Symbolic classification methods applied to the intervals of quantile estimates of production costs

Dominique Desbois

AgroParisTech
Talents for a sustainable planet

INRAE
la science pour la vie, l'humain, la terre

27e Rencontres de la Société francophone de Classification (SFC) – 14 Septembre 2022, Lyon

Méthodes de classification symbolique appliquées aux intervalles d'estimations quantiles des coûts de production

Semoir à engrais (19e siècle) *Musée du Vivant*, AgroParisTech photothèque

Dominique Desbois
Paris Saclay Applied Economics, INRAE/AgroParisTech, Université Paris Saclay

Sommaire

Introduction à l'estimation quantile des coûts

- I) Classification ascendante hiérarchique
- II) Classification divisive
- III) Procédures de validation

Perspectives

Estimation des coûts en engrais: méthodologie « intrants-extrants »

- **Modélisation économétrique des coûts de production agricoles :**

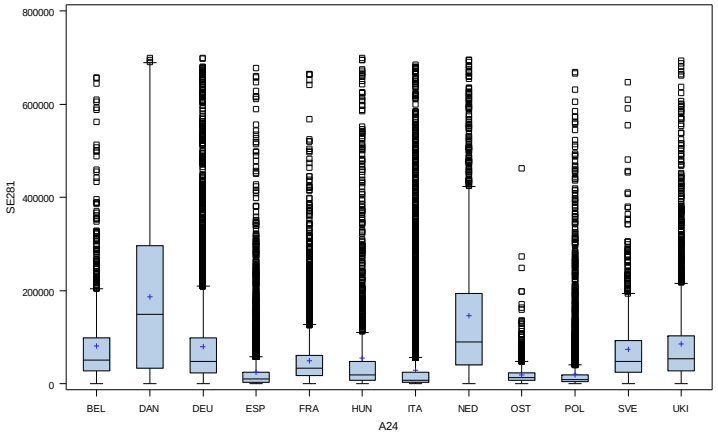
Modèle à coefficients constants

$$X_{ih} = \sum_{k=1}^{K} \alpha_{ih}^{k} Y_{kh} + \varepsilon_{ih} \quad with \quad \varepsilon_{ih} \quad i.i.d.$$

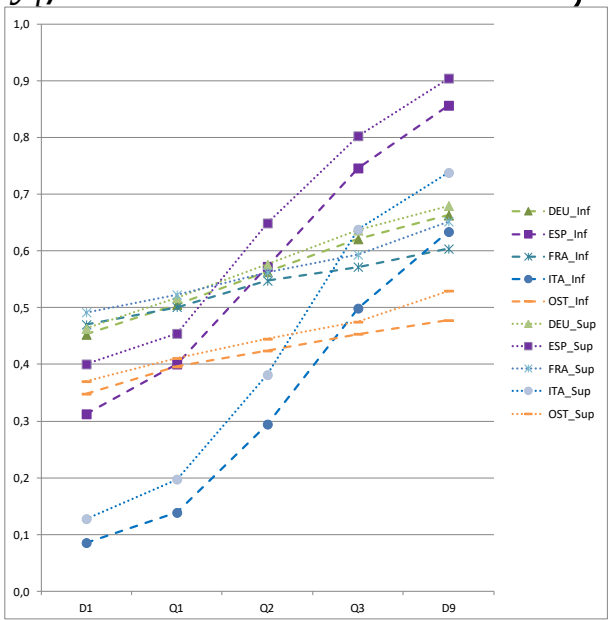
PRODUITS / CHARGES	Y_{1h}	…	Y_{kh}	…	Y_{Kh}	TOTAL CHARGE
X_{1h}	a_{1h}^{1}	…	a_{1h}^{k}	…	a_{1h}^{K}	$\sum X_{1h}$
⋮	⋮		⋮		⋮	⋮
X_{ih}	a_{ih}^{1}	…	a_{ih}^{k}	…	a_{ih}^{K}	$\sum X_{ih}$
⋮	⋮		⋮		⋮	⋮
X_{Ih}	a_{Ih}^{1}	…	a_{Ih}^{k}	…	a_{Ih}^{K}	$\sum X_{Ih}$
TOTAL PRODUIT	$\sum Y_{1h}$	…	$\sum Y_{kh}$	…	$\sum Y_{Kh}$	$\sum_{k} Y_{kh} = \sum_{i} X_{ih}$

(D.Desbois, 2015), projet FACEPA, programme européen FP7

Critère d'optimisation de l'estimation quantile des coûts de production agricoles



$$\underset{\beta}{Min}\left\{\sum_{\vec{i}\,x_i \geq y_i^{'}\beta} q\left|x_i - y_i^{'}\beta\right| + \sum_{\vec{i}\,x_i \leq y_i^{'}\beta} (1-q)\left|x_i - y_i^{'}\beta\right|\right\}$$



(D.Desbois, 2015), , projet FACEPA, programme européen FP7

Coûts de fertilisation pour 1 000 € de produit brut en grandes cultures : 12b pays européens

(D.Desbois, FADN-UE 2006)

id	D1I	D1S	Q1I	Q1S	Q2I	Q2S	Q3I	Q3S	D9I	D9S
Bel	0,009	0,019	0,023	0,030	0,038	0,047	0,056	0,080	0,082	0,110
Dan	0,018	0,024	0,035	0,035	0,056	0,056	0,094	0,094	0,140	0,140
Deu	0,004	0,009	0,025	0,033	0,082	0,082	0,140	0,140	0,181	0,181
Esp	0,013	0,017	0,025	0,033	0,058	0,058	0,103	0,103	0,169	0,169
Fra	0,023	0,028	0,053	0,065	0,125	0,125	0,182	0,182	0,232	0,232
Hun	0,020	0,038	0,056	0,071	0,093	0,110	0,138	0,164	0,197	0,197
Ita	0,007	0,011	0,019	0,022	0,041	0,041	0,078	0,078	0,121	0,121
Ned	0,001	0,004	0,004	0,006	0,009	0,012	0,017	0,022	0,026	0,029
Ost	0,000	0,029	0,043	0,057	0,068	0,086	0,106	0,127	0,155	0,179
Pol	0,024	0,032	0,052	0,059	0,088	0,099	0,146	0,165	0,215	0,228
Sve	-0,007	0,016	0,003	0,038	0,100	0,100	0,215	0,215	0,293	0,293
Uki	0,006	0,029	0,036	0,047	0,088	0,088	0,137	0,137	0,171	0,171

Belgique :

- de 9 à 19 € de fertilisants pour 1 000 € de produit brut en grandes cultures, décile D1
- € 23 to 30 de fertilisants pour 1 000 € de produit brut en grandes cultures, quartile Q1
- € 38 to 47 de fertilisants pour 1 000 € de produit brut en grandes cultures, médiane Q2
- € 56 to 80 de fertilisants pour 1 000 € de produit brut en grandes cultures, quartile Q3
- € 82 to 110 de fertilisants pour 1 000 € de produit brut en grandes cultures, décile D9

```
syearcrop2<-read.sym.table("~/FERTI/syearcrop2b.txt",header=TRUE,sep='\t',dec=',',row.names=1)
print.data.frame(syearcrop2)
```

I) Classification ascendante hiérarchique

- Hausdorff
- Ichino
- Gowda-Diday
- Wasserstein

La dissimilarité d'Hausdorff : définition pour ls données d'intervalle

Pour des intervalles d'estimation, la dissimilarité d'Hausdorff estcalculée comme suit :

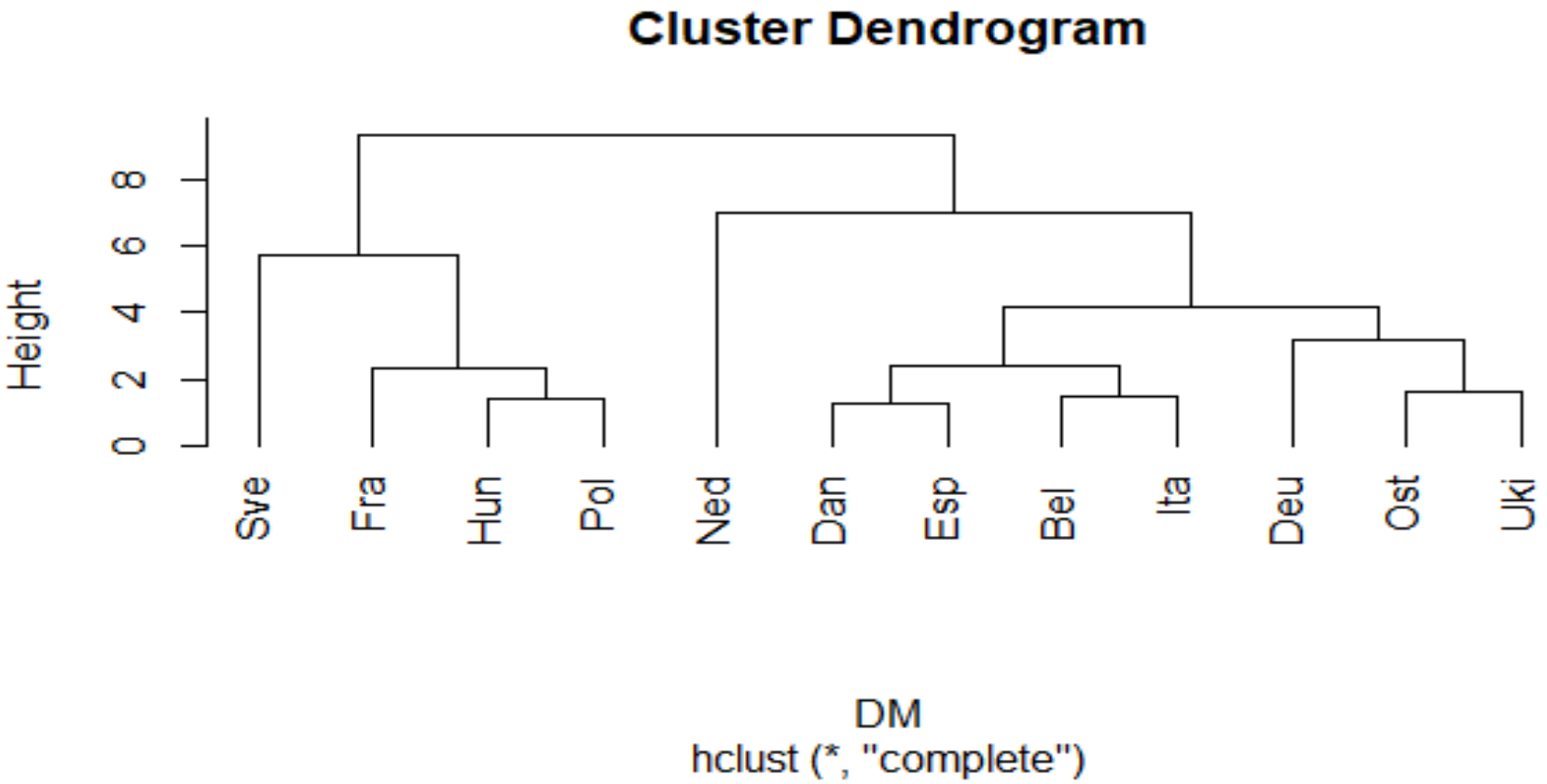
$$\delta(z_l, z_{l'}) = \sum_{q=1}^{Q} max\left\{\left|\underline{z_l^q} - \underline{z_{l'}^q}\right| ; \left|\overline{z_l^q} - \overline{z_{l'}^q}\right|\right\}$$

avec $\underline{z_l^q}$, inf. d'intervalle d'estimation du quantile d'ordre ***q*** pour le pays ***l***

et

$\overline{z_l^q}$, sup. d'intervalle d'estimation du quantile d'ordre ***q*** pour le pays ***l***

Résultats de la CAH symbolique (option «Hausdorff ») : le dendrogramme hiérarchique pour 12 pays européens




```
DM <- sym.dist.interval(sym.data = syearcrop2[,2:6],gamma = 0.5,method = "Hausdorff",
                        normalize = FALSE, SpanNormalize = TRUE, euclidea = TRUE, q = 2)
model <- hclust(DM)
plot(model, hang = -1)
```

L'option « Hausdorff « de la fonction sym.dist.interval du logiciel RSDA est adaptée de Carvalho F., Souza R.,Chavent M., and Lechevallier Y. (2006) Adaptive Hausdorff distances and dynamic clustering of symbolic interval data., *Pattern Recognition Letters, v*olume 27(1): 167-179.

La dissimilarité de Gowda – Diday : définition pour les intervalles

- Pour les intervalles d'estimation la dissimilarité de Gowda – Diday est calculée comme suit pour les intervalles d'estimations :

$$\delta(z_l, z_{l'}) = \sum_{q=1}^{Q} \delta\left(z_l^q, z_{l'}^q\right)$$

où $\delta\left(z_l^q, z_{l'}^q\right) = \delta_p\left(z_l^q, z_{l'}^q\right) + \delta_s\left(z_l^q, z_{l'}^q\right)$

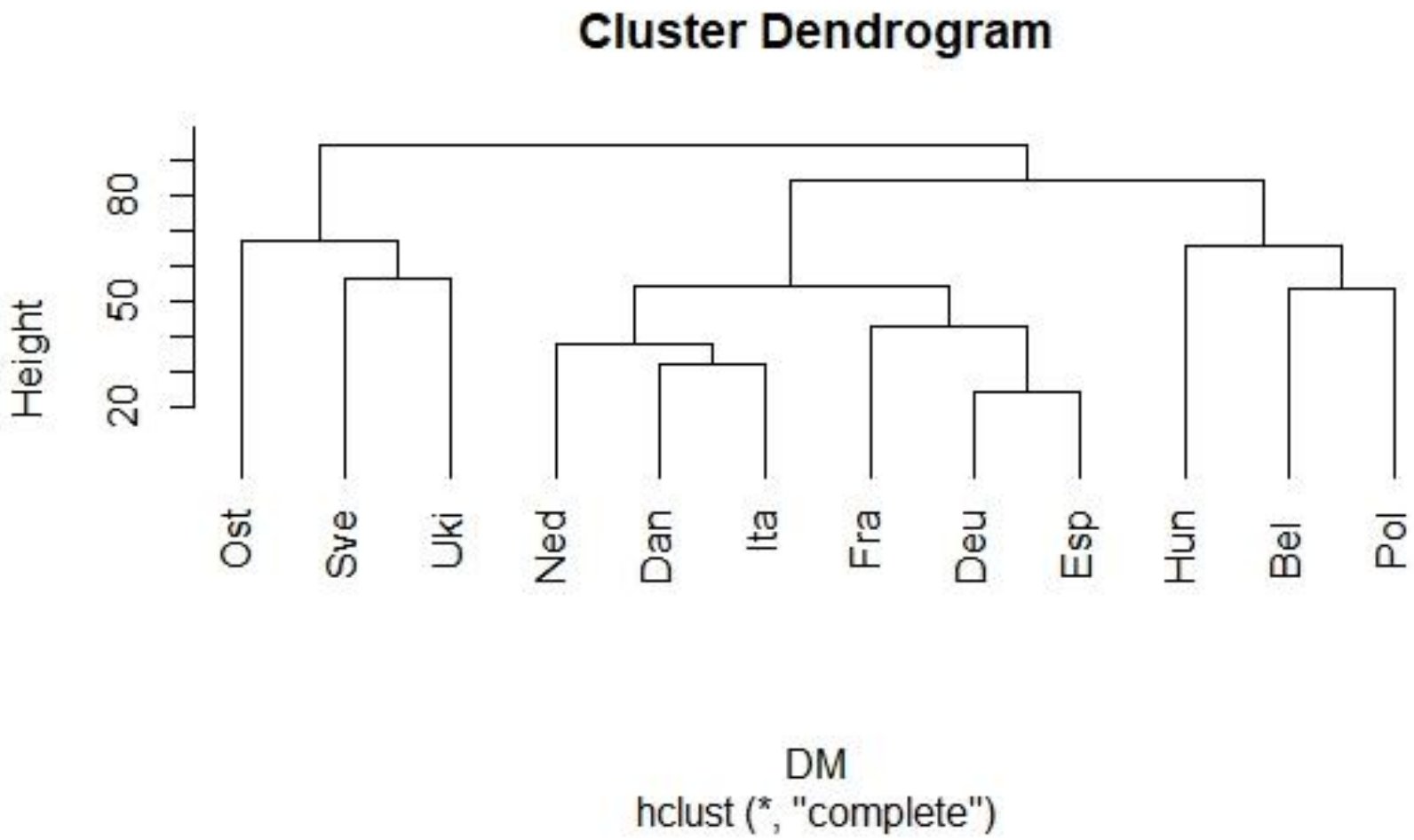
avec $\delta_p\left(z_l^q, z_{l'}^q\right) = cos\left[90\left(1 - \frac{\left|\underline{z_l^q} - \underline{z_{l'}^q}\right|}{u^q}\right)\right]$,

u^q étant la largeur du plus grand intervalle d'estimation for the q^e quantile;

et $\delta_s\left(z_l^q, z_{l'}^q\right) = cos\left[45\left(\frac{\left|\overline{z_l^q} - \underline{z_l^q}\right| + \left|\overline{z_{l'}^q} - \underline{z_{l'}^q}\right|}{\max\left(\overline{z_l^q}, \overline{z_{l'}^q}\right) - \min\left(\underline{z_l^q}, \underline{z_{l'}^q}\right)}\right)\right]$.

Résultats de la CAH symbolique (option «Gowda-Diday ») :

le dendrogramme hiérarchique pour 12 pays européens




```
DM <- sym.dist.interval(sym.data =  syearcrop2[,2:6],  method = "Gowda.Diday")
model <- hclust(DM)
plot(model, hang = -1)
```

La distance d'Ichino : définition pour les intervalles d'estimation

Pour les intervalles d'estimation, la dissimilarité d'Ichino est calculée comme suit :

$$\delta(z_l, z_{l'}) = \sum_{q=1}^{Q} \delta_I\left(z_l^q, z_{l'}^q\right)$$

où

$$\delta_I\left(z_l^q, z_{l'}^q\right) = \begin{cases} \mu\left(z_l^q - z_{l'}^q\right)/\mu\left(z_l^q \cup z_{l'}^q\right) \; if \; \mu\left(z_l^q \cup z_{l'}^q\right) > 0 \\ 0 \; if \; \mu\left(z_l^q \cup z_{l'}^q\right) = 0 \end{cases}$$

soit avec une mesure μ continue

$$\delta_I\left(z_l^q, z_{l'}^q\right) = \frac{\int_{\mho q} \left| f\left(z_l^q\right) - f\left(-z_{l'}^q\right) \right| d\mu}{\int_{\mho q} sup\left(f\left(z_l^q\right); f\left(-z_{l'}^q\right)\right) d\mu}$$

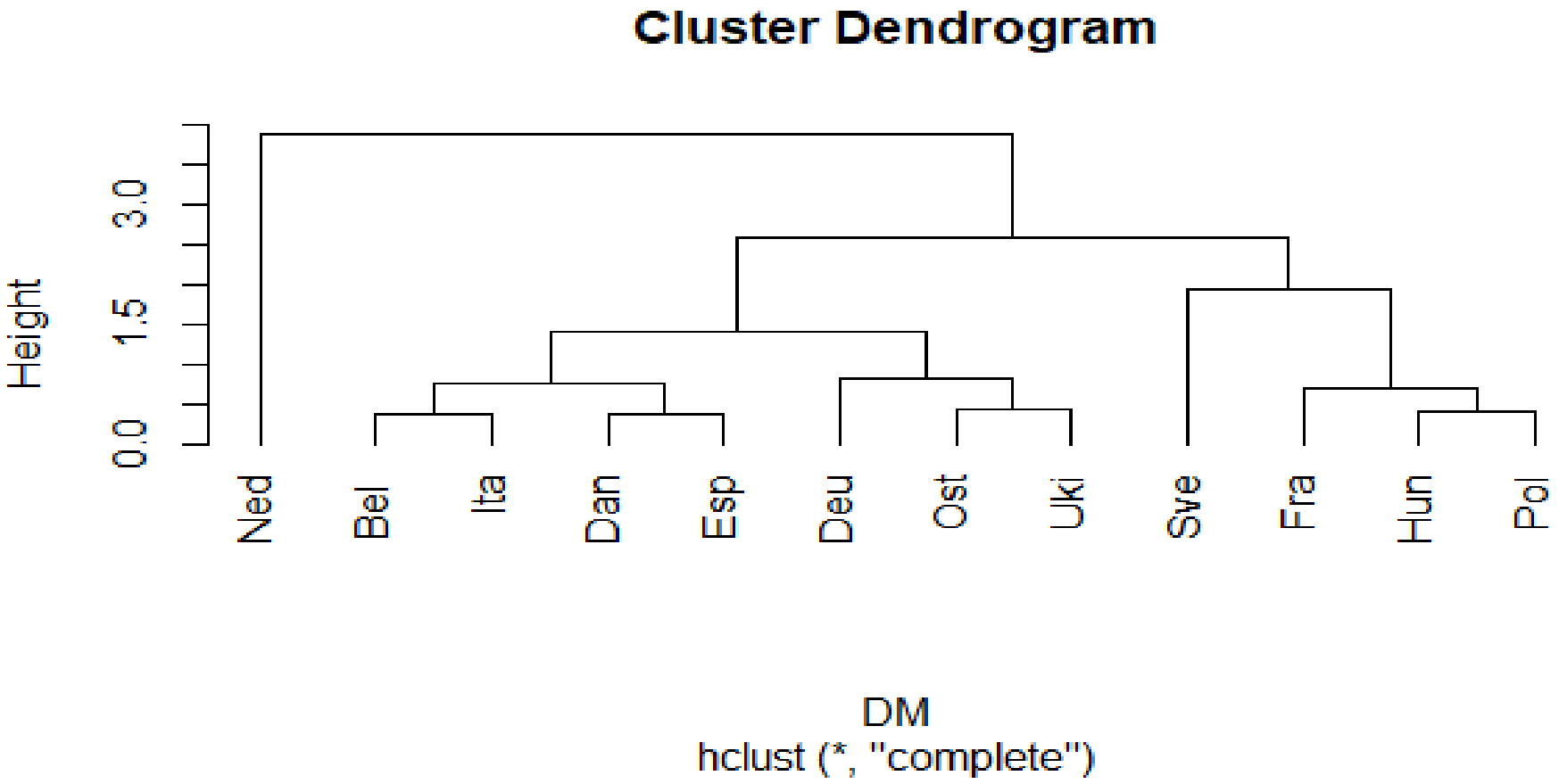
NB: relié à la distance d'Ichino et à la distance de Yaguchi entre ensemble, avec ($\gamma = 0$), basé sur la différence symétrique entre ensembles

$$\delta_I\left(z_l^q, z_{l'}^q\right) = \frac{\pi\left(z_l^q \oplus z_{l'}^q\right) - \pi\left(z_l^q \cap z_{l'}^q\right) + \gamma\left|2\pi\left(z_l^q \cap z_{l'}^q\right) - \pi\left(z_l^q\right) - \pi\left(z_{l'}^q\right)\right|}{R}$$

avec R comme terme de normalisation, soit *R=1,* potentiel du domaine du q^e quantile,
soit $R = \pi(z_l \oplus z_{l'})$, le produit cartésien .

CAH symbolique (option «Ichino ») pour intervalles d'estimation quantiles

Arbre hiérarchique pour 12 pays européens




```
DM <- sym.dist.interval(sym.data= syearcrop2[,2:6], method = "Ichino")
model <- hclust(DM)
plot(model, hang = -1)
```

L'option « Ichino » fonction symbolique sym.dist.interval du logiciel RSDA est adaptée de Ichino, M. and Yaguchi, H. (1994): Generalized Minkov metrics for mixed feature type data analysis. *IEEE Transactions on Systems, Man and Cybernetics*, 24 (4), 698–708.

La distance L2 de Wasserstein :
définition pour les intervalles d'estimation

Soit, les fonctions de répartition inverses F_l^{-1} et $G_{l'}^{-1}$ des distributions *F* et *G* respectivement pour les pays *l* et *l'* :

$$\delta^2(z_l, z_{l'}) = \int_0^1 \left(F_l^{-1}(t) - G_{l'}^{-1}(t)\right)^2 dt$$

Irpino et Romano (2007) utilisent les moments μ_F et μ_G avec les écarts-types σ_F et σ_G de ces distributions pour décomposer cette distance en trois composantes :

$$\delta^2(z_l, z_{l'}) = \underbrace{(\mu_F - \mu_G)^2}_{localisation} + \underbrace{(\sigma_F - \sigma_G)^2}_{taille} + \underbrace{2\sigma_F\,\sigma_G\left(1 - \rho_{QQ}(F,G)\right)}_{forme}$$

La distance de Wasserstein représente ainsi les écarts entre les distributions en termes de localisation du niveau global mais aussi des caractéristiques de taille et de forme.
Ils. montrent que la distance quadratique de Wasserstein entre deux histogrammes est estimable par les distances entre les centres et le tiers des distances entre les rayons des intervalles.
Irpino et Verde (2008) proposent un algorithme de classification pour les données d'intervalle basée sur la distance de Wasserstein.

Irpino A., Romano E., Optimal histogram representation of large data sets: fisher vs piecewise linear approximation, *EGC* 1 (2007) 99–110.
Irpino A. , Verde R. , Dynamic clustering of interval data using a Wasserstein-based distance*, Pattern Recogn. Lett.* 29 (11) (2008) 1648–1658.

II) Classification divisive

La méthode de classification divisive de Chavent :

- générée par la réponse binaire (*oui*/*non*) à la question $\Psi = [Is\ z^q \leq c\ ?\ ]$, notons $\{A_k, \overline{A_k}\}$ la bipartition induite de la classe C_k formée de n_k objets ;

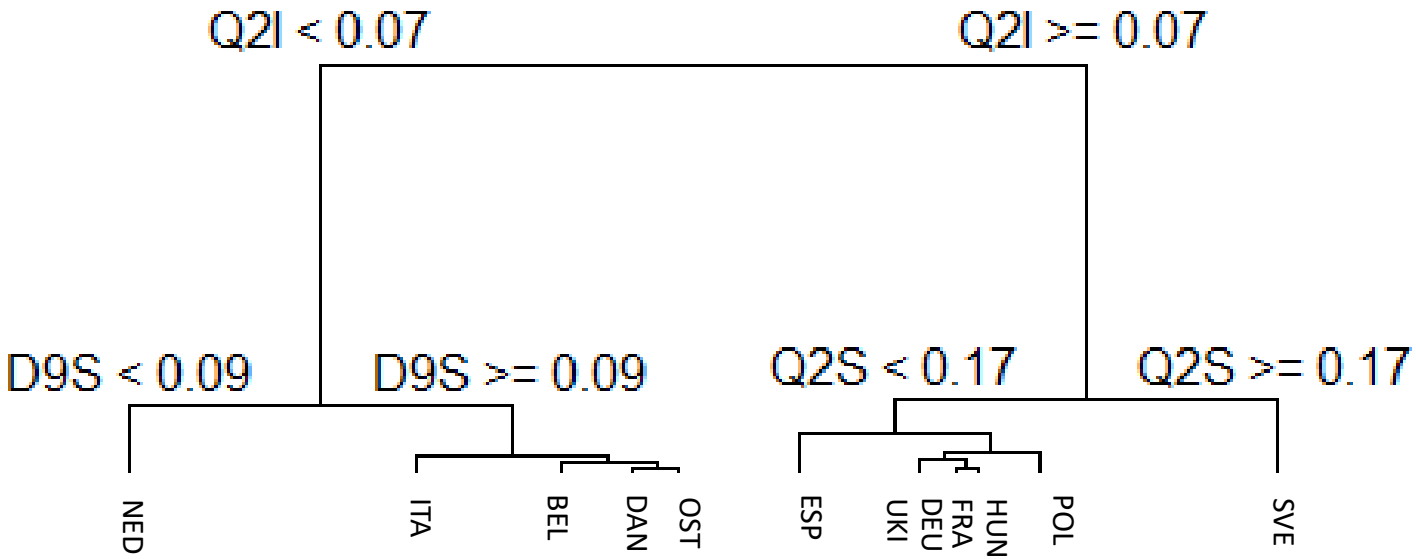
- comme dans la méthode de Ward, la "*hiérarch*ie supérieure" de la partition P_K est indexée par l'indice *h* de la classe C_K , défini par leur inertie inter :

$$h(C_k) = B\left(A_k, \overline{A_k}\right) = \frac{\mu(A_k)\mu\left(\overline{A_k}\right)}{\mu(A_k) + \mu\left(\overline{A_k}\right)} d^2\left(g(A_k), g\left(\overline{A_k}\right)\right)$$

- l'algorithme DIVCLUS−T divise la classe C_K^* qui maximise $h(C_K)$, assurant que la nouvelle partition $P_{K+1} = P_K \cup \{A_K, \overline{A_K}\} - C_K^*$ possède l'inertie intra minimale, respectant l'équation suivante

$$W(P_{K+1}) = W(P_K) - h(C_K^*).$$

Résultats de l'algorithme DIVCLUST de classification divisive : arbre hiérarchique divisif des 12 pays européens




```
syear2crop<-read.delim2("~/FERTI/syear2crop.txt", row.names=1, stringsAsFactors=FALSE)
tree <- divclust(syear2crop[,4:13])
plot(tree)
```

La classification divisive est opérée par- la fonction Divclust de R library, basée sur Chavent M., Lechevalier Y., Briant O. (2007) DIVCLUS-T: A monothetic divisive hierarchical clustering method. *Computational Statistics & Data Analysis*, 52, 2, 687-701.

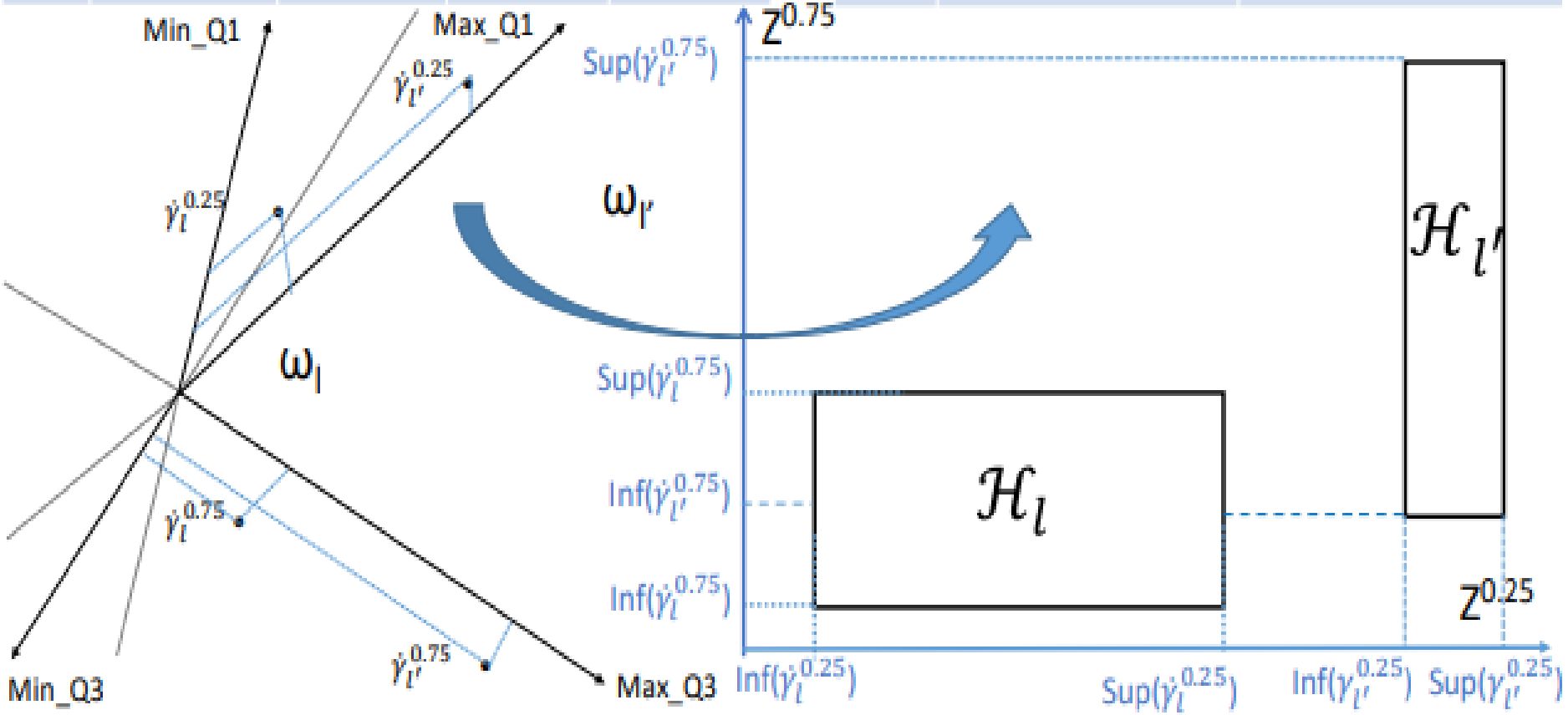
III) Validation

- ACP symbolique
- Analyses de variance*
- Partitions optimales

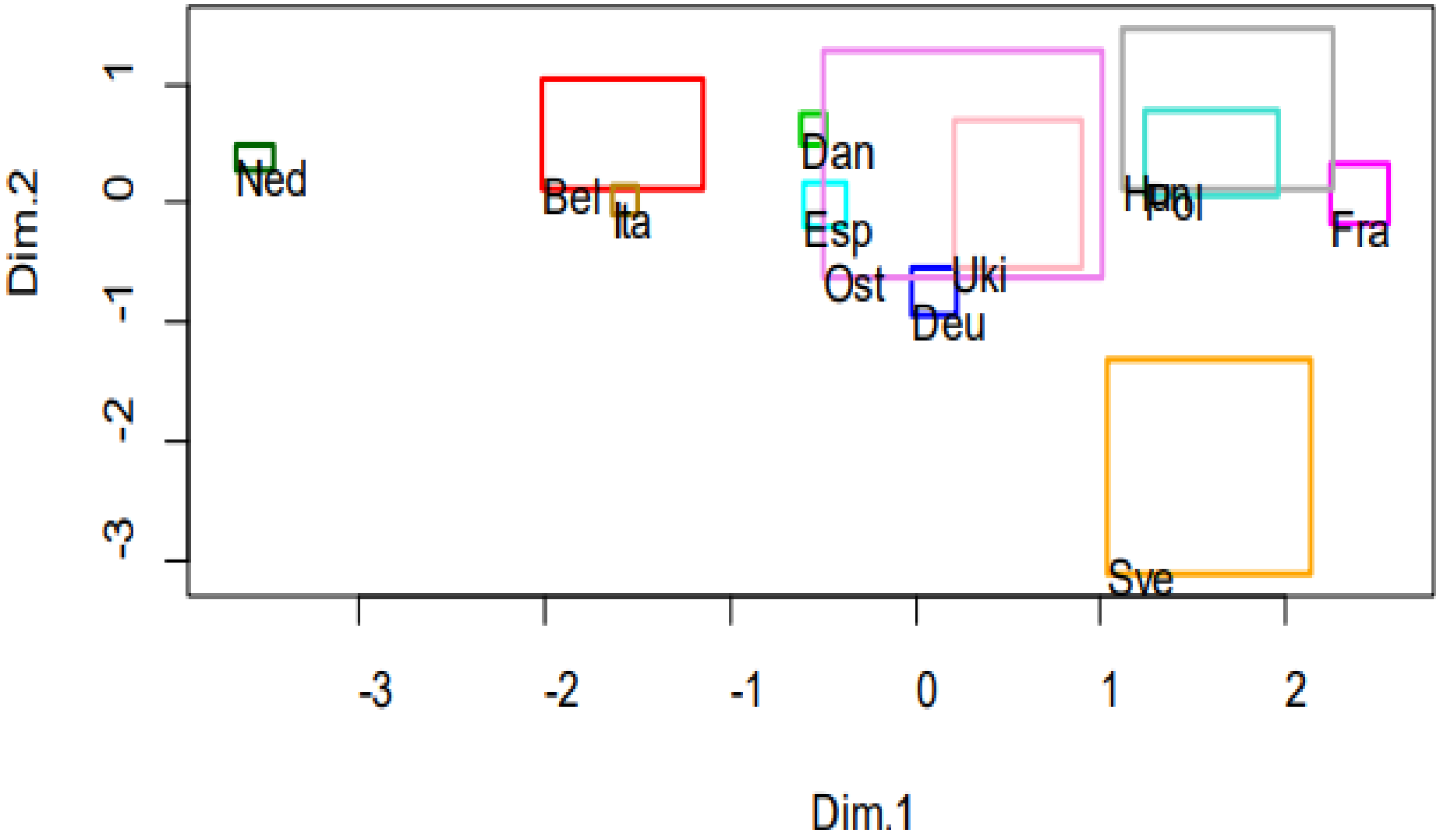
Codage symbolique des intervalles d'estimation pour les quartiles Q1 et Q3

	Min_Q1	Max_Q1	Min_Q3	Max_Q3
...	...	...	...	...
ω_l	$\text{Inf}(\dot{y}_l^{0.25})$	$\text{Sup}(\dot{y}_l^{0.25})$	$\text{Inf}(\dot{y}_l^{0.75})$	$\text{Sup}(\dot{y}_l^{0.75})$
...	...	...	...	...
$\omega_{l'}$	$\text{Inf}(\dot{y}_{l'}^{0.25})$	$\text{Sup}(\dot{y}_{l'}^{0.25})$	$\text{Inf}(\dot{y}_{l'}^{0.75})$	$\text{Sup}(\dot{y}_{l'}^{0.75})$
...	...	...	...	...

	$Z^{0.25}$	$Z^{0.75}$
...	...	...
ω_l	$[\text{Inf}(\dot{y}_l^{0.25}); \text{Sup}(\dot{y}_l^{0.25})]$	$[\text{Inf}(\dot{y}_l^{0.75}); \text{Sup}(\dot{y}_l^{0.75})]$
...	...	...
$\omega_{l'}$	$[\text{Inf}(\dot{y}_{l'}^{0.25}); \text{Sup}(\dot{y}_{l'}^{0.25})]$	$[\text{Inf}(\dot{y}_{l'}^{0.75}); \text{Sup}(\dot{y}_{l'}^{0.75})]$
...	...	...



Premier plan principal (sym.pca) : option variance optimisée



Fonction sym.pca, option « optimized variance », logiciel RSDA v3.0, O. Rodriguez, 2022

Classification symbolique divisive pour les intervalles d'estimation quantiles

partition C2 : test de Wilcoxon sur la première composante principale (CP)

G1=X	PC1	Center(PC1)
BEL	[-1.80 : -1.22]	-1,51
DAN	[-0.95 : -0.31]	-0,63
ITA	[-1.57 : -1.19]	-1,38
NED	[-3.16 : -2.96]	-3,06
OST	[-0.60 : 0.47]	-0,07

G2=Y	PC1	Center(PC1)
DEU	[-0.33 : 0.41]	0,04
ESP	[-0.79 : -0.21]	-0,50
FRA	[1.56 : 2.10]	1,83
HUN	[0.58 : 1.62]	1,10
POL	[0.91 : 1.47]	1,19
SVE	[-0.62 : 3.43]	1,40
UKI	[-0.18 : 0.89]	0,36

>wilcox.test(X,Y)

Wilcoxon rank sum exact test

data: X and Y

W = 1, p-value = 0.005051

Classification symbolique divisive pour les intervalles d'estimation quantiles

partition C4 : test manova sur les 1ere et 2de composantes principales

CNTY	W=center(PC1)	Z=center(PC2)	G
NED	-3,06	-0,45	C1
ITA	-1,38	-0,26	C2
BEL	-1,51	0,31	C2
DAN	-0,63	0,51	C2
OST	-0,07	0,10	C2
ESP	-0,50	-0,02	C3
UKI	0,36	0,33	C3
DEU	0,04	-0,65	C3
FRA	1,83	0,67	C3
HUN	1,10	1,08	C3
POL	1,19	0,87	C3
SVE	1,40	-1,61	C4

```
>res.man<-manova(cbind(W,Z))
>summary (resman)
```

	Df	Pillai	approx F		num Df	den Df	Pr(>F)
G	3	1.4369	6.8055		6	16	0.0009996
Residuals	8						
Response W		**Sum Sq**	**Mean Sq**		**F value**	**Pr(>F)**	
G	3	17.1179	5.7060		9.0294	0.006011	
Response Z		**Sum Sq**	**Mean Sq**		**F value**	**Pr(>F)**	
G	3	3.7136	1.23786		4.1957	0.04654	

Détermination de la partition optimale

- L'indice de chaque partition P_K est le logarithme du rapport des déterminants: $$\varkappa_K = N \log\left(\frac{det(T)}{det\left(WG^{(K)}\right)}\right)$$

  où $T = Z'Z$ est la matrice totale de dispersion(*N* fois la matrice totale de variance-covariance)

  et $WG^{(K)} = \sum_{k=1}^{K} W^{(k)}$ la somme des matrices intraclasses de dispersion $W^{(k)}$ pour chaque groupe C_k de la partition P_K en *K* groupes.

- Le score optimal de l'indice satisfait la règle de decision *min_diff* :

$$K^* = \arg\_\min_K\{\partial_K - \partial_{K-1}\}$$

avec $\partial_K = \varkappa_{K+1} - \varkappa_K$,

en utilisant la procédure *ClusterCrit* (Desgraupes, 2018).

Validation des typologies : Indices des partitions, optimum

Criterion : c(i)	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
Ball Hall max di	0,006605078	0,002982239	0,002061369	0,000744251	0,000472673	0,000379032	0,000221	0,000140056	6,80E-05	5,28E-05
Banfeld Raftery min	-60,01052	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
C index min	0,173622	0,1286931	0,0547564	0,01570826	0,01520788	0,02368761	0,002616917	0,003335737	0	0,03623327
Calinski Harabasz max	14,66531	13,20103	16,23248	26,61088	24,73984	24,29827	31,6854	29,31638	32,40501	17,08513
Davies Bouldin min	0,6735817	0,5730269	0,5592225	0,4833014	0,4291948	0,4498311	0,3964471	0,3284583	0,2468822	0,1900374
Det Ratio min di	84,07927	7,59E+14	1,41E+29	-6,02E+45	3,30E+60	1,74E+79	6,03E+98	1,78E+115	3,07E+133	Inf
Dunn max	0,1835137	0,212243	0,2993189	0,5552686	0,5552686	0,567683	0,9208916	0,9208916	1,073963	0,6597068
Gamma max	0,5974265	0,6565854	0,825	0,941358	0,9392857	0,9111111	0,983871	0,978836	1	0,8769231
G plus min	0,1020979	0,08205128	0,03263403	0,008857809	0,007925408	0,007459208	0,000932401	0,000932401	0	0,00186480
GDI max	0,1835137	0,212243	0,2993189	0,5552686	0,5552686	0,567683	0,9208916	0,9208916	1,073963	0,6597068
Sub-Total C	10	0	0	6	0	0	0	0	11	6

Difference : d(i)=c(i)-c(i+1)	D2	D3	D4	D5	D6	D7	D8	D9	D10	
Ball Hall max di	-0,003622839	-0,000921087	-0,001317118	-0,000271579	-9,36408E-05	-0,000158032	-8,09444E-05	-1,52E-05	-1,51796E-05	
Det Ratio min di	7,59322E+14	1,41253E+29	-6,02063E+45	3,29523E+60	1,73602E+79	6,028E+98	1,7791E+115		-6,02063E+45	
Ksq DetW max di	-3,44761E-37	-8,58941E-50	-8,20859E-64	3,00916E-80	-7,91706E-95	-2,0455E-113	-7,694E-133	-2,36E-167	3,00916E-80	
Log Det Ratio min di	357,98028	394,2828			517,299	539,927	455,084			
Log SS Ratio min di	0,6933183	0,729967	0,91552	0,304388	0,346634	0,642745	0,343504	1,58E-01	0,158403	
Trace W max di	-0,03018137	-0,02257986	-0,01584769	-0,003082663	-0,00261525	-0,003082667	-0,001015	-1,98E-04	-0,000198	
Trace WiB max di	4,30809E+29	2,40249E+31	5,19098E+33	2,56285E+34	6,5834E+36	4,0348E+37	-4,69335E+37	0,00E+00	4,0348E+37	
Sub-total D	2	0	1	1	0	1	0	2	1	0
Total (C+D)	**12**	0	1	**7**	0	1	0	2	**12**	**6**

```
intIdx <- intCriteria(myear2crop[,4:13],cl4$cluster,"all")
intIdx
```

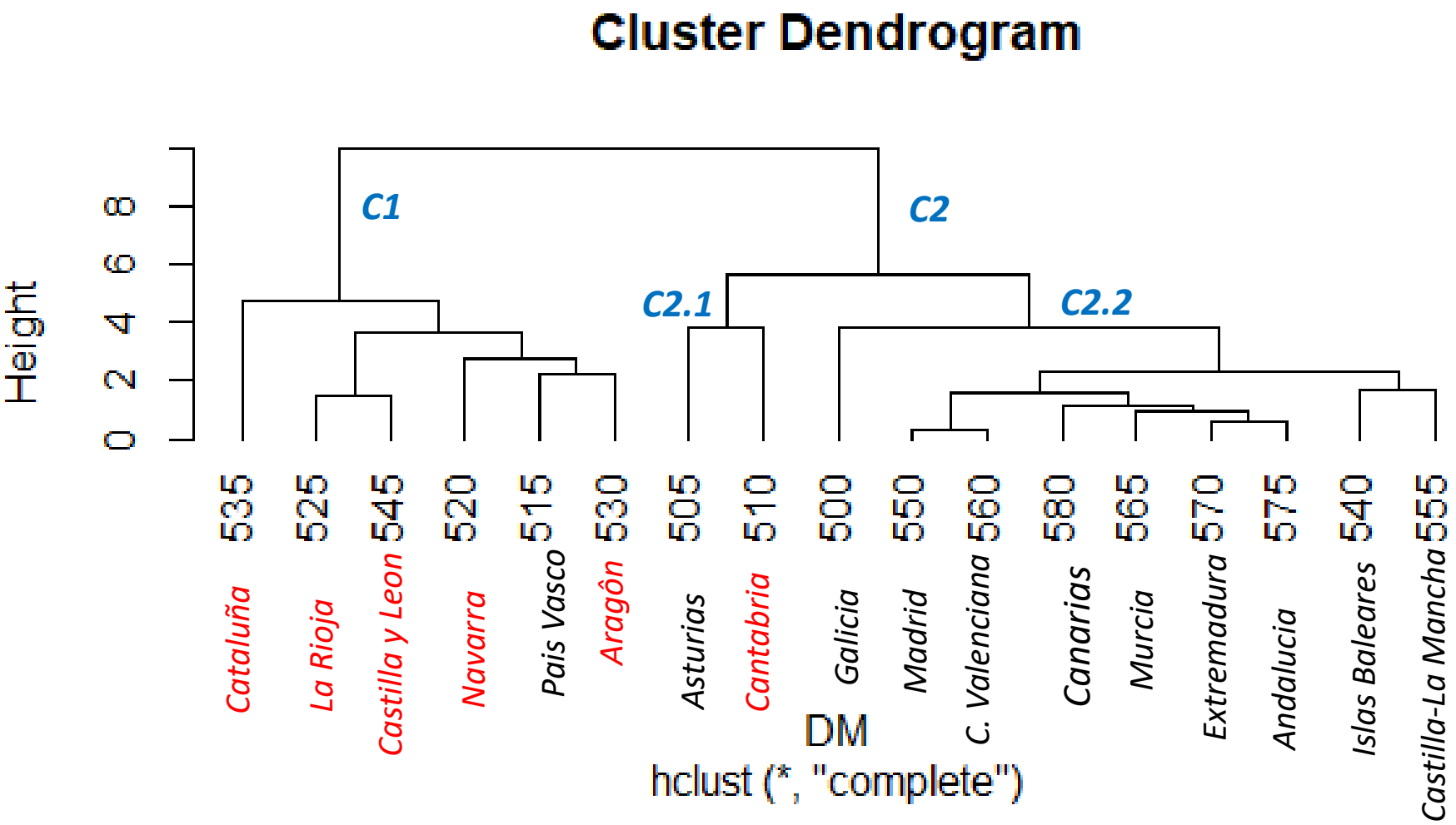
Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *J. Intell. Inf. Syst.*, 17(2-3):107{145, 2001.

Perspectives

- Analyses régionales
- Validation empirique

Résultats de la CAH symbolique (option «Hausdorff ») : le dendrogramme hiérarchique des régions espagnoles

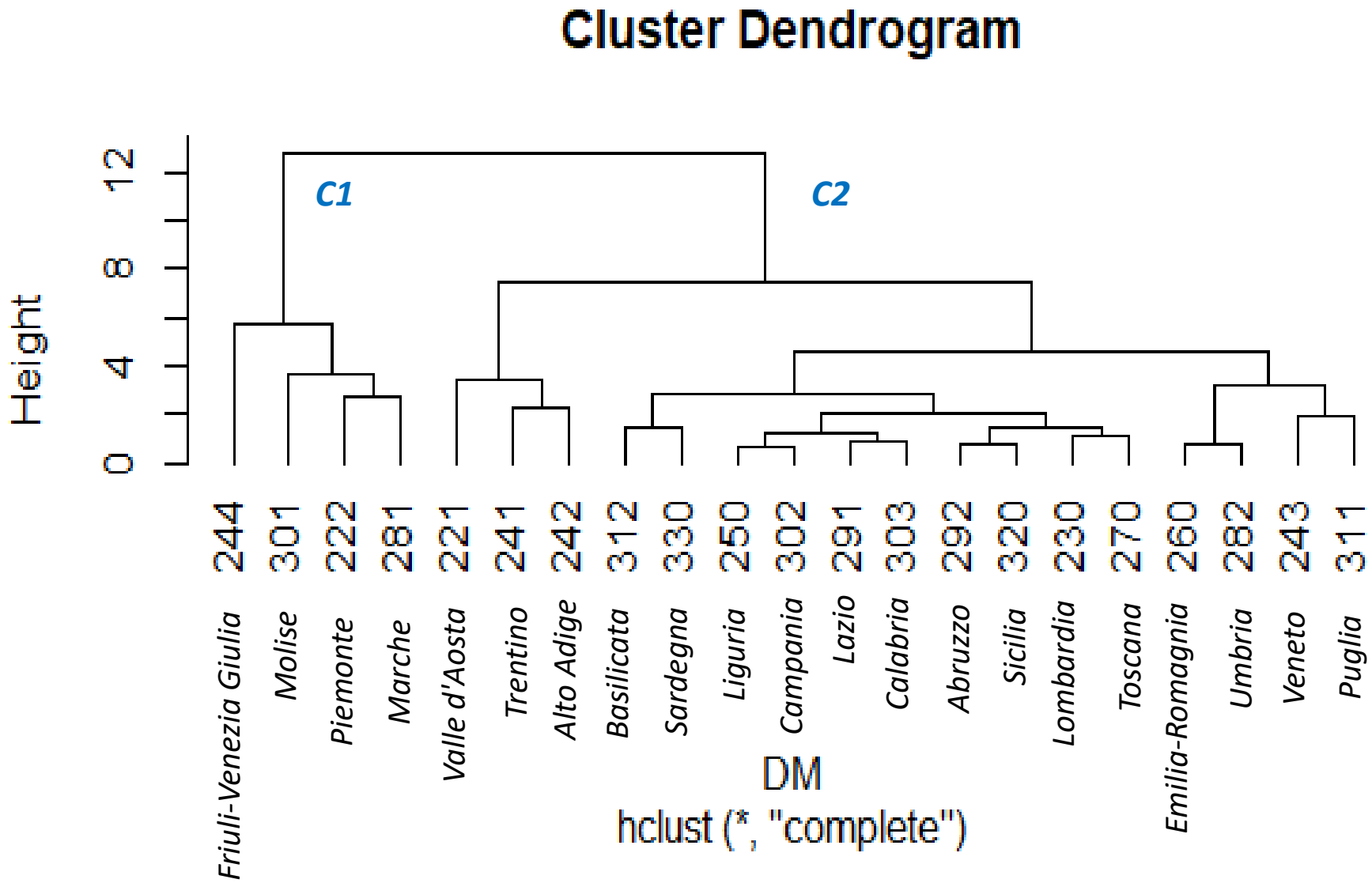
Hausdorff Distance



Cantabria: région espagnole appartenant au bassin de l'Ebre

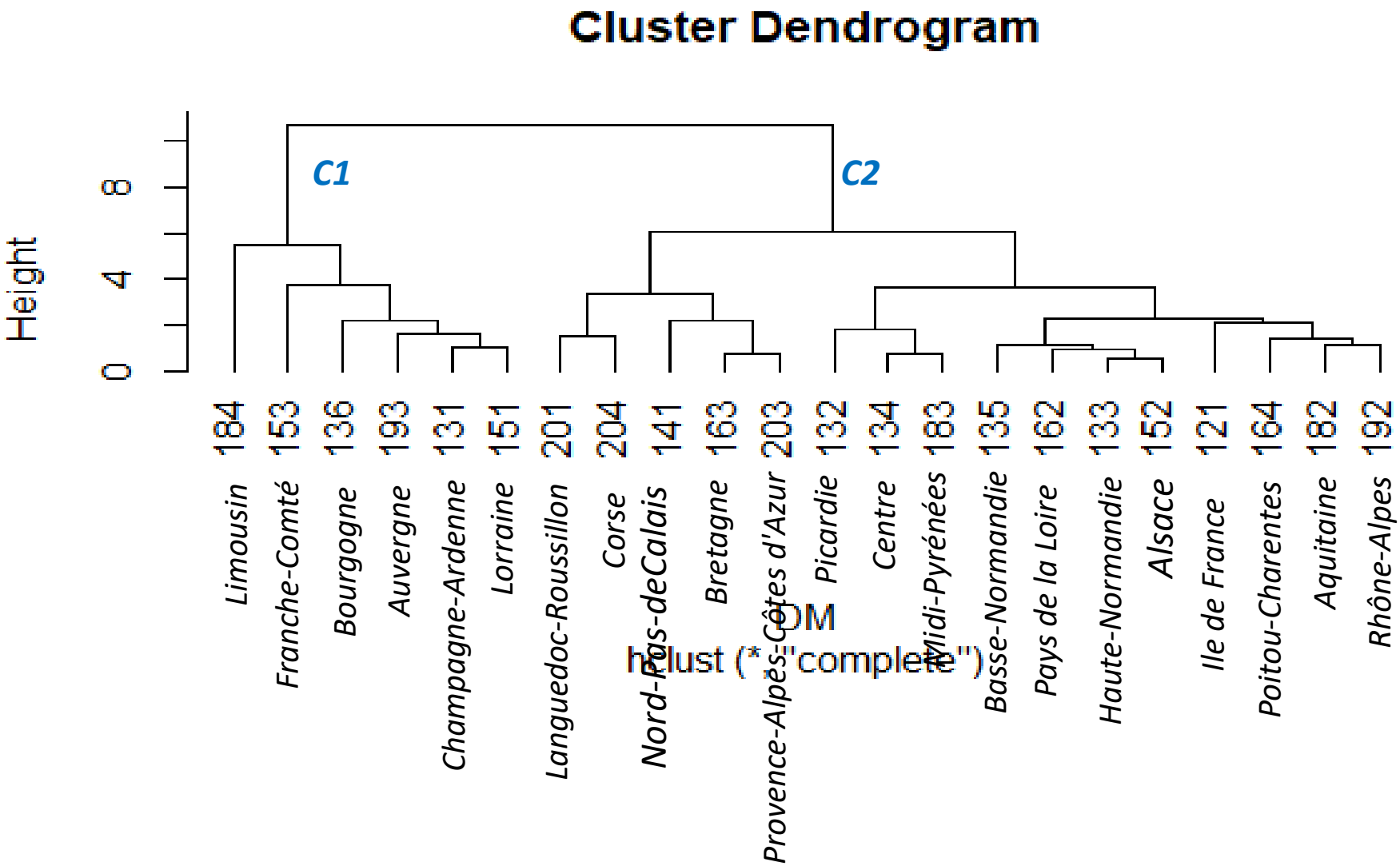
(D.Desbois, FADN-UE, 2006)

Résultats de la CAH symbolique (option «Hausdorff ») :
le dendrogramme hiérarchique des régions italiennes



Résultats de la CAH symbolique (option «Hausdorff ») : le dendrogramme hiérarchique des régions françaises

Hausdorff Distance



Références

- Afonso F., Diday E. and Toque C. (2018) *Data science par analyse des données symboliques*, Technip, Paris, 444 p.
- Billard L., Diday E. (2006) *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, 321 p.
- Cazes P., Chouakria A., Diday E., Schektman Y. (1997) Extensions de l'analyse en composantes principales à des données de type intervalle. *Revue de Statistique Appliquée*, n°24, pp. 5-24.
- Carvalho F., Souza R.,Chavent M., and Lechevallier Y. (2006) Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. *Pattern Recognition Letters*, Volume 27, Issue 3, pp. 167-179.
- Chavent M., Lechevalier Y., Briant O. (2007) DIVCLUS-T: A monothetic divisive hierarchical clustering method. *Computational Statistics & Data Analysis*, 52, 2, 687-701.
- Desbois D. (2015) *Estimation des coûts de production agricoles : approches économétriques*. PhD dissertation directed by J.C. Bureau and Y. Surry, ABIES-AgroParisTech, Paris, 2015.
- Desbois D., Butault J.-P., Surry Y. (2013) Estimation des coûts de production en phytosanitaires pour les grandes cultures. Une approche par la régression quantile*, Economie Rurale*, n° 333. pp.27 49.
- Desgraupes B. (2018) An R Package for Computing Clustering Quality Indices, https://cran.r-project.org.
- Desbois, D., Butault J.-P. and Surry Y. (2017). Distribution des coûts spécifiques de production dans l'agriculture de l'Union européenne : une approche reposant sur la méthode de régression quantile, *Économie rurale*, 361, 3-22.
- Garro J.A., Rodrigues Rojas O. (2019) Optimized Dimensionality Reduction Methods for Interval-Valued Variables and Their Application to Facial Recognitions, *Entropy* 2019, 21(10), 1016.
- Halkidi M., Batistakis Y., and Vazirgiannis Mi. On clustering validation techniques*. J. Intell. Inf. Syst*., 17(2-3):107{145, 2001.
- Ichino M., Yaguchi, H. (1994): Generalized Minkowski metrics for mixed feature type data analysis. *IEEE Transactions on Systems*, *Man and Cybernetics*, 24 (4), 698–708.
- Irpino,A., Romano E. (2007) Optimal histogram representation of large data sets: fisher vs piecewise linear approximation, *EGC* 1, 99–110.
- Irpino A. , Verde R. (2008) Dynamic clustering of interval data using a Wasserstein-based distance, *Pattern Recogn. Lett.* 29 (11) 1648–1658
- Koenker R. and Bassett G. (1978) Regression quantiles. *Econometrica*, 46, 3350, 1978.
- Lauro C.N. and Palumbo F. (2000) Principal component analysis of interval data: a symbolic data analysis approach. *Computational Statistics*, 15, 1, 73-87.
- MAINT.Data: Modelling and Analysing Interval Data in R PDF download
- Pedro Duarte Silva A. , Brito P., Filzmoser P. et Dias J. G. (2021) *The R Journal* 13:2, pages 336-364.
- Rodrigues Rojas O. (2019) *R to Symbolic Data Analysis: Package 'RSDA'*, Version 3.0, October 21, 2019

Tribut personnel à mon directeur de thèse : le **Professeur émérite Yves Surry (SLU, Académie royale suédoise d'Agriculture**) disparu lors durant la crise sanitaire de la Covid 19.