Les méthodes de classification

I/ La méthode de Classification Hiérarchique des données :

La méthode de classification est une technique qui consiste à regrouper les observations en classes de telle sorte que les éléments d'une même classe soient plus proches entre eux que d'un élément quelconque d'une autre classe.

L'objectif de l'analyse hiérarchique est d'aboutir à des classes d'équivalence portant sur les observations (ou les variables) au regard des dimensions retenues. Deux types de modèles peuvent être utilisés à cette fin :

- Les modèles de partition ou modèles de hiérarchie se réfèrent essentiellement au principe de ressemblance ou de regroupement ;
- Les modèles de segmentation utilisent quant à eux le principe de séparation.

On se réfère en général à la « taxinomie » qui est la science des lois de la classification des formes vivantes (notamment animales et végétales) et qui a été étendue au domaine statistique et économique afin de distinguer différentes classes ou familles au sein d'une population quelconque (régions, départements, branches…) matérialisée par une matrice quelconque. Le taxinomiste cherche avant tout à faire passer les observations dans des filtres successifs qui transforment la réalité première, inconnue ou difficile à lire, en une autre réalité plus facile à déchiffrer, mais obtenue au prix de certains déformations. En effet, le statisticien confronté à une masse énorme de données doit trier les données, les classer et les codifier pour une meilleure transparence en vue de les commenter et d'en tire des conclusions. Néanmoins, ces transformations aboutissent indubitablement à une diminution de l'information disponible. Il s'agit donc de concilier les opérations de transformation des données avec l'exigence d'une perte minimale de l'information.

La classification automatique peut être appréhendée sous deux angles différents : soit la partition, soit la hiérarchisation.

Section 1 : Les méthodes de partition

Nous partons toujours d'un tableau de données E * M.

(E : l'ensemble des variables [x_1, ….xi,….xn] ; M : l'ensemble des observations [1….j….m])

Tout couple (xi,j) $\in$ E$^{\times}$ M est une symbolisé par une grandeur Yij ; (xi,j) $\in$ E$^{\times}$ M $\rightarrow$ Yij

L'objectif est de partitionner E en un certain nombre restreint de classes différentes autant que possible les unes des autres, en fonction de leurs caractères j $\in$ M. En outre, ces classes doivent présenter une caractéristique liée à l'homogénéité de chaque classe vis-à-vis de ces caractères. Le critère appliqué peut être la « similarité » pour inclure une grandeur Yij dans une classe p, ou inversement un certain degré de « différenciation » pour l'en exclure. On abouti ainsi à regrouper les diverses grandeurs Yij présentant des ressemblances dans des classes distinctes et homogènes.

Avant de préciser la méthode qui permet de mesurer l'homogénéité des classes, nous allons définir la notion de partition.

1- définition d'une partition :

Soit un ensemble E : P(E)= [A, B, C….G]

Une partition sur un ensemble [P(E)] est un ensemble de classes qui satisfait deux conditions :

a) deux classes A et B sont soit disjointes, soit confondues :

$$\left.\begin{matrix}\forall A \in P(E) \\ \forall B \in P(E)\end{matrix}\right\} A \cap B = \emptyset \; ou \; A \cup B = B$$

b) l'union de toutes les classes correspond à l'ensemble E :

$$A \cup B \cup C \cup D \cup F \cup G = E$$

2- La mesure de l'homogénéité des classes :

Le raisonnement se fait soit en termes de ressemblance (regroupement des éléments) soit en termes de dissemblance (séparation des éléments). La différence entre ces deux principes selon Grigg « qu'au lieu de rechercher des similarités, on observe des différences et au lieu de construire, on sépare ».

Supposons que l'on dispose d'un tableau de données E*M,

E étant l'ensemble des observations, par exemple les 39 provinces marocaines (n= card (E)= 39) ; M un ensemble de caractères relatifs à la population par âge, par sexe, à la production par secteur… (m= card (M)).

M E	1……..j ….....m	
1 x_h x_i x_n	y_{hj} y_{ij}	y_h y_i
	Y_j	y

Yij est la mesure j de la variable observée dans la région i.

La tâche du statisticien sera d'abord de tirer les données, de partitionner les observations sous forme de classes homogènes sur la base de leur ressemblance, c'est-à-dire de leur « dimensions connues ». La ressemblance entre les observations peut être mesurée par leur distance en regard des diverses variables retenues.

On dit que d (application de E*E dans IR^+) est une distance sur E si les trois axiomes suivants sont satisfaits :

$d(xi,xh) = 0 \Leftrightarrow xi = xh \quad \forall\,(xi\,,xh) \in E_2$ (Axiome de séparation)

$d(xi,xh) = d(xh,xi) \quad \forall\,(xi\,,xh) \in E_2$ (Axiome de symétrie)

$d(xi,xh) \leq d(xi,xe) + d(xe,xh) \quad \forall\,(xi\,,xh,xe) \in E_3$.

(Axiome d'inégalité triangulaire : c'est-à-dire qu'un côté est inférieur ou égal à la somme des deux autres).

Dans le cas où seuls les deux premiers axiomes sont vérifiés, nous sommes plutôt en présence d'un indice de distance.

Généralement, les données sont d'abord centrées et réduites, sauf dans le cas où les données se présentent sous forme d'un tableau de fréquences.

A partir du tableau de données brutes ci-dessus, on calcule les formules suivantes :

$$Y.j = \sum_{i \in E} yij \qquad Yi. = \sum_{j \in M} yij$$

$$\bar{Y}j = \frac{1}{n} \sum_{i \in E} Yij \qquad \bar{Y}i = \frac{1}{m} \sum_{j \in E} Yij$$

$$\sigma ij_2 = \frac{1}{n} \sum_{i \in E} [Yij - \bar{Y}j]_2$$

$$\sigma i_2 = \frac{1}{m} \sum_{j \in E} [Yij - \bar{Y}i]_2$$

Les principaux indices de distance et les distances utilisées dans les classifications sont :

$$d^2(i,h) = \sum_{j=i}^{m} (Yij - Yhj)^2 \ distance \text{ euclidienne.}$$

$$d^2(i,h) = \sum_{j=i}^{m} \frac{1}{\sigma^i} (Yij - Yhj)^2 \ distance \text{ euclidienne normée.}$$

$$d(i,h) = \sum_{j=i}^{m} \frac{Yij - Yhj}{Yij - Yhj} \ Bray\ (1957) et\ William\ (1966).$$

$$d^2\ (i,h) = \sum_{j=1}^{m} \frac{Y..}{Y.j} \left(\frac{Yij}{Yi.} - \frac{Yhj}{Yh.} \right)^2 \ Benzécri\ (1965).$$

$$d(i,h) = \sum_{j=1}^{m} p(i) |Yij - Yhj| \ Wall\ (1969) \qquad Jonhnson (1969).$$

$$d(i,h) = (\sum_{j=1}^{m} p(i)|Yij - Yhj|^{1/\gamma})^{\gamma} \text{ distance de Minkowski.}$$

$\gamma \in (0,1)$

$$\text{pour } \gamma = \frac{1}{2}, \text{on obtient la distance euclidienne.}$$

$$\text{pour } \gamma = 1 \text{ et } p(i) = 1, \text{on obtient la formule de Wall et Johnson}$$

$$d(i,h) = \sum_{j=i}^{m} (Yij - Yhj) \text{ distance euclidienne.}$$

$$d(i,h) = \frac{1}{m}\sum_{j=i}^{m} (Yij - Yhj) \text{ Indice des différences absolues}$$

$$d(i,h) = 1 - \frac{\sum_{j=1}(Yij - \bar{Y}\imath)(Yhj - \bar{Y}h)}{\left[\sum_{j=1}^{m}(Yij - Yh)\right]^2 \left[\sum^{m}(Yhj - \overline{Yh})\right]^2}$$

$$d^2(i,h) = \sum_{j=i}^{m} \frac{1}{\sup(Yij - Yhj)^2}(Yij - Yhj)^2$$

Avec Jambu, nous pouvons effectivement noter que la quasi-totalité de ces formules sont basées sur la différence (Yij- Yhj), pondérée par un coefficient qui dépend de la nature des mesures.

Dans le cas d'un tableau de fréquence :

M E	1……j…….m	
1 h i n	 P_{hj} P_{ij}	 P_i
	P_j	

$$Nous\ avons\ pi. = \sum_{j=1}^{m} \frac{Yij}{Y..} = \frac{Yi.}{Y..}$$

$$ph. = \sum_{j=1}^{m} \frac{Yhj}{Y..} = \frac{Yh.}{Y..}$$

$$p.j = \sum_{j=1}^{m} \frac{Yij}{Y..} = \frac{Y.j}{Y..}$$

$$p = \sum_{i=1}^{n} \sum_{i=1}^{m} Pij$$

la distance utilisée ici peut être la distance du $\boldsymbol{x2}$ *de Benzécri* :

$$d^2\ (i,h) = \sum_{j=1}^{m} \frac{Y..}{Y.j} \left(\frac{Yij}{Yi.} - \frac{Yhj}{Yh.}\right)^2\ Benzécri\ (1965).$$

Cette multiplicité des formules de distance soulève la nécessité du meilleur choix :

« L'utilisateur doit choisir celle qui mettra le mieux en évidence les ressemblances et dissemblances dans le cadre de son étude ».

Au terme de ces opérations, on parvient à construire un tableau matriciel représentant les indices de distances respectifs entre les différentes paires d'observation classées selon un ordre croissant (ou décroissant) des distances. C'est une matrice carée, symétrique, avec une diagonale nulle.

Matrice des indices des distances

E / E	……x_h……x_i……x_n
.	0
.	
.	
x_h	0 d_{hi}
.	0
.	
x_i	d_{ih} 0
.	
.	
.	0
.	
x_n	

Dans le cas d'une matrice booléenne (ou tableau de description logique) où tout xij est associé à une valeur 0ou 1 (xij aura la valeur 1 lorsque l'observation i possède le caractère j et la valeur 0 si elle ne la possède pas), la ressemblance des observations est mesurée par le biais de leur similarité en regard des variables prises en considération.

M E	1……m	
x_1		
.		
.		
.		
x_h	0 1 0 0 1 1 0 0 0 1 1 1	
.		
.		
x_i	1 1 1 0 0 1 1 0 1 0 0 1	
.		
.		
.		
.		
x_n		

Matrice Booléenne

Les indices de similarité les plus utilisés sont :

Nom	**Expression algébrique**	**Intervalle de variation**
Dice et Sorensen	$2m_{ih} / 2m_{ih}+u$	0-1
Jaccard	$m_{ih} / m_{ih}+u$	0-1
Sneath et Sokal	$m_{ih} / m_{ih}+2u$	0-1
Kulczynski	m_{ih} / u	0-1
OCHIAI	$m_{ih} / m_i m_h$	0-1
Sokal et Michener	c / m	0-1
Sneath et Sokal	2c / (c+m)	0-1
Rogers et Tanimoto	c / (c+2u)= c / m+u	0-1
Sneath et Sokal	c / u	0-1

Comme dans le cas précédent, on obtient un tableau matriciel d'indices de similarité entre les différentes paires d'observations ordonnées selon le degré décroissant de leur similarité. Il s'agit d'une matrice carrée, symétrique, dont la diagonale est unitaire (S_{ii}= S_{hh}= 1).

E E	x_1……..xh…….xi…… ...xn
x_1 . . . x_h . . . x_i . . . xn	1 1 1 1 1

Matrice des indices de similarité

La formule suivante liant les deux types d'indices permet de passer d'un tableau d'indices de similarité à un tableau d'indices de distance, ou inversement :

D_{ih}= 1- S_{ih}

Parmi les méthodes qui utilisent les techniques de partition, on relève une méthode perfectionnée et d'un intérêt pratique remarquable : il s'agit de la méthode des classifications ascendantes hiérarchiques qui fera l'objet des développements qui suivent.

Section II : La méthode des classifications ascendantes hiérarchiques

Le point de départ est toujours un tableau de données E*M (ou E*E), avec E : L'ensemble des observations ; M : L'ensemble des variables.

Le but est de partitionner E (ou M) en un nombre réduit de classes différentes les unes des autres, mais présentant chacune une certaine homogénéité en regard des variables retenues. Pour cela, on choisit un indice de distance (ou de similarité) pour mesurer les ressemblances entre les observations (ou les variables). On obtient ainsi un tableau carré E*E d'indices de distance ou de similarité, qui permet de regrouper (ou d'agréger) les éléments de E qui présentent une ressemblance entre eux, d'une manière ascendante. On aboutit alors à des « classes » ou « partitions emboîtées » sur E (ou M) qui caractérisent l'arbre hiérarchique.

En fait, on peut retenir trois types de hiérarchie :
- la hiérarchie simple : caractérisée par l'ensemble des parties composées par un ou plusieurs éléments de E. Exemple : E [A, B, C, D, F]

L'inconvénient de cette méthode réside dans la difficulté de détermination de l'ordre des regroupements pour la construction de l'arbre.

- La hiérarchie stratifiée : dans ce cas, l'indétermination précédente est levée, puisque l'ordre correspond aux étapes successives de regroupement progressif des éléments et des parties de E. Plus on s'élève sur l'axe vertical, plus les classes deviennent de moins en moins fines.
- Hiérarchie indicée : ici on fait intervenir un indice pour mesurer l'ordre d'apparition des nœuds. L'indice varie selon la position du nœud sur l'axe vertical et croît au fur et à mesure que le regroupement a lieu tardivement. On se réfère habituellement à la

distance ultramétrique, notée d, qui outre l'ordonnancement qu'elle réalise des éléments de E, fournit également une mesure des écarts interclasses des éléments de E.

d définit sur E est une distance ultramétrique, si les trois conditions suivantes sont vérifiées :

1- condition de séparation : $d(xi,xi) = 0$

2- condition de symétrie : $d(xi,xh) = d(xh,xi)$

3- condition de Krassner : $d(xh,xi) \leq \sup d(xh,xe), d(xe,xi) \qquad (xe,xh,xi) \in E$

L'indice respectant les conditions (1) et (2) est un indice de distance. Si en plus la condition d'inégalité triangulaire est aussi respectée, l'indice est alors une distance.
La condition de Krassner ou axiome d'inégalité triangulaire signifie que « tout triangle est soit équilatéral, soit isocèle ; et s'il est isocèle c'est la base qui est le plus petit côté ».
On peut distinguer globalement deux procédés essentiels de regroupements hiérarchiques des éléments de E :

1- Les regroupements progressifs :

Nous avons dit précédemment que les observations sont regroupées en classes suivant leur ressemblance vis-à-vis des variables. La question qui se pose maintenant est de savoir comment relier ces différentes classes pour aboutir à des partitions de moins en moins fines et en définitive à une hiérarchie. Les regroupements se font généralement de trois manières :

- soit sur la base de la distance minimale qui sépare une observation d'un groupe d'observations. C'est ce qu'on appelle la « loi de liaison simple ».

$$d_{ea} = \min (d_{ei}, d_{eh})$$

- soit sur la base du degré minimum de la distance maximale qui sépare une observation de toutes les observations d'une classe. C'est ce qu'on appelle la « loi de liaison complète ».

$$d_{ea} = \max (d_{ei}, d_{eh})$$

- soit enfin sur la base du minimum de la distance entre une observation et un groupe d'observations.

$$d_{ea} = \frac{1}{2}(d_{ei}, d_{eh})$$

Ou en généralisant : $d_{Moy}(A,B) = \frac{1}{card\ A.cardB} \sum_{\substack{xj \in A \\ xe \in B}} d(xi, xe).$

2- Les passages directs à l'ultramétrique :

Il s'agit ici de trouver un moyen pour passer d'un triangle quelconque caractéristique des distances normales à un triangle isocèle (ou équilatéral) associé à la distance ultramétrique (sans déformation majeure) c'est-à-dire qui respecte la condition de Krassner.

Considérons trois observations x_e, x_h, x_i avec des indices de distance respectifs entre elles die, d_{eh} et d_{ih} ; $d_{ie} > d_{eh} > d_{ih}$

Ces trois observations peuvent former le triangle suivant :

Le passage à l'ultramétrique peut se faire selon trois procédés :

- le procédé de Roux : l'ultramétrique inférieure maximale. Ce procédé consiste à réduire le plus grand indice pour le rendre égal au plus grand des deux autres :

- les deux procédés de Benzécri : les ultramétriques supérieures minimales qui consistent :

⇒ soit à élever le plus faible indice pour le rendre égal au plus grand :

⇒ soit à élever l'indice intermédiaire pour l'égaliser avec le plus élevé :

Le choix de tels procédés obéit à la nécessité de respecter l'éloignement ou la proximité selon le cas. Il va de soi que l'on n'aboutit pas à une classification hiérarchique unique, car le recours à des techniques différentes peut modifier la disposition de l'arbre hiérarchique. Néanmoins, la perte d'information à travers les étapes de transformation des données peut être un moyen susceptible de guider le choix des techniques les plus appropriées. Du reste, « en comparant tous les résultats de ces techniques, apparaissent des regroupements stables qui constituent des noyaux forts. Les éléments qui changent souvent de classes, suivant les méthodes, constituent des éléments marginaux ». C'est pourquoi il est recommandé de recourir à différentes méthodes pour mieux approcher la réalité. On peut conclure avec Jambu : « à partir du moment où on a admis que les méthodes de classification ascendante hiérarchique ne peuvent fournir un optimum absolu, et qu'une méthode n'exprime qu'un point de vue déformé de la réalité, l'art et la sagesse du statisticien seront de confronter, synthétiser plusieurs de ces réalités, en enrichissant ces confrontations de

différents calculs de déformation, imparfaits certes, mais qui, en tout état de cause, permettent quand même d'approfondir la connaissance des données.

II/ La classification hiérarchique en tant qu'instrument de mesure multidimensionnelle de l'évolution des inégalités :

Nous avons vu que le statisticien qui opère dans un champ donné, délimité dans l'espace et dans le temps, procède par étapes chronologiques :

1- les données recueillies qui traduisent des phénomènes observés sont disposées et codifiées dans un tableau E*E (ou E*M) ;
2- ensuite, le statisticien choisit une distance qui exprime la ressemblance (ou la similarité) des observations ou des variables ;
3- la mesure des ressemblances entre des éléments étudiés permet d'obtenir des indices de distance (ou de similarité) que l'on dispose sous forme d'une matrice carrée E*E.
4- on agrège (ou en regroupe) les éléments de E qui présente une ressemblance entre eux, d'une manière ascendante. On obtient ainsi des « partitions emboîtées » sur E qui caractérisent l'arbre hiérarchique.

Mais comment l'arbre hiérarchique peut-il servir de mesure de l'inégalité ?

D'abord il faut que certaines conditions concernant les données soient réunies et cela par le biais de certaines transformations manuelles de la part du statisticien. Ces transformations concernent l'homogénéisation des données d'une part et leur signification d'autre part. sur la base de ces deux conditions, il est possible d'avoir deux appréciations de l'inégalité :

- une mesure statique de l'inégalité,
- une mesure dynamique relative à l'évolution de l'inégalité au cours d'une période donnée.

Section I : les conditions préalables

Ces conditions sont au nombre de deux :

1- la première condition se situe au niveau du codage des données : les données doivent être homogènes pour éviter les distorsions susceptibles de fausser les mesures. Il faut s'attacher à ce que les grandeurs des données ne soient pas trop disparates pour qu'il n'y pas une prédominance écrasante d'une seule variable sur toutes les autres. Cela permet par

conséquence de conférer à toutes les variables une importance uniforme afin de tenir compte de l'effet de chacune d'entre elles dans l'ordonnancement des observations et donc d'éviter de masquer la contribution de certaines variables.

var obs	V1	V2	V3
1	113	145300	2559000
2	88	6225	443000

Le tableau homogène qui sera retenu peut être le suivant :

var obs	V1	V2	V3
1	113	145	256
2	88	6	44

2- il faut également que les variables aient la même signification, c'est-à-dire qu'elles soient toutes caractéristiques du seul phénomène étudié pour que la mesure ait elle-même un sens. En effet, si les variables sont opposées, le résultat sera biaisé dans la mesure où il ne pourra s'identifier à aucun phénomène précis. Par exemple, lorsqu'on cherche) étudier les disparités sociales, les variables doivent toutes traduire des structures d'ordre représentatives soit de la richesse, soit de la pauvreté. L'appréciation qui en résulte représentera dans le premier cas les écarts de richesse entre les groupes sociaux et dans le deuxième cas des écarts de pauvreté. Dans l'une ou l'autre situation, on peut avoir une mesure de l'inégalité, à condition de ne pas les mélanger. En vue de respecter cette exigence, il s'avère souvent nécessaire de recourir à l'inverse des valeurs de certaines variables. C'est ainsi que si l'on opte pour la mesure des inégalités de richesse.

Ces deux conditions étant remplies, il est possible d'aboutir à une mesure de l'inégalité.

Section II : La mesure de l'inégalité

Tout d'abord, il faut souligner que l'inégalité absolue est une notion qui comporte une connotation subjective et il convient de ce fait de l'écarter de notre champ d'investigation. Par contre, l'inégalité relative nous paraît d'une objectivité notable dans la mesure où elle concerne des différences qui caractérisent des situations réelles relatives soit à des catégories sociales, soit à des unités régionales.

L'intérêt de la méthode de classification hiérarchique est de nous fournir une mesure multidimensionnelle de l'inégalité. De surcroît, cette mesure peut être soit statique, soit dynamique.

1- La mesure multidimensionnelle de l'inégalité sur le plan statique :

Nous avons précisé précédemment que la méthode de classification hiérarchique consiste à agréger les éléments de E les plus semblables dans des classes homogènes, c'est-à-dire que les éléments dont les sommes des distances en regard des n variables sont voisines, sont regroupées dans une même classe. Or lorsqu'on sait que les variables représentent les composantes de l'inégalité, on comprend que la somme de ces distances matérialise en fait l'écart entre deux éléments i et j de E sur la base des variables retenues.

L'homogénéité d'une classe peut être mesurée par l'inertie intra-classe. Si E est partitionnée en 1classe : C_1, C_2…..C_E…..C_1, avec $E= U_{e=1}\, C_e$, l'inertie intra-calsse par rapport au centre de gravité g se présente comme suit :

$$Ice = \sum_{xi \in Ce} Pi\, d^2(xi, ge)$$

g_e : le centre de gravité de la classe Ce ;
P_i : le poids de Ce.

Cette inertie intra-classe doit être minimale pour que la classe soit homogène. Chaque classe homogène peut être repérée par son nœud de regroupement qui correspond à un certain

niveau de l'indice de distance ultramétrique porté par l'axe vertical de l'arbre hiérarchique. Etant donné que le classement hiérarchique se fait de façon ascendante, les regroupements les plus homogènes seront situés au bas de l'échelle. Plus on s'élève sur l'axe vertical, plus l'indice de distance croît. Par conséquent, la distance ultramétrique qui sépare deux classes homogènes constituera l'inégalité multidimensionnelle entre ces deux classes. Elle peut être également mesurée par l'inertie intra-classes des centres de gravité g_e des classes Ce par rapport au centre de gravité g de l'ensemble E :

$$Iinter - classe = \sum_{e=1}^{1} P(Ce)d^2(ge, g)$$

L'indice de distance ultramétrique peut donc être considérée comme une mesure des disparités multidimensionnelles inter-classes.

Exemple :

Supposons que l'on ait huit catégories socio- professionnelles (C.S.P) :

[C_1, C_2……C_8] et n variables [V_{01}, V_{02}…….V_N] représentant le revenu par ménage, la taille du ménage, la consommation, l'éducation...

On obtient l'arbre hiérarchique suivant :

Le groupe le plus homogène est représenté par C_5-C_8 car l'indice de distance est le plus bas (0,074). L'inégalité entre la classe C_5-C_8 et la catégorie C_6 peut être mesurée par la différence

(0,220- 0,074= 0,146) entre les deux premiers indices. Il est ainsi possible d'évaluer avec précision l'écart qui sépare les diverses C.S.P. nous constatons par exemple que l'inégalité entre C_7 et C_3 (0,146) est moins grande que celle entre C_3 et C_4 (2,903).
Enfin, on peut même calculer l'inertie totale de E, en faisant la somme de l'inertie intra-classe et de l'inertie inter-classes.

$$IE(g) = \sum_{i \in u} Pi\ d^2(xi, ge) + \sum P\ (\ Ce)\ d^2\ (ge, g).$$

En fait, l'inertie totale n'a un intérêt pratique que lorsqu'on compare l'évolution de l'inégalité dans le temps.

2- La mesure dynamique de l'inégalité multidimensionnelle :

Pour apprécier l'évolution de l'inégalité dans le temps, il suffit de construire deux (ou plusieurs) arbres hiérarchiques à deux (ou divers) points dans le temps et de comparer les distances ultramétriques entre les classes. Selon que celles-ci réduisent ou s'accroissent, on aura une diminution ou une augmentation de l'inégalité au cours de la période étudiée.
En effet, si entre t_1 et t_2 :

$I(j)^{t}{}_1 < I(j)^{t}{}_2$: les inégalités ont augmenté ;
$I(j)^{t}{}_1 > I(j)^{t}{}_2$: il y a eu resserrement des disparités entre t_1 et t_2;
$I(j)^{t}{}_1 = I(j)^{t}{}_2$: l'éventail est resté inchangé.

En outre l'arbre hiérarchique ne se limite pas seulement à mesurer l'inégalité multidimensionnelle. Il permet également de renseigner sur l'évolution structurelle des catégories sociales ou des unités spatiales dans le temps. En effet, si la structure des C.S.P par exemple se modifie entre t_1 et t_2, cela signifie qu'il y a eu en plus de la variation d'inégalité une mobilité sociale au cours de la période étudiée, étant donné que les regroupements entre C.S.P sont différents en t_1 et en t_2. Par contre, lorsque les C.S.P ne se déplacent pas dans le temps, la mobilité sociale n'est pas vérifiée, mais l'inégalité sociale s'est réduite. Enfin, on peut avoir aussi le cas où la mesure de l'inégalité sociale est inchangée alors que la mobilité sociale se traduit par une substitution entre les C.S.P. Néanmoins, cette dernière éventualité est peu vraisemblable car la modalité sociale conduit généralement à une variation de l'inégalité sociale.

Conclusion :

Ainsi, l'intérêt de la méthode de classification hiérarchique se trouve cerné. Le passage suivant de VOLLE le résume bien : « regrouper les éléments d'un ensemble en 'paquets', c'est simplifier cet ensemble et faciliter sa description. En ce sens, les classifications sont un instrument de la statistique descriptive. Mais elles jouent également un autre rôle, plus fondamental peut-être. Elles impliquent une schématisation et un appauvrissement de la réalité que l'on peut déplorer, mais sans une telle schématisation il n'est pas de statistique possible, ni même de construction intellectuelle : le réel est toujours trop divers et trop mouvant pour être appréhendé dans toute sa richesse ».

Outre la visualisation et l'ordonnancement des observations (ou variables), la méthode de classification permet de prévenir les erreurs qui peuvent découler de l'analyse factorielle. En effet, deux observations voisines dans l'espace IR^m se projettent côte à côte sur un plan factoriel à deux dimensions. Néanmoins, la proximité de deux observations dans la projection ne signifie pas forcément qu'elles sont proches dans l'espace IR^m. Par conséquent, l'analyse hiérarchique permet d'éviter une telle confusion grâce à l'ultramétrique qui regroupe les éléments affectés d'un indice de distance voisin, c'est-à-dire les éléments dont la ressemblance est la plus forte. On obtient ainsi des classes homogènes qui nous éclairent sur la rectification à apporter aux erreurs subséquentes à la projection.

Cette méthode de classification nous donne également une indication claire de la discrimination entre les groupes d'observations, que l'on peut mesurer grâce aux indices de distance ultramétrique.

Elle constitue donc un instrument performant de la mesure de l'inégalité puisqu'elle l'appréhende sous son aspect réel, c'est-à-dire en respectant sa multidimensionnalité. C'est aussi une mesure objective de l'inégalité dans la mesure où elle se départit de toute intuition, grâce évidemment à l'usage de l'ordinateur. Cette remarque de Benzécri est significative à cet effet : « on imagine qu'un naturaliste qui travaille sans machine, en surveillant pas à pas le progrès d ses constructions, soumettra volontiers les résultats des calculs à des transformations que lui inspire son intuition ; la machine ne permet guère cela ».

Néanmoins, cette méthode ne va pas sans soulever certains problèmes, voire des inconvénients. On peut relever le fait que le choix de la distance et du procédé reste arbitraire.

De même, à la suite de Bertier et Bouroche, on peut soutenir :

- qu'elles nécessitent le calcul des distances interpoints [n (n-1)/ 2];

- et qu’à toutes les étapes, elles procèdent en n-1 itérations. Lorsque n est grand, ces méthodes deviennent lourdes et coûteuses.

Exercice 1 :

Soit un ensemble E= (a, b, c, d, e, f) et le tableau des distances sur E.

Construire l’arbre hiérarchique correspondant en utilisant l’ultramétrique inférieure maximale.

Tableau des indices de distances

	a	b	c	d	e	f
a	0	8	7	8	7	8
b	8	0	5	4	5	7
c	7	5	0	7	3	2
d	8	4	7	0	8	6
e	7	5	3	8	0	1
f	8	7	2	6	1	0

Réponses :

Pour trouver une solution à ce problème, il convient de classer tous les triangles que l’on peut former à l’aide des éléments de E. Puis on vérifie que chaque triangle est soit isocèle avec comme base le petit côté, soit équilatéral. Dans le cas contraire, on réduit le plus grand côté pour le rendre égale au plus grand des deux autres. Souvent, la correction d’un triangle modifie le caractère isocèle ou équilatéral d’un autre triangle. C’est la raison pour laquelle il est nécessaire de vérifier plusieurs fois les triangles jusqu’au moment où aucune modification des indices ne soit requise pour rendre un triangle isocèle ou équilatéral.

Les triangles	Les côtés		
abc	cb	ca	ba
abd	db	da	ba
abe	cb	ea	ba
abf	fb	fa	ba
acd	dc	da	ca
ace	ec	ea	ca
acf	fc	fa	ca
ade	ed	ca	da
adf	fd	fa	da
aef	fe	fa	ea
bcd	dc	db	cb
bce	ec	eb	cb
bcf	fc	fb	cb
bde	ed	eb	db
bdf	fd	fb	db
bef	fe	fb	eb
cde	ed	ec	dc
cdf	fd	fc	dc
cef	fe	fc	dc
def	fe	fd	ed

Indices de distance					
ba	8	fc	7	fb	cb
ca	7	ed		eb	db
cb	5	fd	7	fb	db
db	8	fe	5	fb	eb
dc	4	ed		ec	dc
eb	7	fd	2	fc	dc
ec	5	fe	7	fc	dc
ed	3	fe	5	fd	ed
fa	8				
fb	8				
fc	7				
fd	2		5		
fe	6				
	1				

Tableau des ultramétriques inférieures maximales

	a	**b**	**c**	**d**	**e**	**f**
a	0	7	7	7	7	7
b		0	5	4	5	5
c			0	5	2	2
d				0	5	5
e					0	1
f						0

Une hiérarchie indicée correspond au tableau des ultramétriques inférieures maximales ci-dessus :

On remarque que les nœuds de la hiérarchie sont ordonnés à partir de Card n+1, soit de 7 à 11.

Exercice 2 : Application des méthodes d'analyse des données aux provinces du Maroc (1960-1970)

A- Liste des provinces :

01- Agadir
02- Ouarzazate
03- Tarfaya
04- Marrakech
05- Safi
06- Béni Mellal
07- Casablanca
08- Tanger
09- Tétouan
10- Kénitra
11- Al Hoceima
12- Fes
13- Taza
14- Nador
15- Oujda
16- Meknes
17-Ksar – Es- Souk

B- Liste des variables

- Population

V02- population urbaine
V03- population rurale
V04- taux d'urbanisation
V05- population active occupée
V06- population inactive
V07- demandeurs d'emploi

- Répartition de la population active selon le secteur d'activité

V08- dans l’agriculture

V09- dans l’industrie

V10- dans les services

- Agriculture

V11- superficie des terres agricoles

V12- superficie des terres irriguées

V14- céréales- production

V15- légumineuses- production

V16- rendement en quintaux par hectares- céréales

V17- rendement en quintaux par hectares- légumineuses

V18- arboriculture et plantations fruitières

V19- élevage- cheptel

- Industrie

V20- production d’électricité hydraulique

V21- production d’électricité thermique

- Tourisme : capacité hôtelière

V23- 4 et 5 étoiles

V24- 1,2 et 3 étoiles

V25- villages de vacances et centres balnéaires

- Infrastructure :

+ Réseau routier :

V27- principal

V29- tertiaire

+ Ports : Trafic des marchandises

V30- débarquement

V31- embarquement

+Aéroport :

V32- principaux (trafic- passagers)

V33- secondaire (nombre)

+ Télécommunication :

V34- nombre de lignes téléphoniques

+ Education :

V35- effectif total du primaire

V36- effectif total du secondaire

V37- effectif total du supérieur

+ Santé :

V38- capacité hospitalière

V39- nombre de médecins

V40- nombre de pharmaciens

V41- nombre de dentistes

V42- nombre de sages femmes

V43- nombre d'infirmières

V44- nombre de vétérinaires

+ Jeunesse et sports :

V45- maisons de jeunesse et foyers culturels

V46- centres d'accueil

V47- camps de vacances

V48- colonies urbaines

V49- foyers féminins

V50- terrains de sport

V51- piscines

+ Etablissements de projection cinématographique :

V52- nombre de places

Interprétation : l'analyse hiérarchique

Au seuil d'indice 4,8, on distingue six sous groupes de régions :

- Le premier sous-groupe, constitué des régions 14, 11,03, 02, 13, 05, 17, 08 et 09, couvre le nord, l'est et l'extrême sud du pays.

Ce groupe de régions ne présente pas un haut degré d'homogénéité puisque les regroupements se font assez tardivement, ce qui souligne les disparités interrégionales en son sein. Néanmoins, par rapport aux autres régions, cet ensemble présente une certaine homogénéité qui tient au fait qu'il réunit des régions rurales arriérées, pratiquant une agriculture traditionnelle et extensive. Là où l'activité agricole est très pauvre, un secteur tertiaire se développe et concerne essentiellement le petit commerce ; c'est le cas de Tarfaya ou Tanger, par exemple, comme le montre l'axe II. Dans ces régions l'exode rural accentue le taux de chômage qui semble subir une atténuation par rapport à 1960 (il a été réduit de moitié au cours de la période étudiée à Tanger).

L’appartenance de la région 08 à ce groupe souligne l’échec des initiatives publiques en vue de développer cette région par le truchement, notamment des avantages du code des investissements et la prime d’équipement pour encourager l’industrialisation de la région de Tanger. Néanmoins, cela a permis la naissance de certaines industries agro-alimentaires, du textile et des conserves de poisson, et c’est la raison pour laquelle cette région ne se regroupe avec les autres qu’au niveau d’indice 4,459.

- Le sous-groupe constitué de Fes, Meknès, Marrakech et Agadir.

Il s’agit de régions à vocation principalement agricole. Le secteur urbain commence à se développer avec l’apparition d’activités industrielles ‘textile, chaussure et huile d’olives à Fes ; industries alimentaires et cimenteries à Meknès ; l’huile d’huile d’olive à Agadir), mais l’essentiel de la population active demeure dans le secteur primaire (54.6% à Fes, 66.6% à Agadir, 62.3% à Marrakech et 47.5% à Meknès).

Dans l’agriculture, les cultures sont assez diversifiées et les rendements sont très moyens. D’autre part, l’infrastructure touristique (V23, V24, V25 et V27) et sanitaire s’est nettement développée par rapport à 1960. Il faut également noter des réalisations dans les domaines de la culture, du sport (V45, V46, V47, V48, V49, V50, V51 et V53) et de l’enseignement (V35, V36, et V37). Enfin le secteur de l’artisanat a gagné en importance par rapport à t1.

Oujda : c’est une région où l’agriculture ne joue pas un rôle primordial. En effet, la population active primaire (37,2%) est bien en deçà de la moyenne nationale (qui est de l’ordre de 51,7%). Il en découle que les secteurs secondaire et tertiaire sont assez développés. L’industrie concerne essentiellement les industries alimentaires, des métaux, de la chimie, du bois et du coton. Quant à l’agriculture elle présente une diversité des cultures (V14, V15, V18), mais la spécialité de la région est l’élevage (ovin et caprin). Les activités touristiques, de même que l’infrastructure routière, sanitaire, culturelle et sportive sont peu développées.

- Béni Mellal : c’est une région riche qui possède la plus grande superficie du pays (V12). Les rendements sont élevés surtout dans les cultures céréalières et l’arboriculture est très importante (V18). Par contre, l’industrie est extrêmement réduite. L’artisanat est orienté principalement vers l’autoconsommation et les infrastructures socio-économiques (routes, santé, culture, enseignement et tourisme) ne sont pas très développées.
- Kénitra : c’est région agriculture riche avec un secteur urbain assez développé essentiellement avec les villes de Rabat, Salé et Kénitra qui ont connu une croissance vertigineuse de la population (65,4% et 35,1% respectivement) au terme de la période 1960-1970.

L’agriculture est très importante (V08, V11 et V12) surtout grâce à la plaine du Gharb dont la fertilité des terres est élevée.

Les cultures sont intensives et diversifiées (V14, V15 et V18) et l’élevage représente une activité appréciable.

L’industrie est tournée vers la transformation (cellulose, sucrerie, huile, engrais ect.), le textile et quelques usines de poisson.

Le tourisme est peu important, mais les autres infrastructures sont assez développées : par exemple le réseau routier (V27, V28 et V29, le domaine de la santé, du loisir, ect. Rabat constitue le principal centre universitaire du pays et la région 10 vient en deuxième position (après Casablanca) en ce qui concerne l’effectif de l’enseignement primaire et secondaire.

Enfin Casablanca : c’est la région la plus peuplée et la plus développée du Maroc. Le taux d’urbanisation est très élevé (58,8%). Les trois secteurs de l’activité économique (primaire, secondaire et tertiaire) se partagent la population active avec une légère supériorité dans le secteur primaire. L’agriculture, rentable et diversifiée (associant les cultures et l’élevage), couvre la superficie la plus étendue du pays. Le secteur industriel, qui emploie la plus forte proportion de la main-d’œuvre du pays, concentre la majeure partie des établissements industriels, énergétiques et miniers (notamment les phosphates à Khouribga et à Youssoufia). C’est une région très équipée (infrastructure routière, hôtelière, sanitaire, scolaire, culturelle, etc).

Au niveau d’indice 6, on ne distingue plus que trois sous-groupes : Casablanca, Kénitra et le reste du pays.

Le regroupement de la région Kénitra avec les autres régions ne se fait qu’au niveau d’indice 8,738, ce qui atteste de grandes disparités interrégionales. Quant à Casablanca, elle surclasse l’ensemble des régions du pays qu’elle ne rejoint qu’au niveau ultime, c'est-à-dire au niveau d’indices 16,117.