Analysis of Free Sorting Data

Overview of statistical techniques for representing a set of partitions

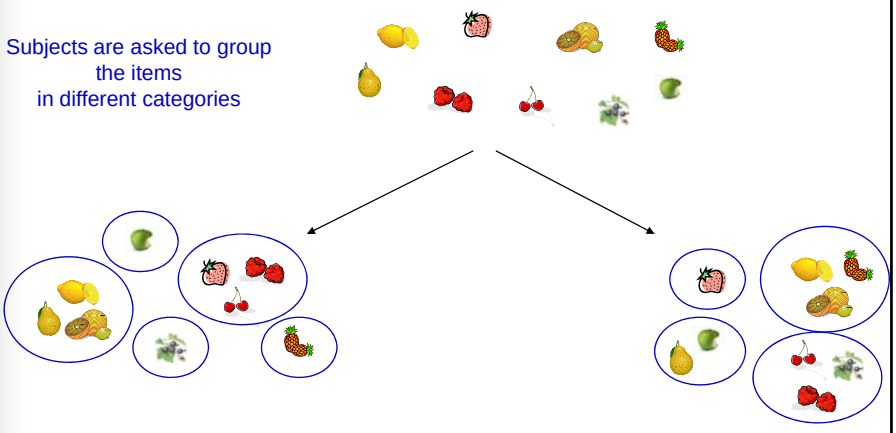
Clustering of subjects

Philippe Courcoux *Pauline Faye* *El Mostafa Qannari*

Unité de Sensométrie et Chimiométrie
ONIRIS Nantes
philippe.courcoux@oniris-nantes.fr

Free Sorting task

Subjects are asked to group the items in different categories

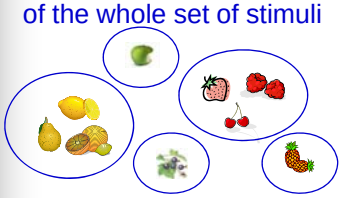


Holistic and non-verbal technique
(often used with un-trained subjects)

+ verbalisation task

Free Sorting Task

Each subject gives a partition of the whole set of stimuli



Individual outcomes

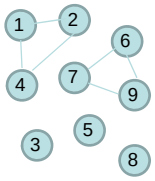
Disjoint classes

{1, 2, 4} {6, 7, 9} {3} {5} {8}

Disjonctive coding

gr1	gr2	gr3	Gr4	gr5
1	0	0	0	0
1	0	0	0	0
0	0	1	0	0
1	0	0	0	0
0	0	0	1	0
0	1	0	0	0
0	1	0	0	0
0	0	0	0	1
0	1	0	0	0

Graph



Dissimilarity matrix

0	0	1	0	1	1	1	1	1
0	0	1	0	1	1	1	1	1
1	1	0	1	1	1	1	1	1
0	0	1	0	1	1	1	1	1
1	1	1	1	0	1	1	1	1
1	1	1	1	1	0	1	1	1
1	1	1	1	1	1	0	1	1
1	1	1	1	1	1	1	0	1
1	1	1	1	1	1	1	1	0

Analysis of sorting data

Analysis of a set of partitions on the same set of objects

Two families of statistical procedures

✓ **Techniques of classification**

- **Trees**

  Hierarchical Ascendant Classification
  Additive Distance Trees

- **Partitions**

  Median partition

✓ **Factorial techniques** (assumption of latent variables)

- **Multidimensionnal Scaling (MDS)**

Proximities between objects : Co-occurrence matrix

K subjects

$$c_{ij} = \sum_{k=1}^{K} c_{ij}^{k}$$

Number of subjects
who put objects *i* and *j* in the same group

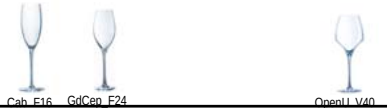
	CabCou30	CabF16	CabT47	CabBea38	GdCepF24	CabVJV35	AromUOA47	CabBal47	OpenSV165	LinV45	AromUOA41	OpenUV40	LinV19	OpenUF20	GdCepVR62	GdCepVB47
CabCou30	41	16	2	6	13	1	0	6	6	1	1	3	4	15	0	3
CabF16	16	41	2	2	37	0	0	2	3	0	1	0	1	29	0	1
CabT47	2	2	41	4	1	5	14	7	4	8	15	3	3	0	16	9
CabBea38	6	2	4	41	0	9	9	13	30	8	13	7	14	1	3	7
GdCepF24	13	37	1	0	41	2	2	0	1	1	1	0	0	30	2	3
CabVJV35	1	0	5	9	2	41	19	5	9	24	17	23	19	2	13	24
AromUOA47	0	0	14	9	2	19	41	4	6	19	31	8	13	0	23	19
CabBal47	6	2	7	13	0	5	4	41	9	11	5	3	7	0	4	6
OpenSV165	6	3	4	30	1	9	6	9	41	8	10	11	18	3	2	7
LinV45	1	0	8	8	1	24	19	11	8	41	14	12	23	2	17	21
AromUOA41	1	1	15	13	1	17	31	5	10	14	41	8	12	0	17	12
OpenUV40	3	0	3	7	0	23	8	3	11	12	8	41	18	7	9	20
LinV19	4	1	3	14	0	19	13	7	18	23	12	18	41	2	9	18
OpenUF20	15	29	0	1	30	2	0	0	3	2	0	7	2	41	0	2
GdCepVR62	0	0	16	3	2	13	23	4	2	17	17	9	9	0	41	21
GdCepVB47	3	1	9	7	3	24	19	6	7	21	12	20	18	2	21	41

Dissimilarity between objects

$$\delta_{ij} = K - c_{ij}$$

Number of subjects
who didn't gather the objects *i* and *j*

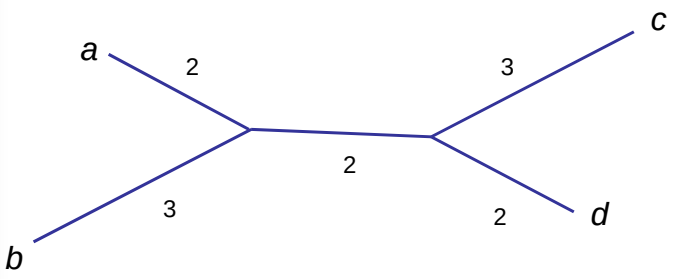
	CabCou30	CabF16	CabT47	CabBea38	GdCepF24	CabVJV35	AromUOA47	CabBal47	OpenSV165	LinV45	AromUOA41	OpenUV40	LinV19	OpenUF20	GdCepVR62	GdCepVB47
CabCou30	0	25	39	35	28	40	41	35	35	40	40	38	37	26	41	38
CabF16	25	0	39	39	4	41	41	39	38	41	40	41	40	12	41	40
CabT47	39	39	0	37	40	36	27	34	37	33	26	38	38	41	25	32
CabBea38	35	39	37	0	41	32	32	28	11	33	28	34	27	40	38	34
GdCepF24	28	4	40	41	0	39	39	41	40	40	40	41	41	11	39	38
CabVJV35	40	41	36	32	39	0	22	36	32	17	24	18	22	39	28	17
AromUOA47	41	41	27	32	39	22	0	37	35	22	10	33	28	41	18	22
CabBal47	35	39	34	28	41	36	37	0	32	30	36	38	34	41	37	35
OpenSV165	35	38	37	11	40	32	35	32	0	33	31	30	23	38	39	34
LinV45	40	41	33	33	40	17	22	30	33	0	27	29	18	39	24	20
AromUOA41	40	40	26	28	40	24	10	36	31	27	0	33	29	41	24	29
OpenUV40	38	41	38	34	41	18	33	38	30	29	33	0	23	34	32	21
LinV19	37	40	38	27	41	22	28	34	23	18	29	23	0	39	32	23
OpenUF20	26	12	41	40	11	39	41	41	38	39	41	34	39	0	41	39
GdCepVR62	41	41	25	38	39	28	18	37	39	24	24	32	32	41	0	20
GdCepVB47	38	40	32	34	38	17	22	35	34	20	29	21	23	39	20	0



Clustering : additive distance tree (*Sattah & Tversky, 1977*)

A X-tree (additive distance tree) is an unrooted tree

- with the stimuli as leaves
- with internal vertices (nodes) of degree 3



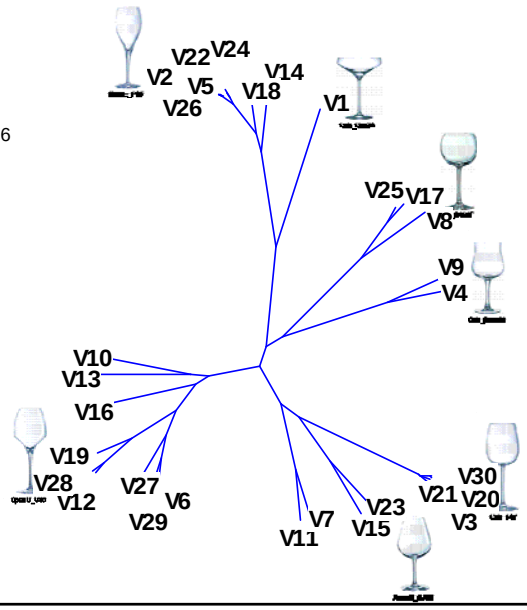
Topological criterion (« four points condition ») :

$$\forall a,b,c,d \quad d(a,b)+d(c,d) \le \max(d(a,c)+d(b,d)\,,\, d(a,d)+d(b,c))$$

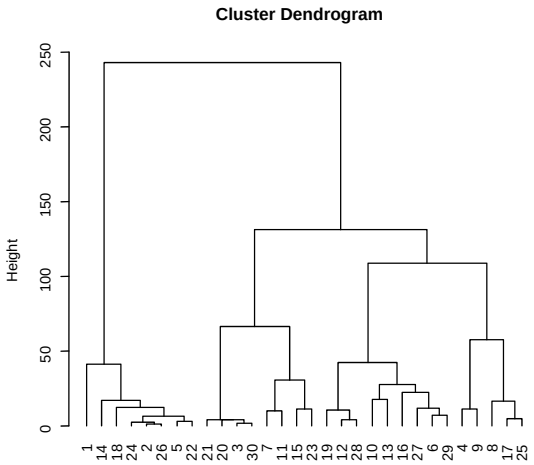
« Pauline's data » : connoisseurs

Statistics :
Least-squares coefficient
Sum (Dij-ADij)^2 = 2703.574590
Average absolute difference
Sum |Dij-ADij|/(n(n-1)/2) = 1.889076

ρ_s= 0.835
R^2=0.940



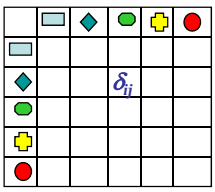
Alternative : the hierarchical clustering

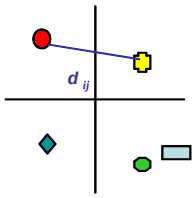


Less intuitive…
easier to cut…

Multidimensional Scaling (1)

(Kruskal, 1964 ; Borg & Groenen, 1997)





Observed dissimilarities

(number of times the pairs of products have not been grouped)

Euclidean representation in a low dimensional space

(defining latent variables)

Solution is obtained by minimizing a Stress Index (Kruskal)

$$\sqrt{\sum_{i,j}\left(\delta_{ij}-d_{ij}\right)^2}$$

Lack of fit

$$\sqrt{\frac{\sum_{i,j}\left(\delta_{ij}-d_{ij}\right)^2}{\sum_{i,j} d_{ij}^{\,2}}}$$

Multidimensional Scaling (2)

Ordinal interpretation of dissimilarities

Questionning the metric interpretation of observed dissimilarities !

We seek to represent more the ranking of dissimilarities than their values

Non metric MDS

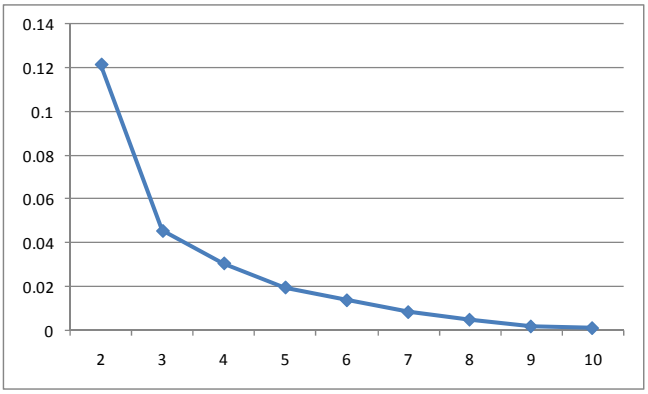
Non-metric Stress Index

$$\sqrt{\frac{\sum_{i,j}\left(f\left(\delta_{ij}\right)-d_{ij}\right)^2}{\sum_{i,j} d_{ij}^{\ 2}}}$$

Euclidean representation of $f(\delta_{ij})$: a monotone increasing transformation of δ_{ij}

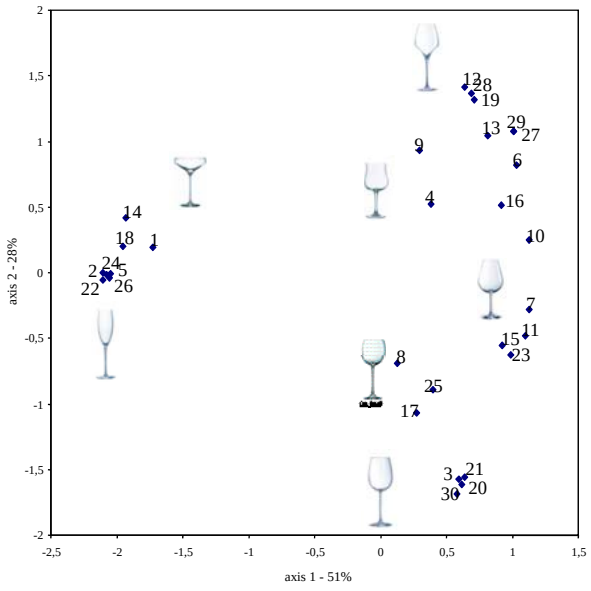
Choice of the dimensionality

Stress plot *(Shepard plot)*

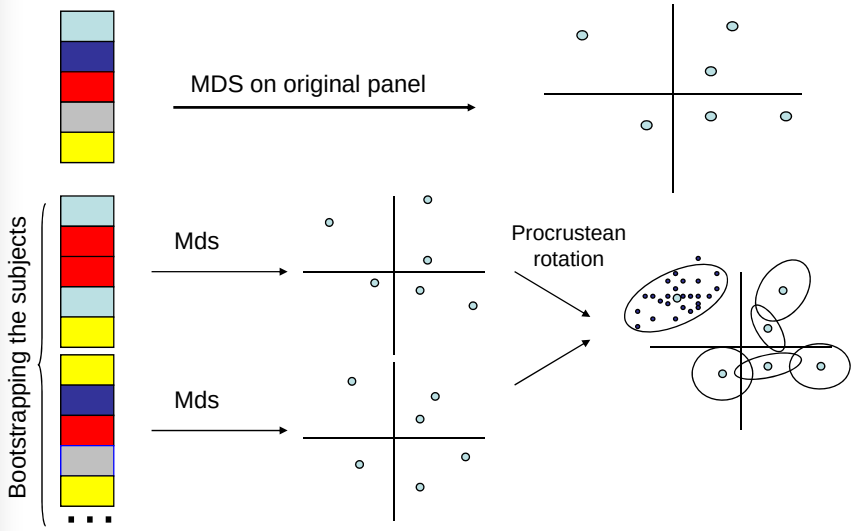


Stress Index (3 dim solution) = 0.046

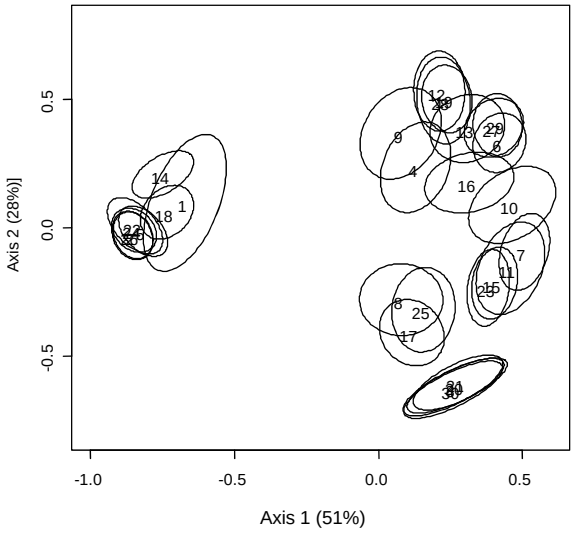
MDS solution (3 dimensions) for connaisseurs



Robustness of the configuration : resampling the subjects



Robustness of the consensus between subjects



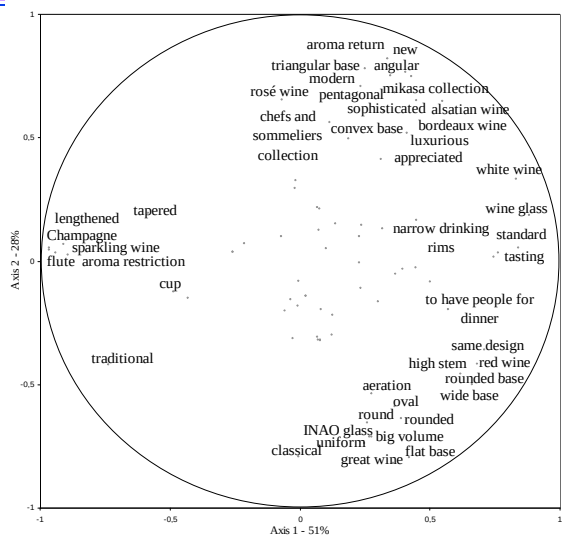
Bootstrapped MDS : ellipsoids (80%)

Interpretation of the configuration

Description of the groups by the subjects
(verbalization task)

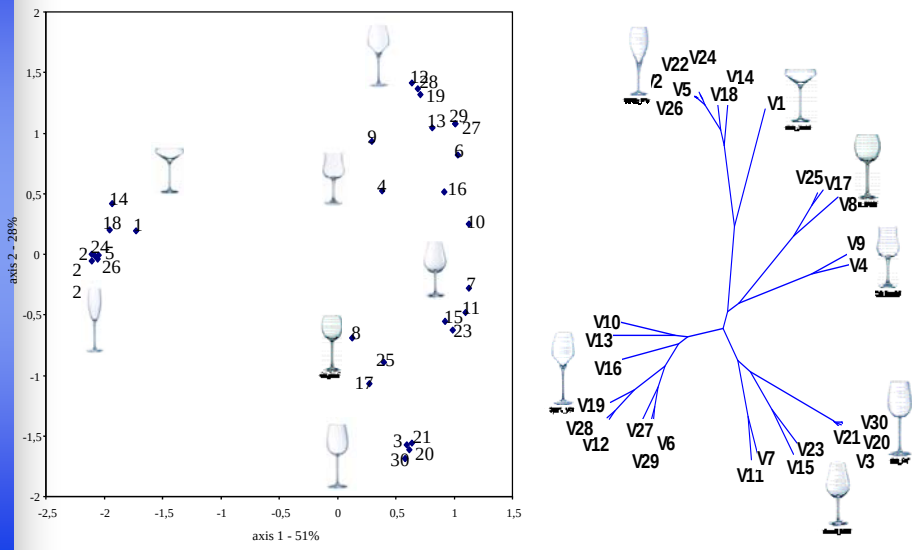
Correlations between frequency of a term and coordinates of MDS solution

Verbatims used by connoisseurs



Physical description + Usage

Coherence MDS / Additive tree



Factorial representation : alternatives

✓ **MDS with individual weighting**

non metric INDSCAL : *Takane (1977)*
DISTATIS : *Abdi et al. (2007)*

✓ **Methods using Khi² metric**

- **Multiple Correspondance Analysis (Homogeneity Analysis)**

  MCA : *Van der Kloot & Van Herk (1991)*
  MDSort : *Takane (1981)*
  FAST : *Cadoret et al. (2009)*

- **Individual weighting**

  IDSort : *Takane (1982)*
  CCSort : *Qannari et al. (2009)*

✓ **Is it possible to find a central partition of a group of subjects ?**

✓ **Is there any difference between the categorization realized by different groups of subjects ?**

✓ **Can we buid clusters of subjects sharing a common perception of stimuli ?**

Comparing sorts

Agreement (or distance) between partitions

Agreement between two partitions

N stimuli:
N(N-1)/2 pairs of objects

Partition $\mathbf{P_1}$ \ Partition $\mathbf{P_2}$	$=$	$\neq$
$=$	a	c
$\neq$	b	d

Pairs of objects :
a+ d : agreement
b + c : disagreement

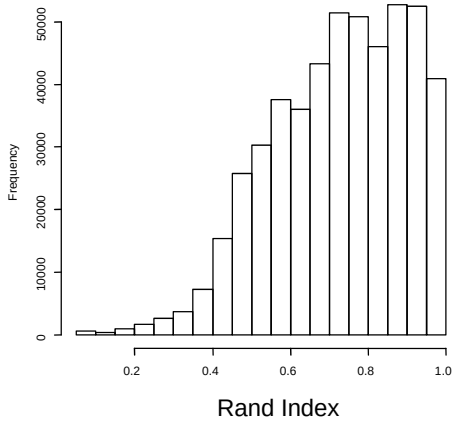
Rand Index
(Rand, 1971)

$$R(P_1, P_2) = \frac{a+d}{a+b+c+d} = \frac{a+d}{N(N-1)/2}$$

link with the symmetric difference : $1 - R(P_1, P_2) = \frac{S(P_1, P_2)}{N(N-1)/2}$

Agreement between two partitions

Simulation : 1000 independent random partitions of 30 items

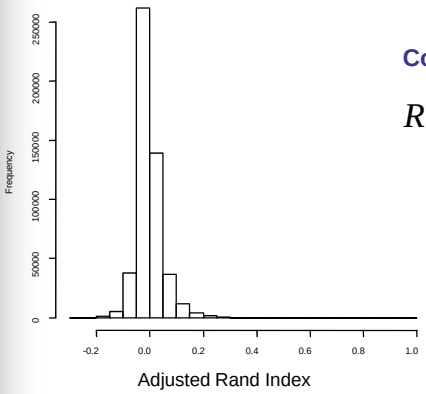


Mean = 0.726
$Q_{0.95}$ = 0.968

Agreement between two partitions

Adjusted Rand Index

(Hubert & Arabie, 1985)



Correction for chance agreement

$$R^a(P_1, P_2) = \frac{R(P_1, P_2) - R_{exp}}{R_{max} - R_{exp}}$$

mean= -0.000019

$Q_{0.95}$ = 0.0853

Distance between partitions P_1 and P_2
= 1 – Adjusted Rand between P_1 and P_2

Consensus partition

Let *P* be a set of partitions $P_1..P_K$

Objective : build a consensus partition *C*
maximizing the agreement with the elements of *P*
(i.e. minimizing the distance to the elements of *P)*

Régnier (1965) ; Marcotorchino & Michaud (1982) ; Barthélémy & Leclerc (1995), Gordon & Vichi (1998) ; Krieger & Green (1999) ; Guénoche (2011)

Consensus partition … or Central partition or Median partition

$$\underset{C}{Arg\,\min} \sum_{k=1}^{K} d(C, P_k)$$

Finding the consensus partition *C* with *T* classes maximizing the agreement with individual partitions $P_1..P_K$

$$\underset{C}{Arg\,\min}\sum_{k=1}^{K} d(C,P_k) \equiv \underset{C}{Arg\,\min}\sum_{k=1}^{K}\left(1-R^a(C,P_k)\right) \equiv \underset{C}{Arg\,\min}\left[K-\sum_{k=1}^{K}R^a(C,P_k)\right]$$

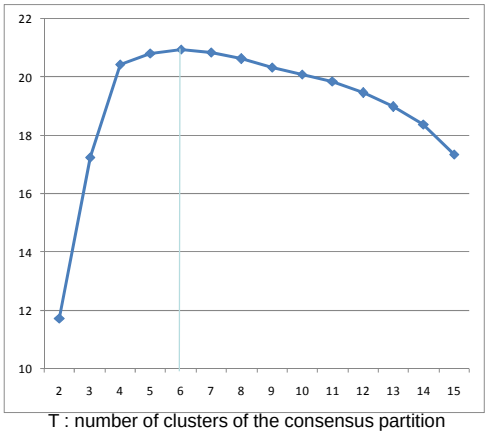
$$\equiv \underset{C}{Arg\,\max}\sum_{k=1}^{K}R^a(C,P_k)$$

- *Step 1*. Start with an initial partition *C* with *T* classes
- *Step 2*. Alternatively, try to merge the stimuli to each of the *T-1* other groups of *C*
  Move each stimulus to the group maximizing the criterion
  (transfert)
- *Step 3*. If no improvement of criterion is observed, stop. Otherwise, go to *step 2*.

Connoisseurs

6 groups
Criterion = 20.842
Mean of AdjRand = 0.508

Mean number of groups
5.95

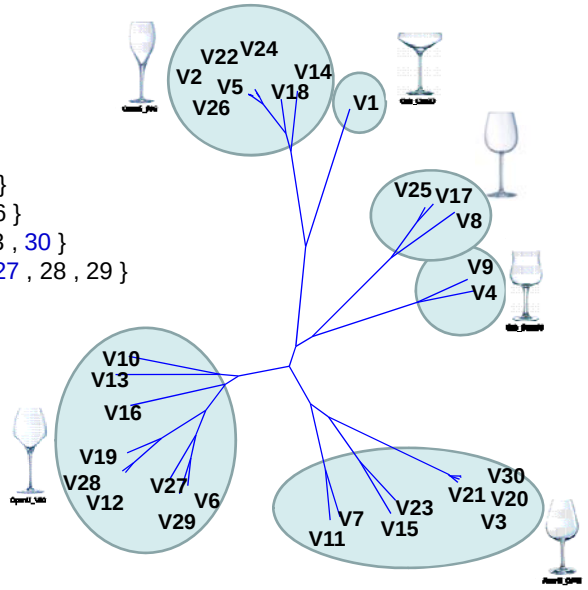


T : number of clusters of the consensus partition

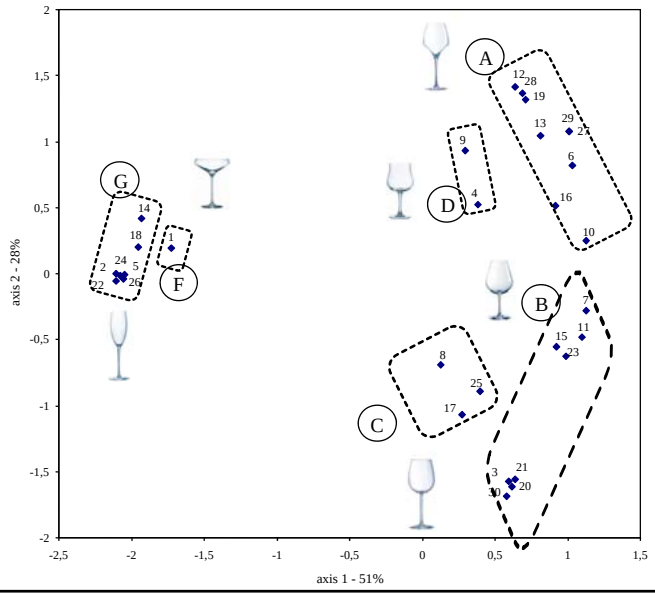
{ 1 } { 2 , 5 , 14 , 18 , 22 , 24 , 26 } { 4 , 9 } { 8 , 17 , 25 }
{ 3 , 7 , 11 , 15 , 20 , 21 , 23 , 30 }
{ 6 , 10 , 12 , 13 , 16 , 19 , 27 , 28 , 29 }

Consensus Partition / Additive tree

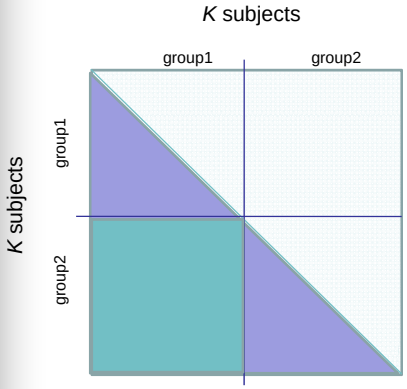
{ 1 } { 4 , 9 } { 8 , 17 , 25 }
{ 2 , 5 , 14 , 18 , 22 , 24 , 26 }
{ 3 , 7 , 11 , 15 , 20 , 21 , 23 , 30 }
{ 6 , 10 , 12 , 13 , 16 , 19 , 27 , 28 , 29 }



Consensus Partition / MDS



Comparing groups of subjects



Decomposition of the total sum of squares of distances

$$SS_T = SS_B + SS_W$$

$$SS_T = \frac{1}{K}\sum_{i<j} d_{ij}^2 \qquad SS_W = \sum_{g=1}^{G} \frac{1}{n_g} \sum_{i<j \in G_g} \delta_{ij}^2$$

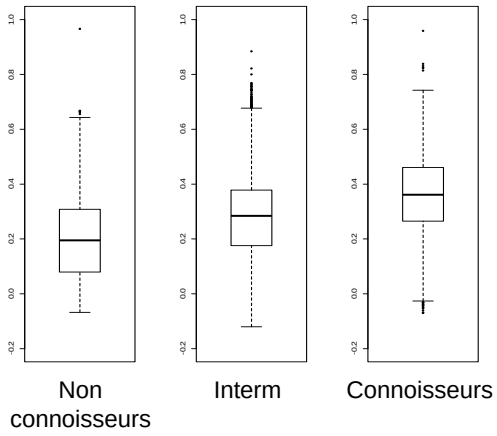
PERMANOVA : *McArdle & Anderson (2001)* MRPP : *Mielke & Berry (2001)*

Measure of the difference between groups

$$Pseudo - F = \frac{SS_B / (G-1)}{SS_W / (K-G)}$$

The higher this Pseudo-F, the greater the difference between groups

If there are no differences between groups,
the pseudo-F is likely to be close to 1



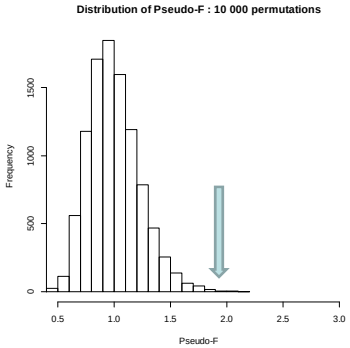
Adjusted Rand Index between subjects

	Non connoisseurs	Interm	Connoisseurs
Non connoisseurs	0.204		
Interm	0.265	0.279	
Connoisseurs	0.260	0.308	0.356

Mean Adjusted Rand Index between subjects

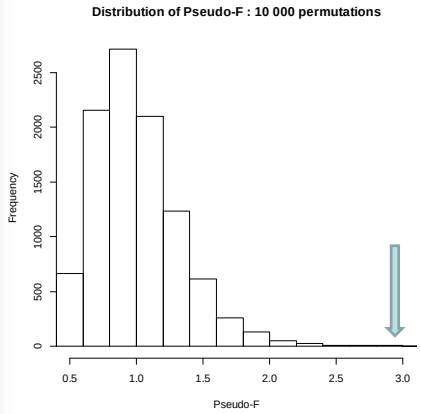
Connoisseurs / Interm / Non connoisseurs

Source	Df	SumsOfSqs	MeanSqs	F.Model	Pr(>F)
Class	2	1.056	0.52796	1.9329	0.00099
Residuals	206	56.269	0.27315		
Total	208	57.325			



Connoisseurs / Non connoisseurs

Source	Df	SumsOfSqs	MeanSqs	F.Model	Pr(>F)
Class	1	0.8192	0.81915	2.9813	0.0002
Residuals	80	21.9811	0.27476		
Total	81	22.8003			

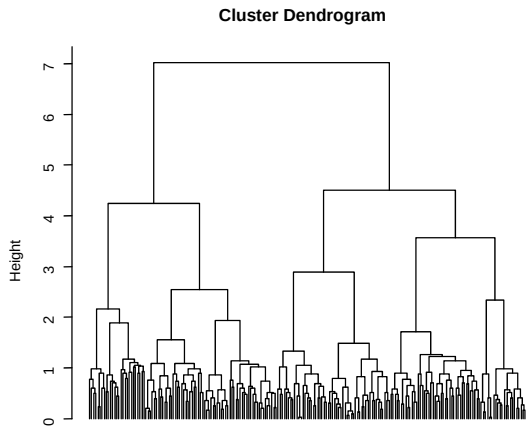


Partition of partitions : Clustering of partitions around consensus

- Step 1. Start with an initial partition of partitions
- *Step 2*. Computation of the consensus partition in each class
  Affectation of each partition to the class with the higher adjusted Rand.
- *Step 3*. If no improvement of criterion is observed, stop. Otherwise, go to *step 2*.

Initial partition

Hierarchical clustering of subjects
using Adjusted Rand Index distance



Dendrogram suggesting a partition with two clusters

Final solution…

➢Class 1.
123 subjects
Consensus : 6 groups
{ 1 } { 2 , 5 , 14 , 18 , 22 , 24 , 26 }
{ 3 , 7 , 11 , 15 , 20 , 21 , 23 , 30 } { 4 , 9 }
{ 6 , 10 , 12 , 13 , 16 , 19 , 27 , 28 , 29 }
{ 8 , 17 , 25 }

	Class 1	Class 2
Novices	17	24
Interm	74	53
Connoisseurs	32	9

➢Class 2.
86 subjects
Consensus : 10 groups
{ 1 } { 2 , 5 , 18 , 22 , 24 , 26 } { 3 , 20 , 21 , 30 }
{ 4 , 9 } { 6 , 10 , 13 } { 7 , 11 , 15 , 23 }
{ 8 , 17 , 25 } { 12 , 19 , 28 , 29 }
{ 14 } { 16 , 27 }

Many techniques…

… different representations

…but very coherent results

Softwares used :
- R Packages smacof and vegan
- T-Rex
- « home made » R functions