

REVUE DE STATISTIQUE APPLIQUÉE

A. BACCINI

J. R. MATHIEU

A. M. MONDOT

Comparaison, sur un exemple, d'analyses des correspondances multiples et de modélisations

Revue de statistique appliquée, tome 35, n° 3 (1987), p. 21-34

http://www.numdam.org/item?id=RSA_1987__35_3_21_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/>

COMPARAISON, SUR UN EXEMPLE, D'ANALYSES DES CORRESPONDANCES MULTIPLES ET DE MODÉLISATIONS

A. BACCINI, J.R. MATHIEU et A.M. MONDOT

*Laboratoire de Statistique et Probabilités,
U.A.-C.N.R.S. 745, Université Paul Sabatier
118, route de Narbonne, 31062 Toulouse Cedex*

RÉSUMÉ

Dans cet article, on montre comment des Analyses Factorielles des Correspondances Multiples (A.F.C.M.) et certaines modélisations peuvent être utilisées conjointement sur des données volumineuses. Dans une phase exploratoire, diverses A.F.C.M. permettent de simplifier sensiblement les données et de définir des sous-populations « homogènes ». Sur ces dernières, une modélisation permet de quantifier les liaisons entre variables et de procéder, dans certains cas, à de nouvelles simplifications. Une méthodologie du traitement de données volumineuses (pouvant inclure d'autres méthodes telles que la segmentation) est ainsi esquissée, dans laquelle un contrôle mutuel des différentes techniques peut être réalisé.

Mots clés : *Analyse des Correspondances, Modélisations, Comparaison de méthodes, Données volumineuses.*

SUMMARY

In this article, we illustrate how Multiple Correspondence Analyses (M.C.A.) and modelling methods can be jointly used on a large data set. In an exploratory stage, several M.C.A. allow to reduce the data (thus obtaining a smaller set of variables and of levels) and to define homogeneous sub-populations of interest. Then, each subpopulation being treated separately, modelling methods enable to express the links between the different variables and lead, in some cases, to further reductions. This provides some clues to set a general methodology for the statistical treatment of large data sets, of which the mutual control of the different technics (M.C.A. and modelling) is an interesting feature.

1. Introduction

Le Secrétariat Général de l'Université Paul Sabatier (U.P.S.) de Toulouse a constitué, depuis 1971, une banque de données relative aux étudiants de l'Université, à laquelle nous avons pu avoir accès. La population étudiée dans un premier temps était l'ensemble des étudiants de l'U.P.S. :

- inscrits dans l'une des trois U.E.R. (Unités d'Enseignement et de Recherche) scientifiques : M.I.G. (Mathématiques, Informatique et Gestion), P.C.A. (Physique, Chimie et Automatique), Sciences Naturelles;
- ayant pris leur première inscription à l'U.P.S. entre 1971 et 1979;
- ayant achevé ou arrêté leurs études au plus tard en 1981 (c'est-à-dire inscrits à l'U.P.S. ni en 1981-1982, ni en 1982-1983).

Après un premier tri du fichier (suppression des données manquantes ou aberrantes), nous avons considéré 9 232 étudiants.

Les variables trop particulières (renseignements sur des unités de valeur,...) n'ayant pas été prises en compte, nous disposions au départ de 13 variables, toutes catégorielles, comportant un total de 77 modalités :

- sexe (2);
- nationalité (6) : française, européenne non française, asiatique, africaine non maghrébine, maghrébine, autre;
- C.S.P. (catégorie socio-professionnelle des parents, 10) : agriculteurs exploitants, ouvriers agricoles, patrons de l'industrie et du commerce, professions libérales et cadres supérieurs, cadres moyens, employés, ouvriers, personnels de service, autres catégories, sans profession;
- série de bac (baccalauréat) (6) : séries A, séries B, G et H (une seule modalité), séries C et E (même chose), séries D, séries F, équivalence du bac;
- mention au bac (5) : très bien, bien, assez bien, passable, sans objet (cette dernière modalité concerne non seulement les étudiants ayant une équivalence du bac, mais aussi ceux qui ayant le bac n'en ont pas précisé la mention);
- âge d'obtention du bac (12) : 16 ans, 17 ans, 18 ans, 19 ans, 20 ans, 21 ans, 22 ans, 23 ans, 24 ans, 25 ans, 26 ans et plus, sans objet;
- lieu d'obtention du bac ou de l'équivalence (4) : Haute-Garonne, départements périphériques, reste de la métropole, étranger;
- U.E.R. de rattachement (3) : M.I.G., P.C.A., Sciences Naturelles;
- niveau atteint en quittant l'U.P.S. (7) : aucun diplôme, première année du premier cycle, premier cycle, licence, maîtrise, D.E.A. (Diplôme d'Etudes Approfondies), doctorat;
- durée d'obtention du premier cycle (5) : 1 an (équivalence de la première année), 2 ans, 3 ans, 4 ans et plus, sans objet;
- durée d'obtention de la licence (5) : 1 an, 2 ans, 3 ans, 4 ans et plus, sans objet;
- durée d'obtention de la maîtrise (4) : 1 an, 2 ans, 3 ans et plus, sans objet;
- durée totale des études à l'U.P.S. (8) : 1 an, 2 ans, 3 ans, 4 ans, 5 ans, 6 ans, 7 ans, 8 ans et plus.

Compte-tenu de l'importance de ce fichier, et dans la mesure où il s'agissait d'étudier, de façon très générale, les facteurs de la réussite et de l'échec, la part des analyses exploratoires a été très importante dans ce travail.

2. Analyses exploratoires préliminaires

L'objectif était, au moyen d'une simplification des données (suppression de variables peu influentes, regroupement de modalités voisines d'une même variable, ...), d'essayer d'y trouver une certaine structure. En particulier, le niveau atteint et les variables de durée (durée totale, durée du premier cycle, ...) pouvant, de façon naturelle, jouer le rôle de variables à expliquer, nous souhaitions savoir dans quelle mesure il était possible de les expliquer au moyen des autres.

Nous avons dans cette phase utilisé des Analyses Factorielles des Correspondances Multiples (A.F.C.M.), cette technique nous paraissant la plus appropriée à ce type de situation. Sauf indications contraires, les A.F.C.M. réalisées

ont consisté à faire l'Analyse des Correspondances usuelle sur le tableau de Burt associé aux variables considérées. Le logiciel utilisé est le logiciel E.P.A.D. (Enseignement et Pratique de l'Analyse des Données), sur le Centre Interuniversitaire de Calcul de Toulouse. Ce logiciel, d'utilisation assez commode, comporte les programmes classiques d'analyse des données (analyse en composantes principales, analyse des correspondances, analyse discriminante, classification hiérarchique) ainsi qu'un programme de tri et de recodification de données et un programme de préparation à l'analyse des correspondances multiples construisant tableau de Burt ou tableau disjonctif complet.

2.1. Analyses des données complètes

Les premières analyses nous ont permis de réduire à 55 le nombre total des modalités des 13 variables. Les trois premiers axes de l'A.F.C.M. des données ainsi réduites représentent respectivement 27 %, 13 % et 6 % de l'inertie totale (les valeurs propres étant 0,103 0,053 et 0,022). Le premier axe est très lié au niveau atteint et aux variables de durée; le deuxième axe est essentiellement expliqué par la position très à part des étudiants de nationalité étrangère et des modalités qui leur sont liées (équivalence du bac, ...); enfin, l'axe 3, d'interprétation plus délicate, oppose, d'un côté, les mentions très bien ou bien, les séries C ou E, les U.E.R. M.I.G. et P.C.A. et, d'un autre côté, les filles, les bacs D, la mention passable et l'U.E.R. de Sciences Naturelles. Le fait important à noter est l'orthogonalité des variables pouvant jouer le rôle de variables à expliquer (niveau atteint et variables de durée), liées à l'axe 1, avec l'ensemble des 8 autres variables. Un modèle de type explicatif pour l'ensemble des données ne semble donc pas pouvoir être envisagé.

2.2. Autres analyses

Compte-tenu de la position très particulière des étudiants de nationalité étrangère et du fait que, le plus souvent titulaires d'une équivalence du baccalauréat, ils ont fréquemment le même profil de réponse aux 4 variables relatives à ce diplôme, nous avons été amenés à faire deux analyses distinctes :

- A.F.C.M. des étudiants de nationalité française et titulaires du baccalauréat (l'effectif est de 8142) : les deux premiers axes résument respectivement 28 % et 7 % de l'inertie totale (valeurs propres : 0,102 et 0,027); l'axe 1 représente, comme dans l'analyse précédente, le niveau atteint et les variables de durée, tandis que l'axe 2 a sensiblement la même signification que l'axe 3 de l'analyse précédente. Si l'on supprime les variables de durée, le premier plan, représentant 35 % de l'inertie totale, est difficile à interpréter, tandis que le troisième axe (9 %) représente le niveau atteint, cette variable étant orthogonale à toutes les autres;

- A.F.C.M. des étudiants de nationalité étrangère : les résultats de ces analyses, très confus, n'ont pas permis de dégager une quelconque structure sur cette sous-population de 1025 étudiants. Pour cette raison, nous n'avons pas poursuivi l'étude sur les étudiants étrangers.

Pour les étudiants de nationalité française et titulaires du baccalauréat, il nous a paru souhaitable, à ce stade, de faire des analyses distinctes selon le niveau atteint. Nous avons pour cela réalisé quatre séries d'analyses :

- pour les étudiants n'ayant obtenu aucun diplôme, ou seulement la première année du premier cycle : les résultats de ces analyses, également très confus, n'ont permis aucune interprétation, sans doute à cause du mélange des étudiants en situation d'échec et de ceux des classes préparatoires, inscrits à l'Université et ne passant pas les examens de premier cycle car reçus à un concours;

- pour les étudiants ayant obtenu uniquement le premier cycle; cette série d'analyses et la suivante sont présentées en détails dans les paragraphes 3 et 4;

- pour les étudiants ayant obtenu uniquement le deuxième cycle;

- pour les étudiants ayant obtenu un diplôme de troisième cycle : peu nombreux (310), dans des situations souvent très différentes, ces étudiants forment une population très hétérogène; les analyses les concernant n'ont pu être interprétées clairement et n'ont pas été poursuivies.

3. Analyses des étudiants de nationalité française, titulaires du baccalauréat, et ayant obtenu le DEUG

Le DEUG (Diplôme d'Enseignement Universitaire Général) est le diplôme sanctionnant la fin du premier cycle. Dans un premier temps, les analyses ont été faites sur la sous-population issue de celle décrite plus haut; dans un second temps, elles ont été recommencées en rajoutant les étudiants ayant achevé leur cursus en 1982 ou en 1983 (nous avons réutilisé la banque de données après mise à jour). Ces deux ensembles d'analyses ayant donné sensiblement les mêmes résultats, nous présentons les dernières, plus complètes.

3.1. Analyses exploratoires

Nous avons tout d'abord considéré les étudiants ayant eu le DEUG et aucun autre diplôme à l'U.P.S. Ils sont 813, et 9 variables à 35 modalités ont été considérées au départ (les variables inopérantes sur cette sous-population sont la nationalité, le niveau atteint, la durée d'obtention de la licence et la durée d'obtention de la maîtrise). Les A.F.C.M. effectuées successivement nous ont conduits à ne retenir finalement que 4 variables à 14 modalités :

- série de bac (3) : séries A, B, F, G et H (codées BFGH dans les graphiques), séries C et E (bCbE), séries D (bac D);
- mention au bac (4) : très bien ou bien (TBBI), assez bien (ABIE), passable (PASS), inconnue (me??; ces étudiants, titulaires du bac, n'ont pas déclaré de mention; la plupart d'entre eux ont vraisemblablement obtenu la mention passable);
- âge d'obtention du bac (4) : avant 18 ans ($a < 18$), 18 ans ($a = 18$), 19 ans ($a = 19$), après 19 ans ($a > 19$);
- durée d'obtention du premier cycle (3) : 2 ans (1cy2), 3 ans (1cy3), 4 ans ou plus (1cy4); on notera que les étudiants ayant obtenu le premier cycle en un an, peu nombreux et en situation particulière, ont été retirés de cette série d'analyses.

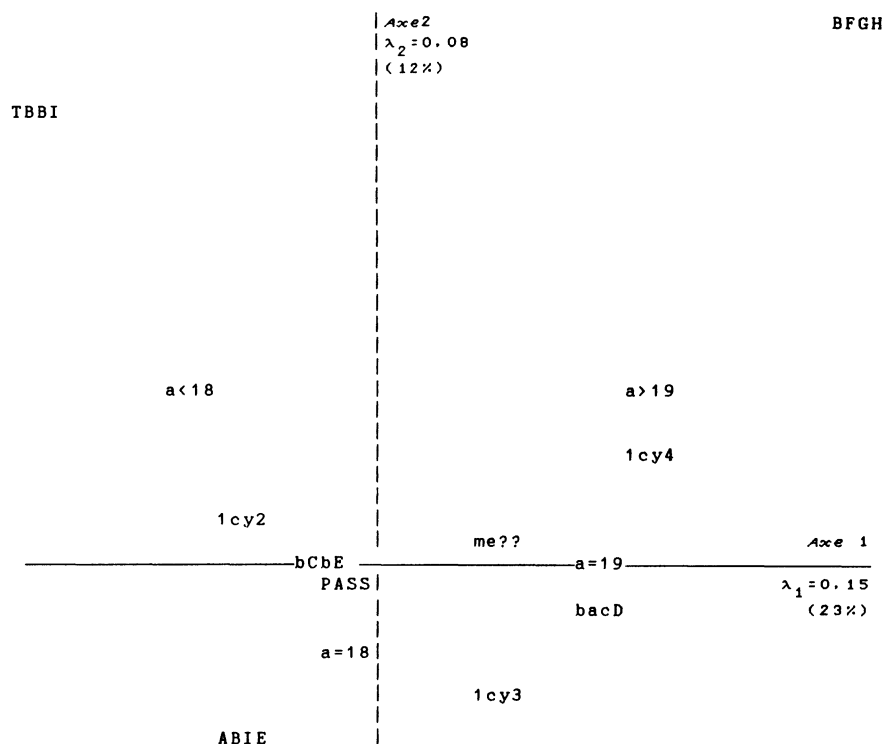


FIGURE 1

Analyse des Correspondances Multiples des étudiants de nationalité française titulaires du baccalauréat et ayant obtenu le DEUG : 4 variables (série de bac, mention, âge d'obtention du bac, durée d'obtention du premier cycle) à 14 modalités; 813 étudiants.

La figure 1 donne le plan principal fourni par l'A.F.C.M. de ces données. Le premier axe (associé à une valeur propre égale à 0,15 et résumant 23 % de l'inertie totale) est facilement interprétable : il représente un effet taille lié à chaque variable et ordonne la durée du premier cycle (de 2 ans à 4 ans), l'âge au bac (des plus jeunes aux plus âgés), la mention au bac (de la meilleure à la moins bonne) et la série de bac (des plus scientifiques aux moins scientifiques). Le deuxième axe (valeur propre : 0,08; inertie : 12 %) semble opposer les extrêmes en haut (mentions très bien ou bien et bac à moins de 18 ans ainsi que bac A, B, F, G ou H et bac à plus de 19 ans) aux moyens en bas (mention assez bien, bac à 18 ans et premier cycle en 3 ans). Le troisième axe (valeur propre 0,07; inertie : 11 %) est, lui, ininterprétable. On trouve ainsi dans le premier plan principal ce que l'on appelle l'effet Guttman (voir BENZECRI, 1973, ou VOLLE, 1985), correspondant bien à la structure ordonnée de chaque variable. La représentation graphique des individus à partir du tableau disjonctif complet montre que la plupart d'entre eux se situent autour de la « parabole » définie par les modalités des variables (nous ne représentons pas ce graphique car il est trop confus); en particulier, on ne trouve aucun étudiant dans la zone située vers le

« sommet » de l'axe 2, ce qui aurait pu dénoter chez certains un comportement en opposition avec l'ensemble. Au vu de la régularité de ces résultats, on peut donc s'attendre à expliquer correctement la durée d'obtention du premier cycle au moyen des trois autres variables.

Afin de confirmer les résultats obtenus par l'A.F.C.M., nous avons souhaité utiliser un modèle log-linéaire sur ces données. La première constatation est que cela n'est pas directement possible, compte-tenu du faible effectif des séries A, B, F, G et H (16 étudiants) et du grand nombre de cellules vides que cela entraîne. Nous avons donc décidé de ne pas considérer ces étudiants et de traiter uniquement les 797 étudiants restants. Une A.F.C.M. sur ces nouvelles données conduit aux mêmes conclusions que la précédente.

Dans la modélisation log-linéaire (effectuée sous le logiciel GLIM) nous avons recherché un modèle convenable $\mathcal{M}$ en prenant comme critère de choix la déviance de ce modèle, qui est la statistique du test du rapport de vraisemblance, c'est-à-dire deux fois la différence entre le logarithme du maximum de vraisemblance dans le modèle saturé (où il y a autant de paramètres que d'observations) et le logarithme du maximum de vraisemblance dans le modèle $\mathcal{M}$. On sait que si le modèle $\mathcal{M}$ est vérifié par les observations, cette statistique est asymptotiquement distribuée suivant une loi du khi-deux centrée dont la moyenne est son degré de liberté; par suite, si la déviance d'un modèle est de l'ordre du degré de liberté, on considère que ce modèle est acceptable.

En l'occurrence, le modèle log-linéaire où figurent les quatre variables et leurs interactions d'ordre 1 (les seules prises en compte par l'A.F.C.M.) a une déviance de 66 pour un degré de liberté de 57. L'analyse des résidus standardisés confirme l'adéquation de ce modèle aux données : seuls 2 résidus sur 96 ont des valeurs absolues dépassant 2 (2,04 et 2,86). De plus, aucune des quatre variables ne semble devoir être omise, ce qui confirme que les A.F.C.M. ont bien mis en évidence les variables structurant les données considérées.

Par ailleurs, afin de simplifier encore les données, nous avons souhaité regrouper les modalités de la variable mention au bac en seulement deux niveaux : « avec mention » (ment : regroupement de TBBI et de ABIE) et « sans mention » (ssme : regroupement de PASS et de me??); ces regroupements semblent, a priori, naturels, ne sont pas contradictoires avec les coordonnées sur l'axe 1 de l'A.F.C.M. et permettent d'obtenir un modèle log-linéaire avec interactions d'ordre 1 très acceptable : déviance de 24 pour 23 degrés de liberté, les deux plus importants résidus standardisés valant 1,28 et 1,80, tous les autres étant inférieurs à 1,2. Nous avons donc décidé de faire ces regroupements et la figure 2 donne le premier plan principal de l'A.F.C.M. de ces données simplifiées. Son interprétation demeure analogue aux précédentes.

Enfin, les modélisations qui suivent (voir paragraphe 3.2) ayant pour objectif d'expliquer la variable durée du DEUG en fonction des trois autres, nous avons achevé les analyses exploratoires en faisant une Analyse des Correspondances simple sur le tableau croisant la durée du DEUG avec ces trois variables (ce tableau est extrait du tableau de Burt traité dans l'A.F.C.M. précédente). Les résultats dans le plan principal en sont présentés dans la figure 2bis et ne font que confirmer les liaisons entre variables apparues dans les analyses précédentes.

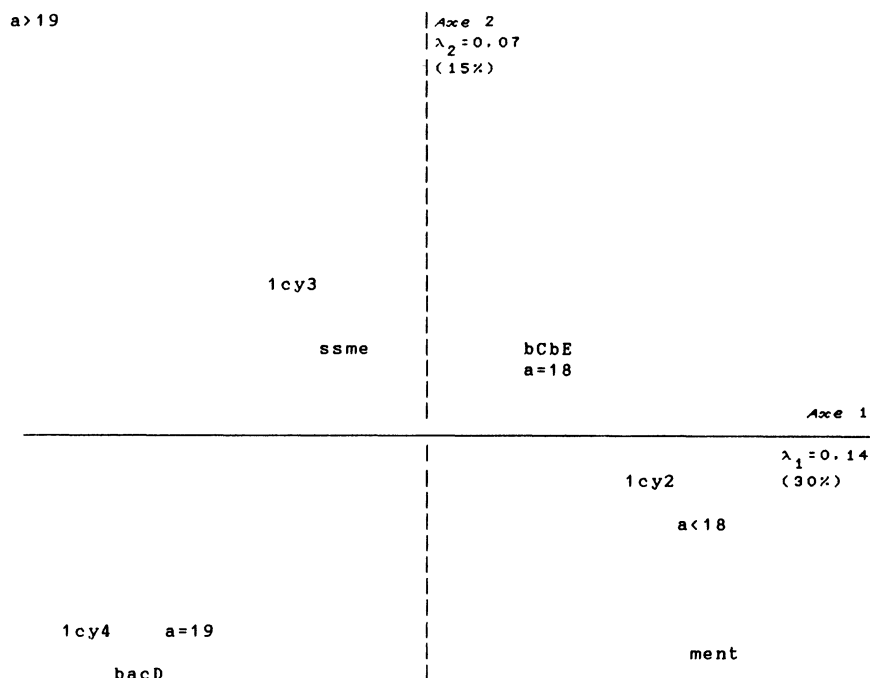


FIGURE 2

Analyse des Correspondances Multiples des étudiants de nationalité française titulaires du baccalauréat et ayant obtenu le DEUG : 4 variables (série de bac, mention, âge d'obtention du bac, durée d'obtention du premier cycle) à 11 modalités; 797 étudiants.

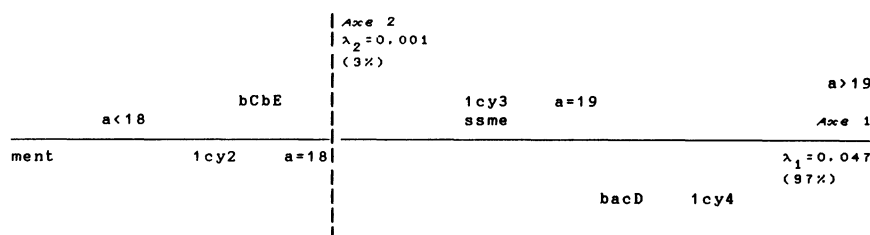


FIGURE 2bis

Analyse des Correspondances simple de la même population (797 étudiants) croisant la durée d'obtention du premier cycle (3 modalités) avec la série du bac, la mention et l'âge d'obtention du bac (8 modalités en tout).

On notera que la contraction de l'échelle de l'axe 2 fait disparaître dans ce cas l'effet Guttman.

3.2. Modélisations

Nous avons souhaité, dans cette partie, expliquer la durée d'obtention du DEUG par les conditions d'obtention du baccalauréat.

Dans la catégorie des étudiants ayant obtenu le DEUG après avoir obtenu le bac à l'âge a ($a = 1, 2, 3, 4$), dans la série s ($s = 1, 2$) et avec la mention m ($m = 1, 2$), on note p_{asm}^2 (resp. p_{asm}^3, p_{asm}^4) les probabilités d'obtention du DEUG en 2 ans (resp. 3 ans, 4 ans), et on se propose d'expliquer p_{asm}^2 et p_{asm}^3 selon des modèles logistiques, c'est-à-dire d'expliquer

$$\text{Log} \frac{p_{asm}^2}{1 - p_{asm}^2} \text{ et } \text{Log} \frac{p_{asm}^3}{p_{asm}^4}$$

suivant des modèles simples (autant que possible additifs).

Remarque : Le modèle logistique est en fait celui qui est apparu le plus approprié après divers essais sous GLIM. Par ailleurs, le choix de $\frac{p_{asm}^3}{p_{asm}^4}$ plutôt que $\frac{p_{asm}^3}{1 - p_{asm}^3}$ tient au fait que $1 - p_{asm}^3$ vaut $p_{asm}^2 + p_{asm}^4$, somme des probabilités de deux événements (obtention du DEUG en 2 ans et obtention en 4 ans) qu'il est difficile d'opposer ensemble à l'obtention en 3 ans.

3.2.a. Explication de $\text{Log} \frac{p_{asm}^2}{1 - p_{asm}^2}$

Avec le modèle sans interactions, où les effets des facteurs s'additionnent les uns aux autres pour reconstituer $\text{Log} \frac{p_{asm}^2}{1 - p_{asm}^2}$, on obtient une explication plutôt satisfaisante des données (déviante de 14 pour 10 d.d.l.).

Dans ce modèle // « de référence », on peut éprouver une hypothèse H_0 avec le test du rapport de vraisemblance dont la statistique T est deux fois la différence des logarithmes des maxima de la vraisemblance d'une part dans le modèle //, d'autre part sous H_0 ; si H_0 est vraie, cette statistique est asymptotiquement distribuée comme une loi du khi-deux dont le degré de liberté est le nombre de contraintes imposées par H_0 .

En l'occurrence, éprouvons chacune des hypothèses exprimant l'absence d'effet d'un facteur :

pour le facteur a : $T_a = 14,77$ pour 3 degrés de liberté;

pour le facteur s : $T_s = 19,78$ pour 1 degré de liberté;

pour le facteur m : $T_m = 24,04$ pour 1 degré de liberté.

On est donc amené à conclure à l'effet significatif de chaque facteur. D'après les estimations des paramètres mesurant les effets de ces 3 facteurs, $\frac{p_{asm}^2}{1 - p_{asm}^2}$ a une valeur de 3,50 (correspondant à 7 succès en 2 ans pour 2 succès en 3 ou 4 ans) pour les étudiants ayant obtenu le bac C ou E, à moins de 18 ans et avec mention; pour les autres catégories, la valeur de $\frac{p_{asm}^2}{1 - p_{asm}^2}$ se déduit de la valeur précédente en la multipliant par un « coefficient d'abattement » ayant pour valeur :

0,70 pour les bac à 18 ans;

0,48 pour les bac à 19 ans;

0,34 pour les bac à plus de 19 ans;

0,47 pour les bac D;

0,43 pour les bac « sans mention ».

Ainsi, pour les étudiants se trouvant dans la situation la moins favorable (bac D à plus de 19 ans et sans mention), la valeur de $\frac{p_{asm}^2}{1 - p_{asm}^2}$ est $3,5 \times 0,34 \times 0,47 \times 0,43 \approx 0,24$, soit un succès en 2 ans pour 4 succès en 3 ou 4 ans.

3.2.b. Explication de $\text{Log} \frac{p_{asm}^3}{p_{asm}^4}$

Dans le modèle additif, on a une bonne explication des données (déviante de 7 pour 10 d.d.l.), les statistiques T ayant ici pour valeurs :

$T_a = 0,35$ pour 3 degrés de liberté;

$T_s = 8,86$ pour 1 degré de liberté;

$T_m = 4,41$ pour 1 degré de liberté.

On est donc amené à considérer que l'âge n'a plus d'influence, mais que les facteurs s et m conservent leur effet. D'après les estimations des paramètres, le rapport théorique du nombre de reçus en 3 ans au nombre de reçus en 4 ans s'obtient pour chaque niveau des facteurs s et m comme le produit d'un taux de base de 5,19 (pour les bac C ou E avec mention) par les coefficients d'abattement suivants :

0,51 pour les bac D;

0,53 pour les bac « sans mention ».

3.2.c. Remarque

Pour contrôler la cohérence des deux modélisations ci-dessus, on a recherché une modélisation de $\frac{p_{asm}^4}{1 - p_{asm}^4}$; un modèle convenable est celui où $\text{Log} \frac{p_{asm}^4}{1 - p_{asm}^4}$ est expliqué par les deux facteurs série de bac et mention, dont les effets s'interprètent facilement (modèle sans interactions). De ces différentes modélisations, on déduit des estimations $\hat{p}^2$, $\hat{p}^3$ et $\hat{p}^4$ des vecteurs p^2 , p^3 et p^4 de $\mathbf{R}^{16}$. On examine ensuite si le vecteur $\hat{p}^2 + \hat{p}^3 + \hat{p}^4$ est voisin du vecteur $\mathbf{1}$. En l'occurrence, ce vecteur a toutes ses composantes comprises entre 0,95 et 1,06, ce qui nous semble assez cohérent.

3.3. Analyses complémentaires

3.3.a. Segmentation

Nous avons, toujours sur les mêmes données, réalisé une segmentation (voir BACCINI et LAMBERT, 1986, pour la méthode et le programme) avec la durée d'obtention du DEUG comme variable à expliquer et l'âge au bac, la série et la mention comme variables explicatives. La variable apparaissant comme la plus liée à la durée est la série de bac (khi-deux de 35,7 pour 2 degrés de liberté, très significatif), suivie de la mention; conditionnellement aux deux variables précé-

dentes, l'âge au bac apparaît comme peu explicative. Tout en confirmant les résultats précédents, cette méthode permet d'obtenir une hiérarchie des 3 variables explicatives (prises séparément) dans leur liaison avec la durée d'obtention du DEUG.

3.3.b. Etude des mêmes variables sur la sous-population des étudiants ayant obtenu la maîtrise

Les études précédentes ayant été réalisées sur les seuls étudiants ayant obtenu uniquement le DEUG à l'U.P.S., nous avons, à titre de comparaison, analysé les résultats au DEUG des étudiants titulaires de la Maîtrise (diplôme de fin du second cycle). Les A.F.C.M. mettent en évidence les mêmes variables que précédemment et, après simplifications analogues, on obtient des résultats tout à fait voisins.

En ce qui concerne la modélisation de $\text{Log} \frac{p_{asm}^2}{1 - p_{asm}^2}$, on obtient une très bonne explication des observations avec un modèle sans interactions (déviante de 2 pour 10 d.d.l.). Les trois facteurs ont un effet, mais l'on peut regrouper les trois premières catégories de l'âge d'obtention du bac (inférieur à 18 ans, 18 ans et 19 ans). On obtient alors les valeurs suivantes pour $\text{Log} \frac{p_{asm}^2}{1 - p_{asm}^2}$: le taux de base est 3,42 pour les plus jeunes qui ont obtenu un bac C ou E avec mention; les coefficients d'abattement, avec un effet multiplicatif, sont alors :

0,51 pour les bac à plus de 19 ans;

0,36 pour les bac D;

0,60 pour les bac « sans mention ».

Comme pour les données précédentes, une modélisation convenable de $\text{Log} \frac{p_{asm}^3}{p_{asm}^4}$ est obtenue avec un modèle additif comportant les deux seuls facteurs séries de bac et mention (déviante de 14 pour 13 d.d.l.). La valeur de base de $\frac{p_{asm}^3}{p_{asm}^4}$ est 7,48, avec les coefficients d'abattement suivants :

0,81 pour les bac D;

0,55 pour les bac « sans mention ».

On remarque que la série de bac intervient peu dans ce cas.

4. Analyses des étudiants de nationalité française, titulaires du baccalauréat, et ayant obtenu la maîtrise

Rappelons que la maîtrise est le diplôme sanctionnant la fin des études en second cycle. Là encore, nous avons procédé en deux temps et nous ne présentons que les résultats relatifs au fichier le plus complet (1 245 étudiants ayant obtenu la maîtrise mais pas de diplôme de troisième cycle à l'U.P.S.). Les variables initiales étaient a priori les mêmes que dans l'analyse précédente (9 variables à 35 modalités, voir 3.1), avec en plus la durée du second cycle à 3 modalités (deux ans, trois ans, quatre ans et plus).

4.1. Analyses exploratoires

Les différentes A.F.C.M. réalisées sur ces données nous ont finalement conduits à ne retenir que 5 variables comportant un total de 15 modalités. Il s'agit de l'âge d'obtention du bac (4 modalités), de l'U.E.R. de rattachement (3), de la C.S.P. (2 : C.S.P. « favorisées », psup, et C.S.P. « moins favorisées », pinf), de la durée du premier cycle (3) et de la durée du deuxième cycle (3). La figure 3 présente le résultat de l'A.F.C.M. de ces données dans le plan principal (les valeurs propres valent 0,078 et 0,064 et résument respectivement 19 % et 15 % de l'inertie totale).

On retrouve la liaison déjà observée entre l'âge au bac et la durée du premier cycle, et l'on note, en outre, une durée du deuxième cycle plus courte dans les U.E.R. M.I.G. et Sciences Naturelles pour les étudiants des C.S.P. « favorisées » et, au contraire, une durée plus longue pour l'U.E.R. P.C.A. et les autres C.S.P.

Si l'on cherche à expliquer la durée du second cycle au moyen des 4 autres variables, on constate ici une liaison moins nette que dans le cas des étudiants ayant obtenu seulement le DEUG; en effet, la durée du second cycle apparaît

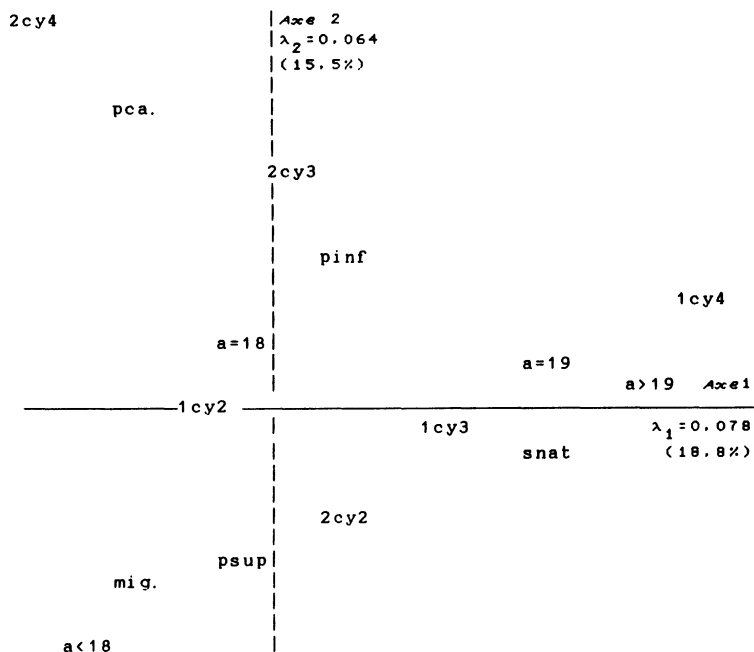


FIGURE 3

Analyse des Correspondances Multiples des étudiants de nationalité française, titulaires du baccalauréat, et ayant obtenu la maîtrise : 5 variables (âge d'obtention du bac, U.E.R. de rattachement, C.S.P., durée du premier cycle, durée du second cycle) à 15 modalités; 1 245 étudiants.

surtout liée au deuxième axe, alors que durée du premier cycle et âge au bac sont davantage liés au premier axe.

Signalons que l'on a également réalisé ici une Analyse des Correspondances simple croisant la durée d'obtention de la maîtrise avec les quatre autres variables, mais que les résultats ne nous ont fourni aucune information complémentaire sur les liaisons entre ces variables.

4.2. Modélisations

Les diverses modélisations réalisées ici ont encore nécessité certains regroupements de modalités. Ces regroupements ont été faits de façon à ne pas conserver des effectifs trop petits et en tenant compte de la position des modalités dans le plan principal de l'A.F.C.M. (regroupement de modalités suffisamment proches). On a finalement considéré :

- la durée du premier cycle, à 2 modalités : 2 ans, 3 ans ou plus;
- l'U.E.R., à 3 modalités;
- la C.S.P., à 2 modalités;
- l'âge d'obtention du bac, à 3 modalités : moins de 18 ans, 18 ans, plus de 18 ans;
- la durée du second cycle, à 2 modalités : 2 ans, 3 ans ou plus.

Un modèle log-linéaire sans interactions du second ordre (effets principaux des facteurs et interactions du premier ordre) n'est pas très satisfaisant : la déviance vaut 67 pour 45 d.d.l.; il semble donc qu'il existe des interactions du second ordre, en particulier entre la durée du second cycle, l'U.E.R. et l'âge au bac, ces interactions pouvant avoir perturbé l'A.F.C.M.

On a ensuite recherché une modélisation de la proportion p_{asm}^2 de maîtrises obtenues en 2 ans relativement au nombre total de succès en maîtrise; avec un modèle logistique sans interactions et faisant intervenir les facteurs durée du premier cycle, U.E.R. et C.S.P., on a une explication assez satisfaisante : la déviance vaut 39 pour 31 d.d.l. (l'effet du facteur âge est donc négligeable).

La valeur de base de $\frac{p_{asm}^2}{1 - p_{asm}^2}$ est 2,34 (pour l'U.E.R. M.I.G., le premier cycle obtenu en 2 ans et les C.S.P. « moins favorisées »), avec les coefficients multiplicateurs suivants :

- 0,67 si la durée du premier cycle est supérieure à 2 ans;
- 0,56 pour l'U.E.R. P.C.A.;
- 1,99 pour l'U.E.R. Sciences Naturelles;
- 1,52 pour les C.S.P. « favorisées ».

Globalement, on constate que ce modèle explicatif (le plus approprié) est moins satisfaisant que ceux retenus à l'issue des analyses présentées au paragraphe 3. Ceci est à rapprocher du fait que les résultats de l'A.F.C.M. sont, du point de vue de l'explication de la variable durée d'obtention du diplôme, moins clairs dans le cas de la maîtrise que dans le cas du DEUG.

Pour terminer, signalons qu'une segmentation a mis en évidence la prépondérance du facteur U.E.R. dans l'explication de la durée de la maîtrise.

5. Conclusions

Au-delà de l'exemple lui-même, il nous paraît important d'essayer de dégager les grandes lignes d'une méthodologie du traitement statistique, tel qu'on peut l'envisager face à ce type de données. Lorsque les données sont volumineuses (et c'est souvent le cas dans la pratique), l'utilisation préalable de méthodes exploratoires est indispensable, surtout si les objectifs ne sont pas tout à fait précisés, comme c'était le cas ici. Nous avons utilisé des Analyses Factorielles des Correspondances Multiples, mais il aurait également été possible d'utiliser des méthodes de classification automatique; nous avons préféré les premières car elles nous paraissent plus efficaces pour déterminer les variables les plus influentes dans la structuration des données. On notera que les unes comme les autres sont, sur de gros fichiers, coûteuses en temps de calcul.

Une fois les données simplifiées, au besoin en extrayant des sous-fichiers comme cela a été fait ici, diverses méthodes de modélisation peuvent alors être envisagées. La modélisation log-linéaire est assez naturelle si les variables considérées jouent des rôles symétriques; c'est en ce sens la plus proche de l'A.F.C.M. et elle peut aussi permettre de contrôler les résultats obtenus dans cette dernière. Elle permet également, le cas échéant, de nouvelles simplifications des données, comme on l'a vu précédemment. Si l'on souhaite ensuite expliquer une variable au moyen des autres, le modèle linéaire généralisé fournit des outils appropriés pour quantifier les liaisons entre variables, comme cela a été vu ici dans chaque cas.

Signalons par ailleurs que des techniques telles que la segmentation peuvent aussi fournir des informations complémentaires sur les données étudiées (par exemple hiérarchie des variables explicatives dans l'explication d'une variable déterminée).

Enfin, on notera, dans l'optique de la comparaison des méthodes, que lorsque les résultats de l'A.F.C.M. étaient très clairs (cas du DEUG), on a obtenu un modèle très satisfaisant, alors que dans le cas de la maîtrise, les résultats de l'A.F.C.M. étaient moins nets et aucun modèle vraiment satisfaisant n'a pu ajuster les données. Ainsi, les deux approches (analyses exploratoires et méthodes de modélisation), bien que de natures très différentes, apparaissent en fait très liées.

Remerciements

Nous tenons à remercier Monsieur J.P. PRINEAU, Secrétaire Général de l'Université Paul Sabatier, pour nous avoir donné accès à la banque de données relative aux étudiants.

Bibliographie

- D.F. ANDREWS (1978). — *Data analysis, exploratory*. In International Encyclopedia of Statistics. Ed. W.H. Kruskal and J.M. Tanur, Collier Macmillan Publishers, The Free Press, New York.
- A. BACCINI (1984). — *Etude comparative des représentations graphiques en Analyses Factorielles des Correspondances simples et multiples*. Publications du Laboratoire de Statistique et Probabilités, Université Paul Sabatier, N° 02 — 84.
- A. BACCINI et E. LAMBERT (1986). — Algorithmes et programme de segmentation aux moindres carrés. *R.A.I.R.O. — Rech. Op.*, Vol. 20, n° 4.
- A. BACCINI, J.R. MATHIEU et A.M. MONDOT (1986). — Utilisation conjointe de l'Analyse des Correspondances Multiple et de modélisations. Dans Comparaison et évaluation des approches française et britannique de l'analyse de données complexes. *Publications du Laboratoire de Statistique et Probabilités*, Université Paul Sabatier, N° 02 - 86.
- J.P. BENZECRI (1973). — *L'Analyse des Données*; vol. II : l'Analyse des Correspondances. Dunod, Paris.
- Y.M. BISHOP, J.E. FIENBERG et P.W. HOLLAND (1975). — *Discrete Multivariate Analysis : Theory and Practice*. The M.I.T. Press, Cambridge, U.S.A.
- R. COPPI (1982). — Theoretical and practical problems in the analysis of multidimensional contingency tables. *Metron*, Vol. XL, n° 1-2.
- J.J. DAUDIN et P. TRE COURT (1980). — Analyse Factorielle des Correspondances et modèle log-linéaire : comparaison des deux méthodes sur un exemple. *Revue de Statistique Appliquée*, Vol. 28, n° 1.
- J.M. DEVAUD (1985). — Discrimination et description sur variables qualitatives : un exemple comparatif sur des données réelles. *Revue de Statistique Appliquée*, Vol. 33, n° 2.
- Y. ESCOUFIER (1985). — L'Analyse des Correspondances : ses propriétés et ses extensions. Proceedings of the 45h session, *I.S.I.*, Amsterdam.
- G.W. KASS (1980). — An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, Vol. 29, n° 2.
- L. LEBART, A. MORINEAU et N. TABARD (1977). — *Techniques de la description statistique*. Dunod, Paris.
- A. LECLERC, A. CHEVALIER, D. LUCE et M. BLANC (1985). — Analyses des Correspondances et modèle logistique : possibilités et intérêt d'approches complémentaires. *Revue de statistique appliquée*, Vol. 33, n° 1.
- M. McCULLAGH and J.A. NELDER (1983). — *Generalized linear models*. Chapman and Hall, London.
- J.R. MATHIEU (1981). — Tests of χ^2 in generalized linear model. *Math. Operationsforsch. Statist., ser. Statistics*, Vol. 12, n° 4.
- M. VOLLE (1985). — Analyse des Données. *Economica*, Paris.