

REVUE DE STATISTIQUE APPLIQUÉE

M. AITKIN

B. FRANCIS

N. RAYNAL

**Une étude comparative d'analyses des correspondances
ou de classifications et des modèles de variables
latentes ou de classes latentes**

Revue de statistique appliquée, tome 35, n° 3 (1987), p. 53-81

http://www.numdam.org/item?id=RSA_1987__35_3_53_0

© Société française de statistique, 1987, 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/>

UNE ÉTUDE COMPARATIVE D'ANALYSES DES CORRESPONDANCES OU DE CLASSIFICATIONS ET DES MODÈLES DE VARIABLES LATENTES OU DE CLASSES LATENTES

M. AITKIN⁽¹⁾, B. FRANCIS⁽¹⁾ et N. RAYNAL⁽²⁾

(1) Centre for Applied Statistics,
University of Lancaster, cartmel College, Lancaster, LA1 4YG, Royaume-Uni

(2) Laboratoire de Statistique et Probabilités, U.A. — C.N.R.S. 745,
Université Paul Sabatier, 118, route de Narbonne,
31062 Toulouse Cedex, France

RÉSUMÉ

L'analyse Factorielle des Correspondances Multiples (A.F.C.M.) et les modèles de variables latentes ont pour objectif commun une réduction de la dimension des données. Dans cet article, nous comparons ces différentes méthodes sur des données qualitatives réelles préalablement simplifiées, lors d'une étude exploratoire, au moyen d'A.F.C.M. A l'issue de cette étude on voit de façon nette que le premier axe de l'A.F.C.M. et la variable continue déduite du modèle pour une seule variable latente ont une interprétation similaire. Nous mettons d'autre part en évidence la similitude des résultats d'une méthode des nuées dynamiques et de ceux du modèle de classes latentes (cas particulier du modèle pour une variable latente qualitative). Ce travail montre donc que ces techniques, bien que présentées sous des formes différentes, donnent des résultats très voisins.

Mots clés : Analyse des Correspondances, Méthode des Nuées Dynamiques, Modèle de variables latentes, Modèle de classes latentes, Comparaison de méthodes.

SUMMARY

Reduction of dimensionality is a common aim of Correspondence Analysis (C.A.) and of latent variable models. In this paper, we compare both methods on a real qualitative data set, previously simplified by means of C.A. It appears that the results related to the first C.A. factor and those of the continuous scale obtained for the one latent variable model lead to similar interpretations. It also appears that latent class analysis (a latent variable model particular case) and the « Nuées dynamiques » clustering method have similar interpretations. Thus this paper shows the apparent similarity of results arising from seemingly unrelated methods.

1. Introduction

Pour le traitement de gros fichiers de données sur variables qualitatives on peut utiliser soit des techniques d'analyse de données, soit des modèles probabilistes permettant une quantification des relations entre variables.

Nous nous sommes particulièrement intéressés ici à l'Analyse Factorielle des Correspondances Multiples (A.F.C.M.) et à la Méthode des Nuées Dynamiques

pour ce qui est des techniques d'Analyse des Données et aux modèles de variables latentes et de classes latentes pour ce qui est de la modélisation. En fait l'objectif du travail décrit ici est la comparaison de ces deux approches sur un exemple réel complexe dont la description est donnée ci-dessous.

L'A.F.C.M., bien adaptée à une étude générale et exploratoire nous a servi d'appui dans un premier temps (voir paragraphe 2) pour :

- une simplification des données ayant permis une meilleure lisibilité des résultats,
- la détermination de sous-groupes de variables liés aux objectifs fixés en début de traitement.

Cela a rendu ensuite possible, d'une part une modélisation sur des sous-groupes de variables et sur le fichier de données ainsi simplifié, d'autre part la comparaison de différentes approches. Cette étude comparative est présentée dans les paragraphes 3 à 6 de cet article.

Présentation de l'enquête

Les données proviennent d'une enquête sur la « protection sanitaire des personnes âgées » faite par le service du Professeur Albarède (Centre Hospitalier Toulouse Purpan). Le questionnaire a été proposé à 1 057 personnes âgées de plus de 60 ans dans 10 départements français; il se composait de 222 questions recouvrant de nombreux domaines :

le statut matrimonial, le statut professionnel, le niveau d'études, le mode d'habitat, la santé, les loisirs, l'autonomie, le support social et le « bien-être psychologique » des personnes.

Après élimination, codage et regroupement de variables, le fichier de données qui nous a été soumis comportait 78 variables qualitatives (comportant en tout 286 modalités et pour lesquelles il n'y avait que 371 données complètes).

La difficulté a priori de l'étude était la complexité du questionnaire. En conséquence, la réduction du nombre de variables a constitué la première partie du travail.

Objectifs

Nous nous sommes plus particulièrement intéressés aux variables de « satisfaction » ou de « bien-être psychologique » (12 variables) souhaitant :

- connaître les liens existant entre elles et déterminer les plus représentatives du « bien-être psychologique » des personnes, d'où une étude sur ces 12 variables,
- étudier les relations entre ces variables et les autres variables du fichier de données.

2. Approche exploratoire par l'analyse des correspondances

Dans la même optique que BACCINI, MONDOT et MATHIEU (1987), nous avons réalisé diverses A.F.C.M. lors de la phase exploratoire; pour cela, on a systématiquement utilisé un programme usuel d'Analyse des Correspondances sur les tableaux de BURT préalablement construits (toutes les analyses de données ont été faites sur le Centre Interuniversitaire de Calcul de Toulouse, à l'aide des logiciels SICLA et SPAD). Cette étude s'est divisée en deux parties :

- des analyses pour différents groupes de variables (nous présentons les résultats de l'A.F.C.M. relative aux variables de « satisfaction »);
- l'analyse d'un tableau réduit, obtenu à partir du tableau initial par création d'indices et regroupement de variables à l'issue d'A.F.C.M.

2.1. A.F.C.M. du tableau relatif aux variables de « satisfaction »

Nous avons 12 variables de « satisfaction » dont on trouvera la description en annexe 1; il est important de noter que chacune possède 3 modalités dont la signification est la suivante :

MOD 1 : représente l'insatisfaction relativement au sujet abordé.

MOD 2 : représente l'indécision.

MOD 3 : représente la satisfaction.

Dans un premier temps, les données manquantes ont été codées comme des modalités non-réponse (on a 932 données complètes sur ces variables). Dans ces conditions, ces modalités prennent beaucoup d'inertie sur le premier axe factoriel et rendent l'interprétation du premier plan inintéressante. Dans un second temps, nous n'avons donc traité que les données complètes et la figure 1 présente les résultats dans le premier plan factoriel.

On remarque une nette discrimination des modalités en 3 groupes : un groupe formé des modalités « satisfaction » (MOD 3) du côté positif du premier axe, un groupe formé des modalités « indécision » (MOD 2) dans le second quadrant, un groupe formé des modalités « insatisfaction » (MOD 1) dans le troisième quadrant. L'axe 1 oppose donc la « satisfaction » à « l'insatisfaction » et l'axe 2 le groupe « indécision » aux deux groupes extrêmes. Les 12 variables semblent donc mesurer un même comportement et l'analyse fournit dans le premier plan factoriel un « effet GUTTMAN » traduisant ce comportement. Il faut toutefois noter que les choses sont moins nettes au niveau individuel puisque le nuage des individus est très diffus dans le premier plan.

2.2. A.F.C.M. d'un tableau réduit

La réduction du tableau de données initial avait pour objectif une plus grande facilité d'exécution des programmes et lisibilité des résultats en vue d'une étude générale. Suite à différentes A.F.C.M. faites sur plusieurs groupes de

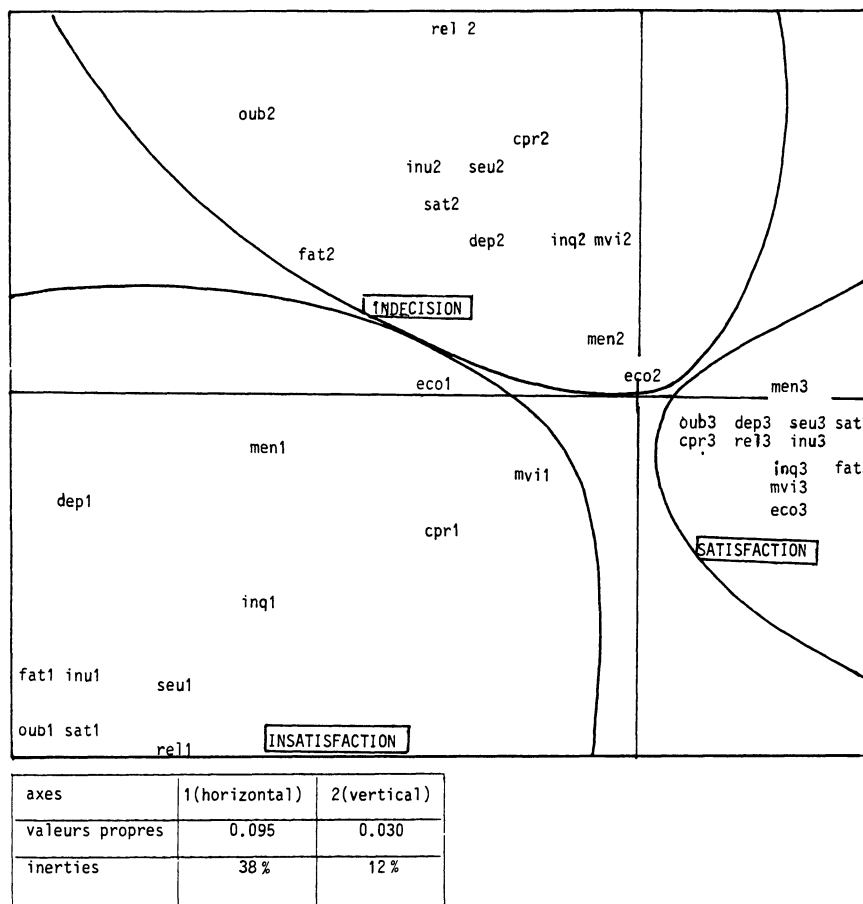


FIGURE 1

Premier plan factoriel de l'A.F.C.M. du sous-groupe des variables de « satisfaction »

variables, nous avons construit un tableau de données simplifié grâce à la création d'indices et au regroupement de variables et de modalités. Le nouveau tableau comporte 48 variables avec un total de 130 modalités (voir liste en annexe 2). Pour la même raison qu'au 2.1, seuls les individus ayant des données complètes ont été considérés dans l'analyse.

Les résultats dans le premier plan factoriel sont donnés sur la figure 2. Les modalités des variables de « satisfaction » y sont entourées, celles de variables concernant la santé et les hospitalisations y sont reliées en traits discontinus. S'il est difficile de tirer des conclusions précises de cette représentation, on peut cependant noter des tendances. De la même façon que pour l'étude précédente, les modalités des variables de « satisfaction » ont une structure en 3 groupes. Le premier axe factoriel oppose la « satisfaction » (partie négative) à l'« indécision » et à l'« insatisfaction » (partie positive). Très liées au groupe « insatisfaction »,

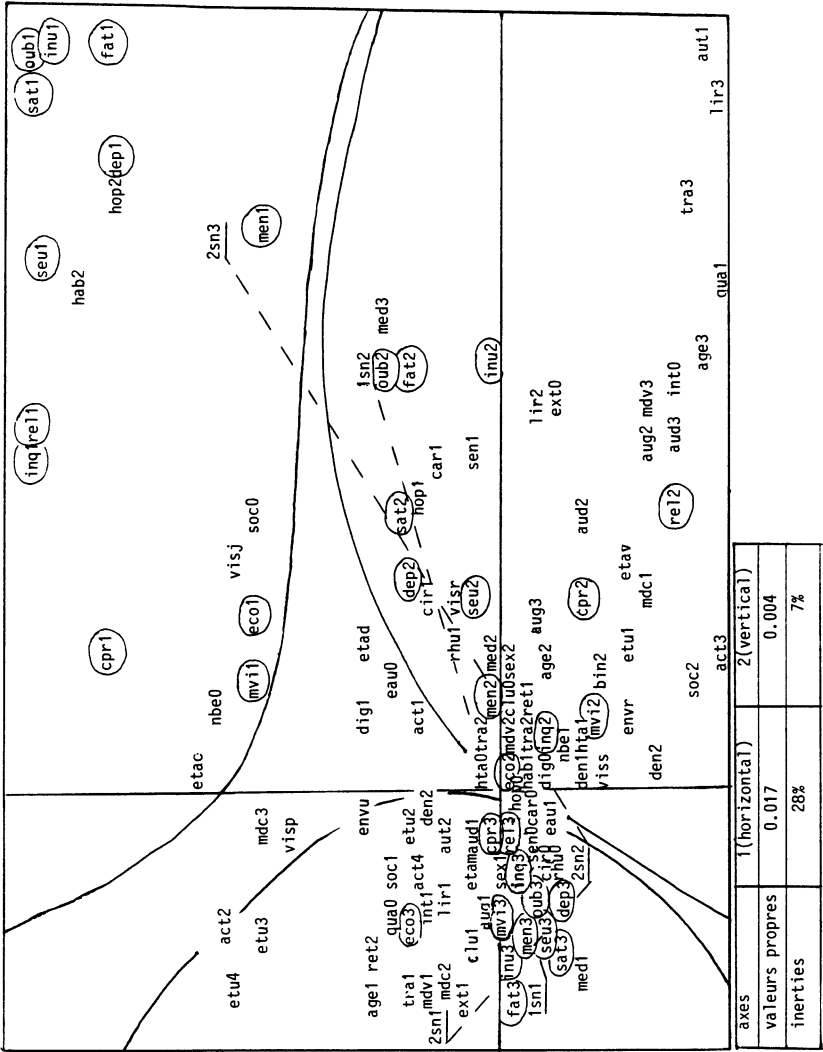


FIGURE 2
Premier plan de l'A.F.C.M. sur le tableau de données réduit

on retrouve les modalités témoignant d'une mauvaise santé, d'un âge avancé, de l'isolement et de l'inactivité. Sont corrélées avec le groupe « satisfaction » les modalités correspondant à une bonne santé et à une vie active. Le deuxième axe oppose, d'une part les modalités correspondant à un environnement rural, à de nombreuses difficultés d'environnement et à un support social insuffisant (partie négative), d'autre par les modalités traduisant un environnement urbain, peu de difficultés d'environnement et un bon support social (partie positive).

2.3. Conclusions

Un des objectifs de l'enquête était de savoir comment les variables de « satisfaction » variaient entre elles et par rapport aux autres. Les analyses précédentes ont montré que les variables de « satisfaction » semblent mesurer une même attitude dont les pôles sont bien discriminés par le premier facteur de l'A.F.C.M. Les variables sur la santé, l'âge et les loisirs semblent suivre cette même tendance. Au vu de ces résultats, nous avons tenté diverses modélisations que nous présentons dans les paragraphes suivants.

3. Cadre général des modélisations réalisées

3.1. Introduction

Le logiciel GLIM (Generalized Linear Interactive Model), très largement répandu au Royaume-Uni, permet le traitement de nombreux modèles. C'est à l'aide de cet outil que nous avons tenté, dans un premier temps, une modélisation de nos données. La modélisation implique une structure des variables en deux types : les variables réponses (ou à expliquer) et les variables explicatives.

Dans notre exemple, ce sont les 12 variables de « satisfaction » auxquelles nous avons fait jouer le rôle de variables réponses, les autres variables tenant lieu de variables explicatives. Or, actuellement, une modélisation dans GLIM est impossible dans le cas de plusieurs variables réponses. Si l'on ne s'intéresse qu'à une seule variable à expliquer sur les 12, le modèle log-linéaire s'avère également inutilisable vu les grandes dimensions des tables de contingence et leur taux de cellules vides. Notre but a donc été la réduction du nombre de variables réponses en une seule variable continue pouvant être expliquée par un modèle de régression sur les variables explicatives.

Nous avons vu au paragraphe 2 que les variables de « satisfaction » semblaient mesurer une même attitude, d'où l'idée de l'existence d'une variable latente ou d'une structure en classes latentes sous-jacentes.

Une analyse en classes latentes et une analyse factorielle binaire (cas particuliers des modèles de variables latentes) ont chacune permis la construction d'un score continu résumant les variables réponses initiales. Dans chaque cas, le score a été construit en ignorant l'effet des variables explicatives; toutefois, un modèle de régression a été ensuite utilisé pour relier le score à ces dernières et les résultats obtenus ont été comparés à ceux d'une A.F.C.M. Cette étude est présentée dans les paragraphes 4 et 6. Dans le paragraphe 5, nous avons comparé

les résultats du modèle de classes latentes à ceux de la méthode des nuées dynamiques.

3.2. Présentation générale du modèle de variables latentes

On trouvera une présentation simple et résumée des modèles de variables latentes dans EVERITT (1984).

Etant données p variables observées, corrélées entre elles, le but du modèle de variables latentes est l'extraction d'un plus petit nombre de variables non observables, appelées latentes, et expliquant la dépendance entre les variables observées.

On note

$$\underline{X} = (X_1, X_2, \dots, X_p)$$

le vecteur des variables observées et

$$\underline{Y} = (Y_1, Y_2, \dots, Y_q)$$

le vecteur des variables latentes; on suppose $q < p$ pour que le modèle ait un intérêt.

Si $\Phi(\underline{X}/\underline{Y})$ est la densité conjointe des variables observées conditionnellement aux variables latentes et $h(\underline{Y})$ la densité de $\underline{Y}$, la densité de $\underline{X}$ s'écrit :

$$f(\underline{X}) = \int \Phi(\underline{X}/\underline{Y}) h(\underline{Y}) d\underline{Y} \quad (1)$$

Remarque : le modèle de classes latentes est un cas particulier du modèle ci-dessus, lorsque les variables observées sont qualitatives et que l'on ne considère qu'une seule variable latente, elle-même qualitative. L'équation (1) s'écrit alors comme un mélange fini de produits de lois multinomiales (voir paragraphe 4).

La densité de $\underline{Y}$ conditionnellement à $\underline{X}$ s'écrit :

$$h(\underline{Y}/\underline{X}) = \frac{h(\underline{Y}) \Phi(\underline{X}/\underline{Y})}{f(\underline{X})}$$

Le problème est alors, connaissant ou supposant connue la densité de probabilité des variables observées, de déterminer $\Phi(\underline{X}/\underline{Y})$ et $h(\underline{Y})$ afin de connaître les relations entre variables latentes et variables observées (on s'intéresse dans la pratique à une quantité telle que $E(\underline{Y}/\underline{X})$ pour $\underline{Y}$ quantitative).

Pour ce faire on utilise les hypothèses suivantes :

1) Conditionnellement aux variables latentes, les variables observées sont indépendantes :

$$\Phi(\underline{X}/\underline{Y}) = \prod_{i=1}^p \Phi_i(X_i/\underline{Y})$$

Cette hypothèse est une hypothèse fondamentale du modèle, appelée « hypothèse d'indépendance conditionnelle ».

2) Les variables latentes ont des distributions identiques et sont indépendantes entre elles.

$$h(\underline{Y}) = \prod_{j=1}^q h(Y_j)$$

Cette hypothèse n'est en fait qu'une convention permettant une simplification et une meilleure interprétabilité de l'analyse.

3) Les formes des distributions des variables latentes et des variables observées conditionnellement aux variables latentes sont données a priori.

Le modèle étant ainsi spécifié, il reste à estimer les paramètres inconnus associés aux distributions.

4. Application du modèle de classes latentes

Le lecteur trouvera une présentation détaillée du modèle de classes latentes dans LAZARSELD et HENRY (1968) ou se reportera à EVERITT (1984) pour une approche plus récente.

4.1. Présentation du modèle

Nous reprenons dans ce paragraphe la présentation de HINDE (1985). Le modèle de classes latentes est obtenu dans le cas d'une seule variable latente et d'observations sur des variables qualitatives; il consiste alors à répartir la population en classes homogènes, le nombre k de classes étant fixé a priori. Nous noterons X_i ($i = 1, \dots, p$) les variables qualitatives observées et ℓ_i le nombre de modalités de X_i . Les k différentes classes latentes considérées seront notées C_r ($r = 1, \dots, k$).

Notons θ_{ij} ($i = 1, \dots, p, j = 1, \dots, \ell_i$) la probabilité pour que la variable X_i prenne la modalité j . Par commodité d'écriture, ces probabilités seront écrites sous la forme d'un vecteur $\theta : \theta = (\theta_{11}, \dots, \theta_{1\ell_1}, \dots, \theta_{p\ell_p})$. Chaque classe C_r définit une structure de probabilité différente pour les variables X_i et sera caractérisée par le vecteur de probabilités :

$$\underline{\theta}_r = (\theta_{r11}, \dots, \theta_{rij}, \dots, \theta_{rp\ell_p})$$

où

$$\theta_{rij} = P[X_i = j / C_r] \quad \text{avec} \quad \sum_{j=1}^{\ell_i} \theta_{rij} = 1.$$

La probabilité pour que le vecteur des variables observées $\underline{X} = (X_1, \dots, X_i, \dots, X_p)$ prenne la valeur $\underline{x} = (x_1, \dots, x_i, \dots, x_p)$ peut s'écrire :

$$P[\underline{X} = \underline{x}] = \sum_{r=1}^k \lambda_r P[\underline{X} = \underline{x} / C_r]$$

où $\lambda_1, \dots, \lambda_r, \dots, \lambda_k$ sont les probabilités a priori de chaque classe $\left(\sum_{r=1}^k \lambda_r = 1 \right)$.

L'hypothèse d'indépendance conditionnelle s'écrit ici :

$$\forall r = 1, \dots, k \quad P[\underline{X} = \underline{x}/C_r] = \prod_{i=1}^p P[X_i = x_i/C_r] \\ = \prod_{i=1}^p \prod_{j=1}^{\ell_i} \theta_{rij}^{\delta(x_i, j)} ;$$

où

$$\delta(x_i, j) = \begin{cases} 1 & \text{si } x_i = j \\ 0 & \text{sinon} \end{cases}$$

Le modèle s'écrit finalement :

$$P[\underline{X} = \underline{x}] = \sum_{r=1}^k \lambda_r \left[\prod_{i=1}^p \prod_{j=1}^{\ell_i} \theta_{rij}^{\delta(x_i, j)} \right]$$

Si on a N observations $\underline{x}_u = (x_{u1}, \dots, x_{ui}, \dots, x_{up})$ ($u = 1, \dots, N$) de $\underline{X}$, la vraisemblance sera :

$$L = \prod_{u=1}^N \left[\sum_{r=1}^k \lambda_r P[\underline{X} = \underline{x}_u/C_r] \right]$$

Les estimations des paramètres λ_r et θ_{rij} [($r = 1, \dots, k, i = 1, \dots, p, j = 1, \dots, \ell_i$)] seront obtenues à l'aide de l'algorithme E.M. qui est présenté au 4.2; nous pourrions ainsi estimer les probabilités d'assignation des individus aux classes à l'aide du théorème de BAYES :

$$P[C_r/\underline{X} = \underline{x}_u] = \frac{\lambda_r P[\underline{X} = \underline{x}_u/C_r]}{\sum_{s=1}^k \lambda_s P[\underline{X} = \underline{x}_u/C_s]}$$

Ces probabilités seront notées par la suite :

$$P_{ru} = P[C_r/\underline{X} = \underline{x}_u], \quad r = 1, \dots, k, \quad u = 1, \dots, N.$$

L'obtention des probabilités d'assignation des individus aux classes est un des principaux objectifs de cette méthode.

4.2. Estimation des paramètres à l'aide de l'algorithme E.M.

Les paramètres à estimer sont les λ_r et les θ_{rij} . Leur estimation à l'aide de la méthode du maximum de vraisemblance serait simple si l'on connaissait les probabilités P_{ru} d'assignation des individus aux classes. Par ailleurs, si l'on connaissait les paramètres, on pourrait calculer les probabilités d'assignation des individus aux classes (cf. 4.1).

On doit donc avoir recours à une procédure alternée d'estimation. L'algorithme utilisé ici est l'algorithme E.M., dont la formalisation, dans un contexte général, a été donnée par DEMPSTER, LAIRD et RUBIN (1977).

Algorithme E.M. dans le cas du modèle de classes latentes

En début de programme, on se donne des probabilités initiales d'assignation des individus aux classes (tirages aléatoires ou choisies a priori par l'utilisateur), ce qui tient lieu d'étape E pour la première itération ($t = 0$). Chaque itération de l'algorithme est alors formée de deux étapes, l'étape E (Expectation step) et l'étape M (Maximisation step). Etant données $(\lambda_r^{(t)}, \theta_{rij}^{(t)})$, valeurs des paramètres trouvées à l'itération t , l'itération $(t + 1)$ peut être décrite comme suit :

Etape E : A l'aide des paramètres $(\lambda_r^{(t)}, \theta_{rij}^{(t)})$ on détermine les probabilité $P_{ru}^{(t+1)}$:

$$P_{ru}^{(t+1)} = \frac{\lambda_r^{(t)} \prod_{i=1}^p \prod_{j=1}^{\ell_i} [\theta_{rij}^{(t)}]^{\delta(x_{ui,j})}}{\sum_{s=1}^k \lambda_s^{(t)} \prod_{i=1}^p \prod_{j=1}^{\ell_i} [\theta_{sij}^{(t)}]^{\delta(x_{ui,j})}} \quad r = 1, \dots, k, \quad u = 1, \dots, N$$

Etape M : Les estimateurs de maximum de vraisemblance des paramètres sont déterminés en fonction des $P_{ru}^{(t+1)}$ trouvés lors de l'étape E :

$$\lambda_r^{(t+1)} = \frac{\sum_{u=1}^N P_{ru}^{(t+1)}}{N} \quad r = 1, \dots, k$$

$$\theta_{rij}^{(t+1)} = \frac{\sum_{u=1}^N P_{ru}^{(t+1)} \delta(x_{ui}, j)}{\sum_{u=1}^N P_{ru}^{(t+1)}} \quad \begin{matrix} i = 1, \dots, p \\ j = 1, \dots, \ell_i \end{matrix}$$

Ces deux étapes sont répétées jusqu'à convergence de l'algorithme; on sait en effet que l'algorithme converge vers un maximum local de la vraisemblance (voir LAIRD, DEMPSTER et RUBIN, 1977). Les résultats dépendent des valeurs initiales. On a utilisé dans les applications une macro instruction du logiciel GENSTAT écrite par HINDE et PATE (Université de Lancaster, U.K.) ajustant le modèle de classes latentes à l'aide de l'algorithme E.M.

Ce programme nous donne en sortie :

- Les estimateurs des paramètres :

$$\hat{\lambda}_r, \hat{\theta}_{rij} (r = 1, \dots, k, i = 1, \dots, p, j = 1, \dots, \ell_i)$$

- Les assignations probabilistes des individus aux classes :

$$\hat{P}_{ru} (r = 1, \dots, k, u = 1, \dots, N)$$

- La valeur de la vraisemblance utilisée pour comparer les résultats obtenus pour diverses valeurs initiales (le nombre de classes étant fixé, on choisira les résultats correspondant à une déviance $D = -2 \log L$ minimum).

Remarques :

- Dans le cadre de la justification du nombre de classes latentes utilisées, AITKIN *et al.* (1981) montrent que l'on ne peut pas appliquer le test du χ^2 usuel à la statistique du rapport de vraisemblance pour le cas du modèle de classes latentes.

- Le programme de classes latentes peut être modifié dans le cas de données manquantes; dans ce cas, ne sont pas prises en compte, lors de l'estimation de $\theta_{rij}^{(t+1)}$, les observations ayant une donnée manquante pour la variable X_i , et l'estimation de $P_{ru}^{(t+1)}$ se fait en ignorant les variables sur lesquelles l'observation $\underline{x}_u$ a une valeur manquante. Pour plus de précisions, on se reportera à HINDE (1985).
- On a noté une extrême lenteur d'exécution de ce programme pour le traitement de gros fichiers de données.

4.3. Application aux variables de « satisfaction » et construction d'un score continu

Compte tenu des résultats d'A.F.C.M. sur les variables de « satisfaction », nous avons appliqué le modèle de classes latentes à ces variables, en considérant les modalités non-réponse comme des données manquantes. Nous avons fait plusieurs essais pour différents nombres de classes de 1 à 5; les résultats nous ont montré une grande chute de la déviance à partir de 2 classes. Il était évident qu'il fallait utiliser un modèle à 2 classes ou plus. Les A.F.C.M. nous ont naturellement amenés à considérer le modèle à 3 classes latentes. Les valeurs des $\hat{\theta}_{rij}$ et des $\hat{\lambda}_r$ sont présentées en annexe 3 (l'initialisation est faite par le programme). Au vu des résultats sur les $\hat{\theta}_{rij}$ et les $\hat{P}_{ru}$ nous avons noté un ordre naturel des classes, allant de la classe 1 représentant les individus « insatisfaits » ($\hat{\lambda}_1 = 15,5\%$) à la classe 3 représentant les individus « satisfaits » ($\hat{\lambda}_3 = 61,5\%$), la classe 2 étant une classe intermédiaire représentant les individus « indécis plutôt insatisfaits » ($\hat{\lambda}_2 = 23\%$). Cette remarque, suggérant l'existence d'une variable continue sous-jacente, nous a amenés à construire un « score de satisfaction » permettant l'évaluation de la « satisfaction » de chaque individu et à appliquer, dans un deuxième temps, un modèle de régression.

Un score continu discriminant deux classes r et r' peut être déduit, ainsi que le montrent AITKIN *et al.* (1981), de la forme des probabilités d'assignation des individus aux classes; il s'écrit :

$$d_{r,r'}(\underline{x}) = \sum_{i=1}^p \sum_{j=1}^{\ell_i} \log \frac{\theta_{rij}}{\theta_{r'ij}} \delta(x_i, j)$$

où $\underline{x} = (x_1, \dots, x_i, \dots, x_p)$ est une observation de $\underline{X} = (X_1, \dots, X_i, \dots, X_p)$.

Le « score de satisfaction » calculé dans notre analyse est $d_{3,1}(\underline{x}_u)$ ($u = 1, \dots, N$) discriminant les classes 3 et 1 représentant respectivement les individus « satisfaits » et « insatisfaits ».

Les valeurs des $\left[\log \frac{\hat{\theta}_{3ij}}{\hat{\theta}_{1ij}} \right]$ ($i = 1, \dots, 12, j = 1, \dots, 3$) sont données en annexe 4.

4.4. Application d'un modèle de régression au score obtenu

Nous avons utilisé un modèle de régression du score obtenu ci-dessus sur les variables explicatives. Il s'agit en fait d'une analyse de la variance à 36 facteurs

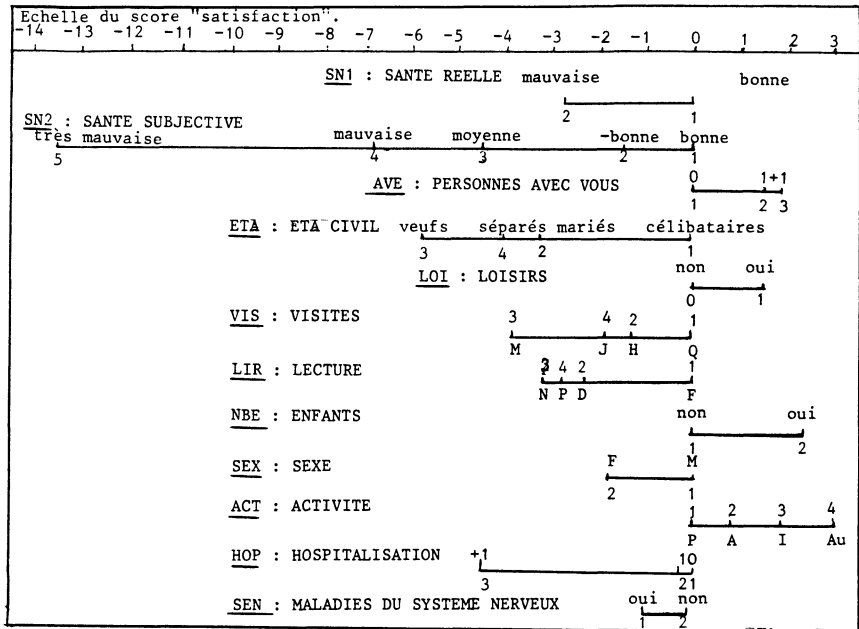


FIGURE 3

Estimateurs des effets principaux, des variables explicatives du modèle de régression sur l'échelle du « score de satisfaction »

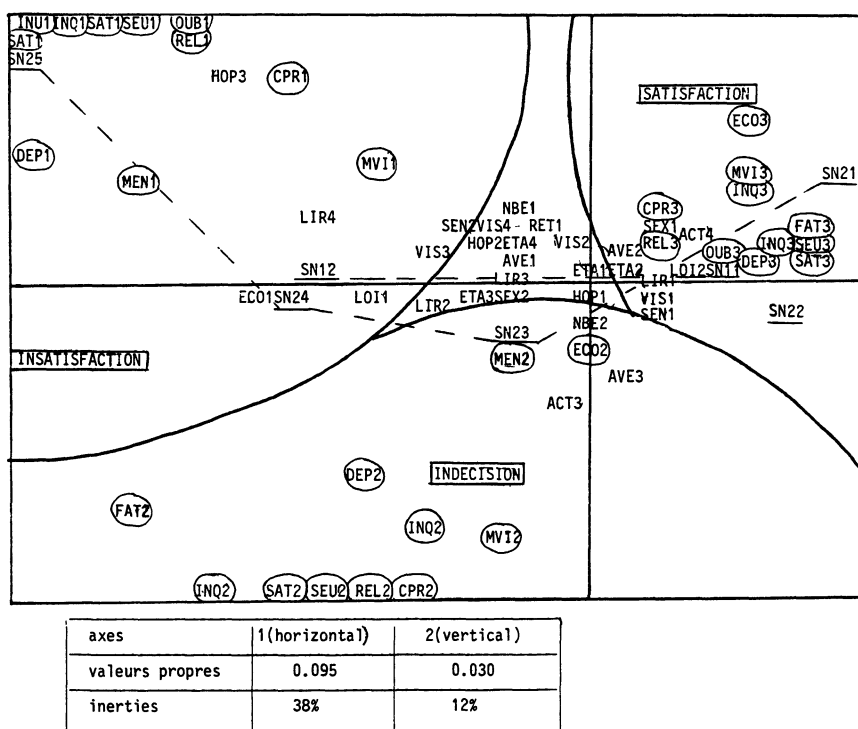
(voir 2.2.), dans laquelle seuls les effets principaux ont été considérés. Le nombre de variables explicatives étant important, nous n'avons retenu que celles dont les effets principaux entraînaient les plus grandes chutes de la déviance relativement à la diminution du degré de liberté (statistique F). Ce sont les douze variables suivantes :

- 1 — Santé réelle : SN1
- 2 — Santé subjective : 2SN
- 3 — Nombre de personnes vivant avec vous : AVE
- 4 — Statut matrimonial : ETA
- 5 — Loisirs : LOI
- 6 — Visites : VIS
- 7 — Lecture : LIR
- 8 — Enfants : NBE
- 9 — Sexe : SEX
- 10 — Activité professionnelle : ACT
- 11 — Hospitalisations : HOP
- 12 — Maladies du système nerveux : SEN

Les valeurs des estimateurs des paramètres du modèle ajusté (effets des facteurs) ont été représentées suivant l'échelle du score (voir Fig. 3). On visualise ainsi le sens et l'ampleur des influences des variables explicatives sur le score. L'effet de la variable « santé subjective » est le plus important.

4.5. Comparaison avec les résultats de l'A.F.C.M.

On a réalisé ici une A.F.C.M., les variables de « satisfaction » étant prises comme variables « actives » et les 12 variables explicatives du modèle de régression comme variables supplémentaires. La figure 4 donne le résultat de cette A.F.C.M. dans le premier plan principal. Les modalités de la variable SN2 sont reliées entre elles sur le graphique, et les modalités des variables de « satisfaction » ont été entourées. L'interprétation que l'on peut donner de ce graphique en ce qui concerne l'influence des modalités des variables explicatives est comparable à celle faite avec le modèle de régression. De plus, sur la figure 5, on a reporté les coordonnées des variables explicatives sur l'échelle du premier axe factoriel, ce qui nous a donné un graphique tout à fait semblable à la figure 3 réalisée à partir de l'échelle du score. On constate que les deux variables sous-jacentes obtenues par deux techniques différentes, le score à partir du modèle de classes latentes et le premier axe de l'A.F.C.M., sont en réalité très proches.



remarque : les modalités INU1, INQ1, SAT1, SEU1, OUB1, INQ2, SAT2, SEU2, REL2, CPR2 ont été ramenées dans le cadre du graphique.

FIGURE 4

Premier plan de l'A.F.C.M. pour les variables de « satisfaction » actives et les variables explicatives du modèle de régression représentatives

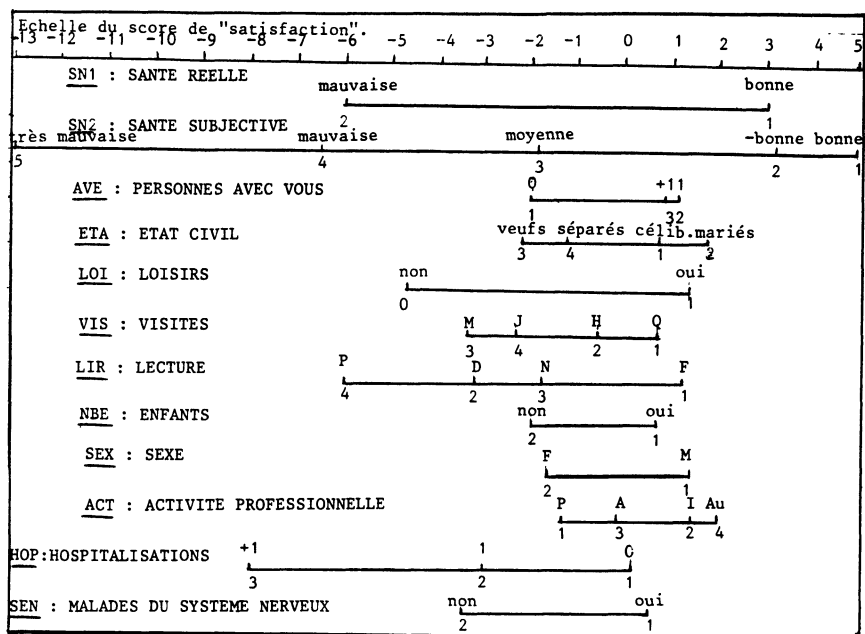


FIGURE 5

Coordonnées des modalités des variables explicatives sur le 1^{er} axe factoriel de l'A.F.C.M.

Remarques :

— On a cependant entre les figures 3 et 5 des inversions des modalités 1 et 2 de la variable ETA et 2 et 3 de la variable LIR.

— Les modalités de la variable « santé subjective » (SN2) ont la même allure que les modalités des variables de « satisfaction », ceci étant sans doute lié à la nature « subjective » de la définition.

5. Comparaison entre le modèle de classes latentes et la méthode des nuées dynamiques

La méthode des nuées dynamiques (M.N.D.) et le modèle de classes latentes ayant pour objectif commun la répartition de la population en classes homogènes, nous avons comparé leurs résultats sur le groupe des variables de « satisfaction ».

Nous avons utilisé la M.N.D. (voir DIDAY et al (1979)) dans le cas où les données sont les tableaux de profils de réponse des individus relativement aux modalités des variables, la métrique étant la métrique du χ^2 et les noyaux les centres de gravité des classes. Le but de la méthode est la répartition de la population en un nombre fixé de classes. Le critère à maximiser est l'inertie inter-classe.

La M.N.D. a été appliquée sur les 932 individus ayant des données complètes sur les variables de « satisfaction ». Trois classes ont été demandées et les résultats ont été comparés au modèle de classes latentes (cf. 4.2).

Comparaison des résultats

Le pourcentage d'inertie expliquée dans la M.N.D. est de 17 %. Les interprétations des classes et leur pourcentage sont comparables pour les deux méthodes. On a en effet :

	classes latentes	M.N.D.
classe 1	individus « insatisfaits »	
classe 2	individus « indécis plutôt insatisfaits »	
classe 3	individus « satisfaits »	
classe 1	$\hat{\lambda}_3 = 15,5 \%$	pct = 11
classe 2	$\hat{\lambda}_2 = 23 \%$	pct = 21,3
classe 3	$\hat{\lambda}_3 = 61,5 \%$	pct = 67,7

Dans l'annexe 3, sont reportés les $\hat{\theta}_{rij}$ estimés du modèle de classes latentes ainsi que les fréquences conditionnelles f_{rij} calculées à partir des résultats de la classification :

$$f_{rij} = \frac{\text{eff}[r; (i, j)]}{\text{eff}[r]} \quad \begin{array}{l} r = 1, \dots, 3 \\ i = 1, \dots, 12 \\ j = 1, \dots, 3 \end{array}$$

où $\text{eff}[r; (i, j)]$ est l'effectif de la modalité j de la variable X_i dans la classe r , et $\text{eff}[r]$ est l'effectif de la classe r .

On voit que ces valeurs sont du même ordre de grandeur.

Assignations probabilistes des individus aux classes

Il est possible d'obtenir, à la suite d'une classification, des assignations probabilistes des individus aux classes, à l'aide du théorème de BAYES, en supposant les variables indépendantes à l'intérieur des classes :

$$P[C_r / \underline{X} = \underline{x}] = \frac{\alpha_r \prod_{i=1}^p P[X_i = x_i / C_r]}{\sum_{s=1}^k \alpha_s \prod_{i=1}^p P[X_i = x_i / C_s]}$$

où $\underline{x}$ est une observation du vecteur des variables X_i ,

α_r les proportions des classes C_r et k le nombre de classes.

Les $P[X_i = j / C_r]$ étant estimées par f_{rij} , il vient :

$$P[C_r / \underline{X} = \underline{x}] = \frac{\alpha_r \prod_{i=1}^p \prod_{j=1}^{\ell_i} f_{rij}^{\delta(x_i, j)}}{\sum_{s=1}^k \alpha_s \prod_{i=1}^p \prod_{j=1}^{\ell_i} f_{sij}^{\delta(x_i, j)}}$$

Les probabilités d'assignation trouvées par cette méthode et celles obtenues par le modèle de classes latentes sont données, pour quelques individus pris au hasard dans le fichier de données, dans le tableau ci-après.

Probabilités d'assignation de quelques individus aux classes pour deux méthodes différentes

Profils	Classes latentes			Nuées dynamiques		
	1	2	3	1	2	3
333333332333	0	0.001	0.999	0	0.002	0.998
233322332332	0.005	0.620	0.375	0.002	0.361	0.637
332322332332	0.009	0.849	0.142	0.004	0.746	0.250
312223222331	0.138	0.862	0	0.015	0.985	0
32223222322	0.006	0.992	0.002	0.003	0.994	0.003
33122332332	0.075	0.863	0.062	0.026	0.888	0.086

On notera que, dans les deux cas, les résultats dépendent des tirages initiaux (assignations probabilistes des individus aux classes pour le modèle de classes latentes et partition initiale pour la M.N.D.). Les algorithmes de la M.N.D. étant beaucoup plus performants que ceux du modèle de classes latentes, on pourrait utiliser les fréquences conditionnelles calculées à la suite d'une M.N.D., comme valeurs initiales pour le programme de classes latentes. Des développements sur la comparaison des modèles mathématiques sous-tendant les deux méthodes sont actuellement étudiés.

Enfin, signalons que des programmes de classification utilisant la M.N.D. ont été élaborés pour l'analyse de mélanges de distributions, le critère à minimiser étant le produit des vraisemblances associées aux échantillons définis par la partition. Ceci a été présenté dans le cas gaussien par SHROEDER (1976) et des exemples ont été traités dans le cas où l'on ne tenait pas seulement compte de la partition mais aussi des degrés d'appartenance des individus aux classes par CAZES (1976). L'algorithme de la M.N.D. présenté par SHROEDER (1976) appliqué au cas d'un mélange de distributions produits de lois multinomiales est similaire, si l'on considère les probabilités d'assignation des individus aux classes, à l'algorithme E.M. du modèle de classes latentes.

6. Application de l'analyse factorielle binaire

Le score obtenu à partir du modèle de classes latentes (cf. 4.3) est un peu artificiel, dans la mesure où il représente une quantification d'une variable qualitative. Une « analyse factorielle qualitative » (dont le modèle est un cas particulier du modèle de variables latentes lorsque ces dernières sont continues et les données qualitatives) pour un seul facteur (variable latente continue) serait plus naturelle pour arriver à un tel résultat. On est cependant limité par un problème de mise en œuvre informatique puisque le programme n'est actuellement opérationnel que sur un nombre réduit de variables binaires (programme « FACONE » d'analyse factorielle binaire pour un facteur, B.L. SHEA, LONDON SCHOOL OF ECONOMICS, 1984). Pour l'application à notre exemple, une réduction et un recodage des variables (d'où une perte d'information) étaient nécessaires; cela nous a néanmoins paru intéressant dans la perspective de comparaison des

méthodes qui était la nôtre. Pour plus de détails sur l'analyse factorielle binaire, on pourra se reporter à BARTHOLOMEW (1980) et BOCK et AITKIN (1981).

Remarque :

Un nouveau programme d'analyse factorielle qualitative pour un facteur (les variables observées pouvant avoir plus de deux modalités) sera bientôt disponible : programme « POLYFAC » (B.L. SHEA, LONDON SCHOOL OF ECONOMICS).

6.1. Modèle

On note $X_1, \dots, X_i, \dots, X_p$ les variables aléatoires binaires observées ($X_i = 0$ ou $1, i = 1, \dots, p$).

BARTHOLOMEW (1980) choisit une variable latente Y suivant une loi uniforme sur $[0, 1]$.

En notant $\Pi_i(y) = \text{Prob}[X_i = 1/Y = y]$, il montre que, devant vérifier certaines propriétés, ces probabilités conditionnelles sont telles que :

$$G(\Pi_i(y)) = \alpha_{i0} + \alpha_{i1} H(y) \quad (2)$$

où α_{i0} et α_{i1} sont les paramètres du modèle associés à la variable X_i ($i = 1, \dots, p$), G et H sont des fonctions logit ou probit (si $0 < x < 1$, $\text{logit}(x) = \log \frac{x}{1-x}$ et $\text{probit}(x) = \Phi^{-1}(x)$, où Φ est la fonction de répartition de la loi normale réduite).

Les variables X_i ($i = 1, \dots, p$) étant observées sur N individus, les observations peuvent s'écrire sous la forme matricielle suivante :

$$X = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ \vdots & & \vdots \\ X_{N1} & \dots & X_{Np} \end{pmatrix} = \begin{pmatrix} \underline{X}_1 \\ \vdots \\ \underline{X}_N \end{pmatrix}$$

On suppose qu'il existe s différents profils de réponse sur les observations ($s \leq \min(N, 2^p)$) notés

$$\underline{x}_k \quad (k = 1, \dots, s) \quad (\underline{x}_k = (x_{k1}, \dots, x_{ki}, \dots, x_{kp}))$$

et de fréquences respectives $r_k \left(\sum_{k=1}^s r_k = N \right)$.

On note $P_k = P[\underline{X} = \underline{x}_k] \quad (k = 1, \dots, s)$.

Les fréquences r_k sont les observations d'un loi multinomiale de paramètres N et P_k dont le logarithme de la vraisemblance s'écrit :

$$\text{Log } L = C + \sum_{k=1}^s r_k \text{Log } P_k$$

(C ne dépend pas des paramètres α_{i0} et α_{i1} ($i = 1, \dots, p$) à estimer).

On a :

$$P_k = P [\underline{X} = \underline{x}_k] = \int_0^1 P [\underline{X} = \underline{x}_k / y] dy \quad (3)$$

Dans les applications, on a choisi $G = \text{logit}$ et $H = \text{probit}$; pour ce modèle, l'intégrale (3) s'écrit, en posant $y = \Phi(\theta)$ et en notant φ la densité de la loi normale réduite :

$$P_k = P [\underline{X} = \underline{x}_k] = \int_{-\infty}^{+\infty} P [\underline{X} = \underline{x}_k / \theta] \varphi(\theta) d\theta \quad (4)$$

D'après l'hypothèse d'indépendance conditionnelle, on a :

$$P [\underline{X} = \underline{x}_k / \theta] = \prod_{i=1}^p [\Pi_i(\theta)]^{x_{ki}} [1 - \Pi_i(\theta)]^{1-x_{ki}}$$

En approchant l'intégrale (4) par la formule de quadrature de Gauss Hermite, il vient :

$$\begin{aligned} \tilde{P}_k &\cong \sum_{m=1}^q L_k(Z_m) A(Z_m) \\ \text{où } L_k(Z_m) &= \prod_{i=1}^p [\Pi_i(Z_m)]^{x_{ki}} [1 - \Pi_i(Z_m)]^{1-x_{ki}} \end{aligned} \quad (5)$$

Les Z_m ($m = 1, \dots, q$) sont appelés points de quadrature et les $A(Z_m)$ ($m = 1, \dots, q$) poids de quadrature; ces valeurs dépendent, toutes deux, de la densité $\varphi(\theta)$ de la variable latente θ et sont données par le programme.

Nous nous intéressons aux valeurs prises par la variable latente. Cette variable n'étant pas observée, les valeurs qu'elle prend pour les individus u ($u = 1, \dots, N$) seront remplacées par leur espérance conditionnelle aux observations $\underline{x}_u$ ($u = 1, \dots, N$); φ étant la densité de la variable latente θ , on a, grâce au théorème de BAYES :

$$\varphi(\theta / \underline{x}_u) = \frac{P[\underline{X} = \underline{x}_u / \theta] \varphi(\theta)}{P[\underline{X} = \underline{x}_u]}$$

On peut alors vérifier que l'espérance de la variable latente conditionnellement aux profils observés $\underline{x}_k$ ($k = 1, \dots, s$) peut être approchée de la façon suivante (voir BOCK et AITKIN, 1981) :

$$E(\theta / \underline{x}_k) \cong \frac{\sum_{m=1}^q Z_m L_k(Z_m) A(Z_m)}{\tilde{P}_k} \quad k = 1, \dots, s$$

$E(\theta / \underline{x}_k)$ ($k = 1, \dots, s$), est appelée le « θ -score ».

Dans la pratique, on doit estimer les paramètres α_{i0} et α_{i1} ($i = 1, \dots, p$); on en déduit des estimations des $\tilde{P}_k$ ($k = 1, \dots, s$) à partir de (2) et (5), ainsi que du θ -score ci-dessus.

Estimation des paramètres à l'aide de l'algorithme E.M.

On cherche donc à estimer les α_{i0} et α_{i1} ($i = 1, \dots, p$). La procédure de l'algorithme E.M. pour l'analyse factorielle binaire est pour l'essentiel décrite dans BOCK et AITKIN (1981).

Si l'on suppose que les individus forment q groupes homogènes représentant chacun les valeurs $Z_1, \dots, Z_m, \dots, Z_q$ de la variable latente, on peut écrire les équations de vraisemblance des paramètres α_{i0} et α_{i1} ($i = 1, \dots, p$) en fonction des valeurs r_{im} et N_m ($i = 1, \dots, p$; $m = 1, \dots, q$) qui sont définies de la façon suivante :

— r_{im} est la fréquence des observations représentant la valeur Z_m de la variable et pour lesquelles la variable X_i prend la valeur 1. Connaissant x_{ki} et $\Pi_i(Z_m)$, on a :

$$r_{im} = \sum_{k=1}^s \frac{r_k x_{ki} L_k(Z_m) A(Z_m)}{\tilde{P}_k} \quad (6)$$

$$\left[\sum_{m=1}^q r_{im} = r_i, \text{ nombre d'observations pour lesquelles } X_i = 1 \right]$$

— N_m est la taille de l'échantillon représentant la valeur Z_m de la variable latente. On a :

$$N_m = \sum_{k=1}^s \frac{r_k L_k(Z_m) A(Z_m)}{\tilde{P}_k} \quad (7)$$

$$\left[\sum_{m=1}^q N_m = N, \text{ taille totale de l'échantillon} \right]$$

Pour débiter l'algorithme E.M. on donne aux paramètres α_{i0} et α_{i1} des valeurs initiales, ce qui tient lieu d'étape E pour la première itération ($t = 0$). Les étapes E et M de l'algorithme peuvent être décrites comme suit :

Etape E :

Connaissant les valeurs courantes des paramètres α_{i0} et α_{i1} , on calcule :

$$\tilde{P}_k = \sum_{m=1}^q L_k(Z_m) A(Z_m) \quad k = 1, \dots, s$$

$\hat{r}_{im}$: la valeur estimée de r_{im} (voir (6)).

$\hat{N}_m$: la valeur estimée de N_m (voir (7)).

Etape M :

Connaissant $\hat{r}_{im}$ et $\hat{N}_m$ ($i = 1, \dots, p$, $m = 1, \dots, q$) calculés à l'étape E, on estime les paramètres α_{i0} et α_{i1} satisfaisant les équations de vraisemblance (voir BOCK et AITKIN, 1981).

Afin de légitimer le modèle, on utilise la statistique de test :

$$G^2 = 2 \left(\sum_{k=1}^s \hat{r}_k \text{Log} \frac{\hat{r}_k}{N\tilde{P}_k} \right) \text{ qui est à comparer à un } \chi_v^2$$

$$\begin{aligned} \text{ou } v &= s - 2p & \text{si } 2p < s < 2^p \\ &= s - 2p - 1 & \text{si } s = 2^p \end{aligned}$$

Remarque :

Lorsque le nombre de variables p devient grand, la statistique de test devient instable et donc inutilisable.

Programme :

On utilise le programme FACONE qui ajuste le modèle (2) en utilisant l'algorithme E.M. décrit ci-dessus dans les cas où :

$$\begin{array}{ll} G = \text{logit} & \text{et} \quad H = \text{logit : modèle logit,} \\ G = \text{logit} & \text{et} \quad H = \text{probit : modèle logit/probit,} \\ G = \text{probit} & \text{et} \quad H = \text{probit : modèle probit.} \end{array}$$

On a utilisé dans les applications le modèle (logit/probit).

Ce programme nous donne en sortie :

- les fréquences estimées des profils : $\hat{r}_k$ ($k = 1, \dots, s$),
- les « θ -score » (θ : variable latente normale réduite) défini par les espérances conditionnelles : $\theta_k = E(\theta/\tilde{x}_k)$ ($k = 1, \dots, s$),
- le « y -score » (Y : variable latente uniforme sur $[0, 1]$) défini par les espérances conditionnelles : $E(Y/\tilde{x}_k)$ ($k = 1, \dots, s$),
- la statistique de test : $G^2 = 2 \left(\sum_{k=1}^s \hat{r}_k \text{Log} \frac{\hat{r}_k}{N\tilde{P}_k} \right)$.

6.2. Application à une partie des variables de satisfaction

Nous souhaitons appliquer le modèle précédent aux 12 variables de satisfaction qui possèdent chacune 3 modalités (voir annexe 1). Dans ce but, ces variables ont été recodées en variables binaires par regroupement de deux de leurs modalités. Notre choix a été guidé par les résultats de classifications des individus, caractérisés par les 12 variables, en deux classes (résultats non reportés ici) soit à l'aide du modèle de classes latentes, soit à l'aide de la méthode des nuées dynamiques. L'interprétation des classes est alors la suivante : une des classes représente les individus « satisfaits » et l'autre représente les individus « indécis insatisfaits ». Nous avons donc choisi le regroupement des modalités indécision (MOD 2) et insatisfaction (MOD 1), comme étant la transformation la plus fidèle aux données. Il n'a pas été possible d'utiliser le programme FACONE sur l'ensemble des 12 variables de satisfaction puisqu'on avait alors $2^{12} = 4\,096$ profils possibles dont seulement 424 sont observés. Nous avons donc fait des essais sur plusieurs sous-groupes de variables. Nous n'avons retenu que 5 variables, ce nombre étant

le nombre maximum pour lequel on était susceptible d'obtenir un bon ajustement du modèle; les variables conservées sont les suivantes :

- CPR : carrière professionnelle,
- ECO : situation économique,
- REL : relations humaines,
- MVI : recommenceriez-vous la même vie ?
- INQ : inquiétude par rapport à l'avenir.

29 profils différents ont été observés. On a appliqué le programme FACONE sur ces variables, dans le cas du modèle logit/probit. La statistique de test G^2 vaut 15,05 pour 12 degrés de liberté, le modèle est donc acceptable. Les résultats sont présentés en annexe 5.

6.3. Application d'un modèle de régression au score obtenu

De la même façon qu'au 4.4., on peut relier le score obtenu à l'issue du programme FACONE ($E(\theta/\underline{x}_k)$) ($k = 1, \dots, s$) aux variables explicatives en utilisant un modèle de régression (analyse de la variance), prenant par exemple en compte les seuls effets principaux des variables :

- la variable réponse est le score $\theta_k = E(\theta/\underline{x}_k)$ $\theta \sim N(0, 1)$,
- les variables explicatives retenues (critères analogues à ceux du paragraphe 4.4) sont les suivantes :
 - 1 — SN1 : santé réelle,
 - 2 — TRA : travaillez-vous ?
 - 3 — ACT : activité professionnelle,
 - 4 — AFF : êtes-vous affecté ?
 - 5 — LOI : avez-vous des loisirs ?
 - 6 — QUA : sortez-vous du quartier ?
 - 7 — ENV : environnement.

De façon identique au 4.4, on reporte sur l'échelle du score les estimations des effets principaux des variables explicatives (voir Fig. 6). Les variables « santé

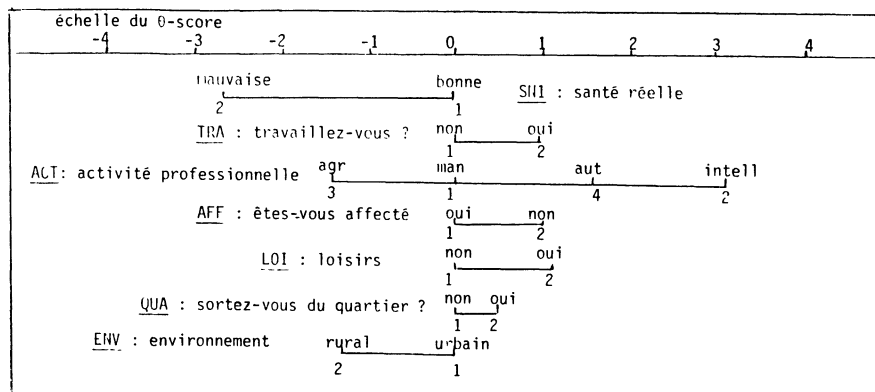


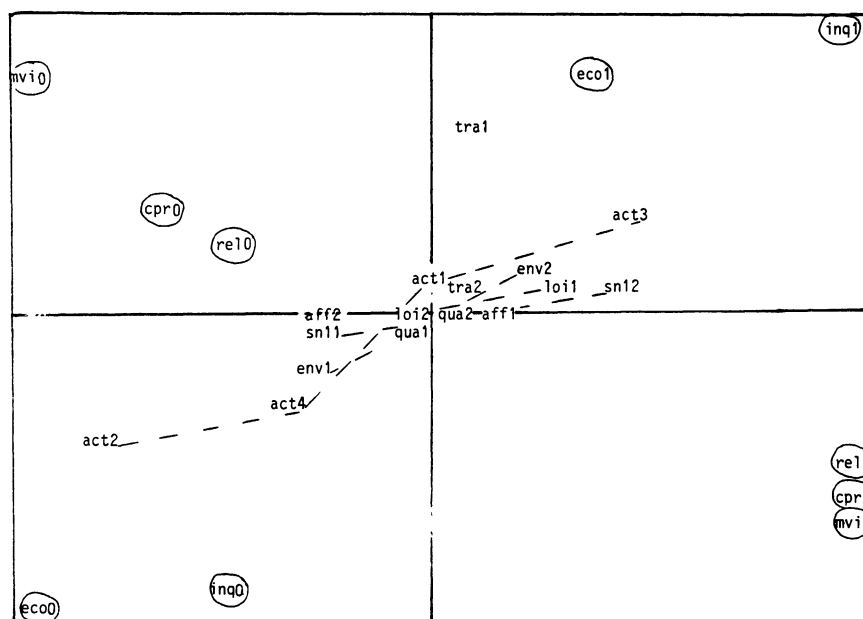
FIGURE 6

Estimateurs des effets principaux des variables explicatives du modèle de régression sur l'échelle du θ -score

réelle » et « activité professionnelle » sont celles dont l'influence sur le score semble la plus importante.

6.4. Comparaison avec l'A.F.C.M.

On a réalisé une A.F.C.M. avec les 5 variables de satisfaction précédemment choisies, recodées en binaires et prises comme variables actives, et les 7 variables du modèle de régression en supplémentaires. Le premier plan principal est représenté sur la figure 7. On constate que le premier axe principal discrimine bien les modalités 0 (insatisfaction-indécision) des modalités 1 (satisfaction) des variables de satisfaction, ainsi que les modalités de la variable « santé réelle : SN1 » et « activité professionnelle : ACT » ; on retrouve donc des résultats voisins de ceux fournis par le modèle de régression appliqué à la variable latente. Afin d'illustrer cela, on a calculé le coefficient de corrélation entre la variable SCORE



axes	1(horizontal)	2(vertical)
valeurs propres	0.32	0.16
inerties	34%	19%

Remarque : les modalités : mvi0, eco0, inq1, rel1, cpr1, mvi1 ont été reportées dans le cadre du graphique.

FIGURE 7

Représentation du premier plan factoriel de l'A.F.C.M. pour les 5 variables de satisfaction actives et les variables explicatives du modèle de régression prises en représentatives

Remarque : les modalités des variables de satisfaction sont entourées
mod 0 : indécision ou insatisfaction
mod 1 : satisfaction.

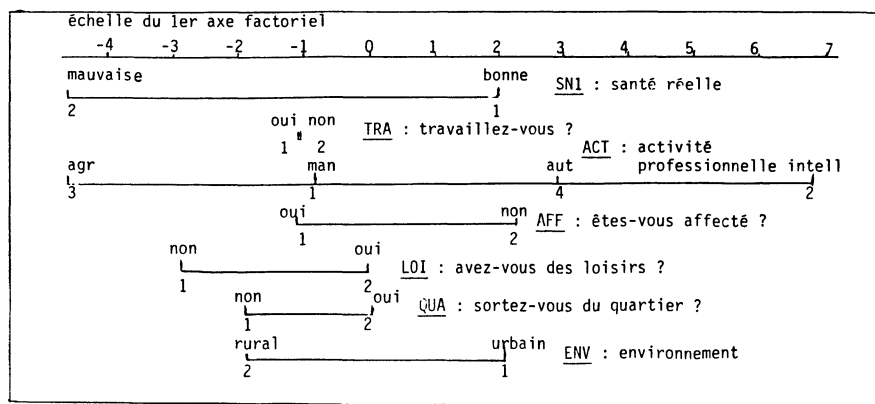


FIGURE 8

Coordonnées des variables explicatives du modèle de régression sur le premier axe de l'A.F.C.M.

(déterminée par ses valeurs pour les 29 profils observés sur les 5 variables de satisfaction) et le premier axe de l'A.F.C.M. FACT (déterminé par les coordonnées de ces 29 profils observés sur cet axe). On a : $\text{corr}(\text{SCORE}, \text{FACT}) = 0,98$.

Ces deux variables sont donc très fortement corrélées.

7. Conclusion

D'une façon générale, les méthodes d'Analyse des Données, telles que l'A.F.C.M. ou la classification automatique, sont les mieux adaptées à l'étude globale et exploratoire de gros fichiers de données. Rapides et efficaces, elles permettent le tri des données, le recodage et le regroupement des variables dans le cadre d'une première approche.

Dans un fichier simplifié, les méthodes d'Analyse de Données et de Modélisation deviennent comparables, voire concurrentes.

Dans le cas particulier qui a été présenté ici, la comparaison entre l'A.F.C.M. et les modèles de classes et de variable latente fait apparaître une grande similitude entre le premier axe de l'A.F.C.M. et les scores obtenus à partir des différents modèles (des conclusions similaires ont été obtenues sur un autre exemple par SAINT-PIERRE, FRANCIS et AITKIN, 1986). Cette conclusion n'a pu être étendue au cas de plusieurs variables latentes, faute de programmes disponibles pour ce type de méthodes.

Sur les données relatives à la « satisfaction », la classification automatique et le modèle de classes latentes nous donnent également des résultats très comparables.

Les objectifs de l'Analyse des Données et de la Modélisation étant souvent les mêmes, l'utilisateur préférera la méthode la plus rapide et la plus commode. Pour le type de données que l'on vient de décrire, les modèles de variables latentes

manquent pour l'instant d'outils de calcul adéquats, mais cette situation semble devoir évoluer dans un avenir très proche et les modèles probabilistes pourront alors être préférés, le cas échéant, à des techniques descriptives de l'Analyse des Données.

Remerciements

Nous tenons à remercier le service du Pr. ALBRARÈDE pour nous avoir transmis les données sur la Protection Sanitaire des personnes âgées.

Nous remercions également A. BACCINI, H. CAUSSINUS et J. SAINT-PIERRE pour leur collaboration dans la rédaction de cet article.

Bibliographie

- M. AITKIN, A.J. ANDERSON and J. HINDE (1981). — Statistical modelling of data in teaching styles. *J. Roy. Statistical Soc. série A*, Vol. 1-44, pp. 443-459.
- M. AITKIN, B. FRANCIS et J. ST-PIERRE (1985). — Comparaison de modélisations et d'analyses de correspondances sur un questionnaire relatif à la camaraderie. *Publication du laboratoire de Statistique et Probabilités*, n° 02.86 Mai 1986, Université Paul Sabatier (Toulouse).
- D. BARTHOLOMEW (1980). — Factor Analysis for categorical data. *J. Roy. Statistic. Soc. série B*, Vol. 42, pp. 293-321.
- D. BOCK and M. AITKIN (1981). — Marginal Maximum Likelihood estimation of item parameters : an application of an E.M. algorithm. *Psychometrika*, n° 46, pp. 443-459.
- A. BACCINI, A.M. MONDOT et J.R. MATHIEU (1987). — Comparaison sur un exemple de l'A.F.C.M. et de modélisations. *Revue de Statistique Appliquée*, Vol. 35, n° 3, 21-34.
- P. CAZES (1976). — Décomposition d'un histogramme en composantes gaussiennes. *Revue de Statistique Appliquée*, Vol. XXIV, n° 1.
- A. DEMPSTER, N. LAIRD and D. RUBIN (1977). — Maximum likelihood from incomplete data via the E.M. algorithm. *J. Roy. Statistical Soc., série B*, Vol. 39, pp. 1-38.
- DIDAY et collaborateurs (1979). — Optimisation en classification automatique. Publié par l'INRIA (2 tomes). Rocquencourt.
- B. EVERITT (1984). — *An introduction to latent variable models*. Chapman & Hall, Londres.
- J. HINDE (1985). — *Descriptive classification of Cetacea : a latent class solution*. Data Analysis in real life environment : ins and outs of solving problems. J.F. Marco-torchino, J.M. Proth and J. Janssen (Eds), Elsevier Science Publishers B.V. (North Holland).
- P. LAZARSFELD and N. HENRY (1968). — *Latent Structure Analysis*. Houghton Mufflin company, Boston.
- N. RAYNAL (1987). — Contribution à l'étude comparative de diverses méthodes statistiques d'analyse exploratoire et de modélisation. Thèse de 3^e cycle, Université Paul Sabatier, Toulouse.
- A. SCHROEDER (1976). — Analyse d'un mélange de distributions de probabilité de même type. *Revue de Statistique Appliquée*, Vol. XXIV, n° 1.

Annexe 1

Description des variables de satisfaction

- 1 — men : souffrez-vous d'un manque d'énergie ?
- 2 — dep : vous sentez-vous déprimé ?
- 3 — seu : vous sentez-vous seul ?
- 4 — oub : vous sentez-vous oublié ?
- 5 — inu : vous sentez-vous inutile ?
- 6 — sat : êtes-vous satisfait de votre situation actuelle ?
- 7 — fat : êtes-vous fatigué de la vie ?
- 8 — inq : êtes-vous inquiet pour l'avenir ?
- 9 — éco : êtes-vous satisfait de la situation économique ?
- 10 — cpr : êtes-vous satisfait de votre carrière professionnelle ?
- 11 — rel : êtes-vous satisfait de vos relations humaines ?
- 12 — mvi : recommenceriez-vous la même vie ?

Annexe 2

Description des variables de l'A.F.C.M. faite sur le tableau de données réduit

<i>identificateur</i>	<i>libellé</i>	<i>modalités</i>
sex	sexe	<i>sex 1</i> : M. <i>sex 2</i> : F
âge	âge	<i>âge 1</i> : 60/69; <i>âge 2</i> : 70/79; <i>âge 3</i> : 80/89
eta	état civil	<i>eta c</i> : <i>celib.</i> ; <i>eta m</i> : marié; <i>eta v</i> : veuf; <i>eta d</i> : séparé.
etu	niveau d'études	<i>etu 1</i> : inf. primaire; <i>etu 2</i> : primaire <i>etu 3</i> : secondaire; <i>etu 4</i> : universitaire.
act	activité professionnelle	<i>act 1</i> : manuel; <i>act 2</i> : intell. <i>act 3</i> : agricult.; <i>act 4</i> : autre.
ret	retraité	<i>ret 1</i> : oui; <i>ret 2</i> : non.
tra	travaillez-vous ?	<i>tra 1</i> : oui; <i>tra 2</i> : non.
eau	eau courante ?	<i>eau 0</i> : non; <i>eau 1</i> : oui.
sn1	santé	<i>sn11</i> : bonne; <i>sn12</i> : mauvaise.
2sn	santé subjective	<i>2sn1</i> : bonne; <i>2sn2</i> : moyenne; <i>2sn3</i> : mauvaise.
sen	malade système nerveux	<i>sen 0</i> : non; <i>sen 1</i> : oui.
hta	hypertension	<i>hta 0</i> : non; <i>hta 1</i> : oui.
cir	pbs circulatoires	<i>cir 0</i> : non; <i>cir 1</i> : oui.
car	pbs cardiaques	<i>car 0</i> : non; <i>car 1</i> : oui.
dig	pbs digestifs	<i>dig 0</i> : non; <i>dig 1</i> : oui.
rhu	maladies osteoarticulaires	<i>rhu 0</i> : non; <i>rhu 1</i> : oui.
aff	êtes-vous affecté ?	<i>aff 1</i> : oui; <i>aff 2</i> : non.
ave	nombre de personne vivant avec vous	<i>ave 1</i> : aucune; <i>ave 2</i> : une. <i>ave 3</i> : plus d'une

Annexe 2 (Suite)

aud	entendez-vous ce que dit une personne ?	<i>aud 1</i> : oui; <i>aud 2</i> : difficilement; <i>aud 3</i> : non;
aug	entendez-vous dans un groupe ?	<i>aug 1</i> : oui; <i>aug 2</i> : difficilement; <i>aug 3</i> : non.
lir	pouvez-vous lire ?	<i>lir 1</i> : facilement; <i>lir 2</i> : difficilement; <i>lir 3</i> : ne lit pas; <i>lir 4</i> : ne peut pas lire.
mdv	nombre de visites du médecin par semaine	<i>mdv 1</i> : 0; <i>mdv 2</i> : 1 à 5; <i>mdv 3</i> : + de 5
mdc	nombre de visites chez le médecin par semaine	<i>mdc 1</i> : 0 <i>mdc 2</i> : 1 à 5; <i>mdc 3</i> : + de 5
hop	hospitalisations par an	<i>hop 0</i> : aucune; <i>hop 1</i> : une; <i>hop 3</i> : plus d'une.
med	nombre de médicaments par semaine	<i>med 1</i> : aucun; <i>med 2</i> : un; <i>med 3</i> : deux ou plus.
hab	vous habillez-vous seul ?	<i>hab 0</i> : non; <i>hab 1</i> : oui.
clu	adhérez-vous à un club ?	<i>clu 0</i> : non; <i>clu 1</i> : oui.
nbe	avez-vous des enfants ?	<i>nbe 0</i> : non; <i>nbe 1</i> : oui.
env	type d'environnement	<i>env u</i> : environnement urbain; <i>env r</i> : environnement rural.
qua	sortez-vous du quartier ?	<i>qua 0</i> : non; <i>qua 1</i> : oui.
den	difficulté d'environnement	<i>den 0</i> : pas de difficulté; <i>den 1</i> : qqes difficultés; <i>den 2</i> : bcp de difficultés.
soc	support social	<i>soc 0</i> : pas de support; <i>soc 1</i> : support moyen; <i>soc 2</i> : bon support.
aut	autonomie	<i>aut 0</i> : pas autonome; <i>aut 1</i> : peu autonome; <i>aut 2</i> : autonome.
ext	loisirs extérieurs	<i>ext 0</i> : non; <i>ext 1</i> : oui.
int	loisirs intérieurs	<i>int 0</i> : non; <i>int 1</i> : oui.
vis	fréquences des visites	(Dans certains cas, ces deux variables ont été regroupées en une seule : <i>loi</i>) <i>viss</i> : souvent; <i>visp</i> : parfois; <i>visr</i> : rarement; <i>visj</i> : jamais.

Les 12 variables de « satisfaction » dont la signification est donnée en Annexe 1, sont aussi prises en compte pour cette analyse. Elles possèdent 3 modalités chacune :

MOD 1 : insatisfaction

MOD 2 : indécision

MOD 3 : satisfaction.

Annexe 3

Valeurs des $\hat{\theta}_{rij}$ et des $\hat{\lambda}_r$ ($r = 1, \dots, 3, i = 1, \dots, 12, j = 1, \dots, 3$) de l'analyse en classes latentes, et des f_{rij} ($r = 1, \dots, 3, i = 1, \dots, 12, j = 1, \dots, 3$) obtenus par la méthode des nuées dynamiques

		classes latentes ($\hat{\theta}_{rij}$)			nuées dynamiques ($\hat{f}_{rij}$)		
classes variables		1	2	3	1	2	3
men	1	0.54	0.16	0.	0.56	0.20	0.08
	2	0.13	0.25	0.14	0.11	0.25	0.15
	3	0.33	0.59	0.78	0.33	0.56	0.76
dep	1	0.50	0.11	0.02	0.51	0.17	0.03
	2	0.17	0.27	0.08	0.14	0.28	0.08
	3	0.33	0.63	0.90	0.35	0.55	0.89
seu	1	0.63	0.06	0.04	0.60	0.10	0.04
	2	0.29	0.48	0.11	0.22	0.55	0.11
	3	0.18	0.45	0.86	0.17	0.35	0.84
oub	1	0.25	0.	0.	0.30	0.01	0.
	2	0.28	0.36	0.06	0.26	0.43	0.01
	3	0.47	0.64	0.94	0.44	0.56	0.99
fnu	1	0.39	0.06	0.01	0.44	0.08	0.01
	2	0.31	0.45	0.10	0.25	0.53	0.06
	3	0.31	0.50	0.89	0.31	0.39	0.93
sat	1	0.50	0.02	0.	0.63	0.03	0.01
	2	0.33	0.53	0.04	0.26	0.58	0.11
	3	0.17	0.45	0.96	0.11	0.40	0.88
fat	1	0.30	0.03	0.04	0.35	0.04	0.
	2	0.51	0.55	0.23	0.45	0.63	0.05
	3	0.19	0.43	0.73	0.20	0.33	0.94
inq	1	0.29	0.06	0.05	0.31	0.07	0.04
	2	0.29	0.55	0.21	0.27	0.54	0.23
	3	0.42	0.39	0.24	0.42	0.39	0.73
eco	1	0.18	0.14	0.05	0.20	0.15	0.06
	2	0.65	0.73	0.71	0.63	0.71	0.71
	3	0.17	0.13	0.24	0.17	0.14	0.23
cpr	1	0.15	0.07	0.05	0.17	0.10	0.05
	2	0.15	0.32	0.10	0.13	0.30	0.11
	3	0.70	0.60	0.84	0.71	0.61	0.93
rel	1	0.19	0.03	0.02	0.24	0.04	0.02
	2	0.11	0.25	0.04	0.09	0.25	0.04
	3	0.69	0.72	0.94	0.67	0.72	0.94
mvi	1	0.40	0.26	0.19	0.43	0.29	0.18
	2	0.25	0.43	0.23	0.22	0.42	0.23
	3	0.34	0.31	0.58	0.35	0.78	0.58
pourcentage des classes		15.5 = $\hat{\lambda}_1$	23 = $\hat{\lambda}_2$	61.5 = $\hat{\lambda}_3$	11	21.3	67.7

Annexe 4Tableau des log $\left[\frac{\hat{\theta}_{3ij}}{\hat{\theta}_{1ij}} \right]$

Variables	Modalités	log $[\hat{\theta}_{3ij}/\hat{\theta}_{1ij}]$
men	1	-1.841
	2	0.048
	3	0.866
dep	1	-3.017
	2	0.772
	3	0.989
seu	1	-2.674
	2	-0.999
	3	1.569
oub	1	4.415
	2	-3.474
	3	0.744
inu	1	-7.149
	2	-1.602
	3	1.116
sat	1	-3.701
	2	-1.253
	3	1.672
fat	1	-4.293
	2	-2.635
	3	1.614
inq	1	-1.953
	2	-0.257
	3	0.565
eco	1	-1.198
	2	-0.084
	3	0.347
cpr	1	-1.087
	2	-0.356
	3	0.191
rel	1	-2.209
	2	-1.092
	3	0.305
mvi	1	-1.212
	2	-0.104
	3	0.524

Annexe 5

Résultats du programme FACONE sur les 938 données complètes relatives
aux 5 variables de satisfaction : CPR, REL, MVI, INQ, ECO

On obtient 29 profils de réponse différents

Fréquence observée du profil	Fréquence estimée	score theta	score Y	nbre de 1	profil
34	35.583	-1.7230	0.0717	0	00000
21	19.705	-1.4477	0.1078	1	00010
1	1.625	-1.3564	0.1227	1	00001
29	25.236	-1.2238	0.1470	1	10000
4	4.462	-1.1961	0.1525	1	00100
1	1.163	-1.0961	0.1736	2	00011
21	19.841	-0.9640	0.2043	2	10010
3	3.579	-0.9361	0.2112	2	00110
45	41.947	-0.8748	0.2268	1	01000
1	1.844	-0.8734	0.2272	2	10001
4	6.265	-0.7092	0.2721	2	10100
39	42.458	-0.6052	0.3026	2	01010
2	1.868	-0.6037	0.3031	3	10011
6	4.314	-0.5087	0.3321	2	01001
8	7.161	-0.4304	0.3567	3	10110
83	91.370	-0.3628	0.3784	2	11000
18	17.256	-0.3314	0.3886	2	01100
9	5.719	-0.2161	0.4263	3	01011
131	134.975	-0.0587	0.4785	3	11010
27	26.093	-0.0248	0.4898	3	01110
2	1.151	-0.0231	0.4903	4	10111
15	15.725	0.0520	0.5152	3	11001
2	3.066	0.0869	0.5266	3	01101
96	0.981	0.2575	0.5818	3	11100
37	31.652	0.3921	0.6239	4	11011
4	6.334	0.4302	0.6354	4	01111
184	190.084	0.6164	0.6897	4	11110
19	26.230	0.7479	0.7254	4	11101
92	88.872	1.1526	0.8193	5	11111
938	936.559				