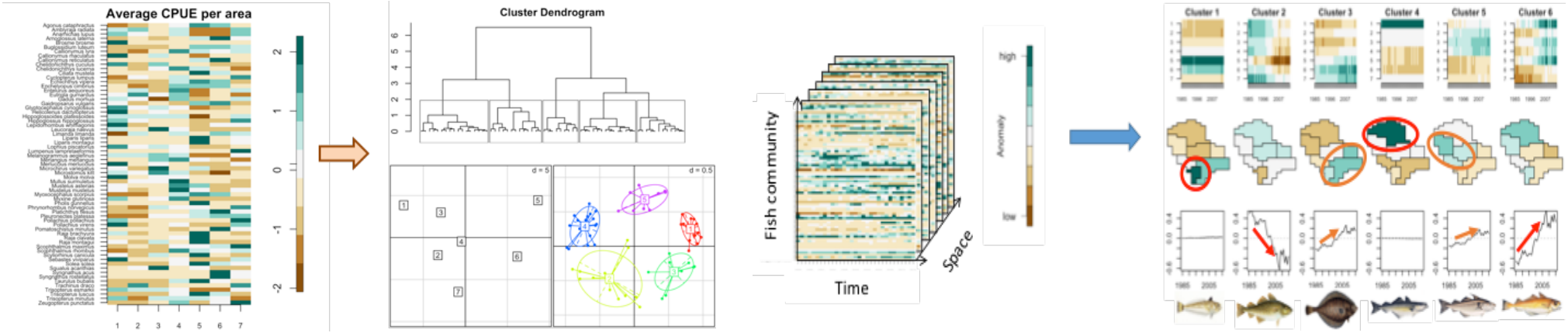
Universität Hamburg

MARmaED

Multivariate analysis: from 2D PCA to 3D PTA



zmt bremen

20th October 2016

Romain Frelat

Objectives

- Run a Principal component analysis (PCA) on a matrix (2D)
- Interpret the Principal Components (PC)
- Run a Principal tensor analysis (PTA) on a array (3D)
- Interpret the Principal Tensors (PT)
- Run clustering analysis with Hierarchical Clustering

- **Understand what is a multivariate analysis, and when can it be useful.**

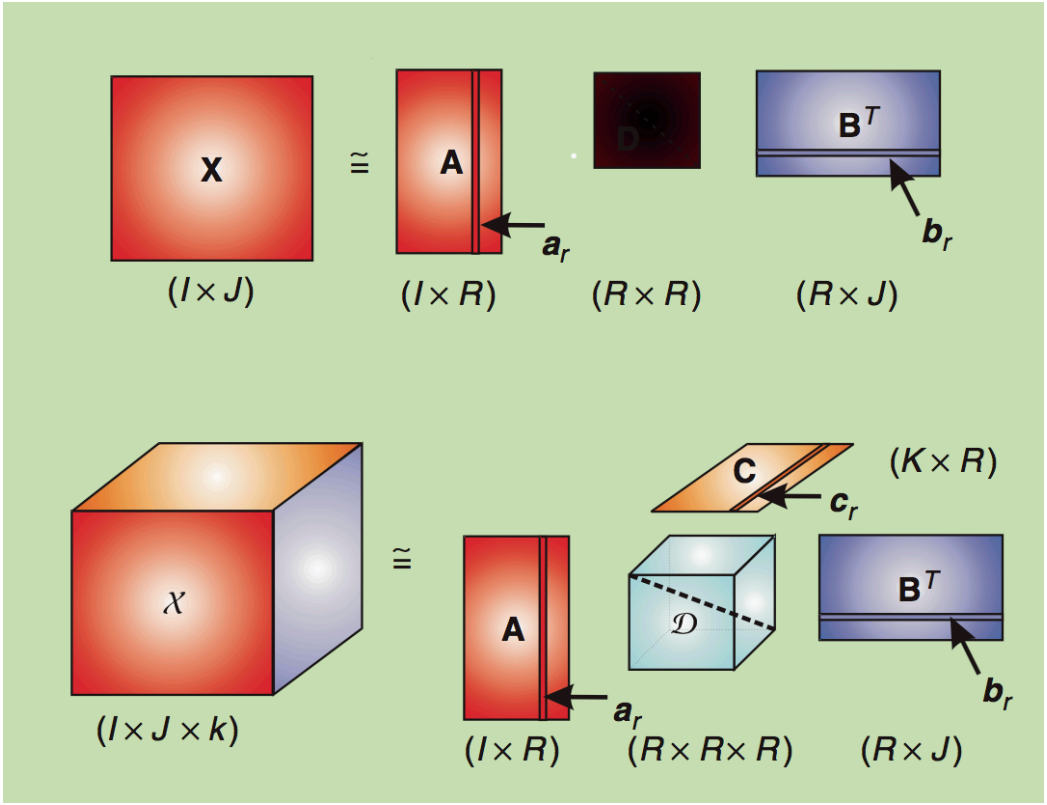
Multivariate analysis

- Simplification of the data - reduce the dimensionality of the data set
- Reveal the pattern ‘hidden’ in the data

- Trade-off between explain the maximum of the variability within the dataset in the minimum number of variables

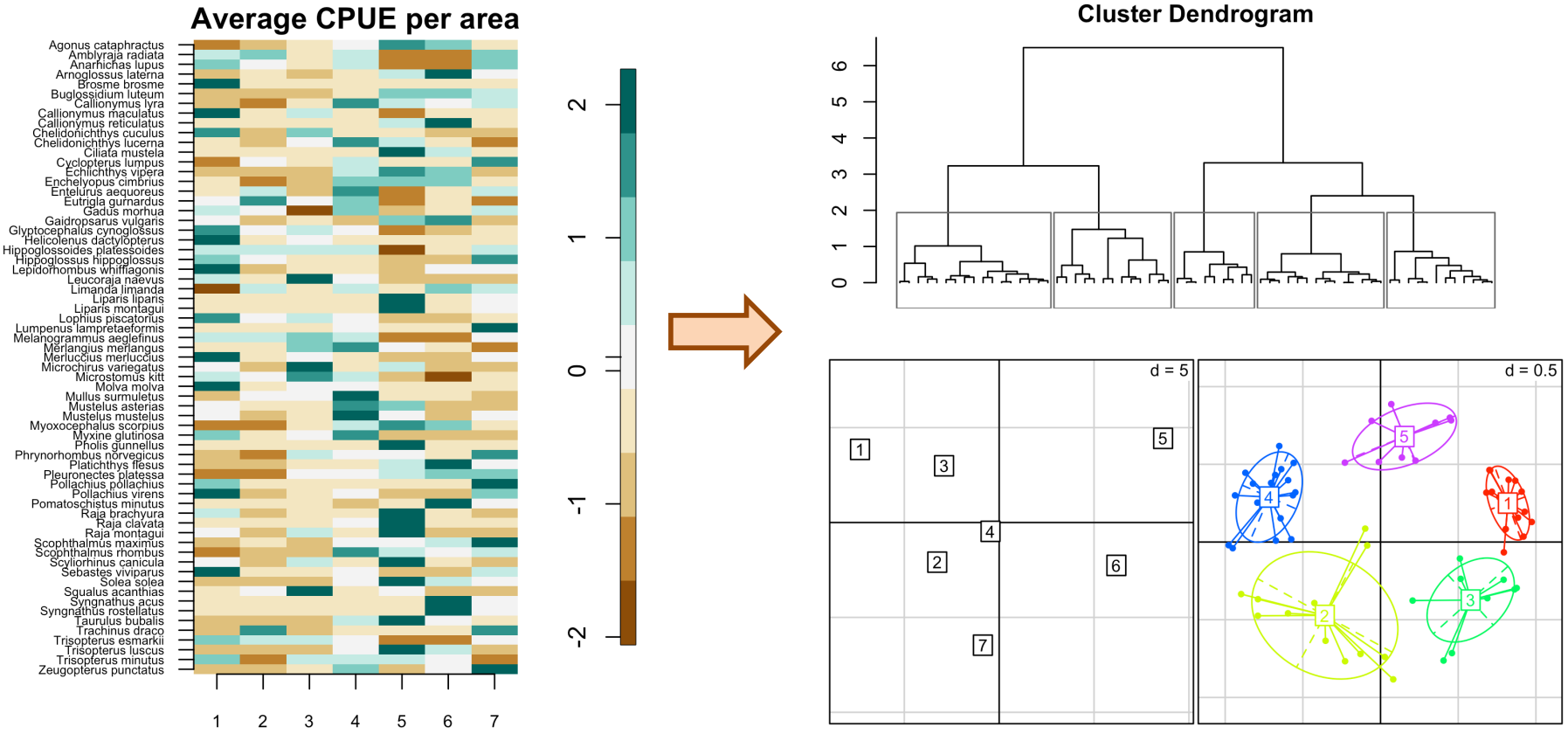
- Data mining : “let the data speaks”
  => no model, no prediction

Multivariate analysis



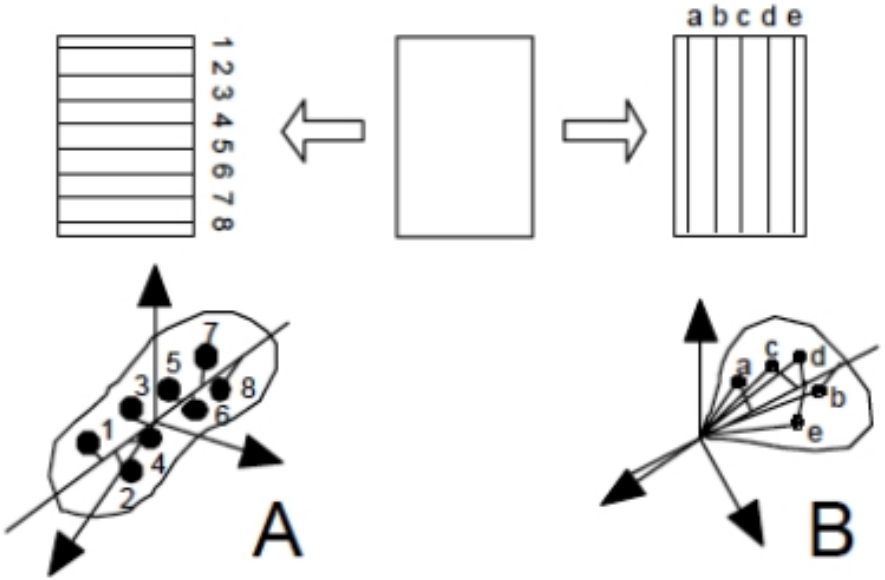
1. Select the variables and check the distribution
2. Scale the variables
3. Run the analysis + selection # components (scree test)
4. Interpret the selected components
5. (optional) Cluster analysis

2D – Principal Component Analysis



Principal component analysis

Projection of data on orthogonal axes (Principal Components) to maximized the "projected inertia", i.e. to separated as much as possible the data on this axes



Let X be a matrix with *p* variables on *n* individual.

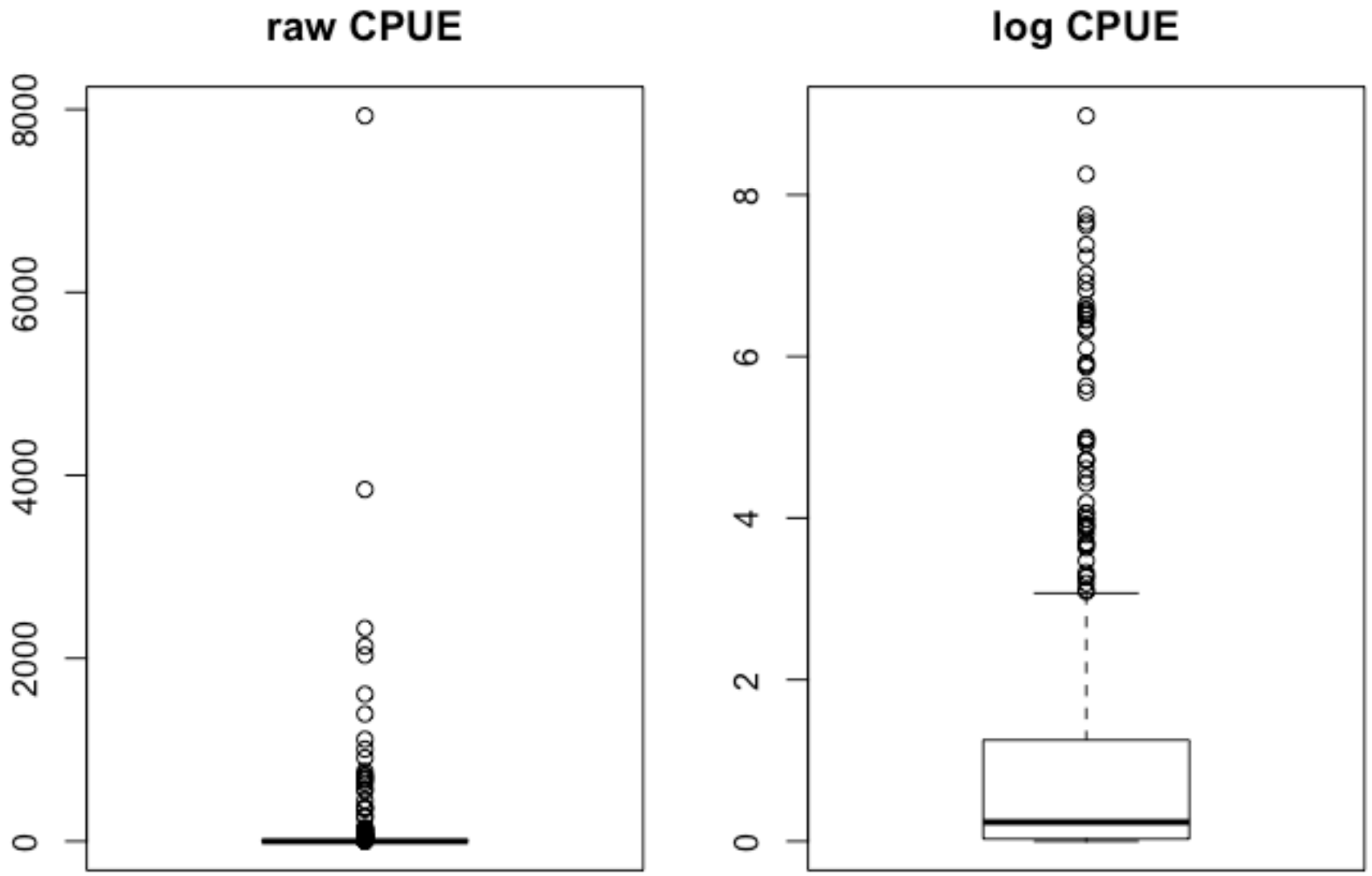
Principal axes are the eigen vectors of the variance-covariance matrix of X.

Source : A.B. Dufour

You can find a complete tutorial on PCA at : https://pbil.univ-lyon1.fr/R/pdf/course2.pdf

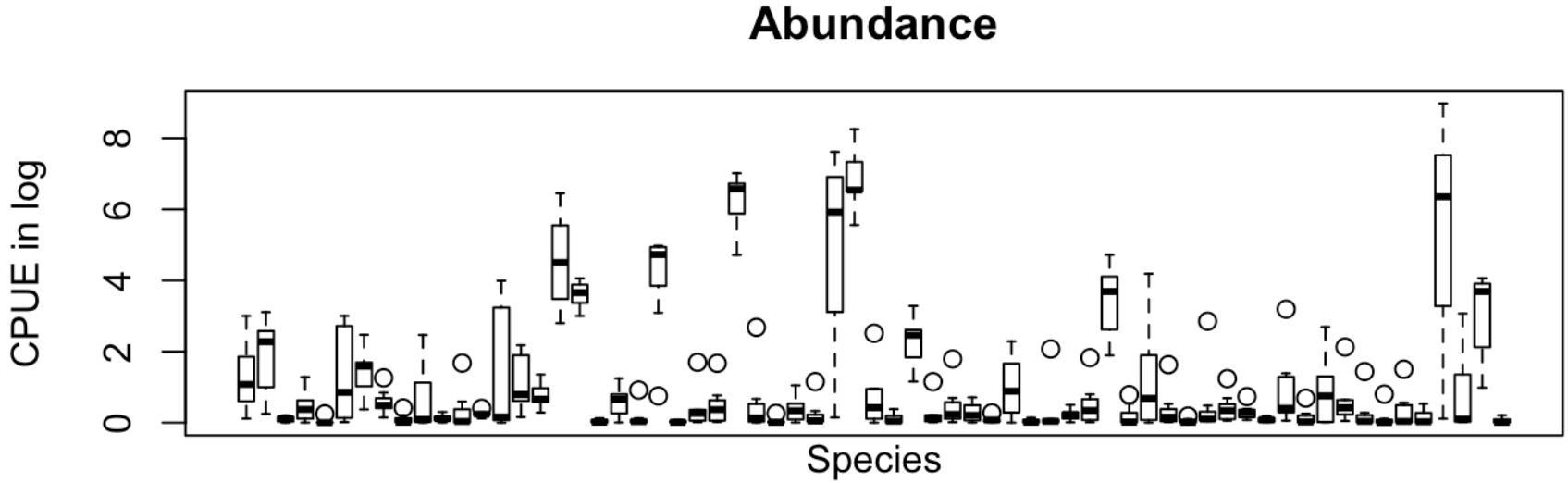
1. **Select variables and check the distribution**
2. Scale the variables
3. Run the analysis + selection # components (scree test)
4. Interpret the selected components
5. (optional) Cluster analysis

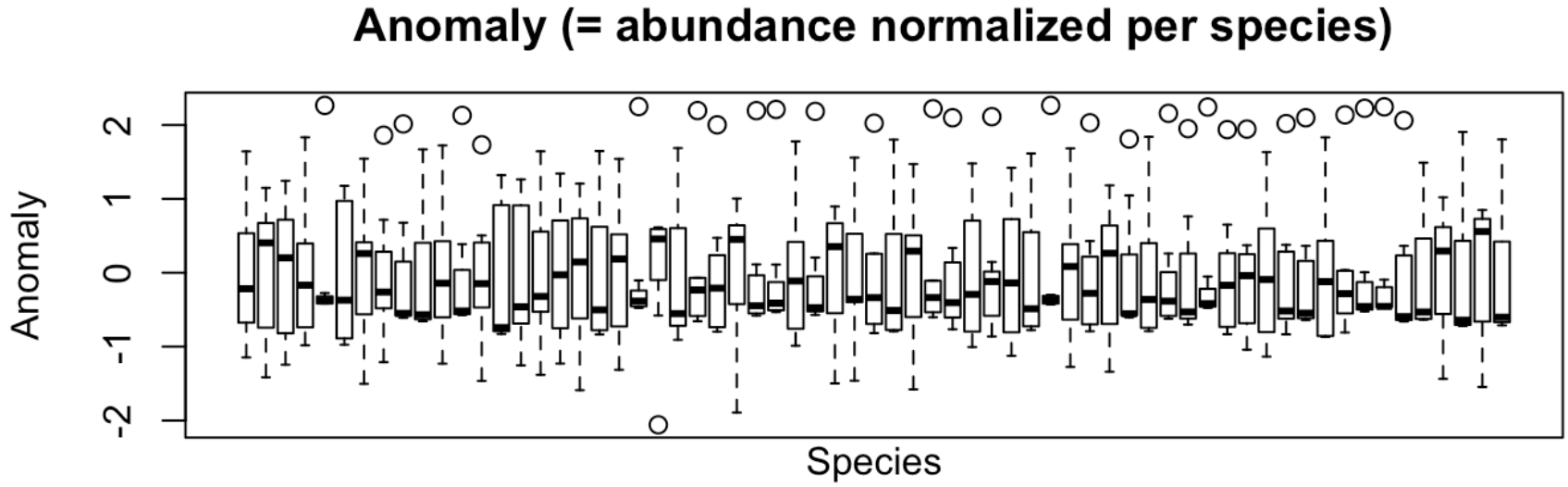
CPUE distribution



1. Select variables and check the distribution
2. **Scale the variables**
3. Run the analysis + selection # components (scree test)
4. Interpret the selected components
5. (optional) Cluster analysis

Variables are scaled

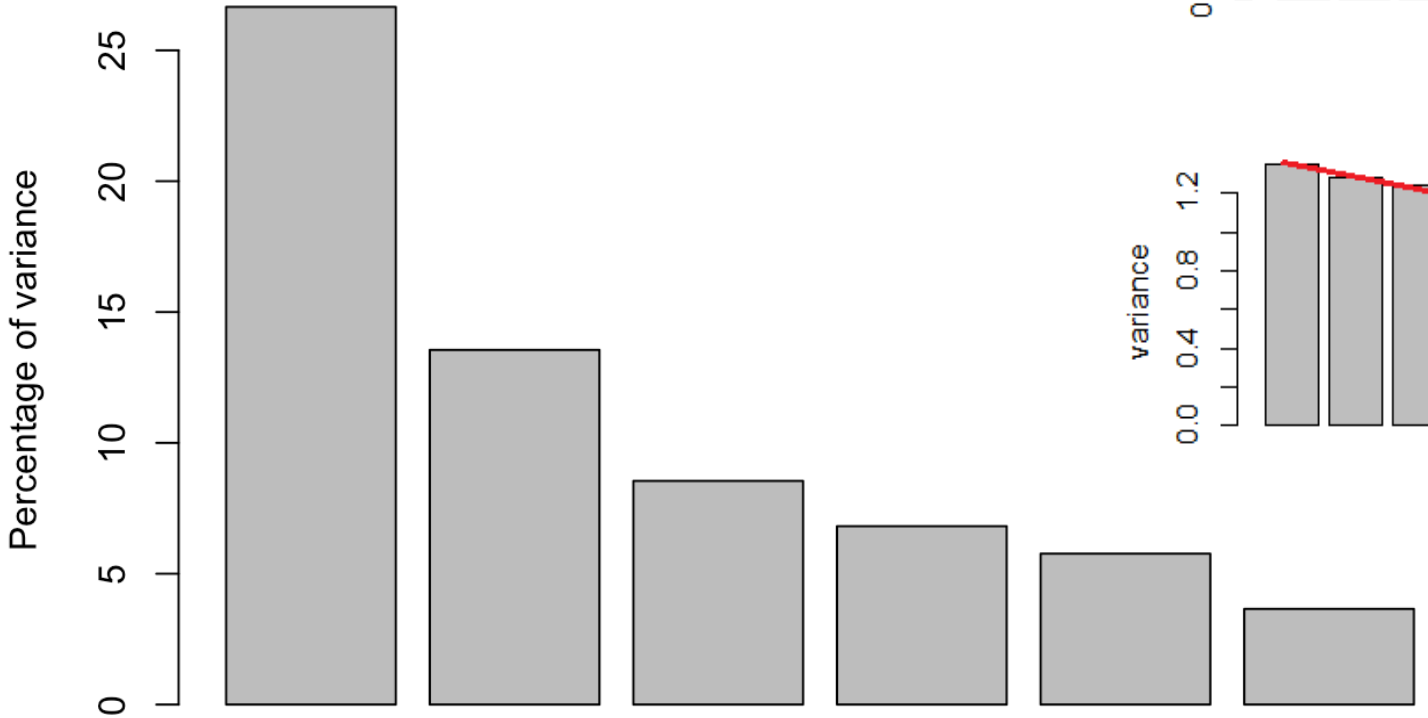


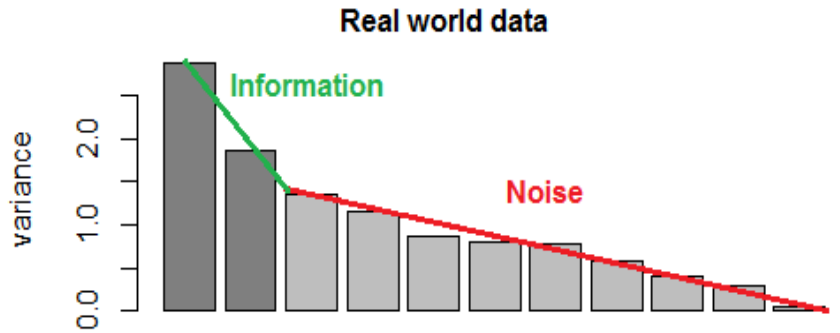


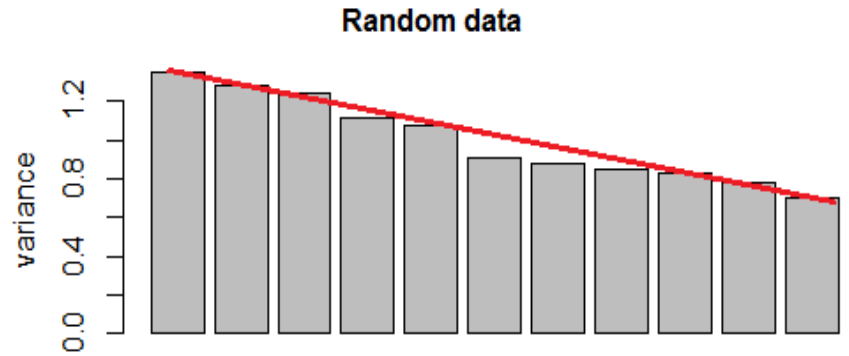
1. Select variables and check the distribution
2. Scale the variables
3. **Run the analysis + selection # components (scree test)**
4. Interpret the selected components
5. (optional) Cluster analysis

Eigen values

Select the number of axes:

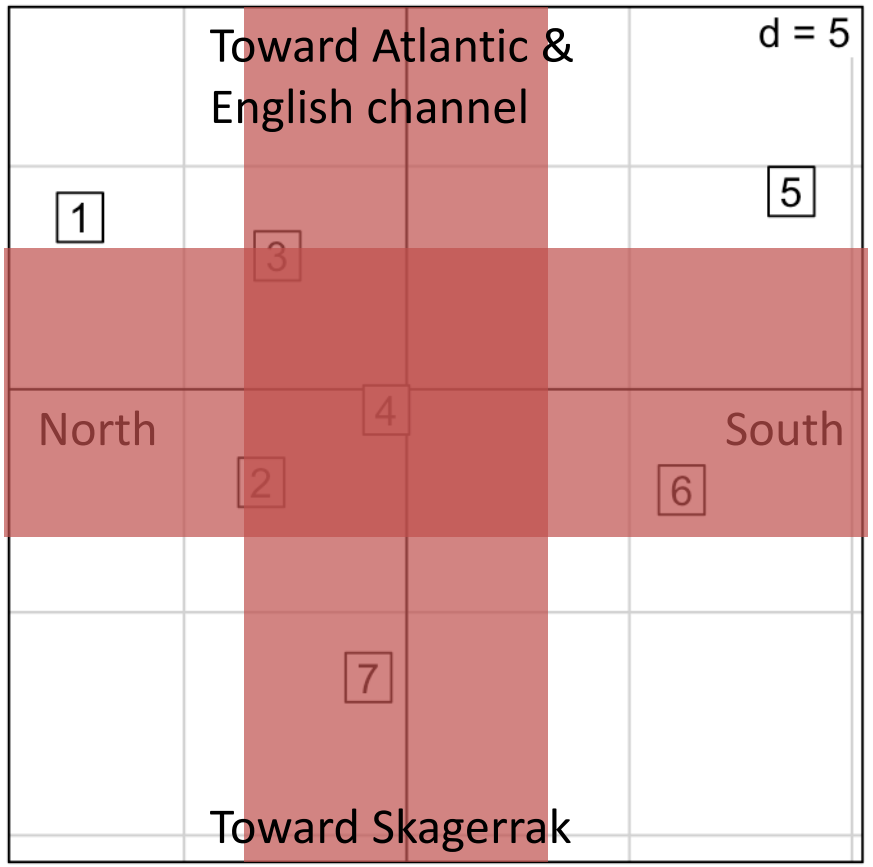


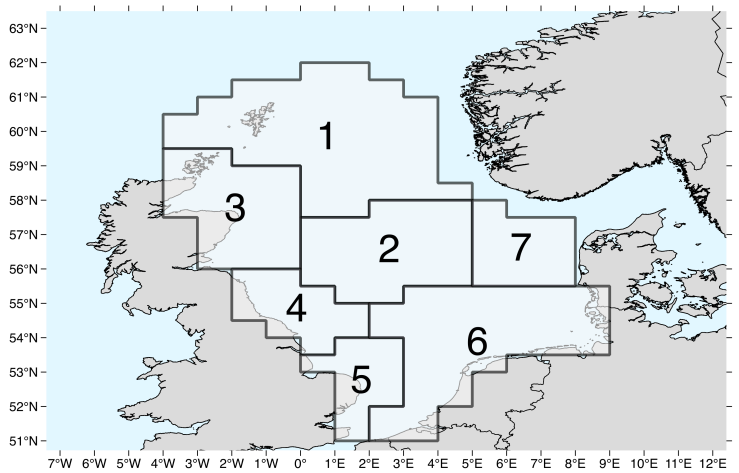




1. Select variables and check the distribution
2. Scale the variables
3. Run the analysis + selection # components (scree test)
4. **Interpret the selected components**
5. (optional) Cluster analysis

Interpret PC

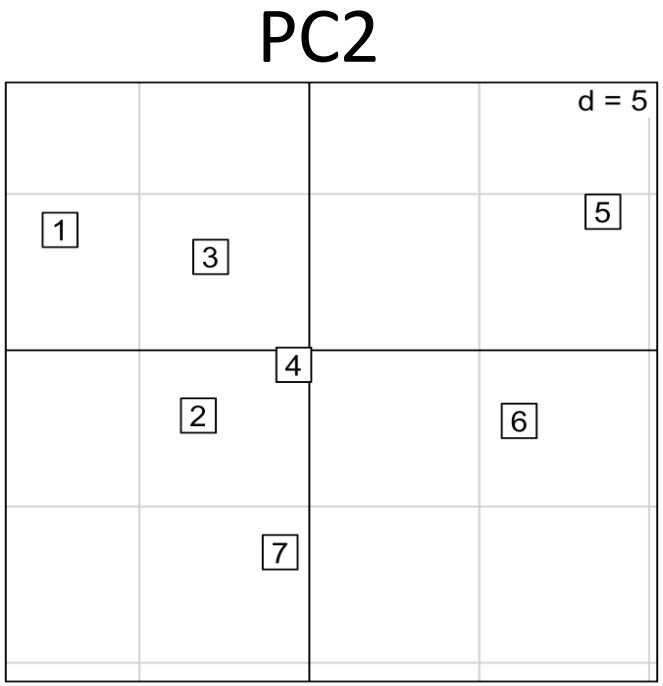


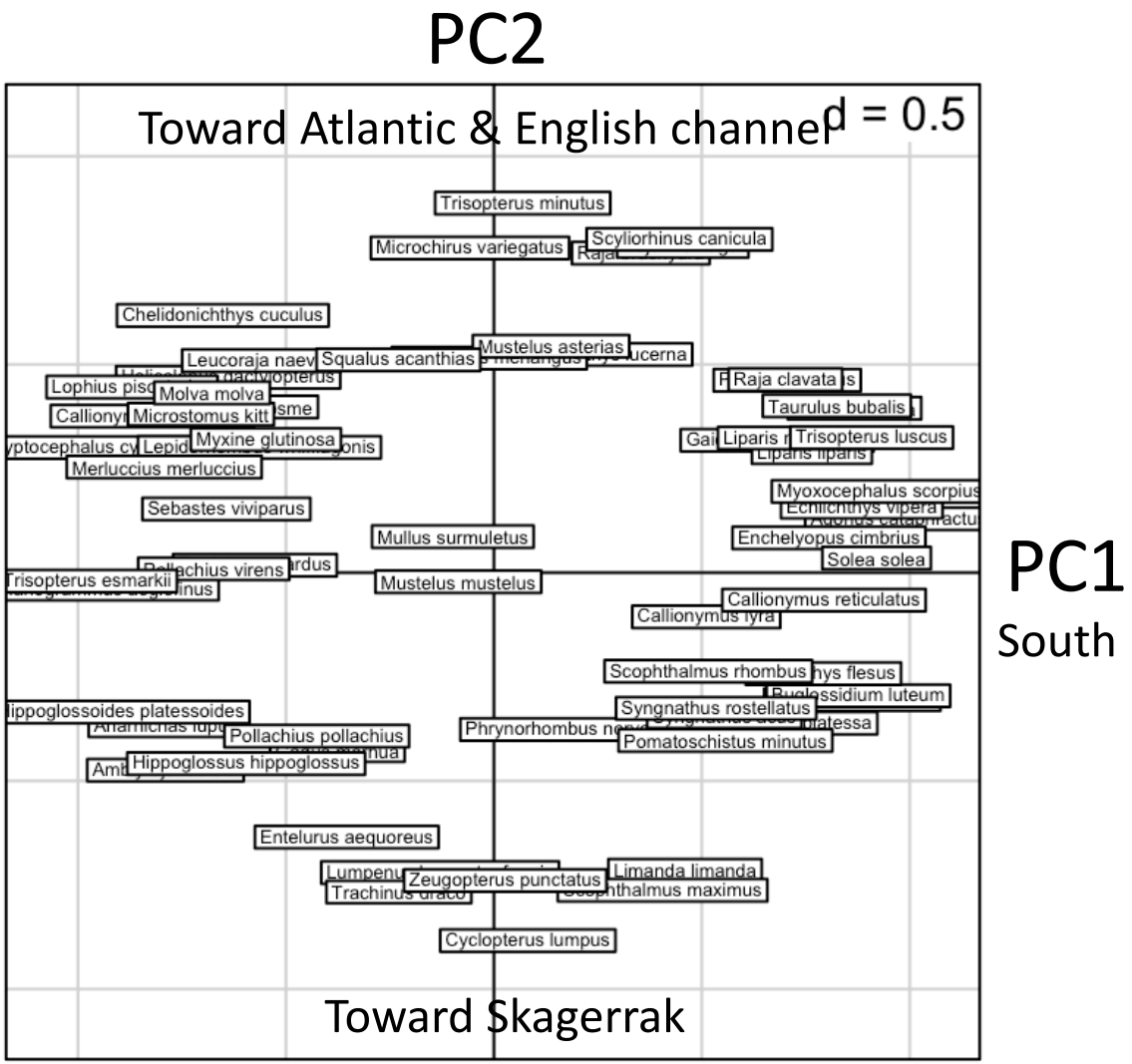


	Axis1	Axis2
1	**-7.3353689**	**3.8519039**
2	-3.2640990	-2.0883788
3	-2.9030102	2.9885736
4	-0.4560783	-0.4644314
5	**8.6404760**	**4.4337734**
6	**6.1761023**	-2.2590725
7	-0.8580218	**-6.4623682**

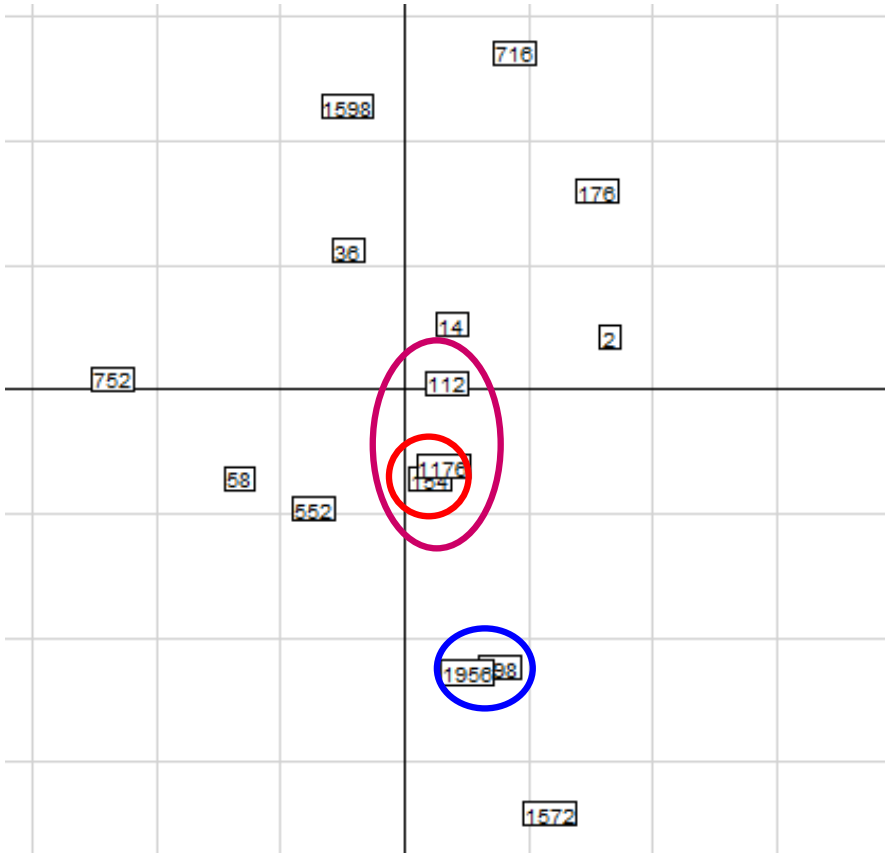
1. Select variables and check the distribution
2. Scale the variables
3. Run the analysis + selection # components (scree test)
4. **Interpret the selected components**
5. (optional) Cluster analysis

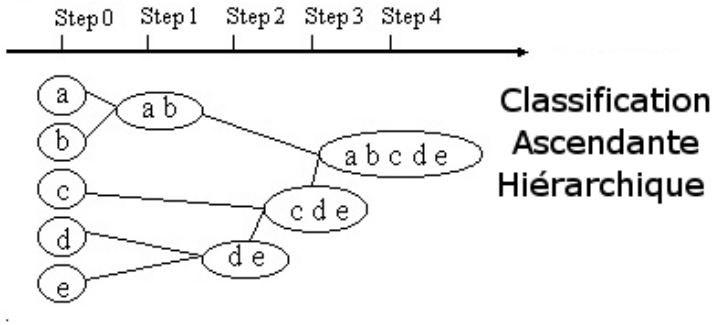
Interpret PC





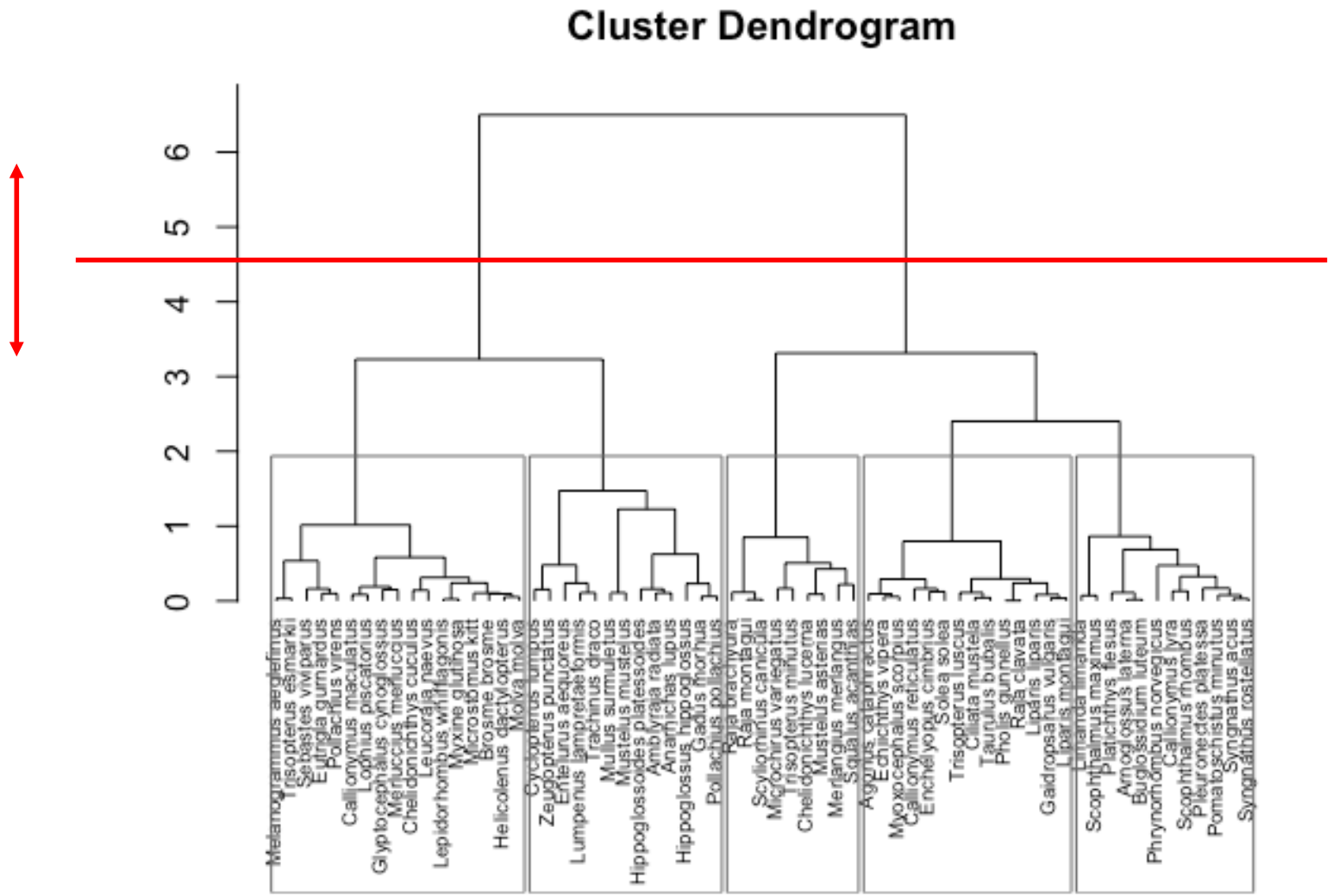
Ascendant Hierarchical Clustering





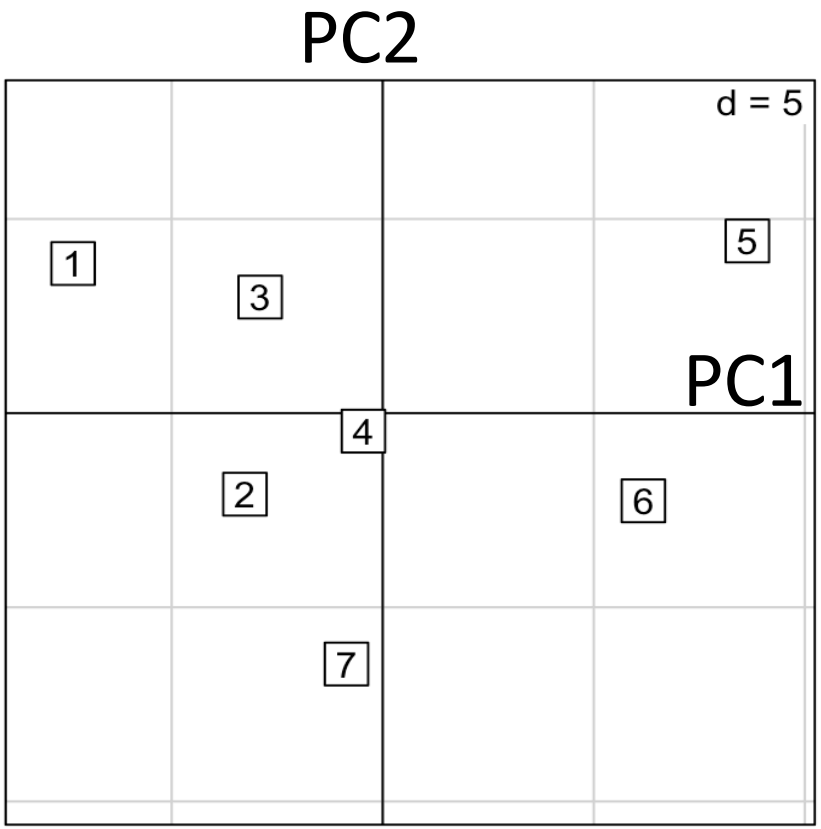
1. Select variables and check the distribution
2. Scale the variables
3. Run the analysis + selection # components (scree test)
4. Interpret the selected components
5. **(optional) Cluster analysis**

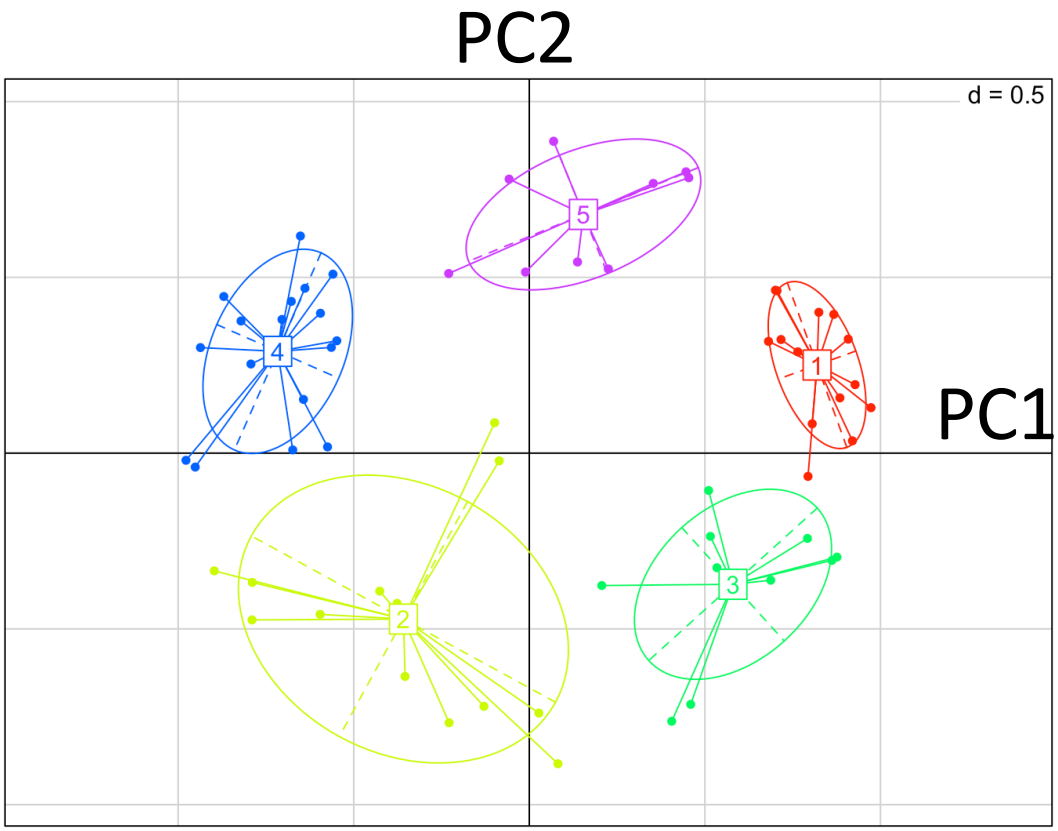
How many clusters ?



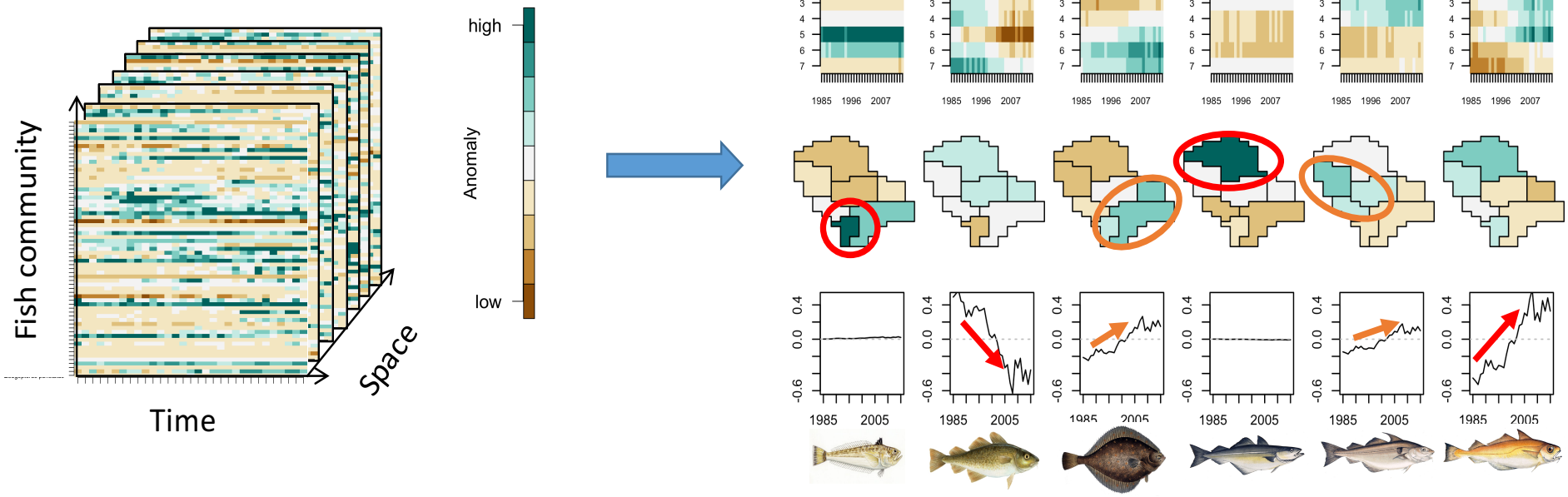
1. Select variables and check the distribution
2. Scale the variables
3. Run the analysis + selection # components (scree test)
4. Interpret the selected components
5. **(optional) Cluster analysis**

Clustering





3D – Principal Tensor Analysis



1. Select variables and check the distribution
2. Scale the variables
3. **Run the analysis + selection # components (scree test)**
4. Interpret the selected components
5. (optional) Cluster analysis

Output of PTA-k

```
> PTA3(IBTS_logscale, nbPT = 3, nbPT2 = 3)
```

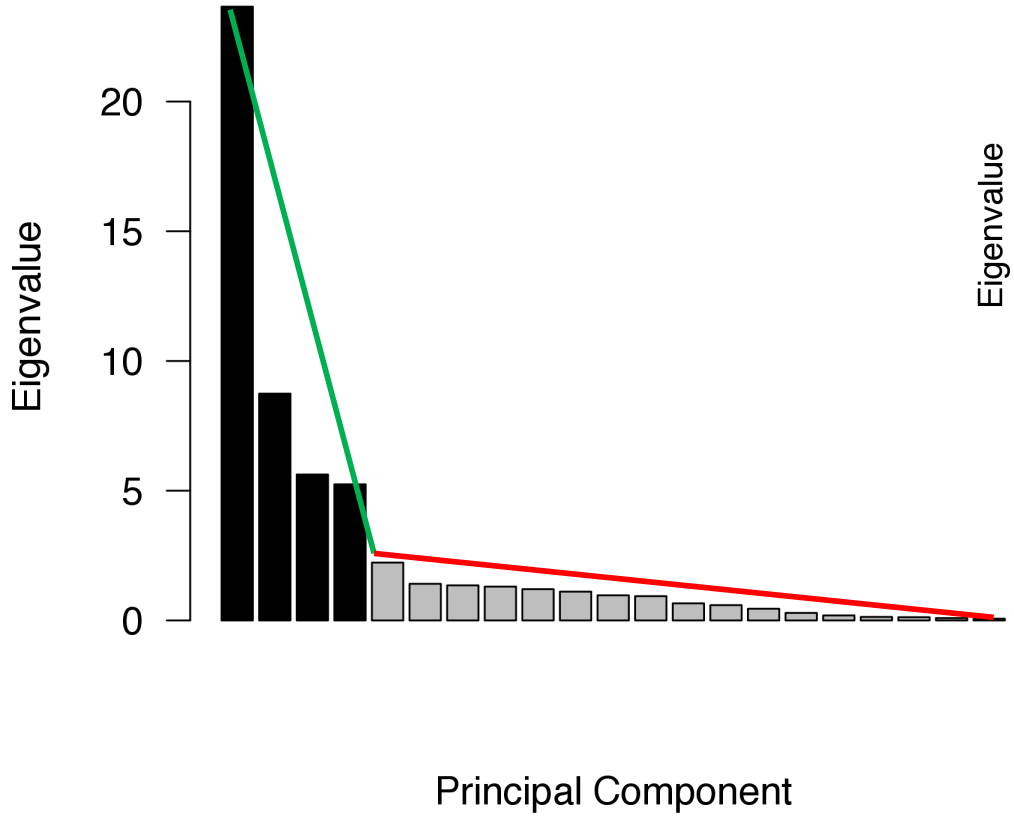
First principal tensor

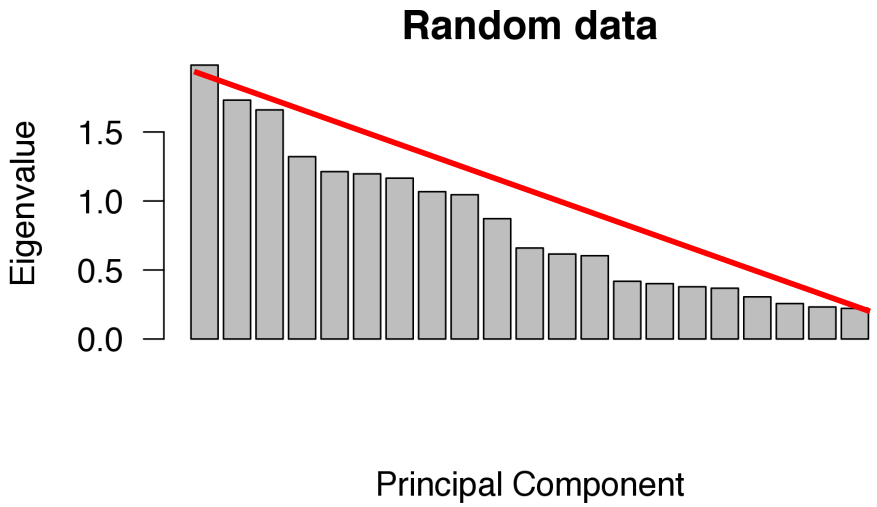
PT associated to vs111 in different mode :
65 species mode
31 temporal mode
7 spatial mode

```
++++ PTA-  3 modes ++++
                data=  IBTS_logscale   65 31 7
                ------Percent Rebuilt---- 56.37922 %
                ------Percent Rebuilt from Selected ---- 56.37922 %
               -no- --Sing Val--  --ssX-- --local Pct-- --Global Pct--
vs111                     1      57.6306 14040.00       23.6559       23.655875
65 vs111 31 7             3       9.0978  3464.06        2.3894        0.589533
65 vs111 31 7             4       5.2054  3464.06        0.7822        0.192991
31 vs111 65 7             6      35.0333  6511.03       18.8500        8.741668
31 vs111 65 7             7      28.0998  6511.03       12.1271        5.623930
7 vs111 65 31             9      17.6930  4600.42        6.8047        2.229654
7 vs111 65 31            10      11.4570  4600.42        2.8533        0.934920
vs222                    11      27.1374  6107.06       12.0588        5.245304
65 vs222 31 7            13       4.1761   789.39        2.2093        0.124215
65 vs222 31 7            14       3.5684   789.39        1.6131        0.090693
31 vs222 65 7            16      13.7724  1313.69       14.4387        1.350994
31 vs222 65 7            17      12.4827  1313.69       11.8612        1.109821
7 vs222 65 31            19      14.0761  2068.85        9.5771        1.411229
7 vs222 65 31            20      13.0077  2068.85        8.1785        1.205129
vs333                    21      13.5278  3408.02        5.3697        1.303431
65 vs333 31 7            23       4.3630   220.69        8.6256        0.135581
65 vs333 31 7            24       3.0676   220.69        4.2639        0.067022
31 vs333 65 7            26       7.9591   336.79       18.8094        0.451196
31 vs333 65 7            27       6.3762   336.79       12.0715        0.289568
7 vs333 65 31            29      11.6585   996.61       13.6383        0.968089
7 vs333 65 31            30       9.6144   996.61        9.2751        0.658381
```

1. Select variables and check the distribution
2. Scale the variables
3. **Run the analysis + selection # components (scree test)**
4. Interpret the selected components
5. (optional) Cluster analysis

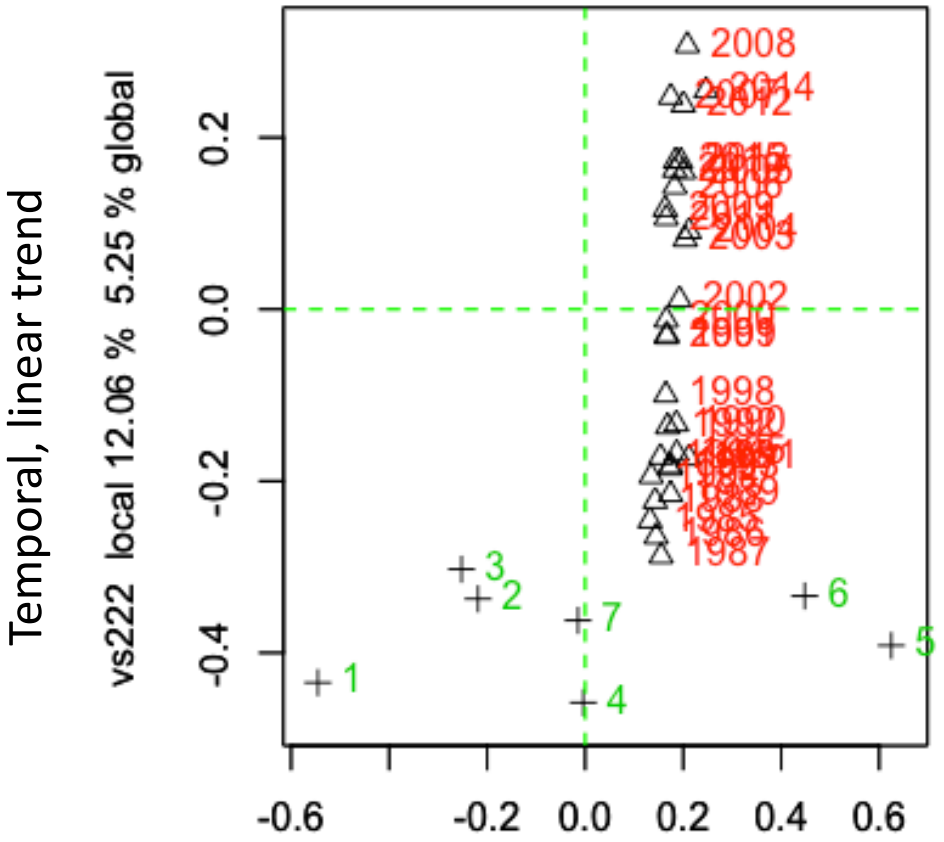
Selection of Principal Tensor



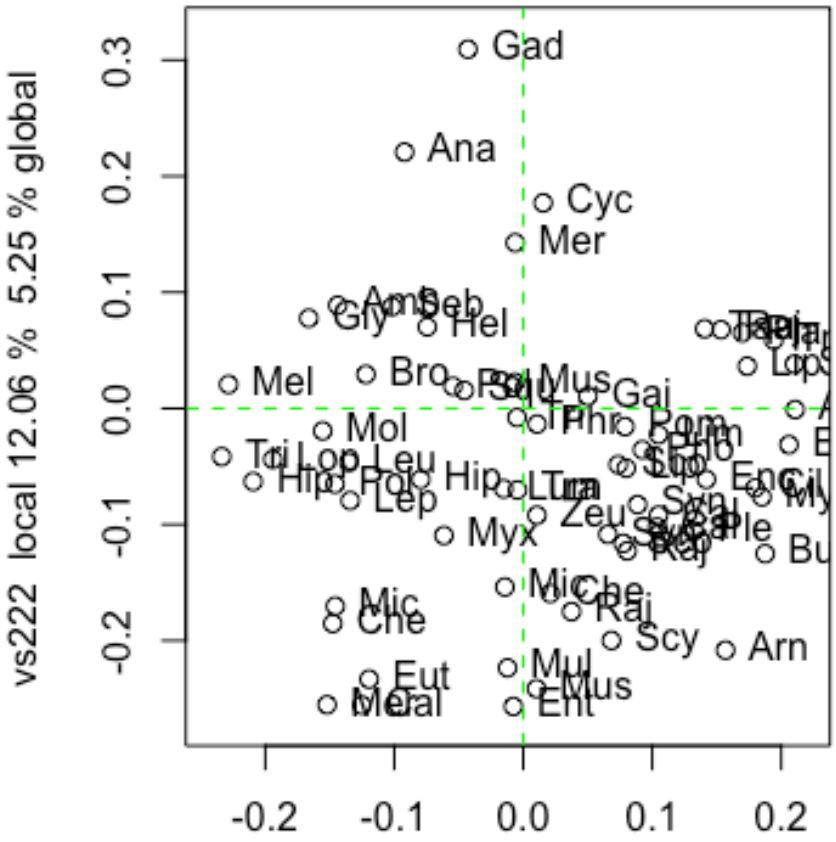


1. Select variables and check the distribution
2. Scale the variables
3. Run the analysis + selection # components (scree test)
4. **Interpret the selected components**
5. (optional) Cluster analysis

Interpret PT

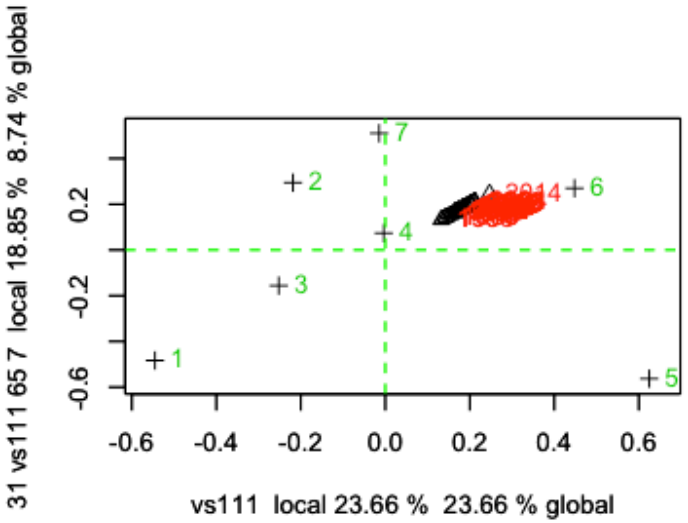


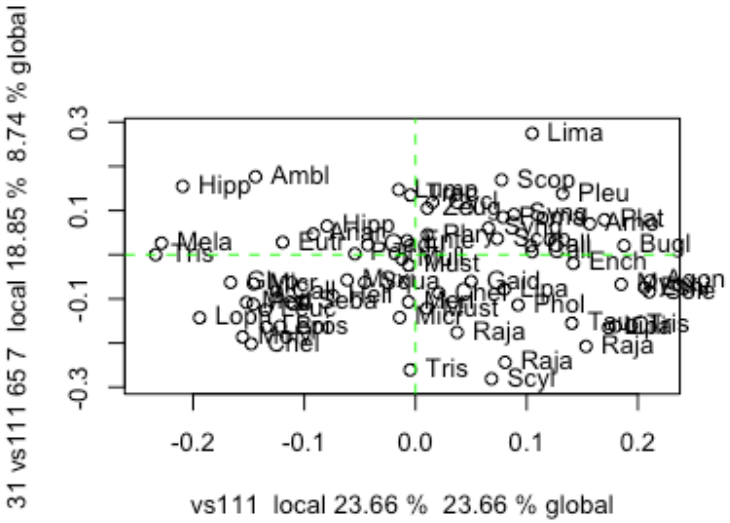
Spatial, North - South gradient

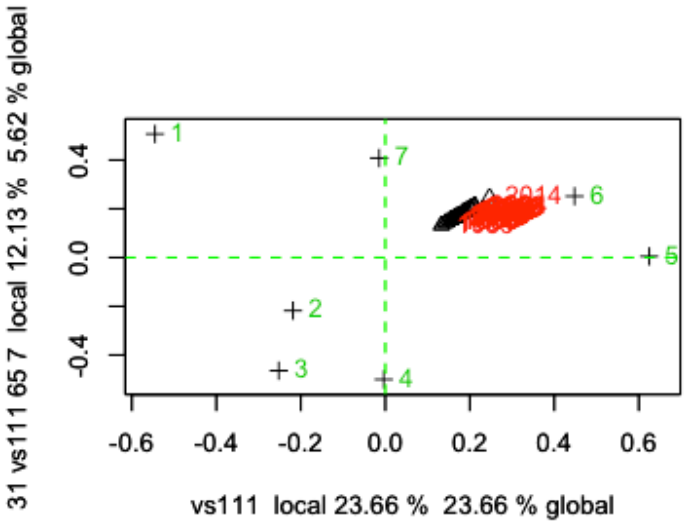


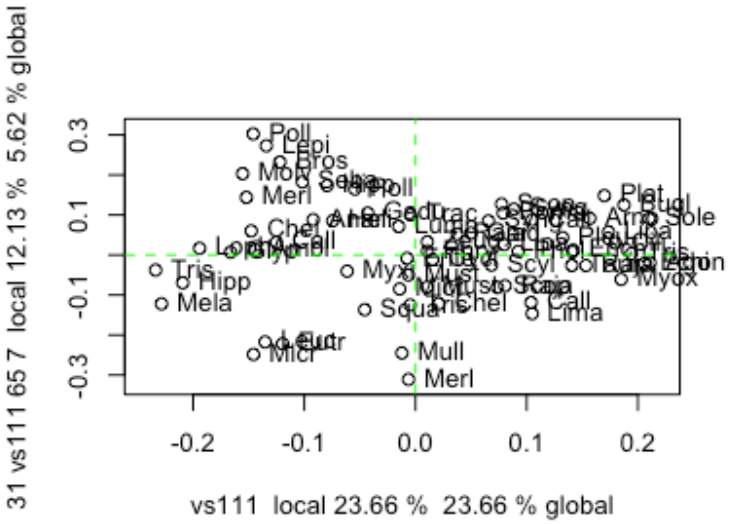
1. Select variables and check the distribution
2. Scale the variables
3. Run the analysis + selection # components (scree test)
4. **Interpret the selected components**
5. (optional) Cluster analysis

Interpret PT



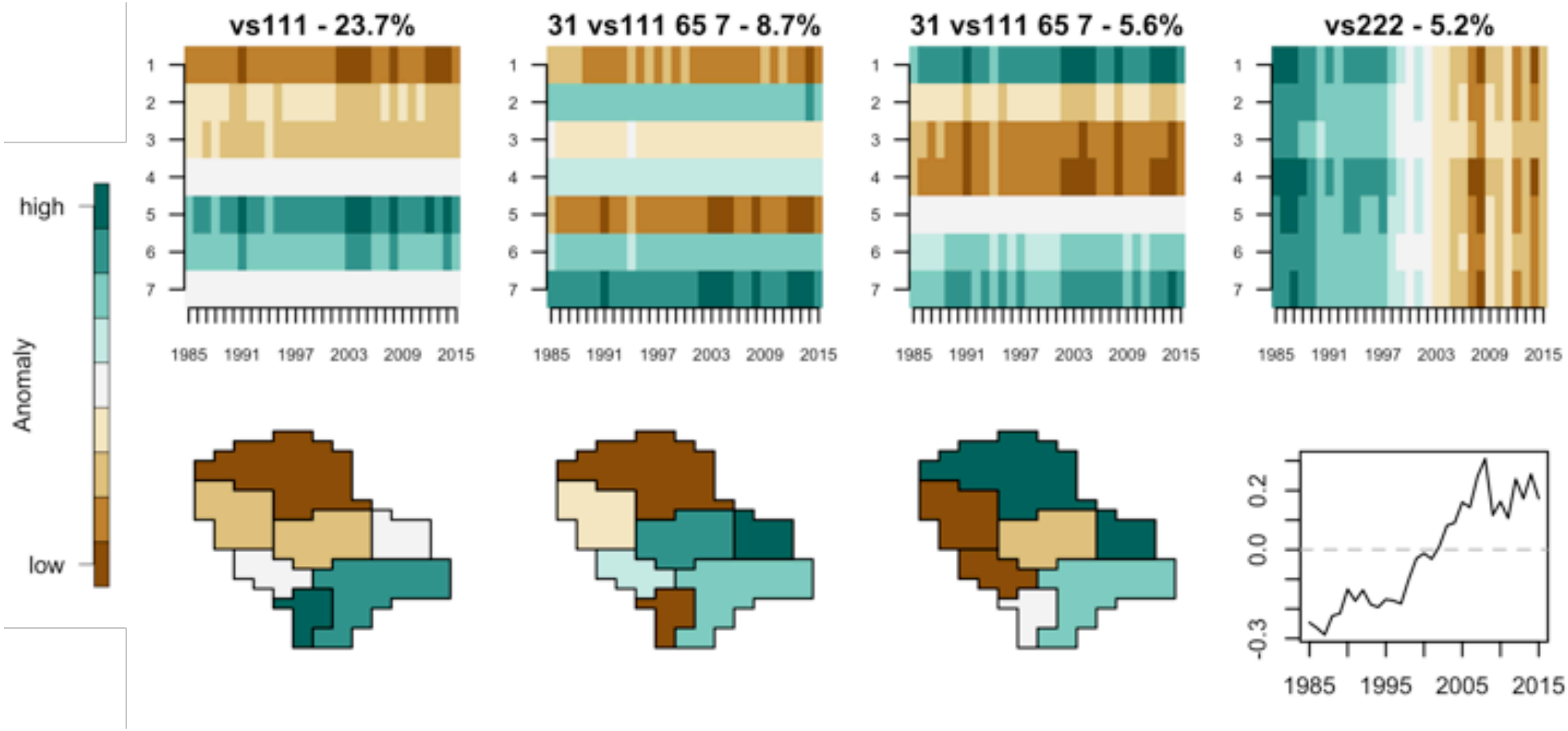






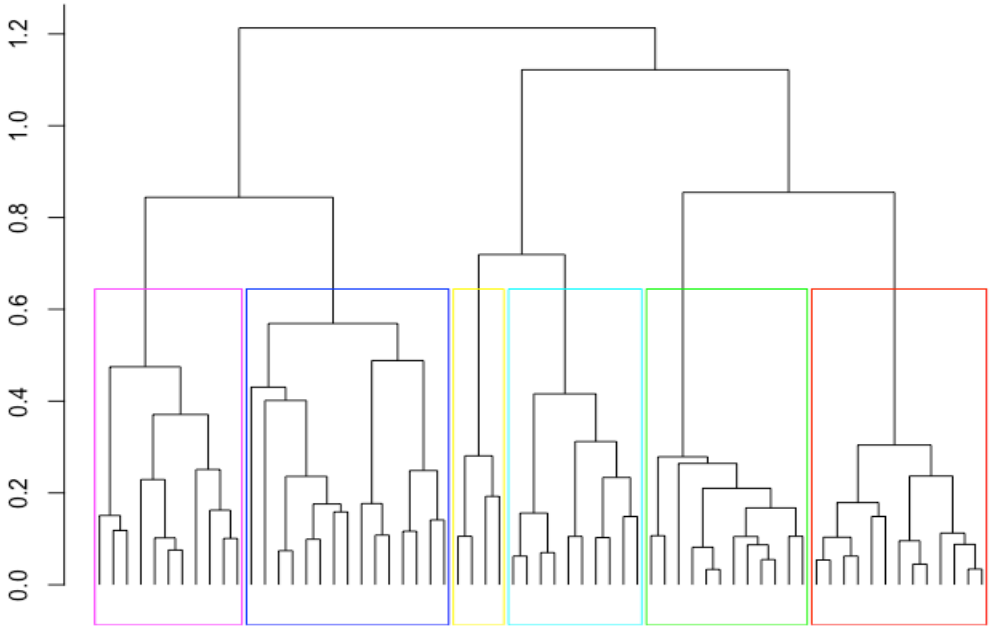
1. Select variables and check the distribution
2. Scale the variables
3. Run the analysis + selection # components (scree test)
4. **Interpret the selected components**
5. (optional) Cluster analysis

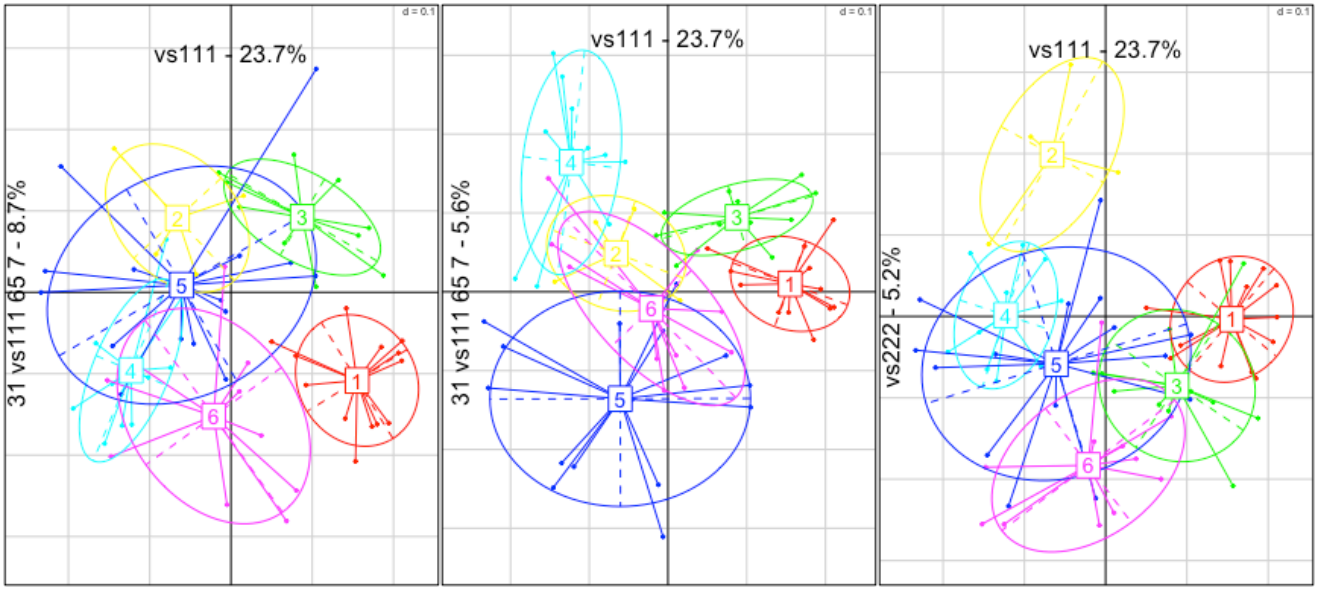
Interpret PT



1. Select variables and check the distribution
2. Scale the variables
3. Run the analysis + selection # components (scree test)
4. Interpret the selected components
5. **(optional) Cluster analysis**

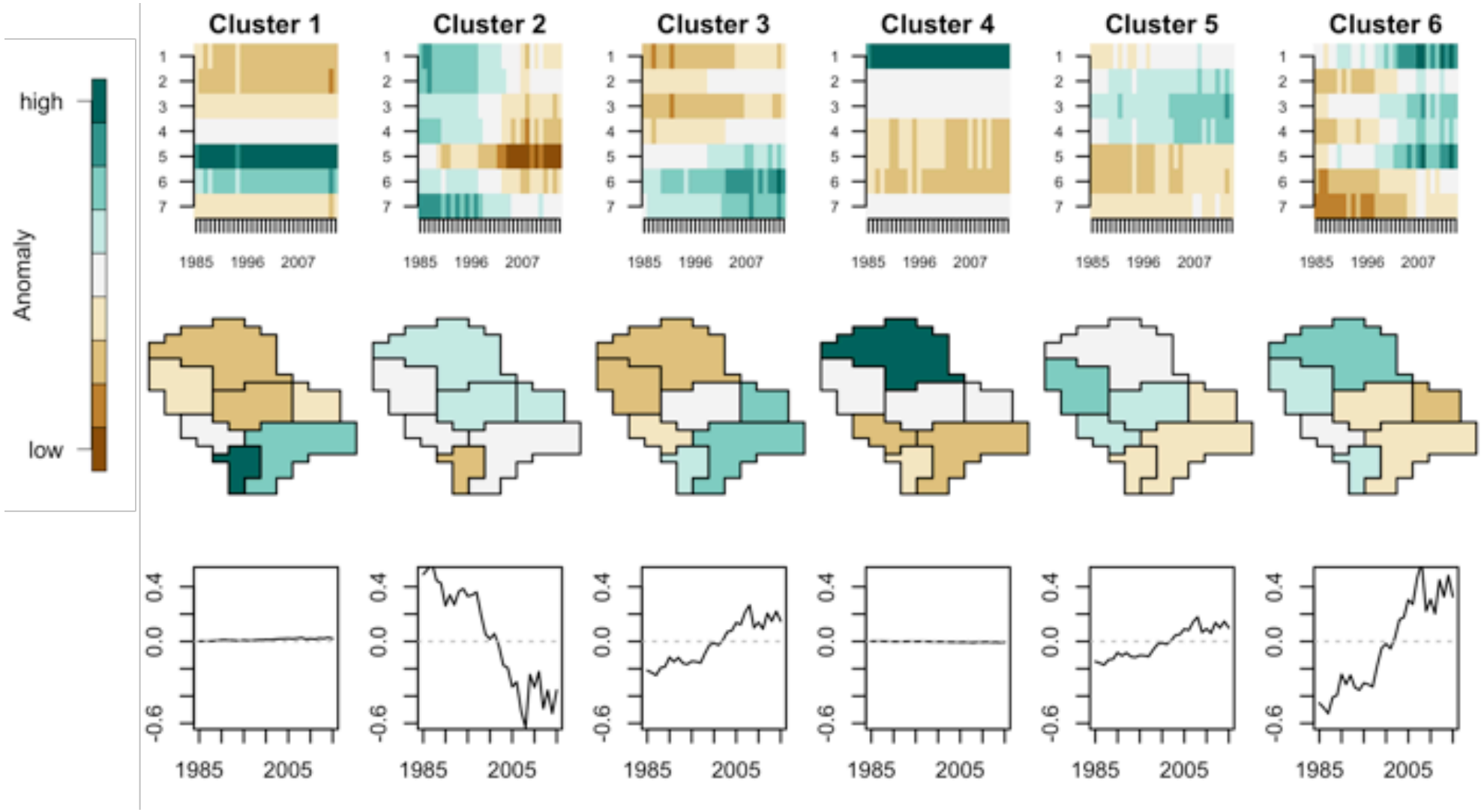
Clustering





1. Select variables and check the distribution
2. Scale the variables
3. Run the analysis + selection # components (scree test)
4. Interpret the selected components
5. **(optional) Cluster analysis**

Clustering

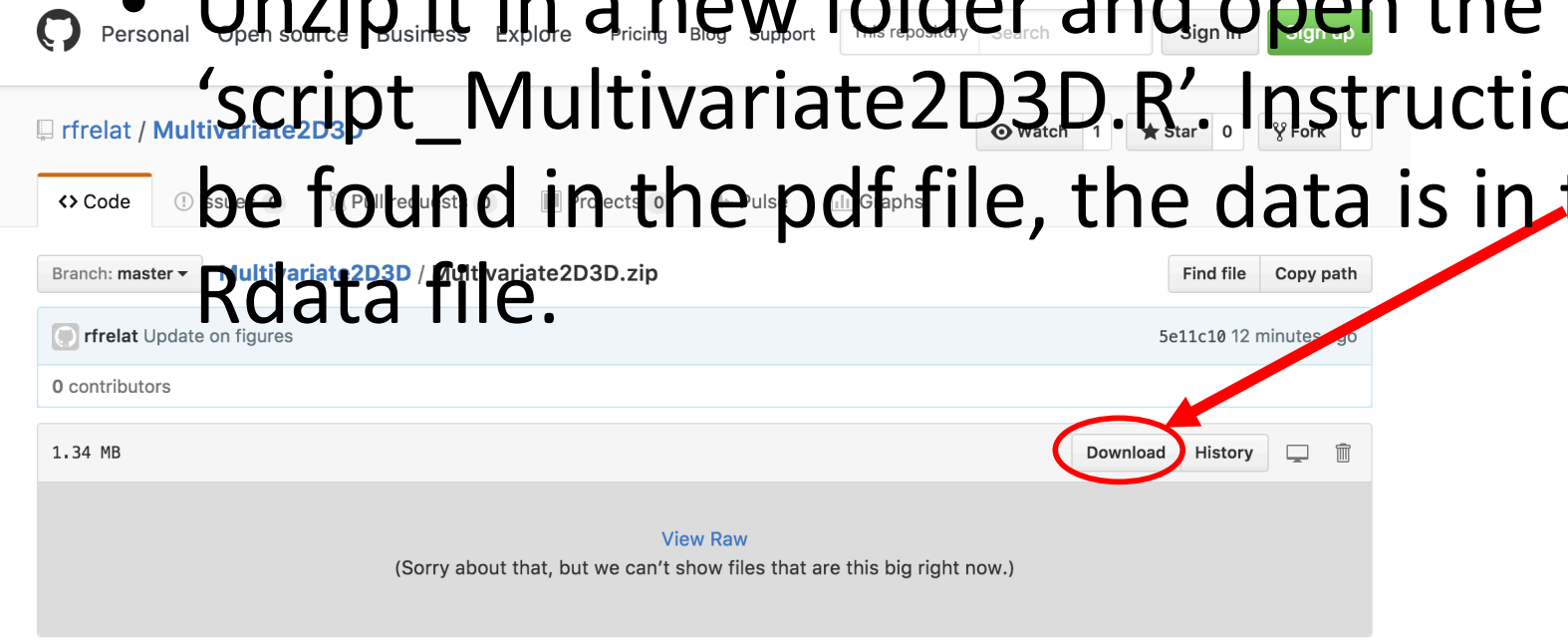


Summary

Steps	2D matrix	3D tensor
1. Check distribution	If too skewed, log or square root transform log(), sqrt()	
2. Scale	Automatically done in dudi.pca()	Depending on each dataset
3. Run the analysis	dudi.pca()	PTA3() or PTAk()
and select # of components with scree plot	Automatically	summary.PTAk()
4. Interpret	s.corcircle() s.label() scatter()	plot(... , mod=, nb1= , nb2=)
5. Cluster (if needed)	dist() : compute the distance between individuals hclust() : create the dendogram cutree() : create the clusters s.class() : visualize the clusters on the components	
Package :	ade4	PTA-k

Your turn

- Download the zip file Multivariate2D3D.zip in : https://github.com/rfrelat/Multivariate2D3D/
- Unzip it in a new folder and open the file 'script_Multivariate2D3D.R'. Instructions can be found in the pdf file, the data is in the Rdata file.



Any question, feel free to write me an e-mail :
romain.frelat@uni-hamburg.de