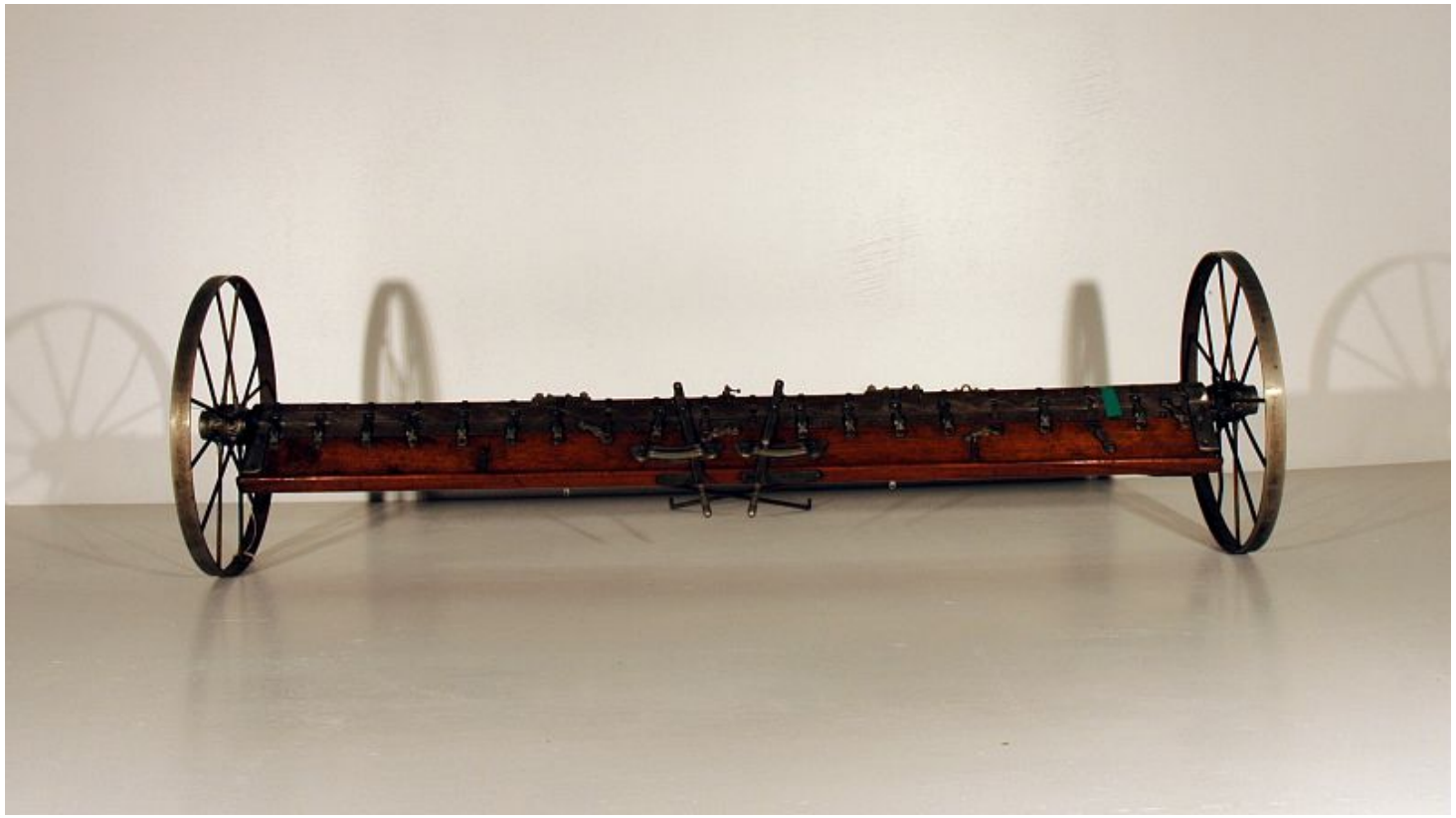
15èmes Journées de Recherche en Sciences Sociales (JRSS) – 9 et 10 Décembre 2021 - Toulouse

Symbolic clustering methods with applications to interval estimates of production costs



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

UMR Economie Publique, INRAE/AgroParisTech, Université Paris Saclay

Estimates of Fertilizer Costs: an Input-Output Methodology

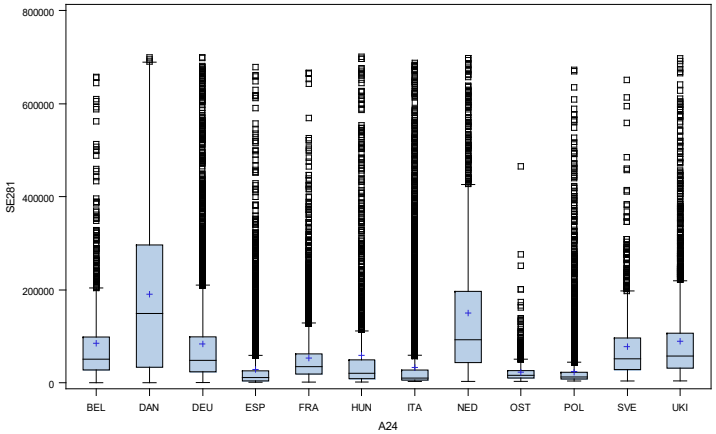
- **Econometric modeling of agricultural production costs : *choice of a model with constant coefficients***

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

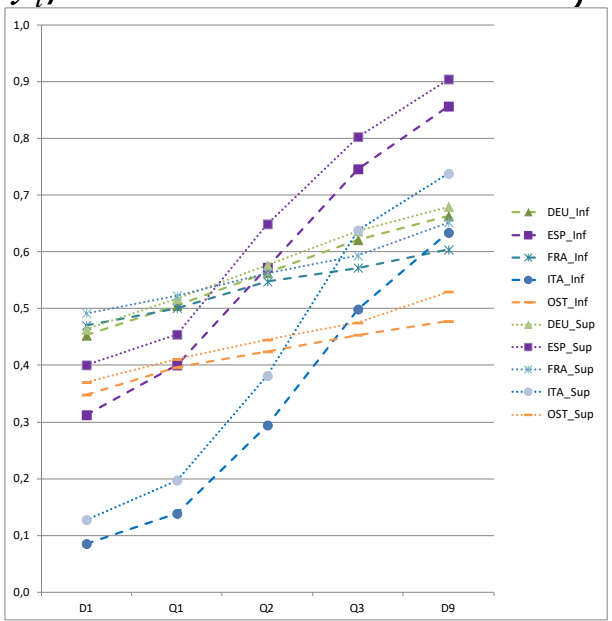
PRODUITS / CHARGES	Y_{1h}	...	Y_{kh}	...	Y_{Kh}	TOTAL CHARGE
X_{1h}	a_{1h}^{1}	...	a_{1h}^{k}	...	a_{1h}^{K}	$\sum X_{1h}$
⋮	⋮		⋮		⋮	⋮
X_{ih}	a_{ih}^{1}	...	a_{ih}^{k}	...	a_{ih}^{K}	$\sum X_{ih}$
⋮	⋮		⋮		⋮	⋮
X_{Ih}	a_{Ih}^{1}	...	a_{Ih}^{k}	...	a_{Ih}^{K}	$\sum X_{Ih}$
TOTAL PRODUIT	$\sum Y_{1h}$	...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

(D.Desbois, 2015)

Quantile Estimates of Production Costs in Agriculture



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



(D.Desbois, 2015)

12 Europeans Countries, Fertilisation Costs for 1K € of Gross Product for Yearly Crops

(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

Belgium:

- € 9 to 19 of fertilizers for a € 1,000 € gross product of yearly crops for D1 decile of fertilizer inputs
- € 23 to 30 of fertilizers for a € 1,000 gross product of yearly crops for Q1 quartile of fertilizer inputs
- € 38 to 47 of fertilisers for a € 1,000 € gross product of yearly crops for median of fertilizer inputs
- € 56 to 80 of fertilizers for a € 1,000 € gross product of yearly crops for Q3 quartile of fertilizer inputs
- € 82 to 110 of fertilizers for a € 1,000 gross product of yearly crops for D9 decile of fertilizer inputs

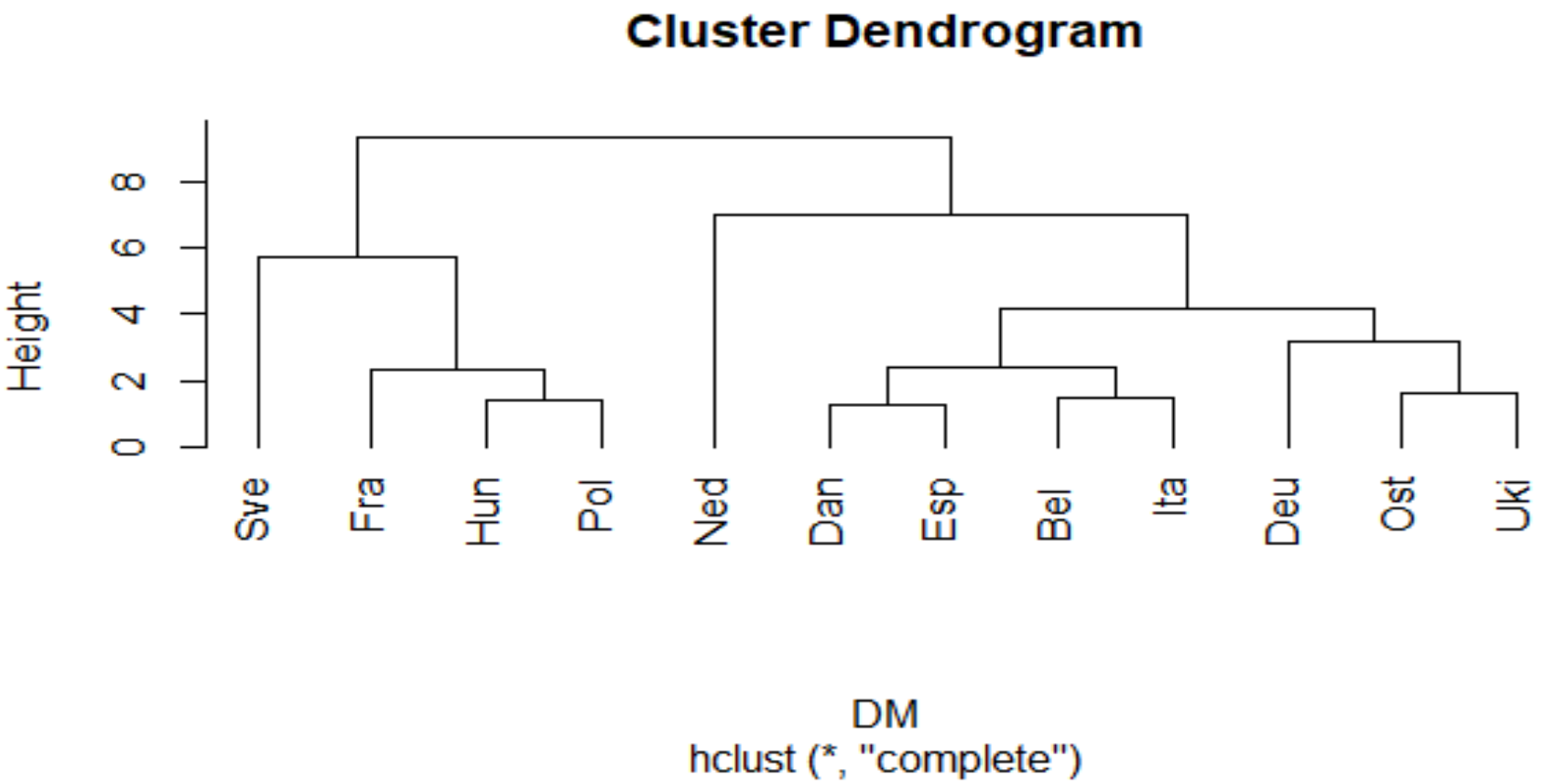
```
syearcrop2<-read.sym.table("~/FERTI/syearcrop2b.txt",header=TRUE,sep='\t',dec=',',row.names=1)
print.data.frame(syearcrop2)
```

The Hausdorff dissimilarity: definition for interval data

- for interval estimates Hausdorff dissimilarity is computed as follows :

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

Symbolic HCA («Hausdorff » option) for Quantile Estimates
The Hierarchical Dendrogram for Countries : **Hausdorff Distance**




```
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)
```

The Hausdorff option of the RSDA sym.dist.interval function is adapted from 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, February 2006, pp.167-179.

The Gowda – Diday dissimilarity: definition for interval data

- The Gowda – Diday dissimilarity is computed as follows for interval estimates :

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

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

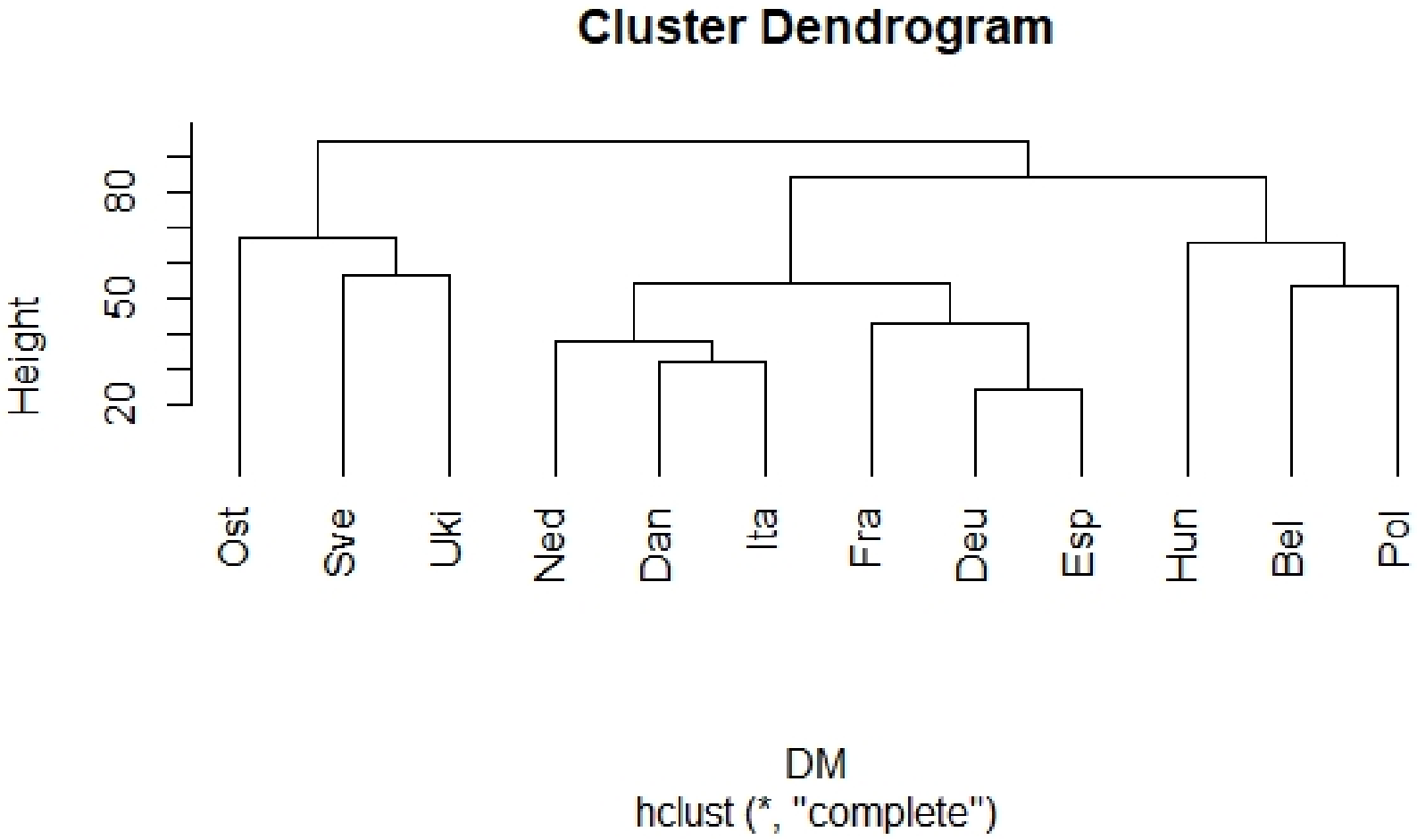
with $\delta_p\left(z_l^q, z_{l'}^q\right) = cos\left[90\left(1 - \frac{\left|\underline{z_l^q} - \underline{z_{l'}^q}\right|}{u^q}\right)\right]$,

u^q being the length of the maximum interval for the q^{th} quantile;

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

Symbolic HCA («Gowda-Diday » option) for Quantile Estimates

The Hierarchical Dendrogram for Countries




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

The Ichino dissimilarity: definition for interval data

- for interval estimates, Ichino dissimilarity is computed as follows :

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

where

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

i.e.

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

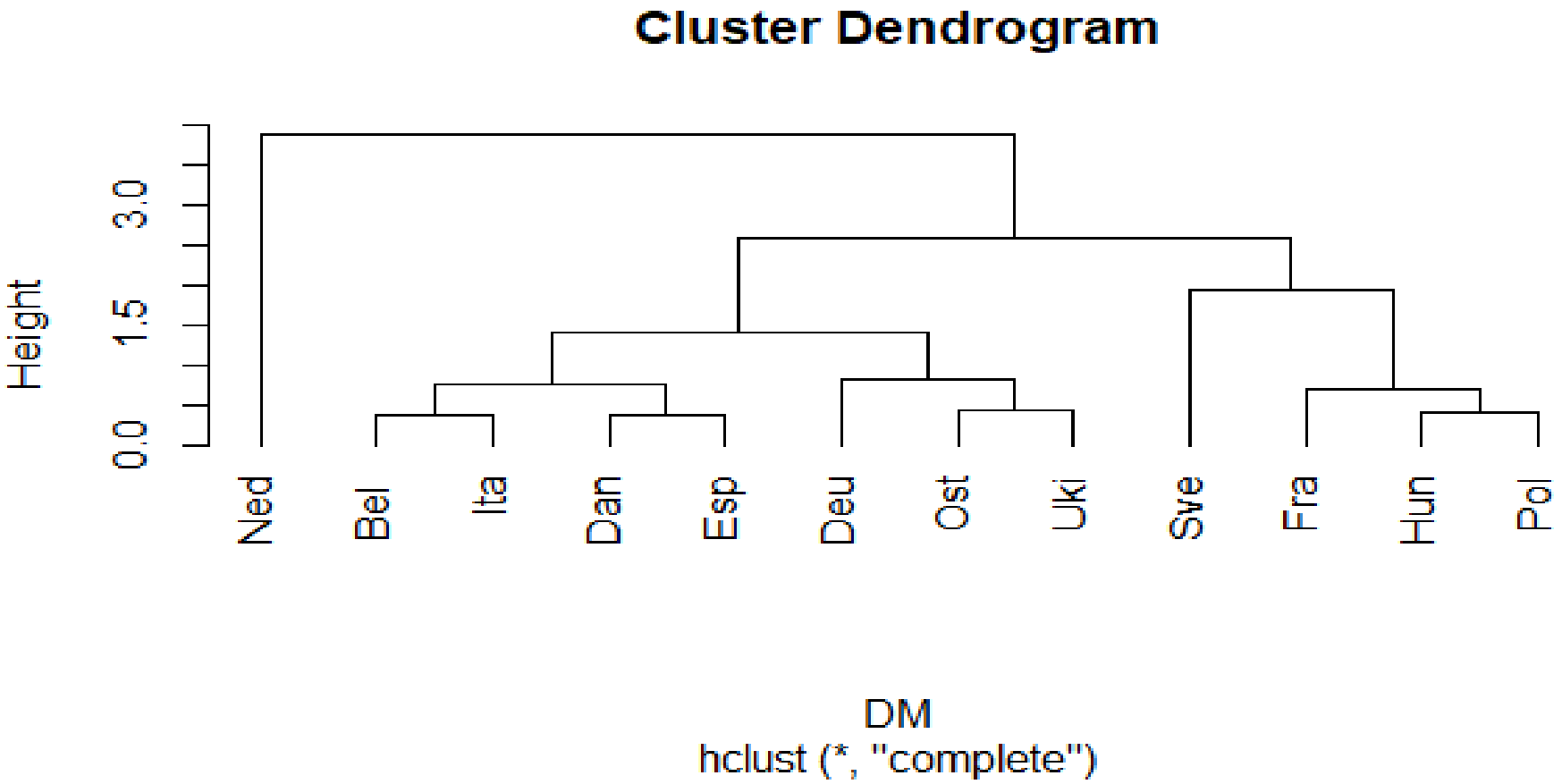
NB: related to Ichino & Yaguchi's distance between sets, with ($\gamma = 0$), based on symmetric difference between sets

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

with R as a normalization term, as *R=1* or $R = \pi(z_l \oplus z_{l'})$, or the overall potential of the q^{th} quantile domain.

Symbolic HCA («Ichino » option) for Quantile Estimates

The Hierarchical Dendrogram for Countries




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

The « Ichino » Option of the RSDA sym.dist.interval function is adaptded from Ichino, M. and Yaguchi, H. (1994): Generalized Minkowski metrics for mixed feature type data analysis. *IEEE Transactions on Systems, Man and Cybernetics*, 24 (4), 698–708.

The Chavent divisive algorithm: definition for interval data

- Generated by the logical binary choice (*yes*/*no*) to a numerical binary question $\Psi = [Is\ z^q \leq c\ ?\ ]$, let us denote $\{A_k, \overline{A_k}\}$ the induced bipartition of a cluster C_k formed of n_k objectss follows :

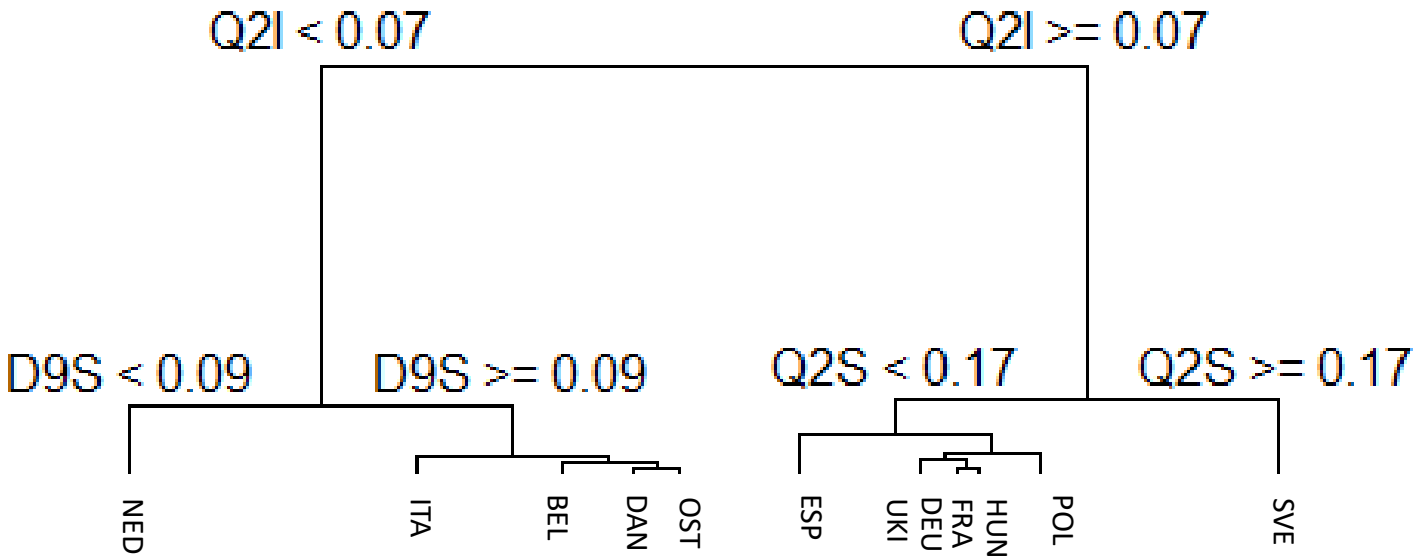
- As in Ward method, the "*upper hierarchy*" of partition P_K is indexed by the height *h* of a cluster C_K , defined by its inter inertia as follows:

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

- The DIVCLUS−T algorithm splits the cluster C_K^* that maximises $h(C_K)$, ensuring that the next partition $P_{K+1} = P_K \cup \{A_K, \overline{A_K}\} - C_K^*$ has the minimum intra inertia value, with respect to the rule

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

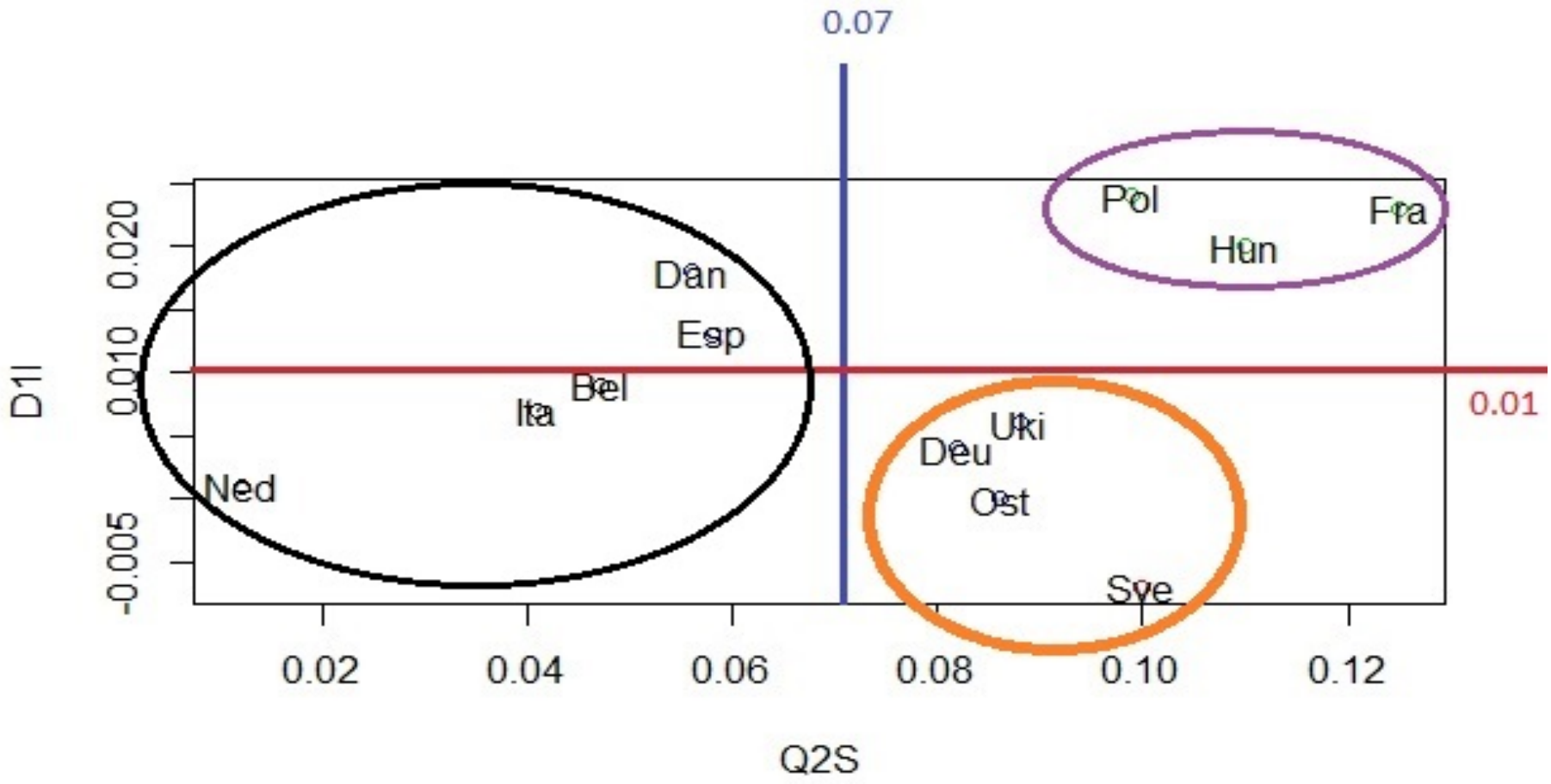
The DIVCLUT divisive clustering results :




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

The divisive clustering divclust function is issued the divclust R library, based on Chavent M., Lechevalier Y., Briant O. (2007) DIVCLUS-T: A monothetic divisive hierarchical clustering method. *Computational Statistics & Data Analysis*, 52, 2, 687-701.

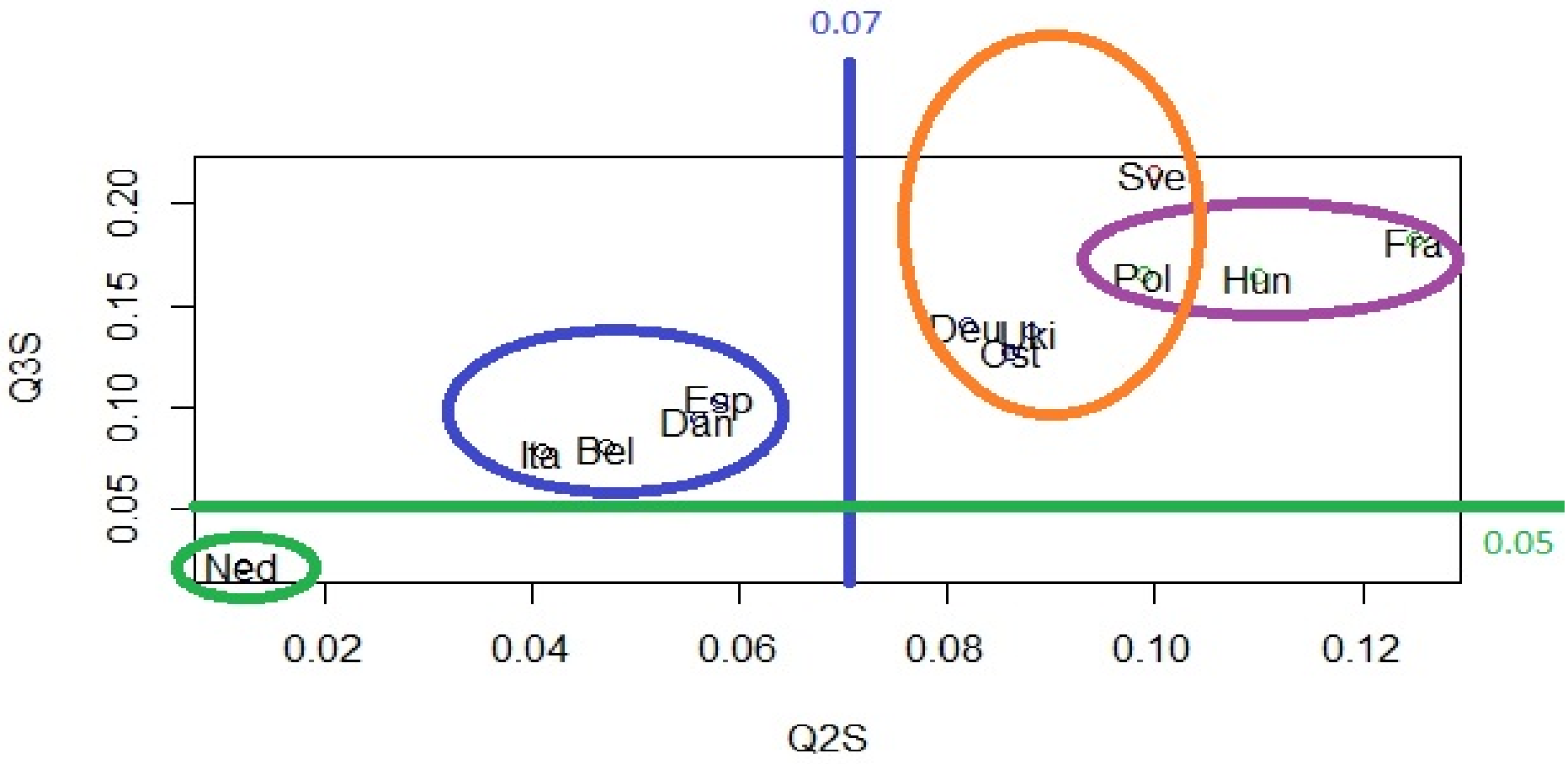
Symbolic Divisive Clustering for Quantile Estimates
The Divisive Tree for Countries : **C3 partition**




```
plot(syear2crop[,9:9],syear2crop[,4:4],col=cl4$cluster, xlab="Q2S",ylab="D1I");
text (syear2crop[,9:9], syear2crop[,4:4], row.names(syear2crop))
```

The divisive clustering divclust function is issued the divclust R library, based on Chavent M., Lechevalier Y., Briant O. (2007) DIVCLUS-T: A monothetic divisive hierarchical clustering method. *Computational Statistics & Data Analysis*, 52, 2, 687-701.

Symbolic Divisive Clustering for Quantile Estimates
The Divisive Tree for Countries : **C4 partition**




```
plot(syear2crop[,9:9],syear2crop[,11:11],col=cl4$cluster, xlab="Q2S",ylab="Q3S");
text (syear2crop[,9:9], syear2crop[,11:11], row.names(syear2crop))
```

C4 is one of the best partitions, accordingly with the criterion of the Determinant Ratio Index = det (T) / det(WG) ,
where WG is the within-group covariance matrix, and Tis the total covariance matrix, BG being the between-group covariance matrix.
BG being the between group covariance matrix.

Determine the optimal clustering

- In order to determine an optimal clustering, we use as the internal quality index for each partition P_K , the log of the determinant ratio computed as follows

  $\varkappa_K = N \log\left(\frac{det(T)}{det(WG^{(K)})}\right)$

- where $T = Z'Z$ is the total scatter matrix (*N* times the total variance-covariance matrix)

  and $WG^{(K)} = \sum_{k=1}^{K} W^{(k)}$ the sum of the within-group scatter matrices,

  $W^{(k)}$ for each group C_k of the partition P_K in *K* groups.

- The optimal score for the quality index is given by the *min_diff* decision rule:

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

  with $\partial_K = \varkappa_{K+1} - \varkappa_K$, using procedure *ClusterCrit*

Clustering Validation : Internal Indices, 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	0,001864802
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	-0,00092087	-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.

References

- 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.
- 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.
- 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.
- Rodrigues Rojas O. (2019) *R to Symbolic Data Analysis: Package 'RSDA'*, Version 3.0, October 21, 2019