MATHÉMATIQUES ET SCIENCES HUMAINES

C. ARNAL
A. BATTINI
J. P. MARCIANO

Le problème des descripteurs classifiants : une illustration sur la structure du tiers monde

Mathématiques et sciences humaines, tome 91 (1985), p. 23-55
<http://www.numdam.org/item?id=MSH_1985__91__23_0>

Math. Sci. hum. (23e année, n°91, *1985*, p.23-55)

LE PROBLEME DES DESCRIPTEURS CLASSIFIANTS : UNE ILLUSTRATION SUR LA STRUCTURE DU TIERS MONDE

Par C. ARNAL, A. BATTINI, J.P. MARCIANO*

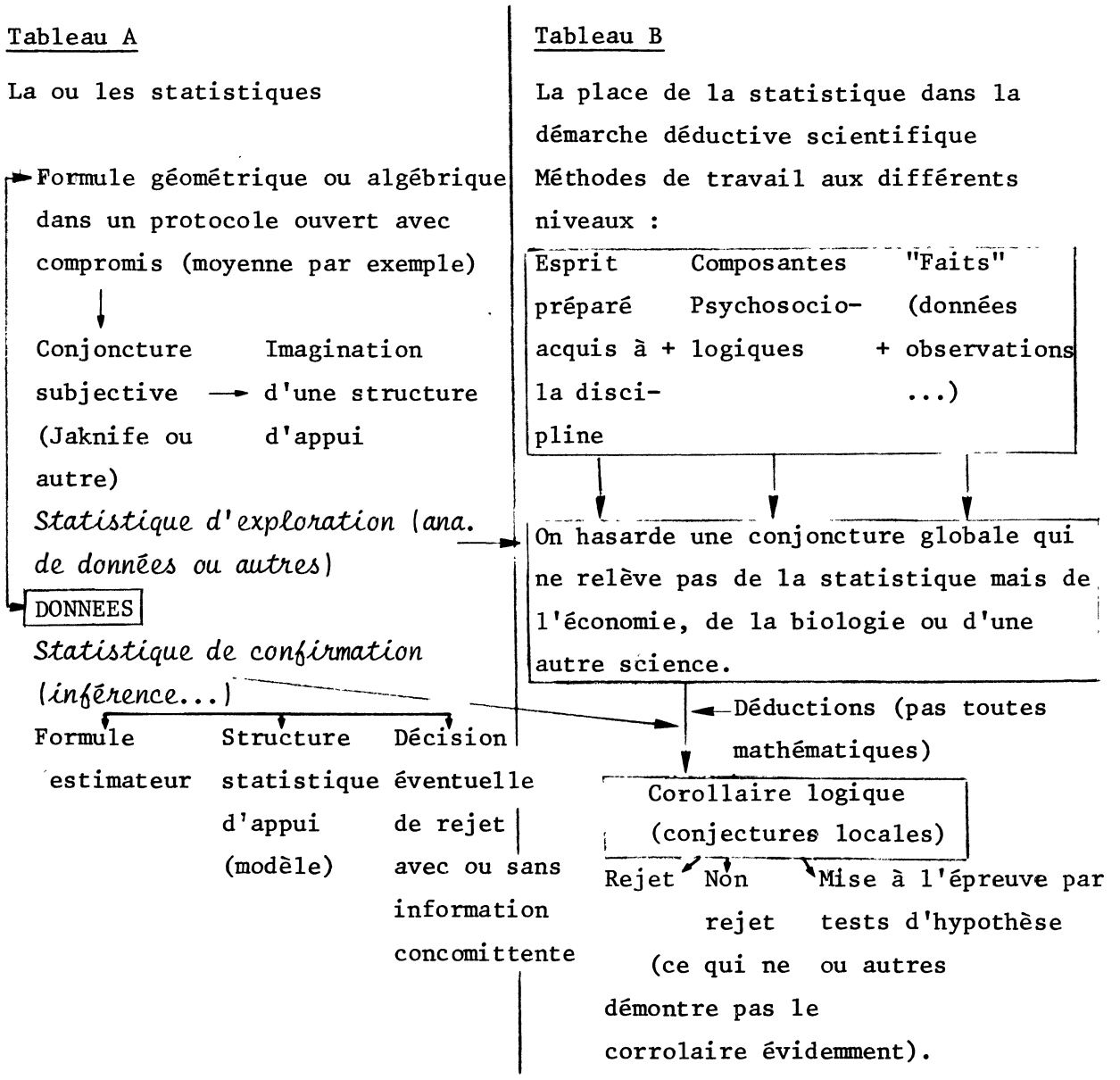
PROLOGUE : D'UN PARADIGME A L'AUTRE

Les statisticiens sont très sensibles aujourd'hui au "partage" entre statistique exploratoire (analyse de données, classification, etc.) et statistique confirmatoire classique selon le tableau A suggéré par J.R. Barra.

Les spécialistes d'analyse de données ont réussi à en faire une des facettes de la statistique et on accepte de plus en plus l'analyse de données sinon comme un paradigme au moins comme une branche d'une statistique dont peut être affirmée l'unicité dans le cadre de la méthodologie scientifique. Il s'agit dans cette étude non pas d'abord de contribuer à la construction d'un éventuel paradigme de l'analyse exploratoire des données incluant la classification mais d'aborder un autre problème qui s'inscrit plutôt en amont, l'analyse de système, rapport d'un objet et de l'ensemble de ses descripteurs, sorte d'alphabet de variables disponibles pour le décrire. L'étude recherche un principe et une métarègle pour extraire de cet alphabet un mot conduisant à la discrimination la plus pertinente entre les classes d'objets, ce mot de variables prenant une valeur différente pour chaque objet décrit.

Il est nécessaire auparavant de replacer non seulement la Statistique y compris l'analyse de données, mais aussi l'économétrie, le calcul économique, etc. dans une démarche scientifique comme celle de Popper ou de Kuhn dont l'ouvrage vient d'être repris en français (tableau B).

* Resp. Assistant, Chercheur et Directeur.
Atelier de Prévision de la Faculté d'Economie Appliquée, Université d'Aix-Marseille III.



Les deux facettes citées plus haut de la statistique apparaissent alors comme souvent complémentaires car intervenant à des niveaux différents de la démarche.

Selon le tableau B, un économiste du développement aura par exemple à priori quelque idée sur la classification des pays (développés, en voie de développement, riches en énergie, sous-développés) ; il y ajoutera les "faits", c'est-à-dire un ensemble de descripteurs, séries prises dans un annuaire disponible pour ces pays, objets de son étude.

Dans un premier niveau, les méthodes plus ou moins automatiques de classification peuvent guider la phase exploratoire ce qui permettra d'affiner les conjectures globales ; ainsi, par exemple, la construction d'un

modèle économétrique expliquant le taux de croissance de la production nationale en fonction du taux de croissance de la production industrielle apparaîtra mieux adapté à une certaine classe de pays qu'à l'ensemble. Les deux facettes sus-citées de la statistique font apparaître là nettement leur caractère complémentaire. Mais l'arbre ne doit pas cacher la forêt. Un problème méthodologique au moins aussi important est celui de l'identification de l'objet (le pays) à l'ensemble des descripteurs disponibles (l'annuaire) et des biais qu'elle peut entraîner ; le problème de l'existence du paradigme de l'analyse de données risque d'être effacé par celui du paradigme systémique (tableau C) d'E. Morin (B.9).

Tableau C. Définition et descripteurs d'un objet

		Fonctionnelle	: ce qu'il fait (il fonctionne)
Objet-Acteur (ici pays) sous l'angle économique	Définitions	Ontologique	: ce qu'il est (il se stabilise)
		Génétique	: ce qu'il devient
Métarègle de discrimination Le meilleur filtre est un "mot de variables"			
Annuaire de descripteurs (alphabet de variables)	Est-ce une description exhaustive du patrimoire génétique ?		

Dans ce cas, le problème n'est plus d'opposer statistique inférentielle et statistique sans modèle (au sens probabiliste) mais en amont, de prendre conscience qu'un Annuaire -Alphabet de variables- est déjà un modèle. De même, une de ses parties, le mot de variables jugé le plus discriminant, résultat d'une systémographie initiale sera un modèle plus ou moins pertinent avec en tout cas, un champ de validité à définir. A défaut, d'un automate formel ou d'une métarègle quelques règles automatisables tout à fait originales sont ici établies pour un meilleur choix des descripteurs des acteurs, préalablement à une classification hiérarchique puis pour l'interprétation des résultats de cette dernière.

1. UNE METHODE TYPOLOGIQUE SUPPORT (B2) ET LE ROLE DES DESCRIPTEURS DES ACTEURS

1.0. Introduction

Pour cette étude a été choisie, parmi d'autres, une méthode exploratoire

de classification automatique ou algorithmique. L'automatisation d'une grande partie de la démarche typologique atténue le risque d'une personnalisation excessive des résultats et peut sembler plus objective.
L'étude est assortie de méthodes également automatiques d'aide à l'interprétation des classes obtenues.

Il apparaît qu'une excessive diversité ou l'existence de rendondances dans l'ensemble des descripteurs nuisent à la détermination de types dans l'ensemble des acteurs. On ne peut donc éviter un problème généralement mal maîtrisé de sélection des descripteurs les plus pertinents.

Une solution proposée ici, dans l'optique d'une classification ultérieure des objets, semble susceptible de guider efficacement cette opération inévitable de choix des descripteurs dans ce que l'on a appelé une analyse structurale des descripteurs.
La mesure des parts respectives des différents descripteurs dans la séparation des classes, jointe à la possibilité de réintroduction de descripteurs supplémentaires, permet d'améliorer l'interprétation de la classification.

La démarche générale proposée ici fait intervenir à la base une méthode ascendante de classification hiérarchique notée par la suite C. H. qui sera rappelée brièvement en 1.1. (voir ex. B.4). Les éléments méthodologiques plus originaux utilisés pour éclairer les résultats de la C.H. sont en grande partie déjà développés dans la thèse de C. Arnal, (B.2. 1980).

1.1. La classification des acteurs

1.1.1. La constitution d'une chaîne binaire de partitions

On définit une mesure de dissimilarité δ entre n objets-acteurs à classifier ainsi qu'entre deux ensembles disjoints ou non de ces objets : On supposera δ additivement décomposable selon p descripteurs quantitatifs : entre deux objets r et q on a

$\delta_{rq} = \sum_{j=1}^{j=p} \delta^j_{rq}$, avec δ^j_{rq} composante imputable au descripteur D^j dans la similarité δ_{rq}.

Tant que tous les objets ne sont pas réunis dans une seule et même classe, les deux objets les moins dissemblables sont agrégés et leur couple

constitue un nouvel objet remplaçant ses deux composants.

Remarque : *Lorsque plusieurs paires de classes réalisent la valeur minimale de δ le choix se porte arbitrairement ici sur la première paire rencontrée dans le programme informatique.*

Au bout de n-1 agrégations binaires, tous les objets sont réunis. On n'obtient pas directement une partition de l'ensemble des objets mais une suite de partitions en n-1, n-2, n-3,..., 3, 2 classes représentées traditionnellement par un arbre hiérarchique.

Quelle que soit la structure de la matrice des dissimilarités entre les objets, la C.H. produit un arbre hiérarchique, même dans le cas d'une répartition uniforme. Il est donc fondamental d'évaluer la séparation des classes dans les partitions obtenues.

Remarque : *On peut utiliser aussi, préalablement à toute classification, des méthodes d'appréciation de l'aptitude d'un ensemble à être classifié, Lerman (B.13).*

1.1.2. Une mesure originale de la pertinence de la partition des acteurs

On a réuni les n objets par n-1 agrégations binaires successives.
Les dissimilarités correspondantes entre objets agrégés sont notées :

$$\delta_1, \delta_2, \delta_3, \ldots\ldots\ldots, \delta_{n-1} .$$

La dissimilarité totale dans l'ensemble des objets est mesurée par la somme des "longueurs" des n-1 "liens" successivement établis pour les réunir ;

$$\delta_{tot} = \sum_{i=1}^{i=n-1} \delta_i .$$

A l'étape n-k de la C.H., après n-k agrégations binaires successives, on obtient une partition en k classes de l'ensemble des objets. Comment mesurer le degré f(k) de séparation entre les classes de cette partition ?

- La mesure proposée ici est le rapport entre la longueur moyenne δ INTER(k) des k-1 liens restant à établir entre les k classes pour une

réunion complète et la longueur moyenne $\bar{\delta}$ INTRA(k) des n-k liens déjà établis pour constituer les k classes :

$$f(k) = \frac{\bar{\delta}\ \text{INTER}(k)}{\bar{\delta}\ \text{INTRA}(k)} = \frac{\dfrac{\sum_{i=n-k+1}^{i=n-1} \delta_i}{k-1}}{\dfrac{\sum_{i=1}^{i=n-k} \delta_i}{n-k}} \quad .$$

La ou les partitions de l'arbre hiérarchique, retenues comme supports de la typologie, correspondront aux maximas absolus ou relatifs de l'indicateur f(k) fonction du nombre k des classes.

Remarque : Dans le cas particulier où l'on choisit comme mesure de dissimilarité δ la "distance de l'inertie" on retrouve avec f(k) la "variance ratio criterion" présentée par Calinski et Harabasz (J.4) en 1974 et justifiée par des propriétés géométriques directement liées à l'inertie.
D'autres mesures ont été proposées pour la pertinence des partitions (voir notamment Lerman (B.13) et Jambu (B.12)).

Quoiqu'il en soit avant d'obtenir une C.H. satisfaisante des objets-Acteurs étudiés, il est généralement nécessaire de procéder à un choix judicieux de leurs descripteurs.

1.2. L'analyse structurale des descripteurs

Cette étape du traitement des données n'est généralement pas automatique et toujours sujette à un certain arbitraire. Elle repose ici néanmoins sur quelques principes et instruments originaux d'orientation des choix susceptibles d'améliorer les résultats typologiques.

1ère étape :

On se borne à recueillir de façon extensive les descripteurs disponibles selon la nature (économique, géographique, climatique,...) de la typologie recherchée, l'ensemble des séries d'un annuaire, par exemple.

Soit p le nombre de ces descripteurs initialement retenus. Généralement avec cet ensemble brut, riche en redondances et trop disparate, la

combinaison dans la C.H. des p répartitions des objets-Acteurs induites isolément par chaque descripteur aboutit plutôt à l'uniformité qu'à l'émergence de classes séparées et ceci d'autant plus que p est grand.

2ème étape :

Une classification des descripteurs précède la recherche dans chaque classe obtenue d'un élément représentatif sous la contrainte de l'induction d'une répartition en classes similaires de l'ensemble des objets-Acteurs. Le respect strict de cette contrainte peut parfois amener l'abandon complet de certaines classes de descripteurs.

Ainsi, d'un nombre p de descripteurs, on passe à un nombre m inférieur du fait de l'abandon de certaines classes de descripteurs et de la sélection d'un seul élément représentatif par classe retenue.

Il faut alors se demander si l'appauvrissement de la description des objets qu'a entraîné la sélection de descripteurs dits co-classifiants n'est pas de nature à trop atténuer la portée des conclusions typologiques et si en conséquence la poursuite des opérations se justifie.

1.2.1. La classification des descripteurs

Deux descripteurs sont équivalents s'ils engendrent entre les objets des dissimilarités proportionnelles.

Avec ce critère, on classifie les descripteurs en utilisant une variante originale de la C.H. (voir (B.2)) 2e partie, ch. 2, algorithme TYPVA) qui permet d'obtenir sous la forme d'un arbre une chaîne binaire de partitions à fortes corrélations intra-classes et faibles corrélations inter-classes en valeurs absolues.

Comme pour les objets se pose alors le problème du choix d'une partition dans l'arbre hiérarchique de leurs descripteurs. Soit m le cardinal de la partition choisie.

Reste alors à sélectionner dans chaque classe un descripteur au maximum de façon à éliminer les redondances et les classes qui perturberaient la classification des objets.

1.2.2. Un principe fondamental : la recherche d'un ensemble de descripteurs peu corrélés et co-classifiants des acteurs

Il va falloir rechercher un compromis entre deux objectifs : d'une part représenter un nombre maximum des m classes de descripteurs pour ne pas trop appauvrir la description des objets, d'autre part assurer le caractère co-classifiant des descripteurs représentatifs retenus. C'est-à-dire l'induction d'une répartition en classes similaires de l'ensemble des objets.

Le choix de descripteurs faiblement liés, à partir de la classification de ces derniers est expérimenté par ailleurs, notamment par Geffrault (B15) cependant le caractère co-classifiant des descripteurs sélectionnés n'y semble pas assuré.

D'autre part, Lerman (B13, chap.3) propose une mesure de la neutralité "classificatoire" des variables et l'élimination des plus neutres de façon à obtenir une classification des objets plus nette. Là encore cette façon de procéder ne nous semble pas assurer la convergence vers une seule et même partition de l'effet "classifiant" des variables ainsi indépendamment sélectionnées. Les variables les plus "classifiantes" ne sont pas nécessairement co-classifiantes.

Un descripteur est classifiant s'il induit dans R^1 une répartition des objets en intervalles séparés. Une C. H. des objets décrits par cet unique descripteur donnerait un indicateur $f(k)$ possédant un maximum ou des maxima relatifs. Supposons maintenant qu'une C.H. portant sur les mêmes objets mais décrits par un autre descripteur unique produise un $f(k)$ atteignant un maximum pour un même nombre k_o de classes. Cela signifie-t-il pour autant que les deux descripteurs soient co-classifiants ? Non, car rien n'assure que les deux partitions en k_o classes issues de deux C.H. distinctes soient composées des mêmes classes.

Le principe fondamental de co-classification :

Pour résoudre ce problème on effectue une C.H. préliminaire des objets décrits simultanément par les p descripteurs disponibles, avec le calcul pour chaque descripteur D^j d'un indicateur $f_j(k)$ construit sur le même modèle que l'indicateur global $f(k)$.

L'indicateur $f(k)$ est une moyenne pondérée des différents $f_j(k)$. En effet,

$$f(k) = \sum_{j=1}^{j=p} f_j(k) \times \frac{\delta^j \text{ INTRA}(k)}{\delta \text{ INTRA}(k)}$$

avec

$$\delta^j \text{ INTRA}(k) = \sum_{i=1}^{i=n-k} \delta_i^j \quad ; \quad \delta_i = \sum_{j=1}^{j=p} \delta_i^j \text{ , et}$$

$$f_j(k) = \frac{\overline{\delta^j \text{ INTER}(k)}}{\overline{\delta^j \text{ INTRA}(k)}} .$$

Soient deux descripteurs D^j et $D^{j'}$. Si $f_j(k)$ et $f_{j'}(k)$ possèdent un maximum pour un même nombre k_o de classes, il s'agit bien cette fois des mêmes classes puisque $f_j(k)$ et $f_{j'}(k)$ sont issus d'une seule et même C.H. et les deux descripteurs D^j et $D^{j'}$ sont co-classifiants.

Bien entendu si l'indicateur global $f(k)$ produit par la C.H. préliminaire des objets indique clairement une bonne participation, la sélection des descripteurs devient inutile, on les conserve tous, mais ce n'est généralement pas le cas.

On dispose alors de deux instruments pour le choix des descripteurs co-classifiants :

- premièrement la composition des m classes issues de la classification des descripteurs,
- deuxièmement du tableau des $f_j(k)$ issus de la classification préliminaire des objets.

Si les données utilisées se prêtent à la classification des objets on repère dans le tableau des $f_j(k)$ des zones de valeurs de k dans lesquelles de nombreux descripteurs voient leur $f_j(k)$ atteindre un maximum au moins relatif.

L'idéal est alors la présence d'au moins un de ces descripteurs co-classifiants dans chacune des m classes de descripteurs.

Par tâtonnements dûs par exemple à la présence de plusieurs intervalles intéressants dans les valeurs de k, on est généralement amené à retenir ℓ

descripteurs co-classifiants avec $\ell < m$. De plus, le prélèvement de chacun de ces descripteurs dans une classe différente obtenue par C.H., limite la possibilité de corrélations élevées entre les descripteurs sélectionnés et permet donc d'éviter les redondances.

Si les répartitions induites par les différents descripteurs ne convergent pas suffisamment vers une classification des objets, ℓ est très nettement inférieur à m et l'investigation typologique dans le domaine de description choisi initialement est alors sujette à caution. Par contre, si la représentation de ℓ classes seulement sur m n'entraîne pas un appauvrissement excessif de la description des objets on passe à la classification de ces derniers caractérisés par les ℓ descripteurs co-classifiants sélectionnés.

1.3. L'interprétation automatique de la classification des acteurs

1.3.0. La classification des acteurs et les fondements de son interprétation automatique

Le comportement de $f(k)$ quand on parcourt l'arbre produit par la C.H. amène à choisir une partition. L'interprétation mathématique de la partition choisie aura deux aspects, l'un à dominante descriptive l'autre plus explicatif.

1ère étape :
On caractérise la structure de la partition à l'aide des dissimilarités entre objets, entre classes, entre l'ensemble global des objets et ces dernières. Cette première étape est de nature essentiellement descriptive, consistant à faire apparaître notamment l'agencement des classes les unes par rapport aux autres, leur homogénéité respective.

2ème étape (plus explicative) :
On développe les dissimilarités entre objets en sommes de composantes, faisant ainsi apparaître les actions d'intensités diverses exercées par les différents descripteurs dans la détermination de la partition analysée. L'appréciation des parts imputables aux différents descripteurs, dans la séparation des classes de la partition constitue au moins un élément d'explication du phénomène de discontinuité se manifestant dans l'ensemble des objets.

Les méthodes présentées ici restent cependant purement quantitatives. Se ramenant à l'étude de l'organisation d'un ensemble de dissimilaritées et à la mesure des contributions des différents descripteurs à celles-ci, ces méthodes sont indépendantes de la discipline scientifique à laquelle se rattachent les

données. Elles devraient être accompagnées d'une interprétation déductive selon le schéma B déjà cité, plus subjective ou théorique, propre au chercheur et à la discipline à laquelle appartiennent les objets-acteurs décrits.

Dans les ouvrages et travaux de recherche récents, la place accordée aux algorithmes d'obtention de classes est en général considérable par rapport à celle faite aux méthodes d'aide à l'interprétation des classes obtenues. Dans cette optique la présentation synthétique tentée dans les pages suivantes pour cette catégorie de méthodes constitue selon nous une contribution à la méthodologie typologique. Certaines des méthodes réunies ici se retrouvent sous diverses formes spécifiques dans les ouvrages de Benzecri (B10), Jambu (B12), Lerman (B13)(B14), Hardouin, Chantrel (B16), Geffrault (B15), Chandon, Pinson (B17), d'autres à notre connaissance sont originales comme les mesures proposées pour la dissimilarité interne des classes ou la séparation entre classes.

1.3.1. La structure géométrique de la partition obtenue

1.3.1.1. Deux règles complémentaires pour interpréter l'agencement des classes

Dans une partition obtenue en coupant un arbre hiérarchique à une hauteur donnée les dissimilarités entre classes voisines peuvent être très diverses selon l'éloignement du niveau de l'arbre auquel intervient leur agrégation. Il est donc important pour déceler les particularités de la partition de positionner les classes les unes par rapport aux autres et par rapport à l'ensemble des objets.

1ère règle : évaluer l'excentricité des classes

Les dissimilarités δ entre les différentes classes et l'ensemble des objets (on rappelle que δ doit être définie également entre ensembles non-disjoints) permettent de juger des degrés de différenciation par rapport à une référence moyenne, de distinguer éventuellement de grosses classes centrales et des classes périphériques, ou au contraire de constater une équirépartition des classes.

2ème règle : évaluer le positionnement des classes dans leur environnement

La matrice des dissimilarités δ entre classes permet d'apprécier au niveau binaire la différenciation entre les classes. On peut, avec cette matrice, savoir si les classes à leur tour, se regroupent en amas locaux ou sont plus ou moins uniformément réparties, repérer les classes voisines et les classes extrêmes, noter ainsi les ressemblances ou contrastes essentiels, etc...

Remarque : *Certaines dissimilarités entre classes doivent être recalculées car*

la C.H. ne retient que n-1 dissimilarités "minimales".

1.3.1.2. La règle de comparaison des degrés de dissimilarité interne des classes

Une classe N de x objets obtenue à un niveau quelconque de l'arbre hiérarchique a toujours été construite par x-1 des n-1 agrégations binaires que comporte la C.H. complète. La dissimilarité interne δ_N d'une telle classe N est mesurée par la somme des x-1 dissimilarités correspondant aux x-1 agrégations binaires nécessaires à sa constitution.

En notant A_N le sous-arbre obtenu en remontant l'arbre hiérarchique de la classe N vers ses éléments terminaux, on a avec i comme indice des noeuds de l'arbre :

$$\delta_N = \sum_{i \in A_N} \delta_i \ .$$

Pour pouvoir comparer les degrés d'hétérogénéité de classes inégales en effectif, il faut prendre les dissimilarités internes moyennes, soit pour la classe N :

$$\overline{\delta}_N = \frac{\delta_N}{x - 1} \ .$$

1.3.1.3. L'appréciation binaire de la séparation des classes

La combinaison des informations apportées par les dissimilarités δ entre classes d'une part, par les dissimilarités internes de ces dernières d'autre part, permet d'apprécier le degré de séparation entre les classes prises deux à deux. Le degré de séparation de deux classes N_v et N_w est fonction croissante, à dissimilarités internes données, de la dissimilarité entre elles $\delta(N_v, N_w)$ et fonction décroissante à $\delta(N_v, N_w)$ fixée de leur degré moyen de dissimilarité interne.

La séparation entre N_v et N_w est mesurée par la différence $s(N_v,N_w)$ entre leur dissimilarité binaire $\delta(N_v,N_w)$ et la moyenne pondérée de leurs dissimilarités internes moyennes. Soient x_v et x_w les effectifs respectifs de N_v et N_w. On a alors :

$$s(N_v,N_w) = \delta(N_v,N_w) - \left(\frac{x_v - 1}{x_v + x_w - 2} \ \overline{\delta N}_v + \frac{x_w - 1}{x_v + x_w - 2} \ \overline{\delta N}_w \right).$$

Lorsque les classes N_v et N_w sont des singletons, leur degré moyen de dissimilarité interne est conventionnellement pris égal à zéro. On pourrait envisager de mesurer la séparation par un rapport et non une différence mais dans le cas de deux singletons le choix du degré arbitraire moyen de dissimilarité serait indéterminé, zéro ne convenant évidemment plus et un non plus puisque dans le cas où N_v et N_w ne sont pas des singletons, on peut avoir un degré moyen de dissimilarité interne inférieur à un.

1.3.2. Vers une métarègle pour l'identification des descripteurs les plus explicatifs de la partition des acteurs

1.3.2.1. Les descripteurs déterminants de chaque classe

La règle d'identification des descripteurs produisant l'excentricité d'une classe :

Cette analyse se retrouve dans certains travaux, notamment (B13,14,15,16) avec la possibilité de grouper les descripteurs pour évaluer le degré de responsabilité de classes de descripteurs dans la formation d'une classe d'objets, ou encore, la possibilité d'évaluation inverse du degré de responsabilité des objets dans la formation des classes de descripteurs.

On recherche les descripteurs engendrant l'essentiel de la dissimilarité entre une classe quelconque N_r et l'ensemble complet N_o des objets. dissimilarité notée,

$$\delta(N_r, N_o) = \sum_{j=1}^{j=\ell} \delta^j (N_r, N_o)$$

et pour laquelle la contribution relative d'un descripteur D^m notée $C_{D^m}(\delta(N_r, N_o))$, n'est autre que :

$$C_{D^m}(\delta(N_r, N_o)) = \frac{\delta^j(N_r, N_o)}{\delta(N_r, N_o)} .$$

Au moment de l'interprétation d'une classe donnée, les contributions permettent de séparer nettement les descripteurs singularisants et les autres. Cependant, si l'on peut associer de cette façon, à chaque classe, un classement des descripteurs en fonction de la contribution à son excentricité, cela n'indique pas si les valeurs prises sont relativement élevées ou au contraire relativement faibles. Il faut donc encore comparer par exemple les valeurs moyennes des descripteurs singularisants sur la classe N_r d'une part sur l'ensemble complet N_o d'autre part.

La règle d'identification des descripteurs produisant l'homogénéité d'une classe :

Les descripteurs assurant le plus de ressemblance entre les objets composant une classe ne sont pas nécessairement les mêmes que ceux donnant à cette dernière l'essentiel de sa singularité dans l'ensemble N_o des objets.

Les descripteurs recherchés sont ceux qui contribuent le moins à la dissimilarité interne (voir 1.3.1.2.) d'une classe N_r par exemple.

La dissimilarité interne de N_r, δ , peut se décomposer selon les ℓ descripteurs.

On a $\delta_{N_r} = \sum_{i \in A_{N_r}} \delta i$ et $\forall i$, $\delta_i = \sum_{j=1}^{j=\ell} \delta_i^j$;

on peut donc écrire :

$$\delta_{N_r} = \sum_{j=1}^{j=\ell} \sum_{i \in A_{N_r}} \delta_i^m \cdot$$

La contribution $C_{D^m}(\delta N_r)$ d'un descripteur D^m à la dissimilarité interne de N_r est alors, avec $\delta^m N_r = \sum_{i \in A_{N_r}} \delta_i^m$,

$$C_{D^m}(\delta_{N_r}) = \left\{ \frac{\delta^m N_r}{\delta N_r} \right. \cdot$$

1.3.2.2. La règle d'identification des descripteurs les plus discriminants entre deux classes

Parfois, pour deux classes distinctes N_r et N_ℓ, la hiérarchie des contributions relatives fait apparaître les mêmes descripteurs les plus déterminants. Dans ce cas, il est particulièrement important de préciser ce qui différencie les deux classes. Cette différence peut provenir tout simplement de valeurs opposées pour les descripteurs déterminants communs. Mais il peut arriver aussi que les descripteurs les plus déterminants communs exercent leur action dans le même sens. L'explication de la différenciation entre les deux classes est alors à chercher dans l'action des autres descripteurs.

La contribution relative d'un descripteur D^m à la dissimilarité entre N_r et N_s est :

$$C_{D^m}(\delta(N_r, N_s)) = \frac{\delta^m(N_r, N_s)}{\delta(N_r, N_s)} \cdot$$

Les descripteurs par lesquels N_r et N_s se différencient globalement le plus, ne sont pas nécessairement ceux séparant le plus efficacement les deux classes. En effet, un descripteur D^m contribuant fortement à $\delta(N_r, N_s)$ peut très bien aussi engendrer de telles dissimilarités internes dans N_r et N_s que les deux classes en arrivent à s'interpénétrer par les valeurs de D^m.

Pour mettre en évidence les descripteurs les plus séparateurs de N_r et N_s, il faut calculer leurs contributions relatives à la séparation des deux classes (voir 1.3.1.3. et 1.3.2.1.).

Pour un descripteur D^m une telle contribution s'écrit :

$$C_{D^m}(s(N_r, N_s)) = \frac{\delta^m(N_r, N_s) - \left(\frac{x_r - 1}{x_r + x_s - 2}\,\overline{\delta^m_{N_r}} + \frac{x_s - 1}{x_r + x_s - 2}\,\overline{\delta^m_{N_s}}\right)}{s(N_r, N_s)}$$

avec x_r et x_s respectivement effectifs de N_r et N_s.

Pour les descripteurs les moins séparateurs, pouvant amener isolément une interpénétration de N_r et N_s, les contributions prennent parfois des valeurs négatives.

1.3.2.3. La règle d'identification des descripteurs globalement les plus discriminants

La dissimilarité totale δ_{tot} dans l'ensemble des objets peut être décomposée selon les ℓ descripteurs :

$$\delta_{tot} = \sum_{j=1}^{j=\ell} \sum_{i=1}^{i=n-1} \delta^j_i \;;$$

$\delta^m_{tot} = \sum_{i=1}^{i=n-1} \delta^m_i$ est alors la dissimilarité totale imputable au descripteur D^m

Pour la partition de l'ensemble des objets en k classes par exemple, δ^m_{tot} peut être décomposée en dissimilarité inter-classes et dissimilarité intra-classes selon la formule :

$$\delta^m_{tot} = \delta^m{}_{INTRA}(k) + \delta^m{}_{INTER}(k).$$

Plus le taux de dissimilarité $t^m(k) = \frac{\delta^m_{INTER}(k)}{\delta^m_{tot}}$

est grand, plus la partition de l'ensemble des objets en k classes explique la dissimilarité totale imputable à D^m, ou formulé différemment, plus D^m est discriminant pour cette partition. Ce type de mesure de l'importance discriminante d'un descripteur se retrouve dans d'autres travaux, (B13, chap.3) (B14), avec notamment la possibilité de grouper les descripteurs pour évaluer leur coefficient de discrimination conjointe.

Toutes les méthodes d'identification des descripteurs les plus explicatifs de la partition faisant l'objet de la subdivision 1.3.2. peuvent être mises en oeuvre après réintroduction des $p - \ell$ descripteurs abandonnés lors de la sélection de descripteurs co-classifiants. Il n'est pas exclu non plus d'introduire pour caractériser les classes certains descripteurs totalement nouveaux.

Par ces moyens, on arrive à enrichir et nuancer l'interprétation de la partition obtenue de la C.H. des objets.

2. ILLUSTRATION : UNE PARTITION DE PAYS ACTEURS

2.1. Introduction

Nous avons vu dans le prologue (Tableau B de la démarche scientifique) que l'esprit acquis à la discipline était nécessaire avant la phase exploratoire d'où la nécessité dès cette introduction, d'une analyse qualitative du problème.

Créée dans les années cinquante par A. Sauvy pour caractériser un vaste ensemble de pays dont le trait dominant était la faiblesse du développement économique, l'expression "Tiers Monde" a eu la fortune que l'on sait.

Longtemps a prévalu, pour des raisons de commodité, de simplicité et de rareté des informations statistiques, la classification ordinale fondée sur le critère unique et tant critiqué (quand il n'est pas condamné) du P.N.B. par tête d'habitant exprimé en dollars des Etats-Unis.

Notre classification algorithmique notamment, suffisamment enrichie de méthodes d'aide à l'interprétation des résultats peut faire apparaître sur un ensemble de Pays Acteurs, chacun décrit par n descripteurs quantitatifs, une typologie qui ne résulte pas pour l'essentiel de critères arbitraires choisis à priori généralement orientés dans le sens que l'on veut donner à l'étude, en

provenance des inclinations conscientes ou non, avouées ou non de celui qui l'entreprend.

2.2. L'analyse structurale des descripteurs des Pays Acteurs

Les 54 Pays Acteurs décrits par 35 descripteurs socio-économiques sont assimilés à des points $\in R^{35}$ affectés de poids p_i (ici $\forall i = 1, \ldots, 54$; $p_i = 1$) où chacune des composantes est la valeur prise par un des 35 descripteurs ordonnés (tableau F). Les coordonnées du centre de gravité du nuage sont les moyennes des valeurs des descripteurs, $\forall j = 1, 2, \ldots, 35$;

$$\bar{x}^j = \frac{\sum_{i=1}^{i=54} p_i x_i^j}{\sum_{i=1}^{i=54} p_i} .$$

Les écarts types sont les normes des vecteurs de R^{54} représentatifs des variables centrées, $\forall j = 1, 2, \ldots, 35$;

$$\sigma^j = \left[\frac{\sum_{i=1}^{i=54} p_i (x_i^j - \bar{x}^j)^2}{\sum_{i=1}^{i=54} p_i} \right]^{\frac{1}{2}} .$$

Une standardisation des données augmentera la crédibilité des conclusions typologiques en assurant une certaine stabilité aux résultats, notamment par rapport au changement d'unité de mesure des variables, et en empêchant l'influence prépondérante et artificielle de celles dont les ordres de grandeur sont les plus élevés ; pour cela, à partir des valeurs brutes x_i^j des moyennes $\bar{x}^j$ et des écarts-types σ^j, on obtient les valeurs standardisées

$$x_i'^j = \frac{x_i^j - \bar{x}^j}{\sigma^j} .$$

Matrice des corrélations entre les 35 descripteurs

Deux descripteurs, variables quantitatives, sont considérés équivalents si l'un d'eux peut être reconstruit par transformation linéaire de l'autre, avec dans ce cas un coefficient de corrélation unitaire en valeur absolue.

Des coefficients de corrélation élevés signalent des redondances rela-

TABLEAU F : LES 35 DESCRIPTEURS DISPONIBLES DES PAYS ACTEURS

Nom des Variables pour le traitement informatique	SIGNIFICATION
Variables structurelles :	
POPULATI	Population en 1968 (millions d'hab.).
SUPERFIC	Superficie (milliers de km²)
DENSITE	Densité hab/km² (1968)
AGRI/PIB	Pourcentage de l'agriculture dans le PIB en 1970
INDU/PIB	Pourcentage de l'industrie dans le PIB en 1970
MANU/PIB	Pourcentage des industries manufacturières en 1970
COPUBPIB	Pourcentage de la consommation des administrations dans le PIB en 1970
CONS/PIB	Pourcentage de la consommation privée dans le PIB en 1970
FBCF/PIB	Pourcentage de la F.B.C.F. dans le PIB en 1970
X/PIB	Pourcentage des exportations dans le PIB en 1970
M/PIB	Pourcentage des importations dans le PIB en 1970
MANU/X	Pourcentage des produits manufacturés dans les exportations en 1970
PRIPRO/X	Part du principal produit d'exportation dans le total des exportations en 1970
AGRIPOPA	Part de la population active agricole dans la population active totale en 1977
INDUSPOP	Part de la population active industrielle dans la population active totale en 1977
URBANISA	Part de la population active urbaine dans la population totale en 1970 (ou 1965)
PIBPOP60	PIB/tête en $ en 1960 (année de base)
PIBPOP75	PIB/tête en $ en 1975 (année terminale)
ESPERVIE	Espérance de vie à la naissance en 1975
ANALPHAB	Pourcentage d'analphabètes dans la population âgée de plus de 15 ans (73-76)
SCOLARIS	Taux de scolarisation dans l'enseignement primaire et secondaire en 1975
ENSEISUP	Taux de scolarisation dans l'enseignement supérieur (20-24 ans) (73-76)
HABHOPIT	Nombre d'habitants par lit d'hôpital en 1975
HABMEDEC	Nombre d'habitants par médecin en 1975
Variables de tendance :	
CROIPOPU	Taux annuel de croissance démographique (1966-1975)
CROISPIB	Taux annuel de croissance du PIB global (1960-1976)
CRPIBPOP	Taux annuel de croissance du PIB/tête (1960-1976)
CRPRAGRI	Taux annuel de croissance de la production agricole totale (1960-1976)
CRPRINDU	Taux annuel de croissance de la production industrielle totale (1960-1976)
CRPRMANU	Taux annuel de croissance de la production manufacturière (1960-1976)
CRCONSOM	Taux annuel de croissance de la consommation privée (1960-1976)
CROIFBCF	Taux annuel de la F.B.C.F. privée (1960-1976)
CROISSAX	Taux annuel de croissance des exportations de biens et services (1960-1976)
FLUCTUAX	Indice de fluctuation des exportations (1965-1977)
CROISSAM	Taux annuel de croissance des importations de biens et services (1960-76)

Les tableaux D et E des données ne sont pas représentés ici.

tives dans l'ensemble de ces variables. Dans l'ensemble brut des descripteurs certains aspects du phénomène étudié sont généralement sur-représentés par des sous-ensembles de variables fortement corrélées entre elles.

On voit donc bien un des intérêts d'une classification des descripteurs et de la sélection d'un seul d'entre eux par classe obtenue.

Si l'on retient seulement les coefficients de corrélation au moins égaux à 0,68 en valeur absolue (corrélations positives ou négatives), nous pouvons établir le tableau synthétique G suivant :

TABLEAU G : QUELQUES DESCRIPTEURS CORRELES

BONNES CORRELATIONS NEGATIVES ($\rho < -0,7$)	BONNES CORRELATIONS POSITIVES ($\rho > 0,7$)
	1. DENSITE ⟷ M/P.I.B.
	2. X/P.I.B. ⟷ M/P.I.B.
	3. AGRIPOPA ⟷ AGRI/P.I.B.
4. AGRIPOPA ⟷ MANU/P.I.B.	
5. URBANISA ⟷ AGRI/P.I.B.	
6. ANALPHA ⟷ MANU/P.I.B.	
	7. SCOLARIS ⟷ MANU/P.I.B.
	8. ENSEISUP ⟷ MANU/P.I.B.
9. AGRIPOPA ⟷ INDUSPOP	
10. AGRIPOPA ⟷ URBANISA	
11. AGRIPOPA ⟷ PIBPOP 60	
12. AGRIPOPA ⟷ ESPERVIE	
	13. AGRIPOPA ⟷ ANALPHA
14. AGRIPOPA ⟷ SCOLARIS	
	15. INDUSPOP ⟷ SCOLARIS
	16. URBANISA ⟷ INDUSPOP
	17. URBANISA ⟷ PIBPOP 60
	18. URBANISA ⟷ ESPERVIE
	19. URBANISA ⟷ SCOLARIS
	20. URBANISA ⟷ ENSEISUP
	21. PIBPOP 75 ⟷ PIBPOP 60
22. ESPERVIE ⟷ ANALPHAB.	22. PIBPOP 60 ⟷ ESPERVIE
	23. ESPERVIE ⟷ SCOLARIS
24 ANALPHA ⟷ AGRIPOPA	
25. ANALPHA ⟷ SCOLARIS	

En dehors d'évidences inévitables à ce stade de l'analyse, les relations 6, 7, 8, 12, 13, 14, 15 d'une part, et les relations 17, 18, 19, 20, 22, 23, 24 d'autre part viennent corroborer d'une manière indéniable les enseignements les plus classiques pour ne pas dire traditionnels des diverses approches du sous-développement de la théorie du développement, notamment : le

rôle des industries manufacturières dans la formation du P.I.B. est primordial ; il s'accompagne d'une croissance de la scolarisation primaire (7) (avec diminution de l'analphabétisme (6)) et de l'enseignement supérieur (8) dans le cadre d'une urbanisation croissante (19,20) liée à une augmentation de la part de la population occupée dans le secteur industriel (15), phénomènes structurels et qualitatifs allant de pair avec une augmentation de l'espérance de vie (18, 22, 23).

L'arbre hiérarchique des descripteurs

Cet arbre représente la chronologie des fusions de descripteurs ou de classes de descripteurs au cours de la C.A.H.I.. Les premiers regroupements (à gauche sur la figure H) concernent les ressemblances les plus fortes entre les descripteurs.

L'indicateur f(k) de qualité des partitions obtenues par C.A.H.I. des descripteurs a été décrit en détail dans la première partie. On en cherche un maximum. Dans la C.A.H.I. des variables cet indicateur n'indique rien d'autre que les partitions triviales extrêmes et encore de façon très progressive ce qui montre l'absence de partition très nettement dominante dans l'ensemble des variables utilisées. Un indicateur plus fin comme celui de I.C. Lerman (B13) basé non sur la seule suite $\delta_1 \quad \delta_2, \ldots \delta_{n-1}$ mais sur la préordonnance totale aurait peut-être permis de déceler une partition "meilleure" que les autres dans l'ensemble des descripteurs.

De toute façon, l'objectif est moins de découvrir une partition dominante dans l'ensemble des descripteurs que de réduire systématiquement mais de façon pertinente (en minimisant la perte de dissimilarité inter-classes) le cardinal de cet ensemble. On peut donc choisir arbitrairement dans l'arbre hiérarchique une partition de cardinal "raisonnable". On peut aussi s'appuyer sur le comportement d'autres indicateurs t(k) et t'(k) basés comme f(k) sur la suite des dissimilarités $\delta_1, \delta_2, \ldots \delta_{n-1}$

$t(k) = \frac{\delta \text{ INTER}(k)}{\delta \text{ TOTALE}}$ décroît continûment de k=p à k=1

mais de façon accentuée juste après une éventuelle partition géométriquement pertinente. Aucune rupture de cette sorte n'apparaît dans le tableau et ce comportement est cohérent avec celui de f(k).

$t'(k) = \frac{1}{n} k + (1 - \frac{1}{n}) \frac{\delta \text{ INTRA}(k)}{\delta \text{ TOTALE}}$ exprime un compromis entre un petit

FIGURE H : LA C.A.H.I. des descripteurs et la partition optimale en 12 Classes

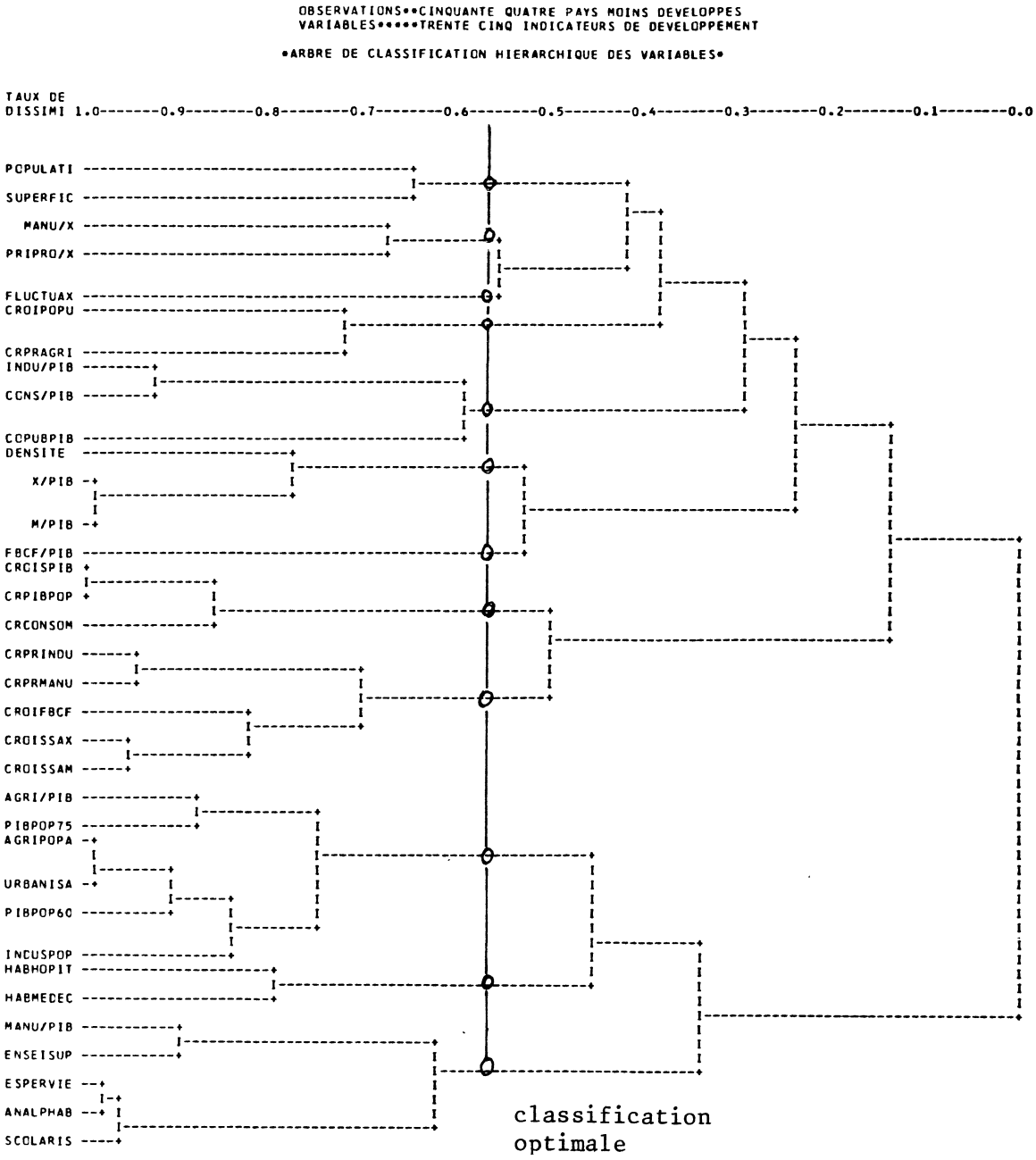


TABLEAU I : INDICATEUR DE QUALITE DES PARTITIONS DE L'ENSEMBLE DES DESCRIPTEURS

NB DE CLASSES	TX DE DISSIM T	INDIC QUAL PARTITION F	INDIC QUAL PARTITION K¹
1	0.000	0.000	1.000
2	0.129	4.928	0.902
3	0.231	4.820	0.832
4	0.291	4.253	0.802
5	0.337	3.825	0.786
6	0.378	3.530	0.775
7	0.418	3.360	0.764
8	0.458	3.265	0.754
9	0.496	3.209	0.745
10	0.527	3.105	0.744
11	0.558	3.037	0.743
12	0.587	2.979	0.743
13	0.615	2.940	0.744
14	0.644	2.923	0.745
15	0.671	2.915	0.748
16	0.697	2.923	0.750
17	0.723	2.941	0.754
18	0.747	2.963	0.759

nombre de classes et homogénéité de ces dernières, on en cherche un minimum. Il convient parfaitement à l'objectif poursuivi dans la C.A.H.I. des descripteurs. Il atteint son minimum pour k=12 et l'on a donc retenu la partition en 12 classes de descripteurs pour la suite de l'étude.

On va sélectionner des variables (en général, une seule) dans chaque classe pour la suite de l'étude. En ne retenant qu'une variable par classe, on perd une certaine information d'autant plus importante que la classe concernée est moins homogène. Si l'on estime dans certaines classes la perte d'information trop importante, on peut éventuellement en conserver plusieurs variables les moins corrélées possibles.

Tableau des $f_j(k)$ pour la recherche de variables co-classifiantes dans les 12 classes retenues

La construction du tableau des $f_j(k)$ a été expliquée dans la première partie. On cherche dans ce tableau des intervalles de valeur de k dans lesquels un grand nombre de $f_j(k)$ atteignent des maxima relatifs. On essaie ensuite de repérer au moins un des descripteurs D^j concernés par ce type de comportement de $f_j(k)$ dans chacune des classes de descripteurs précédemment constituées par C.A.H. Ainsi, après quelques tâtonnements, il est possible de sélectionner un certain nombre de descripteurs co-classifiants.

Dans notre exemple, de nombreux $f_j(k)$ passent par un maximum relatif entre k = 4 et k = 9. On a donc essayé de trouver des descripteurs de ce type dans les 12 classes de la partition retenue et il a été possible de sélectionner ainsi les 12 descripteurs suivant le principe fondamental de co-classification.

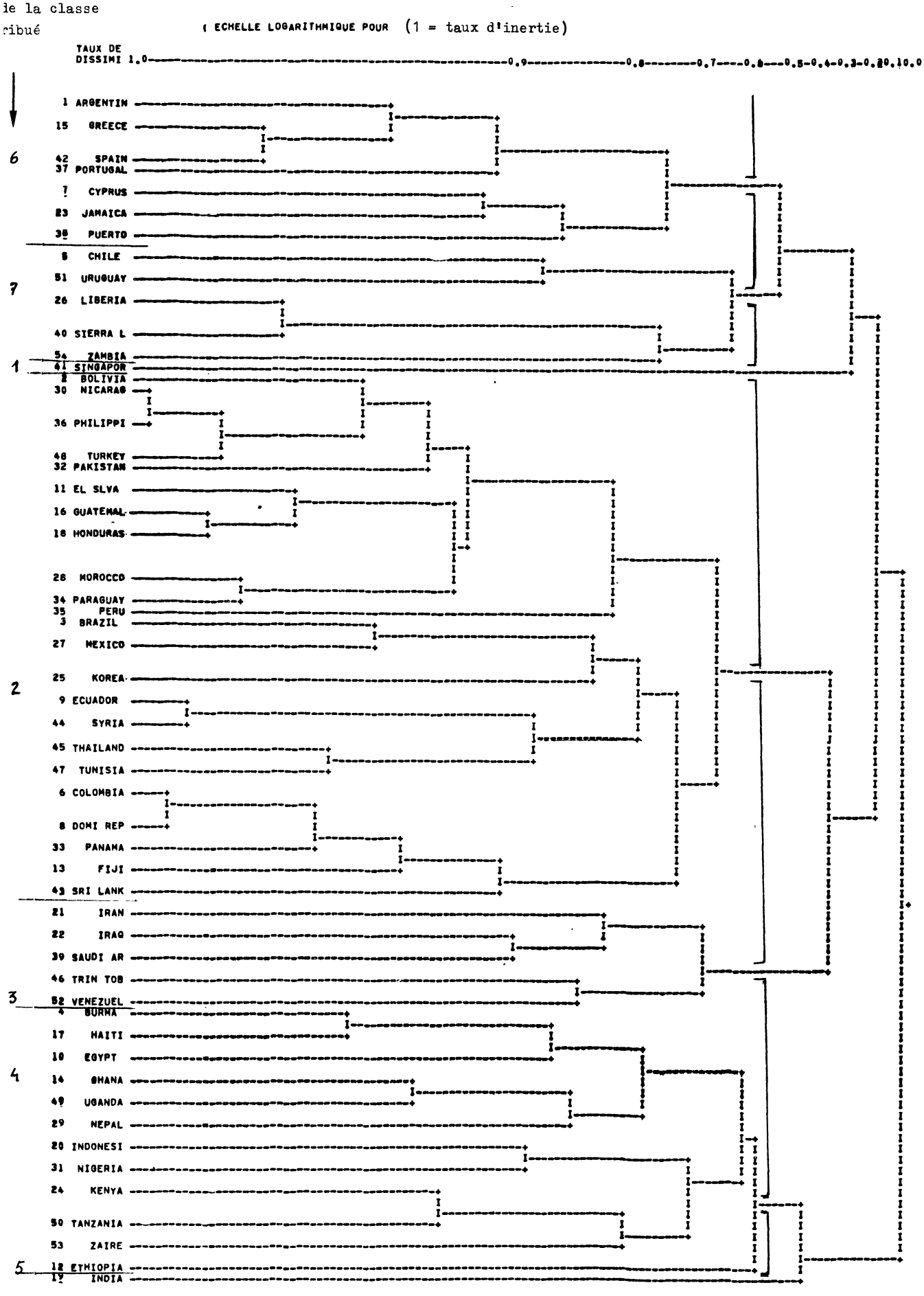
1. Population totale
2. Densité
3. Taux de production agricole
4. Taux de production industrielle
5. Taux brut d'investissement
6. Part des produits primaires dans les exportations
7. Espérance de vie
8. Nombre d'habitants par médecin
9. Taux de croissance démographique
10. Taux de croissance de la consommation
11. Taux de croissance de la formation brute de capital fixe
12. Indice de fluctuation des exportations

2.3. La classification automatique des acteurs

Les dissimilarités sont mises en évidence entre les pays Acteurs décrits par

FIGURE J : ARBRE DE CLASSIFICATION HIERARCHIQUE DES PAYS ACTEURS

le la classe
ribué



Partition optimale

12 descripteurs co-classifiants précédemment sélectionnés.

La construction de l'indicateur f(k) est rappelée dans la 1ère partie, f(k) atteint un maximum relatif pour k = 7.

Aussi, c'est la partition en 7 classes qui est retenue pour élaborer la typologie des pays étudiés.

Les 3 premières règles d'interprétation de la partition : Evaluation de l'excentricité des classes, de leur positionnement dans leur environnement, de leur dissimilarité interne.

La construction des tableaux suivants est exposée dans la première partie.

Le tableau K sur la mesure de l'excentricité des classes et leur positionnement fait apparaître la classe 2 comme centrale dans l'ensemble des 7 classes retenues. En effet, ce groupe rassemblant près de la moitié des pays est le plus proche de la moyenne générale et le plus homogène comme le montre sa faible dissimilarité interne moyenne.

Tableau K. Indicateurs pour les règles d'interprétation des classes (extraits)

CLASSE	EFFECTIF	VALEUR DES INDICATEURS		
		EXCENTRICITE	SEPARATION MOYENNE DES AUTRES CLASSES	DISSIMILARITE INTERNE MOYENNE
1	1	2,758	2,312	0
2	23	0,759	2,147	0,286
3	5	2,454	2,130	o,646

Tableau L. Matrice des séparations entre les classes (extraits)
(On pourrait même dresser un tableau des dissimilarités entre les classes)

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6	Classe 7
Classe 1	0	2,445	1,389	3,443	2,707	1,699	2,190
Classe 2	2,445	0	2,112	3,112	1,144	2,689	1,382
Classe 3	1,389	2,112	0	4,065	1,586	2,142	1,483

Vers une Métarègle

a) Premières identifications de descripteurs caractéristiques :

Dans le tableau M sur les dissimilarités inter-classes par descripteur, les composantes du point moyen sont toutes nulles car les variables ont été centrées. D'autre part, les taux de dissimilarité inter-classes sont égaux aux dissimilarités inter-classes car les variables ayant été réduites et la distance de l'inertie utilisée dans la C.A.H., la dissimilarité totale imputable à chaque variable, est égale à 1.

Le tableau M fait apparaître les variables densité et population appelées DENSITE et POPULATI comme les plus discriminantes entre les 7 classes de la partition étudiée.

Les $t_j(k)$ sont les taux de dissimilarité inter-classes au niveau de chaque descripteur ; pour la partie variable POPULATI, la partition en 7 classes explique 91,4% de la dissimilarité totale. Aucune perte sensible de dissimilarité inter-classes sur POPULATI n'est supportée par passage à la partition entre 6 classes. Par contre, le passage à la partition en 5 classes fait perdre environ 80% de la dissimilarité inter-classes imputable à POPULATI, ensuite les pertes ultérieures sont négligeables.

Le passage de la partition en 5 classes à la partition en 6 classes correspond dans l'arbre hiérarchique à l'isolement de l'Inde. Il est donc clair que POPULATIdiscrimine entre l'Inde et les 6 autres classes considérées globalement et non entre les 7 classes les unes par rapport aux autres. Par contre, PIBOP75 ou AGRI/PIB discriminent bien mieux entre les classes les unes par rapport aux autres.

TABLEAU M : DISSIMILARITES INTER CLASSES
PAR DESCRIPTEUR

SUPERFIC	0,094
COPUAPIB	0,152
CRPRMANU	0,177
HABHOPIT	0,220
MANU/X	0,229
CRPRINDU	0,274
MANU/PIB	0,326
CRPRAGRI	0,328
CROTSSAX	0,333
CONS/PIB	0,354
CROTSSAM	0,358
ENSETSUP	0,360
ANALPHAB	0,393
SCOLAPIS	0,400
EPCE/PIB	0,418

b) La règle d'identification des descripteurs les plus discriminants entre deux classes ou globalement

Dans le tableau N dont la construction est exposée dans la première partie, les contributions des différents descripteurs à l'excentricité des classes de pays sont classées par ordre de valeurs croissantes. Ainsi, par exemple, DENSITE, M/PIB, X/PIB assurent respectivement 32,7%, 15,9% et 12,2% de l'excentricité de Singapour (classe 1).

La classe 2, la plus centrale, est également la plus homogène avec une dissimilarité moyenne de 0,286. Le tableau O désigne les variables qui contribuent le moins à la dissimilarité interne de cette classe, donc le plus à son homogénéité ; ce sont le nombre d'habitants par médecin, la densité, la population totale, le produit par tête en 1975. Par contre, les pays de la classe 2, demeurent assez dissemblables quant à la superficie du territoire (11,7% de la dissimilarité interne), la part des produits manufacturés dans les exportations (7,1% de la dissimilarité interne) etc.

Le tableau non représenté ici de la contribution des variables aux dissimilarités entre les groupes de pays-Acteurs pris deux à deux montrerait la part considérable (61,5%) de la population dans la dissimilarité entre classes 2 et 5 (Inde) et l'influence comparativement négligeable des autres variables, etc.

TABLEAU N : EXCENTRICITE -- DISPERSION DES CLASSES

ET CONTRIBUTIONS DES DESCRIPTEURS A L'EXCENTRICITE

(début)

CLAS 1 (7)		CLAS 2 (7)		CLAS 3 (7)	
EXCENTRI	2,758	EXCENTRI	0,759	EXCENTRI	2.454	EXCENTRI
DISPERST	0.0	DISPERSI	0.286	DISPERST	0.646	DISPERSI
INDU/PIB	0,000	AGRI/PIB	0,000	SCOLARTS	0.000	POPULATT
CONS/PIB	0,000	SUPERFIC	0,000	MANU/PIB	0.000	SUPERFIC
COPUBPIB	0,000	URBANISA	0,000	FBCF/PIB	0.000	FLUCTUAX
FLUCTUAX	0,000	MANU/X	0,000	ESPERVIE	0.000	CRPRAGRI
CRPRAGRI	0,000	ENSEISUP	0,000	ENSEISUP	0.000	DENSITE
ENSEISUP	0,001	AGRIPOPA	0,001	CRPRAGRI	0,001	COPUBPIB
POPULATT	0,001	FBCF/PIB	0,002	DENSITE	0.001	CROIPOPU
SCOLARTS	0,002	POPULATT	0,005	SUPERFIC	0.001	PRIPRO/X
ANALPHAB	0,002	HABHOPTT	0,006	POPULATT	0,002	M/PIB
HABHOPTT	0,002	CROISSAX	0,007	CROIPOPU	0.002	X/PIB
HARMEDEC	0,002	DENSITE	0,008	HABHOPTT	0,002	CONS/PIB
SUPERFIC	0,002	INDUSPOP	0.008	M/PIB	0.002	CRPRMANU
CRCONSOM	0,002	SCOLARTS	0,011	CRPRMANU	0.004	CRPRINDU
MANU/X	0,003	ESPERVIE	0,011	ANALPHAB	0.004	CROIFBCF
MANU/PIB	0,004	CROISSAM	0,014	HABMEDEC	0,007	MANU/X
PRIPRO/X	0,004	M/PIB	0,014	URBANISA	0,009	CROISSAM
CROIPOPU	0,005	FLUCTUAX	0,015	CROIFBCF	0,010	CROISPIB
PIBPOP60	0,007	PIBPOP60	0,016	INDUSPOP	0.013	CROISSAX
CRPRINDU	0,012	MANU/PIB	0,019	AGRIPOPA	0,013	FBCF/PIB
INDUSPOP	0,014	INDU/PIB	0,021	MANU/X	0,014	ENSEISUP
CROISPIB	0,015	CRPIBPOP	0,023	CRPRINDU	0,016	CRPIBPOP
ESPERVIE	0,015	COPUBPIB	0,026	X/PIB	0,032	HABHOPIT
AGRI/PIB	0,017	CONS/PIB	0,030	CRCONSOM	0,032	PIBPOP75
CRPRMANU	0,018	X/PIB	0,032	COPUBPIB	0,035	INDU/PIB
CRPIBPOP	0,019	CRPRMANU	0,034	AGRI/PIB	0,038	INDUSPOP
CROIFBCF	0,019	ANALPHAB	0,034	PIBPOP60	0,041	MANU/PIB
PIBPOP75	0,023	PIBPOP75	0,035	CROISSAX	0,041	PIBPOP60
CROISSAM	0,033	PRIPRO/X	0,036	CROISSAM	0,044	ANALPHAB
AGRIPOPA	0,033	CRPRINDU	0,038	CRPIBPOP	0,062	SCOLARTS
CROISSAX	0,034	HARMEDEC	0,051	CROISPIB	0,062	URBANISA
FBCF/PIB	0,047	CROISPIB	0,069	PRIPRO/X	0,076	CRCONSOM
URBANISA	0,057	CROIPOPU	0,102	PIBPOP75	0,091	AGRIPOPA
X/PIB	0,122	CROIFBCF	0,108	CONS/PIB	0,095	ESPERVIE
M/PIB	0,159	CRPRAGRI	0,109	INDU/PIB	0,108	HARMEDEC
DENSITE	0,327	CRCONSOM	0,116	FLUCTUAX	0,142	AGRI/PIB

TABLEAU 0 : CONTRIBUTIONS DES DESCRIPTEURS

AUX DISSIMILARITES INTERNES DES CLASSES

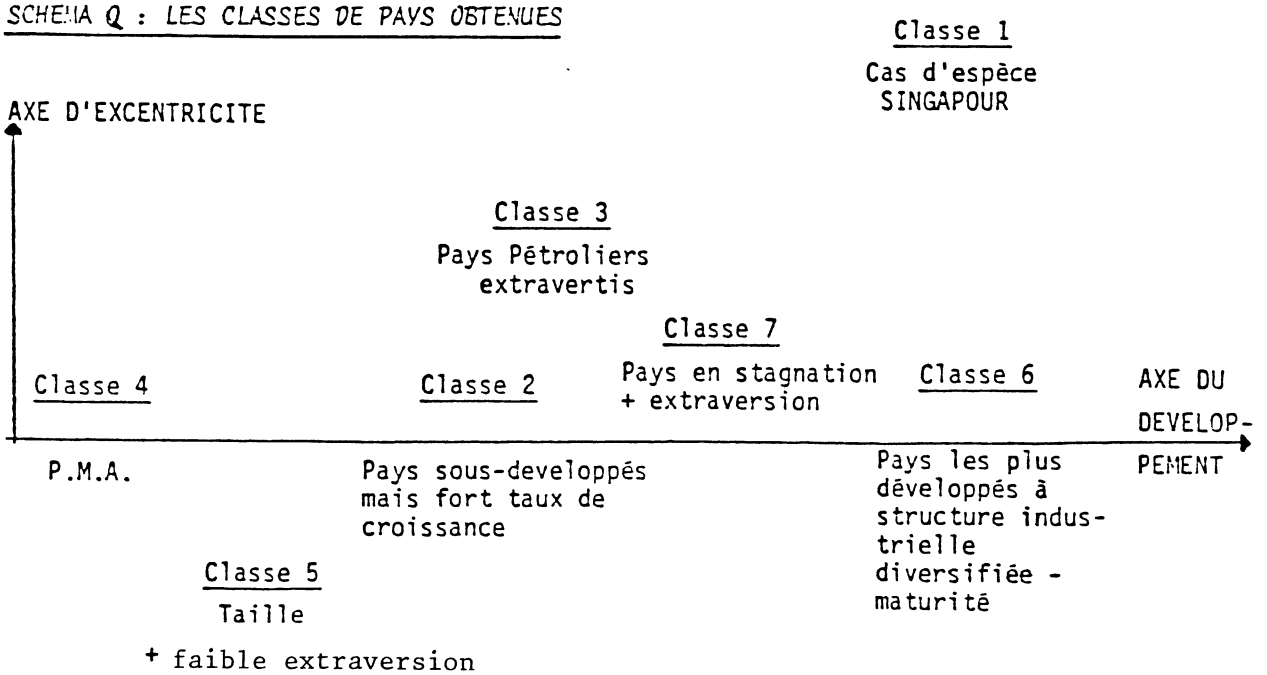
(début)

CLAS 1 (7)		CLAS 2 (7)		CLAS 5 (7)	
DISSIMIL	0.0	DISSIMIL	6.294	DISSIMIL	0.0
POPULATI	0.0	HABMEDEC	0.004	POPULATI	0.0
SUPERFIC	0.0	DENSITE	0.004	SUPERFIC	0.0
DENSITE	0.0	POPULATI	0.006	DENSITE	0.0
AGRI/PIB	0.0	PIBPOP75	0.008	AGRI/PIB	0.0
INDU/PIB	0.0	CONS/PIB	0.010	INDU/PIB	0.0
MANU/PIB	0.0	PIBPOP60	0.011	MANU/PIB	0.0
COPUBPIB	0.0	HABHOPIT	0.012	COPUBPIB	0.0
CONS/PIB	0.0	AGRI/PIB	0.014	CONS/PIB	0.0
FBCF/PIB	0.0	AGRIPOPA	0.015	FBCF/PIB	0.0
X/PIB	0.0	X/PIB	0.018	X/PIB	0.0
M/PIB	0.0	INDU/PIB	0.019	M/PIB	0.0
MANU/X	0.0	CROIPOPU	0.020	MANU/X	0.0
PRIPRO/X	0.0	ESPERVIE	0.020	PRIPRO/X	0.0
AGRIPOPA	0.0	CROISPIB	0.021	AGRIPOPA	0.0
INDUSPOP	0.0	CROIFBCF	0.023	INDUSPOP	0.0
URBANISA	0.0	M/PIB	0.023	URBANISA	0.0
PIBPOP60	0.0	CRCONSOM	0.024	PIBPOP60	0.0
PIBPOP75	0.0	ENSEISUP	0.025	PIBPOP75	0.0
ESPERVIE	0.0	URBANISA	0.025	ESPERVIE	0.0
ANALPHAB	0.0	CRPIBPOP	0.026	ANALPHAB	0.0
SCOLARIS	0.0	INDUSPOP	0.028	SCOLARIS	0.0
ENSEISUP	0.0	CRPRAGRI	0.030	ENSEISUP	0.0
HABHOPIT	0.0	COPUBPIB	0.031	HABHOPIT	0.0
HABMEDEC	0.0	ANALPHAB	0.033	HABMEDEC	0.0
CROIPOPU	0.0	MANU/PIB	0.034	CROIPOPU	0.0
CROISPIB	0.0	FLUCTUAX	0.035	CROISPIB	0.0
CRPIBPOP	0.0	CROISSAM	0.036	CRPIBPOP	0.0
CRPRAGRI	0.0	PRIPRO/X	0.036	CRPRAGRI	0.0
CRPRINDU	0.0	CRPRINDU	0.037	CRPRINDU	0.0
CRPRMANU	0.0	SCOLARIS	0.039	CRPRMANU	0.0
CRCONSOM	0.0	FBCF/PIB	0.040	CRCONSOM	0.0
CROIFBCF	0.0	CRPRMANU	0.047	CROIFBCF	0.0
CROISSAX	0.0	CROISSAX	0.056	CROISSAX	0.0
CROISSAM	0.0	MANU/X	0.071	CROISSAM	0.0
FLUCTUAX	0.0	SUPERFIC	0.117	FLUCTUAX	0.0

TABLEAU P : (extrait : SEPARATION PAR RAPPORT A LA CLASSE 2)
CONTRIBUTIONS DES DESCRIPTEURS AUX SEPARATIONS ENTRE LES CLASSES PRISES DEUX A DEUX

	CLAS 1 (7)		CLAS 2 (7)
SEPARATI	2.445	SEPARATI	0.0
SUPERFIC	-0,011	POPULATI	0,0
MANU/X	-0,005	SUPERFIC	0,0
FLUCTUAX	-0,004	DENSITE	0,0
SCOLARIS	-0,004	AGRI/PIB	0,0
ANALPHAB	-0,004	INDU/PIB	0,0
COPUBPIB	-0,004	MANU/PIB	0,0
CRCONSOM	-0,003	COPUBPIB	0,0
PRIPRO/X	-0,003	CONS/PIB	0,0
ENSEISUP	-0,002	FBCF/PIB	0,0
MANU/PIB	-0,002	X/PIB	0,0
INDU/PIB	-0,002	M/PIB	0,0
POPULATI	-0,000	MANU/X	0,0
HABHOPIT	-0,000	PRIPRO/X	0,0
HABMEDEC	-0,000	AGRIPOPA	0,0
CONS/PIB	-0,000	INDUSPOP	0,0
CRPRAGRI	0,001	URBANISA	0,0
CRPRINDU	0,004	PIBPOP60	0,0
CROISPIB	0,006	PIBPOP75	0,0
CROIFBCF	0,007	ESPERVIE	0,0
CRPRMANU	0,008	ANALPHAB	0,0
PIBPOP60	0,010	SCOLARIS	0,0
ESPERVIE	0,011	ENSEISUP	0,0
CROIPOPU	0,011	HABHOPIT	0,0
CRPIBPOP	0,012	HABMEDEC	0,0
INDUSPOP	0,015	CROIPOPU	0,0
AGRI/PIB	0,017	CROISPIB	0,0
CROISSAM	0,026	CRPIBPOP	0,0
CROISSAX	0,027	CRPRAGRI	0,0
PIBPOP75	0,032	CRPRINDU	0,0
AGRIPOPA	0,036	CRPRMANU	0,0
FBCF/PIB	0,050	CRCONSOM	0,0
URBANISA	0,060	CROIFBCF	0,0
X/PIB	0,150	CROISSAX	0,0
M/PIB	0,187	CROISSAM	0,0
DENSITE	0,374	FLUCTUAX	0,0

SCHEMA Q : LES CLASSES DE PAYS OBTENUES



Dans le tableau P, la colonne correspondant à la classe 2 (par rapport à elle-même) est alors nulle. Dans la colonne correspondant à la classe 1 figure tout d'abord la séparation par rapport à la classe 2.

La classification algorithmique utilisée a eu pour résultat de faire apparaître sept classes, bien distinctes les unes des autres, présentant chacune un degré d'homogénéité satisfaisant ; rappelons que ces deux caractéristiques sont issues d'un calcul basé sur les distances euclidiennes de chaque observation dans un espace $\mathbb{R}^{35}$ et non plus du découpage de l'espace géographique "Tiers-Monde" à partir de critères choisis à priori : il s'agit d'une classification à posteriori.

Faisons l'effort d'imaginer un schéma comportant deux axes (schéma Q) :

- Un axe horizontal orienté de la gauche vers la droite indiquant la marche vers le développement économique (à gauche les pays les moins développés, à droite ceux qui le sont le plus) : il pourrait avoir pour unité le P.I.B./tête en 1975 ;

- Un axe vertical, n'ayant aucune unité, et servant seulement à figurer l'excentricité de la classe par rapport au "noyau de l'échantillon".

Les sept classes obtenues pourraient alors être placées sur le graphique de manière à faire apparaître leur trait dominant dans la course au développement (nous placerons les pays à forte extraversion au-dessus de l'axe horizontal et ceux à faible extraversion au-dessous de cet axe).

3. EPILOGUE

Soit à étudier le modèle inférentiel expliquant l'accroissement relatif de la production intérieure brute Y par celui de la production industrielle X

$$Y = \alpha X + \beta + U$$

avec les hypothèses classiques du modèle linéaire de prévision,

Y variable supposée aléatoire dont on a un vecteur d'observations
X variable exogène dont on a un vecteur d'observations de même dimension
U variable aléatoire résiduelle, normale, centrée, de variance σ^2, sans

autocorrélation d'une valeur à une autre, sans corrélation avec X

Pour l'ensemble des 54 pays (classes 1 à 7), on a le modèle estimé suivant :

$Y = 0,46\ X + 2,10 \qquad r = 0,77$

$(0,05) \quad (0,40) \qquad (0,08)$

(les écarts-types de coefficients sont indiqués entre parenthèses)

Les tests de Student et Fisher conduisent à un rejet de l'hypothèse d'indépendance entre Y et X.

On peut affiner, grâce à l'étude des classes de pays déjà obtenue, encore que dans cette optique, il eût peut-être mieux valu rechercher éventuellement alors des nuages de points allongés comme autour d'une droite de tendance (méthodes de l'arbre de longueur minimale ou de percolation).

Grâce à la classification automatique des pays acteurs, grâce au principe qui permet de choisir le sous-ensemble de descripteurs le plus pertinent, grâce aux indicateurs d'aide au choix du nombre de classes et aux règles d'aide à leur interprétation, on a pu dégager ici 7 classes de pays cohérentes ; on obtient pour la classe 4, celle des pays les moins développés, le modèle estimé

$Y = 0,15\ X + 2,96 \qquad r = 0,52$

$(0,07) \quad (0,53) \qquad (0,27)$

D'après les tests de Student et Fisher, on ne peut rejeter l'hypothèse d'indépendance entre Y et X.

Dans ces pays assez sous-développés, c'est sûrement l'accroissement relatif de production agricole qui explique le mieux les variations de la P.I.B. plutôt que l'accroissement de la production industrielle. Le modèle estimé n'est plus fiable pour la prévision.

La classification préalablement à la statistique inférentielle a permis dans une phase exploratoire d'affiner la phase confirmatoire ; le "bon choix" des descripteurs a permis d'obtenir des sous-ensembles cohérents de pays acteurs.

BIBLIOGRAPHIE

AIVAZIAN, Les méthodes statistiques d'étude de dépendances entre variables classifiantes, T.S.E.M.I., Moscou, 1976.

ANDERBERT M.R., Cluster Analysis for applications, New York, Academic Press, 1973 (J.1).

ARNAL C., Une méthode d'analyse typologique : quelques résultats et application à l'économie des pays de l'O.C.D.E., thèse de doctorat de 3ème cycle du Groupe de Recherche en Analyse de systèmes et calcul économique, Université de Droit, d'Economie et des Sciences d'Aix-Marseille, Faculté d'Economie Appliquée (B.2).

BATTINI A., Rythmes de croissance : Analyse internationale (1950-1975), une critique de la thèse de l'inégalité croissante, thèse de Doctorat d'Etat es Sciences Economiques, Université de Droit, d'Economie et des Sciences d'Aix-Marseille, Faculté d'Economie Appliquée (B.3).

BAKER F.B., "Stability of two hierarchical grouping techniques case 1 : sensitivity to Data Errors", Journal of the American Statistical Association, n°69 p.440-445, 1974 (J.2).

BENZECRI J.P. et collaborateurs, L'analyse des données, Paris, DUNOD, 1973 (B.10).

BERTIER P., BOUROCHE J.M., Analyse des données multidimensionnelles, Paris, Presses Universitaires de France, 1975, (B.11).

CAILLEZ F., PAGES J.P. sous la direction de MORLAT G., "Introduction à l'analyse de données", Société de mathématiques appliquées et de sciences humaines, Paris, 1976 (B.4).

CALINSKI and HARABASZ, "A dendrite method for cluster analysis", Communications in statistics, 3(1), 1-27, 1974 (J.4).

CHANDON J.L. et PINSON S., Analyse typologique, théories et applications, Paris, Masson, 1980, (B.17).

CHENERY H., "Pattern of Development", World Bank Research Publication, Oxford University Press, 1975 (B.5).

C.N.U.C.E.D. (Conférence des Nations Unies sur le Commerce et le Développement), Manuel de statistiques du commerce international et du développement, Nations Unies, New York, 1979 (B.6).

FOWLKES E.B. and MALLOWS C.L., "A method for comparing two hierarchical Clustering", J.A.S.A. n°78 p. 553-584, 1983 (J.3).

GEFFRAULT J.P.,Discrimination de classes et détermination d'ensembles minimaux de mesures pour la classification automatique de formes, application à des données en Archéologie, thèse U.E.R. de mathématiques et informatique, Université de Rennes I, 1982 (B.15).

HARTIGAN J.A., Clustering Algorithms, New York, John Wiley, 1975 (B.10').

JAMBU M., Classification automatique pour l'analyse des données, Paris, DUNOD, 1978 (B.12).

KHUN T.S., La structure des révolutions scientifiques, Paris, Flammarion, 1983 (B.7).

LEMOIGNE J.L., Théorie du système général ; la théorie de la modélisation,Paris, Presses Universitaires de France, 1983 (B.8).

LEONTIEFF W., CARTER A.P., PETRI P., 1999, l'expertise de Wassily Léontieff, Paris, collection DEMAIN, DUNOD, 1977, (B.1.).

LERMAN I.C., Classification et analyse ordinale des données, Paris, DUNOD, 1981 (B.13).

LERMAN I.C. et Collaborateurs, Programme d'analyse des résultats d'une classification automatique, Institut de recherche en informatique et systèmes aléatoires, Rennes, 1982, (B.14).

LERMAN I.C., M. HARDOUIN, T. CHANTREL, Analyse de la situation relative entre deux classifications floues, Laboratoire de statistique, IRISA, Université de Rennes I (B.16).

MORIN E., La nature de la nature, Paris, Edition du Seuil, 1977 (B.9).

MORIN E., La vie de la vie, Paris, Edition du Seuil, 1980 (B.9').

POPPER R., La quête inachevée, Paris, Calmann Lévy, 1981.