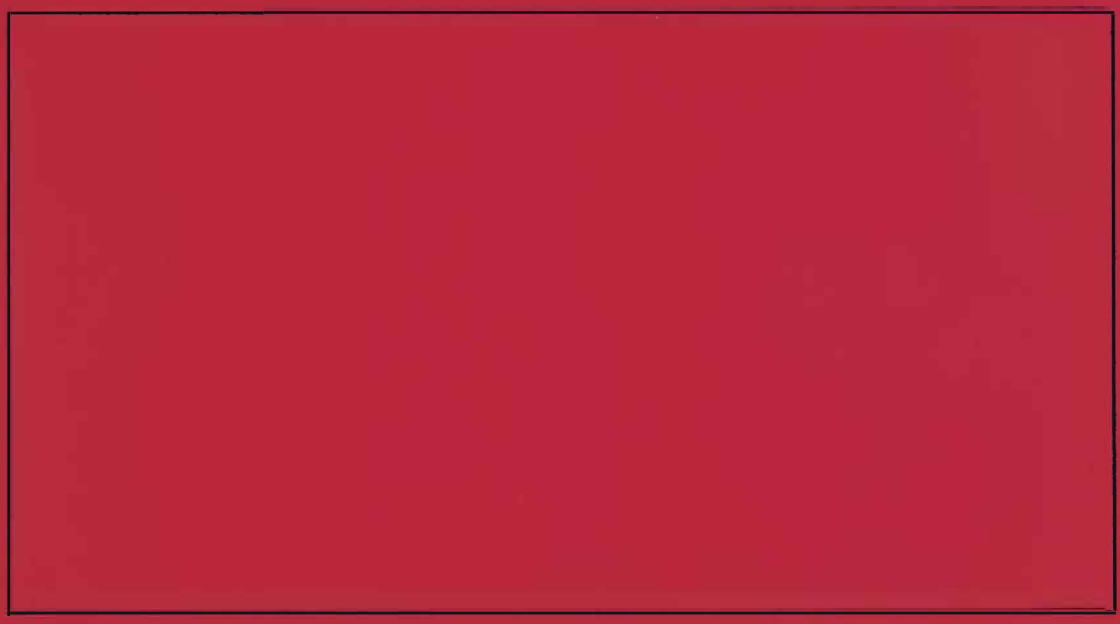
SEGSOFT

Bernard Fustier

I.M.E.

EQUIPE DE RECHERCHE ASSOCIEE AU C.N.R.S.

DOCUMENT DE TRAVAIL



INSTITUT DE MATHEMATIQUES ECONOMIQUES

UNIVERSITE DE DIJON

FACULTE DE SCIENCE ECONOMIQUE ET DE GESTION

4, BOULEVARD GABRIEL — 21000 DIJON

Nº 95

SEGSOFT

Bernard FUSTIER

Janvier 1987

INTRODUCTION

Le champ d'application privilégié de l'analyse des données concerne des tableaux de valeurs exprimées dans le même système d'unité. Disposant le plus souvent d'un ensemble d'observations décrites par des variables de natures différentes, l'économiste est obligé - s'il veut recourir aux techniques d'analyse factorielle descriptive ou de classification automatique - d'effectuer certaines opérations préalables sur les données brutes.

La procédure d'homogénéisation la plus couramment utilisée consiste à normaliser l'information portée par chaque variable. Les variables (ou descripteurs) étant désignées par les colonnes du tableau, il suffit de diviser les données contenues par une colonne, par la plus grande valeur rencontrée dans cette dernière. Cette opération est reproduite pour chaque variable du tableau. Les observations, disposées en lignes, sont alors décrites par des valeurs de l'intervalle $[0, 1]$ commun à chaque colonne.

Cette transformation du tableau présente l'avantage de la simplicité et ne modifie pas les ordres de grandeur constatés initialement dans les colonnes. Observons, cependant, qu'elle ne légitime pas l'emploi de n'importe quelle technique d'analyse des données (analyse factorielle des correspondances, par exemple). Pour limiter les risques d'interprétation des résultats, nous pensons qu'il est préférable d'adapter la technique à la nouvelle forme de l'information.

La technique exposée dans ce document est un algorithme de classification hiérarchique descendante.

Le principe sur lequel il repose est classique :

- disposant d'un ensemble fini I d'observations, le problème de classification consiste à définir une partition sur cet ensemble compte tenu d'un ensemble J de descripteurs également fini.

- la solution retenue propose un système de classes emboîtées ou hiérarchie.

- dans le cas présent, cette hiérarchie est un système dichotomique : les éléments de I sont séparés en deux classes complémentaires ou segments, puis chaque segment est scindé en deux autres sous-ensembles, et ainsi de suite ; la classification des observations s'opère par voie descendante.

La difficulté du problème réside dans la séparation des observations par des facteurs qui ne sont pas dichotomiques. En effet, la description des observations n'est pas d'essence binaire, les données normalisées estiment les degrés d'adhésion exprimés, plus ou moins intensément, par chaque observation aux prédicats sous-entendus par les descripteurs. En d'autres termes, on considère qu'un descripteur donné définit un sous-ensemble flou de I.

Par exemple, si l'on utilise la taille comme descripteur d'un ensemble d'individus, la personne associée à une valeur (taille normalisée) proche de 1 appartient davantage au sous-ensemble flou des "grandes tailles" qu'une personne caractérisée par une valeur proche de 0. Dans ces conditions, on peut décider que la première personne sera affectée au segment caractérisé par le prédicat "grandes tailles" tandis que la seconde appartiendra au segment complémentaire des "non-grands". Bien entendu la décision devient plus hésitante lorsque les valeurs avoisinent 0,50.

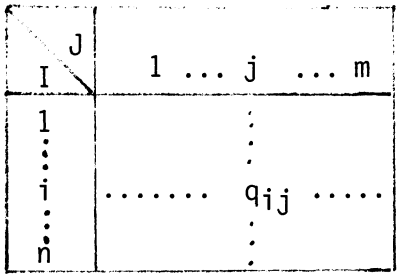
Il faudra donc convenir d'une règle générale pour affecter les observations dans des classes obtenues sur la base de prédicats vagues. Cette règle fait intervenir ici un seuil de sévérité dont le rôle est analogue au seuil de concordance utilisé dans l'analyse des choix multicritères. Elle n'est pas une "relation floue", mais une règle souple (le seuil de sévérité pouvant être modifié au gré du décideur, dans certaines limites cependant).

La référence à la théorie des sous-ensembles flous ne concerne que les prédicats révélés par les descripteurs. L'opération de normalisation ne saurait remettre en cause la nature des données brutes supposées exactes.

Le traitement des tableaux rectangulaires de données imprécises (nombres flous) est un autre problème qui ne sera pas abordé dans le présent travail.

1. OBJECTIF GENERAL.

Considérons le tableau de valeurs numériques :



où :

I représente un ensemble d'observations (ou : objets)

J désigne un ensemble de facteurs (ou : descripteurs)

q_{ij} : est la mesure effectuée sur la ième observation dans le système d'unité correspondant au jème descripteur.

Le problème consiste à définir une suite de partitions sur I en faisant intervenir, à chaque étape de la procédure, les facteur "les plus représentatifs" de J. Procédant par dichotomies successives de l'ensemble à classer, la méthode de résolution proposée ici est un algorithme de classification descendante : des n observations considérées comme formant un tout homogène, on s'achemine progressivement vers des partitions de plus en plus fines de l'ensemble initial.

L'idée est la suivante :

1) Supposons que j° soit le facteur "le plus représentatif" de J (en ce sens qu'il constitue une sorte de résumé des autres descripteurs). Une première subdivision de I est obtenue en rangeant :

- dans une classe, les observations les "mieux caractérisées" par j°.
- dans la classe complémentaire, les observations qui apparaissent "moins bien caractérisées" par ce facteur

N.B. Les classes ainsi définies sont appelées segments et correspondent au niveau 1 de la segmentation.

2) On procède de la même façon sur chaque segment du niveau 1 en choisissant le facteur "le plus représentatif" dans J - $\{j^0\}$.

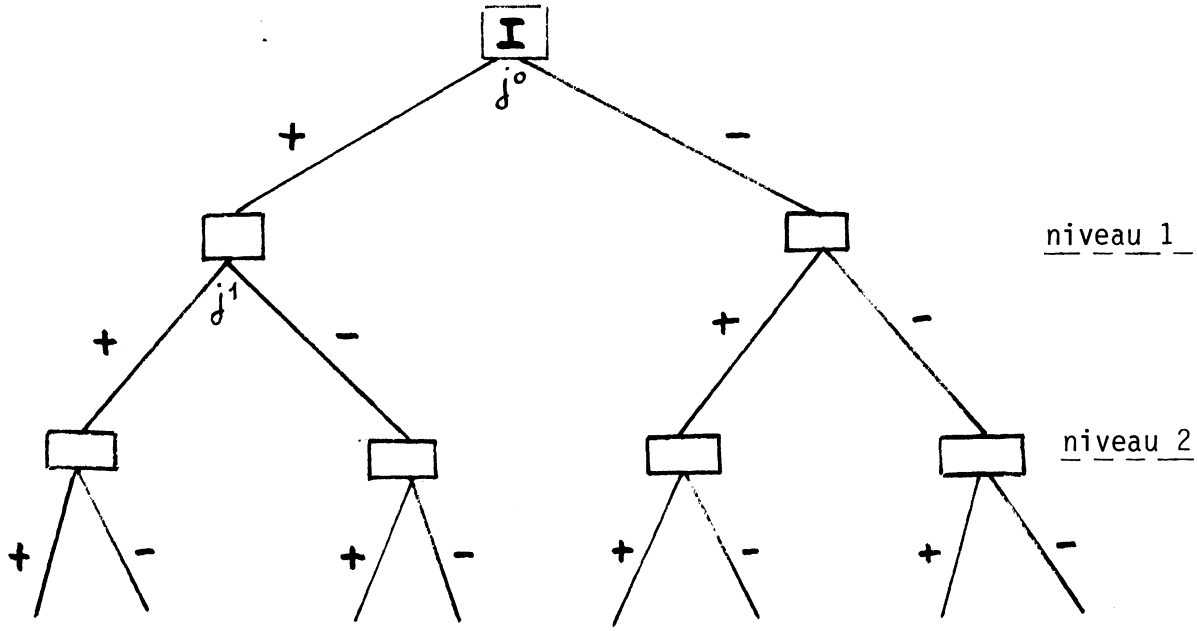
Et ainsi de suite.

N.B. A chaque étape, il faut naturellement exclure de J, le facteur ayant conduit à la dichotomie précédente.

En symbolisant les notions d'observations :

- les "mieux caractérisées" par le signe +
- les "moins bien caractérisées" par le signe -

Le schéma suivant représente les partitions obtenues à chaque niveau de la procédure



La partition sur I est d'autant plus fine que le niveau de la segmentation est numériquement élevé ; les observations appartenant à un même segment présentent un degré d'homogénéité important tandis que les segments sont fortement différenciés entre-eux.

Laissant en suspend, pour l'instant, le problème du choix des facteurs "les plus représentatifs", nous réservons la section suivante à la définition d'une règle d'appartenance des observations aux segments issus de chaque dichotomie.

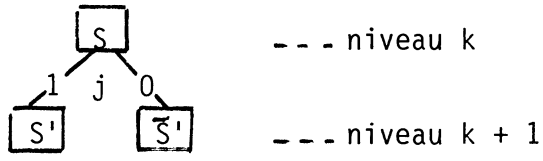
2 - NORMALISATION - SEPARATION

2.1. Cas particulier

La méthode de WILLIAMS et LAMBERT (1959) est l'exemple typique de l'algorithme esquissé ci-dessus. S'appliquant à un tableau de description logique (ne comportant que des 0 ou 1) l'appartenance d'une observation à un segment ou à son complémentaire est définie au sens de la théorie des ensembles, c'est-à-dire sans ambiguité.

En effet, lorsque les facteurs ne comportent que deux modalités exprimant respectivement une propriété donnée et son contraire (présence/absence; oui/non, etc...), la description des éléments de I se prête naturellement à un codage binaire ($q_{ij} = 1$ si i possède la propriété révélée par j; $q_{ij} = 0$ dans le cas contraire).

A n'importe quel niveau de segmentation, la subdivision d'un segment S en deux sous-ensembles complémentaires S' et $\bar{S}'$ est automatique :



la séparation des éléments de S est fondée sur la règle suivante

$$(1)\quad \forall i \in S \begin{cases} i \in S' \Leftrightarrow q_{ij} = 1 \\ \text{ou} \\ i \notin S' \Leftrightarrow q_{ij} = 0 \quad \text{donc } i \in \bar{S}' \end{cases}$$

2.2. Cas général.

L'ensemble J retenu dans la section 1 se réfère à des descripteurs non nécessairement dichotomiques et de natures différentes :

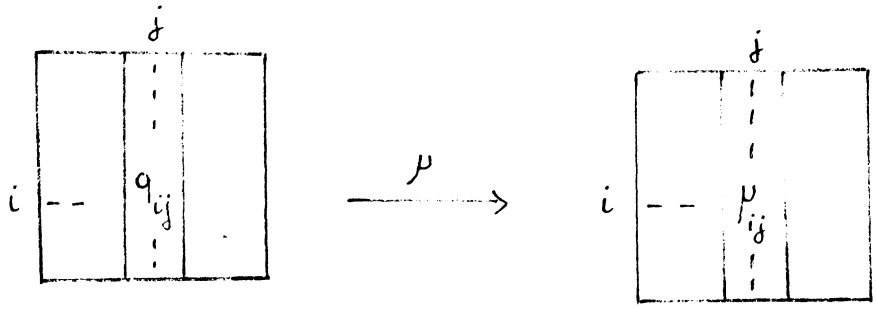
- variables qualitatives pouvant comporter un nombre de modalités supérieur à deux (très bien/bien/passable/mauvais...); les équivalents numériques associés à chaque modalité doivent donc être choisis dans un ensemble plus vaste que $\{0, 1\}$.
- variables quantitatives et continues (revenu par tête, densité de population etc...).

La non-négativité est la seule condition imposée aux valeurs du tableau I x J.

L'homogénéisation des données s'effectue en normalisant chaque colonne du tableau :

$$(2)\quad \forall j \in J : \mu_{ij} = q_{ij} / \max_{i} q_{ij}$$

$$(i = 1...n)$$



données brutes — données normalisées

Par définition :

$$(3)\quad \forall i \in I : u_{ij} \in [0, 1]$$

$$(j = 1...m)$$

(4) $\forall j \in J$: il existe au moins une observation i' de I telle que $\mu_{i'j}=1$

Cette opération de normalisation, adoptée par d'autres auteurs (voir par exemple PONSARD et TRANQUI 1985, ROLLAND-MAY 1985), permet de généraliser l'interprétation des tableaux de description logique en adoptant le point de vue de la théorie des sous-ensembles flous.

En effet, l'application μ définie par (1) peut être considérée comme une fonction d'appartenance dont la spécification dépend du référentiel choisi pour introduire la notion de sous-ensemble flou (s.e.f.)

1) Les auteurs pré-cités conviennent de choisir l'ensemble des descri- esc teurs comme référentiel et assimilent, par conséquent, chaque observation de I à un s.e.f. de J.

En notant $\underset{\sim}{A}_i$ le s.e.f. de J correspondant à la ième observation, on

obtient :

$$(5)\ \underset{\sim}{A}_i = \left\{(j, \mu_{ij})\ ;\ \forall_j \in J\ ;\ \mu_{ij} \in [0,1]\right\}$$

μ_{ij} s'interprète comme le degré d'appartenance du descripteur j au s.e.f. $\underset{\sim}{A}_i$.

Chaque descripteur caractérise "plus ou moins" l'observation i.

2) Dans le cas présent, le référentiel est donné par l'ensemble des observations dans la mesure où l'on cherche à apprécier le degré d'adéquation des éléments de I au prédicat correspondant à chaque descripteur.

Le s.e.f. $\underset{\sim}{A}_j$ associé au jème descripteur a pour expression :

$$(6)\ \underset{\sim}{A}_j = \left\{(i, \mu_{ij})\ ;\ \forall i \in I\ ;\ \mu_{ij} \in [0, 1]\right\}$$

μ_{ij} représente donc le degré d'appartenance de l'observation i au s.e.f. $\underset{\sim}{A}_j$; il exprime numériquement la façon dont le prédicat sous-entendu par le descripteur j s'applique à l'observation i.

Considérons, par exemple, un prédicat tel que "peuplé". Ce prédicat est sous-entendu par la "densité de population (descripteur). Dans un pays composé de quatre régions ($I = \{1, 2, 3, 4\}$)on observe les résultats suivants :

1 : 36h/km^2
2 : 235h/km^2
3 : 128h/km^2
4 : 142h/km^2

auxquels correspond le s.e.f. des "régions peuplées" :

$$\{(1;\ 0,15)\ (2;1)(3;0,54)(4;0,60)\}$$

On observera que l'appartenance d'une région donnée au s.e.f. est relative à l'observation caractérisée par la plus forte valeur du descripteur et dont le degré d'appartenance est, par définition, égal à 1.

Rigoureusement (6) est un s.e.f. normalisé (Prade 1982 p. 76) : d'après (4), il existe au moins un élément du référentiel possédant un degré d'appartenance égal à l'unité ; cet élément - non nécessairement unique - est "le plus fortement" caractérisé par j.

Les observations "les mieux" caractérisées par j peuvent, à un seuil extrême de sévérité, se ramener aux observations "les plus fortement" caractérisées par ce facteur. Dans ce cas, la règle de séparation des éléments d'un segment donné S sur la base d'un facteur j est analogue à (1) :

$$(7)\quad \forall i \in S \begin{cases} i \in S' \Leftrightarrow \mu_{ij} = 1 \\ \text{ou} \\ i \notin S' \Leftrightarrow \mu_{ij} < 1 \qquad \text{donc } i \in \bar{S}' \end{cases}$$

D'une manière plus générale, on conviendra de choisir un seuil de sévérité, noté p, tel que :

$$(8)\quad p \in]0{,}5 \quad 1]$$

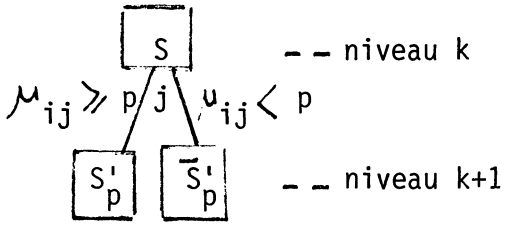
La valeur du seuil étant fixée, la relation (7) s'étend à :

$$(9)\quad \forall i \in S \begin{cases} i \in S'_p \Leftrightarrow \mu_{ij} \geqslant p \\ \text{ou} \\ i \notin S'_p \Leftrightarrow \mu_{ij} < p \qquad \text{donc } i \in \bar{S}'_p \end{cases}$$

où :

S'_p contient les observations de S "les mieux caractérisées" par j, c'est-à-dire dont les degrés d'appartenance sont au moins égaux au seuil de sévérité exigé.

$\bar{S}'_p$ regroupe, par conséquent, les observations les "moins bien caractérisées "dans le contexte de sévérité ainsi défini.



Evidemment, lorsque la valeur de p diminue, les observations les "mieux caractérisées" (respectivement : les "moins bien caractérisées") deviennent plus nombreuses (resp. moins nombreuses) :

$$(10)\quad p_2 < p_1 \Rightarrow S'_{p_2} \supseteq S'_{p_1} \text{ et } \bar{S}'_{p_2} \subseteq \bar{S}'_{p_1}$$

La mise en oeuvre de l'algorithme suppose que la valeur de p soit préalablement fixée. Débutant avec une valeur égale à 1 (ou très proche), la procédure est réitérée en atténuant à chaque fois le seuil de sévérité (voir l'exemple développé en section 4.)

3 - REPRESENTATIVITE - PROXIMITE

Considérons un segment $S \subseteq I$ dont les éléments sont décrits par les facteurs de $K \subseteq J$.

S'agissant de séparer les observations contenues par S sur la base d'un facteur unique, il conviendra de choisir un facteur représentatif de K, un facteur dont le profil résume au mieux les descriptions établies par les autres facteurs.

De nombreux indicateurs peuvent être proposés pour quantifier le degré d'association d'un facteur à tous les autres (WILLIAMS et LAMBERT utilisent le (chi-deux pour les facteurs dichotomiques). On retiendra ici le critère de moindre distance, le facteur le "plus proche" étant supposé introduire le maximum de discrimination entre les observations (pour un seuil de sévérité donné).

La distance entre deux éléments quelconques j et j' de K, notée d(j,j'), a pour expression :

$$(11)\quad d(j,j') = \sum_{i \in S} |\mu_{ij} - \mu_{ij'}|$$

Par extension, la distance, notée D(j), de j à tous les autres élements de K s'écrit :

$$(12)\quad D(j) = \sum_{\substack{j' \in K \\ j' \neq j}} d(j,j')$$

Dans ces conditions, le facteur qui, jusqu'à présent, a été appelé le "plus représentatif", est l'élément de K, noté, j^*, tel que

$$(13)\quad D(j^*) = \min_{j \in K} D(j)$$

En termes équivalents, le s.e.f. $\underset{\sim}{A}_{j^*}$ correspondant à j^* est le s.e.f. "le plus proche" des autres s.e.f. définis sur S (non nécessairement normalisés si $S \subset I$). (11) représente, en effet, la distance de HAMMING généralisée entre $\underset{\sim}{A}_j$ et $\underset{\sim}{A}_{j'}$ (KAUFMANN, 1973). Dans le cas particulier d'un tableau de description logique, d(j, j') mesure la distance entre les sous-ensembles ordinaires A_j et $A_{j'}$.

4 - DESCRIPTION DE L'ALGORITHME A PARTIR D'UN EXEMPLE SIMPLIFIE

4.1. Exemple

I représente les 12 régions (1, 2,...,12) d'un territoire donné que l'on visualise de la façon suivante :

1	2	3	4
5	6	7	8
9	10	11	12

L'ensemble J comprend les descripteurs suivants :

. R : revenu mensuel moyen (milliers de F)
. C : taux de chômage (%)
. V : valeur ajoutée par tête (milliers de F)
. D : densité de population (nombre d'habitants au km^2)
. A : aéroport (facteur dichotomique : oui = 1 non = 0)

Données brutes :

	R	C	V	D	A
1	14	0	120	250	1
2	10	3	110	112	0
3	8	8	80	35	0
4	7	12	60	35	0
5	14	1	110	210	1
6	10	2	100	150	0
7	8	10	90	38	0
8	7	15	50	35	0
9	15	0	140	200	1
10	10	1	100	95	0
11	8	8	70	42	0
12	6	15	40	20	0

Données normalisées :

	R	C	V	D	A
1	0.93	0.00	0.86	1.00	1.00
2	0,67	0,20	0,79	0.45	0,00
3	0.53	0.53	0.57	0.14	0.00
4	0,47	0,80	0,43	0,14	0,00
5	0,93	0.07	0.79	0.84	1.00
6	0,67	0,13	0,71	0.60	0,00
7	0,53	0.67	0,64	0.15	0.00
8	0,47	1,00	0,36	0,14	0.00
9	1.00	0.00	1.00	0.80	1.00
10	0,67	0,07	0,71	0,38	0,00
11	0.53	0,53	0,50	0.17	0.00
12	0,40	1,00	0,29	0,08	0,00

4.2. Résultats

seuil : p = 1

A - Recherche du niveau 1

à partir des données normalisées du tableau I x J :

Etape 1

Calcul des distances d(j,j') pour chaque paire de facteurs de J ; si card J = m, C_m^2 calculs doivent être effectués :

d(R,C) = 6,00 d(C,V) = 6,21 d(V,D) = 3,15 d(D,A) = 0,08
d(R,V) = 0,88 d(C,D) = 7,31 d(V,A) = 5,36
d(R,D) = 3,05 d(C,A) = 7,87
d(R,A) = 5,07

Etape 2 :

Calcul des distances D(j) pour chaque facteur de J :

D(R) = 6,00 + 0,88 + 3,05 + 5,07 = 15,00

D(C) = 27,39 D(D) = 13,59 D(A) = 18,38

Facteur le plus proche : D

Etape 3 :

Segmentation de I selon D :

- sous-ensemble des observations les mieux caractérisées par D :
  segment n° 1 = {1}
- sous-ensemble des observations les moins bien caractérisés par D :
  segment n° 2 = {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}

B - Recherche du niveau 2

B.1 - Le segment n° 1 ne comportant qu'une seule observation l'algorithme se poursuit pour le segment complémentaire.

B.2 - Segmentation des observations du segment n° 2

Le facteur D est exclu de l'ensemble J on considère la sous-matrice :

	R	C	V	A
2	0.67	0.50	0.79	0.00
3	0.53	0.93	0.57	0.00
4	0.47	0.80	0.43	0.00
5	0.93	0.07	0.79	1.00
6	0.67	0.13	0.71	0.00
7	0.53	0.67	0.64	0.00
8	0.47	1.00	0.36	0.00
9	1.00	1.00	1.00	1.00
10	0.67	0.07	0.71	1.00
11	0.53	0.13	0.50	1.00
12	0.40	0.00	0.29	1.00

Etape 1

d(R,C) = 5,07 d(C,V) = 5,36 d(V,A) = 5,21
d(R,V) = 0,80 d(C,A) = 6,87
d(R,A) = 5,00

Etape 2

D(R) = 10,87 D(C) = 17,30 D(V) = 11,37 D(A) = 17,08

facteur le plus proche : R

Etape 3

- sous-ensemble des observations les mieux caractérisées par R :
  segment n° 3 = {9}
- sous-ensemble complémentaire :
  segment n° 4 = {2, 3, 4, 5, 6, 7, 8, 10, 11, 12}

C- Recherche du niveau 3

C.1. Fin de l'algorithme pour le segment n° 3

C.2. Segmentation des observations du segment n° 4

Le facteur R est exclu de J - {D}

On considère la sous-matrice :

	C	V	A
2	0.20	0.79	0.00
3	0.53	0.57	0.00
4	0.80	0.43	0.00
5	0.07	0.79	1.00
6	0.13	0.71	0.00
7	0.67	0.64	0.00
8	1.00	0.36	0.00
10	0.07	0.71	0.00
11	0.53	0.50	0.00
12	1.00	0.29	0.00

Etape 1

d(C,V) = 4,36 d(V,A) = 5,21

d(C,A) = 5,87

Etape 2

D(C) = 10,23 D(V) = 10,57 D(A) = 11,08

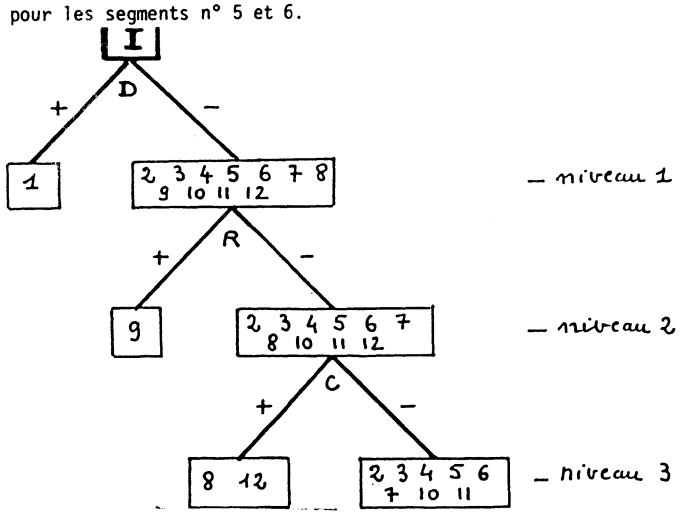
facteur le plus proche : C

Etape 3

- sous-ensemble des observations les mieux caractérisées par C :
  segment n° 5 = {8, 12}
- sous-ensemble complémentaire :
  segment n° 6 = {2, 3, 4, 5, 6, 7, 10, 11}

Fin de l'algorithme : après exclusion de C de J - {D,R}, il ne reste plus que deux facteurs à comparer pour les segments n° 5 et 6.

seuil : 1



seuil : p = 0,75

A - Recherche du niveau 1

S'appliquant aux données normalisées du tableau I x J, les étapes 1 et 2 conduisent au résultat obtenu pour p = 1 (facteur le plus proche : D).

L'étape 3 est modifiée en raison de la nouvelle valeur du seuil :

- sous-ensemble des observations les mieux caractérisées par D :
  segment n° 1 = {1,5,9}
- sous-ensemble complémentaire :
  segment n° 2 = {2, 3, 4, 6, 7, 8, 10, 11, 12}

B - Recherche du niveau 2

B.1. Segmentation des observations du segment n° 1.

Le facteur D est exclu de l'ensemble J

On considère la sous-matrice :

	R	C	V	A
1	0.93	0.00	0.86	1.00
5	0.93	0.07	0.79	1.00
9	1.00	0.00	1.00	1.00

Etape 1 :

d(R,C) = 2,80 d(C,V) = 2,58 d(V,A) = 0,36
d(R,V) = 0,22 d(C,A) = 2,93
d(R,A) = 0,13

Etape 2 :

D(R) = 3,15 D(C) = 8,31 D(V) = 3,16 D(A) = 3,42
Facteur le plus proche : R

Etape 3 :

- sous-ensemble des observations les mieux caractérisées par R :
  segment n° 3 = $\{1, 5, 9\}$
- sous-ensemble complémentaire :
  segment n° 4 = $\{\emptyset\}$

B.2. Segmentation des observations du segment n° 2.
Le facteur D est exclu de l'ensemble J.

On considère la sous-matrice :

	R	C	V	A
2	0.67	0.20	0.79	0.00
3	0.53	0.53	0.57	0.00
4	0.47	0.80	0.43	0.00
5	0.67	0.13	0.71	0.00
7	0.53	0.67	0.64	0.00
8	0.47	1.00	0.36	0.00
10	0.67	0.07	0.71	0.00
11	0.53	0.53	0.50	0.00
12	0.40	1.00	0.29	0.00

Etape 1 :

d(R,C) = 3,20 d(C,V) = 3,64 d(V,A) = 5,00
d(R,V) = 0,66 d(C,A) = 4,93
d(R,A) = 4,93

Etape 2 :

D(R) = 8,79 D(C) = 11,77 D(V) = 9,30 D(A) = 14,86

Facteur le plus proche : R

Etape 3 :

- sous ensemble des observations les mieux caractérisées par R :
  segment n° 5 = $\{\emptyset\}$
- sous-ensemble complémentaire :
  segment n° 6 = $\{2, 3, 4, 6, 7, 8, 10, 11, 12\}$

C. Recherche du niveau 3

C.1. Segmentation des observations du segment n° 3
Le facteur R est exclu de l'ensemble J - $\{D\}$

On considère la sous-matrice :

	C	V	A
1	0.00	0.86	1.00
5	0.07	0.79	1.00
9	0.00	1.00	1.00

Etape 1 :

d(C,V) = 2,58 d(V,A) = 0,36
d(C,A) = 2,93

Etape 2 :

D(C) = 5,51 D(V) = 2,94 D(A) = 3,29
Facteur le plus proche : V

Etape 3 :

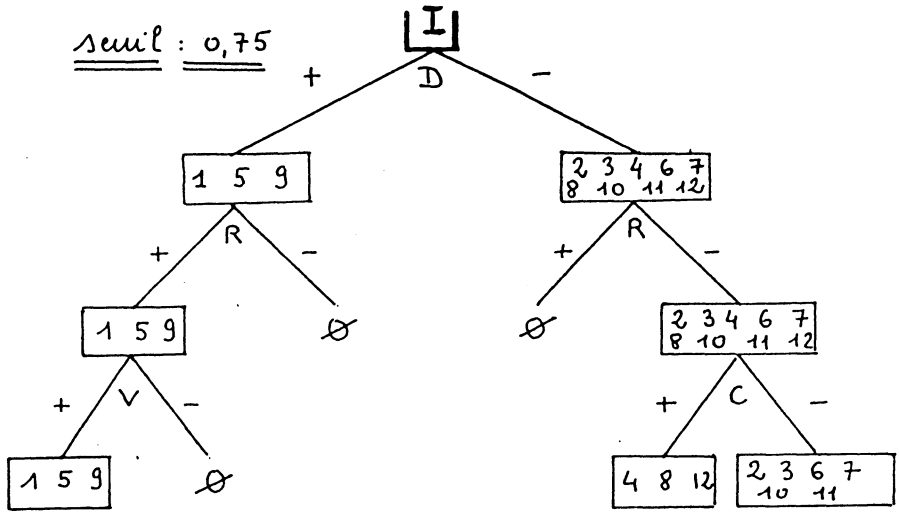
- sous-ensemble des observations les mieux caractérisées par V :
  segment n° 7 = {1, 5, 9}
- sous-ensemble complémentaire :
  segment n° 8 = {∅}

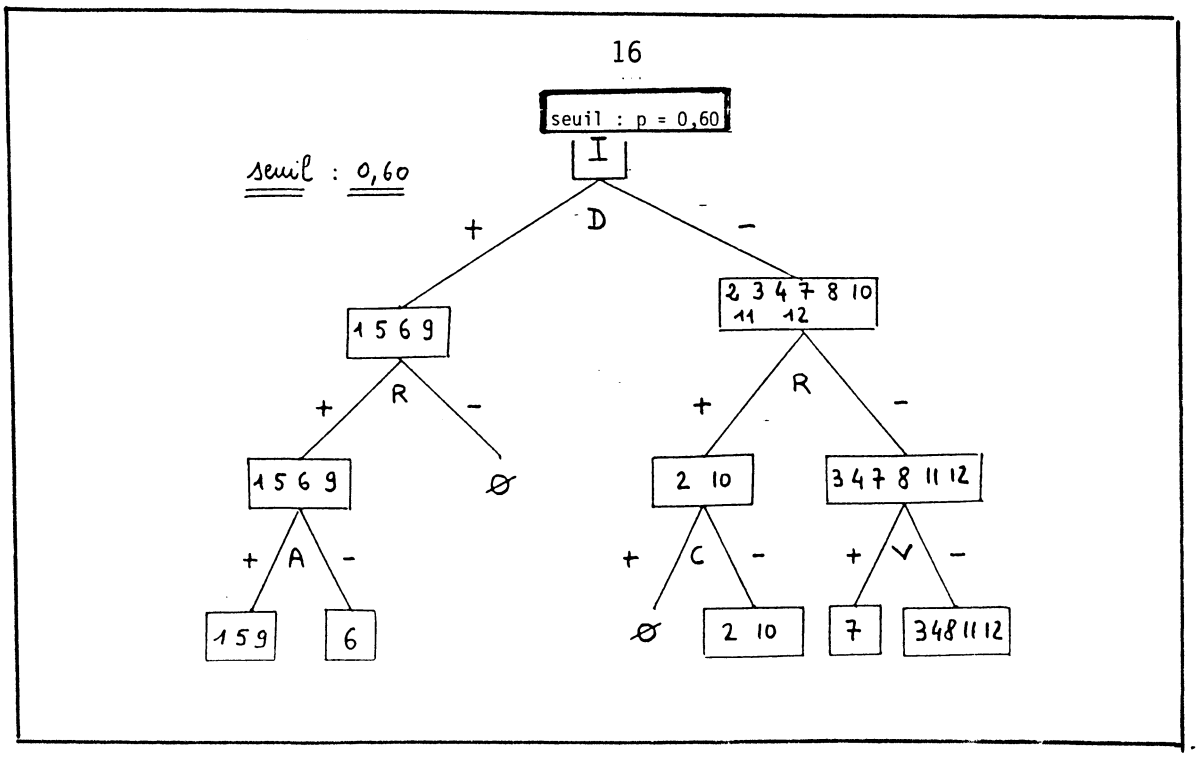
C.2. Segmentation des observations du segment n° 6

	C	V	A
2	0.20	0.79	0.00
3	0.53	0.57	0.00
4	0.80	0.43	0.00
6	0.13	0.71	0.00
7	0.67	0.64	0.00
8	1.00	0.36	0.00
10	0.07	0.71	0.00
11	0.53	0.50	0.00
12	1.00	0.29	0.00

Facteur le plus proche : C
- sous-ensemble des observations les mieux caractérisées par C :
  segment n° 9 = {4, 8, 12}
- sous-ensemble complémentaire :
  segment n° 10 = {2, 3, 6, 7}

Fin de l'algorithme (deux facteurs à comparer)





4.3. Commentaires

Le processus de séparation des éléments de I ou de l'un de ses sous-ensembles, peut-être rapproché de la problématique des choix multicritères. La vérification du surclassement éventuel d'un élément par un autre dépend, en particulier, de l'importance numérique de la majorité des critères qui se prononcent en faveur de cette hypothèse ; lorsque l'unanimité des critères est exigée, l'hypothèse est rarement vérifiée : les éléments sont jugés incomparables. Pour enrichir la relation de surclassement, il convient d'abaisser le seuil de majorité (ou "seuil de concordance").

Le seuil de sévérité est de la même nature que le seuil de majorité. Pour p = 1, les observations les "mieux caractérisées" par le facteur sélectionné sont peu nombreuses ; l'algorithme isole les observations les "plus typées" à chaque niveau de la segmentation :

Région de plus forte DENSITE (niveau 1) →

Région de plus haut REVENU parmi les régions les moins denses (niveau 2) →

1	2	3	4
5	6	7	8
9	10	11	12

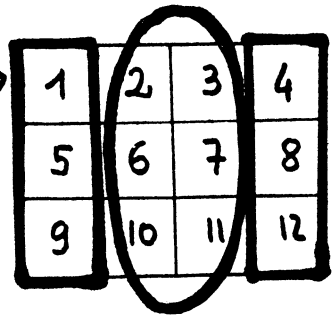
← Régions de plus fort SOUS-EMPLOI parmi les régions les moins denses et les moins riches (niveau 3)

Par définition, le sous-ensemble des observations les "mieux caractérisées" s'élargit lorsque le seuil de sévérité est atténué. De nouvelles dichotomies peuvent donc avoir lieu. Au niveau donné de la segmentation (supérieur à 1)

on constate que la partition sur I devient généralement plus fine au fur et à mesure que p diminue. Les observations d'une même classe présentent alors de nombreux point communs ; l'homogénéité des regroupements obtenus se renforce et la description de l'ensemble initial s'enrichit.

Au niveau 3 de la segmentation réalisée pour p = 0,75, les résultats suivants sont mis en évidence :

Zone Favorisée
(Régions de forte DENSITE ayant un REVENU et une VALEUR AJOUTEE par tête relativement élevés)



Zone Défavorisée
(se distingue de la zone intermédiaire par un un taux de CHÔMAGE relativement élevé).

Zone intermédiaire
(aucun facteur révélé par la segmentation ne caractérise les régions de ce regroupement).

Un découpage en trois classes relativement équilibrées paraît devoir convenir pour des cas aussi simples que l'exemple proposé.

Pour des tableaux faisant intervenir des observations et des descripteurs en nombre plus important, on pourra descendre en-dessous d'un seuil de 0,75. Mais on évitera le seuil 0,50 pour lequel on ne peut plus parler de "sévérité".

La valeur minimale du seuil à ne pas dépasser peut-être justifiée par la notion de sous-ensemble (ordinaire) le plus voisin d'un s.e.f. (KAUFMANN 1973, FUSTIER 1979 p. 597).

En effet, désignons par s le seuil minimum de sévérité : s est une valeur très proche de 0,50 telle que $s > 0{,}50$. Au seuil s, le sous-ensemble des observations de S les "mieux caractérisées" par j^*, et noté S'_s, s'écrit d'après(9):

$$(14)\quad S'_s = \left\{ i \in S / \mu_{ij^*} \geq s \right\}$$

ou d'une façon purement symbolique :

$$(15)\quad S'_s = \left\{ (i, f_{ij^*}) ; \forall i \in S ;\ f_{ij^*} \in \{0,1\} \right\}$$

avec :

$$(16)\begin{cases} f_{ij^*} = 1 & \text{si } \mu_{ij^*} \geqslant s \\ f_{ij^*} = 0 & \text{si } \mu_{ij^*} < s \end{cases}$$

Par définition, S'_s est le sous-ensemble de S le plus voisin du s.e.f. $\underset{\sim}{A}_{j^*}$ défini sur S et correspondant à j^*. D'après (16), on vérifie que S'_s n'est pas flou.

Remarque :

$\underset{\sim}{A}_{j^*}$ est le s.e.f. le "plus proche" de tous les autres s.e.f. définis sur S(voir section 3).

S'_s est le sous-ensemble (ordinaire) de S le "plus voisin" du s.e.f. $\underset{\sim}{A}_{j^*}$.

On veillera à distinguer les deux notions. La notion de "plus proche" s'applique entre des s.e.f. La notion de "plus voisin" concerne un sous-ensemble (au sens de la théorie traditionnelle ensembliste) et un s.e.f.

Bibliographie

KAUFMANN A - "Introduction à la théorie des sous-ensembles flous", Masson et Cie, tome 1, Paris, 1973.

PONSARD C., TRANQUI P. - "Fuzzy Economic Regions in Europe", Environment and Planning A, volume 17, 1985, pp. 873-887.

PRADE H., - "Modèles Mathématiques de l'imprécis et de l'incertain en vue d'applications au raisonnement naturel", Thèse de doctorat d'Etat ès sciences. Université Paul Sabatier, Toulouse, juin 1982.

ROLLAND-May C. - "Fuzzy Geographical Space : Algorithm of Fuzzy Classification and Application to Fuzzy Regionalization". Paper presented at the 7th European Congress on Operational Research. Bologne, juin 1985.

WILLIAMS W.T., LAMBERT J.M. - "Multivariate Methods in Plant Ecology". Journal of Ecology, n° 47, 1959, pp. 83-101.

FUSTIER B. - "Les interactionsspatiales en Economie", Collection de l'I.M.E., n°21, Editions Sirey, Paris, 1979.