L'ORIENTATION DU DÉPOUILLEMENT DE CERTAINES ENQUÊTES PAR L'ANALYSE DES CORRESPONDANCES MULTIPLES

par

Ludovic LEBART

SOMMAIRE

Il n'est pas rare que des enquêtes (1) très coûteuses de par la nature du recueil de données (volume du questionnaire et de l'échantillon, conditions matérielles du questionnement, apurement) soient exploitées de façon incomplète : l'information patiemment recueillie est alors inutilisée pour sa plus grande part. Nous proposons dans cet article une procédure permettant d'orienter certaines phases du dépouillement d'une enquête qui généralise la démarche suivie lors de l'étude classique par tabulations. L'outil utilisé au cours de cette procédure, l'analyse des correspondances multiples, généralisation naturelle de l'analyse des correspondances [3] (2), a déjà

(1) Cet article concerne exclusivement les enquêtes par questionnaires fermés réalisées sur des échantillons importants.

(2) Les chiffres entre crochets renvoient à la bibliographie en fin d'article.

été présenté dans un rapport C.R.E.D.O.C.-C.O.R.D.E.S. paru en 1973 [12] et antérieurement dans une note plus théorique du laboratoire de statistique mathématique de l'Université de Paris VI [2]. Cette généralisation s'appuie notamment sur certains travaux déjà anciens du statisticien britannique Cyril Burt [4]. Cette méthode semble avoir été entrevue par d'autres statisticiens, essentiellement des praticiens, comme le laisse entendre M.O. Hill [9].

Dans une première partie, nous rappellerons brièvement en quoi consiste la phase de dépouillement d'une enquête, en tentant d'insister sur les inconvénients et les lacunes de certaines procédures usuelles.

Dans une seconde partie, nous donnerons un exposé technique de l'analyse des correspondances multiples.

Dans la troisième partie, nous montrerons comment cette méthode permet de guider le choix des tabulations les plus pertinentes, et de procéder à une critique des plans d'exploitations qui auraient été faits *a priori*, en éliminant les tabulations redondantes et en exhibant éventuellement des croisements significatifs qui n'auraient pas été prévus. Nous nous appuierons en partie sur un exemple illustratif emprunté à l'enquête C.N.A.F.-C.R.E.D.O.C. de 1971 [18].

Enfin, la quatrième partie examine les problèmes posés par la mise en œuvre pratique de ces méthodes ; les aspects techniques et les modalités pratiques de calculs feront l'objet d'une prochaine publication.

1. LE DÉPOUILLEMENT D'ENQUÊTE

On peut, en première approximation, classer les variables décrivant une unité statistique lors d'une enquête en deux groupes principaux :

— Le premier groupe comporte les variables que nous qualifierons en bref de *variables socio-administratives* et qui se scindent en deux sous-groupes :

a) les variables *de base*, qui servent à construire le plan de sondage, et qui sont connues avant même l'interview (ce peut être selon les cas : le nombre d'enfants, certains aspects de la localisation géographique, le nombre de pièces du logement, etc.). Ces variables préexistent sur le fichier qui a servi à construire l'échantillon ;

b) les variables que nous appellerons *de structure*, qui sont de même nature que les précédentes, mais qui ne font pas l'objet de contrôle *a priori* : ce sont par exemple l'âge du chef de famille, la composition de la famille, la catégorie socio-professionnelle de chacun des membres actifs de la famille ; il s'agit principalement de variables démographiques, économiques, ou décrivant de façon très générale (dans un contexte administratif), l'insertion sociale de la famille. Il s'agit en somme de variables permettant au

législateur, au politique, à l'administratif, d'identifier la famille à partir des catégories qui leur sont propres.

— Le second groupe est constitué par les *variables relatives au contenu même de l'enquête*, qui peuvent concerner un ou plusieurs thèmes ; on peut les diviser en trois sous-groupes, que distinguent le niveau et la qualité de la mesure :

a) les variables constituées par des réponses à des questions factuelles (possession d'un chauffe-eau, lieu où sont pris les repas), que nous distinguerons des groupes précédents parce que, par exemple, les possesseurs de chauffe-eau ne constituent pas un groupe sociologique ni une entité administrative d'usage courant ;

b) les variables décrivant un comportement de la personne enquêtée ou de ses proches, qui sont encore factuelles, mais qui peuvent faire l'objet de dissimulation, être entâchées d'inexactitudes, être difficiles à coder (par exemple : « regardez-vous la télévision ? ») ;

c) enfin, les variables d'attitudes ou d'opinion, qui jouent un rôle fondamental dans la compréhension et la prévision des phénomènes socio-économiques ; ces variables ne fournissent cependant qu'une information fragile et vulnérable, surtout si on les étudie isolément, en s'abstenant de faire converger plusieurs questions (formulées différemment) sur un même thème.

Le dépouillement de l'enquête consistera le plus souvent à utiliser les variables du premier groupe pour définir des grilles de tabulations, afin de comprendre et d'expliquer le « comportement » des variables du second groupe.

On pourra tout d'abord procéder à un examen minutieux des variables de niveau et à des « tris à plat » (par exemple, calcul d'une moyenne de consommation d'un certain produit par catégorie socio-professionnelle), puis à des « tris croisés » (en effectuant par exemple le tri à plat précédent pour chaque catégorie de commune, ce qui revient à croiser les critères « catégorie socio-professionnelle » et « catégorie de commune »). Les variables socio-administratives jouent en quelque sorte le rôle de prédicteurs, et fournissent des présomptions d'explication.

Si l'enquête n'est pas la première du genre, l'expérience antérieure des statisticiens peut faire espérer que le plan d'exploitation réalisé *a priori* recouvrera l'essentiel des problèmes auxquels l'enquête tente d'apporter des réponses. Si elle concerne un champ original, il est probable que les tabulations prévues seront en partie redondantes et cependant qu'elles s'avèreront insuffisantes. Le nombre de tabulations jugé nécessaire *a priori* peut être considérable (souvent plusieurs milliers de tableaux croisés, si les variables socio-administratives sont nombreuses). De plus, la consultation séquentielle des tableaux croisés ne tient pas compte des relations existant entre les éléments mêmes de la grille de tabulation. Ainsi, étudier

le temps consacré à certains loisirs en fonction de la catégorie socio-professionnelle, puis en fonction du niveau d'instruction, enfin en fonction de classes de revenus est une démarche qui ne tient pas compte des interrelations existant entre ces trois types de critères.

Il semble donc nécessaire, afin d'être le plus exhaustif possible, d'utiliser de façon globale les variables socio-administratives, tout en tenant compte de leur réseau d'interrelations, de façon à éviter un certain piétinement dans la lecture des résultats. L'analyse des correspondances multiples va nous permettre de progresser dans cette voie.

2. L'ANALYSE DES CORRESPONDANCES MULTIPLES

2.1. Généralités

Une partie généralement importante des fichiers d'enquête se compose de réponses à des questions mises sous *forme disjonctive complète*, c'est-à-dire de questions dont les diverses modalités de réponses s'excluent mutuellement, et telles qu'une modalité est obligatoirement choisie.

L'ensemble des r modalités de réponses à une telle question permet de partitionner l'échantillon en r classes, au plus.

Exemple 1 : intitulé de la question : âge du père.

Huit modalités : 1° moins de 25 ans,
2° de 25 à 29 ans,
3° de 30 à 34 ans,
4° de 35 à 39 ans,
5° de 40 à 44 ans,
6° de 45 à 49 ans,
7° 50 ans et plus,
8° sans objet ou non-réponse.

Exemple 2 : intitulé de la question : avez-vous un (ou plusieurs) lave-vaisselle ? »

Deux modalités : 1° oui,
2° non.

La donnée des deux questions mises sous forme disjonctive complète nous permet d'observer deux partitions de l'ensemble des individus enquêtés. L'analyse du tableau de correspondance croisant ces deux partitions peut être généralisée au cas de Q partitions (Q étant un entier supérieur à 2). La généralisation proposée (qui n'est évidemment pas la seule possible) est assez naturelle et conduit à des règles d'interprétation simples.

a) *Notations*

Le cardinal d'un ensemble A (nombre d'éléments de A) sera noté card A.

L'ensemble des questions sera désigné par Q. Une question q consiste en un ensemble J_q de card J_q modalités. $J = \bigcup \{ J_q \mid q \in Q \}$ désigne la réunion de tous ces ensembles de modalités.

On désignera par $I = \prod \{ J_q \mid q \in Q \}$ l'ensemble produit des J_q, c'est-à-dire l'ensemble dont les éléments sont constitués des suites de q modalités, chacune de celles-ci étant prise dans une question différente. Les éléments de I sont donc les réponses possibles des sujets enquêtés.

L'ensemble des individus enquêtés sera désigné par S.

Contrairement aux approches classiques des tables de contingence multiple, où l'on s'intéresse principalement aux fréquences $k(i)$ des individus ayant donné la réponse i ($i \in I$), l'étude qui va suivre reste intéressante si card S est très inférieure à card I, ce qui est souvent le cas dans les applications. Si l'on pose à mille individus douze questions ayant chacune dix modalités : card $S = 10^3$, card $I = 10^{12}$.

Ainsi dans le tableau $k(i)$, il n'y a qu'une proportion infime d'éléments différents de O.

On désignera par Z le tableau (card $S \times$ card J) donnant, pour l'individu $s \in S$ une description booléenne de ses réponses aux card Q questions.

On désignera par $R(s, q)$ la modalité de la question q choisie par le sujet s, ($R(s, q) \in J_q$).

Le tableau des éléments $R(s, q)$ constitue un codage condensé du tableau Z. (Le tableau de terme général $R(s, q)$ n'a que card Q colonnes.)

Les programmes de calcul n'utilisent en fait que ce type de tableaux comme entrée. A l'intérieur de chaque question J_q, les modalités sont indicées de 1 à card J_q. $R(s, q)$ n'est autre que la valeur de cet indice, pour l'individu s et la question q.

b) *Tableau de BURT associé à Z*

Partitionnons les colonnes de la matrice Z de façon à faire apparaître dans un même sous-tableau Z_q les colonnes relatives aux modalités de la question q.

Dans ces conditions : $Z = (Z_1, Z_2, \ldots, Z_{\text{card } Q})$.

Si Z^T désigne la transposée de Z, le tableau

$$B = Z^T Z$$

est appelé « tableau de contingence de Burt » associé au tableau des réponses Z.

Le tableau B est formé de (card Q)^2 blocs.

Le q-ième bloc diagonal $Z_q^T Z_q$ est une matrice diagonale d'ordre (card J_q)^2 (puisque deux modalités d'une même question ne peuvent être choisies simultanément).

Le bloc indicé par (q, q'), d'ordre (card $J_q \times$ card $J_{q'}$), n'est autre que le tableau de contingence croisant les réponses aux deux questions q et q'.

Nous désignerons par D, d'ordre (card $J \times$ card J), la matrice diagonale ayant les mêmes éléments diagonaux que B (ces éléments diagonaux ne sont autres que les effectifs correspondant à chacune des modalités).

La matrice D peut être également considérée comme formée de (card Q)^2 blocs (seuls les card Q blocs diagonaux sont des matrices non nulles, avec, pour le q-ième bloc diagonal : $D_q = Z_q^T Z_q$, matrice diagonale dont les termes diagonaux sont les effectifs correspondant aux diverses modalités de la question q).

2.2. Cas de deux modalités: (card Q = 2) (correspondance binaire)

Le tableau des réponses Z s'écrit alors : $Z = (Z_1, Z_2)$.

Il est alors équivalent, du point de vue de la description des associations entre modalités :

1° d'effectuer l'analyse des correspondances du tableau Z d'ordre (card S, card J) ;

2° d'effectuer l'analyse des correspondances du tableau B d'ordre (card $J \times$ card J) ;

3° d'effectuer l'analyse des correspondances du tableau $Z_1^T Z_2$ d'ordre (card $J_1 \times$ card J_2).

Montrons que les analyses (1) et (2) fournissent les mêmes facteurs (à une normalisation près) :

Les facteurs issus de (1) vérifient l'équation :

$$\frac{1}{\text{card } Q} D^{-1} Z^T Z \varphi = \lambda \varphi. \tag{1}$$

D'autre part, les marges du tableau B sont les éléments diagonaux de la matrice card Q. D.

La relation de transition (1) relative à l'analyse de B s'écrit, puisque $B = Z^T Z$ est symétrique (et définie non négative), pour un facteur ψ relatif à la valeur propre μ :

$$\frac{1}{\text{card } Q} D^{-1} Z^T Z \psi = \sqrt{\mu}\, \psi. \tag{2}$$

Ainsi, $\psi = \varphi$ et $\mu = \lambda^2$.

---

(1) Nous désignerons ici par relation de transition la relation liant les facteurs homologues relatifs aux deux côtés du tableau analysé.

Montrons maintenant que pour tout couple de facteurs (φ_1, φ_2) associé à la même valeur propre β lors de l'analyse du tableau de contingence $Z_1^T Z_2$ correspond un facteur $\varphi = \begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix}$ de l'analyse de B (ou de Z).

Les deux marges du tableau rectangulaire $Z_1^T Z_2$ sont les éléments diagonaux de D_1 et D_2.

Les deux relations de transition s'écrivent :

$$D_1^{-1} Z_1^T Z_2 \varphi_2 = \sqrt{\beta}\,\varphi_1, \tag{3}$$

$$D_2^{-1} Z_2^T Z_1 \varphi_1 = \sqrt{\beta}\,\varphi_2. \tag{4}$$

Ou encore

$$D_1^{-1}(D_1 \varphi_1 + Z_1^T Z_2 \varphi_2) = (1+\sqrt{\beta})\,\varphi_1,$$

$$D_2^{-1}(D_2 \varphi_2 + Z_2^T Z_1 \varphi_1) = (1+\sqrt{\beta})\,\varphi_2$$

qui s'écrit également, après multiplication des deux membres par $1/2 = 1/\text{card } Q$:

$$\frac{1}{2} D^{-1} Z^T Z \varphi = \frac{(1+\sqrt{\beta})}{2}\,\varphi.$$

Ce qui n'est autre que la relation (1), où $\lambda = (1 + \sqrt{\beta})/2$.

Ainsi, les valeurs propres issues des trois analyses sont respectivement λ, λ^2, $(2\lambda - 1)^2$.

On peut faire ici deux remarques :

Remarque 1 : dans l'analyse du grand tableau disjonctif Z, les points représentant les diverses modalités de réponses aux deux questions sont des éléments d'un même ensemble (l'ensemble des colonnes de Z). Alors que dans l'analyse du tableau de contingence $Z_1^T Z_2$, ils se scindent en points lignes et en points colonnes. Le fait que les typologies obtenues dans l'espace des premiers facteurs soient identiques (à une dilatation près due au fait que les valeurs propres ne sont pas les mêmes) nous prouve que la représentation simultanée des points lignes et des points colonnes en analyse des correspondances des tableaux de contingence n'est pas qu'un artifice graphique.

Remarque 2 : ces trois analyses, reposant sur la même information brute, donnent des résultats similaires, mais avec des valeurs propres différentes, donc des taux d'inertie différents. Les relations existant entre les taux d'inertie nous montrent que ceux-ci seront toujours beaucoup plus élevés lors de l'analyse du tableau de contingence $Z_1^T Z_2$ que lors de celle du tableau Z. D'une manière générale, l'analyse des tableaux sous codage disjonctif donne toujours des taux d'inertie faibles, qui donnent une idée beaucoup trop pessimiste de la part d'information extraite.

2.3. Généralisation au cas de plus de deux questions

Le tableau $Z = (Z_1, Z_2, ..., Z_q, ..., Z_{\text{card}\,Q})$ possède card J colonnes, auxquelles correspondent card J points de $R^{\text{card}\,S}$; plaçons-nous dans l'espace $R^{\text{card}\,S}$. Chaque sous-tableau Z_q engendre une variété linéaire $\mathscr{V}_q$ à card J_q dimensions.

Toutes ces variétés linéaires ont au moins en commun la première bissectrice. Le rang du tableau Z est donc au plus égal à card J — (card Q — 1).

Soit φ_q le tableau (à card J_q lignes et une colonne) des composantes d'un point (ou vecteur) $\mathscr{M}_q$ de $\mathscr{V}_q$ dans la base définie par les colonnes de Z_q.

Le carré de la distance de ce point $\mathscr{M}_q$ à l'origine, selon la norme euclidienne usuelle n'est autre que

$$\varphi_q^T Z_q^T Z_q \varphi_q = \varphi_q^T D_q \varphi_q.$$

L'analyse des correspondances du tableau de contingence croisant deux questions q et q' revient à étudier les positions respectives des variétés $\mathscr{V}_q$ et $\mathscr{V}_{q'}$. En effet, dans $R^{\text{card}\,S}$, les opérateurs-projections sur $\mathscr{V}_q$ et $\mathscr{V}_{q'}$ correspondent (relativement aux bases précitées) aux matrices

$$Z_q(Z_q^T Z_q)^{-1} Z_q^T (= Z_q D_q^{-1} Z_q^T) \qquad \text{et} \qquad Z_{q'}(Z_{q'}^T Z_{q'})^{-1} Z_{q'}^T (= Z_{q'} D_{q'}^{-1} Z_{q'}^T).$$

Les relations de transition (3) et (4) (où $q = 1$, $q' = 2$) expriment que les points $\mathscr{M}_q$ et $\mathscr{M}_{q'}$ sont projections l'un de l'autre. Il revient au même de chercher deux points $\mathscr{M}_q$ et $\mathscr{M}_{q'}$ tels que leur moyenne des carrés des distances à l'origine soit constante.

$$\varphi_q^T D_q \varphi_q + \varphi_{q'}^T D_{q'} \varphi_{q'} = 2\,\text{card}\,S \tag{5}$$

et tels que la distance à l'origine du point $\mathscr{M} = \mathscr{M}_q + \mathscr{M}_{q'}$ soit maximale.

$$\|\mathscr{M}\|^2 = \varphi_q^T D_q \varphi_q + \varphi_{q'}^T D_{q'} \varphi_{q'} + 2\varphi_q^T Z_q Z_{q'} \varphi_{q'}, \tag{6}$$

$$\|\mathscr{M}\|^2 = 2\,\text{card}\,S\left(1 + \left[\frac{1}{\text{card}\,S}\varphi_q^T Z_q^T Z_{q'} \varphi_{q'}\right]\right). \tag{7}$$

Remarque : le maximum de $\|\mathscr{M}\|^2$ s'obtient d'ailleurs (et ceci ne sera valable que pour deux questions) pour $\varphi_q^T D_q \varphi_q = \varphi_{q'}^T D_{q'} \varphi_{q'}$. L'expression entre crochets dans le membre de droite de la relation (7) n'est autre que le cosinus de l'angle des vecteurs $(\mathscr{M}_q, \mathscr{M}_{q'})$.

Posé sous cette dernière forme, le problème se généralise aisément au cas de plus de deux questions.

Si $\varphi_1, \varphi_2, ..., \varphi_{\text{card}\,Q}$ désignent respectivement les vecteurs des composantes de card Q points $\mathscr{M}_1, \mathscr{M}_2, ..., \mathscr{M}_{\text{card}\,Q}$ dans les bases $Z_1, Z_2, ..., Z_{\text{card}\,Q}$ avec $\mathscr{M} = \mathscr{M}_1 + \mathscr{M}_2 + ... + \mathscr{M}_{\text{card}\,Q}$ on cherchera à rendre maxi-

male la quantité

$$\|\mathcal{M}\|^2 = \Sigma\{\varphi_q^T Z_q^T Z_{q'} \varphi_{q'} \mid q \in Q,\ q' \in Q\}$$

avec la contrainte

$$\Sigma\{\varphi_q^T D_q \varphi_q \mid q \in Q\} = \text{card } Q \,.\, \text{card } S.$$

Si φ désigne le vecteur à card J composantes tel que

$$\varphi^T = (\varphi_1^T, \varphi_2^T, \ldots, \varphi_{\text{card } Q}^T) \qquad (\text{ou encore } \varphi = \oplus\{\varphi_q \mid q \in Q\})$$

le problème revient à rendre maximal $\varphi^T B \varphi$ avec $\varphi^T D \varphi = 1$.

Les facteurs φ cherchés sont donc les vecteurs propres de $D^{-1}B$ relatifs aux plus grandes valeurs propres, qui sont proportionnels à ceux issus de l'analyse des correspondances du tableau Z (qui coïncident de plus, à une normalisation près, avec ceux issus de l'analyse du tableau B considéré lui-même comme un tableau de données).

Ainsi, l'analyse des correspondances du tableau disjonctif Z, par un programme classique, peut nous fournir les résultats escomptés. Cependant, ceci n'est possible que pour des tableaux de dimensions modestes, car les calculs deviennent vite très coûteux. Avec par exemple trente variables ayant chacune une dizaine de modalités de réponses, on est conduit à diagonaliser une matrice 300×300. La structure particulière du tableau permet en fait d'utiliser des procédures de calcul qui permettent d'éliminer certains obstacles techniques et de diminuer de façon parfois considérable les coûts (*cf.* quatrième partie).

2.4. Propriétés des analyses multiples

Les facteurs φ issus de l'analyse du tableau Z vérifient l'équation

$$\frac{1}{\text{card } Q} D^{-1} B \varphi = \lambda \varphi \tag{8}$$

en faisant apparaître les composantes φ_q de φ relatives à la question q et les blocs des tableaux D et B, cette équation s'écrit :

$$\frac{1}{\text{card } Q} \Sigma\{D_{q'}^{-1} . Z_{q'}^T Z_q \varphi_q \mid q \in Q\} = \lambda \varphi_{q'}.$$

a) Le centre de gravité du sous-nuage de card J_q points dont les coordonnées dans $R^{\text{card } S}$ sont les colonnes du tableau $Z_q D_q^{-1}$ décrivant les profils des réponses à la question q est le même que le centre de gravité général du nuage.

Il s'ensuit que les composantes de φ_q relatives à une question particulière q sont également centrées. (Chaque point-modalité j est muni d'une masse égale à d_{jj}/card Q.card S, où d_{jj} est le terme générique de la matrice D.)

b) La somme des valeurs propres non triviales vaut, d'après la relation (8) : card J/card $Q - 1$.

(La trace sera donc égale à 1 dans le cas des questions à deux modalités, pour lesquelles card $J = 2$ card Q.)

c) Le carré de distance au centre de gravité d'un point-modalité j ($j \in J$) de $\mathbb{R}^{\text{card } S}$ s'écrit :

$$d^2(0, j) = \Sigma\{\text{card}\, S.(Z_{ij}/d_{jj} - 1/\text{card}\, S)^2 \mid i \in S\}$$

soit, compte tenu de la relation : $\Sigma\{Z_{ij} \mid i \in S\} = d_{jj}$:

$$d^2(0, j) = \text{card}\, S\,(1/d_{jj} - 1/\text{card}\, S).$$

La contribution à l'inertie totale de la modalité j vaut donc

$$c(j) = \frac{d_{jj}}{\text{card}\, Q\, \text{card}\, S} d^2(0, j) = \frac{1}{\text{card}\, Q}\left(1 - \frac{d_{jj}}{\text{card}\, S}\right).$$

La contribution de la question q à l'inertie totale vaut

$$C(q) = \Sigma\{c(j) \mid j \in J_q\} = \frac{1}{\text{card}\, Q}(\text{card}\, J_q - 1).$$

On vérifie que

$$\Sigma\{C(q) \mid q \in Q\} = \text{card}\, J/\text{card}\, Q - 1.$$

La forme de $c(j)$ nous prouve que la contribution d'une modalité est d'autant plus forte que l'effectif correspondant est plus faible, sans toutefois pouvoir dépasser 1/card Q.

d) *Réduction des dimensions du tableau à diagonaliser.*

Dans l'espace $R^{\text{card } S}$, les points représentatifs des card J modalités ont pour coordonnées les colonnes de ZD^{-1} ;

Nous avons vu que le rang de Z (donc de ZD^{-1}) est au plus égal à card J — card $Q + 1$; la variété linéaire engendrée par les colonnes de ZD^{-1} contient la première bissectrice. Comme le nuage est dans l'hyperplan D^{-1} orthogonal à la première bissectrice, le nombre de valeurs propres nulles lors de l'analyse du nuage par rapport à son centre de gravité sera de Card Q.

En faisant choix d'une base dans le support du nuage, on se ramènera donc à la diagonalisation d'une matrice symétrique d'ordre (Card J — Card Q) × (Card J — Card Q).

e) Cas particulier : questions à deux modalités.

Dans ce cas, bien que la réduction précédente puisse s'appliquer sans perte notable de temps, on obtient directement la matrice à diagonaliser, symétrique, qui n'est autre que la matrice des corrélations entre variables, celles-ci n'étant représentées que par une seule de leurs modalités — (Card J — Card Q = 1/2 Card J).

Explicitons la relation (8) ci-dessus, où, rappelons-le, D désigne la matrice diagonale ayant les mêmes éléments diagonaux que B.

$$\text{(8 } bis\text{)} \qquad \frac{1}{\text{card } Q} \sum_{j \in J} \frac{b_{ij}}{b_{ii}} \varphi^j = \lambda \varphi^i.$$

L'ensemble J des questions va maintenant être partitionné en deux sous-ensembles de mêmes cardinalités J^1 et J^2 formés respectivement des premières et des deuxièmes modalités de chacune des card Q questions.

$$J = J^1 \cup J^2 \qquad (J_q = \{j_q^1, j_q^2\} \,|\, j_q^1 \in J^1,\ j_q^2 \in J^2,\ q \in Q).$$

Notons les relations, pour tout $q \in Q$:

$$\begin{cases} b_{ij_q^1} + b_{ij_q^2} = b_{ii}, \\ b_{j_q^1 j_q^1} + b_{j_q^2 j_q^2} = \text{Card } S, \qquad b_{j_q^1 j_Q^1} \cdot \varphi_{j_Q^1} = -b_{j_q^2 j_q^2} \varphi_{j_q^2}. \end{cases}$$

Il suffit donc de restreindre la sommation de la relation (8 *bis*) au seul ensemble J^1, dont l'élément courant sera désormais noté j :

$$\frac{1}{\text{card } Q\, b_{ii}} \Sigma \left\{ \left(b_{ij} \varphi^j - \frac{(b_{ii} - b_{ij})\, b_{jj} \varphi_j}{(\text{card } S - b_{jj})} \right) \middle| \, j \in J^1 \right\} = \lambda \varphi^i.$$

Ce qui peut s'écrire :

$$\text{(9)} \qquad \Sigma \left\{ \frac{\text{Card } S \,.\, b_{ij} - b_{ii} \,.\, b_{jj}}{\text{Card } Q\, (\text{card } S - b_{jj})\, b_{ii}} \varphi^j \,|\, j \in J^1 \right\} = \lambda \varphi^i.$$

Calculons les moments empiriques centrés du second ordre des card Q variables caractérisées par leurs premières modalités.

$$\text{cov}(i,j) = \frac{1}{\text{card } S} \left(b_{ij} - \frac{b_{ii}\, b_{jj}}{\text{card } S} \right),$$

$$\text{var}(j) \ = \frac{1}{\text{card } S} \left(b_{jj} - \frac{b_{jj}^2}{\text{card } S} \right).$$

Le terme général de la matrice des corrélations des card Q variables s'écrit

$$\operatorname{cor}(i, j) = \frac{\operatorname{Card} S \cdot b_{ij} - b_{ii}\, b_{jj}}{\left[(\operatorname{Card} S - b_{jj})\, b_{jj} (\operatorname{Card} S - b_{ii})\, b_{ii}\right]^{1/2}}.$$

Il est clair que si (φ, λ) est la solution de l'équation (9) alors (ψ, λ') est la solution de

$$\Sigma \left\{ \operatorname{cor}(i, j)\, \psi^j \,\middle|\, j \in J_1 \right\} = \lambda' \psi^i,$$

avec

$$\begin{cases} \psi^j = \varphi^j \dfrac{(\operatorname{Card} S - b_{jj})^{1/2}}{b_{jj}}, \\ \lambda' = \lambda . \operatorname{card} Q. \end{cases}$$

f) Cas où l'analyse d'une correspondance multiple se ramène à celle d'une correspondance binaire

Le cas d'une correspondance binaire s'est révélé être particulièrement intéressant du point de vue des calculs à mettre en œuvre, car l'analyse du tableau de BURT d'ordre (Card $J \times$ Card J) équivalait à l'analyse des correspondances du tableau de contingence croisant les modalités des deux questions, ce qui conduit à diagonaliser une matrice d'ordre Inf (card J_1, card J_2).

Ce résultat peut être généralisé de diverses façons sous certaines conditions [2].

Nous retiendrons la propriété suivante, utile pour les applications :

— si l'ensemble Q des questions est partitionné en deux sous-ensembles Q_1 et Q_2 à l'intérieur desquels les questions sont indépendantes, l'analyse des card Q questions se réduit à celle d'une correspondance binaire, et donc à la diagonalisation d'une matrice d'ordre Inf (card J^1, card J^2), où $J^i = \{ J_q \mid q \in Q_i \}$.

(Nous dirons ici que deux questions q et q' sont indépendantes si le tableau $Z_q^T Z_{q'}$ est égal à $1/\operatorname{card} S . d_q \otimes d_{q'}$, où les vecteurs d_q et $d_{q'}$ ont respectivement pour composantes les éléments diagonaux de $Z_q^T Z_q$ et $Z_{q'}^T Z_{q'}$, qui sont également les éléments diagonaux de D_q et $D_{q'}$ de par la définition de ces deux matrices ; en notation matricielle $d_q \otimes d_{q'} = d_q . d_{q'}^T$.)

Écrivons de nouveau la relation (8) en partitionnant φ en deux blocs φ_{Q_1} et φ_{Q_2} ($\varphi_{Q_i} = \oplus \{ \varphi_q \mid q \in Q_i \}$) et les matrices B et D en quatre blocs, de façon à faire apparaître la dichotomie de $Q = Q_1 \cup Q_2$:

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, \qquad D = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}.$$

D'où les deux relations :

$$\begin{cases} \dfrac{1}{\text{card } Q}(D_1^{-1} B_{11} \varphi_{Q_1} + D_1^{-1} B_{12} \varphi_{Q_2}) = \lambda \varphi_{Q_1}, \\ \dfrac{1}{\text{card } Q}(D_2^{-1} B_{21} \varphi_{Q_1} + D_2^{-1} B_{22} \varphi_{Q_2}) = \lambda \varphi_{Q_2}. \end{cases}$$

Remarquons que les card Q_1 (resp. card Q_2) blocs diagonaux de $D_1^{-1} B_{11}$ (resp. $D_2^{-1} B_{22}$) sont des matrices unités dont les ordres correspondent aux cardinaux de chacune des questions :

$$(q \in Q_i,\ q' \in Q_i,\ q = q' \Rightarrow D_q^{-1} Z_q^T Z_q = I_{\text{card } q} \mid i \in \{1, 2\}).$$

On a d'autre part, pour $i \in \{ 1, 2 \}$:

$$q \in Q_i,\ q' \in Q_i,\ q \neq q' \Rightarrow D_q^{-1} Z_q^T Z_{q'} = \frac{1}{\text{card } S} D_q^{-1} d_q . d_{q'}^T.$$

En désignant par $(1_{\text{card}q})$ un vecteur dont les card q composantes valent 1,

$$D_q^{-1} Z_q^T Z_{q'} = \frac{1}{\text{card } S} . (1_{\text{card}q}) . d_{q'}^T.$$

Les relations $d_{q'}^T \varphi_{q'} = 0$ impliquent finalement que, pour $i \in \{ 1, 2 \}$:

$$D_i^{-1} B_{ii} \varphi_{Q_i} = \varphi_{Q_i}.$$

Le système ci-dessus s'écrit alors :

$$D_1^{-1} B_{12} \varphi_{Q_2} = (\lambda \text{ card } Q - 1) \varphi_{Q_1},$$
$$D_2^{-1} B_{21} \varphi_{Q_1} = (\lambda \text{ card } Q - 1) \varphi_{Q_2}.$$

D'où par substitution :

$$D_2^{-1} B_{21} D_1^{-1} B_{12} \varphi_{Q_2} = (\lambda \text{ card } Q - 1)^2 \varphi_{Q_2}.$$

Ainsi, φ_{Q_2} est obtenu par diagonalisation d'une matrice d'ordre (card Q_2). On en déduit facilement φ_{Q_1}.

Nous avons en fait implicitement supposé que card $Q_2 \leqq$ card Q_1, en choisissant de calculer φ_{Q_2} avant φ_{Q_1}.

Remarquons que B_{12} est obtenu par juxtaposition des tableaux de contingence (1) croisant l'ensemble des modalités des questions du premier groupe

(1) Sur l'étude de tableaux constitués par juxtaposition de plusieurs tableaux de contingence, on pourra consulter les travaux de A. LECLERC [15].

et celles relatives au second groupe. Les marges du tableau B_{12} sont les éléments diagonaux de card Q_2. D_1 et card Q_1 . D_2.

Les facteurs issus de l'analyse du tableau B_{12} vérifient la relation

$$\frac{1}{\text{card } Q_1 \text{ card } Q_2} D_2^{-1} B_{21} D_1^{-1} B_{12} \psi = \mu\psi.$$

Ils sont donc proportionnels aux facteurs trouvés précédemment.

Remarque : l'analyse des correspondances binaires peut évidemment se généraliser de plusieurs façons. Celle que nous examinons ici, dont on peut faire remonter le principe à C. BURT [4] est une simple extension du domaine d'application de l'analyse binaire. Les calculs mis en jeu sont relativement simples, les règles d'interprétation des représentations sont claires. D'autres types d'extensions ont été proposés par J.P. BENZECRI (1964) dans son cours à la Faculté des sciences de Rennes, par B. ESCOFFIER-CORDIER (1955) [7], et plus récemment par M. MASSON [16] qui s'appuie notamment sur les travaux de J. D. CARROLL [5], P. HORST [10] et J. R. KETTENRING [11].

3. APPLICATION AU DÉPOUILLEMENT D'ENQUÊTE

Ce paragraphe comprend trois parties. Nous allons tout d'abord montrer comment l'analyse des correspondances multiples permet de dresser une *grille socio-administrative* prête à recevoir des informations relatives à certains thèmes de l'enquête, en nous appuyant sur un exemple. Puis, nous rappellerons comment la technique de projection de variables illustratives s'apparente à la théorie de la régression multiple, et nous montrerons comment elle permet de sélectionner certains tableaux. Enfin, nous donnerons un exemple d'illustration de la grille, et de sélection de tableaux. Bien entendu, dans le cadre de cet article méthodologique, nous raisonnerons en dimensions réduites (pour des raisons d'encombrement graphique, en particulier) ; la technique ne prend tout son sens que dans le cas d'une application à grande échelle.

3.1. Un exemple de grille socio-administrative

Nous prendrons l'exemple de l'enquête C.N.A.F.-C.R.E.D.O.C. [18] réalisée en France en 1971 auprès de 2 003 familles dont le chef ou son conjoint sont salariés. Le plan de sondage assez particulier prévoyait une sur-représentation des familles nombreuses, des femmes exerçant une activité rémunérée, et prenait également en compte l'âge de l'aîné des enfants et certaines caractéristiques de la commune de résidence.

La figure 1 ci-après résume la première phase du dépouillement : l'établissement d'une grille (ou d'une carte) socio-administrative. Nous nous limitons à deux facteurs ici pour insister sur la partie de l'opération qui peut être visualisée. Mais la partie automatique du traitement, notamment la sélection des variables illustratives les plus importantes, peut évidemment se faire dans l'espace des k facteurs jugés significatifs après une procédure de simulation.

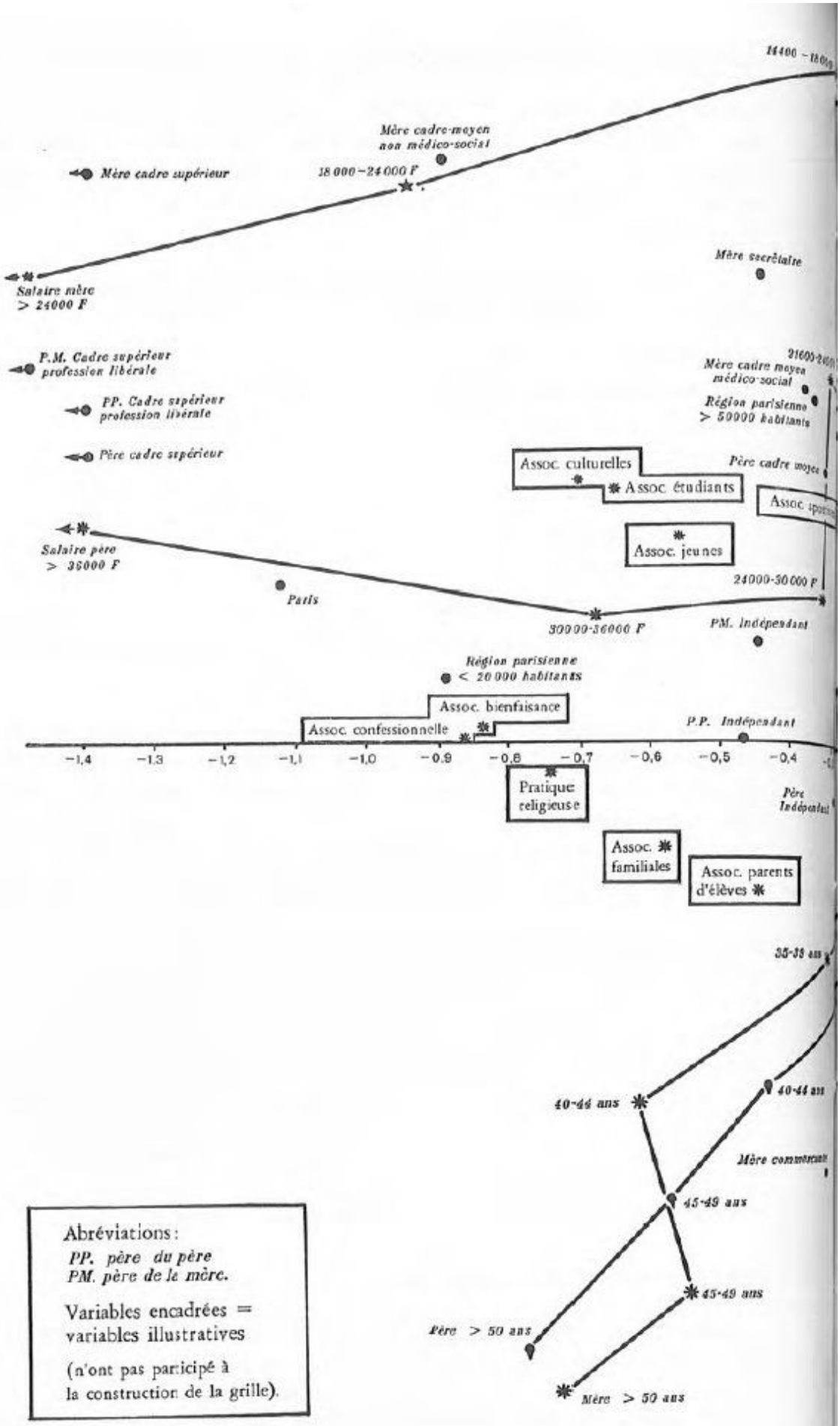
La figure 1 nous donne une typologie plane de 98 modalités de réponses relatives à 11 variables principales :

- profession du père,
- profession de la mère,
- salaire du père,
- salaire de la mère (lorsqu'il existe),
- âge du père,
- âge de la mère,
- nombre d'enfants,
- catégorie de communes,
- profession des ascendants paternels et maternels (deux variables),
- activité de la mère.

Il est clair qu'il existe une part d'arbitraire dans le choix des variables destinées à construire la grille ; il est d'ailleurs possible, si le questionnaire s'y prête, d'envisager l'élaboration de plusieurs types de grilles privilégiant chacune certains aspects des caractéristiques des familles. Il est cependant souhaitable d'obtenir une configuration qui soit la plus stable possible ; dans l'exemple que nous présentons, la grille n'a pas été modifiée de façon notable par la suppression d'un tiers de l'échantillon. De plus, malgré la stratification extrêmement déformante de l'échantillon, le plan des deux premiers facteurs est sensiblement le même, que l'analyse soit faite sur les données brutes ou sur les données redressées. Cette stabilité n'est pas très surprenante, car cette typologie décrit des associations et non des niveaux, qui, eux, sont très sensibles aux systèmes de pondérations. Non seulement les moments d'ordre 2 qui décrivent les associations sont beaucoup plus indépendants des systèmes de pondération que les moments d'ordre 1, mais de plus, le sous-espace engendré par les premières directions propres des matrices de moments d'ordre 2 est lui-même assez stable vis-à-vis des éventuelles fluctations de ces moments.

Interprétation générale de la figure 1

Certaines des variables analysées ont des modalités ordonnées de façon naturelle (par exemple, la variable « âge de la mère » comprend huit modalités ; si l'on excepte la modalité « non-réponse, sans objet », il existe un ordre naturel des différentes classes d'âge). Sur le graphique, ces modalités


Mère cadre-moyen non médico-social
Mère cadre supérieur
18 000-24 000 F
Mère secrétaire
Salaire mère > 24000 F
P.M. Cadre supérieur profession libérale
PP. Cadre supérieur profession libérale
Père cadre supérieur
Mère cadre moyen médico-social
Région parisienne > 50000 habitants
Assoc. culturelles
Assoc. étudiants
Père cadre moyen
Assoc. jeunes
Salaire père > 36000 F
Paris
24000-30000 F
30000-36000 F
PM. Indépendant
Région parisienne < 20 000 habitants
Assoc. bienfaisance
Assoc. confessionnelle
P.P. Indépendant
−1,4
−1,3
−1,2
−1,1
−1,0
−0,9
−0,8
−0,7
−0,6
−0,5
−0,4
Pratique religieuse
Père Indépendant
Assoc. familiales
Assoc. parents d'élèves
35-39 ans
40-44 ans
40-44 ans
Mère commerçant
45-49 ans
45-49 ans
Père > 50 ans
Mère > 50 ans
Abréviations :
PP. père du père
PM. père de la mère.
Variables encadrées = variables illustratives
(n'ont pas participé à la construction de la grille).

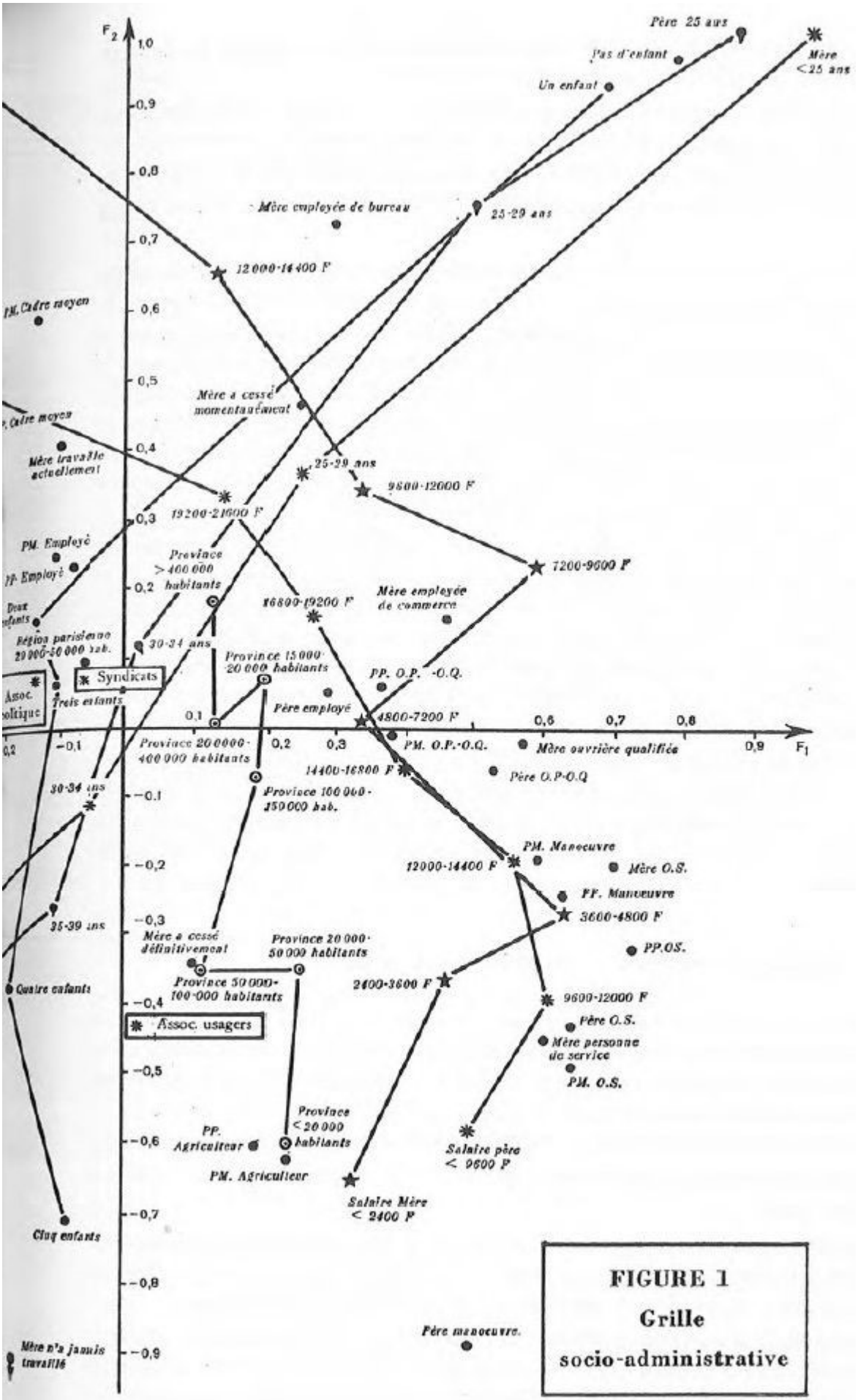


FIGURE 1
Grille socio-administrative

ordonnées sont jointes par un trait, afin de permettre de suivre facilement l'évolution du phénomène continu sous-jacent.

Les variables encadrées n'ont pas participé à l'analyse. Pour l'interprétation de leur position, on se reportera au paragraphe 3.3.

Cet éparpillement de variables, apparemment assez confus, s'ordonne en réalité autour de deux grands thèmes : l'âge de la famille et son statut social.

Suivons en effet les diverses classes d'âge du père, depuis le haut droit du graphique jusqu'au bas gauche, suivant une diagonale assez rectiligne : le long de cet axe, on trouve également les classes d'âge croissantes de la mère, moins dispersées toutefois que celles de leur mari ; on trouve également le nombre d'enfants croissant progressivement, avec un léger infléchissement sur la droite, correspondant au fait que les familles les plus nombreuses se trouvent surtout dans les milieux modestes ; on trouve également, le long de ce même axe, les différentes classes d'âge de l'aîné des enfants, également rangées par ordre croissant dans la même direction.

Assez perpendiculairement à cette direction se trouvent les lignes brisées joignant les différentes classes de salaire du père et de celui de la mère lorsqu'elle travaille. Ces deux lignes brisées sont repliées à leurs extrémités, et tournent leur concavité vers le haut. Les extrémités correspondent aux statuts sociaux précisément les plus extrêmes, ce repliement vers le haut nous montre que ce sont les familles plutôt âgées qui occupent les situations sociales les plus divergentes : aisance ou dénuement.

Le long de ces lignes brisées, les catégories socio-professionnelles du père et de la mère viennent illustrer et confirmer ce trajet le long de l'échelle sociale : aux manœuvres, gens de maison et ouvriers de diverses catégories à droite, s'opposent sur la gauche les professions libérales et les cadres supérieurs.

3.2. Régression visualisée et variables illustratives

Sur la figure 1, auraient également pu figurer les 2 003 familles (qui sont les lignes du tableau Z analysé). Avec ce type de codage, les relations dites de transition ont une interprétation simple : une famille est caractérisée par 11 réponses ; pour obtenir sa position, il suffit de prendre le centre de gravité des 11 points-réponses correspondants sur la figure 1, dont on dilatera les coordonnées (en faisant agir les coefficients « $1/\sqrt{\lambda}$ » relatifs à chaque axe).

De façon analogue, projeter les réponses à une question supplémentaire revient à calculer les centres de gravité des familles concernées par chacune des réponses, et à effectuer sur ces centres la dilatation précédente.

Il est clair que cette procédure est apparentée à la régression multiple dont elle constitue une variante descriptive. On peut en effet considérer les réponses aux variables socio-administratives comme des variables

exogènes, engendrant une certaine variété linéaire. Toutefois, au lieu de projeter les réponses supplémentaires (variables endogènes) directement sur cette variété, on commencera par ajuster celle-ci par un sous-espace de plus faible dimension (deux pour la figure 1). De plus, au lieu de nous intéresser aux coefficients de régression (coordonnées des projections dans la base des variables socio-administratives), on observe simplement les positions relatives des différentes réponses.

La figure 1 constitue donc une trame prête à accueillir différents tissages selon les thèmes de l'enquête.

Précisons que, du point de vue des calculs, il est beaucoup moins onéreux de projeter les modalités de réponse à une question supplémentaire que de croiser cette question avec l'ensemble des variables socio-administratives. De plus, la lecture du graphique est beaucoup plus aisée et sugggestive que celle des onze tableaux croisés correspondants.

Enfin, il n'est pas nécessaire d'observer toutes les réponses illustratives, car celles-ci peuvent en effet être très nombreuses (plusieurs milliers dans le cas de l'enquête citée). Il est en effet possible de sélectionner automatiquement celles qui occupent les positions les plus significatives (sur les k premiers facteurs par exemple).

Supposons qu'une réponse concerne n familles. L'hypothèse nulle sera : n points sont pris au hasard parmi les 2 0003 points repérés par leurs coordonnées dans l'espace des k premiers facteurs. Il est facile, dans ces conditions, de construire des seuils de signification pour la distance de la réponse supplémentaire à l'origine. On peut alors, soit retenir les réponses correspondant à un certain seuil de signification, soit classer ces réponses par ordre de signification croissante.

Bien entendu, ce filtrage automatique étant assez peu coûteux, il est possible de croiser les questions supplémentaires, ce qui permet de détecter certains types d'interactions.

Une extension de la notion de « contribution absolue »

La variance de l'abscisse f_i d'une réponse supplémentaire sur un axe factoriel relatif à la valeur propre λ est : $1/n_i$, si n_i est l'effectif concerné par la réponse : en effet, la variance de la population est λ, et la variance de la moyenne de n_i individus pris au hasard est donc λ/n_i. Comme l'abscisse sur l'axe est obtenue en multipliant la moyenne par $1/\sqrt{\lambda}$, on trouve bien le résultat annoncé.

Ainsi, les quantités $f_i \sqrt{n_i}$ sont, dans l'hypothèse nulle, des réalisations de variables aléatoires ayant même variance, et sont donc comparables entre elles du point de vue de leur signification. Ces mêmes quantités calculées pour les variables ayant participé à l'analyse sont proportionnelles aux racines carrées des classiques contributions absolues. Ce sont elles qui nous permettront de classer et de sélectionner les variables illustratives.

3.3. Un exemple d'illustration de la grille

Sur la figure 1, sont mises en évidence les positions des individus appartenant à des associations, appartenance mesurant ici le degré d'insertion sociale de la famille (variables encadrées). Il s'agit de variables *illustratives*, projetées après réalisation de la trame sous-jacente. Comme on le voit, à l'exception des associations syndicales, d'usagers et politiques, l'appartenance aux divers autres groupes ne semble pas indépendante du statut social tel que celui-ci est décrit par le premier facteur. Le test précédent nous montre que les positions de ces derniers groupes sont toutes significatives d'une répartition non aléatoire dans le plan.

Ainsi, pour prendre un exemple, les associations culturelles concernent 118 familles. L'écart-type de l'abscisse du point représentatif est donc de l'ordre de 0,09, alors que l'abscisse du point observé est de — 0,70 ; celui-ci est donc à plus de sept écarts-types de l'origine qui est la moyenne théorique.

La figure 1 nous suggère alors de prendre comme variable explicative (explication au sens statistique) le salaire du père, qui sera par exemple divisé en deux classes, compte tenu des faibles effectifs correspondant à certaines associations. On obtient ainsi le tableau suivant (qui ne concerne que les familles dont le père est salarié, et dont le salaire a effectivement été déclaré) :

TABLEAU I

Appartenance à des associations selon le salaire du père

Pourcentage (1) d'adhérents	Salaire annuel du père <30000 F.	Salaire annuel du père $\geqslant 30000$ F.
Association familiale	4.3	7.9
Association parents d'élèves	28.8	51.1
Association syndicale	18.3	18.3
Association de bienfaisance	3.4	8.1
Association politique	2.8	3.7
Association confessionnelle	5.0	13.4
Association culturelle	5.8	13.1
Effectif de l'échantillon	1306	255

(1) Pourcentage, après redressement, des familles où le père ou la mère appartiennent à une association.

4. MISE EN ŒUVRE PRATIQUE DES MÉTHODES

Ce paragraphe comprend deux parties : dans une première partie, nous examinerons les problèmes posés par la préparation des données, ainsi que la contribution de la méthode elle-même à la détection des erreurs ou anomalies. Dans la seconde partie, nous passerons brièvement en revue les procédures spécifiques de calcul impliquées par cette méthode.

4.1. Préparation et contrôle des données

Si les données de base se présentent déjà sous la forme de variables divisées en classes, le codage ne pose aucun problème particulier : si la variable q possède card J_q modalités, celles-ci seront repérées par un entier variant de 1 à card J_q. Comme nous l'avons signalé, les programmes de calcul travaillent directement sur ce type de codage, sans nécessiter de transformation en codage disjonctif. Cette transformation se fait de façon implicite à l'intérieur du programme, avec une grande économie d'opérations. Les classes d'effectifs faibles (de façon empirique, on appellera faible un effectif inférieur à 10), devront cependant, soit être agrégées à une classe, soit être réparties aléatoirement ou systématiquement dans les autres classes : la distance impliquée par cette analyse a en effet tendance à surpondérer ce type de classes. Celles-ci pourront de toute façon figurer comme éléments supplémentaires sans aucune restriction concernant les effectifs.

Si les données de base comportent des variables continues, celles-ci devront être divisées en classes. Cette opération, qui fait en théorie perdre de l'information brute (*cf.* par exemple [17]) permet au contraire, lors d'une analyse des correspondances, d'une part de valider *a posteriori* les données (en permettant d'observer l'éventuelle contiguïté des classes voisines), d'autre part de faire apparaître des liaisons non linéaires. L'exemple de la figure est à cet égard assez démonstratif. Les lignes polygonales qui décrivent les classes de salaires et les classes d'âge sont relativement régulières. Ceci prouve une certaine cohérence des données, et la pertinence de la division en classes vis-à-vis du phénomène étudié. De plus, alors que les lignes polygonales qui décrivent les âges du père et de la mère sont plutôt rectilignes, celles qui décrivent les classes de salaires sont incurvées. Nous avons déjà signalé que cela correspondait au fait que les classes les plus extrêmes avaient certaines caractéristiques en commun, en particulier celle de concerner essentiellement les familles les plus âgées. Il s'agit là d'un phénomène qui aurait été indécelable sans une division en classe fines, car les salaires auraient été alors représentés par un seul point, et l'aspect non linéaire de la liaison aurait été entièrement dissout par la contrainte de linéarité imposée aux liaisons par le codage quantitatif.

Codage optimal des variables

Comment construire des limites de classes qui soient les plus naturelles possibles, et combien de classes choisir ? Ce problème n'est évidemment pas purement technique, et nécessite la collaboration des praticiens intéressés par le contenu de l'enquête. La consultation des histogrammes est indispensable. On pourra être aidé dans cette opération par l'utilisation d'un algorithme de calcul dû à W. D. FISHER [8], qui fournit, pour une variable donnée, toutes les partitions optimales exactes (1) en 1,

(1) Il s'agit ici d'optimalité selon des critères additifs sur les classes; par exemple, minimisation de la somme des variances internes des classes.

2, 3, ..., k classes, k étant un entier fixé à l'avance (il s'agit bien d'optima exacts, et non d'optima locaux). Cependant cet algorithme est assez coûteux, et il ne peut s'appliquer raisonnablement que sur les classes d'un histogramme préalablement construit à partir des observations. (On commencera par regrouper, pour fixer les idées, 2 000 observations en 50 classes, que l'on agrégera de façon optimale en tenant compte de leur poids respectif en 2, 3, ..., 10 classes.) On choisira ensuite la partition (et par conséquent le nombre de classes) qui semble la plus adaptée au problème traité parmi les neuf partitions optimales mises en évidence par l'algorithme. Ce choix s'appuiera sur des critères statistiques (pouvoir explicatif) tout en intégrant certaines informations *a priori*.

Détection des erreurs

La construction d'une grille telle que la figure 1 constitue un test de cohérence globale, et permet donc de déceler certaines anomalies ou aberrations. Il est de plus possible de faire apparaître sous forme de variables illustratives certaines *variables techniques* relatives à la construction de l'information (groupes de familles enquêtées par la même personne, heure de l'interview, etc.). Il est assez improbable que des biais systématiques puissent passer inaperçus dans l'espace des premiers facteurs de l'analyse, qui rend compte des principales disparités existantes.

4.2. Caractéristiques de calcul

Il existe plusieurs types de procédures selon la taille des problèmes. Les programmes disponibles actuellement travaillent directement sur le tableau de codage condensé $R(s, q)$. Pour les problèmes de dimension moyenne (moins de 200 modalités de réponses à analyser, sans limite sur le nombre d'observations ni sur le nombre de variables illustratives), le tableau de BURT est calculé à partir du tableau réduit $R(s, q)$ figurant dans un support mémoire auxiliaire. Le nombre d'opérations ne dépend pas du nombre de modalités, mais seulement du nombre de questions (*cf.* [14]). La matrice à diagonaliser est ensuite réduite d'après la propriété signalée au paragraphe 2.4 *a*). La diagonalisation est ensuite effectuée par une procédure classique. L'ensemble des opérations, incluant la projection des variables illustratives, demande trois lectures du tableau condensé $R(s, q)$.

Dans le cas de données de très grandes dimensions (ou dans le cas d'un ordinateur de taille moyenne), il est préférable d'éviter la phase de calcul de la matrice à diagonaliser en utilisant des algorithmes « à lecture directe » [14]. Dans le cas des codages disjonctifs complets, l'adaptation de ces algorithmes permet de faire en sorte que l'ensemble des calculs ne dépendent pratiquement que du nombre de questions, quel que soit le nombre de modalités de réponses de chacune d'entre elles.

Il existe alors deux types de techniques utilisables : l'agrégation préalable ou les algorithmes de diagonalisation directe (puissance itérée décomposée et approximation stochastique).

Techniques procédant par agrégation préalable

Ces techniques utilisent les méthodes d'agrégation autour des centres mobiles ([1], [6]), qui, convenablement adaptées à ce type de codage, permettent de regrouper facilement les individus enquêtés en un petit nombre de classes homogènes (on regroupera par exemple 2 000 individus en 16 classes). L'analyse du tableau agrégé ne nécessite alors qu'une diagonalisation de matrice (16, 16), quel que soit le nombre des modalités de réponses. Les résultats empiriques dont nous disposons sont tout à fait encourageants Compte tenu du coût modeste de l'opération, il est toujours possible d'analyser plusieurs tableaux agrégés de différentes façons, afin d'éprouver la stabilité des résultats. La partition de l'échantillon obtenue est elle-même intéressante à interpréter : elle synthétise en effet les card Q partitions constituées par les questions analysées.

Techniques de diagonalisation directe ([13], [14])

Particulièrement intéressantes dans le cas des codages disjonctifs, ces techniques permettent de traiter des données importantes sur des ordinateurs de dimensions modestes.

RÉFÉRENCES BIBLIOGRAPHIQUES

[1] BALL (G. H.) et HALL (D. J.), *A Clustering Technique for Summarizing Multivariate Data*, Behavioral Sciences, n° 12, 1967, pp. 153-155.

[2] BENZECRI (J. P.), *Sur l'analyse des tableaux binaires associés à une correspondance multiple*, Note multigraphiée du Laboratoire de Statistique mathématique (tour 45-55, Université de Paris VI, 4, place Jussieu, 75005 Paris), 1972.

[3] BENZECRI (J. P.), *L'analyse des données*, Tome 2, L'analyse des correspondances, Dunod, Paris, 619 pages, 1973.

[4] BURT (C.), *The Factorial Analysis of Qualitative Data*, British Journal of statistical psychology, vol. III, n° 3, 1950, pp. 166-185.

[5] CARROLL (J. D.), *Generalisation of Canonical Correlation to Three or More Set of Variables*, Proc. Amer. Psy. Ass., 1968, pp. 227-228.

[6] DIDAY (E.), *La méthode des nuées dynamiques*, Revue de Statistique appliquée, vol. XIX, n° 2, 1970.

[7] ESCOFFIER-CORDIER (B.), *L'analyse des correspondances*. Thèse publiée en 1969 dans les Cahiers du B.U.R.O. n° 13, 1965.

[8] FISHER (W. D.), *On Grouping for Maximum Homogeneity*. Journal of the Amer. Statist. Assoc., n° 53, 1958, pp. 789-798.

[9] HILL (M. O.), *Correspondence Analysis : a Neglected Multivariate Method*, Applied Statist., n° 3, 1974, pp. 340-354.

[10] HORST (P.), *Relation Among m sets of Measures*, Psychometrika, n° 26, 1961, pp. 129-149.

[11] KETTENRING (J. R.), *Canonical Analysis of Several sets of Variables*, Biometrika, 58, n° 3, 1971, pp. 433-450.

[12] LEBART (L.) et TABARD (N.), *Recherches sur la description automatique des données socio-économiques*, Rapport C.O.R.D.E.S.-C.R.E.D.O.C., 234 pages, 1973.

[13] LEBART (L.), *On the BENZECRI's Method for Finding Eigenvector by Stochastic Approximation (The case of Binary Data) ; Proceeding on Computationnal Statistics*. Physica Verlag, WIEN, 1974, pp. 202-211.

[14] LEBART (L.), *Note sur l'application des algorithmes à lecture directe aux données mises sous forme disjonctive complète*, Note multigraphiée du L.S.M. (tour 45-55, Université de Paris VI, 4, place Jussieu, 75005 Paris), 1974.

[15] LECLERC (A.), *Étude de certains types de tableaux par l'analyse des correspondances. Application à une enquête de santé publique*, Thèse de 3e cycle (Université de Paris VI). Extraits publiés en 1974 dans « Proceeding on Computationnal Statistics », Physica Verlag, WIEN, 1973, pp. 212-223.

[16] MASSON (M.), *Analyse non linéaire de données*, C. R. Acad. Sc., t. 278, 11 mars 1974.

[17] NAKACHE (J. P.), *Influence du codage des données en analyse factorielle des correspondances. Étude d'un exemple pratique médical*, Revue de Statistique appliquée, vol. XXI, n° 2, 1973.

[18] TABARD (N.), *Besoins et aspirations des familles et des jeunes*. Collection Études C.A.F. n° 16, 514 pages. (C.N.A.F. : 63, boulevard Haussmann, 75008 Paris), 1974.