XXX Jornadas de Classificação e Análise de Dados

JOCLAD 2023

A comparison of some methods for clustering of variables of mixed types

Ndèye Niang, Mory Ouattara, Gilbert Saporta

le cnam

UNIVERSITÉ DE SAN-PEDRO
CÔTE D'IVOIRE

Outline

1. Introduction and related works
2. RV coefficient, similarities and distances
3. Other clustering strategies
   - 3.1 *p*-values (*likelihood linkage*)
   - 3.2 *z*-scores or normal quantiles (« *valeurs-test* »)
4. Application to indoor air pollution
5. Conclusion and perspectives

1. Introduction and related works

- Clustering variables: a neglected field
  - Few specific methods:
    - SAS ® Proc Varclus, a divisive algorithm
    - Likelihood linkage agglomerative methods (Costa Nicolau, F. and Bacelar-Nicolau, H. , 1998)
  - Even less methods for mixed types variables (categorical and numerical)

- More attention was paid to factor analysis of mixed data:

$$\max_{\mathbf{c}} \left(\sum_{j=1}^{J} r^2\left(\mathbf{c}, \mathbf{x}_j\right) + \sum_{q=1}^{Q} \eta^2\left(\mathbf{c}, \tilde{\mathbf{x}}_q\right) \right)$$

Tenenhaus (1977), Escofier (1979), Saporta (1990), Kiers (1991), Pagès (2004)

- Clustering variables around latent components:
  - ClustVarLV (Vigneau & Qannari, 2003) for numerical variables
  - ClustOfVar (Chavent et al. 2012) for a mix of numerical and categorical variables inspired by PCAMIX (Kiers, 1991)

2. RV coefficient, similarities and distances

- A particular Escoufier operator: projector associated to a data table $\mathbf{X}_j$

$$\mathbf{W}_j = \mathbf{X}_j \left(\mathbf{X}_j' \mathbf{X}_j \right)^{-1} \mathbf{X}_j'$$

- RV between two data sets (Robert & Escoufier, 1976)

$$RV = \frac{Trace\left(\mathbf{W}_1 \mathbf{W}_2 \right)}{\sqrt{Trace\left(\mathbf{W}_1 \right)^2 Trace\left(\mathbf{W}_2 \right)^2}}$$

- Two cases :
  - $\mathbf{X}$ associated to a single numerical variable: a unicolumn matrix $\mathbf{x}$
  - $\mathbf{X}$ associated to a categorical variable with m categories : a binary indicator matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ . & . & . \\ 0 & 1 & 0 \end{pmatrix}$$

- Three RV coefficients:
  - Between two numerical variables: $RV = r^2$
  - Between a numerical and a categorical variable: $RV = \frac{\eta^2}{\sqrt{m-1}}$
  - Between two categorical variables:

$$RV = \frac{\phi^2}{\sqrt{(m_1 - 1)(m_2 - 1)}}$$

Tschuprow's T^2

- RV coefficients provide comparable measurements between numerical and categorical variables
- Matrix of RV coefficients is positive definite like a correlation matrix
- 1- RV is an euclidean distance between variables of mixed types.

  NB all values between 0 and 1

- Hence the possibility to perform euclidean clusterings by *k*-means or Ward's hierarchical algorithm (Qannari, Vigneau & Courcoux, 1998)

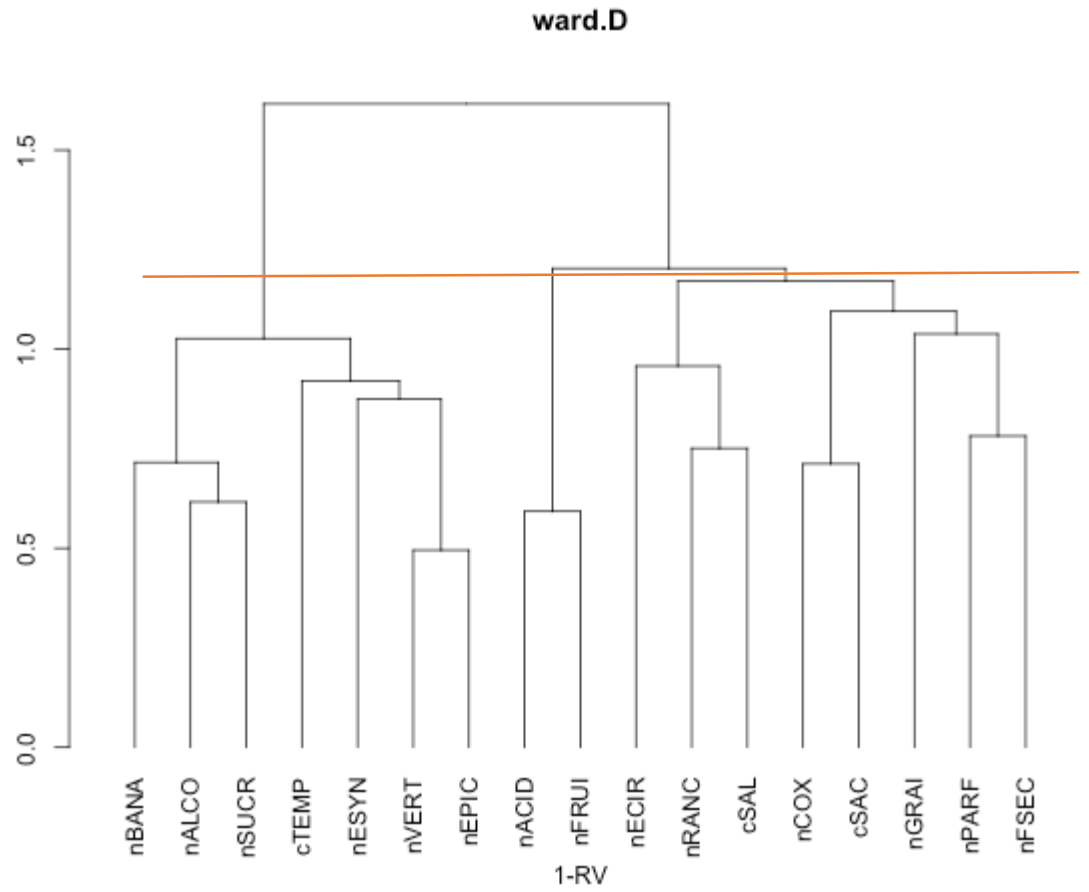
A toy example (Qannari, Vigneau & Courcoux, 1998)

- Sensory evaluation of the aromatic characteristics of 26 lots of apples of the Cox variety (14 quantitative descriptors) under different storage conditions (3 qualitative variables).

Descripteurs d'arôme		Conditions de stockage	
n°	libellé	n°	libellé
1	ALCO alcoolisé	15	SAC stockage atmosph. contrôlée
2	VERT vert		SAC1 0 semaine
3	ACID acide		SAC2 10 semaines
4	PARF parfumé		SAC3 18 semaines
5	COX typique de la variété Cox		SAC4 25 semaines
6	BANA banane		SAC5 31 semaines
7	FRUI fruité	16	SAL stockage à l'air libre
8	ESYN ester synthétique		SAL1 0 jours
9	ECIR ester cireux		SAL2 6-10 jours
10	SUCR sucré		SAL3 15-19 jours
11	EPIC épicé		SAL4 25-32 jours
12	FSEC feuilles sèches		SAL5 39, 48 ou 51 jours
13	GRAI graisseux	17	TEMP température (air libre)
14	RANC rance		TEMP1 11°C
			TEMP2 17°C

The distance matrix of $1\text{-}RV$ coefficients

	nALCO	nVERT	nACID	nPARF	nCOX	nBANA	nFRUI	nESYN	nECIR	nSUCR	nEPIC	nFSEC	nGRAI	nRANC	cSAC	cSAL	cTEMP
nALCO	0,00	0,64	0,98	0,94	1,00	0,69	1,00	0,77	0,96	0,62	0,67	0,96	0,98	0,96	0,83	0,95	0,79
nVERT	0,64	0,00	0,97	0,87	0,98	0,76	1,00	0,74	0,99	0,83	0,50	0,90	1,00	0,93	0,93	0,93	0,89
nACID	0,98	0,97	0,00	0,91	0,98	0,96	0,59	0,92	0,97	0,99	1,00	0,97	1,00	0,94	0,89	0,98	0,99
nPARF	0,94	0,87	0,91	0,00	1,00	1,00	0,86	0,89	1,00	0,96	0,82	0,78	0,98	0,98	0,92	0,96	0,96
nCOX	1,00	0,98	0,98	1,00	0,00	0,90	0,94	0,95	0,97	0,92	1,00	0,82	0,99	0,99	0,71	0,95	0,99
nBANA	0,69	0,76	0,96	1,00	0,90	0,00	0,92	0,93	1,00	0,69	0,86	0,78	1,00	0,99	0,95	0,87	0,80
nFRUI	1,00	1,00	0,59	0,86	0,94	0,92	0,00	0,92	0,76	1,00	0,98	1,00	0,94	0,98	0,86	0,92	0,99
nESYN	0,77	0,74	0,92	0,89	0,95	0,93	0,92	0,00	0,94	0,82	0,82	0,88	1,00	0,94	0,77	0,94	0,89
nECIR	0,96	0,99	0,97	1,00	0,97	1,00	0,76	0,94	0,00	0,98	0,96	0,99	1,00	1,00	0,84	0,81	0,93
nSUCR	0,62	0,83	0,99	0,96	0,92	0,69	1,00	0,82	0,98	0,00	0,88	0,98	1,00	0,92	0,90	0,93	0,95
nEPIC	0,67	0,50	1,00	0,82	1,00	0,86	0,98	0,82	0,96	0,88	0,00	0,82	0,97	0,94	0,91	0,96	0,75
nFSEC	0,96	0,90	0,97	0,78	0,82	0,78	1,00	0,88	0,99	0,98	0,82	0,00	0,97	0,98	0,95	0,95	0,84
nGRAI	0,98	1,00	1,00	0,98	0,99	1,00	0,94	1,00	1,00	1,00	0,97	0,97	0,00	1,00	0,94	0,97	1,00
nRANC	0,96	0,93	0,94	0,98	0,99	0,99	0,98	0,94	1,00	0,92	0,94	0,98	1,00	0,00	0,97	0,75	0,93
cSAC	0,83	0,93	0,89	0,92	0,71	0,95	0,86	0,77	0,84	0,90	0,91	0,95	0,94	0,97	0,00	0,86	0,97
cSAL	0,95	0,93	0,98	0,96	0,95	0,87	0,92	0,94	0,81	0,93	0,96	0,95	0,97	0,75	0,86	0,00	0,95
cTEMP	0,79	0,89	0,99	0,96	0,99	0,80	0,99	0,89	0,93	0,95	0,75	0,84	1,00	0,93	0,97	0,95	0,00


ward.D
1.5
1.0
0.5
0.0
nBANA
nALCO
nSUCR
cTEMP
nESYN
nVERT
nEPIC
nACID
nFRUI
nECIR
nRANC
cSAL
nCOX
cSAC
nGRAI
nPARF
nFSEC
1-RV


Torgerson scaling
nPARF
nFSEC
nEPIC
nVERT
cTEMP
nESYN
nGRAI
nACID
nFRUI
nALCO
nBANA
nSUCR
cSAC
nCOX
nECIR
nRANC
cSAL

Cautions

- In factor analysis and clustering around components, categorical variables with a large number of modalities may have more influence than binary or numerical variables
- RV is less sensitive, but dividing by the square root of the degree of freedom does not completely correct the effect of the number of categories

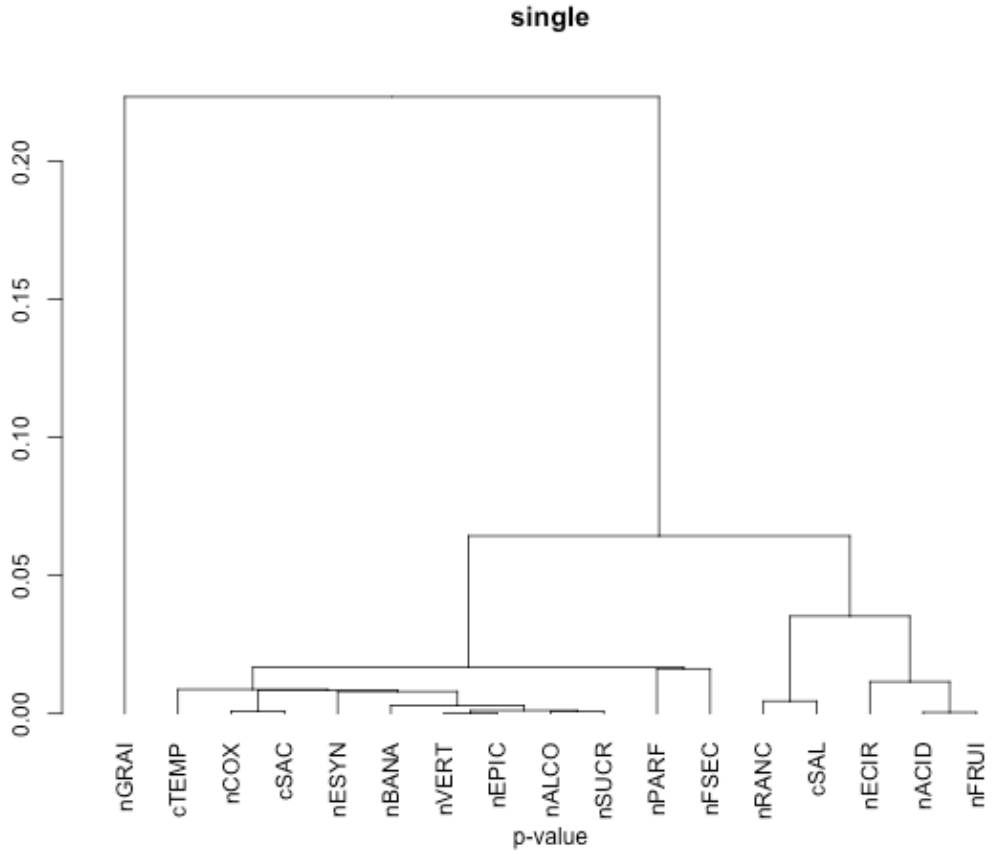
3. Other clustering strategies

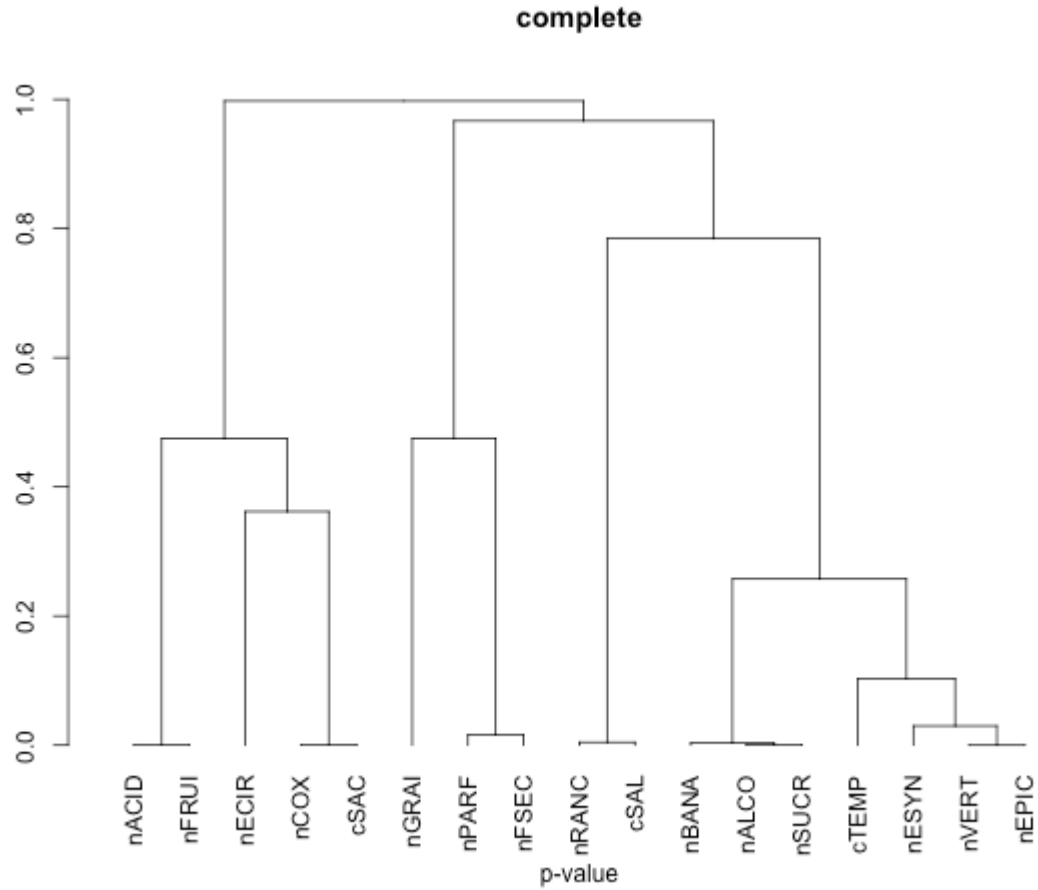
3.1 *p*-values

- Change the $1\text{-}RV$ coefficients into the *p*-values of the independence tests associated to r^2 , χ^2 (for square Tschuprow), or Anova F (for η^2)
- *p*-values no longer depend on the degrees of freedom
- *p*-values provide a dissimilarity matrix
- Complete linkage algorithm is close to the AVL probabilistic approach (Lerman, 1973 and Nicolau & Bacelar-Nicolau, 1998)

P-value matrix

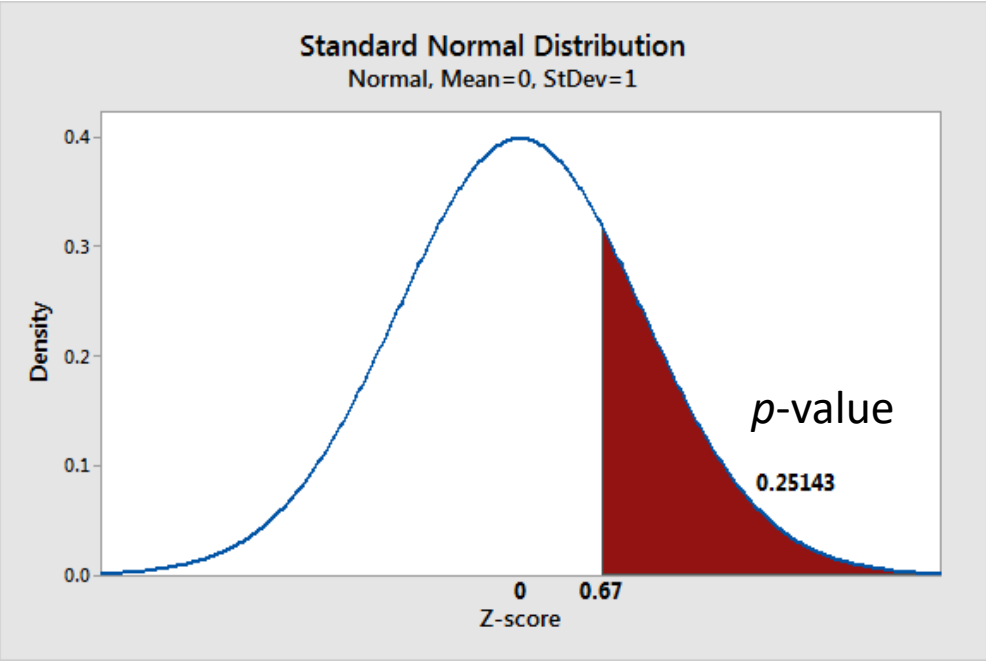
	nALCO	nVERT	nACID	nPARF	nCOX	nBANA	nFRUI	nESYN	nECIR	nSUCR	nEPIC	nFSEC	nGRAI	nRANC	cSAC	cSAL	cTEMP
nALCO	0,00	0,00	0,48	0,22	0,92	0,00	0,98	0,01	0,35	0,00	0,00	0,30	0,53	0,35	0,06	0,70	0,02
nVERT	0,00	0,00	0,38	0,07	0,53	0,01	0,75	0,01	0,56	0,04	0,00	0,12	0,82	0,18	0,50	0,49	0,10
nACID	0,48	0,38	0,00	0,15	0,47	0,32	0,00	0,17	0,44	0,72	0,93	0,39	0,74	0,21	0,24	0,94	0,66
nPARF	0,22	0,07	0,15	0,00	1,00	0,79	0,06	0,10	0,77	0,31	0,03	0,02	0,48	0,52	0,43	0,75	0,35
nCOX	0,92	0,53	0,47	1,00	0,00	0,12	0,21	0,27	0,36	0,16	0,79	0,03	0,68	0,58	0,00	0,64	0,70
nBANA	0,00	0,01	0,32	0,79	0,12	0,00	0,17	0,18	0,85	0,00	0,06	0,02	0,82	0,59	0,64	0,15	0,02
nFRUI	0,98	0,75	0,00	0,06	0,21	0,17	0,00	0,16	0,01	0,78	0,46	0,93	0,22	0,51	0,12	0,40	0,61
nESYN	0,01	0,01	0,17	0,10	0,27	0,18	0,16	0,00	0,24	0,03	0,03	0,08	0,79	0,22	0,01	0,59	0,09
nECIR	0,35	0,56	0,44	0,77	0,36	0,85	0,01	0,24	0,00	0,45	0,35	0,70	0,96	0,97	0,08	0,04	0,18
nSUCR	0,00	0,04	0,72	0,31	0,16	0,00	0,78	0,03	0,45	0,00	0,08	0,46	0,93	0,17	0,27	0,53	0,26
nEPIC	0,00	0,00	0,93	0,03	0,79	0,06	0,46	0,03	0,35	0,08	0,00	0,03	0,38	0,24	0,34	0,78	0,01
nFSEC	0,30	0,12	0,39	0,02	0,03	0,02	0,93	0,08	0,70	0,46	0,03	0,00	0,40	0,46	0,69	0,69	0,04
nGRAI	0,53	0,82	0,74	0,48	0,68	0,82	0,22	0,79	0,96	0,93	0,38	0,40	0,00	0,75	0,61	0,83	0,97
nRANC	0,35	0,18	0,21	0,52	0,58	0,59	0,51	0,22	0,97	0,17	0,24	0,46	0,75	0,00	0,89	0,00	0,20
cSAC	0,06	0,50	0,24	0,43	0,00	0,64	0,12	0,01	0,08	0,27	0,34	0,69	0,61	0,89	0,00	0,56	0,79
cSAL	0,70	0,49	0,94	0,75	0,64	0,15	0,40	0,59	0,04	0,53	0,78	0,69	0,83	0,00	0,56	0,00	0,67
cTEMP	0,02	0,10	0,66	0,35	0,70	0,02	0,61	0,09	0,18	0,26	0,01	0,04	0,97	0,20	0,79	0,67	0,00


single
0.00
0.05
0.10
0.15
0.20
nGRAI
cTEMP
nCOX
cSAC
nESYN
nBANA
nVERT
nEPIC
nALCO
nSUCR
nPARF
nFSEC
nRANC
cSAL
nECIR
nACID
nFRUI
p-value


complete
0.0
0.2
0.4
0.6
0.8
1.0
nACID
nFRUI
nECIR
nCOX
cSAC
nGRAI
nPARF
nFSEC
nRANC
cSAL
nBANA
nALCO
nSUCR
cTEMP
nESYN
nVERT
nEPIC
p-value

3.2 *z*-scores or normal fractiles (« *valeurs-test* »)

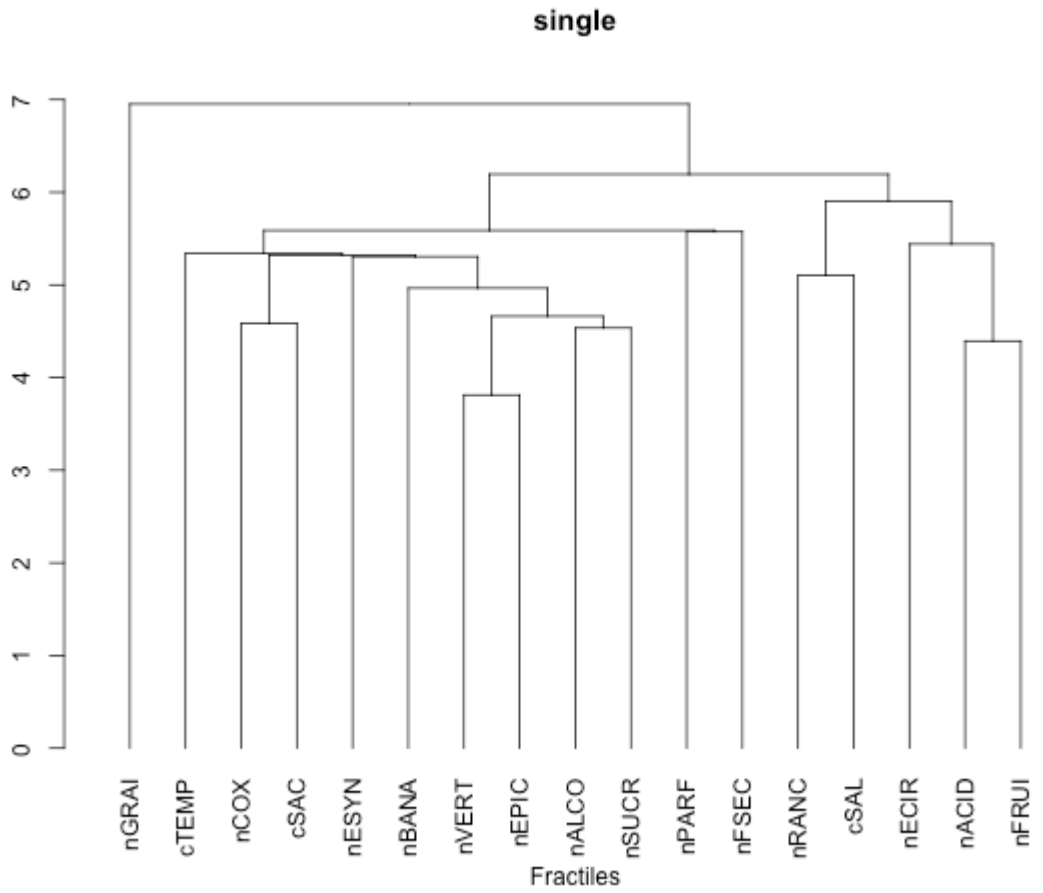
- When the number of observations is very large, all *p*-values are close to zero (paradox of large samples) and are no longer usable.
- Our proposal: replace the *p*-values by the associated normal *z*-score

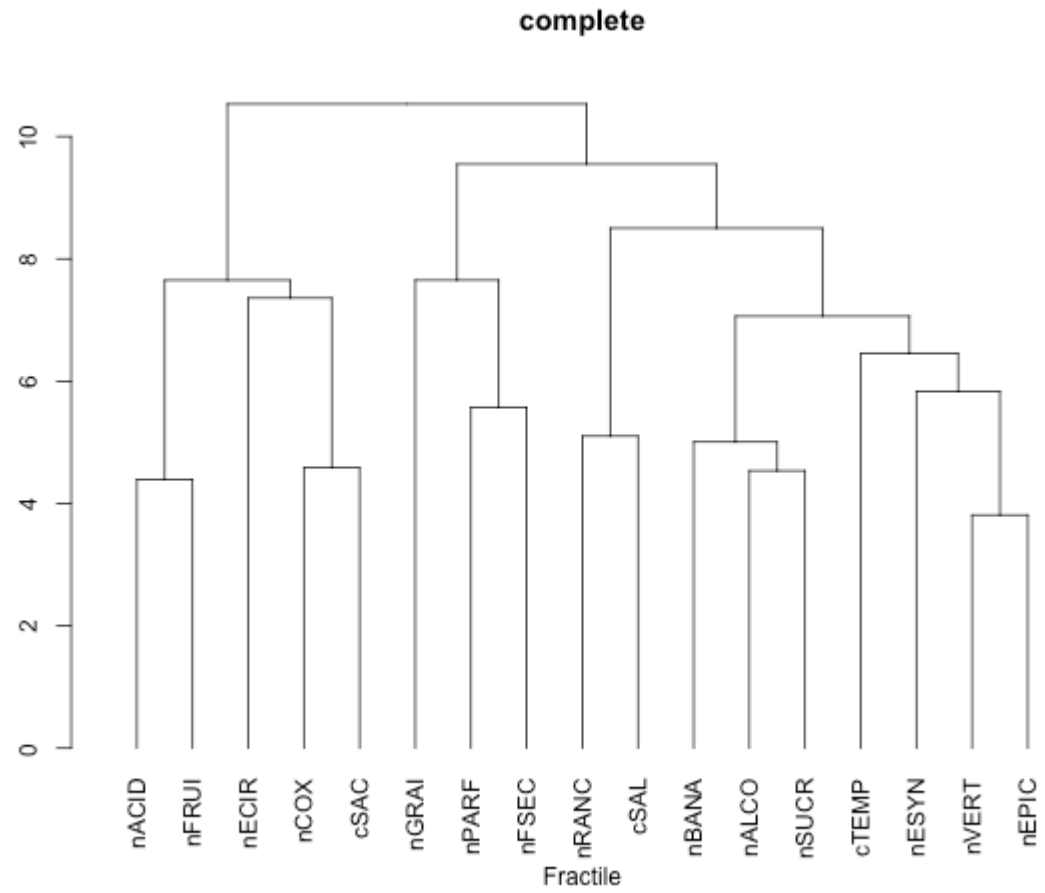


A universal measure of the strength of a link expressed in number of standard deviations
« test value » of SPAD (Morineau,1991)
z-scores are dissimilarities

Z-scores or normal fractiles

	nALCO	nVERT	nACID	nPARF	nCOX	nBANA	nFRUI	nESYN	nECIR	nSUCR	nEPIC	nFSEC	nGRAI	nRANC	cSAC	cSAL	cTEMP
nALCO	0,00	4,66	7,66	6,95	9,11	4,97	9,70	5,52	7,32	4,54	4,84	7,19	7,78	7,32	6,18	8,24	5,63
nVERT	4,66	0,00	7,41	6,23	7,80	5,44	8,39	5,31	7,86	5,95	3,81	6,53	8,63	6,81	7,72	7,69	6,45
nACID	7,66	7,41	0,00	6,66	7,65	7,26	4,39	6,74	7,56	8,31	9,19	7,42	8,37	6,91	6,99	9,29	8,12
nPARF	6,95	6,23	6,66	0,00	10,53	8,52	6,19	6,44	8,44	7,23	5,83	5,57	7,65	7,75	7,53	8,39	7,33
nCOX	9,11	7,80	7,65	10,53	0,00	6,55	6,92	7,12	7,36	6,70	8,52	5,85	8,18	7,93	4,58	8,08	8,25
nBANA	4,97	5,44	7,26	8,52	6,55	0,00	6,76	6,81	8,77	5,01	6,12	5,59	8,62	7,93	8,08	6,67	5,70
nFRUI	9,70	8,39	4,39	6,19	6,92	6,76	0,00	6,73	5,44	8,50	7,62	9,19	6,95	7,74	6,52	7,46	7,99
nESYN	5,52	5,31	6,74	6,44	7,12	6,81	6,73	0,00	7,00	5,87	5,83	6,34	8,51	6,93	5,32	7,95	6,39
nECIR	7,32	7,86	7,56	8,44	7,36	8,77	5,44	7,00	0,00	7,60	7,34	8,24	9,52	9,58	6,32	5,91	6,78
nSUCR	4,54	5,95	8,31	7,23	6,70	5,01	8,50	5,87	7,60	0,00	6,30	7,62	9,21	6,77	7,11	7,80	7,06
nEPIC	4,84	3,81	9,19	5,83	8,52	6,12	7,62	5,83	7,34	6,30	0,00	5,84	7,41	7,01	7,29	8,50	5,34
nFSEC	7,19	6,53	7,42	5,57	5,85	5,59	9,19	6,34	8,24	7,62	5,84	0,00	7,45	7,61	8,21	8,21	5,96
nGRAI	7,78	8,63	8,37	7,65	8,18	8,62	6,95	8,51	9,52	9,21	7,41	7,45	0,00	8,40	8,00	8,66	9,55
nRANC	7,32	6,81	6,91	7,75	7,93	7,93	7,74	6,93	9,58	6,77	7,01	7,61	8,40	0,00	8,92	5,10	6,87
cSAC	6,18	7,72	6,99	7,53	4,58	8,08	6,52	5,32	6,32	7,11	7,29	8,21	8,00	8,92	0,00	7,87	8,51
cSAL	8,24	7,69	9,29	8,39	8,08	6,67	7,46	7,95	5,91	7,80	8,50	8,21	8,66	5,10	7,87	0,00	8,15
cTEMP	5,63	6,45	8,12	7,33	8,25	5,70	7,99	6,39	6,78	7,06	5,34	5,96	9,55	6,87	8,51	8,15	3,17


single
0
1
2
3
4
5
6
7
nGRAI
cTEMP
nCOX
cSAC
nESYN
nBANA
nVERT
nEPIC
nALCO
nSUCR
nPARF
nFSEC
nRANC
cSAL
nECIR
nACID
nFRUI
Fractiles


complete
0
2
4
6
8
10
nACID
nFRUI
nECIR
nCOX
cSAC
nGRAI
nPARF
nFSEC
nRANC
cSAL
nBANA
nALCO
nSUCR
cTEMP
nESYN
nVERT
nEPIC
Fractile

Spearman's correlations between ultrametrics

	Dsingle	Dcomplete	Daverage	Dward.D	PVsingle	PVcomplete	PVaverage	Fsingle	Fcomplete	Faverage
Dsingle	1	0,396	0,802	0,387	0,993	0,396	0,721	0,993	0,396	0,729
Dcomplete	0,396	1	0,607	0,488	0,349	1	0,774	0,349	1	0,778
Daverage	0,802	0,607	1	0,505	0,766	0,607	0,903	0,766	0,607	0,906
Dward.D	0,387	0,488	0,505	1	0,347	0,488	0,395	0,347	0,488	0,399
PVsingle	0,993	0,349	0,766	0,347	1	0,349	0,687	1	0,349	0,692
PVcomplete	0,396	1	0,607	0,488	0,349	1	0,774	0,349	1	0,778
PVaverage	0,721	0,774	0,903	0,395	0,687	0,774	1	0,687	0,774	0,997
Fsingle	0,993	0,349	0,766	0,347	1	0,349	0,687	1	0,349	0,692
Fcomplete	0,396	1	0,607	0,488	0,349	1	0,774	0,349	1	0,778
Faverage	0,729	0,778	0,906	0,399	0,692	0,778	0,997	0,692	0,778	1

Rand index between partitions

	ClDistAver	ClDistCom	ClDistSing	ClDistWar	ClFractileA	ClFractileC	ClFractileS	ClPVAvera	ClPVComp	ClPVSingle
ClDistAverage	1	0,92	0,67	0,89	0,76	0,85	0,67	0,76	0,85	0,67
ClDistComplete	0,92	1	0,75	0,97	0,84	0,93	0,75	0,84	0,93	0,75
ClDistSingle	0,67	0,75	1	0,72	0,82	0,71	1	0,82	0,71	1
ClDistWard	0,89	0,97	0,72	1	0,81	0,93	0,72	0,81	0,93	0,72
ClFractileAverage	0,76	0,84	0,82	0,81	1	0,88	0,82	1	0,88	0,82
ClFractileComplete	0,85	0,93	0,71	0,93	0,88	1	0,71	0,88	1	0,71
ClFractileSingle	0,67	0,75	1	0,72	0,82	0,71	1	0,82	0,71	1
ClPVAverage	0,76	0,84	0,82	0,81	1	0,88	0,82	1	0,88	0,82
ClPVComplete	0,85	0,93	0,71	0,93	0,88	1	0,71	0,88	1	0,71
ClPVSingle	0,67	0,75	1	0,72	0,82	0,71	1	0,82	0,71	1

4. Application to indoor air pollution

Observatoire de la qualité de l'air intérieur

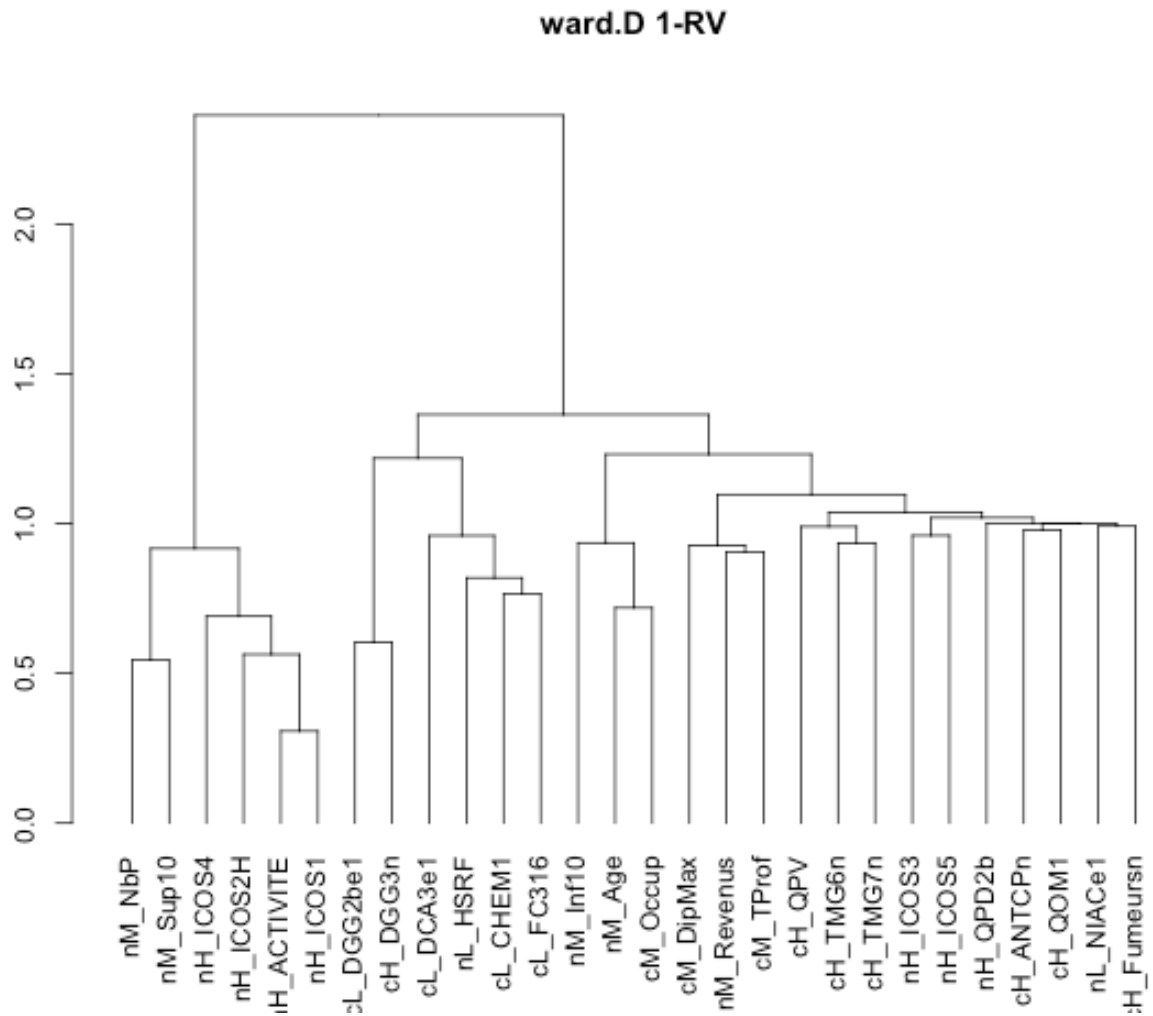
- A sample of 567 dwellings representative of the 24 million primary residences in mainland France collected by The Indoor Air Quality Observatory (OQAI).
- Objective: to draw up a state of the air pollution in the habitat
- 28 variables collected according to 3 criteria (blocks):
  - Housing (L) : 6 variables (2 numerical, 4 categorical)
  - Household (M) : 8 variables (5 numerical, 3 categorical)
  - Habit (H): 14 variables (7 numerical, 7 categorical)
- The names of the variables begin with **n** for the numerical, variables and **c** for the categorical variables followed by the letter of the block (L, M or H).

- Examples:

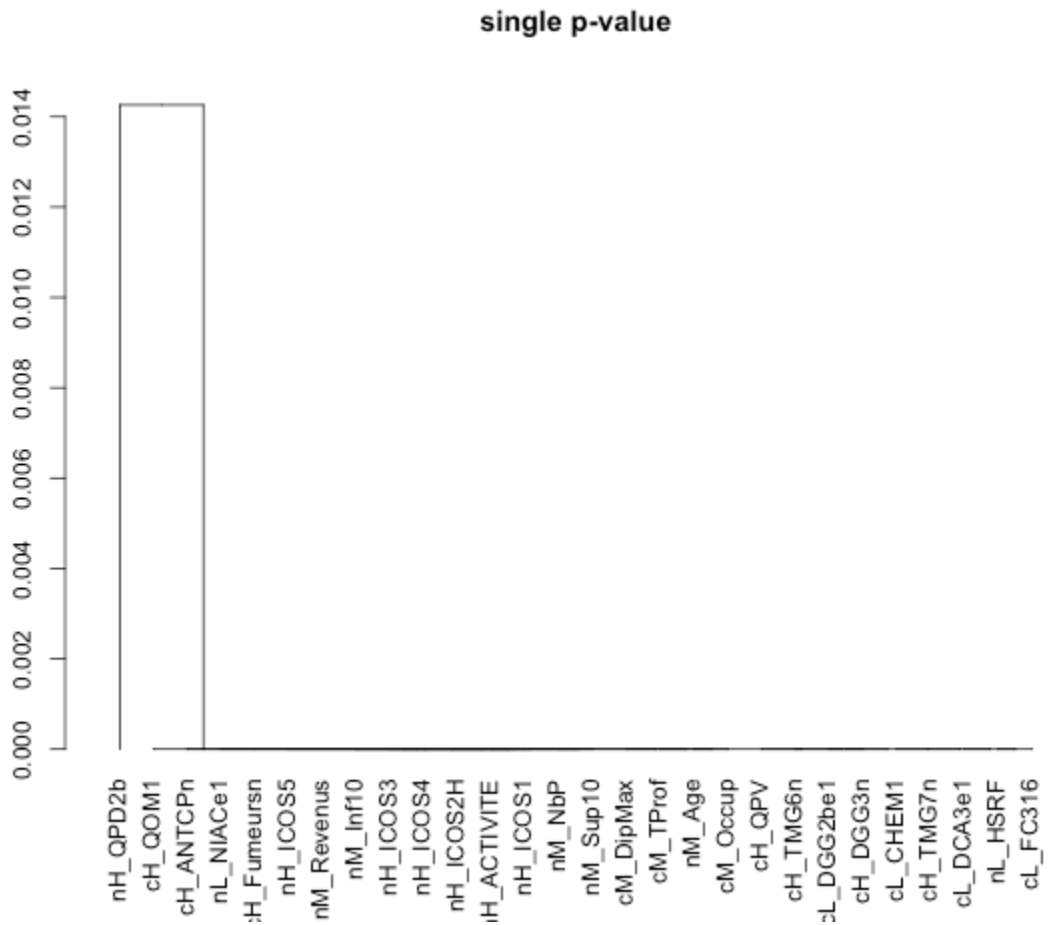
- **nL_HSRF** : What is the average height of the rooms in your dwelling, apart from the professional rooms and the annex rooms? (cm)
- **nL_NIACe1** : Age of the building or dwelling
- **cL_CHEM1** : Does your home have a chimney (Yes or No)…

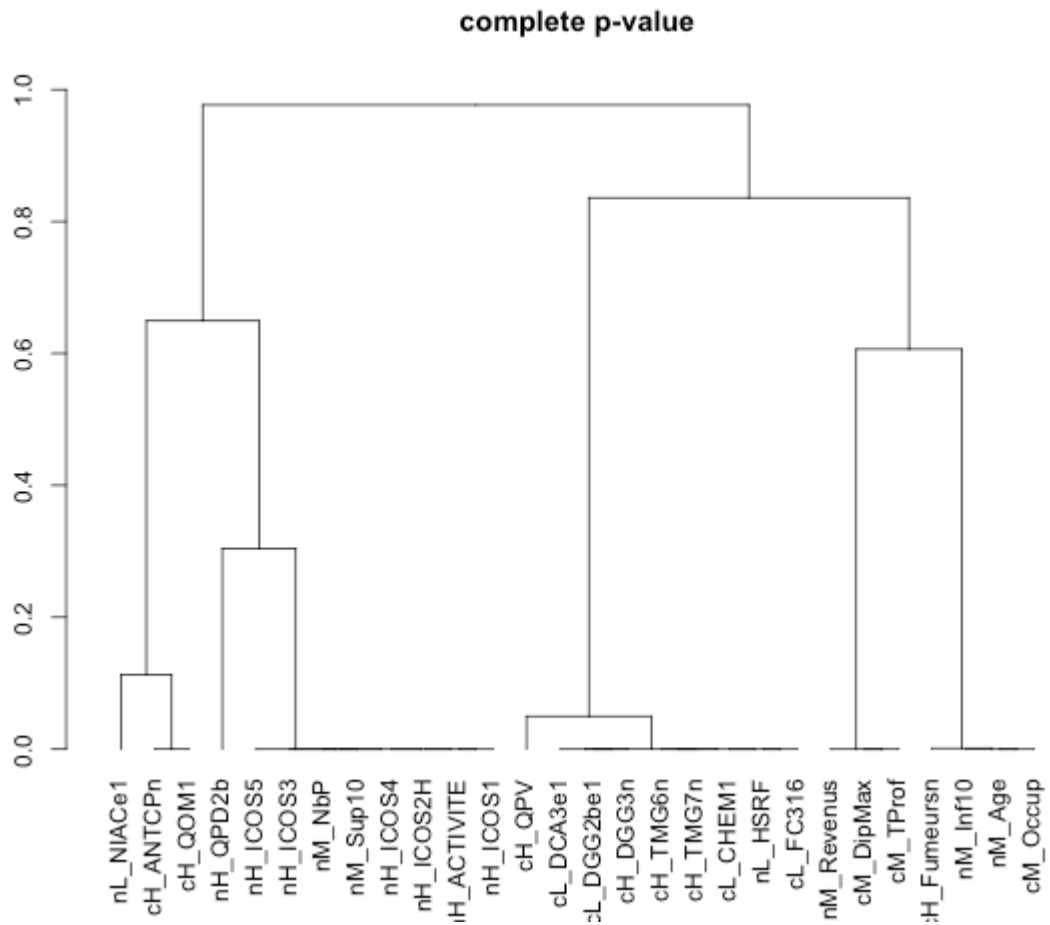
- **nM_Age :** Age of the person (in years)
- **nM_Revenu :** Household income
- **nM_Inf10 :** Number of children under 10 years old
- **nM_NbP :** Number of people
- **cM_DipMax** : Highest degree obtained
- **cM_TProf** : Type of occupation

- **nH_ACTIVITE** : frequency of general cleaning
- **nH_ICOS1 à nH_ICOS5 :** Various forms of cleaning
- **cH_QPD2b** : During the week, did you use any OTHER TYPE OF DEODORIZER in your home?
- **cH_Fumeursn** : Do any members of the household smoke inside your home?
- **cH_QPV** : How many plants do you have in your home?

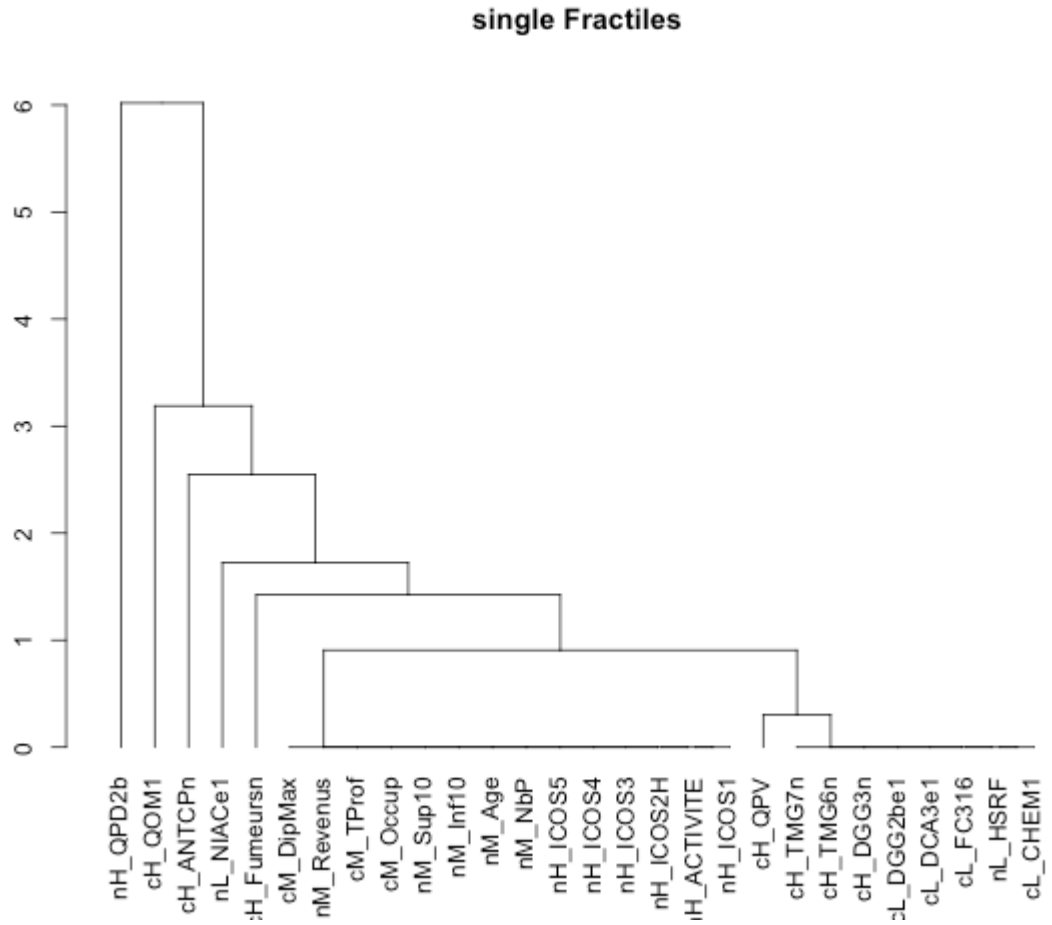

ward.D 1-RV
0.0
0.5
1.0
1.5
2.0
nM_NbP
nM_Sup10
nH_ICOS4
nH_ICOS2H
H_ACTIVITE
nH_ICOS1
L_DGG2be1
cH_DGG3n
cL_DCA3e1
nL_HSRF
cL_CHEM1
cL_FC316
nM_Inf10
nM_Age
cM_Occup
cM_DipMax
nM_Revenus
cM_TProf
cH_QPV
cH_TMG6n
cH_TMG7n
nH_ICOS3
nH_ICOS5
nH_QPD2b
cH_ANTCPn
cH_QOM1
nL_NIACe1
H_Fumeursn

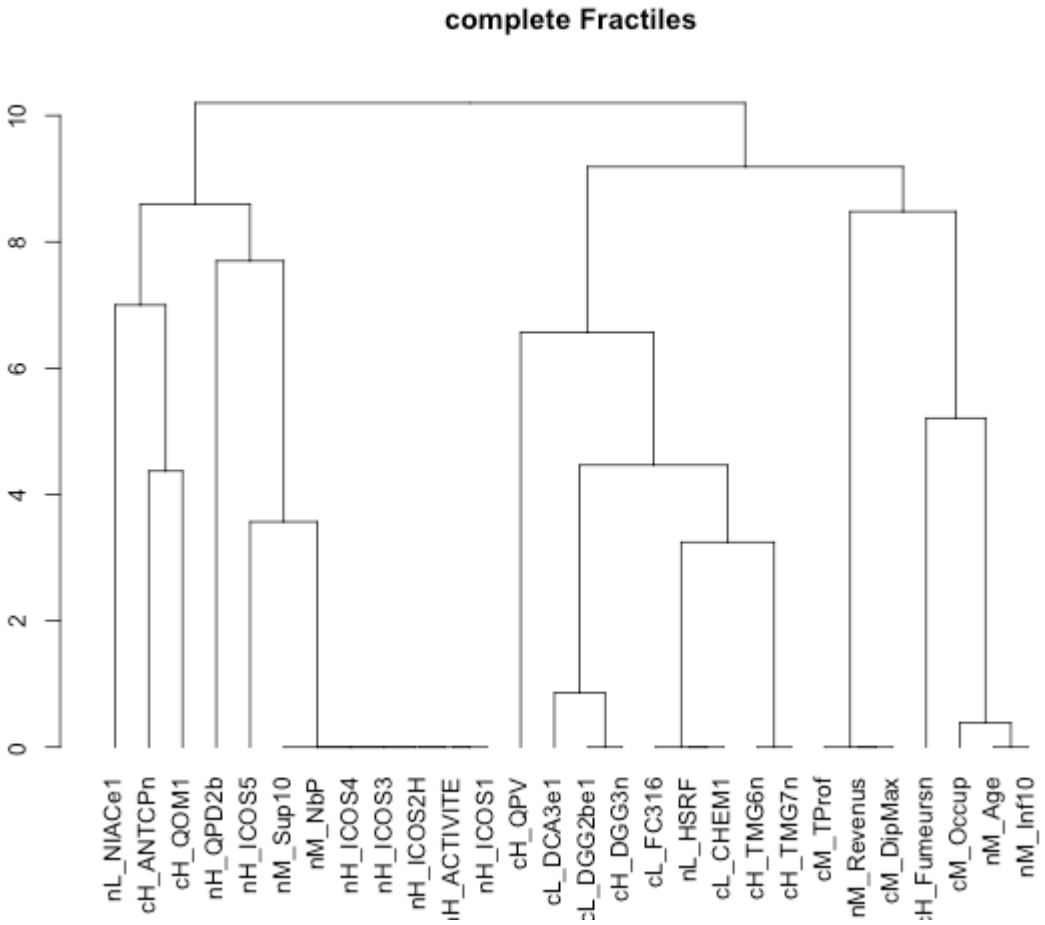
Single and complete linkage on p-values dissimilarities

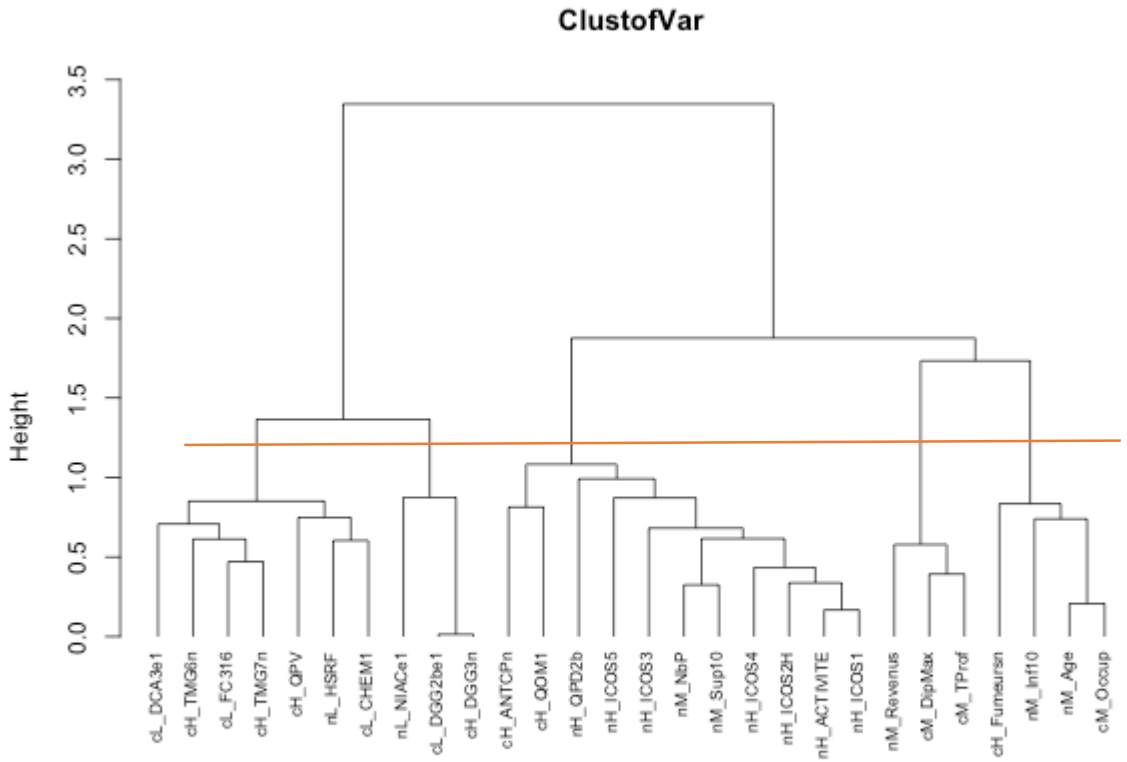




Single and complete linkage on z-scores dissimilarities







Gives a partition in 5 clusters very close to the ones given by complete linkage with p-values or z-scores dissimilarities

Rand index between partitions

	ClDistAve	ClDistCom	ClDistSing	ClDistWar	ClFractileA	ClFractileC	ClFractileS	ClPVAvera	ClPVComp	ClPVSingle	ClClustOfVar
ClDistAverage	1	0,78	0,3	0,66	0,85	0,88	0,3	0,37	0,78	0,3	0,81
ClDistComplete	0,78	1	0,39	0,67	0,83	0,85	0,39	0,46	0,83	0,39	0,87
ClDistSingle	0,3	0,39	1	0,28	0,45	0,25	1	0,93	0,34	1	0,31
ClDistWard	0,66	0,67	0,28	1	0,58	0,7	0,28	0,29	0,66	0,28	0,67
ClFractileAverage	0,85	0,83	0,45	0,58	1	0,79	0,45	0,52	0,75	0,45	0,73
ClFractileComplete	0,88	0,85	0,25	0,7	0,79	1	0,25	0,31	0,9	0,25	0,92
ClFractileSingle	0,3	0,39	1	0,28	0,45	0,25	1	0,93	0,34	1	0,31
ClPVAverage	0,37	0,46	0,93	0,29	0,52	0,31	0,93	1	0,35	0,93	0,33
ClPVComplete	0,78	0,83	0,34	0,66	0,75	0,9	0,34	0,35	1	0,34	0,92
ClPVSingle	0,3	0,39	1	0,28	0,45	0,25	1	0,93	0,34	1	0,31
ClClustOfVar	0,81	0,87	0,31	0,67	0,73	0,92	0,31	0,33	0,92	0,31	1

Spearman's correlations between ultrametrics

	Dsingle	Dcomplet	Daverage	Dward.D	PVsingle	PVcomple	PVaverag	Fsingle	Fcomplete	Faverage	ClustOfVar
Dsingle	1	0,219	0,825	-0,073	0,921	0,261	0,495	0,915	0,261	0,602	0,267
Dcomplet	0,219	1	0,365	0,167	0,272	0,51	0,42	0,271	0,51	0,489	0,638
Daverage	0,825	0,365	1	0,076	0,895	0,42	0,578	0,896	0,42	0,863	0,564
Dward.D	-0,073	0,167	0,076	1	-0,034	0,352	0,248	-0,059	0,347	0,158	0,243
PVsingle	0,921	0,272	0,895	-0,034	1	0,312	0,554	0,984	0,311	0,7	0,374
PVcomple	0,261	0,51	0,42	0,352	0,312	1	0,774	0,294	1	0,469	0,611
PVaverage	0,495	0,42	0,578	0,248	0,554	0,774	1	0,543	0,773	0,647	0,5
Fsingle	0,915	0,271	0,896	-0,059	0,984	0,294	0,543	1	0,294	0,698	0,361
Fcomplete	0,261	0,51	0,42	0,347	0,311	1	0,773	0,294	1	0,469	0,61
Faverage	0,602	0,489	0,863	0,158	0,7	0,469	0,647	0,698	0,469	1	0,788
ClustOfVa	0,267	0,638	0,564	0,243	0,374	0,611	0,5	0,361	0,61	0,788	1

- Dendrograms give partitions with 3 to 5 clusters which come from different blocks and include variables of different nature

- For the partitions with 5 clusters

  - Cluster 2 groups together the existence of an adjoining garage (**n**L_DGG2be) and Age of building from the **Dwellings block** with the number of times you park your car there (**c**H_DGG3n) from the **Habit block**

  - Cluster 4 includes the highest degree (**c**M_dipMax) with the household income (**n**M_Rev) and type of occupation of the household database (**c**M_TProf) of the same **Household block**

5. Conclusion and perspectives

- Ward's hierarchical clustering associated with RV coefficients gives satisfactory results
- Dissimilarities based upon p-values or z-scores should be used with complete linkage and give similar results
- Clustering aroung latent components still a serious competitor
- Validation needed:
  - benchmarks ?
  - simulations

References

- Chavent, M. , Kuentz-Simonet, V. , Liquet, B. and Saracco, J., ClustOfVar: An R package for the Clustering of Variables. *Journal of Statistical Software*, **50**(13):1–16, 2012.
- Costa Nicolau, F. and Bacelar-Nicolau, H., Some trends in the classification of variables. In C. Hayashi, ed., *Data Science, Classification, and Related Methods. Proceedings of the Fifth Conference of the IFCS,* Kobe, 89–98. Springer, 1998.
- Escofier, B., Traitement simultané de variables qualitatives et quantitatives en analyse factorielle, *Cahiers de l'Analyse des Données*, **4**, 137-146, 1979.
- Kiers, H., Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables, *Psychometrika,* **56**, 197-212, 1991.
- Kirchner S, Mandin C, Derbez M, Ramalho O, Riberon J, Dassonville C, Lucas J.P et Ouattara M. Qualité d'air intérieur, qualité de vie. 10 ans de recherche pour mieux respirer, Observatoire de la qualité de l'air intérieur, *CSTB*. 212 p, 2011.
- Lerman, I. C. , Etude distributionnelle de statistiques de proximité entre structures finies de même type; application à la classification automatique. *Cahiers du Bureau universitaire de recherche opérationnelle Série Recherche*, **19**, 3-53, 1973.
- Morineau, A., SPAD.N logiciel pour l'analyse statistique des données, *Modulad-Le Monde des Utilisateurs de L'Analyse des Données*, **6**, 27-60, 1991.

- Qannari, E.M. , Vigneau, E. and Courcoux, Ph., Une nouvelle distance entre variables. Application en classification. *Revue de Statistique Appliquée,* **46**(2):21–32, 1998.
- Pagès, J., Analyse factorielle de données mixtes, *Revue de Statistique Appliquée*, **52**(4), pp. 93-111, 2004.
- Robert, P. and Escoufier, Y., A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Applied Statistics*, **25**(3):257–265, 1976.
- Saporta, G., Quelques applications des opérateurs d'Escoufier au traitement des variables qualitatives. *Statistique et analyse des données*, **1**(1), 38-46, 1976.
- Saporta, G., Simultaneous analysis of qualitative and quantitative data. In *Atti della XXXV Riunione Scientifica, Societa Italiana di Statistica*, Padova, Italy, **1**, 62–72, 1990.
- SAS Institute Inc., The VARCLUS Procedure , *SAS/STAT Software 15.2 User's Guide,* SAS Institute Inc., Cary, NC, 2020.
- Tenenhaus, M., Analyse en composantes principales d'un ensemble de variables nominales ou numériques. *Revue de Statistique Appliquée*, **25**(2):39–56, 1977.
- Vigneau, E. and Qannari, E.M., Clustering of variables around latent components. *Communications in Statistics-Simulation and Computation*, **32**(4):1131–1150, 2003.