

REVUE DE STATISTIQUE APPLIQUÉE

I. C. LERMAN

B. TALLUR

Classification des éléments constitutifs d'une juxtaposition de tableaux de contingence

Revue de statistique appliquée, tome 28, n° 3 (1980), p. 5-28

http://www.numdam.org/item?id=RSA_1980__28_3_5_0

© Société française de statistique, 1980, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

CLASSIFICATION DES ÉLÉMENTS CONSTITUTIFS D'UNE JUXTAPOSITION DE TABLEAUX DE CONTINGENCE

I. C. LERMAN et B. TALLUR

La suite des principaux paragraphes est :

- I – Introduction ; position du problème.
- II – Rappel sur les indices de proximité entre variables descriptives.
- III – Structure de proximité prise en compte par A.V.L.
- IV – Représentation géométrique d'un tableau de contingence.
- V – Indice de proximité entre lignes ou colonnes d'un ou d'une juxtaposition de tableaux de contingences.
- VI – Arbre des classifications directement associé à l'indice.
- VII – Suite donnée à l'indice dans un processus d'analyse par la classification de ce type de données.
- VIII – Application, structure de l'agriculture régionale française.
- IX – Conclusion.

I. INTRODUCTION ; POSITION DU PROBLEME

Le point de référence est le cas classique d'un tableau de données X de dimension $n \times m$, indexé par $E \times V$, définissant la description d'un ensemble E de n individus ou objets par un ensemble V de m variables *numériques*. Le point visé est le *tableau de contingence* donnant les cardinaux des classes du croisement de deux partitions ; puis par extension, le tableau résultant d'une juxtaposition de sous-tableaux qui sont des tables de contingence. Le point de vue développé est celui de la classification des lignes ou colonnes d'un tel tableau. On se servira d'une représentation mathématique, ici géométrique, adéquate du tableau des données.

Nous considérons le tableau de données X comme ayant un caractère essentiellement *dissymétrique*. D'un point de vue appréhension des données, c'est flagrant ; en effet, une même variable descriptive peut être opérationnalisée par un appareil de mesure et à la limite, il y a autant de différence entre une variable et un individu qu'entre un appareil de mesure et l'objet sur lequel est effectuée la mesure. On peut également le voir au niveau de la représentation mathématique qu'on propose en général pour ce tableau en vue d'une analyse factorielle en composantes (cf. [14] pour plus de détails).

Au contraire, le tableau de contingence de dimension $n \times m$ que nous notons ici

$$C = \{k_{ij}/(i, j) \in I \times J\}; \quad (1)$$

a de façon évidente un caractère parfaitement symétrique.

Il est entendu que dans le tableau (1) ci-dessus k_{ij} représente le nombre d'individus ayant la modalité i de la première variable-partition et la modalité j de la seconde variable-partition.

Relativement au tableau X des données, Individus \times Variables, les taxinomistes sont davantage polarisés par la classification de l'ensemble des individus que par celle de l'ensemble des variables de description. Pour cette raison, compte tenu de notre perception de l'espace géométrique, c'est davantage le langage des *distances* qui a été utilisé pour l'expression des critères de réduction de l'information. Si par contre, il s'agit d'organiser par proximité l'ensemble des variables descriptives, c'est le langage de la *corrélation* qui s'impose d'abord, bien qu'on puisse à partir du schéma de dualité définir une distance entre variables.

De toute façon, on peut distinguer deux familles de méthodes de classification ; celle basée sur la notion de *distance* et celle basée, sur la notion de *corrélation*. Dans la première, nous accordons une place de choix au critère de l'inertie (i.e. variance) expliquée ; celui-là même sous-jacent aux méthodes d'analyse factorielle.

Les méthodes de classification hiérarchique de I.C. LERMAN, basées sur la vraisemblance des liens rentrent dans le cadre de la deuxième famille. Elles se réfèrent en effet à une notion très générale de la *corrélation* entre structures finies de même type et permettent l'organisation en classes et sous-classes "significatives" de proximité ; d'abord de l'ensemble des variables, quel que soit le type commun de structure défini par ces dernières ; ensuite de l'ensemble des individus ou objets, en tenant étroitement compte de la dissymétrie du tableau de données Individus \times Variables descriptives.

Il restait, dans le cadre de ces méthodes, l'intégration du problème de la classification des lignes ou colonnes d'un tableau de contingence puis, plus généralement, d'un tableau résultant de la juxtaposition de tables de contingence.

Cette structure parfaitement symétrique du tableau des données est l'argument privilégié de l'analyse des correspondances. La représentation euclidienne de ce tableau donnée par cette méthode, a conduit J.P. BENZECRI et M. JAMBU (cf. [1] et [2]) à développer une méthode de classification hiérarchique des lignes (resp. colonnes) d'un tel tableau qui se réfère à la première famille de méthodes, basée sur le critère de l'inertie expliquée, conformément à la métrique du χ^2 .

Il y avait donc lieu dans notre cadre de définir la méthode de la vraisemblance des liens, basée sur les "corrélations" entre lignes (resp. colonnes) d'un tel tableau, et à comparer les résultats concrets obtenus par les deux méthodes de classification hiérarchique.

II. RAPPEL SUR LES INDICES DE PROXIMITÉ ENTRE VARIABLES DESCRIPTIVES

II.1. DEMARCHE GÉNÉRALE

Nous représentons une variable descriptive par une partie d'un ensemble ou une pondération sur ce dernier. Nous distinguons deux principaux types d'une variable ; celui pouvant être représenté par une partie de l'ensemble E des individus

ou une pondération sur E ; et celui, dont la représentation est une partie de $E \times E$ ou une pondération sur $E \times E$.

Dans la première catégorie, on peut classer l'attribut descriptif a qui peut être représenté par la partie E_a des individus qui le possèdent et la variable numérique v qui peut être regardée comme définissant une pondération sur E en attachant à chaque individu x , le nombre $v(x)$, valeur de la mesure de la variable v sur l'individu x .

Toute variable qualitative discrète qui définit une relation binaire sur E est susceptible d'une représentation dans $E \times E$, par son graphe par exemple. De sorte que dans la deuxième catégorie on peut classer

– La variable "rang" qui définit un ordre total 0 sur E que nous représentons par son graphe

$$R(0) = \{x, y\} \in E \times E / x < y \text{ pour } 0 \} \subset E \times E ; \quad (1)$$

– Le caractère descriptif à l'ensemble totalement ordonné des modalités (i.e. variable qualitative ordinale) qui définit un préordre total w sur E que nous représentons par la partie $R(w)$ de $E \times E$ définie par

$$R(w) = \sum_{i < j} E_i \times E_j \text{ (somme ensembliste) ;} \quad (2)$$

où E_i est la $i^{\text{ème}}$ classe du préordre formée des individus possédant la $i^{\text{ème}}$ modalité du caractère.

– Le caractère descriptif à l'ensemble sans structure des modalités (i.e. variable qualitative nominale) qui définit une partition π sur E que nous représentons dans l'ensemble plus réduit $F = P_2(E)$ des parties à deux éléments de E par

$$R(\pi) = \sum_i P_2(E_i) \text{ (somme ensembliste) ;} \quad (3)$$

où $P_2(E_i)$ est l'ensemble des paires réunies dans la $i^{\text{ème}}$ classe.

– Signalons enfin de façon très générale la variable "pondération" sur $E \times E$ qui peut être représentée par une matrice carrée

$$\{\mu_{xy} / (x, y) \in E \times E\}; \quad (4)$$

où μ_{xy} est la pondération affectée au couple (x, y) , $(x, y) \in E \times E$.

La démarche générale pour l'évaluation de l'association entre deux variables de même type (i.e. définissant le même type de structure sur E) est la suivante :

Soit (a, b) un couple de telles variables et soit (α, β) le couple de structures définies sur E par (a, b) ; ainsi par exemple, (α, β) est un couple de parties de E si (a, b) est un couple d'attributs ; (α, β) est un couple de préordres totaux sur E si (a, b) est un couple de variables qualitatives ordinales. A α (resp. β) nous associons l'ensemble A (resp. B) des structures sur E de même type et ayant les mêmes caractéristiques cardinales que α (resp. β). Ainsi par exemple si α est un préordre total sur E de composition $u = (n_1, n_2, \dots, n_k)$ ($n = n_1 + n_2 + \dots + n_k$) ; A peut être défini comme l'ensemble de tous les préordres totaux sur E de même composition u.

Conformément à ci-dessus, nous représentons (α, β) par un couple de parties $(R(\alpha), R(\beta))$, éventuellement pondérées, de E s'il s'agit de variables de la première catégorie et de $E \times E$ s'il s'agit de variables de la deuxième.

On introduit l'indice "brut" de proximité

$$s = \text{card} (R(\alpha) \cap R(\beta)) ; \quad (5)$$

auquel nous associons les deux variables aléatoires duales

$$S_\alpha = \text{card} (R(\alpha) \cap R(\beta')) \text{ et } S_\beta = \text{card} (R(\alpha') \cap R(\beta))$$

où α' (resp. β') est un élément aléatoire dans l'ensemble A (resp. B) muni d'une probabilité uniformément répartie. On démontre de façon synthétique, pour les différents cas ci-dessus envisagés, que la distribution de la v.a. S_α est la même que celle, S_β ; d'autre part, l'expression formelle des moments d'une telle distribution permet de montrer, sous des conditions assez générales, son caractère asymptotiquement normal (cf. [7], [8], [9], [10]). En désignant par $\mu_{\alpha\beta}$ et $\sigma_{\alpha\beta}^2$ la moyenne et la variance de cette distribution, nous appelons indice "centré réduit"

$$Q(\alpha, \beta) = (s - \mu_{\alpha\beta}) / \sigma_{\alpha\beta} ; \quad (6)$$

C'est cet indice qui prend le sens d'une corrélation généralisée, on s'en rendra compte ci-dessous (§ II.2.) dans le cas de la comparaison d'un couple de variables numériques.

N désignant l'hypothèse d'absence de lien ci-dessus exprimée, l'indice définitif que nous considérons et qui se réfère à une échelle de probabilité, s'écrit

$$P(\alpha, \beta) = \text{Pr}^N \{S < s\} \quad (7)$$

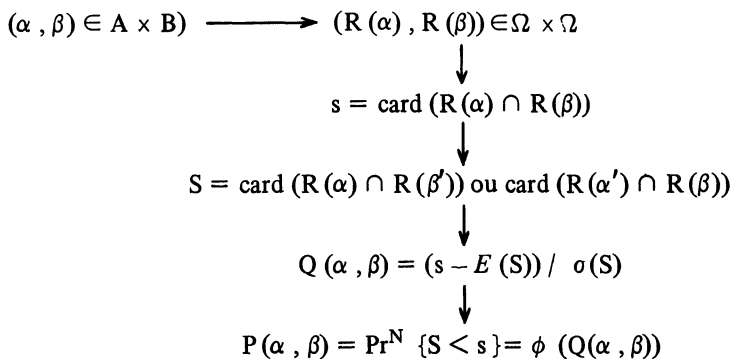
où S est l'une des deux v.a. duales de même loi S_α et S_β . En d'autres termes, les deux variables descriptives a et b ont un degré de ressemblance d'autant plus grand que la valeur de s est invraisemblablement grande, relativement à l'hypothèse d'absence de lien N.

Le passage de la formule (6) à celle (7) se fait, avec une bonne approximation, au moyen de la relation

$$P(\alpha, \beta) = \phi(Q(\alpha, \beta)) ; \quad (8)$$

où ϕ est la fonction de répartition de la loi normale (0, 1).

La démarche générale peut être résumée par le schéma suivant :



Ce schéma intervient dans les méthodes de I.C. LERMAN, à chaque fois qu'il y a lieu de définir un critère de synthèse pour la formation des classes ou la reconnaissance de leur "signification".

II.2. CAS PARTICULIER DE LA COMPARAISON DE DEUX VARIABLES NUMERIQUES

(a, b) est un couple de variables numériques ; ayant codé E par l'ensemble $I = \{1, 2, \dots, i, \dots, n\}$ on est conduit à prendre comme indice brut de proximité

$$s = \sum \alpha_i \beta_i ; \quad (9)$$

où α_i (resp. β_i) est le nombre affecté à l'individu i par la première (resp. seconde) variable.

Dans le cas où la donnée est un couple de pondérations sur E (ou d'ailleurs sur $E \times E$), l'hypothèse d'absence de lien prend nécessairement une forme permutationnelle.

Les deux v.a. duales associées à (9) sont ici

$$S_\alpha = \sum_{i \in I} \alpha_i \beta_{\sigma(i)} \quad \text{et} \quad S_\beta = \sum_{i \in I} \alpha_{\sigma(i)} \beta_i ; \quad (10)$$

où σ est un élément aléatoire dans l'ensemble G_n des $n!$ permutations sur I. On sait que l'étude du comportement asymptotique d'une telle distribution est l'objet du célèbre théorème de Wald, Wolfowitz et Noether ; repris par Hajek.

La moyenne et la variance communes sont respectivement :

$$\mu_{\alpha\beta} = n \mu_\alpha \mu_\beta \quad \text{et} \quad \sigma_{\alpha\beta}^2 = \frac{n^2}{(n-1)} \sigma_\alpha^2 \sigma_\beta^2 ; \quad (11)$$

où μ_α et σ_α^2 (resp. μ_β et σ_β^2) sont la moyenne et la variance de la distribution empirique de a (resp. b) sur E ; c'est-à-dire $\{\alpha_i / i \in I\}$ (resp. $\{\beta_i / i \in I\}$).

L'indice "centré réduit" se met sous la forme

$$Q(a, b) = \sqrt{n-1} \rho(a, b) ; \quad (12)$$

où $\rho(a, b)$ est précisément le coefficient de corrélation entre les deux variables a et b.

Le caractère asymptotiquement normal de la distribution commune de S_α et de S_β conduit à l'indice qui se réfère à une échelle de probabilité sous la forme

$$P(a, b) = \phi((s - \mu_{\alpha\beta}) / \sigma_{\alpha\beta}) ; \quad (13)$$

III. STRUCTURE DE PROXIMITE PRISE EN COMPTE PAR A.V.L. (ALGORITHME DE LA VRAISEMBLANCE DES LIENS)

Dans la pratique, pour des raisons liées à la variance de la représentation des unités de données dans les Sciences Humaines, on centre et on réduit globalement

le tableau symétrique suivant des indices "centrés réduits" $Q(a, b)$ (cf formule (6) § II.1.) entre éléments de l'ensemble V à classier :

$$\{Q(a, b)/(a, b) \in V \times V\}; \quad (1)$$

et ce, avant la référence à une échelle de probabilité. En d'autres termes, on substitue au tableau (1), le tableau

$$\{Q'(a, b)/(a, b) \in V \times V\}; \quad (2)$$

avec

$$(\forall \{a, b\} \in P_2(V)), Q'(a, b) = (Q(a, b) - \bar{Q})/\sigma_Q; \quad (3)$$

où \bar{Q} et σ_Q^2 sont la moyenne et la variance de la distribution de Q sur l'ensemble $P_2(V)$ des paires (i.e. parties à deux éléments) de V .

Il nous suffit ici de supposer que l'ensemble V est celui des variables de description et que l'indice Q n'est, à un coefficient multiplicatif près, autre que le coefficient de corrélation, lorsque V est formé de variables numériques.

La structure de proximité prise en compte dans l'Algorithme de la Vraisemblance des Liens est définie par la table des nombres

$$\{P(a, b)/\{a, b\} \in P_2(V)\}; \quad (4)$$

avec

$$(\forall \{a, b\} \in P_2(V), P(a, b) = \phi(Q'(a, b)); \quad (5)$$

L'arbre des classifications associé à cet algorithme est construit de façon ascendante en démarrant de la partition la plus fine et en réunissant à chaque pas la paire ou les paires de classes les plus "proches" au sens de l'indice de proximité que nous allons ci-dessous exprimer. Pour obtenir un tel arbre, il y a donc lieu d'étendre la notion de proximité entre deux éléments à celle, entre deux parties disjointes C et D de V .

On part de l'indice de base qui joue le rôle de l'indice "brut" :

$$p(C, D) = \max \{P(c, d)/(c, d) \in C \times D\}; \quad (6)$$

où $P(c, d)$ a été défini en (5).

L'indice final qu'on retiendra résulte de la distribution de la v.a. $p(C', D')$ associées à (6), où C' et D' sont respectivement associés à C et D dans le cadre d'une hypothèse d'absence de lien tenant compte des caractéristiques de cardinalité de C et de D . L'indice auquel on aboutit prend la forme suivante :

$$P(C, D) = (p(C, D))^{l \times m}; \quad (7)$$

où $l = \text{card}(C)$ et $m = \text{card}(D)$.

Le passage de l'indice (6) à celui (7) a constitué un progrès décisif pour l'apparition des classes de faible cohésion qui n'en ont pas moins une certaine cohérence interne ; les éléments de telles classes se trouvaient avec l'indice (6) éparpillés en raison de l'attraction de quelques noyaux de forte cohésion.

F. Nicolău (thèse en préparation à la Faculté des Science de Lisbonne) a imaginé d'autres indices de même type : vraisemblance d'un lien ; mais où le lien est mesuré par une autre fonction que le maximum de l'ensemble des valeurs

$$\{P(c, d)/(c, d) \in C \times D\}; \quad (8)$$

par exemple à partir de la loi de la v.a., associée dans l'hypothèse d'absence de lien à

$$\frac{1}{lm} \sum \{P(c, d)/(c, d) \in C \times D\}; \quad (9)$$

Toutefois, les premiers résultats comparatifs obtenus par la classification hiérarchique montre que c'est toujours l'indice (20) qui donne les résultats les plus raffinés et les plus cohérents dans leurs nuances.

IV. REPRESENTATION GEOMETRIQUE D'UN TABLEAU DE CONTINGENCE

Reprenons le tableau C (cf. (7) § I) de contingence

$$C = \{k_{ij}/(i, j) \in I \times J\}; \quad (1)$$

représentant les cardinaux des classes du croisement de deux partitions ; la première à $n = \text{card}(I)$ classes et la deuxième à $m = \text{card}(J)$ classes.

Pour représenter géométriquement ce tableau de données, on fait jouer à l'un des deux ensembles I ou J, le rôle de l'ensemble des points-objets et à l'autre, le rôle de l'ensemble des variables. Si, pour fixer les idées, c'est I qui joue le rôle de l'ensemble des individus ou objets et J celui des variables, la représentation la plus courante de I à travers J est fournie dans le cadre de l'Analyse des Correspondances. En notant, pour tout $(i, j) \in I \times J$,

$$\left. \begin{aligned} k_{i.} &= \sum \{k_{ij}/j \in J\}, k_{.j} = \sum \{k_{ij}/i \in I\} \\ k_{..} &= \sum \{k_{ij}/(i, j) \in I \times J\} \\ f_{ij} &= k_{ij}/k_{..}, p_{i.} = k_{i.}/k_{..}, p_{.j} = k_{.j}/k_{..} \\ f_j^i &= (f_{ij}^i/j \in J) \text{ où } f_j^i = f_{ij}/p_{i.} : \text{ "profil de } i \text{ à travers } J" ; \end{aligned} \right\} \quad (2)$$

On associe à I le nuage $N(I)$ suivant dans R^m muni de la métrique diagonale $(1/p_{.j}/j \in J)$:

$$N(I) = \{(f_j^i, p_{i.})/i \in I\}; \quad (3)$$

où $p_{i.}$ est la masse affectée au point f_j^i , pour tout $i \in I$.

Considérons pour fixer les idées, le problème de la classification de J. Il importe de noter ici une *différence essentielle* entre notre méthode et celle qui domine la "Classification Ascendante Hiérarchique" de J.P. BENZECRI et M. JAMBU pour ce type de données. Cette méthode se conçoit au niveau de R^n où J est représenté par un ensemble de points-objets ou individus, elle utilise les concepts de *distance au carré* et *d'inertie expliquée* conformément à la métrique du χ^2 . Alors que notre méthode se conçoit au niveau de $(R^m)^*$ et utilise le concept de la *corrélation*, conformément à la représentation géométrique ci-dessus définie.

V. INDICE DE PROXIMITE ENTRE COLONNES OU LIGNES D'UN OU D'UNE JUXTAPOSITION DE TABLEAUX DE CONTINGENCES

V.1. Cas d'un seul tableau

Comme nous venons de l'exprimer, soit, pour fixer les idées, le problème de la classification de J , J devant jouer le rôle de variables descriptives, considérons la représentation de I à travers J et prenons celle la plus commune, définie par le nuage $N(I)$ (cf. (3) § IV) où la mesure de la variable j sur l'individu i est définie par f_j^i (cf. 2) § (IV), $1 \leq j \leq m$, $1 \leq i \leq n$.

L'indice de proximité entre les éléments j et h de J , $1 \leq j, h \leq m$, n'est autre que le coefficient de corrélation $\rho(j, h)$ dont nous allons préciser l'expression.

On a

$$\begin{aligned} \text{moy}(j) &= \sum_{i \in I} p_i (f_j^i) = p_{.j} \\ \text{var}(j) &= \sum_{i \in I} p_i (f_j^i - p_{.j})^2 = \sum_{i \in I} \left(\frac{f_{ij}^2}{p_i} \right) - p_{.j}^2 \end{aligned} \quad (1)$$

où moy désigne moyenne et var variance.

D'autre part,

$$\begin{aligned} \text{cov}(j, h) &= \sum_{i \in I} p_i (f_j^i - p_{.j}) (f_h^i - p_{.h}) \\ &= \sum_{i \in I} \left(\frac{f_{ij} f_{ih}}{p_i} \right) - p_{.j} p_{.h} \end{aligned} \quad (2)$$

où cov désigne la covariance.

On obtient finalement

$$\begin{aligned} \rho(j, h) = \text{cor.}(j, h) &= \text{cov.}(j, h) / \sqrt{\text{var.}(j) \text{var.}(h)} \\ &= \left\{ \sum_{i \in I} (f_{ij} f_{ih} / p_i) - p_{.j} p_{.h} \right\} / \left\{ \left[\sum_{i \in I} (f_{ij}^2 / p_i) - p_{.j}^2 \right] \times \right. \\ &\quad \left. \left[\sum_{i \in I} (f_{ih}^2 / p_i) - p_{.h}^2 \right] \right\}^{1/2} \end{aligned} \quad (3)$$

Cet indice dépend exactement de la représentation euclidienne de I à travers J , il est invariant par rapport à la métrique dont peut être muni \mathbf{R}^m pour l'évaluation des distances entre éléments de I .

Pour obtenir directement l'indice de proximité entre éléments de I , il y a de façon analogue lieu de considérer la représentation euclidienne de J à travers I au moyen du nuage dual $N(J)$.

On peut, à partir de (3), écrire l'expression de cet indice en intervertissant les rôles de I et de J ; soit

$$\rho(i, i') = \left\{ \sum_{j \in J} (f_{ij} f_{i'j} / p_{.j}) - p_{i.} p_{i'.} \right\} / \left\{ \left[\sum_{j \in J} (f_{ij}^2 / p_{.j}) - p_{i.}^2 \right] \times \left[\sum_{j \in J} (f_{i'j}^2 / p_{.j}) - p_{i'.}^2 \right] \right\}^{1/2} \quad (3')$$

$1 \leq i, i' \leq n.$

Les valeurs de l'indice (3) (resp. (3')) entre éléments de J (resp. de I) joueront le même rôle que celles de l'indice Q(a, b), conformément à la table (1) du paragraphe III. On applique à ces indices la même transformation (3) (§ III) pour aboutir à la structure de proximité (5) (§ III) prise en compte par A.V.L.

V.2. Structure d'une juxtaposition de tableaux de contingences

Nous distinguons deux principaux types de cette structure :

a) Le tableau global des données est indexé par un ensemble de la forme

$$(I^{(1)}, I^{(2)}, \dots, I^{(k)}, \dots, I^{(K)}) \times (J^{(1)}, J^{(2)}, \dots, J^{(l)}, \dots, J^{(L)}) \quad (4)$$

où le sous-tableau indexé par $I^{(k)} \times J^{(l)}$ est un tableau de contingence

$$(1 \leq k \leq K, 1 \leq l \leq L)$$

et où l'ensemble des variables dont les modalités indexent les lignes est *disjoint* de l'ensemble des variables dont les modalités indexent les colonnes).

b) Le tableau général des données est indexé par un ensemble de la forme

$$(J^{(1)}, J^{(2)}, \dots, J^{(l)}, \dots, J^{(L)}) \times (J^{(1)}, J^{(2)}, \dots, J^{(l')}, \dots, J^{(L)}) \quad (5)$$

où le sous-tableau indexé par $J^{(1)} \times J^{(l')}$ est une vraie table de contingence si $l' \neq l$ et se réduit à un tableau carré $J^{(l)} \times J^{(l)}$ où seulement la diagonale n'est pas chargée de zéros si $l' = l$. La suite des valeurs de cette diagonale est la suite des nombres $k(j^{(l)})$ d'individus possédant la $j^{(l)}$ -ème modalité du caractère ($j^{(l)} \in J^{(l)}$). Il s'agit de ce que l'on appelle communément le tableau de "Burt" ; généralement associé au croisement par lui-même de l'ensemble des questions d'un questionnaire.

Désignons par $J = \cup \{J^{(l)} / 1 \leq l \leq L\}$; le tableau carré $J \times J$, pris globalement est symétrique quant à son contenu :

$$(\forall (j, j') \in J \times J), k(j, j') = k(j', j) \quad (6)$$

un même sous-tableau de contingence se retrouvant deux fois aux blocs

$$J^{(l)} \times J^{(l')} \text{ et } J^{(l')} \times J^{(l)}.$$

Nous considérons cette structure du tableau des données comme plus éloignée, comparativement à a) ci-dessus, d'une structure de tableau de contingence ; cela d'ailleurs apparaît clairement, au niveau de la définition d'un indice de proximité entre lignes ou colonnes.

V.3. Extension de l'indice dans le cas d'une juxtaposition de tableaux de contingences

Nous adoptons pour l'extension de l'indice un principe de parcimonie qui consiste à prendre à chaque fois la solution la plus locale mettant en œuvre le plus petit nombre de tables de contingences.

Supposons, sans restreindre la généralité et pour fixer les idées, qu'il s'agit de définir un indice de proximité entre deux colonnes j et j' . On considère alors la plus "petite" juxtaposition de tables de contingences, dont la structure est celle de a) ci-dessus, et qui contient les deux colonnes j et j' ; dans une telle juxtaposition il y a au plus deux rangées verticales de tables de contingences. A chacune des tables $I^{(k)} \times J^{(l)}$ de cette dernière juxtaposition, on commence par associer la table des fréquences :

$$f_{I^{(k)} \times J^{(l)}} = \{f_{ij}/(i, j) I^{(k)} \times J^{(l)}\} \quad (7)$$

à laquelle on substitue la table

$$f'_{I^{(k)} \times J^{(l)}} = \{f'_{ij} = (f_{ij}/t)/(i, j) I^{(k)} \times J^{(l)}\} \quad (8)$$

où t est le nombre de tableaux de contingences de la dernière juxtaposition.

De la sorte on se ramène à une pondération positive de masse totale 1 sur un produit de deux ensembles de la forme $I' \times J'$ qui définit un sous-tableau du tableau total, correspondant à la juxtaposition retenue. C'est au niveau de ce dernier tableau de nombres que l'indice $\rho(j, j')$ est calculé conformément à la formule (3) ci-dessus. Pour plus de détails on se référera à [14] où les formules sont développées pour chacun des cas. Nous nous limiterons ici à considérer ci-dessous (Figure 1) le schéma concerné par le calcul que nous allons préciser dans le cas particulier d'une juxtaposition selon une direction, horizontale, de tables de contingences; ce qui correspond à $K = 1$. Le tableau global se trouve alors indexé par un ensemble de la forme

$$I \times (J^{(1)}, J^{(2)}, \dots, J^{(l)}, \dots, J^{(L)}) \quad (9)$$

où chacun de ses sous tableaux $I \times J^{(l)}$ est de contingence; $1 \leq l \leq L$.

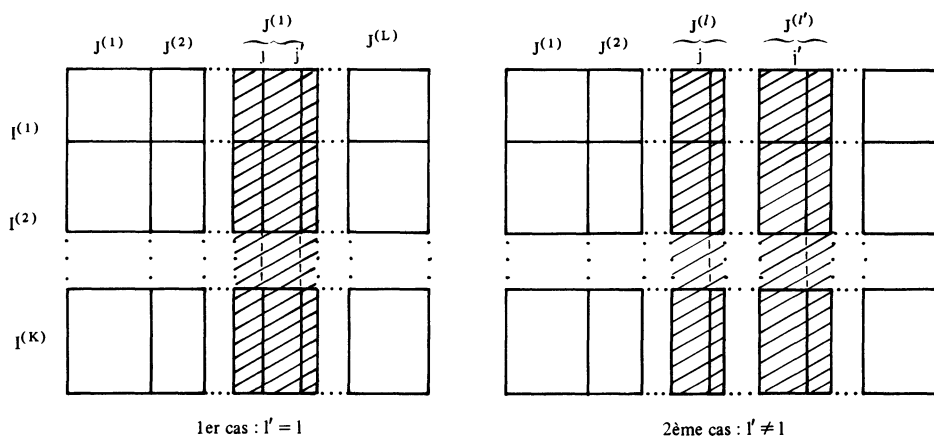


Figure 1

Soit le problème de l'extension de l'indice (3) ci-dessus entre éléments de $J = \cup \{J^{(l)} / 1 \leq l \leq L\}$; considérons pour cela un couple (j, j') d'éléments distincts de J .

Dans le premier cas où il existe l tel que $(j, j') \in J^{(l)} \times J^{(l)}$, on se ramène exactement à la situation V.1. précédente où il y a lieu de considérer l'indice (3) entre j et j' relativement au tableau de contingence $I \times J^{(l)}$.

Dans le deuxième cas où $(j, j') \in J^{(l)} \times J^{(l')}$ avec $l' \neq l$, on considèrera la définition de l'indice par rapport au sous-tableau indexé par $I \times (J^{(l)}, J^{(l')})$ qui est défini par une juxtaposition horizontale d'exactly deux tables de contingence. En désignant par

$$\{f_{ij}^{(l)} / (i, j^{(l)}) \in I \times J^{(l)}\} \text{ et } \{f_{ij}^{(l')} / (i, j^{(l')}) \in I \times J^{(l')}\} \quad (10)$$

les deux matrices de fréquences, respectivement associées aux deux tableaux de contingence $I \times J^{(l)}$ et $I \times J^{(l')}$, on définit le tableau de pondérations, de masse totale 1

$$\{f'_{ij} / (i, j) \in I \times (J^{(l)} \cup J^{(l')})\} \quad (11)$$

en posant

$$f'_{ij} = \frac{1}{2} f_{ij}, \text{ pour tout } (i, j) \in I \times J^{(l)} \cup I \times J^{(l')} \quad (12)$$

De la sorte, relativement à l'ensemble du tableau $I \times J'$ où $J = J^{(l)} \cup J^{(l')}$, on a

$$p'_{.j} = \frac{1}{2} p_{.j}, \text{ pour tout } j \in J^{(l)} \cup J^{(l')}$$

et

$$p'_{i.} = \frac{1}{2} (p_{i.}^{(l)} + p_{i.}^{(l')}), \text{ pour tout } i \in I \quad (13)$$

où $p_{i.}^{(l)}$ (resp. $p_{i.}^{(l')}$) est la masse de i définie au niveau du tableau $I \times J^{(l)}$ (resp. $I \times J^{(l')}$).

Dans ces conditions, on appliquera l'indice (3) du paragraphe V précédent à partir des nombres du tableau (11) entre j de $J^{(l)}$ et j' de $J^{(l')}$. En utilisant les éléments primaires définis au niveau de chacun des tableaux $I \times J^{(l)}$ et $I \times J^{(l')}$, cet indice peut se mettre sous la forme :

$$\frac{\left\{ \sum_{i \in I} (f_{ij} f_{ij'} / p'_{i.}) - p_{.j} p_{.j'} \right\}}{\left\{ \left[\sum_{i \in I} (f_{ij}^2 / p'_{i.}) - p_{.j}^2 \right] \left[\sum_{i \in I} (f_{ij'}^2 / p'_{i.}) - p_{.j'}^2 \right] \right\}^{1/2}} \quad (14)$$

Il importe ici de noter une différence par rapport à la nature de la structure prise en compte dans l'Analyse Factorielle des Correspondances Multiple (cf. [3]). Alors que dans cette méthode on prend directement en compte l'ensemble du tableau résultant de la juxtaposition de toutes les tables de la forme $f_{I^{(k)} \times J^{(l)}}$ (cf. (7) ci-dessus) pour lui substituer un seul tableau définissant une

pondération positive de masse totale 1, obtenue en divisant le contenu de chacune des cases par le nombre de tables de contingences ; il nous suffit quant à nous d'opérer le plus localement possible, de façon qui nous paraît précise et ce, compte tenu de la structure de proximité prise en compte par A.V.L. (cf. § III).

V.4. Cas d'un tableau de "Burt"

Bien que la démarche du paragraphe précédent permette le calcul de la valeur de l'indice de proximité pour chaque couple de lignes ou colonnes d'un tel tableau, dans notre approche de l'analyse par la classification d'un questionnaire, nous nous refusons d'assimiler le tableau de Burt à un tableau de contingence. Dans cette approche, pour l'organisation des variables, il y a deux étapes correspondant à deux niveaux d'analyse différents ; le premier niveau étant le plus global. Chaque question définissant une variable-partition sur l'ensemble E des individus où la *l*-ème a card ($J^{(l)}$) modalités ; la première étape consiste dans la classification de l'ensemble des questions conformément à l'indice général de proximité appliqué au cas des variables partitions (cf. § II.1.). Cette étape permet la réorganisation du questionnaire en chapitres et sous-chapitres correspondant respectivement à des thèmes et sous-thèmes principaux du comportement de la population étudiée à travers l'échantillon défini par l'ensemble E.

La deuxième étape consiste à associer à chaque modalité d'une même question un attribut ; formant de la sorte un ensemble A de card (J) = card ($J^{(1)}$) + . . . + card ($J^{(L)}$) attributs descriptifs ; les attributs associés aux modalités d'une même question étant mutuellement exclusifs et complémentaires. Le tableau de Burt apparaît alors exactement comme le tableau des indices "bruts" s de proximité entre attributs (cf. (5) § II.1.). La classification hiérarchique de A conformément à l'Algorithme de la Vraisemblance des Liens (cf. § III) permet la découverte des profils et sous profils de comportement ; lesquels permettent de "comprendre" la formation des thèmes apparus à travers les classes de questions de l'analyse globale.

Signalons à titre indicatif que deux autres étapes complètent notre analyse pour ce type de données. D'abord une classification hiérarchique de l'ensemble E des individus décrits à travers l'ensemble A des modalités – attributs où dans l'élaboration de l'indice de proximité par rapport à une hypothèse d'absence de lien, on tient compte de la structure formelle et statistique de A (cf. [10]).

La dernière étape consiste en le croisement de couples de classifications où le premier argument d'un couple donné est une classification sur E correspondante à un niveau significatif (repéré à partir d'un indice de signification basé sur l'ordonnance (cf. [7], [8])) de l'arbre des classifications sur E et où le second argument du couple est une classification significative dans le même sens, de l'ensemble A des attributs (cf. [12], [13]).

Il se pose maintenant la question d'analyser la comparaison entre les deux démarches que nous avons esquissées pour établir l'indice de proximité sur l'ensemble des modalités-attributs indexant les lignes (resp. colonnes) d'un tableau de Burt ; celle qu'on vient d'exprimer et celle du paragraphe V.3 ci-dessus.

V.5. Cas d'une famille de distributions sur une suite d'ensembles finis

$(J^{(1)}, \dots, J^{(l)}, \dots, J^{(L)})$ est une suite d'ensembles finis et chaque i de I définit la suite de distributions de probabilités

$$(\phi_j^i(I) / 1 \leq l \leq L) = (\{\phi_j^i / j \in J^{(l)}\} / 1 \leq l \leq L) \quad (15)$$

où pour chaque i , $\phi_j^i(I)$ définit une distribution de probabilité sur $J^{(l)}$.

Pour étendre l'indice qu'on vient de définir (cf. § V.1 & V.3 ci-dessus) on commencera par associer à chaque tableau

$$\phi_J^i(I) = \{\phi_{ij}^i / (i, j) \in I \times J^{(l)}\} \quad (16)$$

à card (I) lignes et card $(J^{(l)})$ colonnes, le tableau des nombres

$$\phi_{I \times J^{(l)}} = \{\phi_{ij}^{(l)} / (i, j) \in I \times J^{(l)}\}$$

où

$$\phi_{ij}^{(l)} = \frac{1}{n} \phi^{(l)} \frac{i}{j} \text{ pour tout } (i, j) \in I \times J^{(l)} \quad (17)$$

La suite des tableaux $\phi_{I \times J^{(l)}}(1)$ est alors assimilée à une juxtaposition de tableaux de fréquences à laquelle on applique le principe exprimé au paragraphe V.3 précédent.

VI. ARBRE DES CLASSIFICATIONS DIRECTEMENT ASSOCIE A L'INDICE

VI.1. Rappel pour le cas d'un critère basé sur les carrés des distances

La structure des données qui nous intéresse ici est celle qu'on peut assimiler à un tableau de contingence. En d'autres termes, la donnée de base que nous considérons ici est celle

$$f_{I \times J} = \{f_{ij} / (i, j) \in I \times J\} \quad (1)$$

définie par une distribution de probabilités sur $I \times J$.

Il s'agit du tableau de fréquences si la donnée est un tableau C de contingence (cf. (7) § I) ; du tableau des nombres

$$f^{(k,l)} / KL \text{ pour tout } (i, j) \in I^{(k)} \times J^{(l)}, 1 \leq k \leq K, 1 \leq l \leq L,$$

s'il s'agit d'une juxtaposition selon les deux directions de $K \times L$ tableaux de contingences. C'est en d'autres termes la représentation mathématique adoptée dans l'Analyse Factorielle des Correspondances Multiple mais où ici, nous n'intégrons pas le cas du tableau de Burt.

Nous avons déjà fait allusion aux paragraphes I et IV à la méthode de formation de l'arbre des classifications, basée sur l'inertie expliquée conformément à la métrique du χ^2 . Considérons pour fixer les idées la classification de J . Cette méthode ne peut se concevoir qu'au niveau de la représentation euclidienne du nuage $N(J)$.

Lorsqu'on utilise la représentation euclidienne du nuage $N(J)$ dans \mathcal{R}^n muni de la métrique diagonale définie par l'inverse des poids p_i , une même classe J_r se trouve représentée par son centre de gravité

$$g_i^r = \frac{1}{p(J_r)} \sum_{j \in J_r} p \cdot j f_i^j = (f_i^{J_r}, \dots, f_i^{J_r}, \dots, f_n^{J_r})$$

où nous avons noté

$$p(J_r) = \sum_{j \in J_r} p_j$$

$$f_i^{J_r} = \frac{f(i, J_r)}{p(J_r)} = \frac{\sum \{f_{ij} / j \in J_r\}}{p(J_r)} \quad (2)$$

A partir d'un niveau donné de l'arbre définissant la partition

$$\{J_1, J_2, \dots, J_t\}$$

on passe au niveau suivant en réunissant les paires $\{J_r, J_s\}$ de classes pour lesquelles la variation du critère de l'inertie est minimum ; cette variation s'exprime par la formule

$$\frac{p(J_r) p(J_s)}{p(J_r) + p(J_s)} \|g_i^r - g_i^s\|^2 \quad (3)$$

D'une étape à l'autre, il y a lieu de réactualiser les centres de gravité des classes ; c'est-à-dire de calculer les centres de gravité des nouvelles classes $J_r \cup J_s$ formées ; au moyen de la formule

$$g_i^{rvs} = \frac{p(J_r) g_i^r + p(J_s) g_i^s}{p(J_r) + p(J_s)} \quad (4)$$

et d'affecter au centre de gravité g_i^{rvs} de la nouvelle classe, le poids $(p(J_r) + p(J_s))$.

Ainsi, cette façon de former l'arbre des classifications "oublie" complètement le tableau $f_{I \times J}$ dès qu'on a "installé" le nuage $N(J)$ dans \mathcal{R}^n muni de la métrique diagonale $\{1/p_i / 1 \leq i \leq n\}$.

Une deuxième façon de concevoir la formation de l'arbre se situe au niveau plus élémentaire de la table $f_{I \times J}$ et ne nécessite que l'expression de la distance du χ^2 entre deux colonnes j et j' d'un même tableau $f_{I \times J}$; à savoir

$$d^2(j, j') = \sum_{i \in I} \frac{1}{p_i} (f_i^j - f_i^{j'})^2 = \sum_{i \in I} p_i \left(\frac{f_{ij}}{p_i \cdot p_j} - \frac{f_{ij'}}{p_i \cdot p_{j'}} \right)^2 \quad (5)$$

En effet, la première étape de cette forme de l'algorithme consiste à déterminer la paire, ou les paires de colonnes s'il y en a plusieurs, pour lesquelles la valeur de l'indice (5) est minimale. Si $\{j, j'\}$ est une telle paire, on fusionne les colonnes j et j' en une seule que nous noterons $j \vee j'$ (représentant la classe $\{j, j'\}$) et nous poserons

$$f_{ij \vee j'} = f_{ij} + f_{ij'} \quad (6)$$

$$f_{j \vee j'} = p_j + p_{j'}$$

Le nouveau tableau des nombres $f_{i\cdot}$ a toujours son ensemble des lignes indexé par I alors que son ensemble de colonnes se trouve indexé par l'ensemble

des classes de J en présence dont le nombre est ici $\text{card}(J)$ moins le nombre de classes formées. La somme des pondérations f_{ir} est toujours égale à 1 ; d'ailleurs il est important de noter que si le tableau initial $f_{I \times J}$ est associé à un tableau de contingence ; c'est-à-dire le tableau des fréquences des classes du croisement de deux partitions sur un ensemble E d'objets ; *le nouveau tableau est également de même nature*, associé au tableau de contingence de croisement de deux partitions dont la première (aux classes indexées par I) est identique à celle correspondante précédente et dont la seconde a ses classes indexées par les classes en présence de J.

Le passage d'un niveau quelconque de l'arbre des classifications au niveau suivant n'est pas plus général dans son articulation. Si $\{J_1, J_2, \dots, J_t\}$ est la partition de J correspondante à un niveau donné de l'arbre ; désignons par

$$\{f_{ir} / i \in I, 1 \leq r \leq t\} \quad (7)$$

le tableau des fréquences des classes du croisement de deux partitions sur E, où les classes de la première sont indexées par I et où les classes de la seconde partition sont respectivement indexées par J_1, J_2, \dots, J_t . On a, en revenant aux partitions initiales,

$$f_{ir} = f_{iJ_r} = \sum_{j \in J_r} f_{ij}$$

$$p_{.r} = p_{.J_r} = \sum_{j \in J_r} p_{.j} \quad (8)$$

On se ramène ainsi à une situation analogue à celle de la première étape, mais avec le tableau (7) de dimension $\text{card}(I) \times t$.

On peut interpréter en termes géométriques cette nouvelle manière de former l'arbre des classifications sur J ; en effet, il s'agit à chaque pas de réunir les paires $\{J_r, J_s\}$ de classes pour lesquelles est minimale la distance au carré entre les centres de gravité de J_r et de J_s ; c'est-à-dire avec les notations (2) ci-dessus

$$\|g_I^r - g_I^s\|^2 \quad (9)$$

Alors qu'on notera que la quantité critère (3), qui correspond à la première procédure pour la formation de l'arbre, pondère cette distance au carré entre les centres de gravité par le coefficient $p(J_r)p(J_s) / p(J_r \cup J_s)$.

VI.2. Arbre des classifications directement associé à l'indice

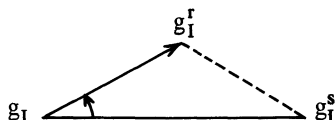
Nous inspirant de la deuxième façon de former l'arbre des classifications dans le cas précédent d'un critère basé sur le carré de la distance ; il est tout à fait naturel de proposer le même algorithme de formation de l'arbre à partir de l'indice de corrélation que nous avons défini (cf. formule (3) § V).

Le passage d'un niveau donné de l'arbre au suivant se fait donc exactement de la même façon que ci-dessus ; à cela près que l'indice de proximité entre colonnes (qui peuvent résulter de regroupements de colonnes initiales) n'est plus (9) ci-dessus, mais celui (3) du paragraphe V précédent :

$$\frac{\sum_{i \in I} (f_{ir} f_{is} / p_{i.}) - p_{.r} p_{.s}}{\{[\sum_{i \in I} (f_{ir}^2 / p_{i.}) - p_{.r}^2][\sum_{i \in I} (f_{is}^2 / p_{i.}) - p_{.s}^2]\}^{1/2}} \quad (10)$$

où nous avons, conformément aux formules (8) ci-dessus, noté f_{ir} pour f_{iJ_r} (resp. f_{is} pour f_{iJ_s}) et $p_{.r}$ pour $p_{.J_r} = p(J_r)$ (resp. $p_{.s}$ pour $p_{.J_s} = p(J_s)$).

D'un point de vue géométrique le critère d'association représente le cosinus pour la métrique du χ^2 de l'angle des deux vecteurs $(g_I^r - g_I)$ et $(g_I^s - g_I)$ où g_I (resp. g_I^r) représente le centre de gravité de la classe J_r (i.e. du sous-nuage $N(J_r)$) (resp. de la classe J_s (i.e. du sous nuage $N(J_s)$)) et g_I est le centre de gravité du nuage total $N(J)$.



VII. SUITE DONNEE A L'INDICE DANS UN PROCESSUS D'ANALYSE PAR LA CLASSIFICATION DE CE TYPE DE DONNEES

Pour simplifier et nous fixer les idées, nous allons, comme au paragraphe VI précédent, considérer que la donnée de base est celle définie par une distribution $f_{I \times J}$ de probabilités sur $I \times J$.

Notre approche consiste à faire suivre l'indice de corrélation établi avec l'Algorithme de la Vraisemblance des Liens ; en d'autres termes, nous référant aux notations du paragraphe III où l'indice sous-jacent à cet algorithme est présenté ; le rôle de l'indice Q sera joué par notre indice de corrélation.

Par conséquent, on peut bâtir un couple d'arbres de classifications le premier sur J et le second sur I .

Notre méthode de classification hiérarchique offre directement une représentation graphique de l'arbre condensé aux niveaux où apparaît un "noeud significatif" (cf. Fig. 2), mais avec toute l'information de l'arbre total puisque chaque association se trouve ponctué par le numéro du niveau de l'arbre détaillé où elle s'est formée. Un "noeud significatif" dont l'interprétation correspond à la complétion le long de la suite des niveaux d'une statistique "locale" des niveaux dont la conception est basée sur l'ordonnance (i.e. ordre total sur l'ensemble des paires de l'ensemble à classier) (cf. [7], [8] chap. 4). Si la statistique "locale" des niveaux juge l'association qui vient de se former ; celle "globale" évalue l'ensemble de la partition produite à un niveau donné. C'est l'examen de la distribution de cette statistique globale le long de la suite des niveaux qui permet de déterminer un "petit" nombre de classifications fortement significatives qui correspondent à des états d'équilibre dans la synthèse. Soit $(\{I_r / 1 \leq r \leq k\}, \{J_s / 1 \leq s \leq h\})$ un couple de telles partitions sur respectivement, I et J .

Considérons l'une des deux classifications, par exemple $\{I_r / 1 \leq r \leq k\}$. On cherchera à contrôler la "signification" de la classification prise globalement ainsi que celle de chacune des classes I_r , $1 \leq r \leq k$; en précisant l'importance du rôle d'un même j de J ou, plus généralement, d'une même classe J_s de la deuxième partition, $1 \leq s \leq h$.

PRODUCTIONS ANIMALES	01 TOUJOURS EN HERBE	>-----*	
		> 10----	*
	08 OVINS	>-----*	I
		> 14	I
	06 CULT.FOURRAGERES	>-----*	I
		> 11----	I
09 PORCS	>-----*	I	
	> 13	I	
07 BOVINS	>-----*	I	
	> 7----	I	
15 20 A 50 HECTARES	>-----*	I	
<hr/>			
SYST. EXTENSIF	03 CULT.MARAICHERS	>-----*	I
		> 8----	I----
	05 PLANTES SARCLES	>-----*	I
		> *12	I
	04 CEREALES	>---	I
		> 1	I
17 > 80 = A 100 H	>---I-----	I	
	> *3	I	
16 50 A 100 HECTARES	>---	I	
<hr/>			
SYST. INTENSIF	02 CULT.PERMENANTES	>-----*	16
		> I	I
	10 INF.A 1 HECTARE	>---*	*6 I
		> 5	I I
	11 1 A 2 HECTARES	>---I---I----	*
		> 2	I
	12 2 A 5 HECTARES	>---*	9
	> I		
13 5 A 10 HECTARES	>---	I	
	> 4----	*	
14 10 A 20 HECTARES	>---	*	

CLASSIFICATION DES VARIABLES 1972 (BOIS ET FORETS EXCLUS)

ALGORITHME: AVL (COR, χ^2)

Figure 2

La mesure de l'importance du rôle de j (resp. J_s) dans la formation des classes I_r , $1 \leq r \leq k$; repose sur la formule de décomposition de l'inertie totale du nuage $N(I)$ relativement à la partition $\{I_r / 1 \leq r \leq k\}$: soit

$$\sum_{i \in I} p_i \cdot \|f_j^i - g_j\|^2 = \sum_{1 \leq r \leq p} \sum_{i \in I_r} p_i \cdot \|f_j^i - g_j^r\|^2 + \sum_{1 \leq r \leq p} p(I_r) \|g_j^r - g_j\|^2 \quad (1)$$

où $p(I_r) = \sum \{p_i / i \in I_r\}$ est le poids de la r -ème classe, $g_j = (p_{.j} / j \in J)$ le centre de gravité du nuage total et où

$$g_j^r = \left\{ \left[\left(\sum_{i \in I_r} f_{ij} \right) / p(I_r) \right] / j \in J \right\}$$

est le centre de gravité de la r -ème classe.

L'inertie retenue ou expliquée par la classification se met sous la forme suivante du rapport de la variance inter-classes sur celle, globale :

$$\sum_{1 \leq r \leq k} p(I_r) \|g_j^r - g_j\|^2 / \sum_{i \in I} p_i \cdot \|f_j^i - g_j\|^2 \quad (2)$$

La discrimination de la classification $\{I_r/1 \leq r \leq k\}$ par j (resp. J_s) est définie comme le rapport des contributions de j (resp. J_s) au numérateur et dénominateur de (2). La discrimination de la seule classe I_r par j (resp. J_s) est également définie comme le rapport des contributions de j (resp. J_s) à

$$p(I_r) \|g_j^r - g_j\|^2 / \sum_{i \in I} p_i \|f_j^i - g_j\|^2 \quad (3)$$

C'est-à-dire

$$C(I_r/j) = \frac{p(I_r) [(f(I_{r,j})/p(I_r)) - p_{.j}]^2}{\sum_{i \in I} p_i [(f_{ij}/p_i) - p_{.j}]^2} \quad (4)$$

où $f(I_{r,j}) = \sum \{f_{ij}/i \in I_r\}$.

Nous ne nous attarderons plus sur l'expression de ces indices de tableaux de contingences, programmés et évalués pour deux corpus de données, dans le cadre de la future thèse de 3^e cycle d'A. Prod'homme.

Ainsi, dans cette approche nous ne lions pas la nature du critère d'“explication” des classes à celui de formation des classes.

Pour mieux fixer les idées, nous allons supposer que $f_{I \times J}$ est la table des fréquences associée à un tableau de contingence défini à partir du croisement d'un couple de partitions sur un ensemble E d'objets ou individus. Dans ce contexte I (resp. J) est l'ensemble des modalités de la première (resp. de la seconde) variable partition.

La donnée d'une partition $\{I_r/1 \leq r \leq k\}$ définit une partition

$$\{E_r/1 \leq r \leq k\}$$

sur l'ensemble E des objets de départ ; la partie E_r étant formée de l'ensemble des individus qui possèdent l'une des modalités i de I_r . Soit dans ces conditions le couple $(\{E_r/1 \leq r \leq k\}, \{F_s/1 \leq s \leq h\})$ de partitions sur E ainsi associé au couple de partitions $(\{I_r/1 \leq r \leq k\}, \{J_s/1 \leq s \leq h\})$ de I et de J .

Une deuxième approche dans l'“explication” de la situation relative d'une des partitions $\{I_r/1 \leq r \leq k\}$ où $\{J_s/1 \leq s \leq h\}$ par rapport à l'autre est la construction de la table des indices

$$\{\chi_{rs} = \sqrt{n} (f_{rs} - p_{r.} p_{.s}) / \sqrt{p_{r.} p_{.s}} / 1 \leq r \leq k, 1 \leq s \leq h\} \quad (5)$$

où

$$f_{rs} = f_{I_r J_s} = \sum \{f_{ij}/(i, j) \in I_r \times J_s\}$$

et où

$$p_{r.} = p(I_r) = \sum \{p_i/i \in I_r\}$$

$$p_{.s} = p(J_s) = \sum \{p_{.j}/j \in J_s\} ;$$

$1 \leq r \leq k, 1 \leq s \leq h$.

Pour (r, s) fixé, l'indice χ_{rs} pris séparément, est dans l'hypothèse d'indépendance entre les deux partitions $\{E_r/1 \leq r \leq k\}$ et $\{F_s/1 \leq s \leq h\}$, une réalisation d'une variable aléatoire normale centrée réduite. χ_{rs} qui représente la contribution

orientée à la statistique du χ^2 du tableau de contingence de ces deux dernières partitions, mesure l' "intensité" du lien entre I_r et J_s ; $1 \leq r \leq k$, $1 \leq s \leq h$.

VIII. APPLICATION. STRUCTURE DE L'AGRICULTURE REGIONALE FRANCAISE

De nombreuses applications de la méthode (couplage de l'indice de corrélation entre lignes (resp. colonnes) d'une juxtaposition de tableaux de contingence et de l'A.V.L., suivis des développements mentionnés au paragraphe précédent) ont été réalisés. Nous avons à chaque fois obtenu les résultats les plus raffinés, notamment par comparaison avec d'autres méthodes.

Nous allons exposer brièvement le premier exemple qui a provoqué cette recherche et nous nous intéresserons à la comparaison des résultats avec la méthode de classification hiérarchique ascendante basée sur le critère de l'inertie expliquée, conformément à la métrique du χ^2 . Pour un exposé complet voir [15].

VIII.1. Bref exposé des résultats

Le support de l'information est une suite de trois tableaux de contingence croisant l'ensemble I de 89 départements français avec

- l'ensemble $J^{(1)}$ des modalités du caractère "structure d'exploitation du sol". Ce caractère a 8 modalités correspondantes à 8 classes de taille de l'exploitation ;

- l'ensemble $J^{(2)}$ des modalités du caractère "mode d'utilisation du sol". Ce caractère a 6 modalités correspondantes aux 6 principales cultures ; et

- l'ensemble $J^{(3)}$ des modalités du caractère "importance du cheptel" qui comprend 3 types d'élevage.

Les données recueillies en 1972 nous ont été fournies par P. Rainelli (maître de recherche à la Station d'Economie Rurale de l'I.N.R.A. de Rennes) qui nous a aidé dans l'interprétation et l'évaluation des résultats.

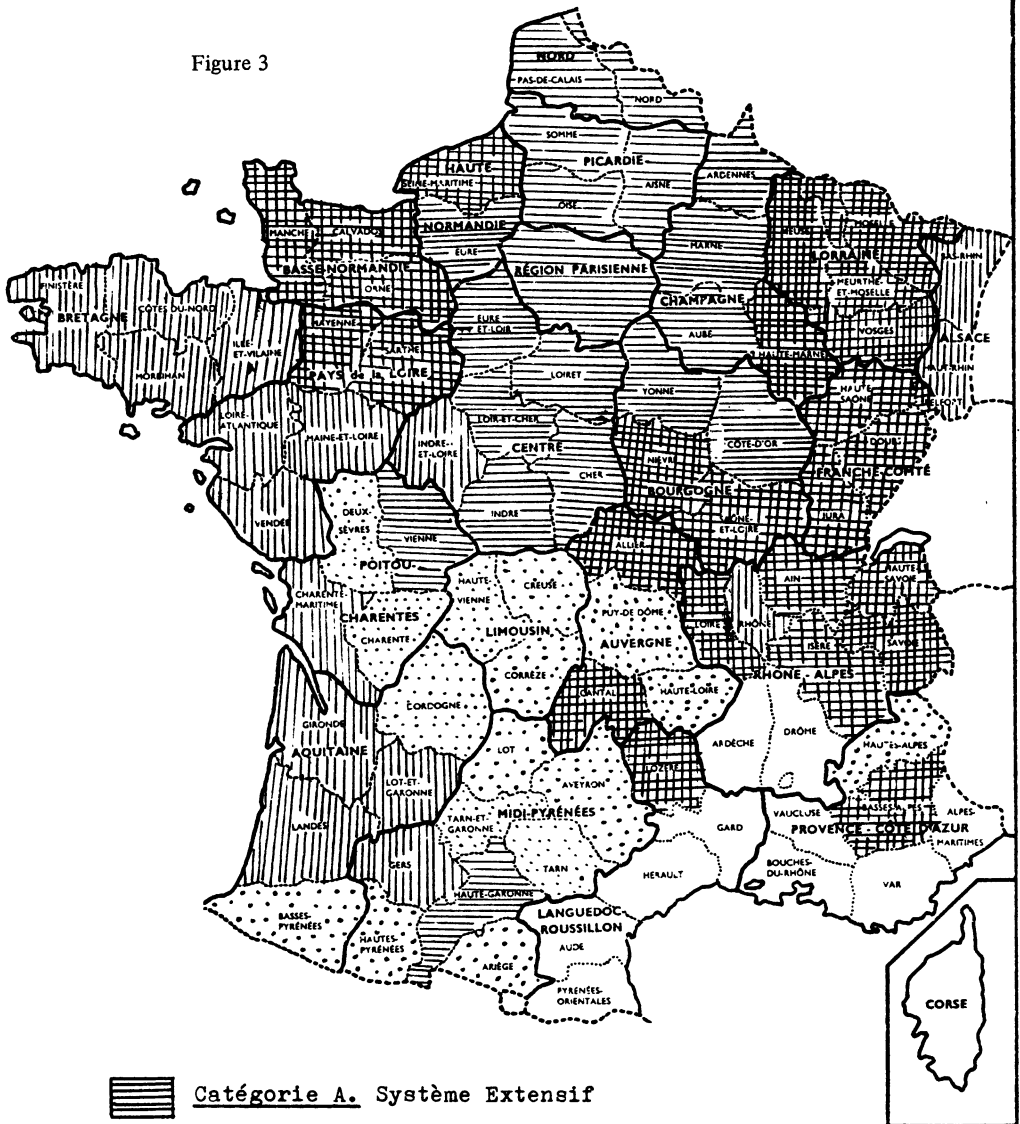
Différentes analyses de ces tableaux de données ont permis de dégager les principaux types d'économie régionale en France. Ainsi, la classification de l'ensemble $J = J^{(1)} \cup J^{(2)} \cup J^{(3)}$ des 17 modalités (cf. Fig. 1), met en évidence l'existence de trois grands systèmes de production agricole :

- Le premier est centré sur les grandes exploitations (de taille supérieure ou égale à 50 ha) associées aux céréales ; ce qui correspond à l'agriculture la plus favorisée.

- Le second est basé sur de petites exploitations où l'on pratique des cultures permanentes (viticulture, arboriculture) ; c'est le système intensif par excellence.

- Enfin, dans le troisième et dernier système, se trouvent associées les productions animales, les cultures fourragères et les surfaces toujours en herbe sur des exploitations de taille moyenne.

Figure 3



Catégorie A. Système Extensif



Catégorie B. Productions Animales, Spéculations Fourragères



Catégorie C. Productions Animales associées à des S.T.H. et aux Céréales



Catégorie D1. Type Extensif avec des ovins; Petites ou moyennes Exploitations



Catégorie D2. Type intensif (Cultures Permanentes) associé avec les ovins

On a d'autre part effectué la classification de l'ensemble I des départements à travers (a) $J^{(1)}$, (b) $J^{(2)}$, (c) $J^{(1)} \cup J^{(2)}$ et enfin (d) $J^{(1)} \cup J^{(2)} \cup J^{(3)}$.

a) Classification des départements selon la structure d'exploitation

Au niveau le plus significatif de l'arbre (niveau 86), on a une partition en trois grandes classes se caractérisant, respectivement, par des exploitations de taille moyenne (par exemple Nord, Pas-de-Calais, Mayenne, Dordogne, Isère, Haute-Savoie, . . .), petites exploitations, d'au plus 5 ha (tels que Gironde, Gard, Hérault, Aude, Pyrénées Orientales, Var, Vaucluse, . . .) et grandes exploitations, d'au moins 50 ha (exemples : Aisne, Cher, Oise, Seine-et-Marne, . . .). La classe des exploitations de taille moyenne comporte quatre sous-classes définissant une échelle de grandeur des exploitations, allant de 5 à 50 ha.

b) Classification des départements selon le mode d'utilisation du sol agricole

On y voit apparaître quatre grandes classes :

– la première composée de 22 départements, que l'on peut qualifier de producteurs de céréales ;

– la deuxième comportant 17 départements, où dominant les cultures permanentes ;

– la troisième classe réunit 20 départements caractérisés par des surfaces toujours en herbe et des cultures fourragères et maraichères ;

– la quatrième, enfin, regroupe des départements qui sont très fortement dominés par des surfaces toujours en herbe (tels que par exemple : Savoie, Alpes-Maritimes et Cantal, où l'herbe occupe respectivement 89 %, 92 % et 80 % des surfaces cultivables).

c) Par la classification des départements à travers la structure d'exploitation et le mode d'utilisation du sol agricole, on retrouve les classes se différenciant par divers systèmes de cultures, tels que : grandes exploitations céréalières, petites exploitations avec des cultures permanentes, exploitations moyennes associées aux cultures fourragères et maraichères et petites ou moyennes exploitations dominées par des surfaces toujours en herbe.

d) Enfin, la classification des départements selon les trois caractères a permis de situer dans l'espace géographique les trois types de production dégagés par la classification de l'ensemble des modalités (voir figure 3 qui correspond à la partition associée au niveau le plus significatif).

VIII.2. Comparaison des résultats avec ceux obtenus à partir du critère de l'inertie expliquée

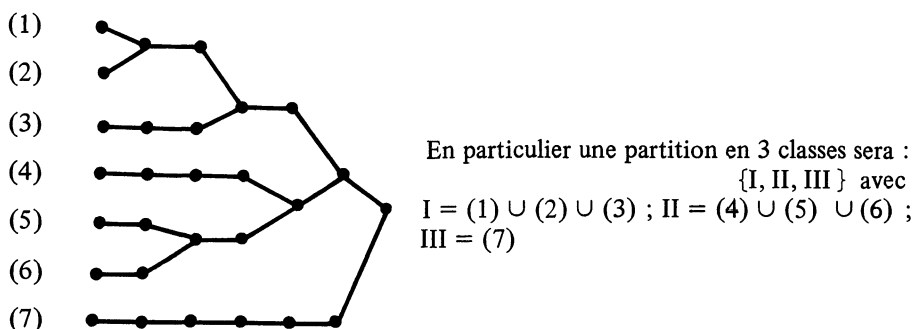
En vue de faciliter et rendre plus claire la comparaison, nous avons retenu la classification des départements selon la structure d'exploitation ; ce qui nous ramène à une vraie table de contingence. L'algorithme basé sur le critère de l'inertie expliquée que nous désignerons par C.A.H. (M^2 , χ^2) donne la partition en sept classes suivantes numérotées de (1) à (7) :

– Les classes (1), (2) et (3) sont formées des départements où dominant les petites et très petites exploitations

– La classe (4) : exploitations de taille moyenne

- La classe (5) : exploitations de taille moyenne ou grande
- La classe (6) : grandes et très grandes exploitations
- La classe (7) : très grandes exploitations

Le schéma ci-dessous permet d'obtenir une partition en moins de 7 classes.



Le tableau suivant le croisement des deux classifications ; obtenues par les algorithmes A.V.L. (C.O.R.) d'une part et celui C.A.H. (M^2 , χ^2) permet d'établir la correspondance entre elles :

A.V.L. \ C.A.H. (M^2, χ^2)	I		II		III			Total
	(7)	(6)	(5)	(4)	(3)	(2)	(1)	
A (grandes expl.)	15	10	2	0	1	0	0	28
B (moyennes expl.)	0	0	23	22	0	0	0	45
C (petites expl.)	0	0	0	2	8	1	5	16
Total	15	10	25	24	9	1	5	89

où on a inscrit à l'intersection d'une ligne et d'une colonne le nombre de départements appartenant à la fois aux deux classes correspondantes.

On remarque la correspondance entre la classe A et (6) U (7) ; entre la classe B et (4) U (5) et entre la classe C et (1) U (2) U (3) à quelques exceptions près.

Au niveau de la partition en 6 ou 7 classes, les deux classifications sont pratiquement équivalentes. Mais dès que l'on cherche une partition en 5 classes ou moins l'algorithme ; C.A.H. (M^2 , χ^2) réunit la classe (6) à la classe (5), alors que l'A.V.L. (COR) la réunit à la classe (7). En effet, la classe (6) comportant de grandes ou très grandes exploitations, se trouve plus naturellement associée à la classe (7) des très grandes exploitations qu'à la classe (5) formée d'exploitations moyennes.

Il apparaît donc pour les données qui nous concernent ici que les deux méthodes ont fourni des résultats quasiment équivalents jusqu'avant les dernières associations de formation des grandes classes où la partition donnée par A.V.L. (COR.) s'explique mieux que celle produite par C.A.H. (M^2 , χ^2).

IX. CONCLUSION

L'indice de proximité que nous avons défini et étendu au paragraphes V ouvre pour notre méthode très générale de classification hiérarchique, la voie à une classe de structures de tableaux de données qui est le domaine privilégiée de l'Analyse Factorielle des Correspondances et de la Classification des moindres carrés conformément à la métrique du χ^2 . Les résultats que nous avons pu obtenir jusqu'à présent sont d'une grande finesse et d'une grande cohérence et la comparaison avec les résultats de méthodes maintenant plus classiques, est pleine d'intérêt (cf. [4], [5], [15], [16]).

Il sera d'autre part intéressant d'étudier la comparaison entre l'indice que nous avons pu proposer entre distributions de probabilités sur un ensemble fini J (cf. § V.5), avec d'autres indices connus ; par exemple celui $\Sigma\{\phi_j^i \phi_j^{i'}/j \in J\}$ entre les deux distributions ϕ_j^i et $\phi_j^{i'}$ sur J , sous jacent à la métrique de Hellinger, qu'on dit également être de Matusita.

Il sera enfin utile d'évaluer l'intérêt des résultats d'un arbre des classifications directement associé à l'indice (cf. § VI ci-dessus) par rapport à la solution que nous avons adoptée et qui, encore une fois, consiste à coupler la table des indices de corrélation entre lignes (resp. colonnes) du tableau initial avec A.V.L.

BIBLIOGRAPHIE

- [1] J.P. BENZECRI et collaborateurs. — *L'Analyse des Données* — Tome 1. "La Taxinomie", Tome 2. Analyse des Correspondances. Dunod, 1973.
- [2] M. JAMBU. — *Classification Automatique pour l'Analyse des Données* — tome 1. Méthodes et Algorithmes. Dunod, 1978.
- [3] L. LEBART, A. MORINEAU et N. TABARD. — *Techniques de la Description Statistique*. Dunod, 1977.
- [4] G. LECRENAIS. — *Classification de Catégories Socio-Professionnelles de Parents relativement aux Projets d'Avenir qu'ils formulent pour leurs enfants*, rapport de D.E.A., Université de Rennes 1, Laboratoire de Statistiques, Sept. 1979.
- [5] A. LELIEVRE. — *Méthodes de Classification Hiérarchique Appliquées à la Répartition des Décès Attribuables à certaines causes dans les communes de deux départements bretons*, rapport de D.E.A., Université de Rennes 1, Laboratoire de Statistiques, Sept. 1979.
- [6] I.C. LERMAN. — *Les Bases de la Classification Automatique*, Gauthier-Villars, 1970.
- [7] I.C. LERMAN. — "Etude Distributionnelle de Statistiques de Proximités entre Structures Finies de même type — Application à la Classification Automatique", *Cahiers du B.U.R.O.* n° 19, ISUP, 1973.
- [8] I.C. LERMAN. — *Reconnaissance et Classification des Structures Finies en Analyse des Données* vol. 1. *Théorie et méthodes*, rapport n° 70, I.R.I.S.A., Rennes 1976-77.

- [9] I.C. LERMAN. — *Formal Analysis of a General Notion of Proximity between Variables* in Proceed. of European Congress of Statisticians, Grenoble, Sept. 1976.
- [10] I.C. LERMAN. — *Etude Formelle et Statistique de la Notion de Ressemblance*, rapport n° 107, I.R.I.S.A. Rennes.
- [11] I.C. LERMAN. — “Méthodes combinatoires et statistiques dans le traitement des données du comportement”. *Bulletin Statistique et Analyse des Données de l’A.S.U.*, 2 — 1978.
- [12] I.C. LERMAN. — Croisement de classifications floues”, *Publications de l’Institut de Statistique*, Université Paris, 1979, XXIV, fasc. 1-2, 13-46.
- [13] I.C. LERMAN, M. HARDOUIN et T. CHANTREL. — “Analyse de la situation relative entre deux classifications floues”, secondes journées internationales “Analyse des Données et Informatique”, *colloque I.R.I.A.*, Versailles, 17-19 Oct. 1979.
- [14] J.C. LERMAN et B. TALLUR. — “Classification des éléments constitutifs d’une juxtaposition de tableaux de contingence”, *rapport I.R.I.S.A.* n° 127, Université de Rennes I, 1980.
- [15] B. TALLUR. — “Etude Régionale l’Agriculture Française”, *rapport I.R.I.S.A.* n° 103, Université de Rennes I, 1978.
- [16] Cl. THELOT et Th. LEMAITRE. — *La transmission du statut social : l’exemple des Pays de la Loire*, rapport interne, Direction régionale de l’I.N.S.E.E. de Nantes, 044 n° 3/NR, 2 Février 1979.