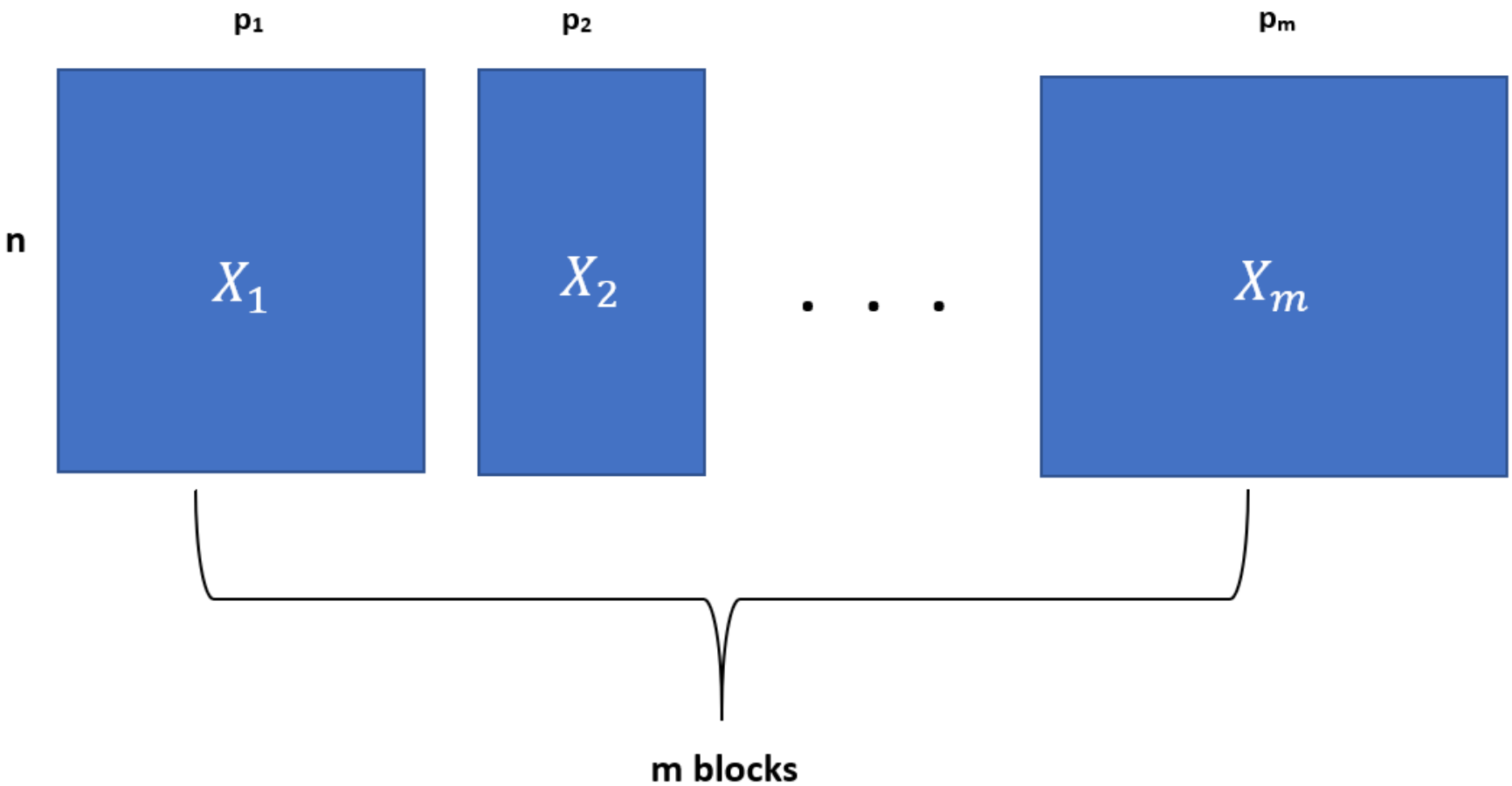
ClustBlock: A package for clustering datasets

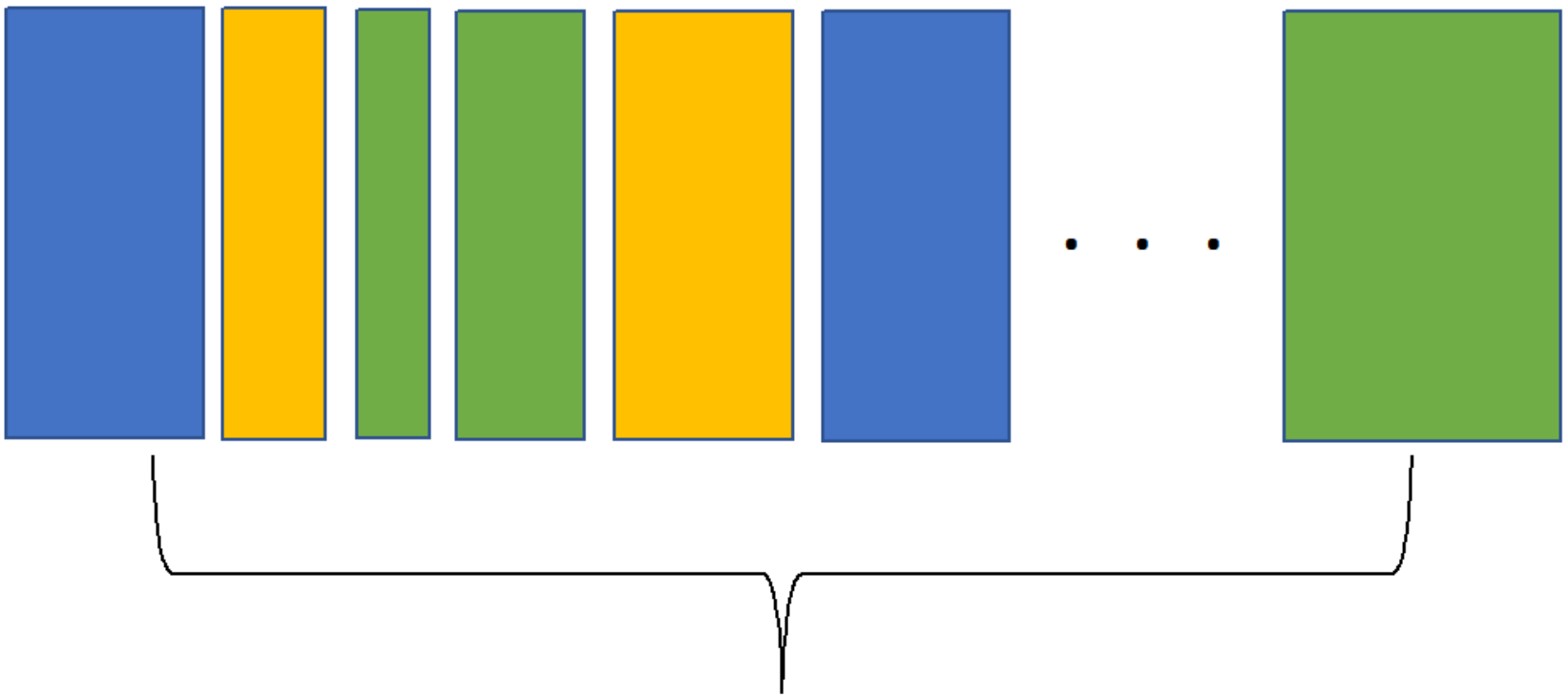
Fabien Llobell, Évelyne Vigneau, Véronique Cariou, El Mostafa Qannari

useR! 2019

Data: Blocks



Aim

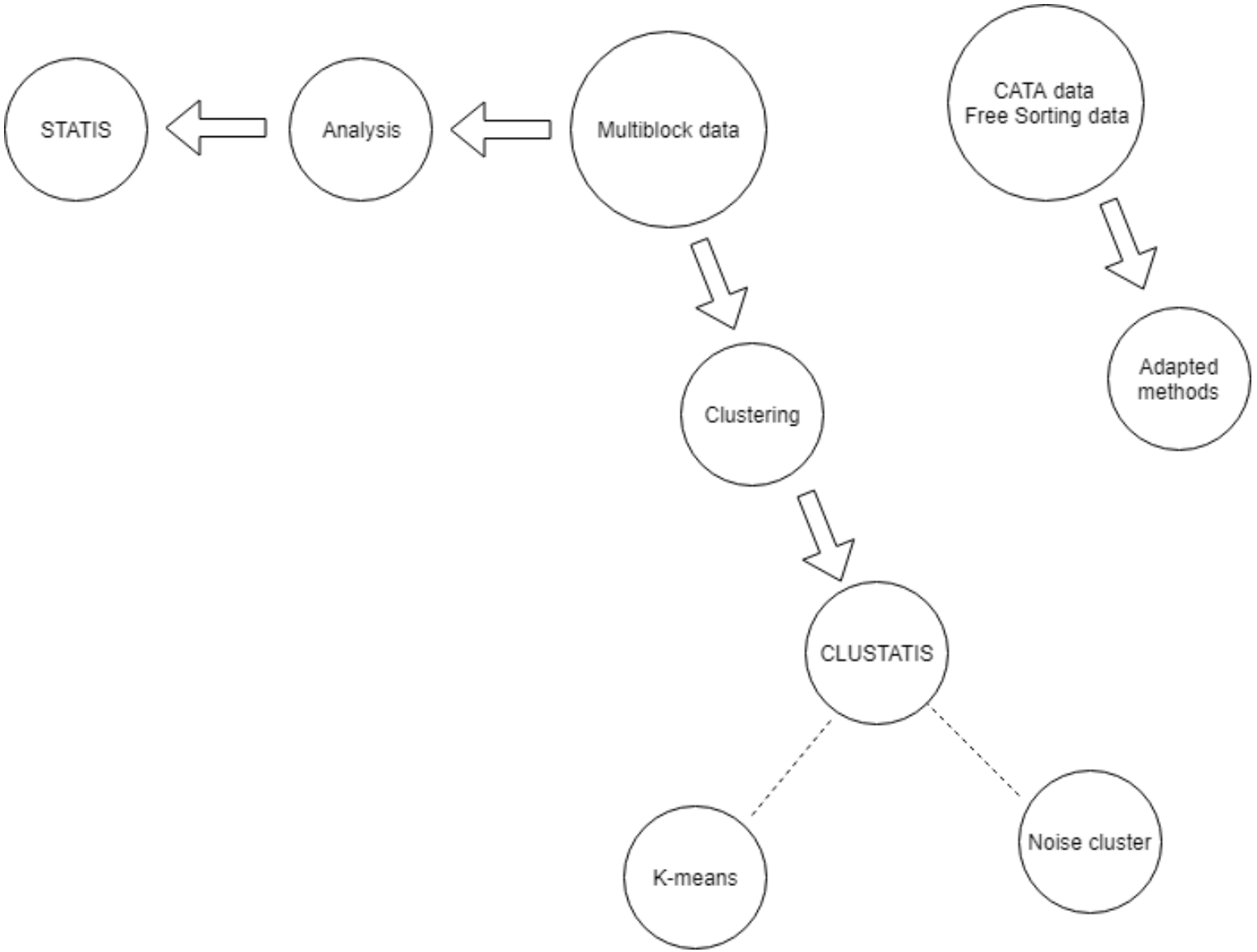


Data: Blocks

There are more and more situations where users have to deal with several blocks of variables

- Datasets repeated in time or space
- Different sources of measurements to characterize the same observations
- In sensory analysis: Projective mapping/Napping, Free choice profiling, Check-All-That-Apply

Package ClustBlock on CRAN



Cluster analysis of blocks of variables

The CLUSTATIS method: Clustering of blocks of quantitative variables decribing the same observations but variables may be different from one block to another

- Test to know if there is more than one cluster
- Recommended number of clusters
- Indices to assess homogeneity of clusters
- Possibility to introduce a noise cluster
- Graphical representations

Criterion of CLUSTATIS

$\mathbf{W_i}$ the scalar product matrix of the dataset i, $\mathbf{W^{(k)}}$ the compromise of the cluster k, m the number of blocks and K the number of clusters

Minimization of:

$$\mathbf{D} = \sum_{\mathbf{k=1}}^{\mathbf{K}} \sum_{\mathbf{i} \in \mathbf{G_k}} ||\mathbf{W_i} - \alpha_{\mathbf{i}}^{\mathbf{(k)}} \mathbf{W^{(k)}}||^{\mathbf{2}}$$

Equivalent to the maximization of:

$$\mathbf{Q} = \sum_{\mathbf{k=1}}^{\mathbf{K}} \sum_{\mathbf{i} \in \mathbf{G_k}} \mathbf{RV^2}(\mathbf{W_i}, \mathbf{W^{(k)}})$$

If $K = 1$, $\mathbf{D} = \sum_{\mathbf{i=1}}^{\mathbf{m}} ||\mathbf{W_i} - \alpha_{\mathbf{i}} \mathbf{W}||^{\mathbf{2}} \Longrightarrow$ STATIS method (multiblock data analysis)

Solutions

- Hierarchical clustering algorithm

  $\longrightarrow$ At each step, the smallest increase of D is taken

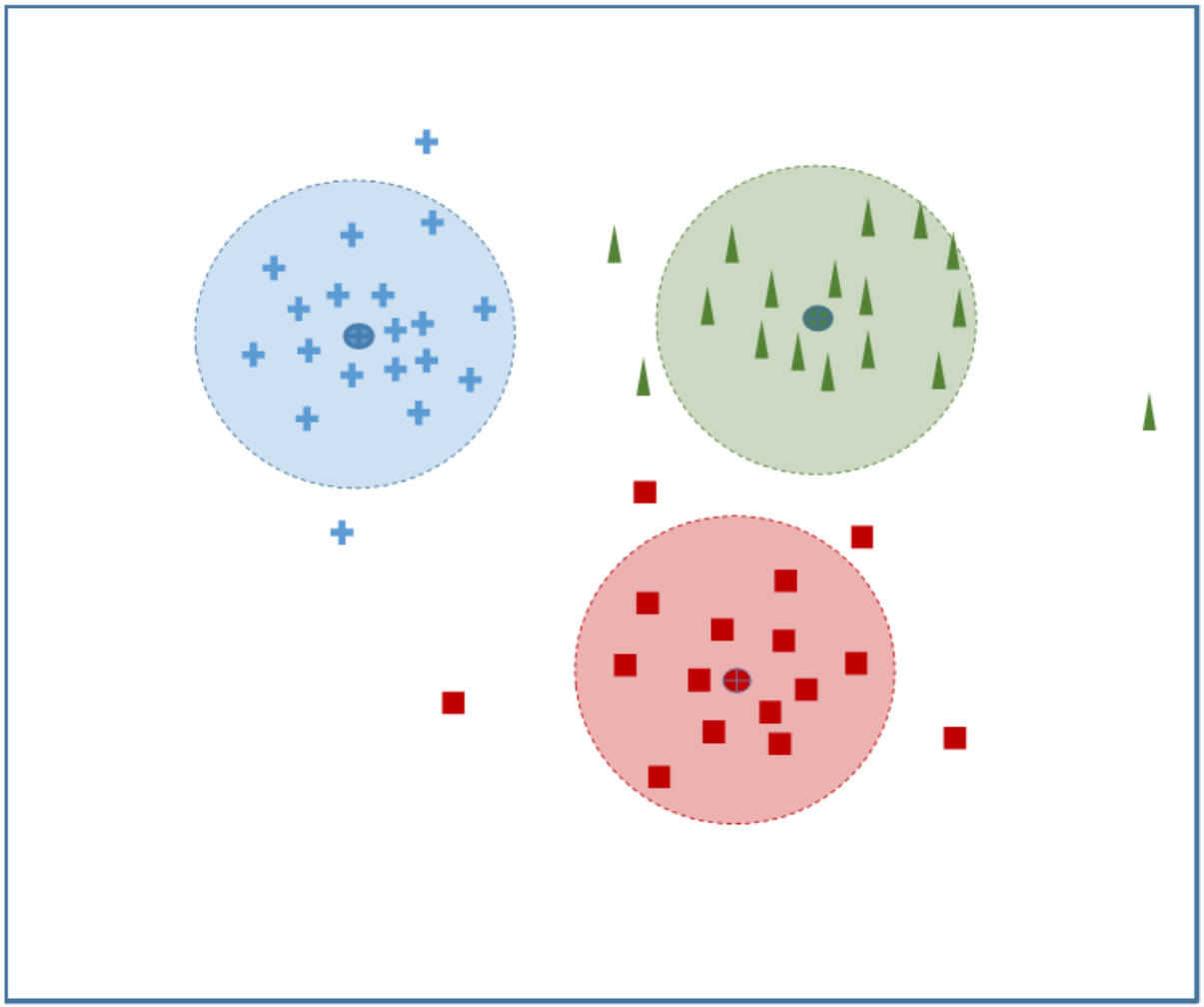
- Partitioning algorithm

  $\longrightarrow$ At each iteration, aggregating each dataset with the nearest compromise (RV coefficient)

Help with the number of clusters

- Permutation test to know if there is more than one cluster
- Recommended number of clusters computed by an adaptation of the Hartigan index

Introduction of a noise cluster



Introduction of a noise cluster: Modification of the criterion

$$\mathbf{Q} = \sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{m}} \sum_{\mathbf{k}=\mathbf{1}}^{\mathbf{K}} (\delta_{\mathbf{ik}} \mathbf{RV^2}(\mathbf{W_i}, \mathbf{W^{(k)}}) + \delta_{\mathbf{i(K+1)}} \rho^{\mathbf{2}})$$

$\Longrightarrow$ Partitioning algorithm

$\longrightarrow$ At each iteration, aggregating each dataset with the nearest compromise (RV coefficient) or to the noise cluster if the similarity with every compromises is lower than ρ (automatically computed)

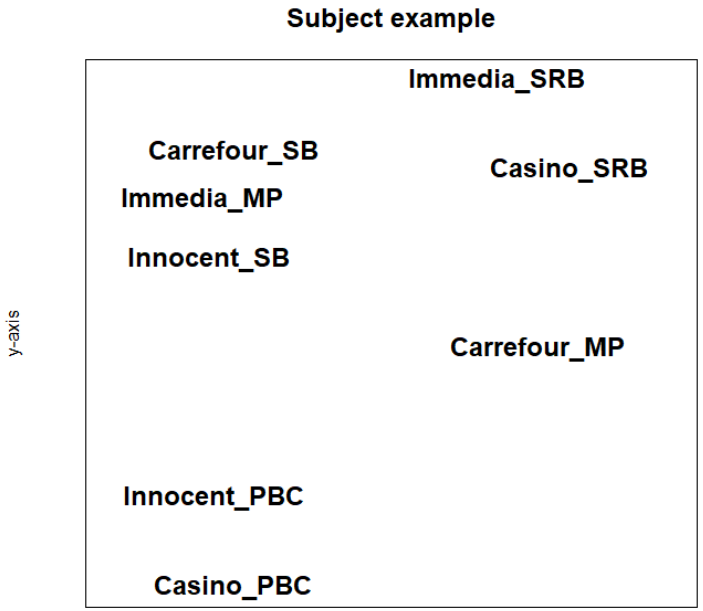
Datasets

$\longrightarrow$ Projective mapping of smoothies

$\longrightarrow$ Concern 8 smoothies and 24 subjects (datasets)

Francois Husson, Sebastien Le and Marine Cadoret (2017). SensoMineR: Sensory Data Analysis. R package version 1.23. https://CRAN.R-project.org/package=SensoMineR

Example of Data Blocks: Projective mapping/Napping



	X	Y
Immedia_MP	23.50	23.50
Carrefour_MP	39.50	21.00
Immedia_SRB	37.50	25.50
Casino_SRB	41.00	24.00
Innocent_PBC	24.00	18.50
Casino_PBC	25.00	17.00
Innocent_SB	23.80	22.50
Carrefour_SB	25.00	23.80

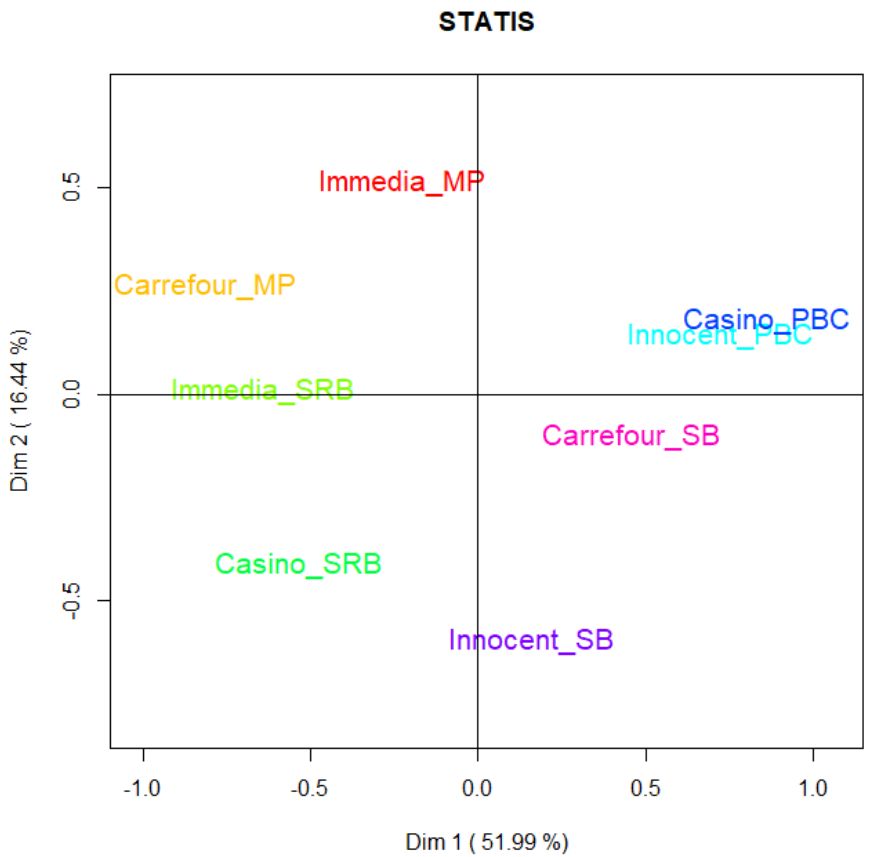
Francois Husson, Sebastien Le and Marine Cadoret (2017). SensoMineR: Sensory Data Analysis. R package version 1.23. https://CRAN.R-project.org/package=SensoMineR

Results

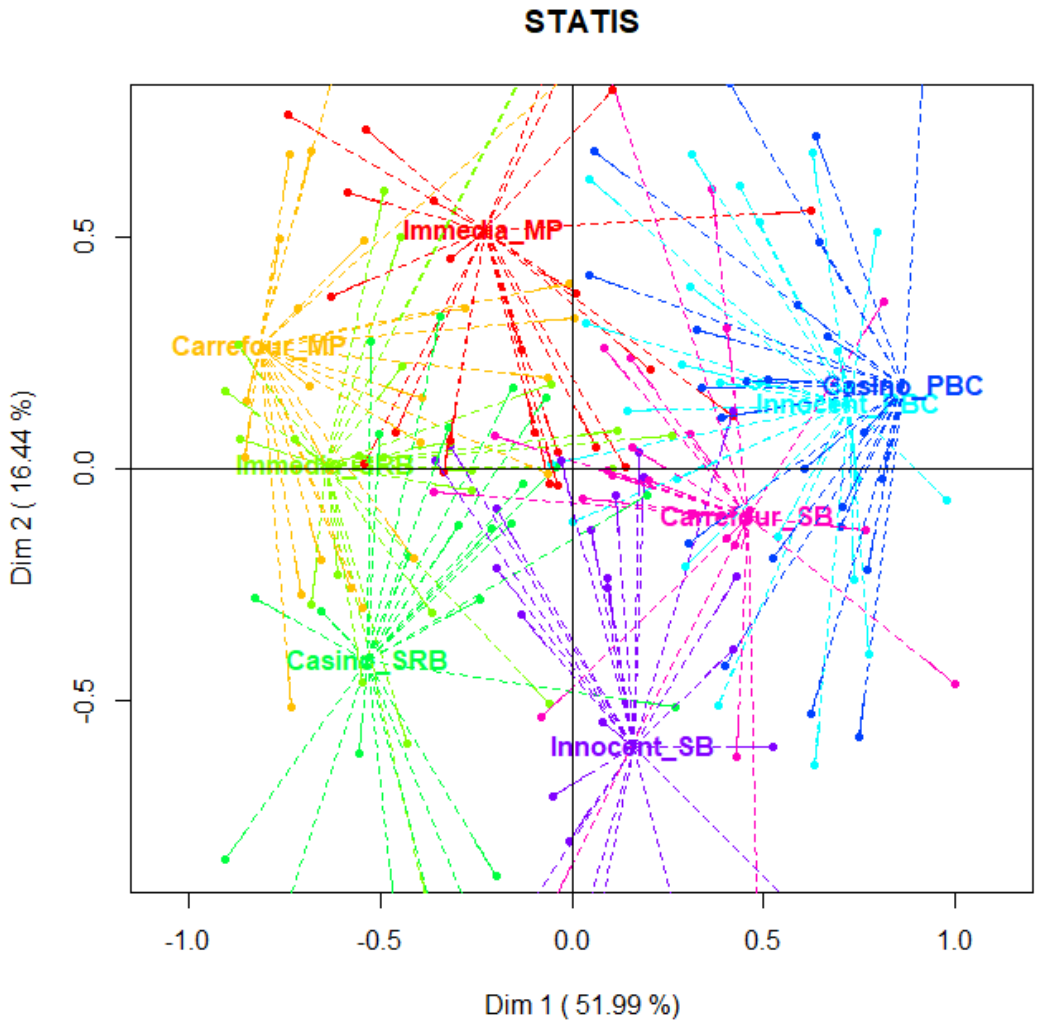
$\longrightarrow$ The test indicates that clustering is not necessary

$\longrightarrow$ We can only perform the STATIS method

```
res.statis=statis(Data=smoo, Blocks=rep(2,24))
```

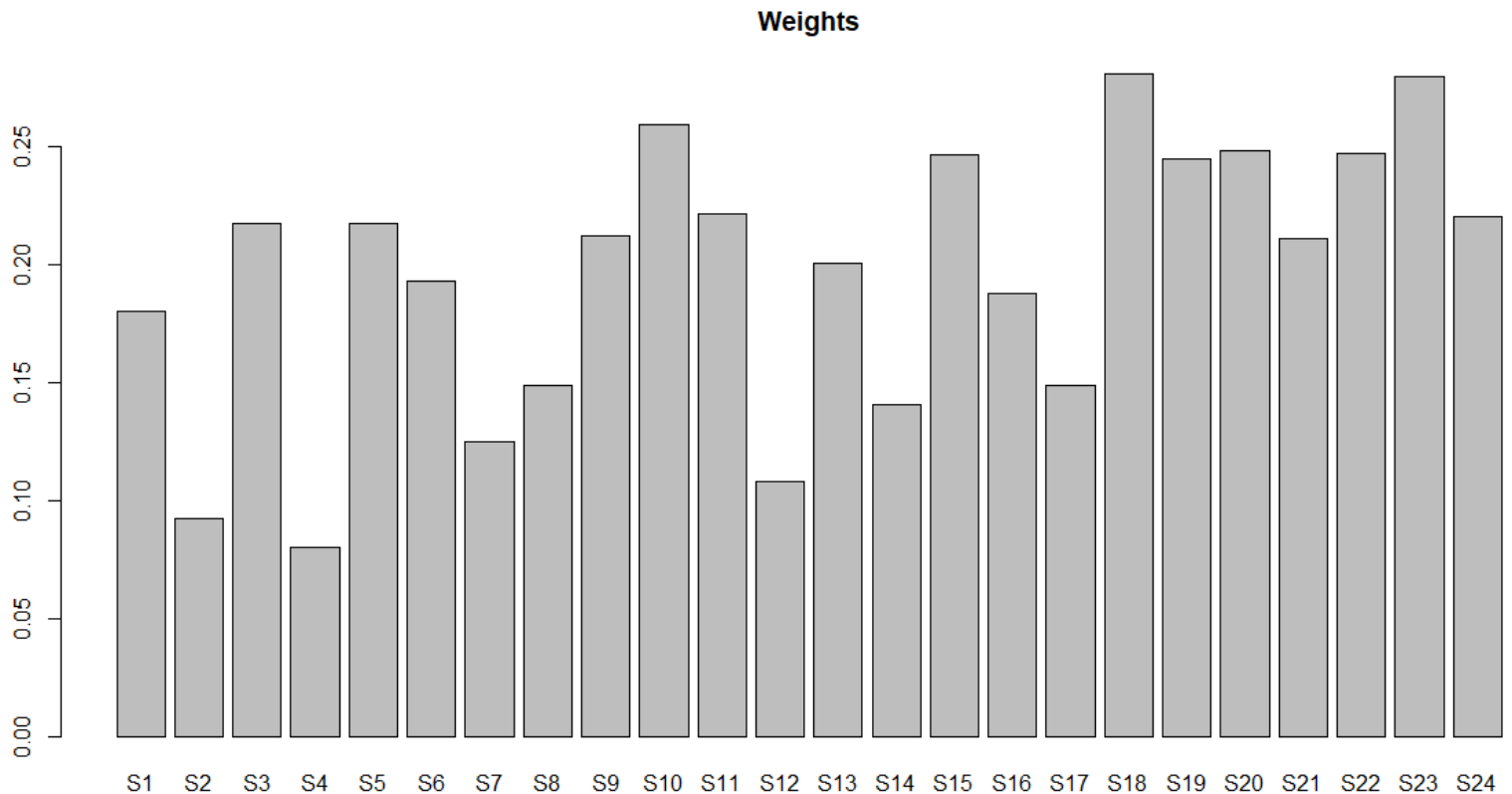


Other graphics



Other graphics

```
plot(res.statis)
```



Datasets

$\longrightarrow$ Projective mapping of yoghurts

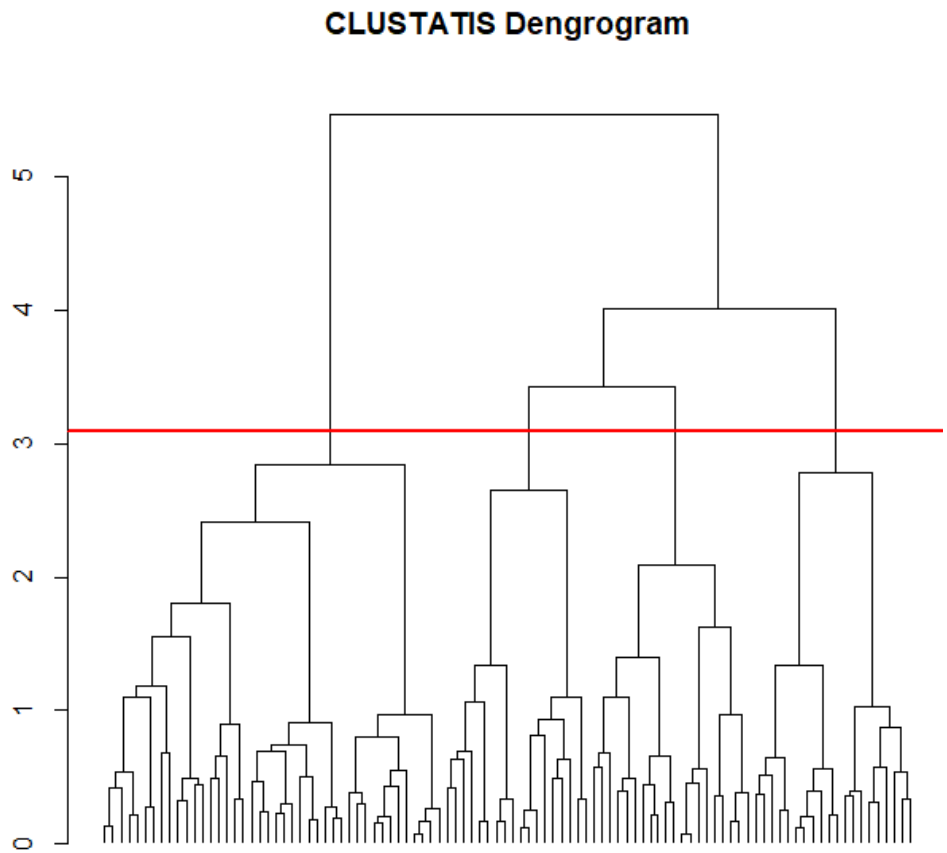
$\longrightarrow$ Concern 12 yoghurts and 100 subjects (datasets)

Berget, I., Varela, P., & Næs, T. (2019). Segmentation in projective mapping. Food Quality and Preference, 71, 8-20.

Application of CLUSTATIS

```
res.clustatis=clustatis(Data, Blocks = rep(2,100),
 Noise_cluster = TRUE)
```

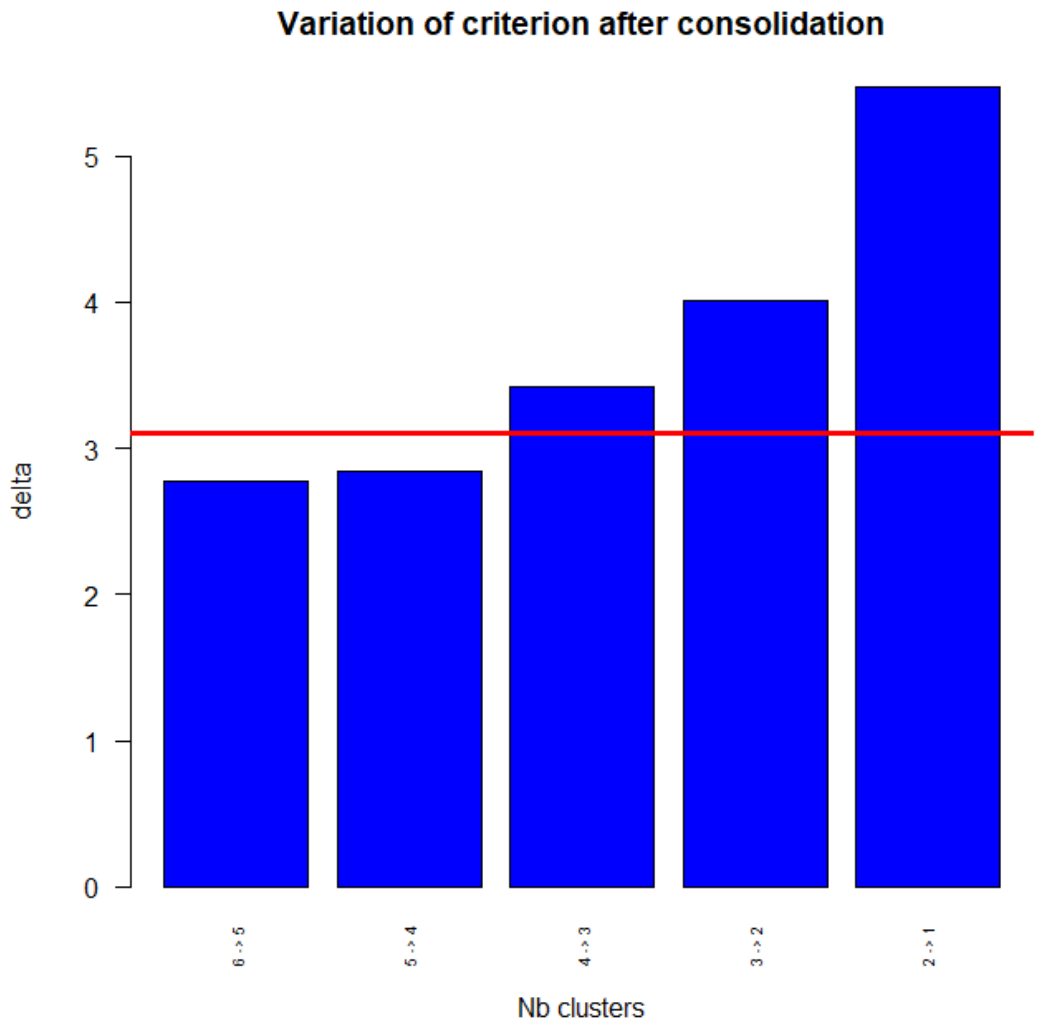
The test indicates that clustering is necessary



Segmentation in 4 clusters

Other help for the choice of the number of clusters

```
plot(res.clustatis)
```



Homogeneity indices

```
summary(res.clustatis)
```

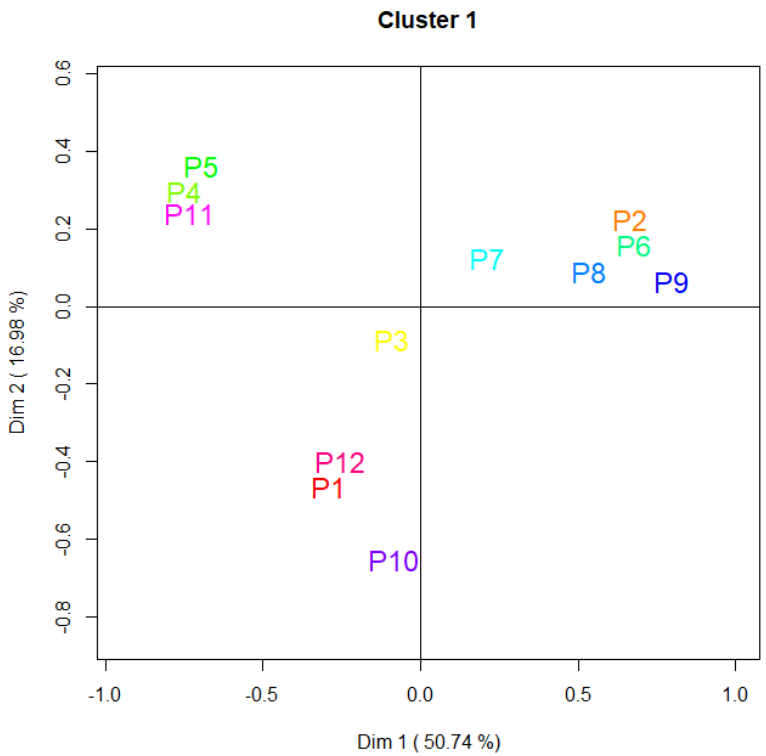
Homogeneity

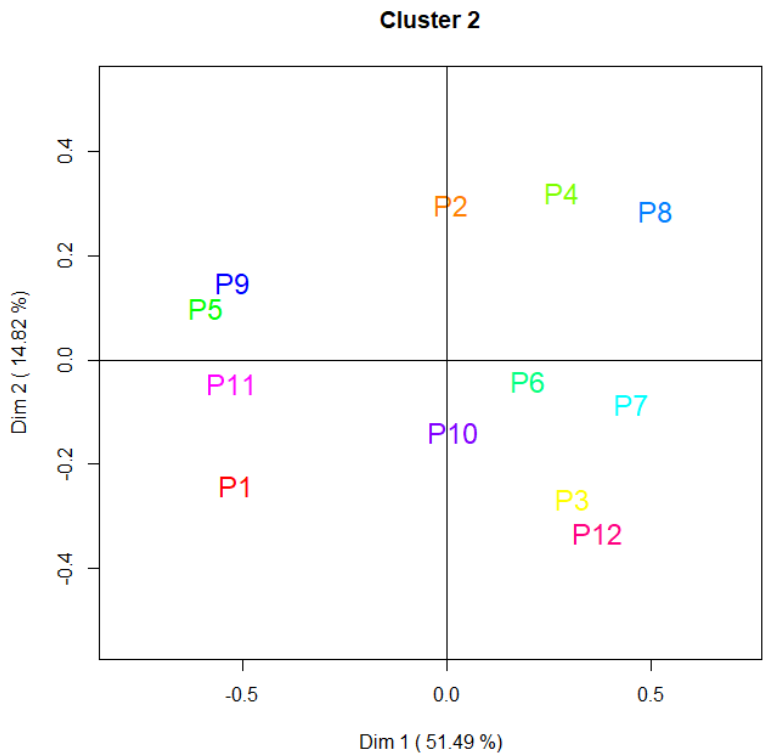
Cluster	Homogeneity	# Subjects
1	47.3%	33
2	46.4%	10
3	42.7%	15
4	51.2%	14
Noise cluster	21.9%	28
Overall homogeneity	47.0%	72
One group	24.3%	100

- Indices of homogeneity for all the clusters, overall and with no clustering
- Subjects who do not fit any cluster are put in an additional cluster

Graphical representations

```
plot(res.clustatis, ngroups=4)
```





Conclusion

- ClustBlock contains clustering methods for multiblock datasets
- Several indices to assess the quality of the cluster solution
- Possibility to discard atypical datasets
- Help for the choice of the number of clusters
- Analysis of each cluster with graphical representations
- Possibility of analysis without clustering
- Multi-start procedures are also available
- Specific methods for CATA and Free Sorting data

References

- Fabien Llobell (2019). ClustBlock: Clustering of Datasets. R package version 2.0.0.
- *Llobell, F., Cariou, V., Vigneau, E., Labenne, A., & Qannari, E. M. (2018). Analysis and clustering of multiblock datasets by means of the STATIS and CLUSTATIS methods. Application to sensometrics. Food Quality and Preference.*
- *Llobell, F., Vigneau, E., & Qannari, E. M. (2019). Clustering datasets by means of CLUSTATIS with identification of atypical datasets. Application to sensometrics. Food Quality and Preference, 75, 97-104.*
- *Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2019). A new approach for the analysis of data and the clustering of subjects in a CATA experiment. Food Quality and Preference, 72, 31-39.*
- *Llobell, F., Giacalone, D., Labenne, A., Qannari, E.M. (2019). Assessment of the agreement and cluster analysis of the respondents in a CATA experiment. Food Quality and Preference, 77, 184-190.*