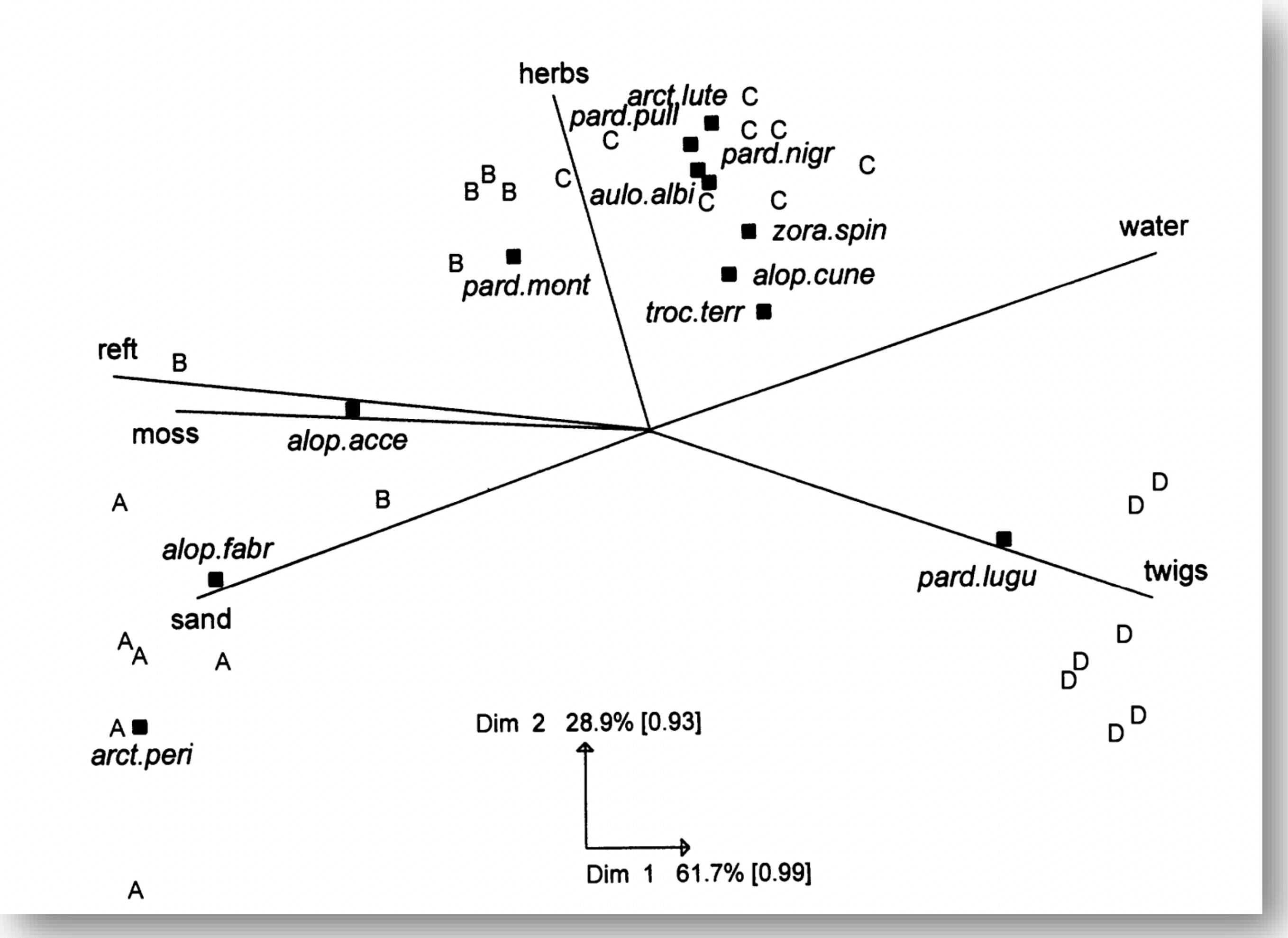
PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

- **Redundancy Analysis (RDA)**

REDUNDANCY ANALYSIS

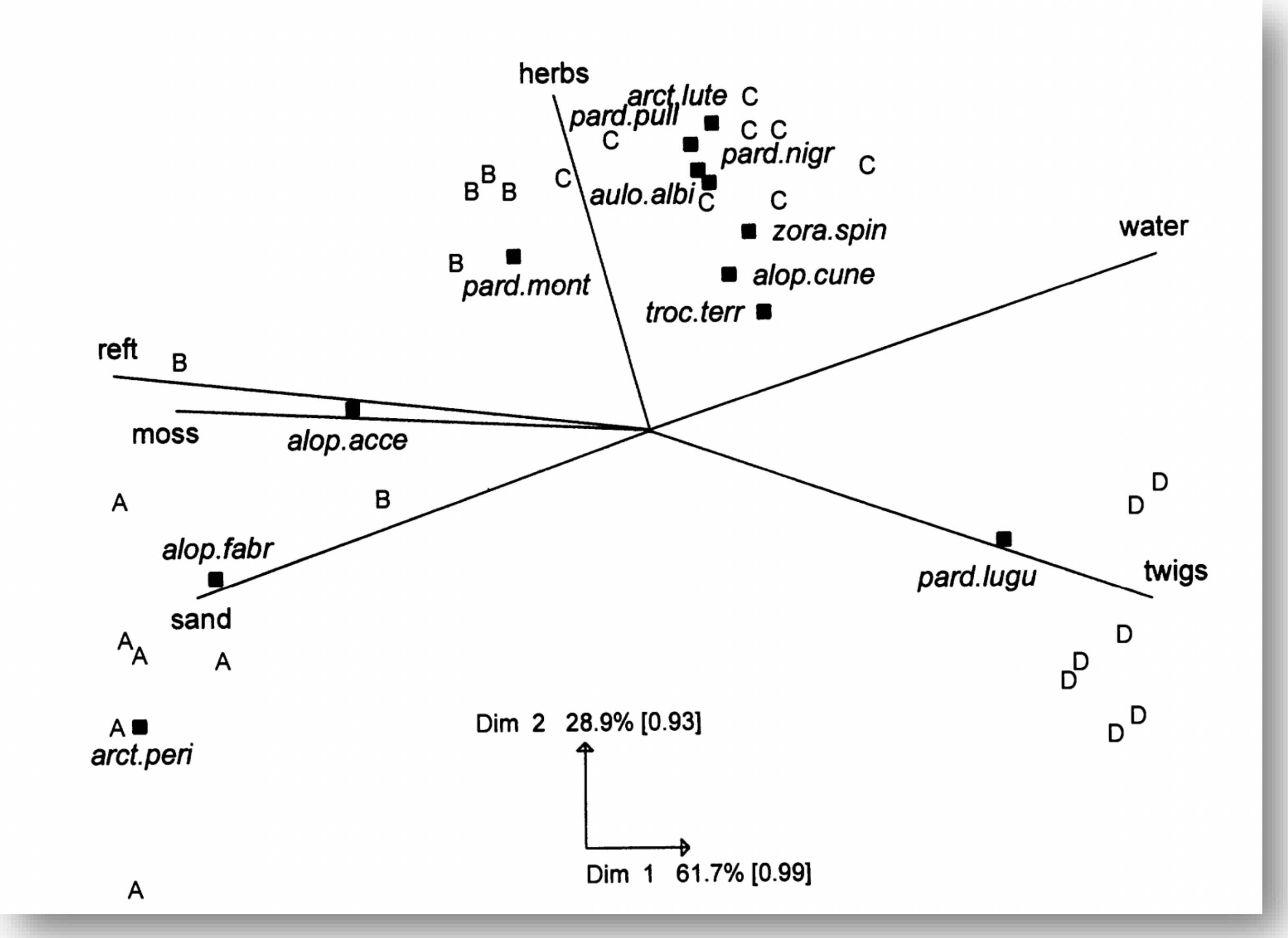
- RDA is a multivariate statistical analysis that examines the relationship between a set of response variables (e.g., species composition) and a set of explanatory variables (e.g., environmental variables).



REDUNDANCY ANALYSIS

- RDA is a multivariate statistical analysis that examines the relationship between a set of response variables (e.g., species composition) and a set of explanatory variables (e.g., environmental variables).

- RDA can be used to identify which explanatory variables are most strongly associated with the response variables, and to visualize the relationship between the response variables and the explanatory variables in a “triplot”.

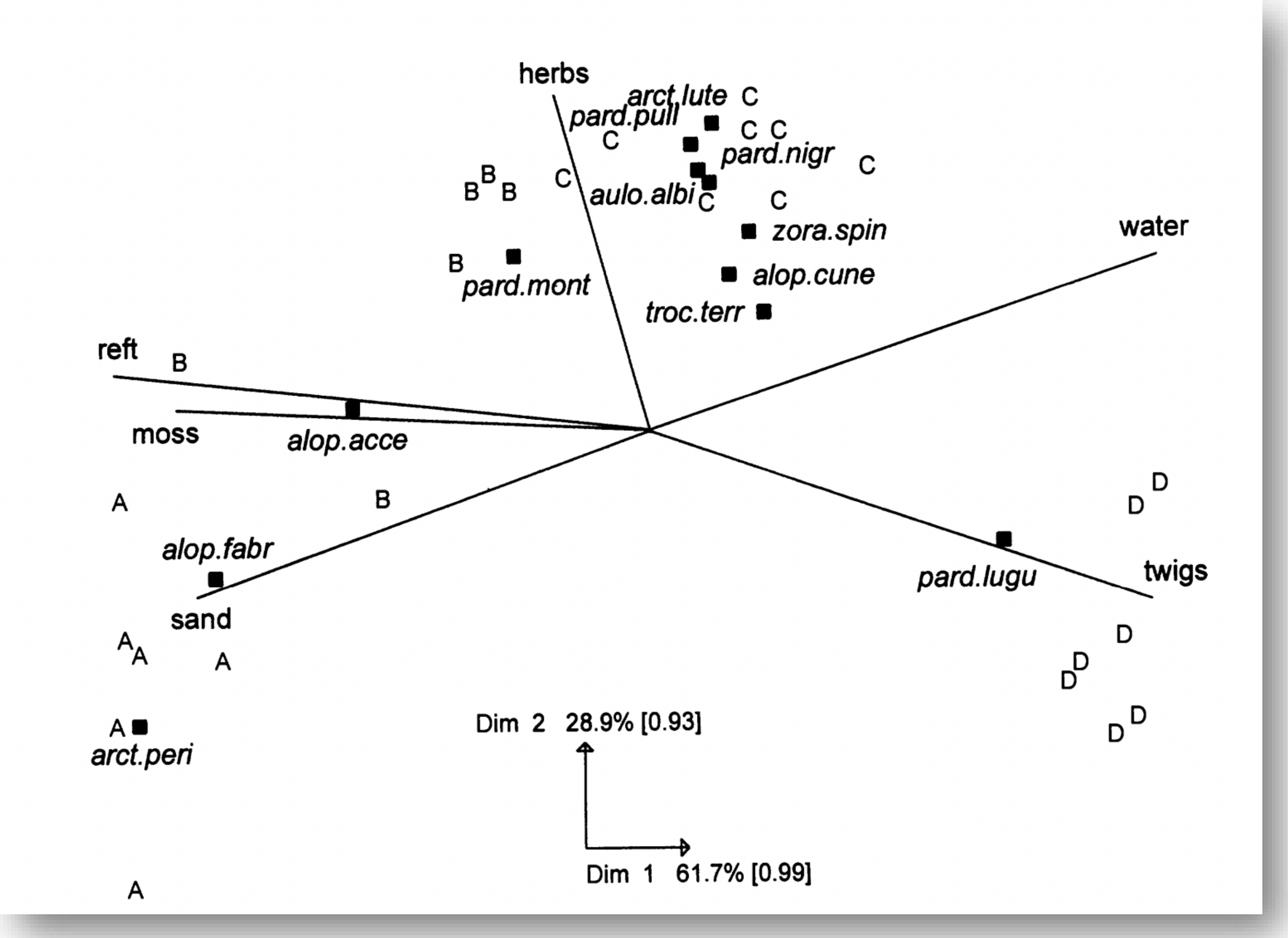


REDUNDANCY ANALYSIS

- RDA is a multivariate statistical analysis that examines the relationship between a set of response variables (e.g., species composition) and a set of explanatory variables (e.g., environmental variables).

- RDA can be used to identify which explanatory variables are most strongly associated with the response variables, and to visualize the relationship between the response variables and the explanatory variables in a “triplot”.

- RDA is commonly used in ecology to study the relationship between species composition and environmental variables, but can also be applied to other fields such as chemistry and genetics.

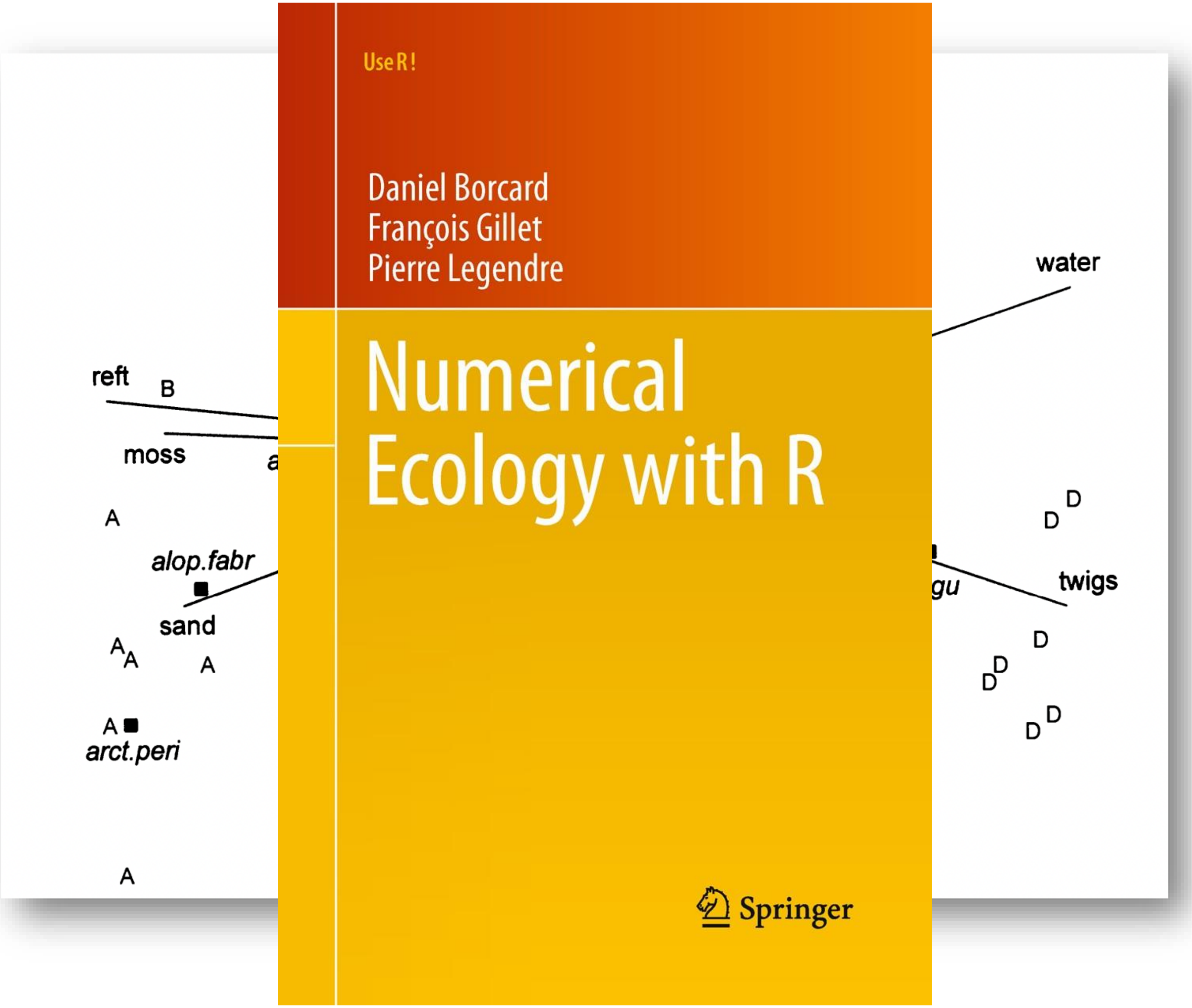


REDUNDANCY ANALYSIS

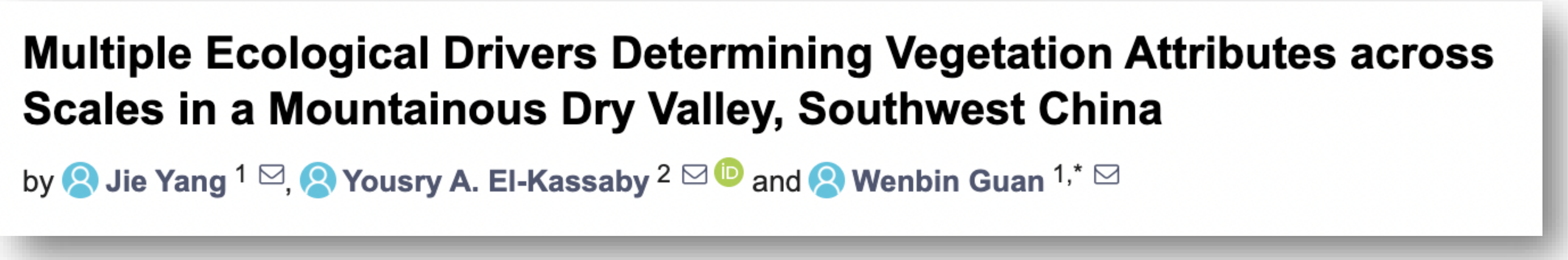
- RDA is a multivariate statistical analysis that examines the relationship between a set of response variables (e.g., species composition) and a set of explanatory variables (e.g., environmental variables).

- RDA can be used to identify which explanatory variables are most strongly associated with the response variables, and to visualize the relationship between the response variables and the explanatory variables in a “triplot”.

- RDA is commonly used in ecology to study the relationship between species composition and environmental variables, but can also be applied to other fields such as chemistry and genetics.



REDUNDANCY ANALYSIS: EXAMPLES



Multiple Ecological Drivers Determining Vegetation Attributes across Scales in a Mountainous Dry Valley, Southwest China

by Jie Yang [1], Yousry A. El-Kassaby [2] and Wenbin Guan [1,*]

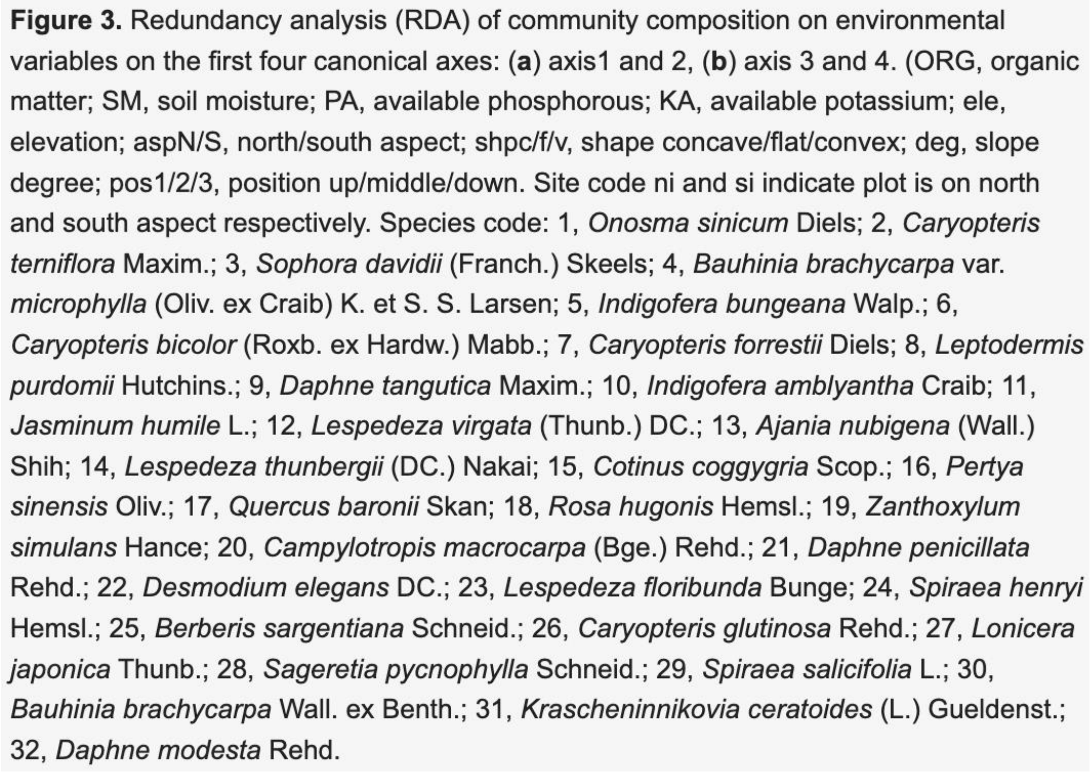
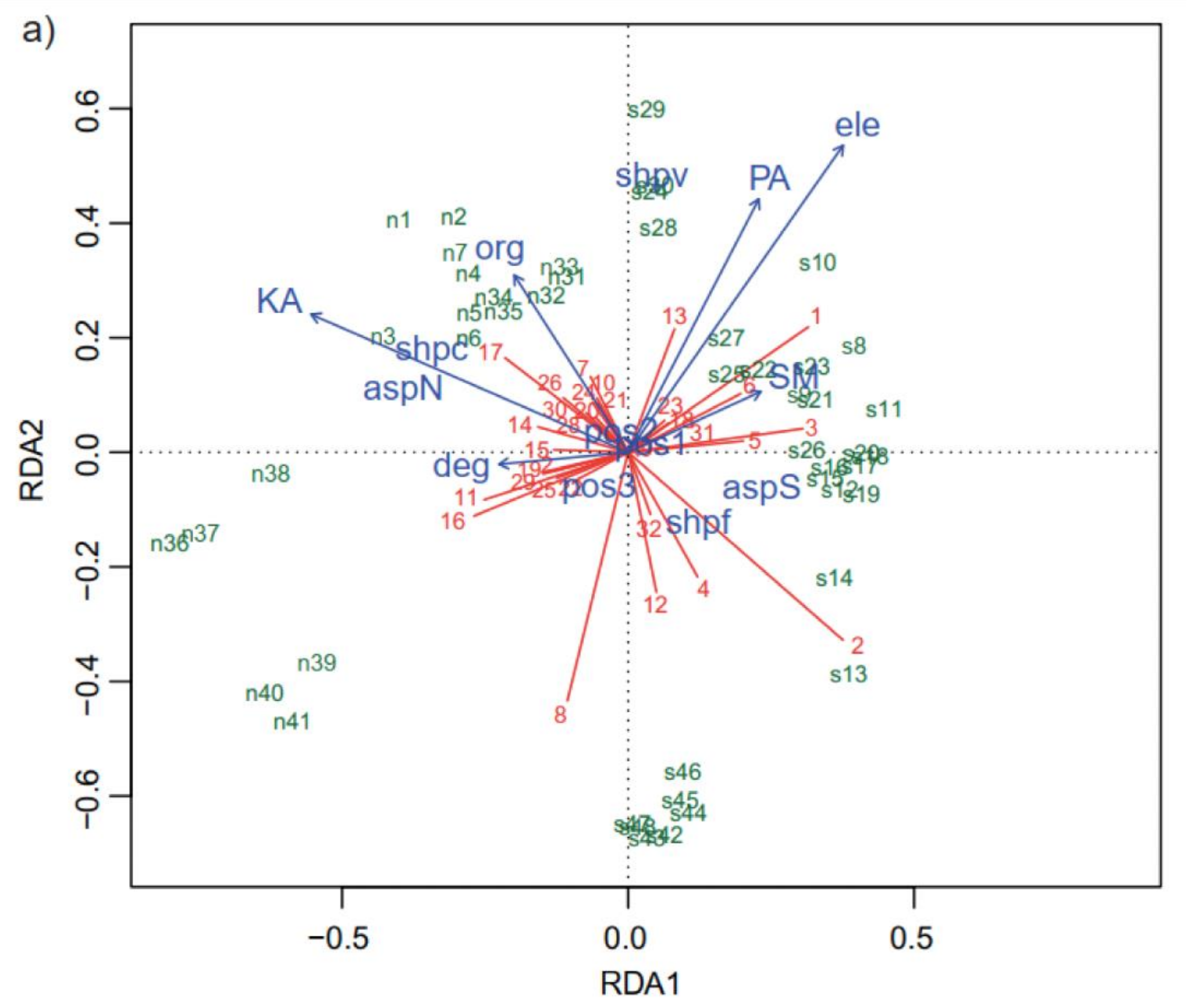
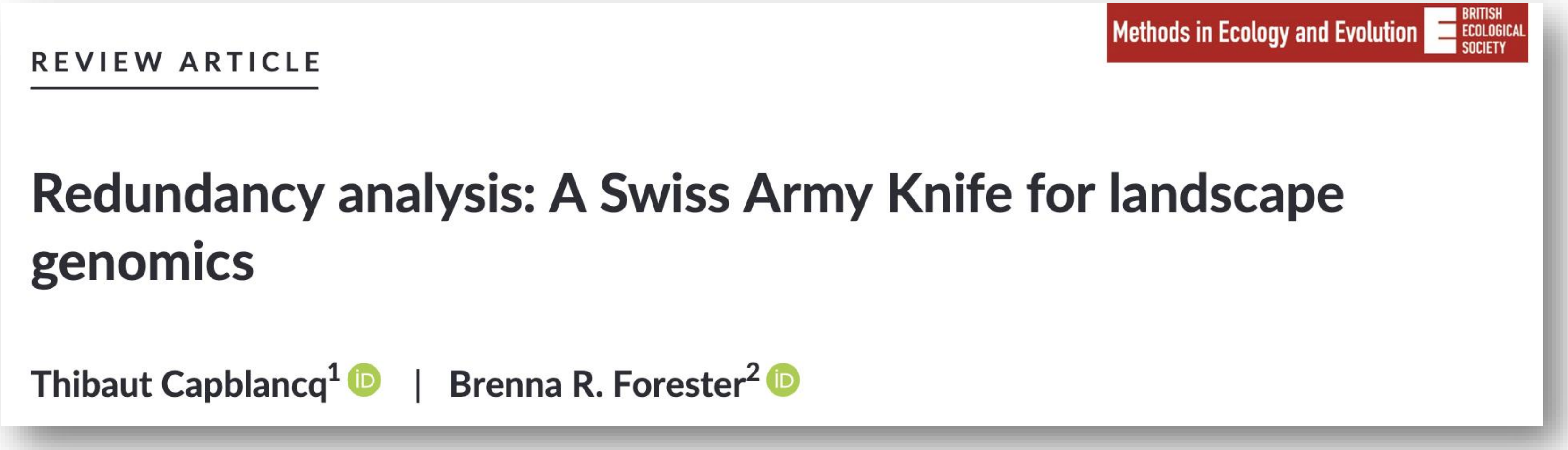


Figure 3. Redundancy analysis (RDA) of community composition on environmental variables on the first four canonical axes: (**a**) axis1 and 2, (**b**) axis 3 and 4. (ORG, organic matter; SM, soil moisture; PA, available phosphorous; KA, available potassium; ele, elevation; aspN/S, north/south aspect; shpc/f/v, shape concave/flat/convex; deg, slope degree; pos1/2/3, position up/middle/down. Site code ni and si indicate plot is on north and south aspect respectively. Species code: 1, *Onosma sinicum* Diels; 2, *Caryopteris terniflora* Maxim.; 3, *Sophora davidii* (Franch.) Skeels; 4, *Bauhinia brachycarpa* var. *microphylla* (Oliv. ex Craib) K. et S. S. Larsen; 5, *Indigofera bungeana* Walp.; 6, *Caryopteris bicolor* (Roxb. ex Hardw.) Mabb.; 7, *Caryopteris forrestii* Diels; 8, *Leptodermis purdomii* Hutchins.; 9, *Daphne tangutica* Maxim.; 10, *Indigofera amblyantha* Craib; 11, *Jasminum humile* L.; 12, *Lespedeza virgata* (Thunb.) DC.; 13, *Ajania nubigena* (Wall.) Shih; 14, *Lespedeza thunbergii* (DC.) Nakai; 15, *Cotinus coggygria* Scop.; 16, *Pertya sinensis* Oliv.; 17, *Quercus baronii* Skan; 18, *Rosa hugonis* Hemsl.; 19, *Zanthoxylum simulans* Hance; 20, *Campylotropis macrocarpa* (Bge.) Rehd.; 21, *Daphne penicillata* Rehd.; 22, *Desmodium elegans* DC.; 23, *Lespedeza floribunda* Bunge; 24, *Spiraea henryi* Hemsl.; 25, *Berberis sargentiana* Schneid.; 26, *Caryopteris glutinosa* Rehd.; 27, *Lonicera japonica* Thunb.; 28, *Sageretia pycnophylla* Schneid.; 29, *Spiraea salicifolia* L.; 30, *Bauhinia brachycarpa* Wall. ex Benth.; 31, *Krascheninnikovia ceratoides* (L.) Gueldenst.; 32, *Daphne modesta* Rehd.



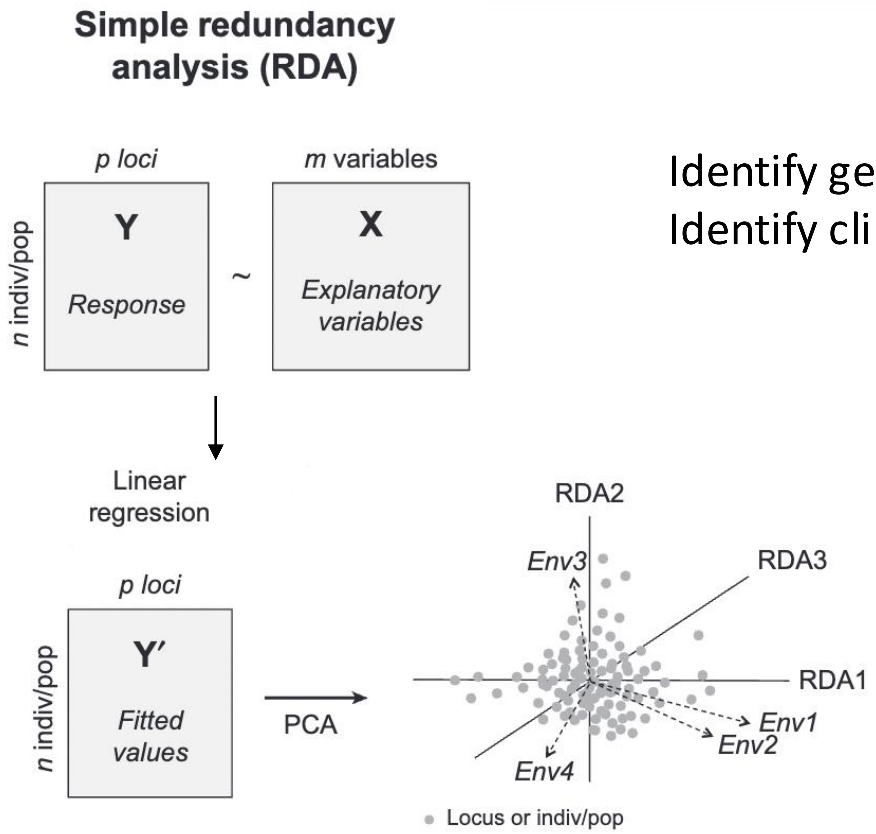


REVIEW ARTICLE

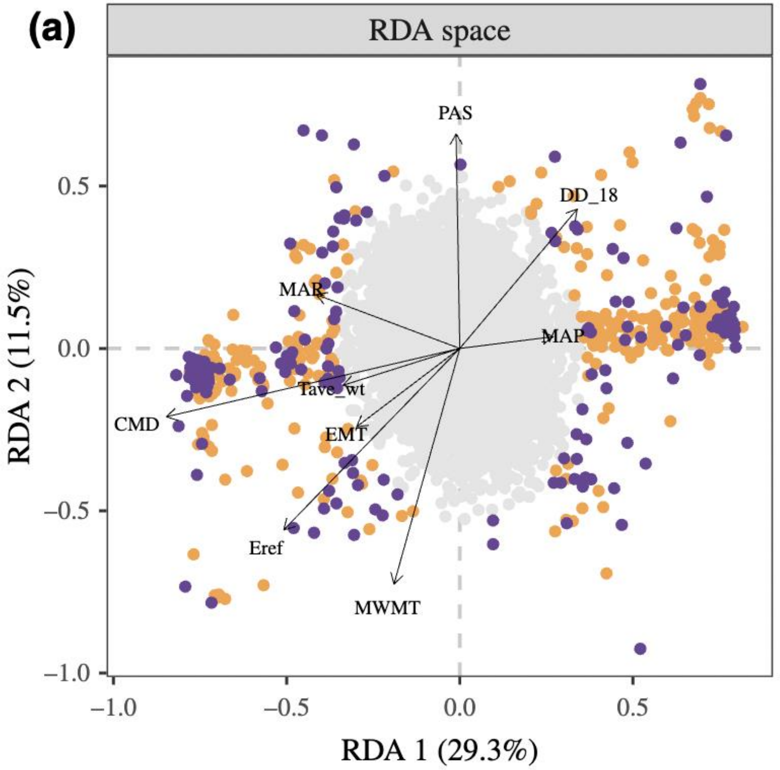
Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

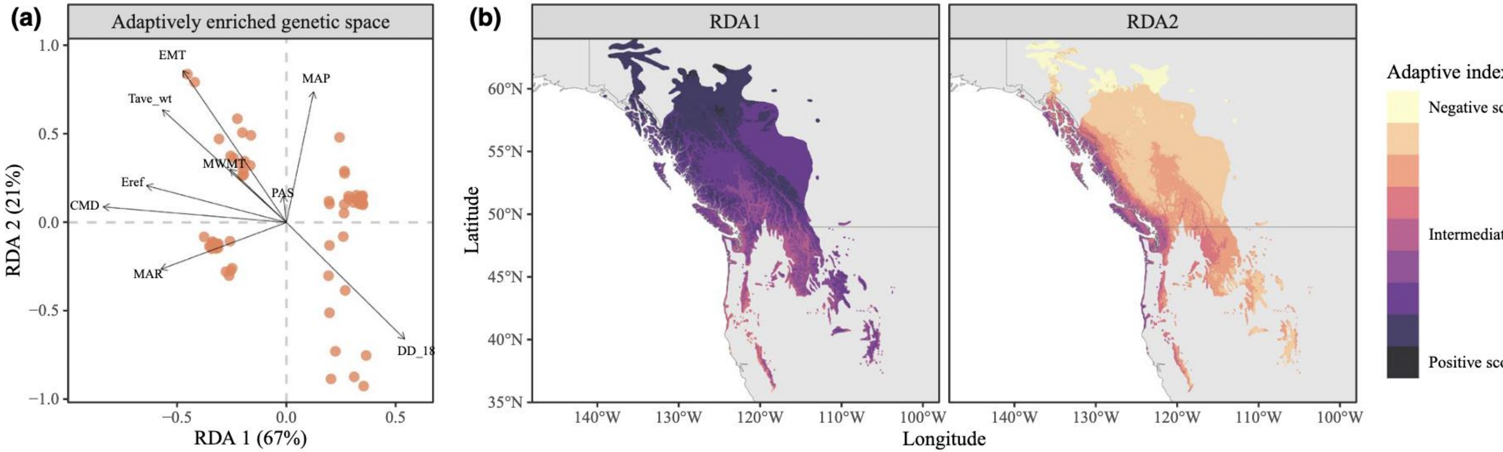
Redundancy analysis: A Swiss Army Knife for landscape genomics

Thibaut Capblancq[1] | Brenna R. Forester[2]



Identify genetic variation associated to climate
Identify climate variables important for genetics

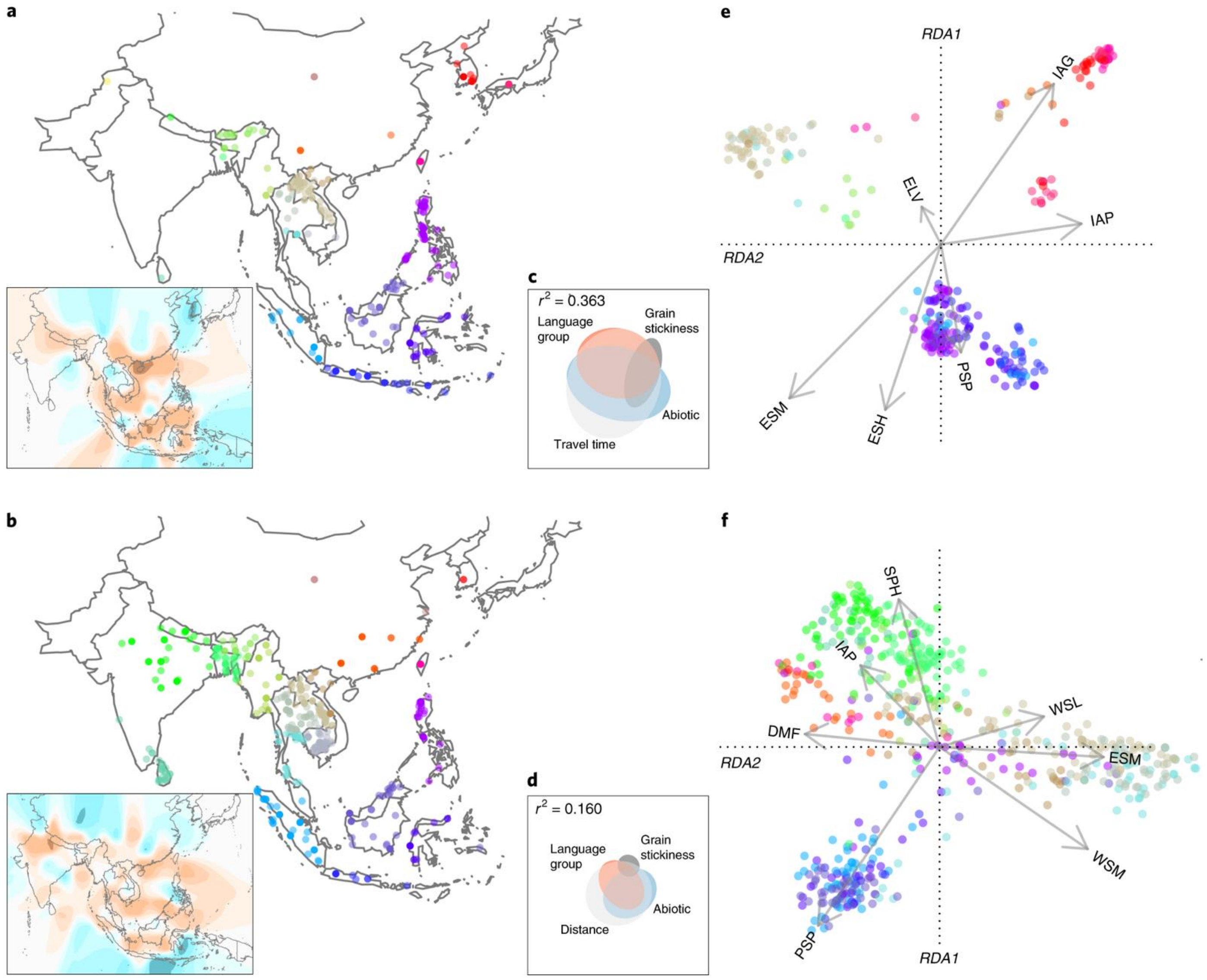




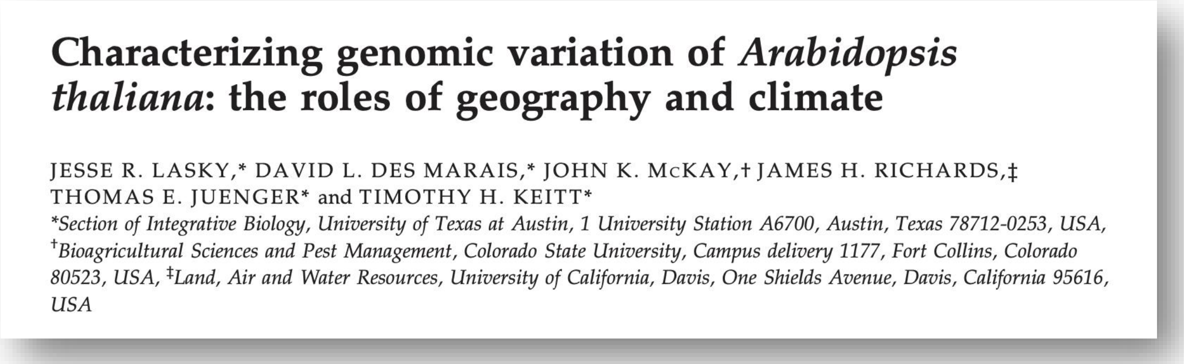
nature > nature plants > articles > article

Article | Published: 15 May 2020

Genomic history and ecology of the geographic spread of rice

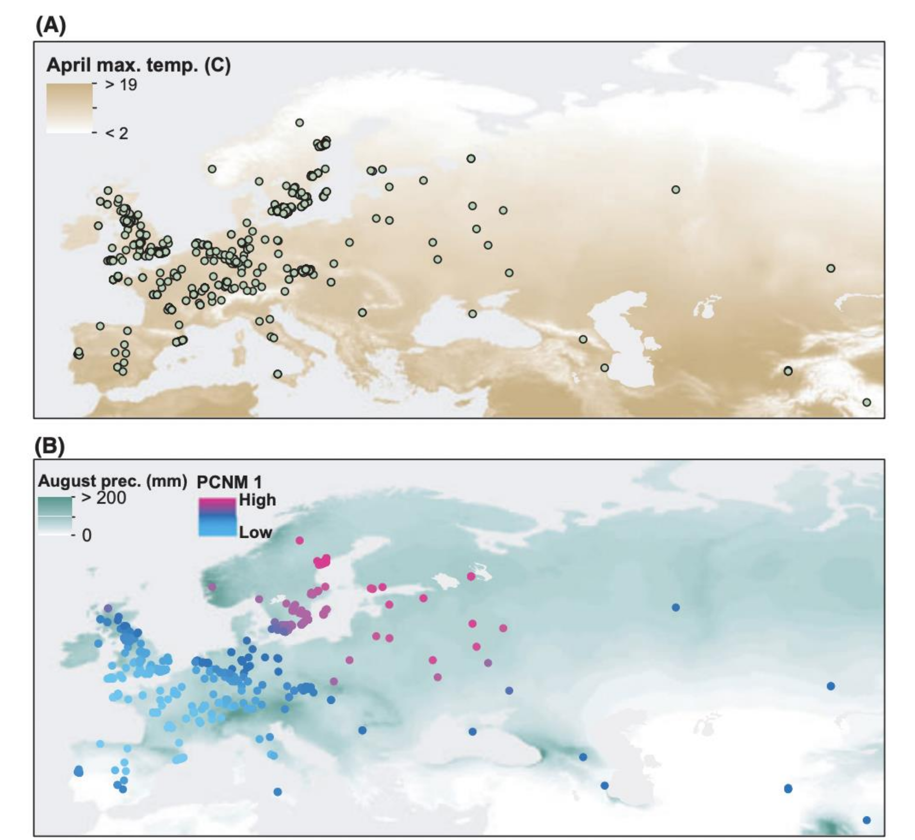


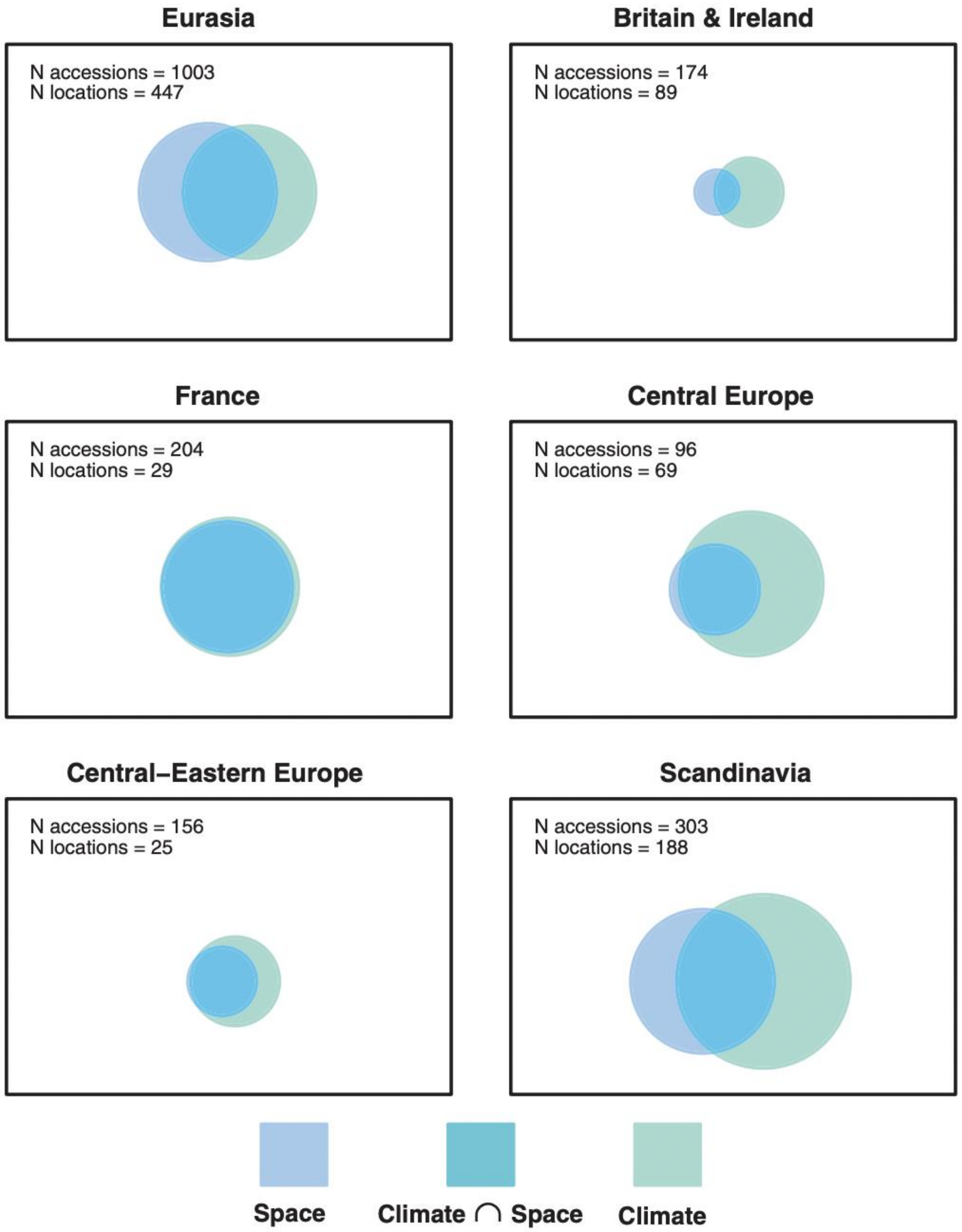
a,**b**, Maps of collection sites for *japonica* (**a**) and *indica* (**b**) landraces used in this study. Colours represent regions of origin. Insets show effective migration surfaces representing migration barriers (orange) and channels (cyan). **c**,**d**, Genomic diversity is best explained by a combination of four factors represented in Euler Plots for *japonica* (**c**) and *indica* (**d**): travel time (migration resistance) or geographic distance, abiotic variables (temperature, moisture and soil characteristics), linguistic group and culinary properties (stickiness). Fields of squares represent total genomic variation, while elliptic shapes represent genomic variation explained by a particular group of variables calculated using variance partitioning with RDA ordination (*japonica*, $n = 317$; *indica*, $n = 656$). **e**,**f**, Genotypes of *japonica* (**e**) and *indica* (**f**) projected on the first two canonical axes of RDA. Arrows represent environmental predictors that strongly correlate with a maximal proportion of variation in linear combinations of SNPs. IAG, interannual coefficient of GDD variation; ESH, end of growing season heat; ELV, elevation; PSP, pre-growing season precipitation; ESM, early growing season minimum temperature; IAP, interannual coefficient of precipitation variation; DMF, distance to major freshwater source; WSM, whole growing-season minimum temperature; WSL, whole growing-season length; SPH, soil pH.



Characterizing genomic variation of *Arabidopsis thaliana*: the roles of geography and climate

JESSE R. LASKY,* DAVID L. DES MARAIS,* JOHN K. McKAY,† JAMES H. RICHARDS,‡
THOMAS E. JUENGER* and TIMOTHY H. KEITT*
**Section of Integrative Biology, University of Texas at Austin, 1 University Station A6700, Austin, Texas 78712-0253, USA, †Bioagricultural Sciences and Pest Management, Colorado State University, Campus delivery 1177, Fort Collins, Colorado 80523, USA, ‡Land, Air and Water Resources, University of California, Davis, One Shields Avenue, Davis, California 95616, USA*

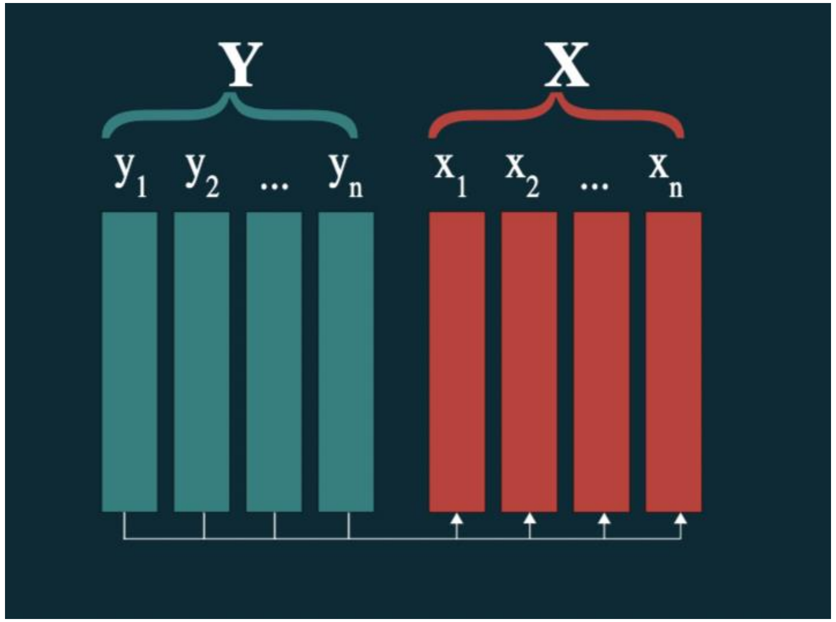




REDUNDANCY ANALYSIS

- RDA is the multivariate (meaning *multiresponse*) technique analogue of regression.
- The method uses a mix of linear regression and principal components analysis (PCA).
- Conceptually, RDA is a multivariate (meaning multiresponse) multiple linear regression followed by a PCA of the table of fitted values.

X a matrix of **explanatory variables**
Y a matrix of **response variables**



RDA procedure works on both **centered** matrices. This simply means that that the average of the variable is subtracted from each observation.

$$\bar{X}_j = \sum_{i=1}^{n} X_{ij} = 0$$

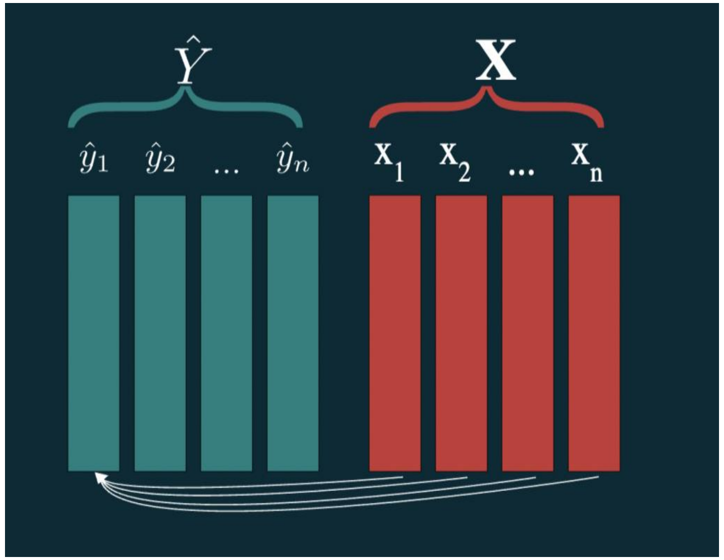
$$\bar{Y}_j = \sum_{i=1}^{n} Y_{ij} = 0$$

These steps are outlined by Borcard, Gillet, and Legendre (2011).

1. Regress each (centered) y variable on explanatory matrix X and compute the fitted (ŷ) and residuals (y_{res}) vectors.
2. Create a new matrix (Ŷ) containing all the fitted vectors (ŷ).
3. Compute a PCA on Ŷ. This will produce a vector of canonical eigenvalues and a matrix U of canonical eigenvectors (principal components).

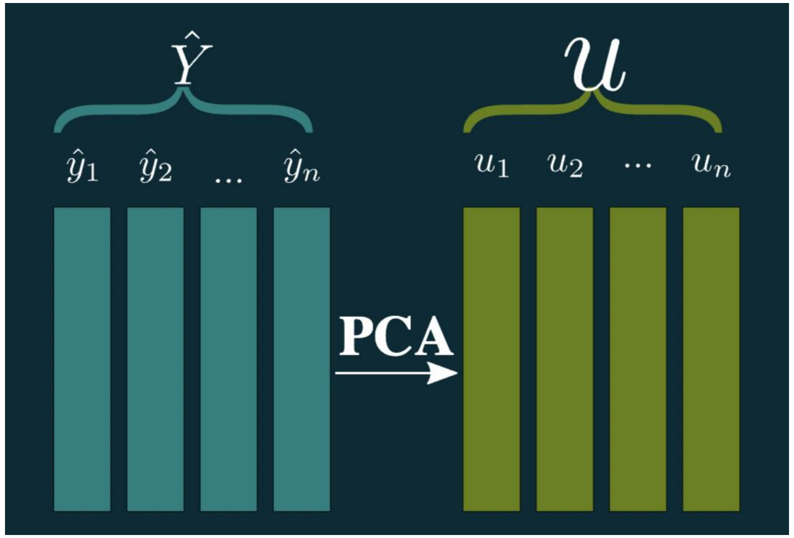
GRAPHICAL VIEW

Ŷ is produced using multiple linear regression between X and each yi.



GRAPHICAL VIEW

A PCA is performed on Ŷ which gives a set of principal component vectors U.

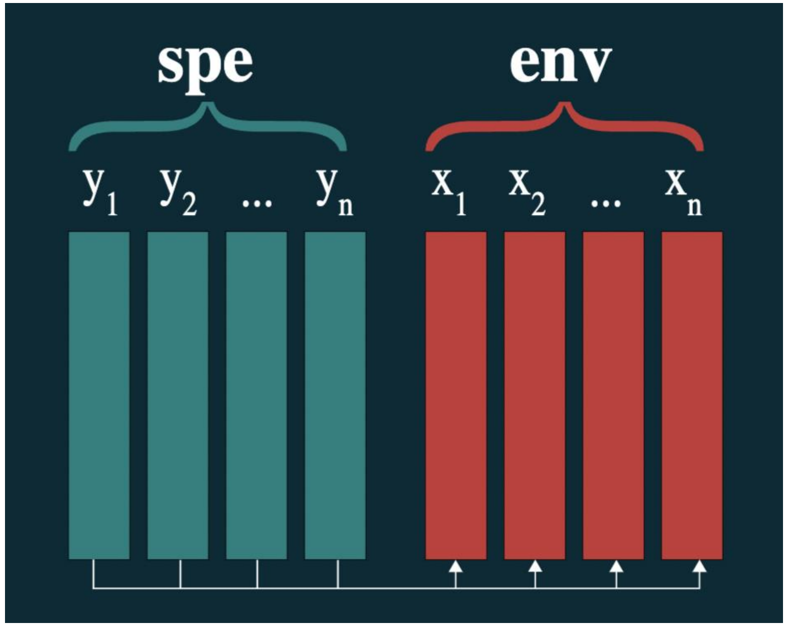


PCA VS RDA

PCA and RDA are very similar:

- PCA is performed on a matrix with explanatory variables.
- RDA is performed on a matrix of **predicted** explanatory variables.

Example in class:
Predicting species abundance in relation to environmental variables



The data come from (Verneaux, 1973).Data have been collected at 30 localities along Doubs river.

The first matrix (**Y**) contains coded abundances of 27 fish species.
The second matrix (**X**) contains 11 environmental variables related to the hydrology, geomorphology and chemistry of the river.

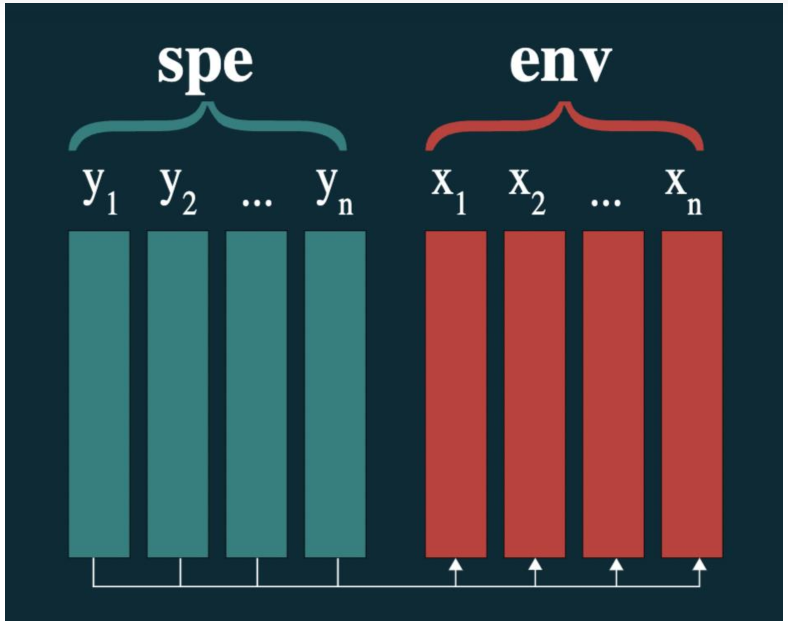
Reference: Verneaux, J. (1973) Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie. Thèse d'état, Besançon. 1–257.

REDUNDANCY ANALYSIS

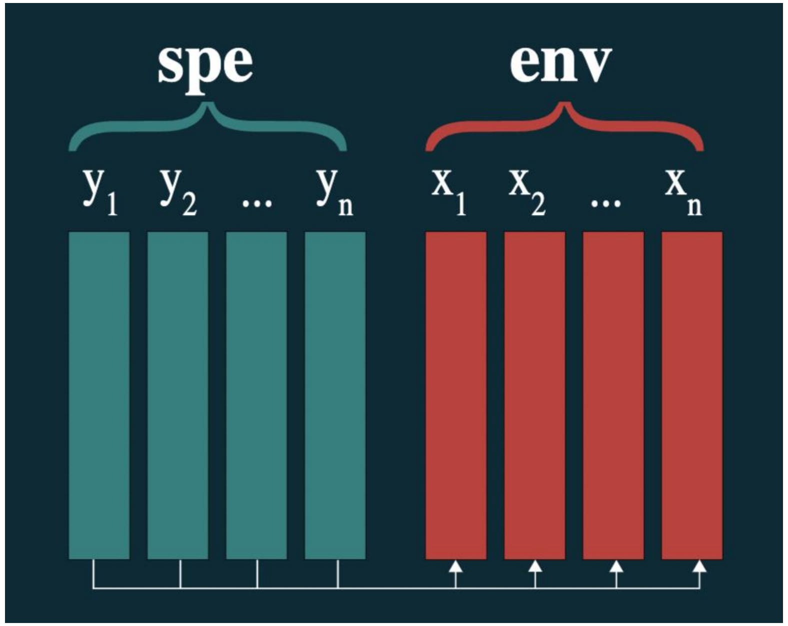
Get the data

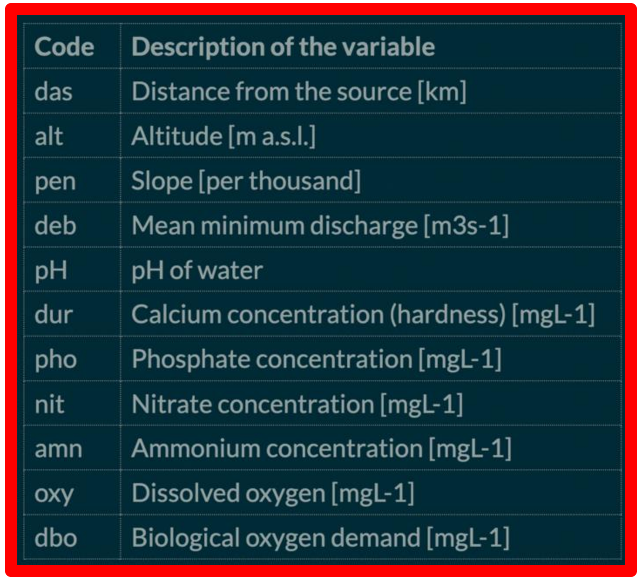
```r
spe <- read.csv("http://goo.gl/hwxKAD", row.names = 1)
```

```r
env <- read.csv("http://goo.gl/HDAWtO", row.names = 1)
```



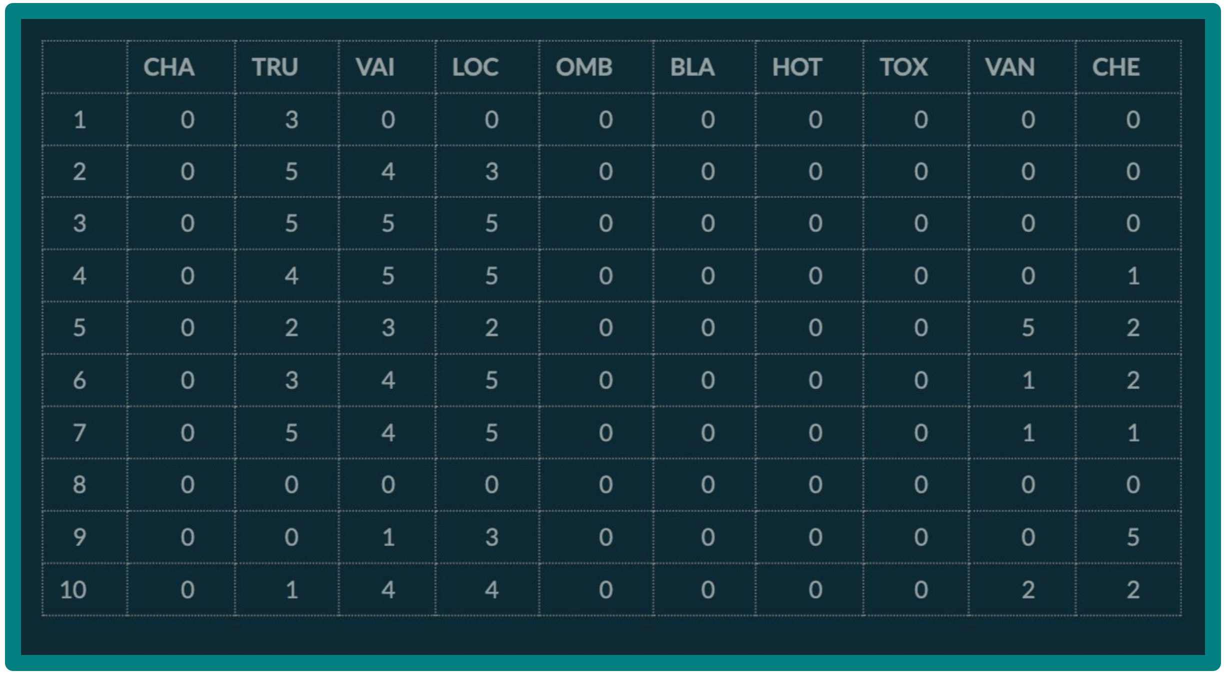
Example in class: Predicting species abundance in relation to environmental variables





Code	Description of the variable
das	Distance from the source [km]
alt	Altitude [m a.s.l.]
pen	Slope [per thousand]
deb	Mean minimum discharge [m3s-1]
pH	pH of water
dur	Calcium concentration (hardness) [mgL-1]
pho	Phosphate concentration [mgL-1]
nit	Nitrate concentration [mgL-1]
amn	Ammonium concentration [mgL-1]
oxy	Dissolved oxygen [mgL-1]
dbo	Biological oxygen demand [mgL-1]

Matrix **Y** fish species abundance.



	CHA	TRU	VAI	LOC	OMB	BLA	HOT	TOX	VAN	CHE
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	0	0	0	0	0	0	0	0
9	0	0	1	3	0	0	0	0	0	5
10	0	1	4	4	0	0	0	0	2	2

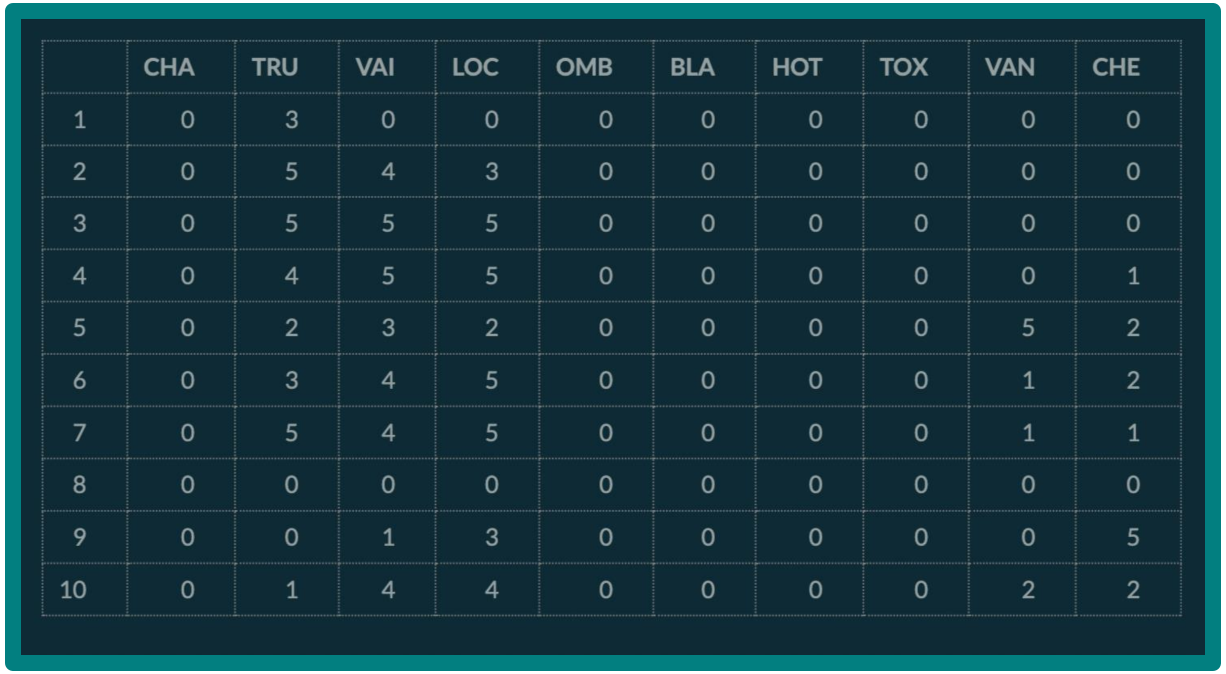
Environmental data (**X**)

	das	alt	pen	deb	pH	dur	pho	nit	amm	oxy	dbo
1	0.30	934	48.00	0.84	7.90	45	0.01	0.20	0.00	12.20	2.70
2	2.20	932	3.00	1.00	8.00	40	0.02	0.20	0.10	10.30	1.90
3	10.20	914	3.70	1.80	8.30	52	0.05	0.22	0.05	10.50	3.50
4	18.50	854	3.20	2.53	8.00	72	0.10	0.21	0.00	11.00	1.30
5	21.50	849	2.30	2.64	8.10	84	0.38	0.52	0.20	8.00	6.20
6	32.40	846	3.20	2.86	7.90	60	0.20	0.15	0.00	10.20	5.30
7	36.80	841	6.60	4.00	8.10	88	0.07	0.15	0.00	11.10	2.20
8	49.10	792	2.50	1.30	8.10	94	0.20	0.41	0.12	7.00	8.10
9	70.50	752	1.20	4.80	8.00	90	0.30	0.82	0.12	7.20	5.20
10	99.00	617	9.90	10.00	7.70	82	0.06	0.75	0.01	10.00	4.30

We will transform abundance data (**Y**) using the **Hellinger-transform** method (Borcard, Gillet, and Legendre 2011; Legendre and Gallagher 2001).

Particularly suited to species abundance data, this transformation gives low weights to variables with low counts and many zeros. The transformation itself comprises dividing each value in a data matrix by its row sum, and taking the square root of the quotient.

Matrix **Y** fish species abundance.

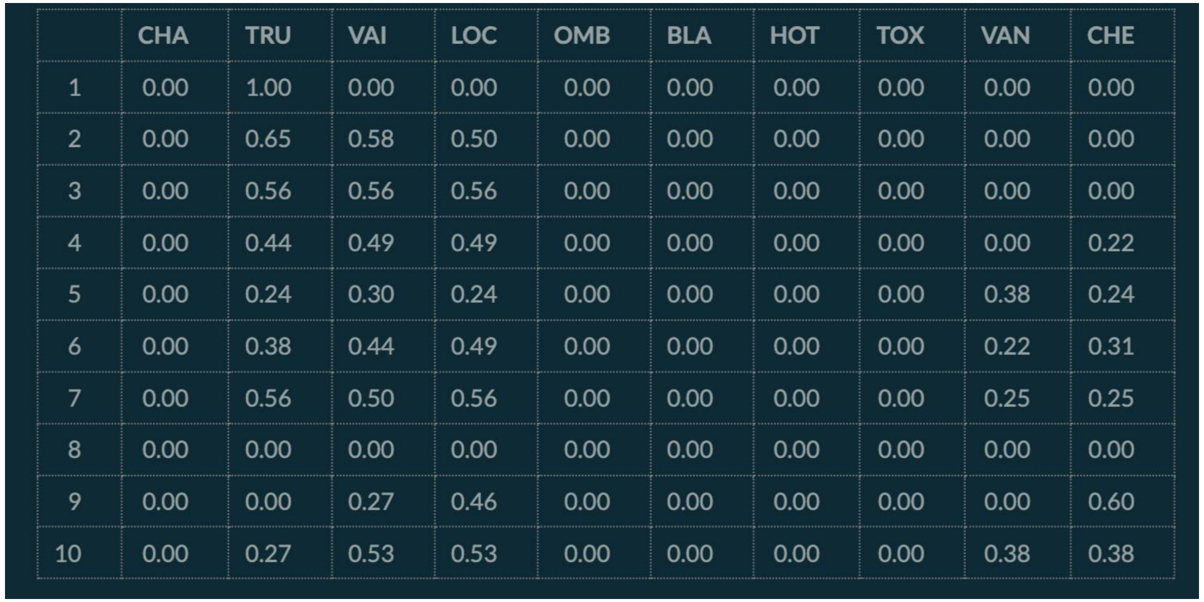


	CHA	TRU	VAI	LOC	OMB	BLA	HOT	TOX	VAN	CHE
1	0	3	0	0	0	0	0	0	0	0
2	0	5	4	3	0	0	0	0	0	0
3	0	5	5	5	0	0	0	0	0	0
4	0	4	5	5	0	0	0	0	0	1
5	0	2	3	2	0	0	0	0	5	2
6	0	3	4	5	0	0	0	0	1	2
7	0	5	4	5	0	0	0	0	1	1
8	0	0	0	0	0	0	0	0	0	0
9	0	0	1	3	0	0	0	0	0	5
10	0	1	4	4	0	0	0	0	2	2

The square root of relative abundance data (if rows are samples)

$$y'_{ij} = \sqrt{\frac{y_{ij}}{y_{i.}}}$$

In R, the **Hellinger-transform** is performed using the decostand() function from vegan package.



	CHA	TRU	VAI	LOC	OMB	BLA	HOT	TOX	VAN	CHE
1	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.65	0.58	0.50	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.56	0.56	0.56	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.44	0.49	0.49	0.00	0.00	0.00	0.00	0.00	0.22
5	0.00	0.24	0.30	0.24	0.00	0.00	0.00	0.00	0.38	0.24
6	0.00	0.38	0.44	0.49	0.00	0.00	0.00	0.00	0.22	0.31
7	0.00	0.56	0.50	0.56	0.00	0.00	0.00	0.00	0.25	0.25
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.27	0.46	0.00	0.00	0.00	0.00	0.00	0.60
10	0.00	0.27	0.53	0.53	0.00	0.00	0.00	0.00	0.38	0.38

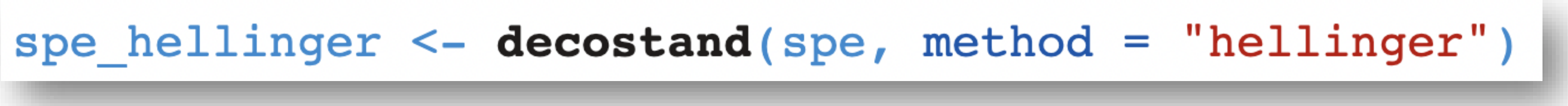
REDUNDANCY ANALYSIS

Get the data

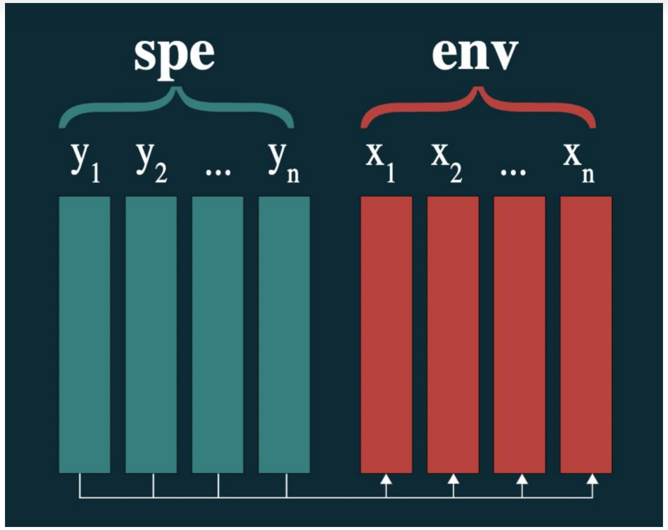
```
spe <- read.csv("http://goo.gl/hwxKAD", row.names = 1)
```

```
env <- read.csv("http://goo.gl/HDAWtO", row.names = 1)
```

Transform *spe*



```
spe_hellinger <- decostand(spe, method = "hellinger")
```



REDUNDANCY ANALYSIS

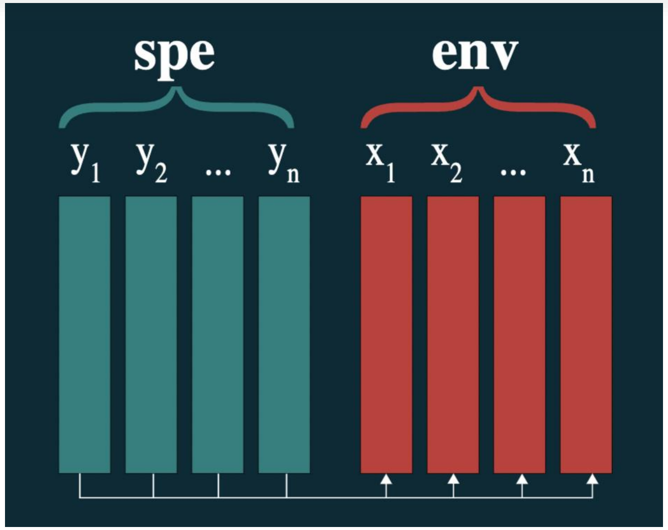
Get the data

```
spe <- read.csv("http://goo.gl/hwxKAD", row.names = 1)
```

```
env <- read.csv("http://goo.gl/HDAWtO", row.names = 1)
```

Transform *spe*

```
spe_hellinger <- decostand(spe, method = "hellinger")
```



Run RDA!

(with *vegan* package)

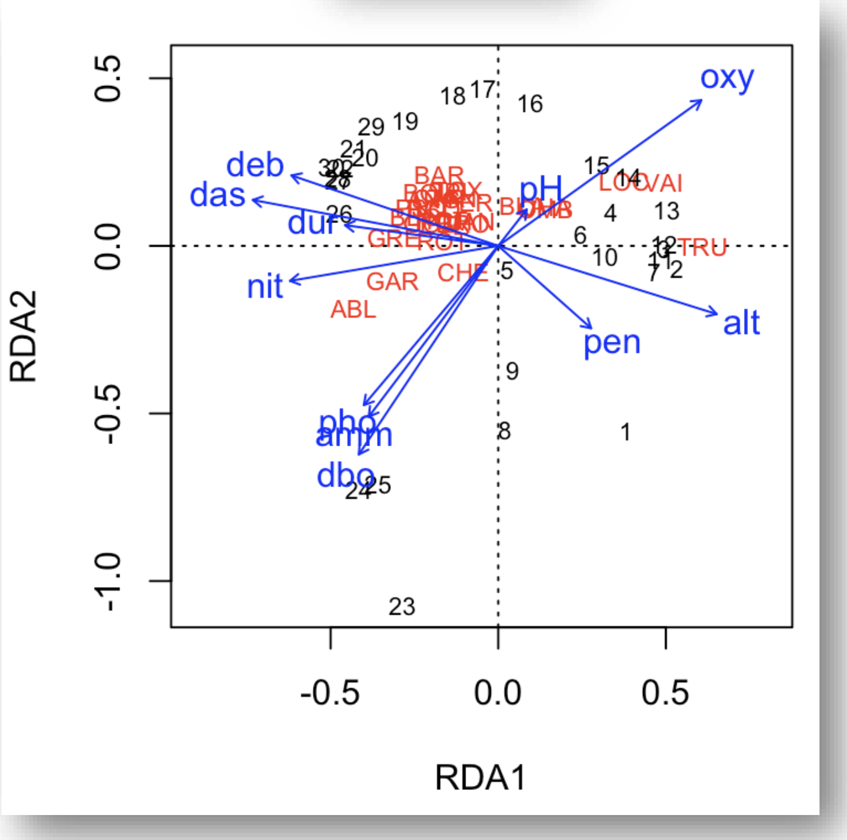
```
# Create the model
my_rda <- rda(spe_hellinger ~ das + alt + pen + deb + pH + dur +
    pho + nit + amm + oxy + dbo, data = env)
```

VISUALISATION

For RDA, the visualization plot is called **triplot** since there are three different entities in the plot: *sites*, *response variables* and *explanatory variables* (Borcard, Gillet, and Legendre 2011).

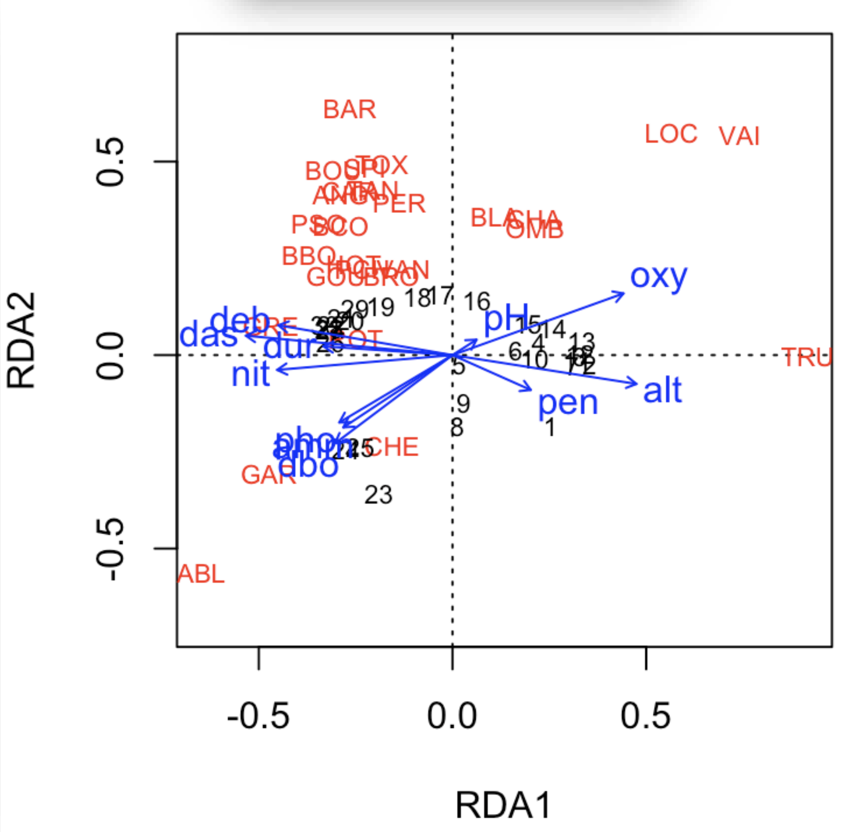
Can plot **without** scaling

`plot(my_rda)`



Can plot **with** scaling

`plot(my_rda, scaling=1)`



RDA: EXPLAINED VARIANCE

Attention: R^2 as the relative contribution of each eigenvectors are unadjusted and are therefore biased. For a proper computation of unbiased, adjusted R^2 one should use the RsquareAdj() function.

```
RsquareAdj(my_rda)  # Total variance explained by the RDA
```

RDA: EXPLAINED VARIANCE

Attention: R^2 as the relative contribution of each eigenvectors are unadjusted and are therefore biased. For a proper computation of unbiased, adjusted R^2 one should use the RsquareAdj() function.

```
RsquareAdj(my_rda)  # Total variance explained by the RDA
```

TESTS OF RDA RESULTS

Most of the time, ecological data are not normally-distributed: **parametric tests are not appropriate**.
For canonical analysis such as RDA, model significance relies on **permutation tests**.

RDA: EXPLAINED VARIANCE

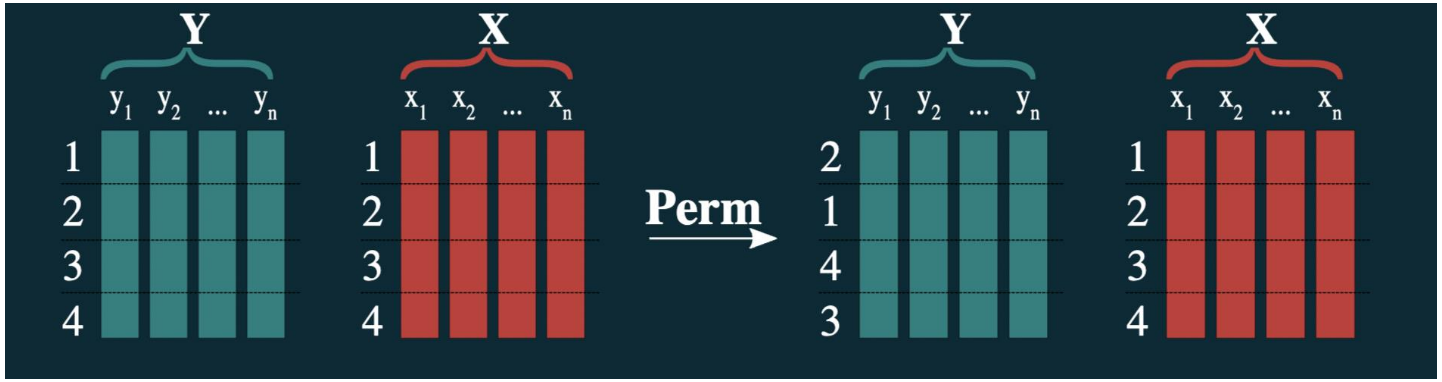
Attention: R^2 as the relative contribution of each eigenvectors are unadjusted and are therefore biased. For a proper computation of unbiased, adjusted R^2 one should use the RsquareAdj() function.

```
RsquareAdj(my_rda)  # Total variance explained by the RDA
```

TESTS OF RDA RESULTS

Most of the time, ecological data are not normally-distributed: **parametric tests are not appropriate**.
For canonical analysis such as RDA, model significance relies on **permutation tests**.

The principle of a permutation test is to generate a reference distribution of the chosen statistic under the null hypothesis H_0 by randomly permuting appropriate elements of the data many times and recomputing the statistic each time. Then, one compares the true value of the statistic to this reference distribution. The p value is computed as the proportion of the permuted values equal to or larger than the true (unpermuted) value of the statistic for a one-tailed test in the upper tail, like the F test used in RDA (Borcard, Gillet, and Legendre 2011).



H_0: Observed results can be produced by random chance.
The null hypothesis is rejected if this p value is equal to or smaller than the predefined significance level α (ex.: $\alpha=0.05$).

PERMUTATION TESTS

For an RDA, you can test for three different things:

1. Global RDA significance (*How much do the explanatory variables explain the RDA axes?*)
2. Axis significance (*How much do the RDA axes explain species?*)
3. Terms (explanatory variables) significance (*Which variables explain the RDAs?*)

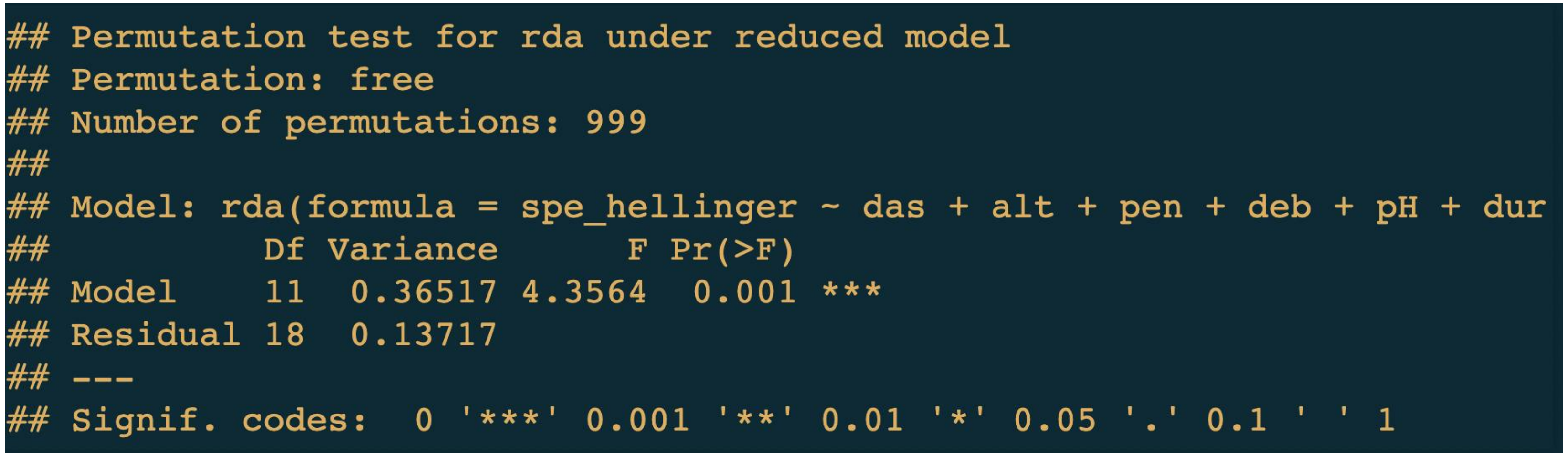
PERMUTATION TESTS

For an RDA, you can test for three different things:

1. Global RDA significance (*How much do the explanatory variables explain the RDA axes?*)
2. Axis significance (*How much do the RDA axes explain species?*)
3. Terms (explanatory variables) significance (*Which variables explain the RDAs?*)

GLOBAL SIGNIFICANCE

We use the function anova.cca() to perform the permutation test. **Do not get confused with the name**, it is different than the classical ANOVA test.



```
## Permutation test for rda under reduced model
## Permutation: free
## Number of permutations: 999
##
## Model: rda(formula = spe_hellinger ~ das + alt + pen + deb + pH + dur
##          Df Variance      F Pr(>F)
## Model    11  0.36517 4.3564  0.001 ***
## Residual 18  0.13717
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since $p = 0.001$ we reject H_0 meaning that the RDA model is significant.

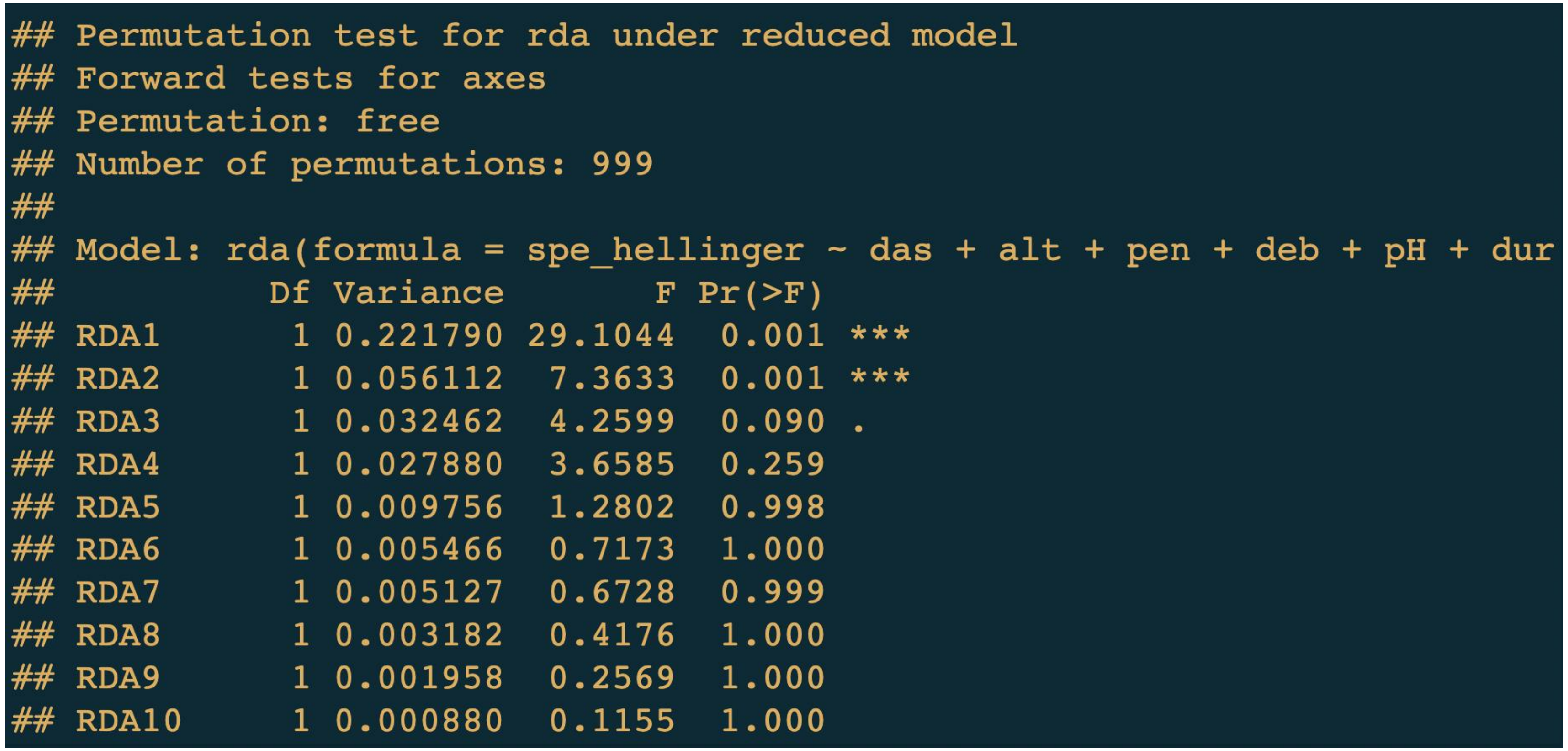
PERMUTATION TESTS

For an RDA, you can test for three different things:

1. Global RDA significance (*How much do the explanatory variables explain the RDA axes?*)
2. Axis significance (*How much do the RDA axes explain species?*)
3. Terms (explanatory variables) significance (*Which variables explain the RDAs?*)

AXIS SIGNIFICANCE

Axis significance can be done by adding by = "axis".



```
## Permutation test for rda under reduced model
## Forward tests for axes
## Permutation: free
## Number of permutations: 999
##
## Model: rda(formula = spe_hellinger ~ das + alt + pen + deb + pH + dur
##          Df Variance       F Pr(>F)
## RDA1      1 0.221790 29.1044  0.001 ***
## RDA2      1 0.056112  7.3633  0.001 ***
## RDA3      1 0.032462  4.2599  0.090 .
## RDA4      1 0.027880  3.6585  0.259
## RDA5      1 0.009756  1.2802  0.998
## RDA6      1 0.005466  0.7173  1.000
## RDA7      1 0.005127  0.6728  0.999
## RDA8      1 0.003182  0.4176  1.000
## RDA9      1 0.001958  0.2569  1.000
## RDA10     1 0.000880  0.1155  1.000
```

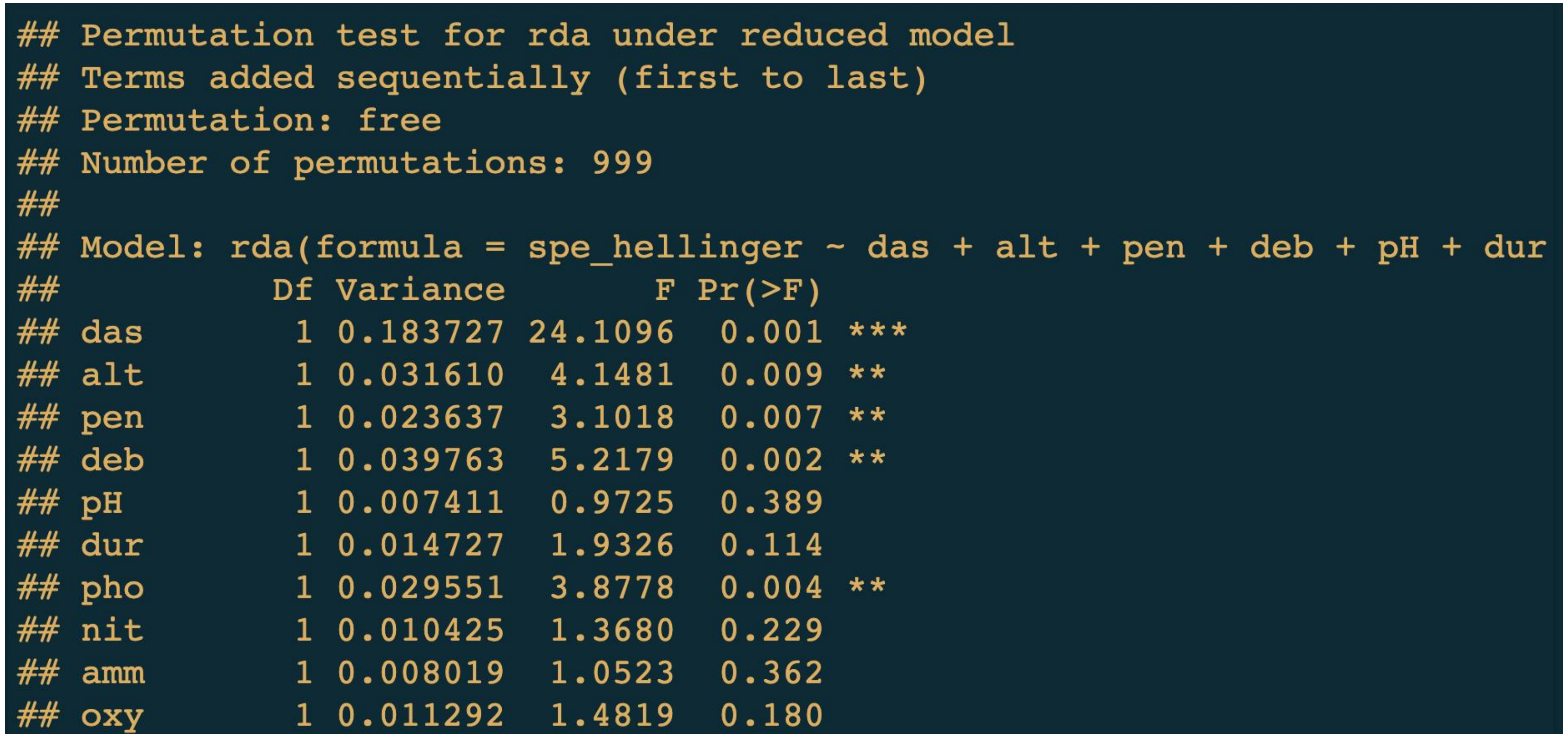
PERMUTATION TESTS

For an RDA, you can test for three different things:

1. Global RDA significance (*How much do the explanatory variables explain the RDA axes?*)
2. Axis significance (*How much do the RDA axes explain species?*)
3. Terms (explanatory variables) significance (*Which variables explain the RDAs?*)

TERMS SIGNIFICANCE

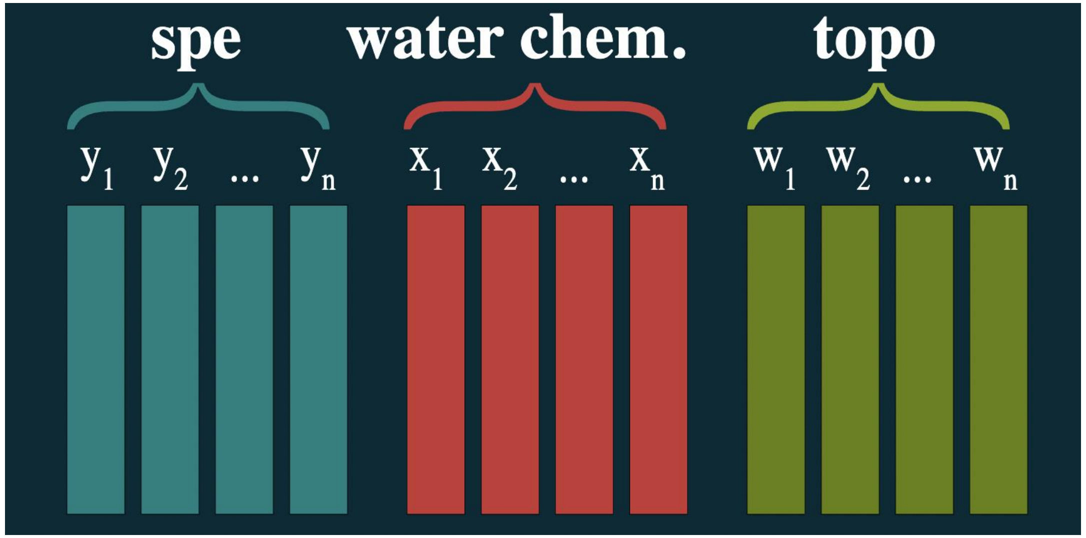
Terms significance can be done by adding by = "terms".



```
## Permutation test for rda under reduced model
## Terms added sequentially (first to last)
## Permutation: free
## Number of permutations: 999
##
## Model: rda(formula = spe_hellinger ~ das + alt + pen + deb + pH + dur
##          Df Variance       F Pr(>F)
## das       1 0.183727 24.1096  0.001 ***
## alt       1 0.031610  4.1481  0.009 **
## pen       1 0.023637  3.1018  0.007 **
## deb       1 0.039763  5.2179  0.002 **
## pH        1 0.007411  0.9725  0.389
## dur       1 0.014727  1.9326  0.114
## pho       1 0.029551  3.8778  0.004 **
## nit       1 0.010425  1.3680  0.229
## amm       1 0.008019  1.0523  0.362
## oxy       1 0.011292  1.4819  0.180
```

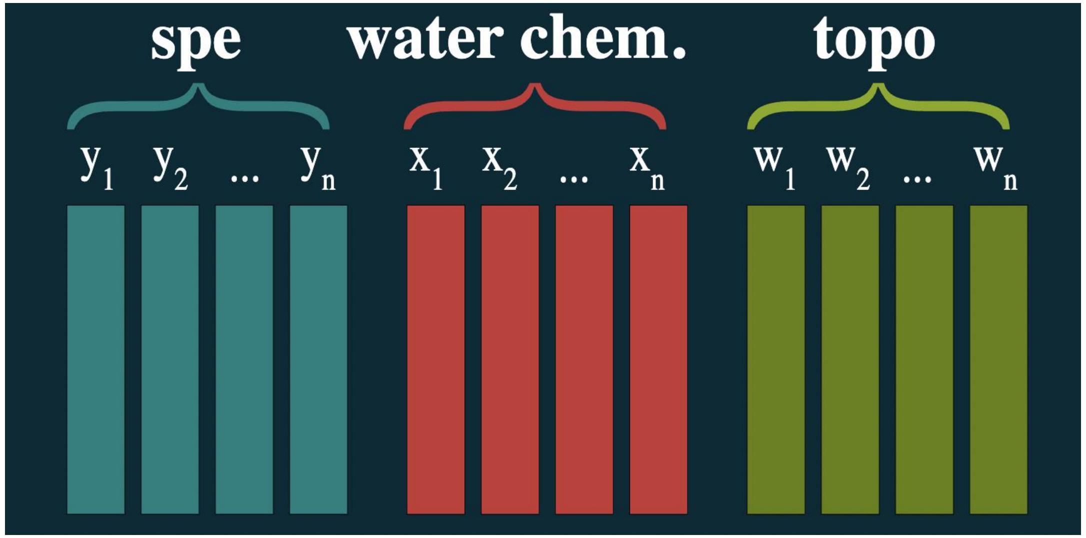
PARTIAL RDA

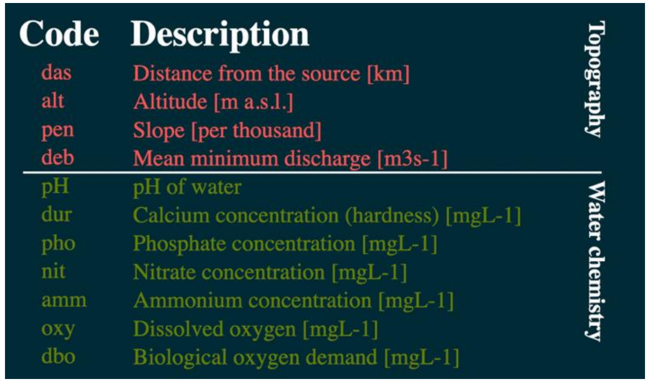
It is possible to run an RDA of a matrix **Y**, explained by a matrix variables **X**, in the presence of co-variable(s) **W**. This analysis allows to display the patterns of the response data (**Y**) uniquely explained by a linear model of the explanatory variables (**X**) when the effect of other covariates (**W**) is held constant (Borcard, Gillet, and Legendre 2011).



PARTIAL RDA

It is possible to run an RDA of a matrix **Y**, explained by a matrix variables **X**, in the presence of co-variable(s) **W**. This analysis allows to display the patterns of the response data (**Y**) uniquely explained by a linear model of the explanatory variables (**X**) when the effect of other covariates (**W**) is held constant (Borcard, Gillet, and Legendre 2011).

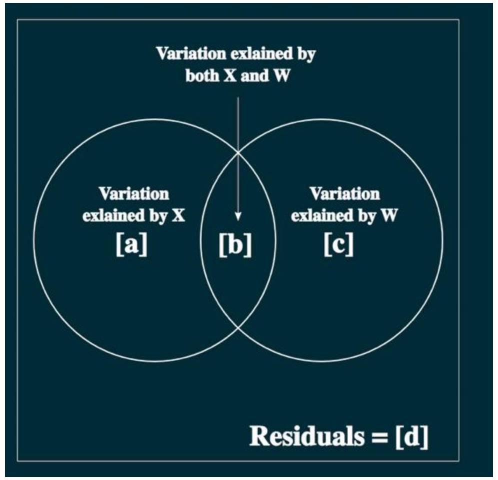




Code	Description	
das	Distance from the source [km]	Topography
alt	Altitude [m a.s.l.]	
pen	Slope [per thousand]	
deb	Mean minimum discharge [m3s-1]	
pH	pH of water	Water chemistry
dur	Calcium concentration (hardness) [mgL-1]	
pho	Phosphate concentration [mgL-1]	
nit	Nitrate concentration [mgL-1]	
amm	Ammonium concentration [mgL-1]	
oxy	Dissolved oxygen [mgL-1]	
dbo	Biological oxygen demand [mgL-1]	

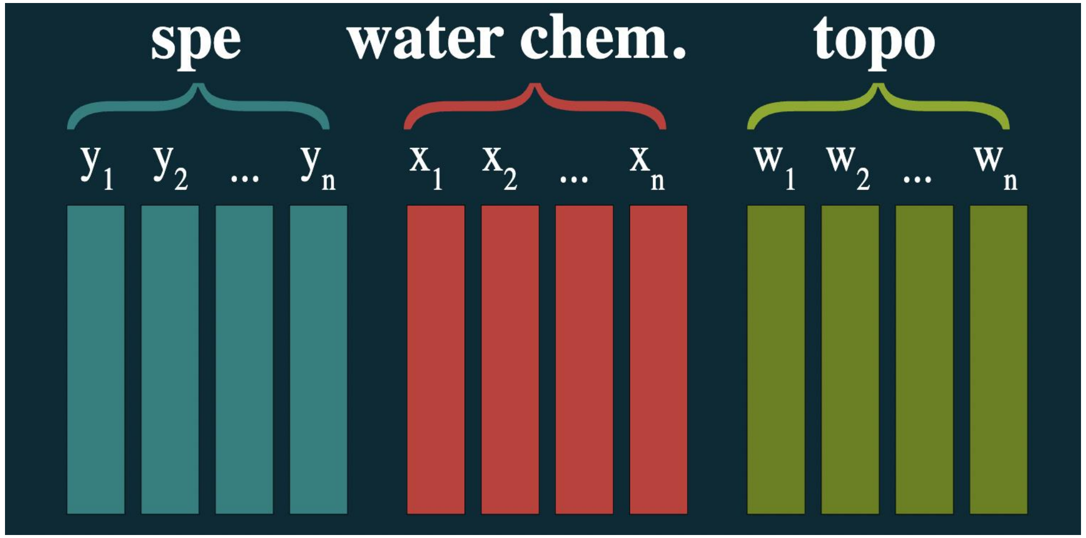
AN EXAMPLE

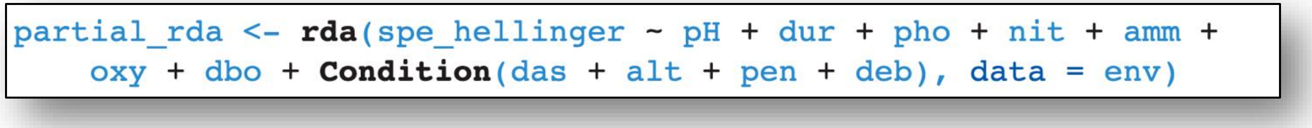
For example, we can test whether water chemistry (**X**) significantly explains the fish species patterns (**Y**) when the effect of the topographical gradient (**W**) is held constant. In other words, we want to model the effect of water chemistry variables on fish abundance once the effect of topographic variables is removed.



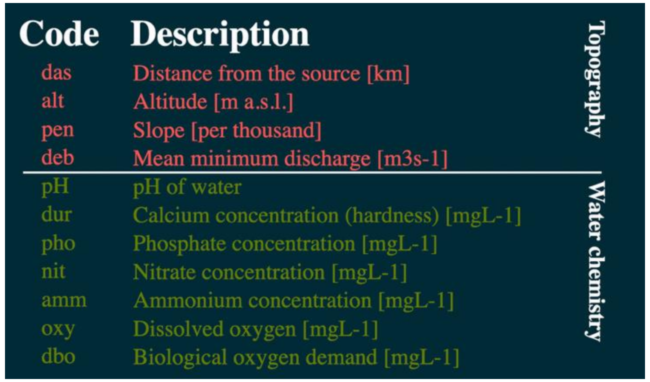
PARTIAL RDA

The construction of a partial RDA is pretty much the same as for a *standard* RDA except that we are using the Condition(…) function to specify which covariate (**W**) we want to remove the effect.





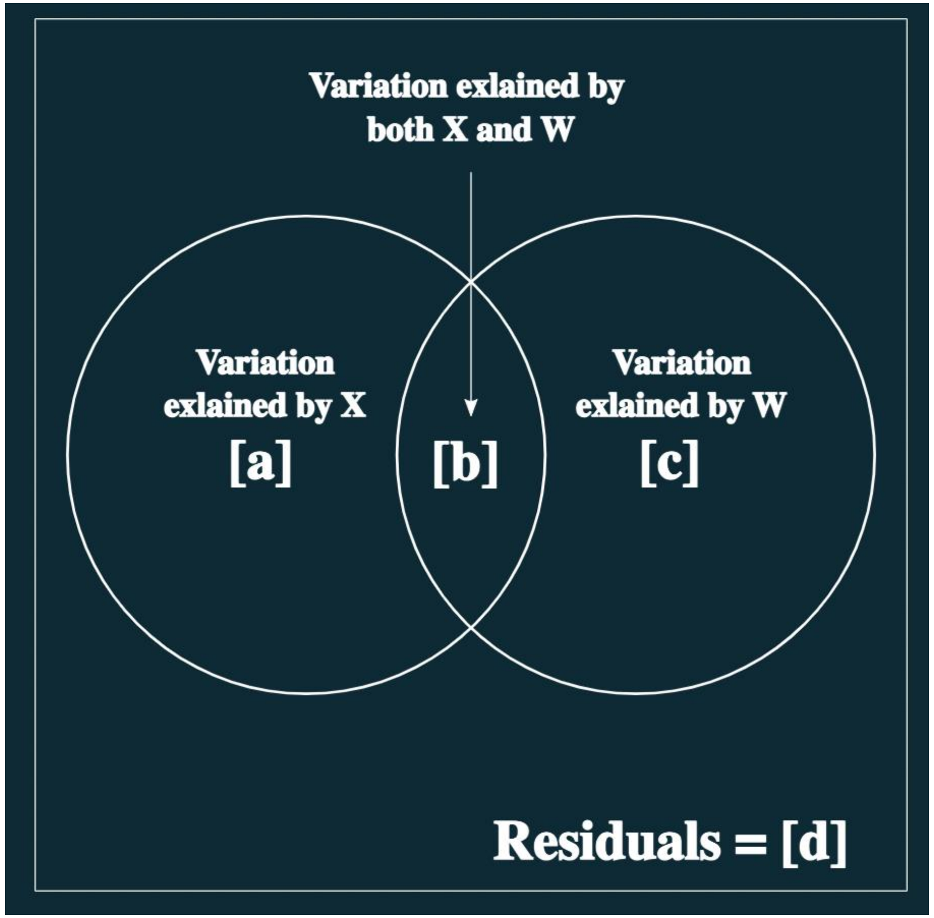
```
partial_rda <- rda(spe_hellinger ~ pH + dur + pho + nit + amm +
    oxy + dbo + Condition(das + alt + pen + deb), data = env)
```



Code	Description	
das	Distance from the source [km]	Topography
alt	Altitude [m a.s.l.]	
pen	Slope [per thousand]	
deb	Mean minimum discharge [m3s-1]	
pH	pH of water	Water chemistry
dur	Calcium concentration (hardness) [mgL-1]	
pho	Phosphate concentration [mgL-1]	
nit	Nitrate concentration [mgL-1]	
amm	Ammonium concentration [mgL-1]	
oxy	Dissolved oxygen [mgL-1]	
dbo	Biological oxygen demand [mgL-1]	

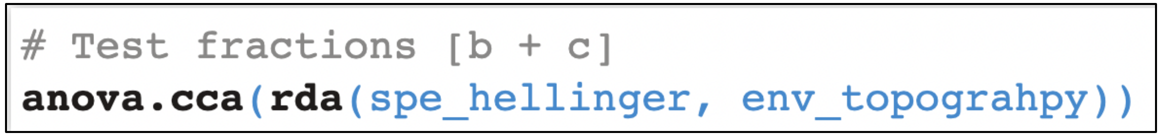
TEST OF SIGNIFICANCE

We have tested the global significance of the RDA. We also need to do it for each fraction: (1) [a + b], (2) [b + c], (3) [a + b + c], (4) [a] and (5) [c]



(RDA)

```
# Test fractions [a + b]
anova.cca(rda(spe_hellinger, env_water_chemistry))
```



```
# Test fractions [b + c]
anova.cca(rda(spe_hellinger, env_topograhpy))
```

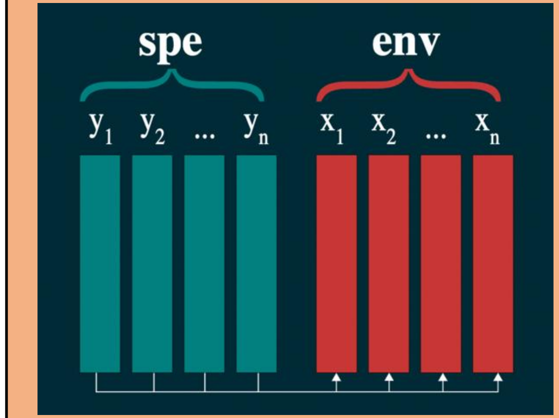
```
# Test fractions [a + b + c]
anova.cca(rda(spe_hellinger, env))
```

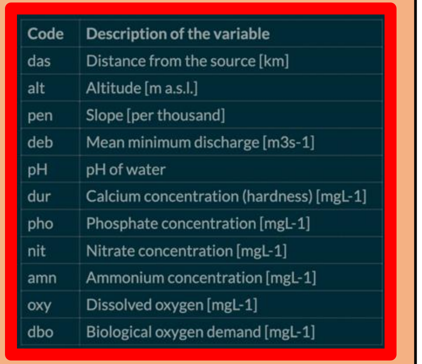
(Partial RDA)

```
# Test fraction [a]
anova.cca(rda(spe_hellinger, env_water_chemistry, env_topograhpy))
```

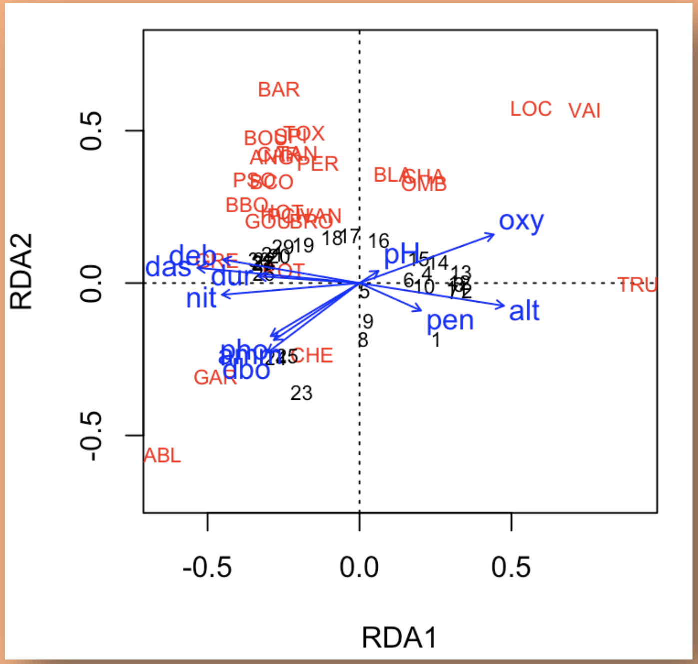
```
# Test fraction [c]
anova.cca(rda(spe_hellinger, env_topograhpy, env_water_chemistry))
```

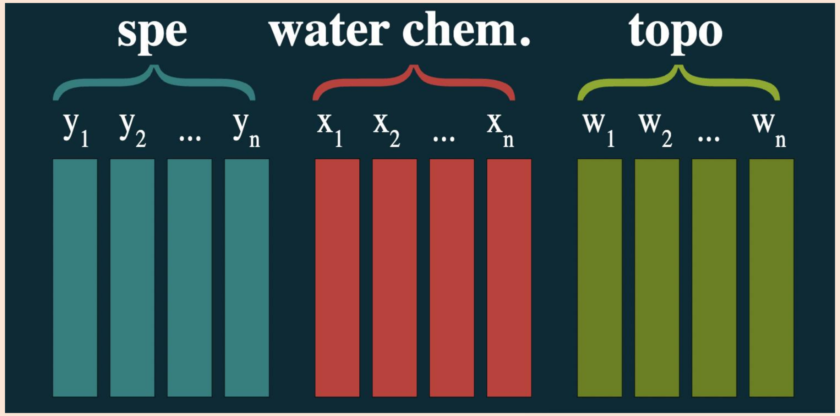
RDA VS PARTIAL RDA

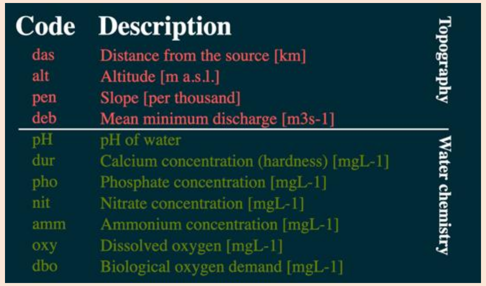




Code	Description of the variable
das	Distance from the source [km]
alt	Altitude [m a.s.l.]
pen	Slope [per thousand]
deb	Mean minimum discharge [m3s-1]
pH	pH of water
dur	Calcium concentration (hardness) [mgL-1]
pho	Phosphate concentration [mgL-1]
nit	Nitrate concentration [mgL-1]
amn	Ammonium concentration [mgL-1]
oxy	Dissolved oxygen [mgL-1]
dbo	Biological oxygen demand [mgL-1]

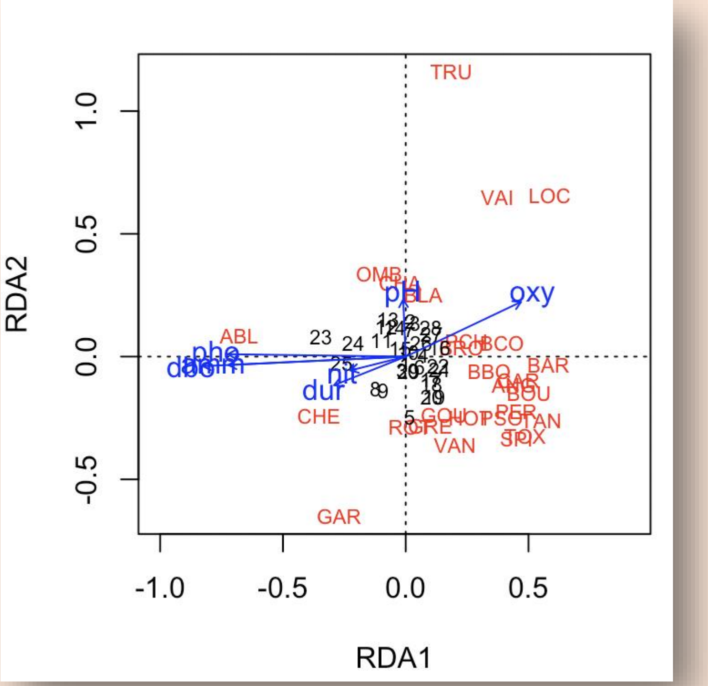






Code	Description	
das	Distance from the source [km]	Topography
alt	Altitude [m a.s.l.]	
pen	Slope [per thousand]	
deb	Mean minimum discharge [m3s-1]	
pH	pH of water	Water chemistry
dur	Calcium concentration (hardness) [mgL-1]	
pho	Phosphate concentration [mgL-1]	
nit	Nitrate concentration [mgL-1]	
amm	Ammonium concentration [mgL-1]	
oxy	Dissolved oxygen [mgL-1]	
dbo	Biological oxygen demand [mgL-1]	

Variables in topo (W, ie. Conditioned variables) are controlled for: not in the triplot



Further constrained ordination methods

Canonical Correspondence Analysis (CCA)

- Widely used
- Extension of (unconstrained) correspondence analysis
- Similar to RDA, but assumes unimodal distribution (χ^2-distance) of species along environmental gradient
- In R: model building as for RDA

`cca() {vegan}`

Constrained additive Ordination (CAO)

- Comparatively new
- derives response of each species to main environmental gradient from data → no linear or unimodal model assumed
- mixture of Generalized Additive Models (GAMs) and Canonical Gaussian Ordination
- computationally demanding
- In R: implemented in extra package

`cao() {VGAM}`