Monothetic divisive clustering with geographical constraints

Marie Chavent[1] Yves Lechevallier[2]
Francoise Vernier[3] Kevin Petit[3]

[1] Université Bordeaux2, IMB, UMR 5251 CNRS, France
chavent@math.u-bordeaux1.fr

[2] INRIA, Paris-Rocquencourt 78153 Le Chesnay cedex, France
Yves.Lechevallier@inria.fr

[3] CEMAGREF-Bordeaux, Unité de recherche ADER 50, France
francoise.vernier,kevin.petit@bordeaux.cemagref.fr

COMPSTAT 2008, **Porto, Portugal**

Introduction

- DIVCLUS-T is a divisive and monothetic hierarchical clustering method which proceeds by optimization of a polythetic criterion. The bipartitional algorithm and the choice of the cluster to be split are based on the minimization of the within-cluster inertia.
- C-DIVCLUS-T is an extension of DIVCLUS-T which is able to take contiguity constraints into account. The new criterion defined to include these constraints is a distance-based criterion.

DIVCLUS-T

DIVCLUS-T algorithm repeats the following two steps :

- splitting a cluster into a bipartition which optimizes a criterion W. The complete enumeration is avoided by using a monothetic approch.
- choosing in the current partition the cluster to be split in such a way that the new partition optimizes the criterion W.

$\Rightarrow$ The process stops after a number of iterations specified by the user.
$\Rightarrow$ The output is an indexed hierarchy (dendrogram) which is also a decision tree.

First : How the bipartitional algorithm works ?

The best bipartition is chosen among the set of bipartitions induced by all possible binary questions.

- On a numerical variable X a binary question is noted "$is \quad X \quad \leq \quad c \quad ?$"
- On a categorical variable X a binary question is noted : $is \quad X \quad \in \quad \mathcal{C} \quad ? \Rightarrow$ Note that for numerical variables with complex descriptions like intervals, is is note possible to answer by yes or no to this binary question.

- On a numerical variable X, the number of binary questions is infinite but these binary questions induce a maximum of $n_\ell - 1$ different bipartitions of a cluster C_ℓ with n_ℓ objects.
- On a categorical variable X of m categories, there will be a maximum of $2^{m-1} - 1$ different bipartitions induced
  $\longrightarrow$ computational problem.

Second : how to choose the cluster to split ?

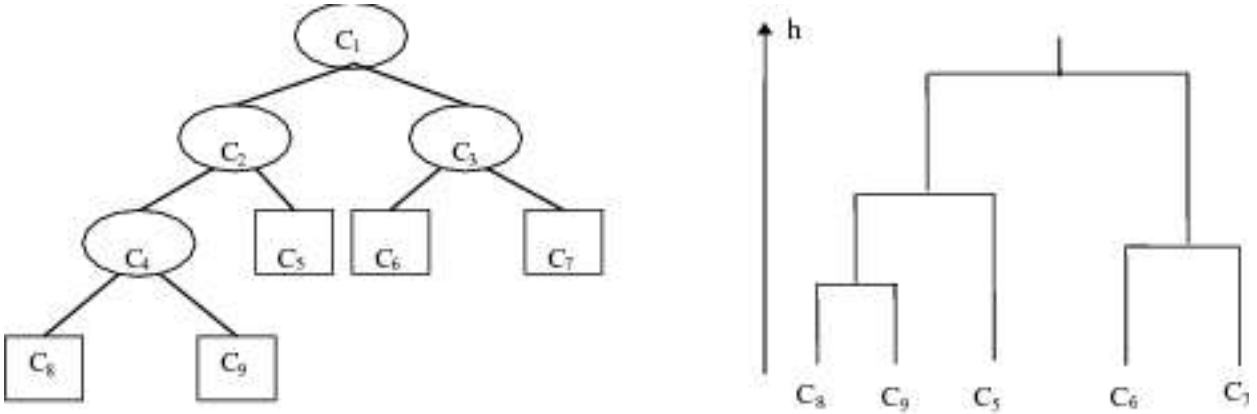
Choose the cluster $C_\ell = A_\ell \cup \bar{A}_\ell$ of P_k such that the partition $P_{k+1} = \{C_1, \ldots, C_{\ell-1}, A_\ell, \bar{A}_\ell, C_{\ell-1}, \ldots, C_k\}$ has the smallest homogeneity criterion $W(P_{k+1})$:

$\Rightarrow$ If the homogeneity criterion $W(P_k)$ is additive :

$$W(P_k) = \sum_{\ell=1}^{k} D(C_\ell)$$

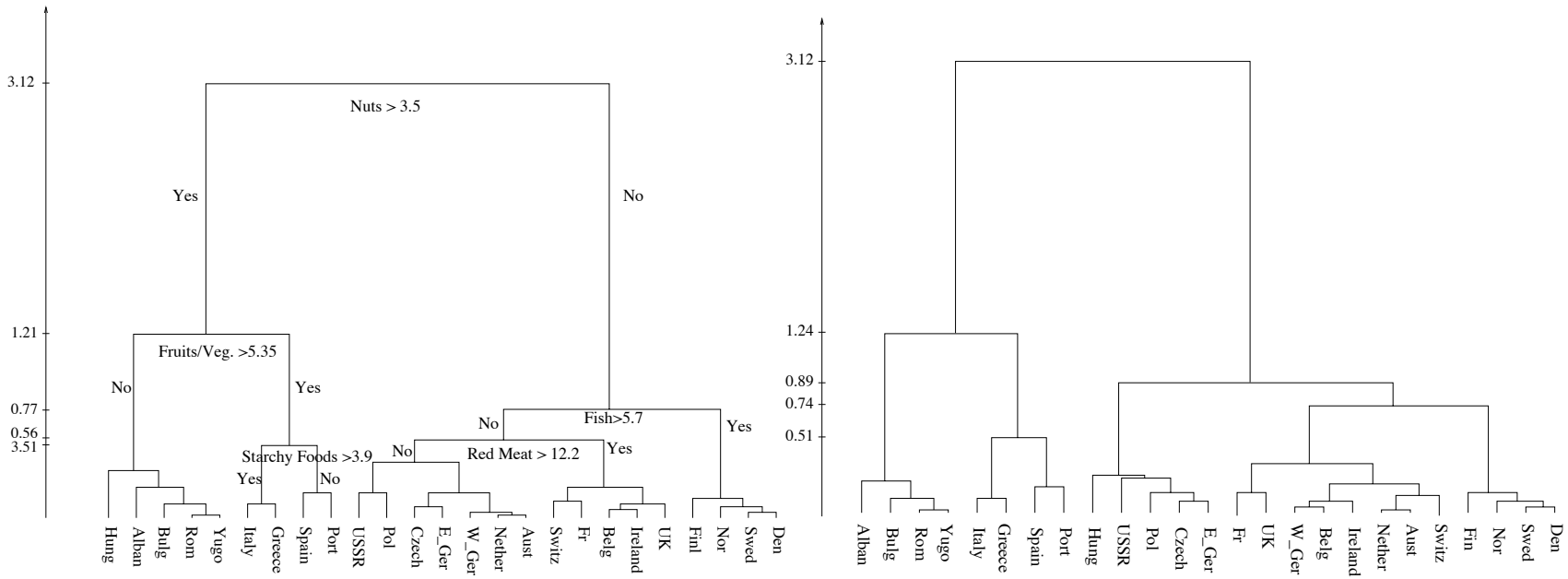
$\Rightarrow$ the cluster C_ℓ chosen maximizes $h(C_\ell) = D(C_\ell) - D(A_\ell) - D(\bar{A}_\ell)$.

Third : how to defined the hierarchical level ?



- The number of divisions is fixed and then the hierarchy is an upper hierarchy.
- The hierarchical level is $h(C_\ell) = D(C_\ell) - D(A_\ell) - D(\bar{A}_\ell)$

DIVCLUS-T : a simple example



DIVCLUS-T : a simple example

What is the price paid in term of inertia for this supplementary monothetic interpretation ?

Proportion of the inertia (in %) explained by the k-clusters partitions obtained with DIVCLUS-T and Ward on the protein data set :

k	2	3	4	5	6	7	8	9	10
DIVCLUS-T	37.1	50.6	59.2	65.5	71.2	73.5	79.3	81.6	84
Ward	34.7	48.5	58.5	66.7	72.4	75.5	79	81.6	84

Chavent, M., Briant, O., Lechevallier, Y. (2007). DIVCLUS-T : a monothetic divisive hierarchical clustering method. *Computational Statistics and Data Analysis*, **32** (2), 687-701.

A distance-based homogeneity criterion

how to define an homogeneity criterion when the data have complex descriptions ?

Let $\mathbf{D} = (d_{ii'})_{n \times n}$ be the distance matrix.

- A distance-based homogeneity criterion D of a cluster C_ℓ can be defined by :

$$D(C_\ell) = \sum_{i \in C_\ell} \sum_{i' \in C_\ell} \frac{w_i w_{i'}}{2\mu_k} d^2_{ii'} \text{ with } \mu_k = \sum_{i \in C_k} w_i$$

A distance-based homogeneity criterion

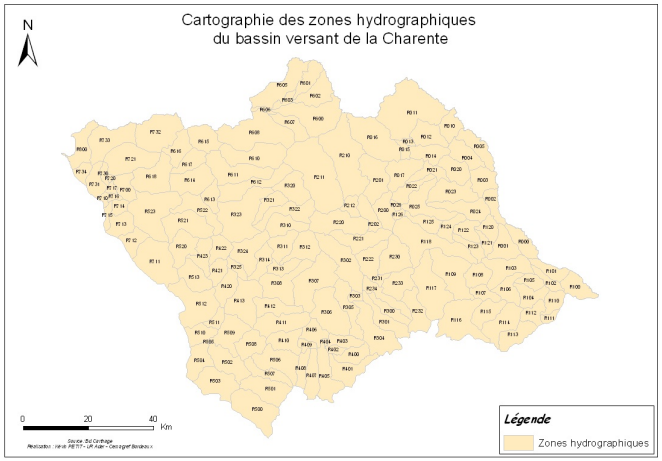
- A distance-based homogeneity criterion W of a partition P_k can be defined by :

$$W(P_k) = \sum_{\ell=1}^{k} D(C_\ell)$$

- $W(P_k)$ is the within-cluster inertia criterion for classical numerical data and the Euclidean distance

Analysis of symbolic data, Ed. H.H.Bock, E. Diday, Springer.

A new distance-based criterion



The geographical constraints are represented in an adjacency matrix $Q = (q_{ii'})_{n \times n}$ where

$$\begin{aligned} q_{ii'} &= 1 \text{ if } i' \text{ is a neighbor of } i \\ q_{ii'} &= 0 \text{ otherwise.} \end{aligned}$$

A new distance-based homogeneity criterion

- We have

$$D(C_\ell) = \sum_{i \in C_\ell} \sum_{i' \in C_\ell} \frac{w_i w_{i'}}{2\mu_k} d^2_{ii'} = \sum_{i \in C_\ell} \frac{w_i}{2\mu_k} D_i(C_\ell)$$

with

$$D_i(C_\ell) = \sum_{i' \in C_\ell} w_{i'} d^2_{ii'}$$

which measures the proximity between the object i and the cluster C_ℓ to which it belongs.

- We define a new homogeneity criterion $\tilde{D}(C_\ell)$ by defining a new criterion $\tilde{D}_i(C_\ell) = \alpha a_i(C_\ell) + (1-\alpha) b_i(C_\ell)$ with $\alpha \in [0, 1]$.
- The new distance-based criterion is $\tilde{W}_\alpha(P_k) = \sum_{\ell=1}^{k} \tilde{D}(C_\ell)$

A new distance-based criterion

In the criterion

$$\tilde{D}_i(C_\ell) = \alpha a_i(C_\ell) + (1-\alpha) b_i(C_\ell),$$

the first part

$$a_i(C_\ell) = \sum_{i' \in C_\ell} w_{i'}(1 - q_{ii'}) d^2_{ii'}$$

measures the coherence or the dissimilarity between i and its cluster C_ℓ. It it small when i is similar to the objects in C_ℓ ($d_{ii'} \approx 0$) and when these objects are neighbor of i ($q_{ii'} = 0$).

A new distance-based criterion

In the criterion

$$\tilde{D}_i(C_\ell) = \alpha a_i(C_\ell) + (1-\alpha) b_i(C_\ell),$$

the second part

$$b_i(C_\ell) = \sum_{i' \notin C_\ell} w_{i'} q_{ii'} (1 - d_{ii'}^2)$$

measures the coherence between i and the objects which are not in C_ℓ. It is small when i is dissimilar from the objects which are not in C_ℓ ($d_{ii'} \approx 1$) and when the objects which are note in C_ℓ are not neighbors of i ($q_{ii'} = 0$). In other words $b_i(C_\ell)$ represents a penalty for the neighbors of i which belongs to other clusters.

Study of the parameter α

The parameter α can be chosen by the user (usually, $\alpha = 0.5$)

- if $\alpha = 1$ then $\tilde{W}_1(P_n) = 0$ and for k we have :

$$\tilde{W}_1(P_k) = \sum_{\ell=1}^{k} \sum_{i \in C_\ell} \sum_{i' \in C_\ell} \frac{w_i w_{i'}}{2\mu_\ell} (1 - q_{ii'}) d_{ii'}^2,$$

- if $\alpha = 0$ then $\tilde{W}_0(P_1) = 0$ and for k we have :

$$\tilde{W}_0(P_k) = \sum_{\ell=1}^{k} \sum_{i \in C_\ell} \sum_{i' \notin C_\ell} \frac{w_i w_{i'}}{2\mu_\ell} q_{ii'} (1 - d_{ii'}^2),$$

Automatic choice of α

The parameter α can be chosen automatically such that $\tilde{W}_\alpha(P_1) = \tilde{W}_\alpha(P_n)$. The parameter α is then equal to :

$$\alpha = \frac{A}{A+B}$$

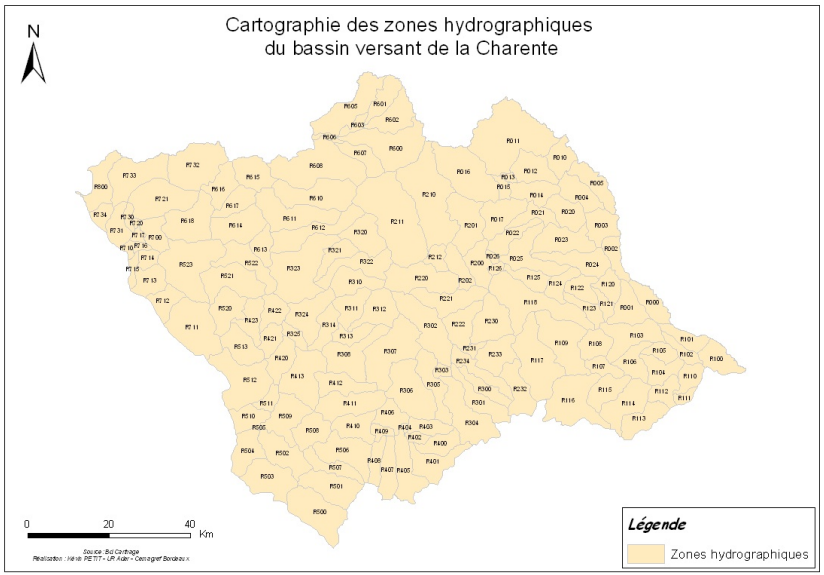
where

$$A = \sum_{i \in \Omega} \sum_{i' \in \Omega, i \neq i'} q_{ii'}(1 - d^2_{ii'}),$$

$$B = \sum_{i \in \Omega} \sum_{i' \in \Omega} (1 - q_{ii'}) d^2_{ii'}.$$

Hydrological areas clustering

- A study is carrying out at Cemagref in the context of the SPICOSA (web site : www.spicosa.eu) project
- The purpose is to define the relevant spatial unit, helpfull for the integrated managment of the “Charente river basin”.
- Find a partition of the 140 hydrological units within the studied area



Hydrological areas clustering

- The 140 hydrological units are characterized on :
  - 14 types of soils,
  - 17 types of soil occupation,
  - 8 main crops, a mean slope and a drainage rate.

Zhydro	Type of soil				Soil occupation				Crope				Mean slope
	S_1	S_2	. . .	S_{14}	O_1	O_2	. . .	O_{17}	C_1	C_2	. . .	C_8	
R000	12	22	. . .	7.8	9.8	12.6	. . .	9.4	12	8.7	. . .	32.1	4.44
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.

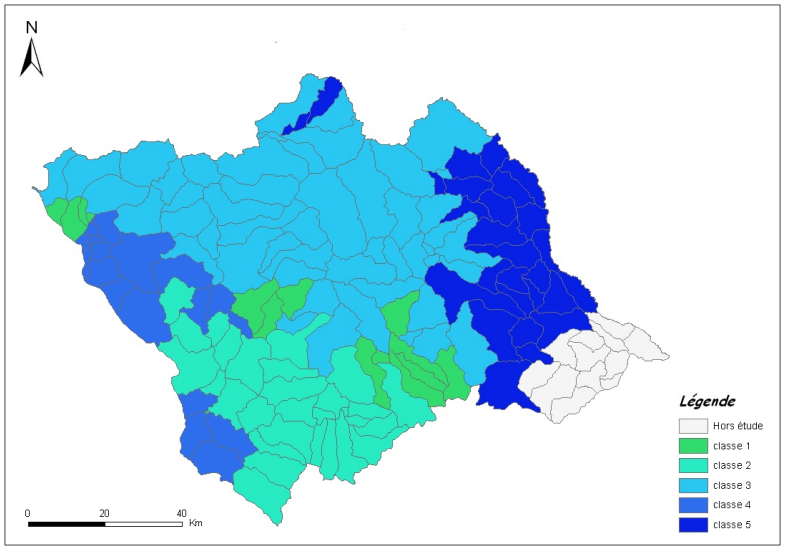
- Two files :
  - the first file includes the descriptions of the 140 hydrological units
  - the second file includes for each hydological area the list of its neighbors

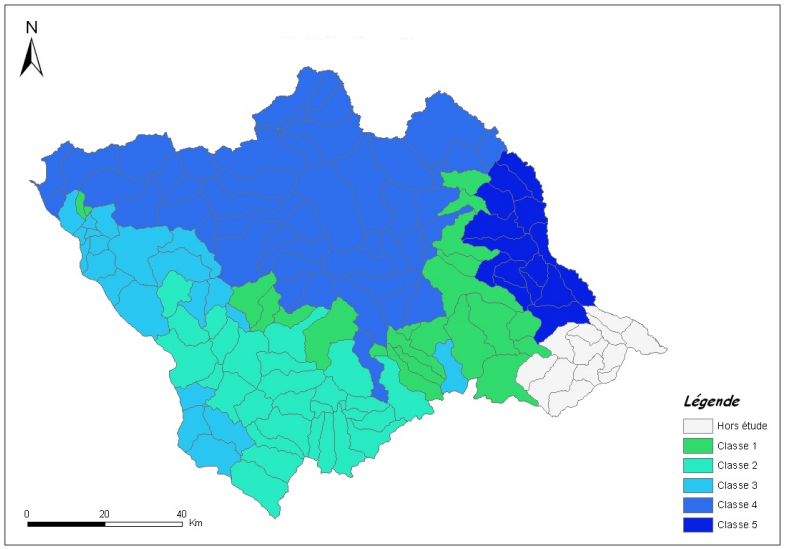
Hydrological areas clustering

- The DIVCLUS-T method has been applied to the first data file
- C-DIVCLUS-T has been applied to the same data file taking into account the contiguity of the data given in the neighbors file
- The five-clusters partition has been retained in both cases

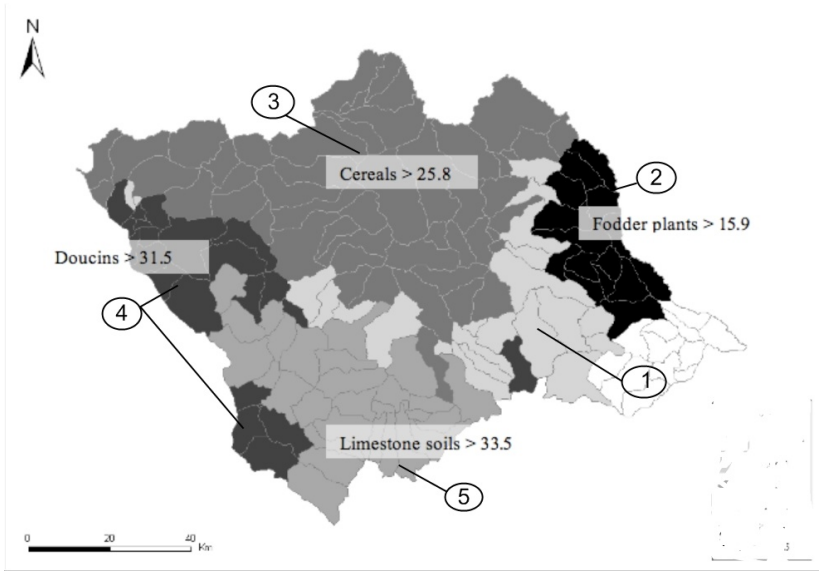
Results with DIVCLUS-T and C-DIVCLUS-T

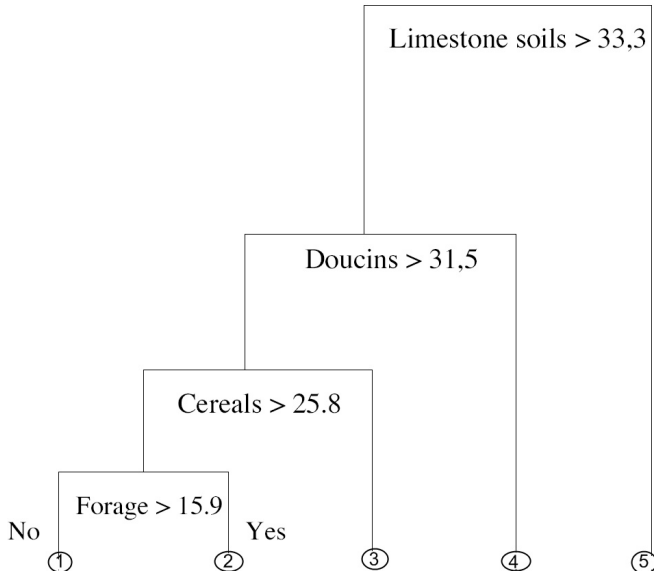
The maps give the clusters obtained by DIVCLUS-T and C-DIVCLUS-T on the Charente basin





Results with C-DIVCLUS-T





- A part of the coastal area can be linked to the presence of Doucins soils (moors).
- In the North of the river basin, an homogeneous area with cereal crops stands out.
- An other relevant area is delimited in the South of the basin with the variable limestone soils : we can find here vineyards and complex cultivation patterns.
- The cluster 1 can be linked to more artificialised areas.

Conclusion

- A first trial of taking contiguity constraints into account in the clustering of this dataset,
- Many other approaches exist and may by used,
- The advantage of C-DICVLUS-T remains its monothetic aspect and the distance based criterion which is able to deal with data having complex descriptions