/11)=0.5544 \dots$  if using a correlation matrix!
  - percentage and cumulative percentage
  - the first principal component explains 55.44% of the variation
  - the second adds an additional 19.70%
  - the total amount of explanation offered by PC1 + PC2 is 75.14%

# Principal Component Analysis (PCA)

- **species scores:** the loadings (a.k.a. scores or scaled eigenvectors) indicate the degree of correlation between the original variables and the new variables known as the principle components (PC1, PC2, etc.)
- they indicate how much each of the original variables contribute to PC1, PC2, etc.
- here, PC1 is made up of uneven contributions from most of the original variables; pH and slope are less important
- given the strength of this principle component (it explains 55.44% of the inertia), one might hypothesise that its constituent variables influence many aspects of the community
- the sign of an eigenvector indicates the polarity of the correlation between the original variable and the new variable

...continue

	PC1	PC2	PC3	PC4	PC5	PC6
dfs	1.0842	0.5150	-0.25749	-0.16168	0.21132	-0.09485
alt	-1.0437	-0.5945	0.17984	0.12282	0.12464	0.14022
slo	-0.5752	-0.5103	-0.55499	-0.80204	0.02764	0.20077
flo	0.9577	0.6412	-0.38654	-0.19427	0.18401	0.03068
pH	-0.0584	0.4820	1.03444	-0.51378	0.14421	0.05821
har	0.9072	0.6183	-0.02280	0.15767	-0.27865	0.50738
pho	1.0460	-0.6093	0.18734	-0.11866	-0.15113	0.04888
nit	1.1432	-0.1290	0.01203	-0.18471	-0.21270	-0.34907
amm	0.9954	-0.6989	0.18597	-0.08277	-0.19234	-0.04979
oxy	-1.0089	0.4578	-0.00918	-0.23450	-0.50552	-0.05764
bod	0.9899	-0.6836	0.11962	0.03646	0.08542	0.21993

continue...

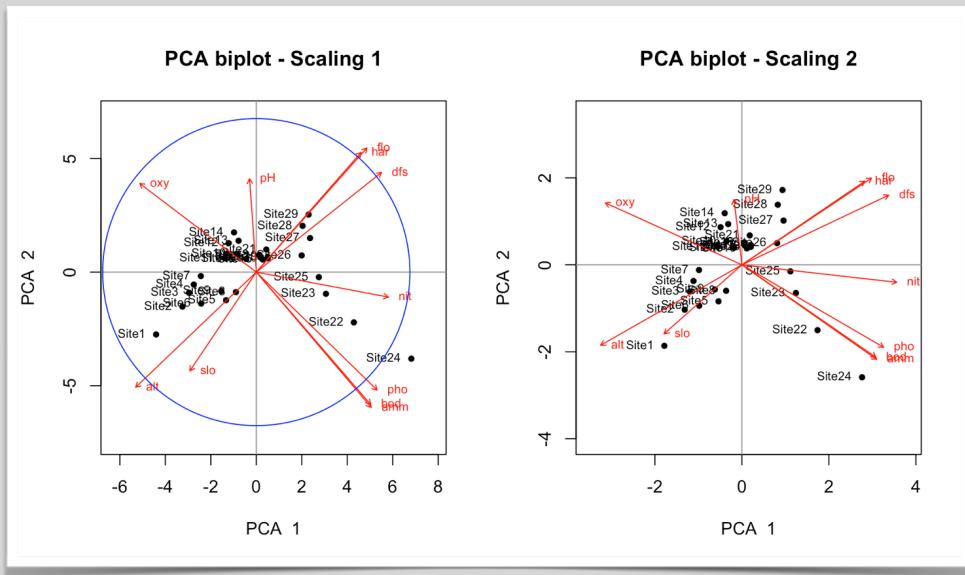
# Principal Component Analysis (PCA)

...continue

	Site scores (weighted sums of species scores)					
	PC1	PC2	PC3	PC4	PC5	PC6
sit1	-1.41243	-1.47560	-1.74593	-2.95533	0.23051	0.49227
sit2	-1.04173	-0.81761	0.34075	0.54364	0.92835	-1.76967
sit3	-0.94881	-0.48823	1.36059	-0.21768	1.05289	-0.69640
sit4	-0.88070	-0.29459	0.21814	0.66428	-0.23934	-0.06427
sit5	-0.42588	-0.66583	0.77631	0.78777	0.62942	1.17850
sit6	-0.77730	-0.74514	-0.06764	0.90839	0.46945	-0.32923
sit7	-0.78155	-0.09448	0.39335	0.23079	-0.45431	1.17306
sit8	-0.28732	-0.47352	0.29471	1.13215	0.69812	1.05344
sit9	-0.49324	-0.44884	-1.31854	0.78932	-0.38574	0.41597
sit10	-0.28009	0.43091	0.12225	-0.11790	-1.07266	0.45807
sit11	-0.44849	0.33200	-0.53896	0.68345	-0.96682	0.11691
sit12	-0.38850	0.68558	0.10462	0.08107	-1.10978	0.84504
sit13	-0.24996	0.74160	0.88642	-0.46709	-0.96946	0.74682
sit14	-0.31329	0.93929	1.93810	-1.27078	0.06283	0.14773
sit15	-0.14329	0.31112	-0.21270	0.24363	-0.61744	-0.52982
sit16	0.08956	0.29836	-0.18601	0.23428	-0.73164	-0.44261
sit17	0.05649	0.34914	-0.22849	0.14198	-0.76039	-0.60408
sit18	0.04513	0.40790	0.12272	-0.20091	-0.49665	-0.87755
sit19	0.16126	0.36126	-0.28794	-0.05345	-0.79294	-1.36200
sit20	0.16079	0.32644	-0.74873	0.46912	0.17568	-0.90766
sit21	0.14178	0.53551	-0.08106	-0.07021	0.58856	-0.24654
sit22	1.37614	-1.19047	0.74766	-0.35075	-0.22921	0.75808
sit23	0.98260	-0.51434	0.01123	0.40978	1.01197	0.84836
sit24	2.18633	-2.04860	0.35817	-0.29583	-1.26009	-0.39052
sit25	0.88257	-0.11921	-0.64734	0.34001	0.85793	-0.14280
sit26	0.63983	0.39438	-0.15997	-0.30089	1.09889	-0.66497
sit27	0.75833	0.80559	0.51815	-0.96863	0.42032	-0.74305
sit28	0.65324	1.09406	-1.68227	0.37796	0.43707	0.65309
sit29	0.73849	1.36252	-0.27161	-0.62819	1.42387	0.88211

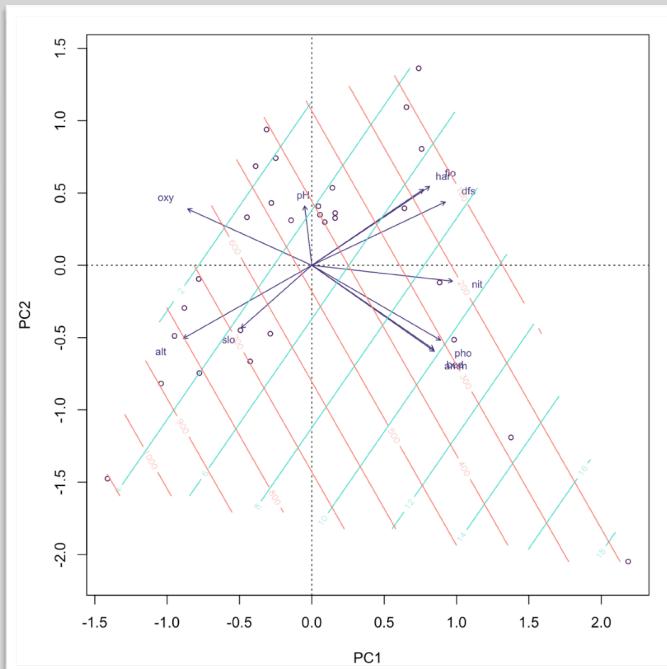
- **site scores:** the (scaled) coordinates of the objects (sites)

# Principal Component Analysis (PCA)



# Principal Component Analysis (PCA)

- the contours form a linear trend surface, i.e. they are perpendicular to the arrow
- this is the main problem of PCA**, as community data are *nonlinear*
- in general, **PCA should not be used for community data**
- PCA is useful for linear data
- can be used to derive new combined variables for other analyses when PCA explains a large proportion of variance
- often displays a ‘horseshoe’ effect, indicating a non-linear response



# Correspondence Analysis (CA)

- also an eigenvector method
- handles nonlinear species responses better than PCA
  - therefore better for community data
- CA is based on similar regression techniques as PCA, but with  $\chi^2$ -standardised data and weights
- then subjected to SVD and the eigenvalues and eigenvectors computed
- the ordination preserves  $\chi^2$ - ( $D_{16}$ ) rather than Euclidian ( $D_1$ ) distance between sites
  - $\chi^2$ -distance is not influenced by double 0s
  - no pre-transformation needed
  - suitable for species counts and presence/absence data

# Correspondence Analysis (CA)

- CA produces one axis fewer than  $\min[n, p]$
- as with PCA, orthogonal axes ranked in decreasing order of importance
- the variation represented is the total inertia, which is the SS of all the values in the  $\chi^2$  matrix
- individual eigenvalues will always be  $< 1$
- the variation represented along an axis is given by dividing the eigenvalues of the axis by the total inertia

# Correspondence Analysis (CA)

- CA approximates a **unimodal** response model
  - *i.e.* matches gradients better (fits better to env. data)
  - the species scores give the species maximum and the abundance decreases in every direction from the centroid of the species score
- in PCA species close to the origin (zero) change little and is poorly presented by the ordination, but in CA it may have its optimum there
- the horseshoe effect is weaker, but still not entirely gone

# Correspondence Analysis (CA)

- scaling of ordination plots
  - **scaling 1**—site scaling (rows are the centroids of columns)
    - *i.e.* sites that plot close together are similar i.t.o. their species relative frequencies
    - any site near a point representing a species will have a relatively large contribution by that species
  - **scaling 2**—species scaling (columns are the centroids of rows)
    - *i.e.* species points that are close together will have relatively similar frequencies among the objects (sites)
    - any species plotting close-by a point that represents an object (site) is more likely to be found at that object, or have a higher frequency there than at objects appearing further away in the ordination plot

# Correspondence Analysis (CA)

```
> spe
# A tibble: 29 x 27
   Cogo Satr Phph Babl Thth Teso Chna Pato Lele Sqce Baba Albi Gogo Eslu Pefl Rham Legi Scer Cyca Titi
   <int> <int>
1     0     3     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
2     0     5     4     3     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
3     0     5     5     5     0     0     0     0     0     0     0     0     0     0     1     0     0     0     0     0
4     0     4     5     5     0     0     0     0     0     1     0     0     0     1     2     2     0     0     0     0
5     0     2     3     2     0     0     0     0     5     2     0     0     0     2     4     4     0     0     2     0
6     0     3     4     5     0     0     0     0     0     1     2     0     0     1     1     1     0     0     0     0
7     0     5     4     5     0     0     0     0     0     1     1     0     0     0     0     0     0     0     0     0
8     0     0     1     3     0     0     0     0     0     0     5     0     0     0     0     0     0     0     0     1
9     0     1     4     4     0     0     0     0     0     2     2     0     0     0     1     0     0     0     0     0
10    1     3     4     1     1     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0
```

# ... with 19 more rows, and 7 more variables: Abbr <int>, Icme <int>, Gyce <int>, Ruru <int>, Blbj <int>, Alal <int>, Anan <int>

```
> (spe.ca <- cca(spe))
Call: cca(X = spe)

Inertia Rank
Total      1.167
Unconstrained 1.167  26
Inertia is scaled Chi-square

Eigenvalues for unconstrained axes:
 CA1   CA2   CA3   CA4   CA5   CA6   CA7   CA8
 0.6010 0.1444 0.1073 0.0834 0.0516 0.0418 0.0339 0.0288
(Showed only 8 of all 26 unconstrained eigenvalues)
```

# Correspondence Analysis (CA)

```
> summary(spe.ca)          # default scaling 2

Call:
cca(X = spe)

Partitioning of scaled Chi-square:
   Inertia Proportion
Total      1.167      1
Unconstrained 1.167      1

Eigenvalues, and their contribution to the scaled Chi-square

Importance of components:
           CA1    CA2    CA3    CA4    CA5    CA6    CA7    CA8    CA9    CA10   CA11   CA12
Eigenvalue 0.601 0.1444 0.10729 0.08337 0.05158 0.04185 0.03389 0.02883 0.01684 0.010826 0.010142 0.007886
Proportion Explained 0.515 0.1237 0.09195 0.07145 0.04420 0.03586 0.02904 0.02470 0.01443 0.009278 0.008691 0.006758
Cumulative Proportion 0.515 0.6387 0.73069 0.80214 0.84634 0.88220 0.91124 0.93594 0.95038 0.959655 0.968346 0.975104
           CA13   CA14   CA15   CA16   CA17   CA18   CA19   CA20   CA21   CA22   CA23
Eigenvalue 0.006123 0.004867 0.004606 0.003844 0.003067 0.001823 0.001642 0.001295 0.0008775 0.0004217 0.0002149
Proportion Explained 0.005247 0.004171 0.003948 0.003294 0.002629 0.001562 0.001407 0.001110 0.0007520 0.0003614 0.0001841
Cumulative Proportion 0.980351 0.984522 0.988470 0.991764 0.994393 0.995955 0.997362 0.998472 0.9992238 0.9995852 0.9997693
           CA24   CA25   CA26
Eigenvalue 0.0001528 8.949e-05 2.695e-05
Proportion Explained 0.0001309 7.669e-05 2.310e-05
Cumulative Proportion 0.9999002 1.000e+00 1.000e+00

Scaling 2 for species and site scores
* Species are scaled proportional to eigenvalues
* Sites are unscaled: weighted dispersion equal on all dimensions
```

continue...

# Correspondence Analysis (CA)

...continue

## Species scores

	CA1	CA2	CA3	CA4	CA5	CA6
Cogo	1.50075	-1.410293	0.268049	-0.307203	0.271777	-0.003465
Satr	1.66167	0.444129	0.57571	0.166073	-0.261870	-0.326590
Phph	1.28545	0.285328	-0.04768	0.018126	0.043847	0.200732
Babl	0.98662	0.360900	-0.35265	-0.009021	-0.012231	0.253429
Tth	1.55554	-1.389752	0.80805	-0.468471	0.471301	0.225409
Teso	0.99709	-1.479902	-0.48835	0.079397	-0.105715	-0.332445
Chna	-0.54916	-0.051534	0.01123	-0.096004	-0.382763	0.134807
Pato	-0.18478	-0.437710	-0.57438	0.424267	-0.587158	0.091866
Lele	0.01337	-0.095342	-0.57672	0.212017	0.126668	-0.389103
Sqce	0.01078	0.140577	-0.34811	-0.538268	0.185284	0.167087
Baba	-0.33363	-0.300682	-0.04929	0.170961	-0.157203	0.103668
Albi	-0.38357	-0.255310	-0.20136	0.374057	-0.385866	0.239001
Gogo	-0.32152	-0.034382	-0.07423	-0.031236	0.014417	-0.156351
Eslu	-0.26165	0.187282	0.00617	0.183771	0.295142	-0.262808
Pefl	-0.28913	0.121044	-0.18919	0.367615	0.218087	-0.163675
Rham	-0.660298	-0.057369	0.20341	0.214299	-0.058977	0.211926
Legi	-0.58669	-0.082467	0.21198	0.050175	-0.120456	0.108724
Scer	-0.61815	0.124733	0.13339	0.147190	0.317736	-0.340380
Cyca	-0.57951	-0.110732	0.20173	0.308547	0.006854	0.153224
Titi	-0.37880	0.138023	-0.07825	0.095793	0.256285	-0.029245
Abbr	-0.70235	0.011155	0.40242	0.211582	0.138186	0.132297
ICme	-0.73238	-0.009098	0.55678	0.321852	0.281812	0.172271
Gyce	-0.69300	0.038971	0.37688	-0.183965	-0.051945	-0.011126
Ruru	-0.44181	0.176915	-0.23691	-0.345104	0.129676	-0.043802
Bibj	-0.70928	0.032317	0.40924	0.030224	0.049050	0.114560
Alal	-0.63114	0.053594	0.15204	-0.661381	-0.414796	-0.206611
Anan	-0.63578	-0.041894	0.30093	0.224044	0.030444	0.203160

continue...

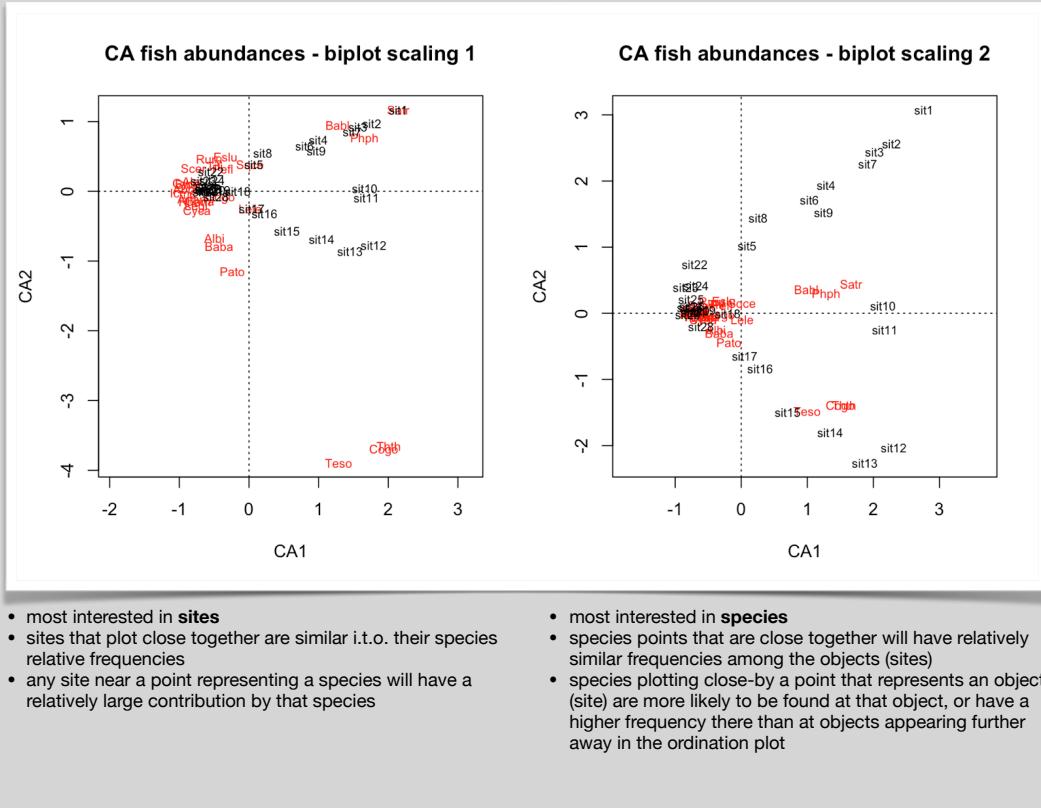
# Correspondence Analysis (CA)

...continue

## Site scores (weighted averages of species scores)

	CA1	CA2	CA3	CA4	CA5	CA6
sit1	2.76488	3.076306	5.3657489	1.99192	-5.07714	-7.80447
sit2	2.27540	2.565531	1.2659130	0.87538	-1.89139	-0.13887
sit3	2.01823	2.441224	0.5144079	0.79436	-1.03741	0.56015
sit4	1.28485	1.935664	-0.2509482	0.76470	0.54752	0.10579
sit5	0.08875	1.015182	-1.4555434	0.47672	2.69249	-2.92498
sit6	1.03188	1.712163	-0.9544059	0.01584	0.91932	0.39856
sit7	1.91427	2.256208	-0.0001407	0.39844	-1.07017	0.32127
sit8	0.25591	1.443008	-2.5777721	-3.41400	2.36613	2.71741
sit9	1.24517	1.526391	-1.9635663	-0.41230	0.69647	1.51859
sit10	2.14581	0.110278	1.6108693	-0.82023	0.53918	1.01153
sit11	2.17418	-0.251649	1.5845397	-0.81483	0.52623	1.05501
sit12	2.30944	-2.034439	1.9181448	-0.60481	0.64435	-0.14844
sit13	1.87141	-2.262503	1.1066796	-0.80840	1.09542	0.11038
sit14	1.34659	-1.805967	-0.6441505	-0.52803	0.76871	-0.67165
sit15	0.70214	-1.501167	-1.9735888	0.98502	-0.93585	-1.27168
sit16	0.28775	-0.836803	-1.2259108	0.73302	-1.57036	0.57315
sit17	0.05299	-0.647958	-0.9234228	0.35770	-0.95401	0.77738
sit18	-0.20584	-0.007252	-1.0154343	0.07041	-1.03450	0.51442
sit19	-0.57879	0.042849	-0.3660551	-0.15019	-0.61357	0.10115
sit20	-0.67320	0.038875	0.1194956	0.17256	-0.14686	-0.12818
sit21	-0.71933	0.014694	0.2204186	0.13598	0.09459	-0.02068
sit22	-0.70438	0.735398	-0.6546250	-6.61523	-2.49441	-1.73215
sit23	-0.83976	0.390120	0.5605295	-4.38864	-2.56916	-0.96702
sit24	-0.68476	0.418842	-0.2860819	-2.88336	-0.37540	-3.93791
sit25	-0.75880	0.210204	0.5894891	-0.70004	-0.01880	-0.10779
sit26	-0.75046	0.100869	0.5531191	-0.12946	0.29164	0.11280
sit27	-0.77878	0.088976	0.7379012	0.05204	0.40940	0.43236
sit28	-0.60815	-0.203235	0.5522726	0.43621	0.15010	0.25618
sit29	-0.80860	-0.019592	0.6686542	0.88136	0.52744	0.16456

# Correspondence Analysis (CA)



# Correspondence Analysis (CA)

## Fit and Plot Smooth Surfaces of Variables on Ordination

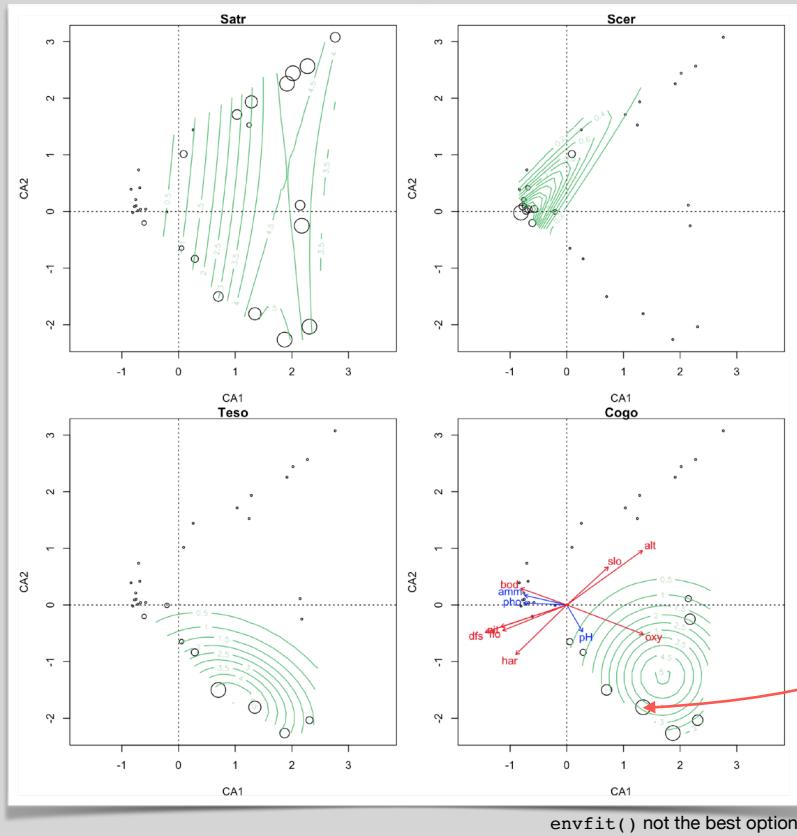
```

require('viridis')
palette(viridis(8))
par(mar = c(4, 4, 0.9, 0.5) + .1, mfrow = c(2, 2))
with(spe, tmp <- ordisurf(spe.ca ~ Satr, bubble = 3,
                           family = quasipoisson, knots = 2, col = 6,
                           display = "sites", main = "Satr"))
abline(h = 0, v = 0, lty = 3)
with(spe, tmp <- ordisurf(spe.ca ~ Scer, bubble = 3,
                           family = quasipoisson, knots = 2, col = 6,
                           display = "sites", main = "Scer"))
abline(h = 0, v = 0, lty = 3)
with(spe, tmp <- ordisurf(spe.ca ~ Teso, bubble = 3,
                           family = quasipoisson, knots = 2, col = 6,
                           display = "sites", main = "Teso"))
abline(h = 0, v = 0, lty = 3)
with(spe, tmp <- ordisurf(spe.ca ~ Cogo, bubble = 3,
                           family = quasipoisson, knots = 2, col = 6,
                           display = "sites", main = "Cogo"))
abline(h = 0, v = 0, lty = 3)

# A posteriori projection of environmental variables in a CA
# The last plot produced (CA scaling 2) must be active
(spe.ca.env <- envfit(spe.ca, env, scaling = 2)) # Scaling 2 is default
plot(spe.ca.env)
# Plot significant variables with a different colour
plot(spe.ca.env, p.max = 0.05, col = "red")

```

biplot scaling 2



- instead of an arrow of increase, the species score is seen as a centre of abundance
- the species scores give the species maximum and the abundance decreases to every direction from the centroid given by the species score

`envfit()` not the best option

## Principal Coordinates Analysis (PCoA)

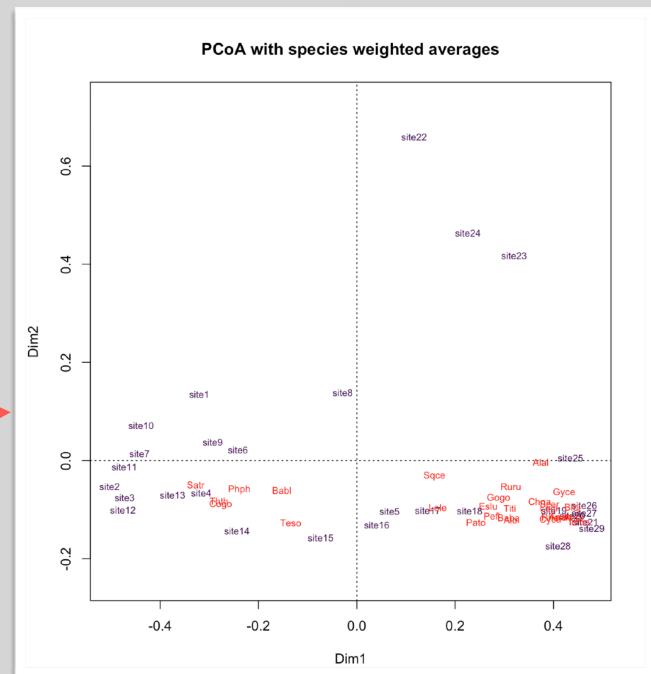
- a.k.a. ‘Classical multidimensional scaling’ (MDS), hence the function name in vegan, `cmdscale()`
- PCoA uses dissimilarities among rows instead of rectangular data of rows and columns
- if dissimilarities are Euclidean distances, then PCoA is equal to PCA
- the strength of the method is that it can be used with other dissimilarities that are better suited for communities of species

# Principal Coordinates Analysis (PCoA)

```

spe.bray <- vegdist(spe)
spe.b.pcoa <- cmdscale(spe.bray, k = (nrow(spe) - 1), eig = TRUE)
# Plot of the sites
# dev.new(title = "PCoA on fish species - Percentage difference")
ordiplotscores(spe.b.pcoa, choices = c(1, 2)), type = "t",
                main = "PCoA with species weighted averages")
abline(v = 0, h = 0, lty = 3)
# Add weighted average projection of species
spe.wa <- wascores(spe.b.pcoa$points[,1:2], spe)
text(spe.wa, rownames(spe.wa), cex = 0.7, col = "red")

```



# Non-Metric Multidimensional Scaling (NMDS)

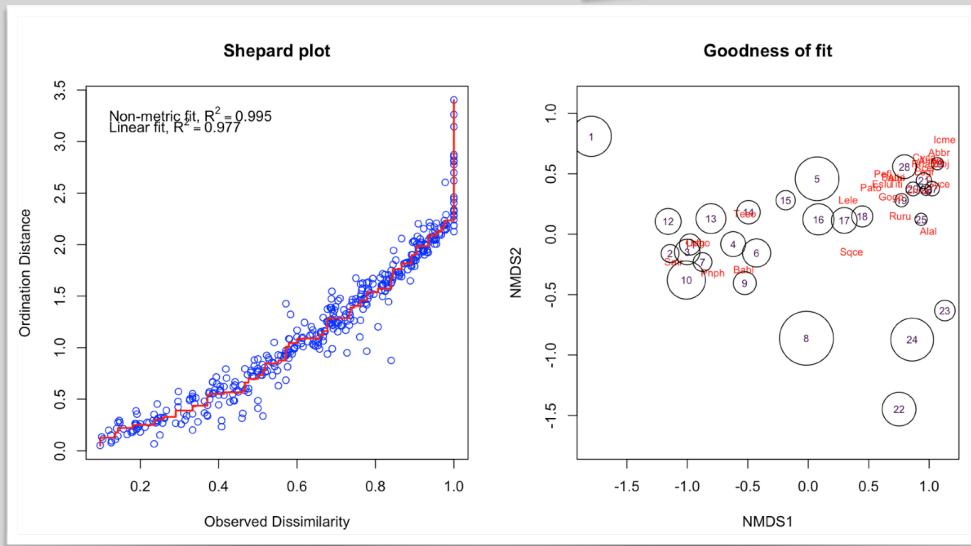
- does not preserve the exact distances among objects in an ordination plot
  - instead, represent as well as possible the ordering relationships among objects in a small and specified number of axes
  - like PCoA, NMDS can produce ordinations of objects from any distance matrix
  - non-linear mapping of dissimilarities onto a low-dimensional ordination
    - i.e. Euclidean distances of points in the ordination space are *rank-order* similar to community dissimilarities
    - the ordination space is metric, but the regression is non-metric
  - rank orders of dissimilarities cannot be exactly preserved by rank-orders of ordination distances in low-dimensional space, and this causes stress
  - stress: scatter of observed dissimilarities against expected monotone regression

# Non-Metric Multidimensional Scaling (NMDS)

```

spe.nmds <- metaMDS(spe, distance = "bray")
spe.nmds
spe.nmds$stress
# dev.new(title="NMDS on fish species - Percentage difference")
plot(spe.nmds, type = "t", main = paste("NMDS/Percentage difference - Stress =", round(spe
.nmds$stress,3)))
# Shepard plot and goodness of fit
# dev.new(title="NMDS - Shepard plot", width=12, height=6)
par(mfrow = c(1, 2))
stressplot(spe.nmds, main = "Shepard plot")
gof <- goodness(spe.nmds)
plot(spe.nmds, type = "t", main = "Goodness of fit")
points(spe.nmds, display = "sites", cex = gof * 300)

```



## Constrained ordinations

- recap:
  - constrained ordination adds a level of statistical testing (next topic)
  - also called direct gradient analysis
  - typically **uses explanatory variables** (in the env. matrix) **to explain the patterns** (or variability) **seen in the spp. matrix**
- (A)...two matrices; a grouping variable (factor), and a multivariate (>>2) response
  - ✓ Discriminant Analysis (**DA**)
  - Analysis of Similarities (**ANOSIM**)
  - Mantel test
  - etc...
- (B)...two matrices; a >>2 explanatory matrix (e.g. env.), and a >>2 response matrix (e.g. spp.)
  - ✓ Redundancy Discriminant Analysis (**RDA**)
  - Canonical Correspondence Analysis (**CCA**)
  - Distance-based redundancy analysis (**db-RDA**)
  - Canonical analysis of principal coordinates (**CAP**)

# Multivariate Techniques

Obs	Group	X-set				Y-set				
1	A	a <sub>11</sub>	a <sub>12</sub>	a <sub>13</sub>	...	a <sub>1p</sub>	b <sub>11</sub>	b <sub>12</sub>	b <sub>13</sub>	... b <sub>1m</sub>
2		a <sub>21</sub>	a <sub>22</sub>	a <sub>23</sub>	...	a <sub>2p</sub>	b <sub>21</sub>	b <sub>22</sub>	b <sub>23</sub>	... b <sub>2m</sub>
3		a <sub>31</sub>	a <sub>32</sub>	a <sub>33</sub>	...	a <sub>3p</sub>	b <sub>31</sub>	b <sub>32</sub>	b <sub>33</sub>	... b <sub>3m</sub>
.		.	.	.	...	.	.	.	.	...
.		.	.	.	...	.	.	.	.	...
n		a <sub>n1</sub>	a <sub>n2</sub>	a <sub>n3</sub>	...	a <sub>np</sub>	b <sub>n1</sub>	b <sub>n2</sub>	b <sub>n3</sub>	... b <sub>nm</sub>
n+1	C	c <sub>11</sub>	c <sub>12</sub>	c <sub>13</sub>	...	c <sub>1p</sub>	<span style="position: absolute; right: -10px; top: -10px; width: 20px; height: 20px; background-color: red; border-radius: 50%;"></span>			
n+2		c <sub>21</sub>	c <sub>22</sub>	c <sub>23</sub>	...	c <sub>2p</sub>	<span style="position: absolute; right: -10px; top: -10px; width: 20px; height: 20px; background-color: red; border-radius: 50%;"></span>			
n+3		c <sub>31</sub>	c <sub>32</sub>	c <sub>33</sub>	...	c <sub>3p</sub>	<span style="position: absolute; right: -10px; top: -10px; width: 20px; height: 20px; background-color: red; border-radius: 50%;"></span>			
.		.	.	.	...	.	<span style="position: absolute; right: -10px; top: -10px; width: 20px; height: 20px; background-color: red; border-radius: 50%;"></span>			
N		c <sub>n1</sub>	c <sub>n2</sub>	c <sub>n3</sub>	...	c <sub>np</sub>	<span style="position: absolute; right: -10px; top: -10px; width: 20px; height: 20px; background-color: red; border-radius: 50%;"></span>			

**(A) Constrained ordination**  
Discriminant Analysis (DA)  
Analysis of Similarities (ANOSIM)  
Mantel test  
etc

## (A) Constrained ordination

## Discriminant Analysis (DA)

### Analysis of Similarities (ANOSIM)

## Mantel test

etc...

<http://www.umass.edu/landeco/teaching/multivariate/schedule/summary.handouts.pdf>

# Multivariate Techniques

## (B) Constrained ordination

## ( ) Redundancy Discriminant Analysis (RDA)

## Canonical Correspondence Analysis (CCA)

## Distance-based redundancy analysis (db-RDA)

## Distance-based redundancy analysis (db-RDA), Canonical analysis of principal coordinates (CAP)

TOPIC 5

## UNCONSTRAINED ORDINATION



TOPIC 6

## CONSTRAINED ORDINATION

