PANEL SOCIO-ECONOMIQUE

"LIEWEN ZU LETZEBUERG"

Document PSELL N° 18

MNDr, partition valuée selon la méthode de ROUBENS et LIBERT

B. Gailly

Document produit par le

CENTRE D'ETUDES DE POPULATIONS, DE PAUVRETE
ET DE POLITIQUES SOCIO-ECONOMIQUES

C.E.P.S./INSTEAD
a.s.b.l.

B.P. 65 L-7201 Walferdange
Tél. (352) 33 25 15

Président: Gaston Schaber

---

1 9 8 9

MNDr, partition valuée selon la méthode de ROUBENS et LIBERT

Cette présentation du programme MNDr, mis au point par ROUBENS et LIBERT, est largement documentée par des publications des auteurs. Elle ne prétend à aucune originalité. D'autres rapports seront présentés ultérieurement. Ils relateront différentes études effectuées à l'aide du MNDr.

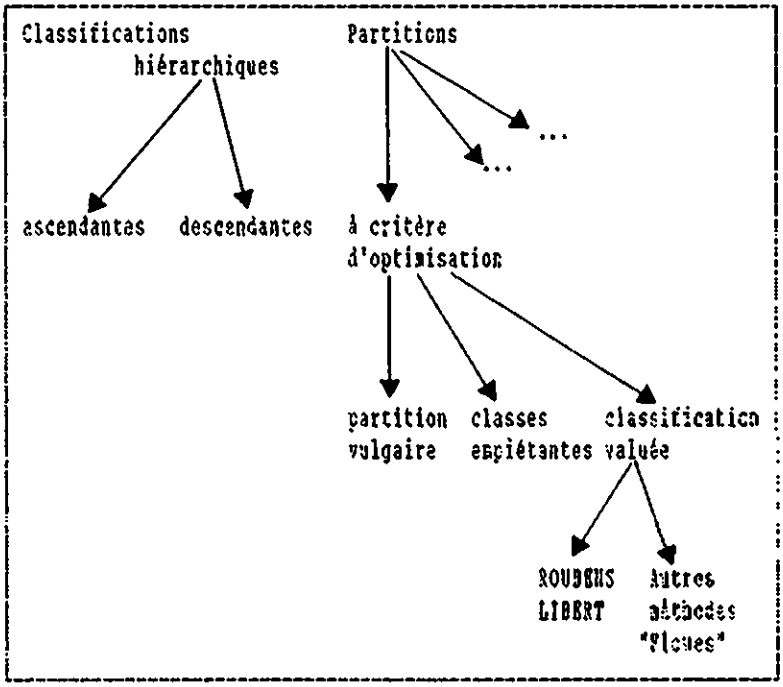
Il faut remercier ici G. LIBERT. Ses qualités pédagogiques, sa clarté, sa précision et sa patience ont facilité notre parcours.

Il faut également remercier Gunther Schmaus qui a implanté le logiciel du MNDr au CEPS-INSTEAD. Il en a rendu l'usage simple et rapide.

S O M M A I R E

I. PARTITION VALUEE OU PARTITION "floue"

La méthode de classification automatique proposée par LIBERT et ROUBENS apporte un certain nombre d'améliorations aux méthodes de classification "valuée". Ces méthodes appartiennent à l'ensemble des méthodes dites "à critère d'optimisation" qui appartiennent elles-mêmes à l'ensemble des méthodes qui construisent des "partitions" sans procéder par voie hiérarchique.



Cet arbre permet de situer la méthode de ROUBENS et de faire état de ses avantages comparatifs par rapport à d'autres méthodes.

Comme toutes les méthodes de classification automatique, la méthode de ROUBENS cherche à distinguer, dans un ensemble d'individus, des sous-groupes d'individus très semblables.

Cette opération de structuration des données appartient à l'ensemble des procédures d'analyse multivariée. Un grand nombre de méthodes ont été développées à cet usage. Il est donc important de préciser les avantages comparatifs de la méthode de ROUBENS.

Partons d'une matrice classique définie par un ensemble d'individus (I) caractérisés par un ensemble d'observations (O).

Soit une matrice M = I x O

L'algorithme de ROUBENS (MNDr) cherche à partitionner cet ensemble d'individus en prenant en compte l'ensemble de leurs caractéristiques. Cette partition s'obtient à partir du critère suivant:

$$\text{MIN} \sum_{k} \sum_{i} \sum_{i'} \mu_{ik}^{r}(i)\, \mu_{i'k}^{r}(i')\, d(i, i')$$

Il s'agit de trouver la valeur minimale d'une somme de classes (k), définies par la somme des valeurs des distances entre chaque individu (i) et tous les autres individus (i'), compte tenu du coefficient d'appartenance de chaque individu i à chaque classe ($\mu_{ik}^{r}(i)$) et du coefficient d'appartenance de chaque autre individu i' à chaque classe ($\mu_{ik}^{r}(i')$).

Ce critère repose sur un certain nombre d'éléments qu'il faut préciser.

1° - le calcul des distances entre les individus ($d(i,i')$)
2° - des fonctions d'appartenance des individus à des classes ($\mu_{ik}^{r}(i)$)
3° - un nombre de classes donné
4° - un algorithme pour estimer les fonctions d'appartenance des individus aux classes.

1. Les distances entre les individus

La méthode de ROUBENS permet de traiter des données qui ne sont pas nécessairement mesurables. Elle n'impose pas les contraintes d'une méthode métrique. Elle ne calcule pas les distances des individus par rapport au centre de gravité des classes.

Elle prend en compte la distance entre chaque individu i et tous les individus i' formant la classe k.

1.1. Les méthodes métriques

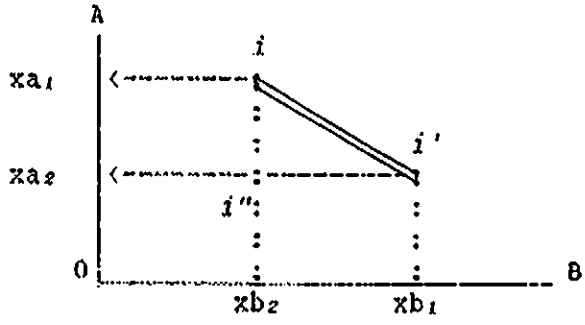
Toutes les méthodes métriques mesurent les distances entre les individus dans l'espace euclidien. De telles méthodes sont utilisées dans les classifications hiérarchiques ascendantes ou descendantes et dans les partitions vulgaires.

Les méthodes métriques mesurent la distance entre l'individu i et l'individu i' par :

$$d(i, i') = \sqrt{d^2(i, i'') + d^2(i', i'')}$$

où i, i' et i'' sont trois individus situés dans l'espace euclidien.

Soit, graphiquement :



La distance entre i et i' peut être mesurée par:

$$d(i, i') = \sqrt{(xa_1 - xa_2)^2 + (xb_1 - xb_2)^2}$$

Quatre critères permettent de décider si une mesure est réellement métrique (ALDENDERFER M.S., 1984):

- la symétrie: $d(i,i') = d(i',i) \geq 0$
- l'inégalité triangulaire: $d(i,i') \leq d(i,i'') + d(i',i'')$
- la possibilité de distinguer les non-identiques:
  si $d(i,i') \neq 0$, alors $i \neq i'$
- l'impossibilité de distinguer les identiques:
  si i et i' étaient identiques, alors $d(i,i') = 0$.

On peut classer parmi les méthodes métriques, toute méthode qui définit:

- un barycentre ou centre de gravité de chaque classe k,

- la variabilité ou la variance de k,

et qui agrège les individus en fonction de l'un de ces critères.

Ces méthodes métriques comportent certains inconvénients: en définissant la distance des individus par rapport au centre de gravité des classes, elles créent des agrégats sphériques. Les données ne se prêtent pas nécessairement à ce mode de structuration.

1.2. Les méthodes non-métriques

Les méthodes non-métriques exigent uniquement la connaissance d'un ordre entre les distances[1]. Elles ne recourent pas à la notion de centre de gravité des classes.

Les distances peuvent être ordonnées à partir d'un noyau d'individus ou à partir de l'ensemble des individus qui appartiennent à une même classe.

Les classifications hiérarchiques ascendantes et descendantes fournissent des exemples:

- La méthode hiérarchique ascendante procède par agrégations successives des individus les plus semblables ou des groupes les plus semblables. Elle peut rechercher l'individu le plus proche de l'un des membres d'une classe ("single linkage") avec les inconvénients que cette méthode comporte. Elle peut, au contraire, inclure dans une classe l'individu qui est simultanément le plus proche de tous les individus appartenant à cette classe ("complete linkage").

- La classification hiérarchique descendante procède par dichotomies successives des classes existantes. La méthode de HUBERT éclate chaque fois la classe qui

---

1. Les individus ne sont pas positionnés dans un espace ayant les propriétés d'un espace métrique. Le principe de l'inégalité triangulaire ne peut donc pas être vérifié.

contient le plus grand diamètre. Le diamètre est la plus grande distance qui sépare deux individus appartenant à une classe (HUBERT, 1972).

1.3. La méthode de ROUBENS

La méthode de ROUBENS mesure la distance entre chaque individu et la classe k formée par la contribution de tous les individus.

Cette mesure des distances ne cherche pas à minimiser les distances des individus par rapport à un centre de gravité (réel ou fictif). Elle évite donc la formation d'agrégats sphériques. Des agrégats non sphériques correspondront généralement mieux à la structure des données.

Ceci demande une précision supplémentaire à propos de la notion d'appartenance ou de "contribution" des individus à une classe.

2. Les fonctions d'appartenance

2.1. Des partitions vulgaires aux partitions floues

D'une manière générale, la distance entre un individu i et une classe k peut s'écrire :

$$(1) \qquad D(i,k) = \sum_{i' \varepsilon k} d(i,i')$$

Soit, i' est un individu quelconque, différent de i et appartenant à la classe k. Ce peut être, par exemple, l'individu le plus proche de i.

Soit, i' est un ensemble d'individus appartenant à la classe k.

On peut donc écrire, de manière équivalente:

(2) $$D(i,k) = \sum_{i'} \mu_{ik}(i')\, d(i,i')$$

où $\mu_{ik}(i')$ désigne une <u>fonction d'appartenance</u> des individus i' à la classe k.

<u>Les partitions vulgaires</u> posent que:

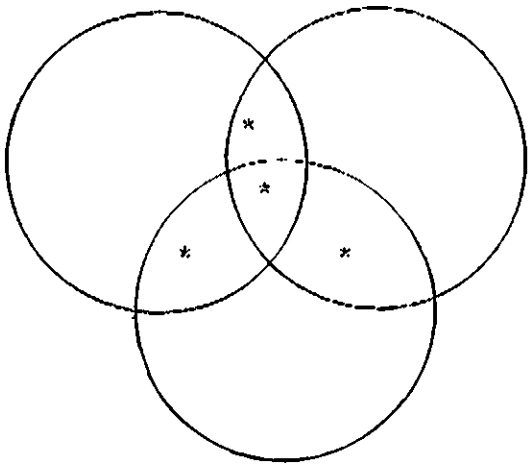
$\mu_{ik}(i')$ vaut 0 <u>ou</u> 1.

Si $\mu_{ik}(i')$ vaut *1*, i' appartient à la classe k.

Si $\mu_{ik}(i')$ vaut *0*, i' n'appartient pas à la classe k.

Dans ce cas, seules les distances entre i et les i' <u>appartenant à k</u> sont prises en compte pour former les classes.

Mais les partitions vulgaires, comme les classifications hiérarchiques, ne sont pas aptes à classer correctement des individus appartenant à deux ou plusieurs classes.



Il existe quelques méthodes permettant de déterminer des "classes empiétantes" mais les résultats sont difficilement interprétables.

Les <u>partitions floues</u> adoptent une autre position et offrent d'autres solutions à ce type de problème.

Elles posent que:

$$\mu_{ik}(i') \in [0,1]$$

et $$\Sigma \mu_{ik}(i') = 1$$

$\mu_{ik}(i')$ varie entre 0 et 1 et n'est plus limité aux seules valeurs 0 et 1.

Chaque individu est affecté d'un poids total égal à 1 et répartit son poids entre les classes.

Les agrégats sont définis en fonction des distances entre chaque individu i et tous les autres individus i', pour peu que $\mu_{ik}(i') > 0$.

Un problème se pose: comment passer d'une fonction d'appartenance dichotomique propre aux partitions vulgaires à cette fonction d'appartenance propre aux partitions floues?

Pour sortir des fonctions d'appartenance dichotomiques, il suffit d'affecter un exposant r à $\mu_{ik}(i')$, soit $\mu_{ik}^{r}(i')$, tel que r soit supérieur à 1.

Dès lors la distance entre un individu i et une classe k devient:

$$(3) \qquad D(i,k) = \sum_{i'} \mu_{ik}^{r}(i')\, d(i,i')$$

Plus r est petit et s'approche de 1, plus la solution s'approche d'une partition vulgaire.

Plus r s'approche de 2 ou dépasse 2, plus les individus répartissent leur poids de manière égale sur toutes les classes. L'ensemble devient de plus en plus flou.

La valeur de r peut être située entre 1 et 2.

En pratique, on fixera r à 1.4. Mais cette valeur peut être modifiée, afin de tester des solutions différentes ou parce que les données l'exigent:

- r sera diminué, si l'on accepte de prendre le risque de mal classer certains individus afin d'obtenir une classification plus nette;

- r sera augmenté, si l'on préfère respecter les appartenances multiples des individus au prix d'une classification plus floue.

2.2. La méthode de ROUBENS

La méthode de ROUBENS appartient à cette famille des partitions floues ou partitions valuées.

Le critère d'agrégation s'écrivait:

$$(4) \qquad MIN \sum_k \sum_i \sum_{i'} \mu_{ik}^r(i)\, \mu_{ik}^r(i')\, d(i,i')$$

avec : $0 \leq \mu_{ik}^r(i) \leq 1$

$$\sum_k \mu_{ik}^r(i) = 1$$

Or, on a noté (3):

$$D(i,k) = \sum_{i'} \mu_{ik}^r(i')\, d(i,i')$$

Cette quantité mesure la distance entre i et le groupe k formé par la contribution de tous les individus i'.

Le programme (4) devient donc:

$$(5) \qquad MIN \sum_k \sum_i \mu_{ik}^r(i)\, D(i,k)$$

La mesure des distances a été précisée. Les fonctions d'appartenance ont été définies. Il reste à fixer la valeur de k (nombre de classes) et à estimer les $\mu_{ik}^r(i)$.

3. Le nombre de classes

Contrairement aux méthodes de classification hiérarchiques, les méthodes de partition exigent que la valeur de k soit fixée a priori. Comme ce nombre est presque toujours inconnu, on réalise donc plusieurs classifications, en variant la valeur de k et on choisit la "meilleure" solution.

Ceci appelle deux commentaires:

3.1. La "meilleure solution"

Le choix de la "meilleure" solution est une opération subjective. LIBERT propose une méthode de repérage de la solution optimale. Elle sera développée plus loin. Il montre néanmoins que la solution optimale dépend de la valeur de l'exposant r; en faisant varier r, on modifie la sensibilité de l'analyse et le nombre optimal de classes varie.

Le choix de la valeur de r explicite le rôle de la subjectivité du chercheur dans le choix du nombre optimal de classes.

Ce rôle de la subjectivité du chercheur reste implicite dans les partitions vulgaires, puisque $\mu_k(i)$ est simplement affecté d'un exposant unitaire invariable.

En explicitant la valeur de r, le chercheur déclare *a priori* le risque qu'il accepte de prendre par rapport aux individus intermédiaires.

3.2. La partition "initiale"

Les partitions s'effectuent à partir d'une partition initiale. La valeur de k peut correspondre à un niveau d'agrégation donné par une classification hiérarchique préalable.

La partition peut également correspondre à une répartition aléatoire des individus dans un nombre de classes déterminé.

On peut montrer que la solution finale est souvent influencée par la partition initiale adoptée.

Par contre, la pratique montre que ces conditions initiales influencent peu la solution dans le cadre de la méthode de ROUBENS,

sauf si: $\mu_k(i) = \frac{1}{q} \quad \forall k$ et $\forall i$.

(Le poids de chaque individu dans chaque classe est inversement proportionnel au nombre de classes).

4. Estimations des fonctions d'appartenance

Les distances étant calculées, le r fixé et le nombre de classes déterminé, le programme doit estimer la fonction d'appartenance de chaque individu.

Il établit des fonctions d'appartenance initiales de chaque individu à chaque classe.

Il calcule, pour le premier individu, la somme de ses distances par rapport à tous les autres individus appartenant à toutes les classes, en prenant en compte leurs coefficients d'appartenance à chacune des classes.

Il obtient ainsi de nouveaux coefficients d'appartenance pour le premier individu.

Il passe au deuxième, troisième, ... individu.

Il reprend au premier si le critère de minimisation des distances n'a pas encore atteint un minimum.

La procédure converge vers un optimum local.

Les itérations s'interrompent lorsque la solution ne s'améliore plus. (Il appartient à l'usager de déterminer le degré de précision avec lequel cette croissance relative doit être calculée).

II. DETERMINER LE NOMBRE OPTIMAL DE CLASSES

Quelle que soit la méthode de classification choisie, le choix du nombre de classes reste un problème extrêmement important.

C'est un problème presque insoluble: établir le nombre exact de classes supposerait que l'on connaisse la structure initiale des données. Ce n'est généralement pas le cas.

Il reste donc à fixer ce nombre de classes de telle manière qu'il reflète le mieux possible une réalité inconnue.

De nombreuses procédures ont été élaborées en vue de déterminer le nombre "optimum" de classes selon la méthode d'agrégation adoptée.

1. Un indice de validité

LIBERT et ROUBENS ont proposé un indice de validité. Après l'avoir présenté d'une manière très formelle, LIBERT en propose une discussion à travers quelques exemples (LIBERT, 1986).

Il le définit comme suit:

$$V = (KF - 1)/(K - 1)$$

où V = indice de validité
K = nombre de classes
$F = ((\sum_x \mu_x / n) + \min_x \mu_x) / 2.$

où $\mu_x = \max_k \mu_k(x)$

Si μ_x est la valeur du plus grand coefficient d'appartenance de l'individu x à l'une des classes,

alors $\sum_x \mu_x$ est la somme des valeurs maximales des

coefficients de tous les individus,

et $(\sum_x \mu_x / n)$ est la moyenne des valeurs maximales des n individus.

Si $\min_x \mu_{x_2}$ est la plus petite de toutes ces valeurs individuelles maximales

alors F est égal à la valeur centrale située entre

- la valeur moyenne des coefficients d'appartenance les plus élevés de tous les individus,
- et le plus petit de ces coefficients.

Pour un nombre de classes donné, K, l'indice de validité, V, sera d'autant plus élevé que la valeur de F augmente.

En rapportant la valeur de F au nombre de classes $(K - 1)$, les valeurs de F peuvent être comparées lorsqu'on fait varier le nombre de classes.

Le nombre optimum de classes correspond à la plus haute valeur de V et donc à la plus haute valeur de F.

2. La valeur minimale de V

$V = 0$, lorsque $\mu_{i_0}(x) = 1/k$, c'est-à-dire lorsque tous les individus affectent à toutes les classes, un poids inversement proportionnel au nombre de classes.

Ex.: - pour 2 classes: 0.50 0.50

- pour 4 classes: 0.25 0.25 0.25 0.25

En d'autres termes, $V = 0$ lorsqu'on se trouve en présence d'une partition totalement floue, telle qu'il n'est plus possible de différencier les degrés d'appartenance des individus aux classes.

Application:

$\cdot \mu_x \cdot$ valeur du plus grand coefficient d'appartenance de x à l'une des classes est identique pour tout

individu, soit $\frac{1}{k}$

. $\sum_x \mu_x/n)$, moyenne de ces valeurs pour l'ensemble des individus, égale

. $MIN_x \mu_x$, la plus petite de ces valeurs, égale

. F, somme de la valeur moyenne et de la plus petite valeur, divisée par 2, égale

. Si $F = \mu_x$, alors $KF = K.\mu_x = K.\frac{1}{K} = 1$

. $V = 0/K - 1 = 0$.

3. La valeur maximale de V

Inversement, $V = 1$ lorsqu'on se trouve en présence d'une partition vulgaire, telle que chaque individu appartient et n'appartient qu'à un seul groupe:

soit $\mu_{A_i}(x) = 0$ ou 1.

Application:

. μ_x, valeur du plus grand coefficient d'appartenance de x à l'une des classes est identique pour tous les individus, soit 1

. $\sum_x \mu_x/n$, moyenne de ces valeurs, égale 1

. $MIN_x \mu_x$, le plus petit coefficient parmi les coefficients les plus élevés, reste égal à μ_x, soit 1

. F, somme de la valeur moyenne et de la plus petite valeur divisée par 2, égale 1.

. Si $F = 1$, alors $KF = K$

. Si $KF = K$, alors $V = (K - 1)/(K - 1) = 1$.

4. Les valeurs de V

L'indice de validation oscille donc entre 0 et 1, selon que la partition est plus floue ou plus nette.

Le nombre optimum de classes correspond à la valeur la plus élevée de l'indice de validité (LIBERT et ROUBENS, 1982).

En-deça et au-delà, la partition s'avère relativement plus floue.

V dépend à la fois de K et de F.

V sera plus faible si le nombre de classes est tel que

- les coefficients d'appartenance maximum sont plus faibles
- le plus petit de ces coefficients est plus faible.

La solution optimum correspond donc au meilleur équilibre entre le nombre de classes et la répartition des coefficients d'appartenance entre les classes.

On perçoit mieux l'utilité de $MIN_x \mu_{xc}$: ce coefficient permet de prendre en compte les individus intermédiaires appartenant à plusieurs classes ou les plus mal classés.

Exemple:

1 - Si la moyenne des coefficients les plus élevés $(\Sigma_x \mu_{xc}/n)$ vaut .70 et le coefficient le plus faible $(MIN_x \mu_{xc})$ vaut .50, $F = .60$

2 - Si la moyenne des coefficients les plus élevés vaut .80, elle traduit une classification plus nette. Mais si le coefficient le plus faible ne vaut que .40, F égale toujours .60 et la validité de la classification n'est pas améliorée.

F a donc la propriété d'être sensible aux individus qui appartiennent à plusieurs classes et répartissent leur poids dans différentes classes.

5. La valeur de l'exposant *r*

Les coefficients d'appartenance des individus aux classes dépendent directement de la valeur fixée pour "*r*" (Cf., I, 2.1.). La valeur de l'indice de validité sera donc liée à la valeur de cet exposant.

LIBERT et ROUBENS (1982) ont montré que, dans un ensemble donné

- plus l'exposant s'approche de 1, plus la solution est proche d'une partition vulgaire et plus le nombre optimal de classes s'élève;

- plus l'exposant s'approche de 2, plus la solution est "floue" et plus le nombre optimal de classes diminue.

- pour un nombre de classes donné, la valeur optimale de *V* diminue systématiquement, lorsque la valeur de *r* s'élève.

En résumé, plus la valeur de *r* est fixée haut, plus la classification devient floue, plus le nombre optimal de classes diminue et plus la valeur optimale du coefficient de validité est faible.

Le tableau suivant illustre cet ensemble de liaisons dans un ensemble de données proposé par LIBERT et ROUBENS.

Tableau 1 : Valeurs de V^1.

IRIS	K=2	3	4	5	6
r=1.2	0.548	0.630	0.639	0.650	0.648
1.4	0.488	0.572	0.590	0.549	0.547
1.6	0.424	0.523	0.510	0.431	0.448
1.8	0.394	0.469	0.402	0.315	0.303
2.0	0.365	0.398	0.282	0.214	0.193
2.2	0.329	0.262	0.188	0.145	0.094

En outre, le tableau 2 présente les valeurs de V, pour le même ensemble de données, en introduisant une nouvelle modification.

Le tableau 1 présente les résultats lorsqu'on utilise le carré de la distance euclidienne pour des variables non standardisées.

Le tableau 2 présente les résultats lorsqu'on utilise le carré de la distance euclidienne pour des variables standardisées.

Tableau 2 : Valeurs de V pour des variables standardisées[1].

IRIS	K=2	3	4	5	6
r=1.2	0.464	0.612	0.653	0.634	0.573
1.4	0.455	0.544	0.519	0.538	0.489
1.6	0.414	0.438	0.373	0.409	0.343
1.8	0.335	0.327	0.225	0.211	0.205
2.0	0.270	0.200	0.121	0.121	0.093
2.2	0.212	0.146	0.103	0.094	0.069

1. Copie de LIBERT (1986)

Lorsque les variables sont standardisées, le nombre optimal de classes est inférieur pour toute valeur de r : les distances entre les individus sont plus petites et les différences semblent moins importantes. Tout se passe comme si l'observateur avait pris "plus de recul" par rapport à l'ensemble des points observés.

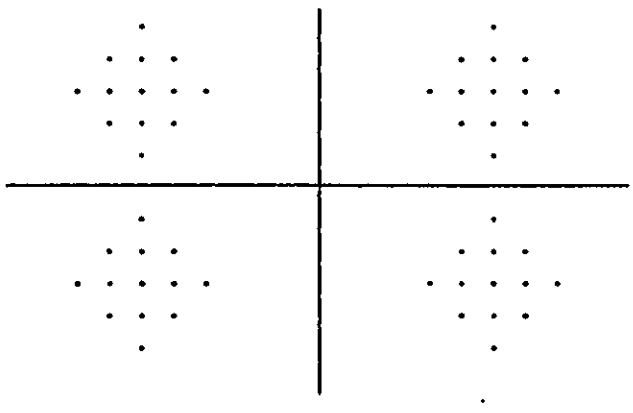
Ces commentaires relatifs à l'influence de la valeur fixée pour l'exposant r et à l'influence de la standardisation des variables attirent l'attention sur un point important:

Le choix du nombre de classes est une opération subjective.

En explicitant la valeur de l'exposant r et le choix de la mesure des distances, le chercheur rend compte de ce facteur "subjectif" ou des objectifs qu'il poursuit ou encore de la variété des solutions acceptables.

Le plus souvent, il existe plusieurs solutions acceptables, à moins qu'un ensemble d'individus s'organise en sous-ensembles très compacts et clairement distincts. LIBERT et ROUBENS ont montré que dans ce dernier cas, le nombre optimal de classes n'est pas sensible à la valeur de r, ni à la standardisation des variables (LIBERT et ROUBENS, 1982).

Les données du fichier RUSPINI illustrent ce cas particulier:



Les données forment quatre groupes compacts et distincts.

Les résultats présentés au tableau 3. montrent que le nombre optimal de classes est indépendant de r et de la standardisation des variables.

Néanmoins la valeur de V reste liée à la valeur de r et à la standardisation des variables:

- la valeur du coefficient de validité diminue lorsque r s'élève: la classification devient plus floue (tableau 3.1.),

- la valeur du coefficient de validité diminue lorsque les variables sont standardisées: l'observateur examine les données "de plus loin" (tableau 3.2.).

Tableau 3.1 : Valeurs de V pour variables non-standardisées et groupes compacts et distincts[1].

RUSPINI	K=2	3	4	5	6
r=1.2	0.897	0.528	0.990	0.783	0.761
1.4	0.705	0.437	0.914	0.671	0.649
1.6	0.556	0.350	0.811	0.646	0.588
1.8	0.436	0.292	0.710	0.527	0.473
2.0	0.342	0.216	0.612	0.469	0.386
2.2	0.265	0.161	0.521	0.399	0.323

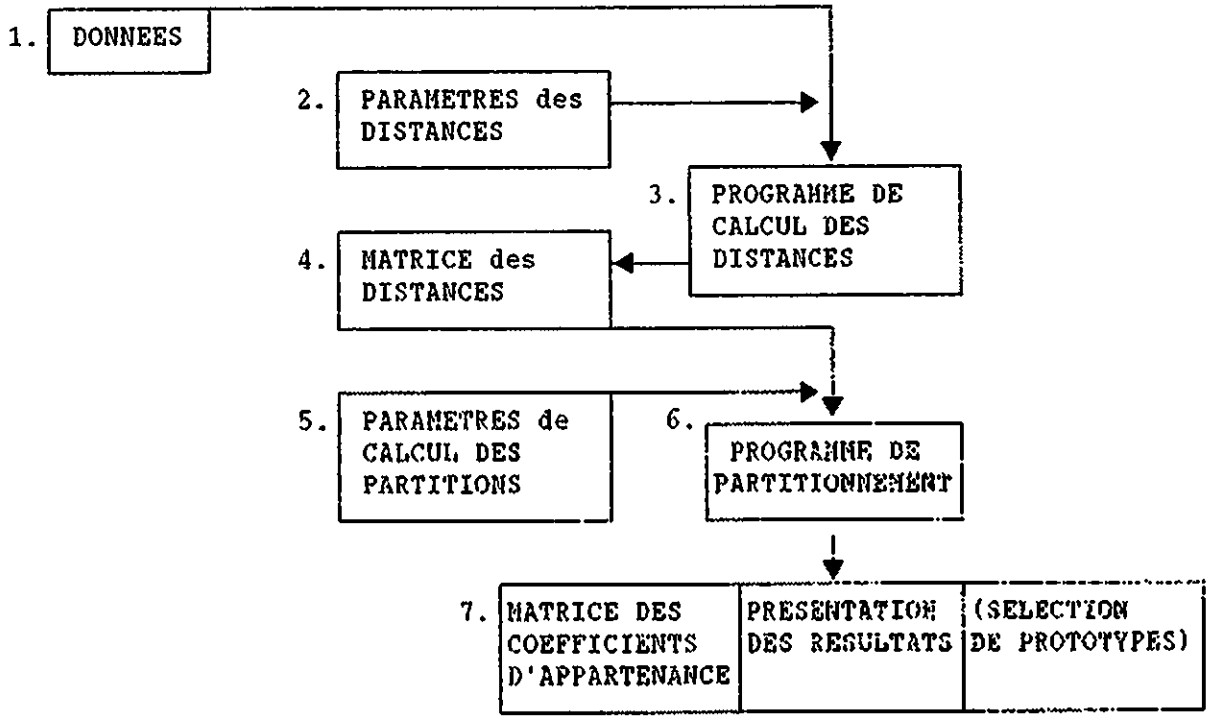
Tableau 3.2 : Valeurs de V pour variables standardisées et groupes compacts et distincts[1].

RUSPINI	K=2	3	4	5	6
r=1.2	0.731	0.581	0.986	0.907	0.787
1.4	0.517	0.523	0.872	0.750	0.692
1.6	0.218	0.469	0.747	0.610	0.562
1.8	0.299	0.373	0.640	0.529	0.495
2.0	0.131	0.127	0.544	0.459	0.407
2.2	0.011	0.012	0.453	0.363	0.303

1. Copie de LIBERT (1986)

III. DESCRIPTION DU LOGICIEL

Le logiciel du MNDr met en relations 7 éléments:



1. Les données doivent être des données brutes, présentées sous forme d'une matrice rectangulaire I X O.

   Les données doivent être précédées du numéro d'identification des individus. Ce numéro ne peut dépasser 3 chiffres.

   Dans son état actuel, le logiciel peut traiter

   - 300 unités d'analyse, au maximum
   - 200 variables, au maximum (200 variables à modalités multiples ou dichotomiques).

2. Le calcul des distances entre les individus est commandé par un certain nombre de paramètres modifiables par l'usager:

1. préciser le nombre d'individus (I)
2. préciser le nombre de variables (O)
3. définir la nature de la mesure des distances (D)
4. définir le format de lecture des données originales.

Exemple:

- ouvrir le fichier DIST PAR

colonnes 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

1e ligne

I	O	D

2e ligne (A 3, 11 F 2.0)

Commentaires:

1) Dans l'exemple ci-dessus O = 11, car le numéro d'identification des individus n'est pas pris en compte dans le calcul du nombre de variables.

2) La mesure des distances (D) sera égale à

0 : s'il s'agit de calculer la distance euclidienne sur des variables réduites ($m = 0$, $\sigma = 1$)

1 : pour calculer la distance euclidienne sur des variables non réduites

2 : pour l'indice de distance de JACCARD (cf. ALDENDERFER, 1984, p.29)

3 : pour mesurer les distances par le Chi^2.

3) Le format définit le numéro individuel (A3) et 11 variables correspondent à des nombres entiers à deux chiffres.

3. Le programme de calcul des distances produit la matrice des distances entre tous les individus en prenant en compte toutes les variables présentes dans le fichier des données initiales.

4. Cette matrice des distances devient, à son tour, l'input du programme de calcul des coefficients d'appartenance des individus aux classes.

5. Ce programme de classification est également commandé par la définition de paramètres modifiables par l'usager:

   1. définir le nombre d'individus (I)
   2. préciser le plus petit nombre de classes initiales, pour lequel la classification est demandée (K_1 ; toujours supérieur à 1)
   3. préciser le plus grand nombre de classes initiales, pour lequel la classification est demandée ($K_1 \leq K_2 \leq 10$)
   4. imprimer ou non la partition vulgaire la plus proche de la partition finale demandée:

      0 = pas d'impression
      1 = impression

   5. nombre maximum d'itérations demandées
   6. déterminer la valeur de l'exposant r ($r \geq 1.2$)
   7. fixer la valeur de la variation relative du critère de validité à partir de laquelle l'algorithme interrompt ses itérations (généralement 1.0 E - 0.6).

[Ce fichier de paramètres prévoit, en outre, un format général pour la lecture de la matrice des distances qui sera traitée par le programme de classification.]

Exemple:

- ouvrir le fichier MNDR PAR

colonnes 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

1e ligne : TITRE du problème traité (20 colonnes maximum)

2e ligne :

I	2	10	0	15

3e ligne :

1 . 4	1 . 0 E - 06

4e ligne : (Format de lecture des matrices de distances).

Commentaires :

Dans cet exemple, le programme

- traitera I individus
- procédera à 9 classifications différentes, prenant successivement comme partition initiale 2, 3, 4, ... 10 classes.
- n'imprimera pas la partition vulgaire correspondante.
- effectuera 15 itérations maximum, avant de s'arrêter
- s'interrompra avant la 15ème itération, si la valeur de la variation relative du critère de validité a atteint le niveau de précision demandé (1.0 E - 06)
- fixera l'exposant *r* à 1.4.

6. Les résultats de la classification sont présentés dans trois fichiers distincts accessibles à l'usager et susceptibles d'être imprimés.

1°/ Un premier fichier (appelons le MNDR RES.), présente les éléments suivants : (voir exemple ci-dessous).

- Rappel
  - du titre du problème traité
  - de la valeur de l'exposant *r*
  - du degré de précision fixé
  - du nombre maximum d'itérations demandé.

- Pour chaque partition (2, 3, 4, ... 10)

  - rappel
    - du nombre de classes demandées
    - du nombre d'unités d'analyse traitées

  - pour chaque itération:

    - valeur de la variation relative du critère de validité

    - indice de validité correspondant (valeur de *V*)

  - présentation de la partition finale:

    - pour 6 classes initiales demandées, la partition finale présente le coefficient d'appartenance de chaque individu à chacune de ces 6 classes.

2°/ Le second fichier (appelé OUT2) isole la matrice des coefficients d'appartenance des individus aux différentes classes. Ce fichier de données pourra être utilisé à différents usages. Il pourra, par exemple, être uni à un "fichier système" SPSSx pour y faire l'objet de différentes procédures de traitement des données.

NOMBRE DE CLASSES DEMANDEES : 6 NOMBRE D'OBJETS : 25

ITERATION	CRITERE	INDICE DE VALIDITE
1	0.4798E+06	0.528
2	0.1723E+08	0.698
3	0.1668E+08	0.635
4	0.1667E+08	0.647
5	0.1667E+08	0.653
6	0.1666E+08	0.657
7	0.1666E+08	0.661
8	0.1666E+08	0.663
9	0.1666E+08	0.665
10	0.1666E+08	0.666
11	0.1666E+08	0.667
12	0.1666E+08	0.668
13	0.1666E+08	0.668

PARTITION FINALE **********

	1	2	3	4	5	6
[1]	0.000	0.000	0.997	0.000	0.002	0.000
2	0.000	0.315	0.004	0.002	0.679	0.000
3	0.000	0.002	0.955	0.000	0.043	0.000
[4]	1.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.350	0.003	0.001	0.646	0.000
6	0.000	0.000	0.990	0.000	0.010	0.000
7	0.000	0.000	0.995	0.000	0.005	0.000
[8]	0.000	0.988	0.000	0.004	0.008	0.000
9	0.000	0.001	0.986	0.000	0.014	0.000
10	0.000	0.012	0.217	0.001	0.771	0.000
11	0.000	0.011	0.241	0.000	0.747	0.000
12	0.000	0.002	0.981	0.000	0.017	0.000
13	0.000	0.012	0.181	0.000	0.807	0.000
14	0.000	0.620	0.002	0.002	0.376	0.000
15	0.000	0.000	0.996	0.000	0.004	0.000
16	0.000	0.958	0.000	0.028	0.014	0.000
17	0.000	0.892	0.001	0.001	0.106	0.000
18	0.000	0.000	0.996	0.000	0.004	0.000
19	0.000	0.000	0.997	0.000	0.003	0.000
[20]	0.000	0.047	0.003	0.000	0.950	0.000
21	0.000	0.011	0.066	0.000	0.923	0.000
22	0.000	0.550	0.003	0.002	0.446	0.000
[23]	0.000	0.000	0.000	1.000	0.000	0.000
24	0.000	0.000	0.997	0.000	0.003	0.000
[25]	0.000	0.000	0.000	0.000	0.000	1.000

3°/ Le troisième fichier (appelé OUT1) isole les individus "prototypiques". Ce fichier se présente sous la forme d'une matrice carrée:

à chaque classe correspond un individu sélectionné parce qu'il possède le coefficient d'appartenance le plus élevé dans cette classe; la matrice est donc formée par les coefficients d'appartenance de ces *n* individus aux *n* classes de la partition finale (encadrés dans la partition finale présentée ci-dessus).

Ce fichier fournit aussi des données utilisables dans d'autres logiciels.

Ces deux derniers fichiers (OUT1, OUT2) sont extraits du premier fichier de résultats (HNDR RES.). Ce premier fichier de résultats présente toutes les partitions correspondant aux différentes solutions demandées (2 classes, 3, 4, ... 10 classes).

Les deux derniers fichiers (OUT1, OUT2) sont TOUJOURS extraits de la partition qui compte le PLUS GRAND NOMBRE DE CLASSES.

Si l'usager a demandé les solutions envisageant de 2 à 7 classes, les fichiers OUT1 et OUT2 seront extraits de la partition en 7 classes.

IV. CONCLUSION

Dans l'ensemble des méthodes de classification "floue" le MNDr (Méthode des nuées dynamiques à exposant *r*), se distingue par quelques caractéristiques importantes:

- le MNDr peut s'appliquer à des données non-métriques;
- il n'impose pas de définition d'un "centre de gravité" des classes, car il mesure la distance entre chaque individu et tous les autres individus appartenant à une classe;
- il n'implique pas la formation d'agrégats sphériques pour des données qui, le plus souvent, ne s'y prêtent pas;
- il permet de respecter les appartenances multiples de certains individus à plusieurs classes;
- la partition finale est peu influencée par la partition initiale, proposée au moment où le nombre de classes est déterminé;
- Le logiciel est muni d'une procédure de calcul de la validité de la solution proposée. Cet indice de validité accorde une importance particulière aux individus appartenant à plusieurs classes (ou, individus difficilement attribués à une classe déterminée).

ORIENTATION BIBLIOGRAPHIQUE

HUBERT L.(1972): "*Some extensions of Johnson's hierarchical clustering algorithms*", in Psychometrika, Vol. 37, 3.

LIBERT G., ROUBENS M.(1982): "*Non metric fuzzy clustering algorithms and their cluster validity*", in: "Approximate Reasoning in Decision Analysis". Eds. GUPTA M., SANCHEZ E., North-Holland, pp. 417-425.

LIBERT G., ROUBENS M.(1983): "*New experimental results in cluster validity of fuzzy clustering algorithms*", in: "New Trends in Data Analysis and Applications". Eds. JANSSEN J., MARCOTORCHINO J.F., PROTH J.N., North-Holland, pp. 205-218.

LIBERT G.: "*Classification automatique*", in: Revue belge de Statistique, d'Informatique et de Recherche Opérationnelle, Vol. 23, N°3, pp. 44-80.

ALDENDERFER M.S., BLASHFIELD R.K.(1984): "Cluster Analysis", Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-044, Beverly Hills and London: Sage Pubns.

MILLIGAN G.W., COOPER M.C.(1985): "*An examination of procedures for determining the number of clusters in a data set*", in: Psychometrika, 50, 159-179.

LIBERT G.(1986): "*Non metric cluster analysis*", in International Encyclopedia of Systems and Control. Ed. M.G. SINGH, Oxford, Pergamon Press.

LIBERT G.(1986): "*Compactness and number of clusters*", in: Control and Cybernetics, Vol.15, n°2, pp. 205-211.