REVUE DE STATISTIQUE APPLIQUÉE

JEAN-MARC DEVAUD

# Discrimination et description sur variables qualitatives : un exemple comparatif sur des données réelles

*Revue de statistique appliquée*, tome 33, n° 2 (1985), p. 5-18

<http://www.numdam.org/item?id=RSA_1985__33_2_5_0>

# DISCRIMINATION ET DESCRIPTION SUR VARIABLES QUALITATIVES : UN EXEMPLE COMPARATIF SUR DES DONNEES REELLES

Jean-Marc DEVAUD
*ESCP - CEGAO*
*79, Avenue de la République, 75011 Paris*

RESUME

L'objet de cet article est de comparer les résultats obtenus en utilisant des techniques statistiques différentes sur un même jeu de données. Ces données, concernant des accidents de travail, sont traitées par deux méthodes d'analyse discriminante et par deux méthodes de description sur variables qualitatives. Après une brève présentation de leurs bases méthodologiques, elles sont appliquées aux données et leurs apports respectifs sont comparés.

ABSTRACT

This paper aims at comparing the results obtained using different statistical methods on a given dataset. These data, concerning industrial accidents, are analyzed by two discriminant analysis methods and two description methods for qualitative variables. After a short survey of their methodological basis, these methods are applied to the data and the respective results are compared.

## INTRODUCTION

Les techniques d'analyse statistique multidimensionnelles (ou multivariées) présentent un intérêt tout particulier dans l'étude de très nombreux phénomènes puisqu'elles permettent d'étudier simultanément plusieurs variables. Généralement, face à un problème donné, plusieurs méthodes peuvent être utilisées. Malgré cela, à notre connaissance, il existe peu de travaux comparatifs mettant en évidence les avantages respectifs de certaines méthodes (ou de certains enchainements de méthodes). La conférence de l'AIM qui s'est tenue à Dijon en 1979(1), dans le but de confronter les différentes approches utilisées dans le cas particulier de l'étude des problèmes de santé a toutefois montré une sensibilisation certaine face à cette question.

L'objet de cet article est de comparer quatre méthodes de traitement de données multidimensionnelles à partir de leur utilisation sur un exemple réel. Les apports respectifs de ces méthodes seront mis en évidence après une brève présentation de leurs base méthodologiques d'une part et des données utilisées d'autre part.

(1) Conférence de l'AIM, Dijon, Mars 1979 : Place et limites de l'analyse des données dans la recherche médicale.



*Mots-clés* : **Analyse des correspondances, modèles loglinéaires, regression logistique, Analyse discriminante qualitative.**

*Key-words* : **Correspondence analysis, Loglinear models, logistic regression, Discrete discriminant analysis.**

## MATERIEL ET METHODES

### Les données utilisées

A l'origine, cette étude a pour cadre les travaux entrepris dans une grande entreprise nationale dans le but de mieux cerner les problèmes de santé de ses personnels. A cette fin, des données ont été recueillies concernant l'absentéisme médical, les invalidités, les décès et les accidents de travail (A.T.). Dans le domaine des A.T., des analyses de nature purement descriptives ont été effectuées – tris à plat et tris croisés sur des variables socio-professionnelles, pour l'essentiel. Elles ont conduit à se poser d'autres questions. Ainsi, en particulier, les accidents de trajet (domicile-travail) et les accidents de travail en service pouvant sembler être le reflet de phénomènes assez différents, il a paru utile d'étudier ce problème plus finement.

Tous les traitements présentés ont été effectués à partir du fichier des A.T. ayant eu lieu en 1978. Celui-ci comprend tous les A.T. suivis d'un arrêt de travail, soit 3 898 accidents. Chaque enregistrement de ce fichier correspond à un accident; il contient des renseignements concernant la victime (âge, sexe, type d'emploi, etc.) et l'accident (date, heure, lésions entrainées, type, circonstances, etc.).

Dans le travail présenté ici, on s'est intéressé à la liaison éventuelle entre la variable type d'accident et les variables âge, sexe et niveau hiérarchique, la variable type d'accident prenant deux modalités (accident de travail en service ou accident de trajet), la variable âge quatre (20-30 ans, 30-40 ans, 40-50 ans, 50-60 ans), la la variable niveau hiérarchique trois (agent d'exécution, agent de maîtrise, cadre) et la variable sexe deux (homme ou femme). Les individus mal codés ont été éliminés ainsi que les accidents de circulation pendant le travail, qui constituaient la troisième modalité de la variable type d'A.T. mais dont la codification paraissait sujette à caution. Il restait alors un fichier initial d'environ 3 000 accidents "utilisables". Le fichier de travail servant de support dans la suite a été constitué par un tirage aléatoire de 50 % des enregistrements du fichier initial, ce qui a conduit à travailler sur une population de 1541 accidents.

Le jeu de données ainsi constitué présente deux avantages. D'une part, sur le plan méthodologique, le problème considéré peut être abordé par deux approches différentes. La première consiste à privilégier la variable type d'A.T. et à l'"expliquer" par les variables âge, sexe et niveau hiérarchique; on se place alors dans le cadre des méthodes "explicatives" sur variables explicatives qualitiatives et, plus précisément, de la discrimination sur variables qualitatives puisque la variable "expliquée" est également qualitative. La seconde approche consiste à mettre toutes les variables sur le même plan et à tirer le maximum d'"information" du tableau de données alors obtenu; on est ainsi dans une optique qu'on peut appeler "description" mutlidimensionnelle sur variables qualitatives. D'autre part, comme on l'a dit, l'analyse porte sur un phénomène connu, ce qui permet donc d'axer l'étude sur l'aspect statistique sans risquer d'introduire de biais issus d'une mauvaise connaissance des données.

### Les méthodes statistiques multidimensionnelles utilisées

Le problème considéré permet donc d'envisager un traitement statistique par une approche "descriptive" ou par une approche "explicative". Les méthodes utilisées sont présentées en insistant davantage sur celles qui sont, semble-t-il, moins connues.

En ce qui concerne les méthodes qualifiées plus haut de "*descriptives*" deux approches émergent : l'analyse des correspondances multiples et les modèles loglinéaires.

C'est en 1935, dans les travaux de HORST [14], qu'est apparue pour la première fois l'approche appelée plus tard analyse des correspondances multiples. Considérablement développée par GUTTMAN [12] et BENZECRI [2], elle peut aussi être considérée comme un cas particulier de l'analyse canonique généralisée (CARROLL [5]). Sans entrer dans les détails de cette méthode, aujourd'hui bien connue, on se contentera de rappeler que son but est de représenter "au mieux" n individus repérés sur p variables qualitatives, ainsi que les modalités de ces variables dans un espace de dimension réduite donnée. Par construction, un point de cet espace correspondant à une modalité sera, à un facteur d'échelle près, le barycentre des points associés aux individus prenant cette modalité. Une présentation plus complète de cette méthode figure dans [20], [22] et [15].

Développés pendant ces vingt dernières années dans les pays anglo-saxons, les modèles loglinéaires permettent d'analyser des tables de contingence multidimensionnelles et, en ce sens, généralisent le test du khi-deux de contingence au cas où l'on étudié simultanément p variables qualitatives ($p > 2$). Le livre de BISHOP, FIENBERG et HOLLAND [3] constitue une synthèse très complète sur cette approche.

Considérons par exemple une table à quatre entrées constituée à partir de quatre variables qualitatives dont aucune ne joue un rôle privilégié par rapport aux autres. L'effectif $x_{ijk\ell}$ observé dans une case quelconque de cette table ($i = 1, \ldots I$; $j = 1, \ldots J$; $k = 1, \ldots K$; $\ell = 1, \ldots L$) peut être considéré comme la réalisation d'une variable aléatoire $X_{ijk\ell}$, dont l'espérance est $m_{ijk\ell}$. Il est possible de décomposer additivement $\log m_{ijk\ell}$, sous la forme :

$$\log m_{ijk\ell} = u_0 + u_1^{(i)} + u_2^{(j)} + u_3^{(k)} + u_4^{(\ell)} + u_{12}^{(ij)} + u_{13}^{(ik)} + u_{14}^{(i\ell)} + u_{23}^{(jk)} + u_{24}^{(j\ell)} + u_{34}^{(k\ell)} + u_{123}^{(ijk)} + u_{124}^{(ij\ell)} + u_{234}^{(jk\ell)} + u_{1234}^{(ijk\ell)}$$

où :

- $u_0$ est la moyenne des $\log m_{ijk\ell}$ ;
- $u_1^{(i)}$ est l'apport de la $i^{ème}$ modalité de la $1^{ère}$ variable à $\log m_{ijk\ell}$ en plus de $u_0$ ; semblablement pour $u_2^{(j)}$, $u_3^{(k)}$, $u_4^{(\ell)}$ ;
- $u_{12}^{(ij)}$ est l'apport de la conjonction des modalités i et j des variables respectives 1 et 2 à $\log m_{ijk\ell}$. Ce terme est appelé terme d'interaction du $1^{er}$ ordre des variables 1 et 2 pour les modalités i et j. L'interprétation des autres termes d'interaction du $1^{er}$ ordre est bien sûr identique ;
- $u_{123}^{(ijk)}$ représente l'apport de la conjonction des modalités i, j et k des variables respectives 1, 2 et 3 à $\log m_{ijk\ell}$. On l'appelle terme d'interaction d'ordre 2, ainsi que les autres termes de même nature. Semblablement, $u_{1234}^{(ijk\ell)}$ sera appelé terme d'interaction d'ordre 3.

La décomposition ci-dessus est évidemment inspirée de l'analyse de la variance et elle est unique si l'on ajoute des contraintes sur les valeurs des paramètres u. En général, on impose : $\sum_i u_1^{(i)} = \sum_j u_2^{(j)} = \sum_k u_3^{(k)} = \sum_\ell u_4^{(\ell)} = 0$

$\sum_i u^{(ij)} = \sum_j u^{(ij)} = 0$, de même pour les termes d'interaction $u_{13}$, $u_{14}$, $u_{23}$, $u_{24}$, $u_{34}$ et, en généralisant, pour les termes d'interaction d'ordres 2 et 3.

L'intérêt de la décomposition de $\log m_{ijk\ell}$ tient à ce que la nullité de certains termes d'interaction s'interprète en termes d'indépendance entre variables de la table. Ainsi, par exemple :

– La nullité de tous les termes d'interaction est équivalente à l'indépendance mutuelle entre toutes les variables constituant la table.

– L'indépendance des variables 1 et 2 conditionnellement aux variables 3 et 4 est équivalente à la nullité des $u_{12}^{(ij)}$ et des $u_{123}^{(ijk)} + u_{124}^{(ij\ell)} + u_{1234}^{(ijk\ell)}$.

Cependant les $m_{ijk\ell}$ sont inconnus et doivent être estimés. Ils seront estimés à partir d'un modèle loglinéaire particulier dont certains paramètres seront supposés nuls, c'est-à-dire sous une hypothèse particulière d'indépendance éventuellement conditionnelle entre les variables de la table. On pourra donc, en estimant tous les $m_{ijk\ell}$ sous cette hypothèse, construire une table théorique associée à l'hypothèse choisie qui sera confrontée à la table observée. Cette confrontation sera faite par un test qui permettra de juger de la nullité de la "distance" entre la table théorique et la table observée. Selon le résultat, on rejettera ou non l'hypothèse nulle, c'est-à-dire celle correspondant au modèle loglinéaire choisi. Plus précisément la procédure utilisée sera celle du test du rapport de vraisemblance – cf. RAO [18]. La décomposition additive de $\log m_{ijk\ell}$, permet donc de traduire une hypothèse et ainsi de construire la table théorique correspondante comme dans un test du khi-deux de contingence usuel.

Pour les méthodes de *discrimination*, le choix est moins aisé car le domaine est encore largement ouvert. A cet égard, on pourra consulter les travaux de synthèse de NAKACHE [17], GOLDSTEIN et DILLON [11] et ceux, plus ponctuels, de SAPORTA [19] et DAUDIN [7].

Les voies choisies ici conduisent à estimer les probabilités conditionnelles d'appartenance aux classes de la variable à expliquer. L'une consiste à trouver ces estimations par une fonction linéaire des variables explicatives (ou de leurs indicatrices), l'autre par une fonction logistique.

L'objectif est d'expliquer l'appartenance d'un individu à l'une des deux classes $E^1$ ou $E^2$, définies par les deux modalités d'une variable $\mathcal{Y}$, à l'aide de p variables $X_1, \ldots X_p$. Les $X_i$ pourront être des variables quantitatives ou les indicatrices associées à des variables qualitatives convenablement choisies pour éviter des redondances. Si le $i^{ème}$ individu de la population observée est caractérisé par le vecteur $(x_{1i}, \ldots, x_{pi})$ des p valeurs prises par $X_1, \ldots, X_p$ sur cet individu, on peut supposer que la probabilité $\pi_i$ pour qu'il appartienne à $E^1$ est fonction de $(x_{1i}, \ldots, x_{pi})$. Or $\pi_i$ peut être interprétée comme le paramètre de la variable de Bernoulli $Y_i^1$ prenant la valeur 1 si le $i^{ème}$ individu appartient à $E^1$ et 0 sinon. Les réalisations de $Y^1$ sur l'échantillon ne sont autres que les valeurs prises par l'indicatrice associée à la première modalité de $\mathcal{Y}$.

Comme $\pi_i = E(Y_i^1)$, on peut envisager d'effectuer une régression linéaire de l'indicatrice $Y^1$ de $\mathcal{Y}$, associée à $E^1$ sur $X_1, \ldots, X_p$, puisque la régression linéaire a pour but d'estimer $E(Y_i^1/X_1 = x_{1i}, \ldots, X_p = x_{pi})$. Cette idée consistant à régresser les indicatrices de la variable expliquée sur les variables explicatives a été proposée en 1973 dans [1] sous le nom de Multivariate Nominal scale Analysis (MNA). Elle permet d'obtenir, pour le $i^{ème}$ individu, une probabilité estimée d'appartenance à $E^1$, $\hat{y}^1$, $\hat{y}_1 = \hat{b}_0^1 + \hat{b}_1^1 x_{1i} + \ldots + \hat{b}_p^1 x_{pi}$.

Plus généralement, lorsqu'on discrimine sur k classes, on obtient, en régressant chacune des k indicatrices associées à $\mathcal{Y}$ les probabilités estimées d'appartenance à chacune des k classes $\hat{y}_i^1, \ldots, \hat{y}_i^k$, avec $\hat{y}_i^h = \hat{b}_0^h + \hat{b}_{1x_{1i}}^h + \ldots + \hat{b}_{px_{pi}}^h$, $h = 1, \ldots, k$.

Par construction, les $\hat{y}_i^h$ vérifient $\sum_h \hat{y}_i^h = 1$, mais il n'est pas certain que tous les $\hat{y}_i^h$ appartiennent à [0, 1]. De plus, l'estimation par les moindres carrés ordinaires n'a plus son caractère optimal puisqu'on constate que, par exemple, l'hypothèse d'homoscédasticité n'est plus vérifiée, les $Y_i^h$ étant bernoulliennnes. Il est toutefois possible de donner une interprétation géométrique à ce problème en représentant chaque individu par un point de coordonnées $(\hat{y}_i^1, \ldots, \hat{y}_i^k)$ dans un espace de dimension $k - 1$ $\left(\text{puisque } \sum_h \hat{y}_i^h = 1\right)$. Ainsi, lorsque $k = 2$, tous les individus peuvent être représentés par des points sur une droite et on peut aussi visualiser une modalité d'une variable explicative qualitative par le point moyen sur la droite, des individus prenant ladite modalité. Il est alors possible de juger du caractère discriminant de certaines modalités, comme on le verra plus loin. Cette approche géométrique, développée dans [10], permet également de donner des règles de réaffectation des individus, les $\hat{y}_i^h$ apparaissent ainsi plutôt comme des indicateurs d'appartenance aux classes, le critère des moindres carrés étant alors un simple critère d'ajustement.

Pour assurer l'appartenance des $\hat{y}_i^h$ à [0, 1] on peut envisager d'estimer ces probabilités à l'aide d'une fonction logistique. Autrement dit, si par exemple $k = 2$, on posera $\pi_i = [1 + \exp - (a_o + a_1 x_{1i} + \ldots + a_p x_{pi})]^{-1} \cdot \pi_i$ appartiendra alors par construction à [0, 1] et $a_o, a_1, \ldots, a_p$ seront des paramètres à estimer. L'estimation se fait en général par le maximum de vraisemblance et on peut tester la validité du modèle à l'aide du logarithme du rapport de vraisemblance – cf. [18] –. Cette approche, connue sous le nom de régression logistique, est développée dans [6] et [16], des justifications sur l'emploi de la fonction logistique sont données dans [8].

## RESULTATS

Les méthodes citées plus haut ont donc été appliquées aux données d'accidents de travail déjà évoquées. On a utilisé le programme MULTM [15] pour le traitement par l'analyse des correspondances, P4F des BMDP [4] pour les modèles loglinéaires, LOGIST de SAS [21] pour la régression logistique et GLM, du même logiciel SAS pour MNA.

Les résultats exposés ci-après sont présentés de façon plus détaillée dans [9].

### Application de l'analyse des correspondances multiples

L'analyse effectuée a porté sur le tableau disjonctif complet associé aux données, les quatre variables étudiées étant ainsi mises sur le même plan, cela afin que les résultats obtenus soient comparables à ceux fournis par les modèles loglinéaires.

TABLEAU 1a

| | Contributions absolues | | | | Contributions relatives | | | |
|---|---|---|---|---|---|---|---|---|
| | F1 | F2 | F3 | F4 | F1 | F2 | F3 | F4 |
| **Age** | | | | | | | | |
| 20-30 ans | 6,7 | 18,9 | 6,9 | 4,5 | 0,15 | 0,34 | 0,11 | 0,07 |
| 30-40 ans | 0,6 | 1,5 | 29,2 | 9,0 | 0,01 | 0,03 | 0,43 | 0,12 |
| 40-50 ans | 0,0 | 17,0 | 5,0 | 27,5 | 0,00 | 0,27 | 0,07 | 0,36 |
| 50-60 ans | 6,6 | 6,6 | 0,3 | 53,8 | 0,12 | 0,10 | 0,00 | 0,64 |
| **Sexe** | | | | | | | | |
| Homme | 2,9 | 1,5 | 0,1 | 0,1 | 0,46 | 0,19 | 0,01 | 0,01 |
| Femme | 27,8 | 14,4 | 0,7 | 0,9 | 0,46 | 0,19 | 0,01 | 0,01 |
| **Catégorie** | | | | | | | | |
| Exécution | 4,9 | 10,1 | 1,4 | 0,1 | 0,22 | 0,37 | 0,04 | 0,00 |
| Maîtrise | 6,5 | 16,9 | 13,0 | 1,0 | 0,14 | 0,29 | 0,20 | 0,01 |
| Cadres | 6,5 | 4,4 | 41,0 | 2,4 | 0,10 | 0,06 | 0,45 | 0,02 |
| **Type d'A.T.** | | | | | | | | |
| Trajet | 29,6 | 6,7 | 1,9 | 0,5 | 0,56 | 0,10 | 0,03 | 0,01 |
| Service | 7,9 | 1,8 | 0,5 | 0,1 | 0,56 | 0,10 | 0,03 | 0,01 |

TABLEAU 1b

| | Valeur propre | Pourcentage | Pourcentage cumulé |
|---|---|---|---|
| 1 | 0,372 | 21,28 | 21,28 |
| 2 | 0,303 | 17,35 | 38,63 |
| 3 | 0,268 | 15,35 | 53,98 |
| 4 | 0,247 | 14,17 | 68,15 |
| 5 | 0,229 | 13,13 | 81,28 |
| 6 | 0,180 | 10,32 | 91,60 |

L'étude des tableaux 1a et 1b et de la figure 1 fait apparaître que le premier axe factoriel rend surtout compte des variables sexe et type d'A.T. et, plus particulièrement, des modalités "femme" et "trajet". Les autres axes rendent essentiellement compte de la catégorie hiérarchique et de l'âge. Globalement, il semble apparaître une liaison entre l'âge et le niveau hiérarchique, l'âge étant d'autant plus élevé que la responsabilité est grande. Ce phénomène est connu puisque la promotion est en partie déterminée par l'ancienneté dans cette entreprise. En outre, l'analyse met également en évidence une forte tendance des femmes à être atteintes par des accidents de trajet lorsqu'elles sont victimes d'accidents.

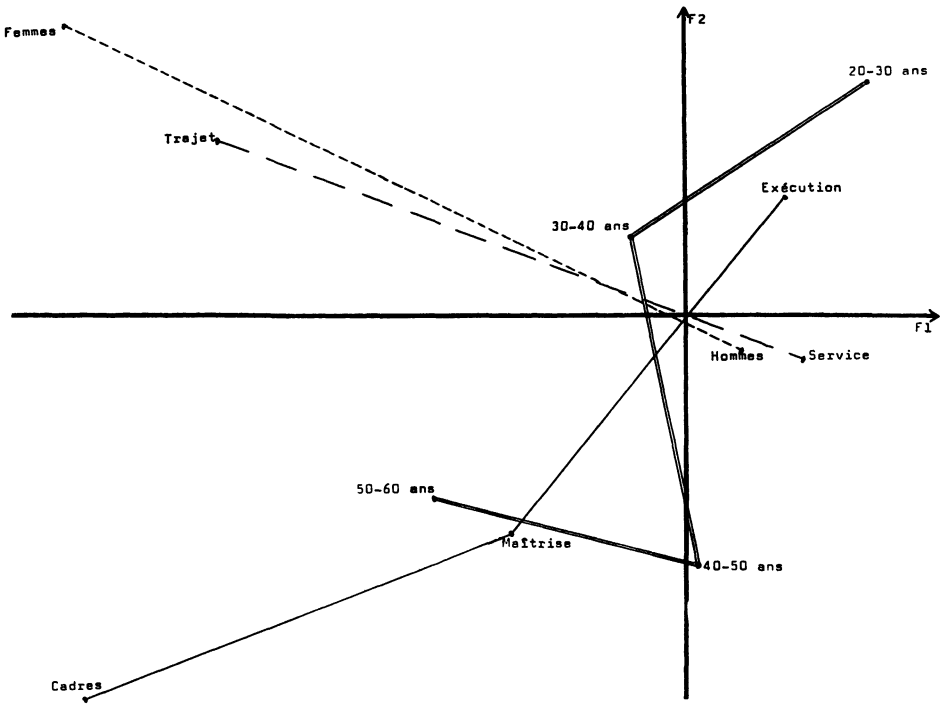


Figure 1. – Premier plan factoriel de l'analyse des correspondances.

## Application des modèles loglinéaires

Avec cette approche, se pose le problème du choix du modèle. En effet, si une hypothèse sur la table de contingence à quatre entrées étudiée peut se traduire par un modèle loglinéaire il n'est pas aisé de formuler une telle hypothèse a priori, lorsque le problème est mal connu. L'usage est alors de construire un modèle par une procédure de type pas à pas et de traduire ensuite le modèle auquel on a abouti en termes d'indépendance entre variables. C'est ce qui a été fait ici. Les paramètres introduits dans le modèle étaient ceux dont la présence, en plus des autres, était la plus significative au sens du test du logarithme du maximum de vraisemblance sur ce paramètre, l'introduction de nouveaux paramètres dans le modèle cessant lorsque celui-ci était globalement satisfaisant, là-encore au sens du test du rapport de vraisemblance.

Le modèle retenu contient tous les paramètres représentant l'apport propre de chaque variable à long $m_{ijk\ell}$ et les termes d'interaction d'ordre 1 reflétant la conjonction des variables âge et catégorie hiérarchique, type d'A.T. et sexe, type d'A.T. et catégorie hiérarchique :

$$\log m_{ijk\ell} = u_o + u_A^{(i)} + u_C^{(j)} + u_S^{(k)} + u_T^{(\ell)} + u_{AC}^{(ij)} + u_{CT}^{(j\ell)} + u_{ST}^{(k\ell)}$$

avec : A : Age, C : Catégorie hiérarchique, S : Sexe, T : Type d'A.T.

Le test du rapport de vraisemblance montre que la table de contingence observée n'est pas significativement différente de la table théorique construite à partir du modèle précédent. On n'a donc pas de raison de rejeter ce modèle.

Ce modèle traduit en particulier que l'âge et le type d'A.T. sont indépendantes conditionnellement au sexe et à la catégorie hiérarchique.

L'observation des valeurs estimées des paramètres d'interaction (cf. tableaux II, III, IV) fait apparaître trois autres constatations – là encore le lecteur trouvera des justifications et des compléments dans [9] –. L'âge et la catégorie hiérarchique sont liés : les personnels les plus jeunes occupent plutôt des fonctions d'agent d'exécution, les plus âgés se trouvent le plus souvent parmi les cadres – cf. $u_{AC}^{(20\text{-}30,EXE)} = 1{,}01$; $u_{AC}^{(50\text{-}60,CAD)} = 0{,}67$. Le sexe et le type d'accident sont liés : les femmes ont une forte propension aux accidents de trajet – $u_{ST}^{(FEM,TRA)} = 0{,}63$. Si ces deux dernières constatations apparaissent dans l'analyse des correspondances, tel n'est pas le cas de la suivante : la catégorie hiérarchique et le type d'A.T. sont liés ; les cadres sont plutôt atteints par des accidents de trajet, les agents d'exécution par des accidents en service – $u_{CT}^{(CAD,TRA)} = 0{,}59$ : $u_{CT}^{(EXE,SER)} = 0{,}41$.

TABLEAU II

Paramètres d'intéraction âge x catégorie hiérarchique

| Age | Catégorie hiérarchique | | |
|---|---|---|---|
| | EXE | MAI | CAD |
| 20-30 ans | 1,01 | 0,02 | – 1,03 |
| 30-40 ans | 0,12 | 0,16 | – 0,28 |
| 40-50 ans | – 0,47 | – 0,17 | 0,64 |
| 50-60 ans | – 0,66 | – 0,01 | 0,67 |

TABLEAU III

Paramètres d'intéraction sexe x type d'A.T.

| Sexe | Type d'accident | |
|---|---|---|
| | TRA | SER |
| Hommes | – 0,63 | 0,63 |
| Femmes | 0,63 | – 0,63 |

TABLEAU IV

Paramètres d'interaction type d'A.T. x catégorie hiérarchique

| Type d'accident | Catégorie hiérarchique | | |
|---|---|---|---|
| | EXE | MAI | CAD |
| TRA | – 0,41 | – 0,18 | 0,59 |
| SER | 0,41 | 0,18 | – 0,59 |

**Application de la régression logistique**

Il s'agit d'estimer la probabilité d'accident de trajet pour un individu de la population étudiée. Le nombre restreint de variables explicatives a permis d'établir le "meilleur" modèle pour prédire la variable expliquée : type d'accident. Le modèle ne contenant que des variables ayant un rôle significatif et restituant au mieux la probabilité d'accident de trajet au sens du rapport de vraisemblance comporte pour variables explicatives le sexe et la catégorie hiérarchique. Pour effectuer les tests, la procédure utilisée a été celle du logarithme du rapport de vraisemblance. Ainsi, si on posait comme hypothèse nulle que l'âge n'avait pas d'influence sur le type d'A.T., en plus du sexe et de la catégorie hiérarchique, le logarithme du rapport des vraisemblances sous l'hypothèse nulle et l'hypothèse alternative valait 3,68, quantité non significativement différente de 0, confrontée à un khi-deux à 3 degrés de liberté (le nombre de degrés de liberté est égal au nombre d'indicatrices non redondantes associées à la variable âge, soit 4-1), ce qui conduit à ne pas rejeter l'hypothèse nulle.

Le modèle contenant les variables sexe et catégorie hiérarchique est significatif au sens où le logarithme du rapport de vraisemblance "global" vaut 210,90, quantité qui, confrontée à u khi-deux à 3 degrés de liberté est significativement différente de 0 ($p < 10^{-5}$). L'hypothèse nulle stipulant que les deux variables n'ont pas d'influence sur le type d'accident doit donc être rejetée.

Il apparait donc que le sexe et la catégorie hiérarchique permettent d'"expliquer" le type d'A.T., mais que l'âge n'apporte pas d'information en plus de ces deux dernières.

Les valeurs estimées des paramètres de chaque indicatrice figurent ci-après :

| | | |
|---|---|---|
| Constante = − 1,9 | Sexe-homme = 0 | Sexe-femme = 2,5 |
| Catégorie-exé = 0 | Catégorie-maî = 0,4 | Catégorie-cad = 2,0 |

Les coefficients associés aux modalités "homme" et "exécution" ont été posés égaux à 0 pour éviter la dégénérescence du problème. On remarque la forte propension des cadres et des femmes pour les accidents de trajet. Les probabilités estimées d'accident de trajet pour les différentes sous-populations sont alors :

| | | | |
|---|---|---|---|
| Hommes-exécution : | 0,13 | Femmes-exécution : | 0,64 |
| Hommes-maîtrise : | 0,19 | Femmes-maîtrise : | 0,73 |
| Hommes-cadres : | 0,50 | Femmes-cadres : | 0,91 |

**Application de MNA**

Pour éviter des inconvénients liés à la grande disparité des effectifs des classes, on a effectué des régressions pondérées où chaque individu avait un poids égal à l'inverse de l'effectif de sa classe – cf. [10]—. MNA ne permettant pas, en toute rigueur, de faire les tests usuels, les trois variables explicatives ont été introduites simultanément dans le modèle, le but étant toujours d'estimer la probabilité d'accident de trajet. Les coefficients estimés figurent ci-après :

CONSTANTE : 0,77

AGE *20-30 ans* : 0,01 *30-40 ans* : − 0,01
*40-50 ans* : − 0,04 *50-60 ans* : 0,04

CATEGORIE HIERARCHIQUE *Exé* : − 0,17 *Maî* : − 0,07 *Cad* : 0,24

SEXE *Homme* : − 0,24 *Femme* : 0,24

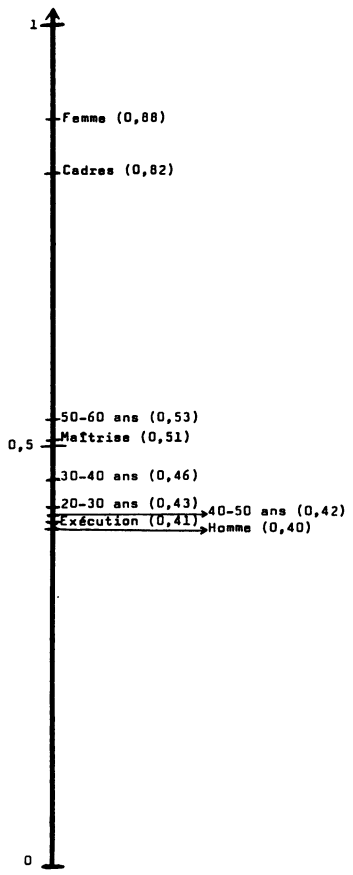


Figure 2.

Pour éviter la dégénérescence, on a imposé à la somme des coefficients des indicatrices associées à une même variable d'être nulle.

Il apparaît que les femmes et les cadres ont une forte tendance à avoir des accidents de trajet. L'âge semble avoir un pouvoir discriminant assez faible, les coefficients de ses indicatrices étant faibles en valeur absolue, en regard des autres paramètres estimés.

Les indicateurs d'appartenance aux classes $\hat{y}_i^1$ ("probabilités" estimées n'appartenant pas forcément à [0,1]) se calculent aisément à partir des valeurs de paramètres ci-dessus. On a ainsi, entre autres valeurs des $\hat{y}_i^1$ : 0,32 pour les hommes-exécution de 40-50 ans, 0,80 pour les femmes-exécution du même âge, 0,73 pour les hommes-cadres et 1,21 pour les femmes-cadres de la même classe d'âge.

On peut représenter les modalités des différentes variables sur une droite rapportée à un repère tel qu'un individu fictif ayant une probabilité estimée nulle d'accident de trajet soit positionné en l'origine et un individu ayant une probabilité estimée d'accident de trajet égale à 1 ait pour abscisse 1 (Fig. 2). Il parait alors naturel d'affecter un nouvel individu à la classe des accidents de trajet (resp. en service) si son abscisse sur cet axe est supérieure (resp. inférieure) à 0,5. Cette règle

possède d'autres justifications –cf. [10] –, elle nous aide également à interpréter le positionnement d'une modalité sur l'axe. En effet, lorsque, par exemple, une modalité est représentée par un point d'abscisse proche de 1, cela signifie que, en moyenne, les individus prenant cette modalité ont des abscisses proches de 1, donc sonnt affectés aux accidents de trajet.

Une telle modalité sera donc sans doute liée à l'apparition d'un accident de trajet. Ainsi, les positions des modalités "femmes" et "cadres" sur l'axe traduisent une forte propension de ces populations à être atteintes par des accidents de trajet. En revanche, les modalités de la variable âge sont représentées par des points d'abscisse proche de 0,5, ce qui traduit un caractère peu discriminant de cette variable.

## DISCUSSION-CONCLUSIONS

Les différences d'ordre méthodologique que présentent les quatre méthodes utilisées sont synthétisées dans le tableau V.

TABLEAU V

| Méthode | Type de problématique | Type de variables | Présence de tests pour le type de problème posé | Présence de représentations graphiques |
|---|---|---|---|---|
| MNA | Explication | Expliquée : qualitative | non | oui |
| Régression logistique | Explication | Explicatives : qualitatives ou quanti - tatives | oui | non |
| Analyse des correspond. | Description | Toutes qualitatives | non | oui |
| Modèles loglinéaires | Description | | oui | non |

En ce qui concerne l'application étudiée, l'analyse des correspondances (ACM) a fait apparaître une liaison entre le type d'A.T. et le sexe (surtout la propension des femmes à avoir des accidents de trajet) et une autre entre l'âge et le niveau hiérarchique (niveau hiérarchique croissant avec l'âge). Les modèles loglinéaires ont mis en évidence les mêmes phénomènes mais en plus, ils ont mis en lumière une liaison entre le type d'A.T. et la catégorie hiérarchique (surtout la propension des cadres aux accidents de trajet). Ils ont également permis de tester l'hypothèse d'indépendance entre l'âge et le type d'A.T., conditionnellement au sexe et à la catégorie hiérarchique. Cette procédure nous a conduit à ne pas rejeter cette hypothèse.

Le fait que, contrairement aux modèles loglinéaires, l'ACM conduise à analyser les données à partir des liaisons entre variables deux à deux explique peut-être en partie les différences entre les résultats obtenus par les deux méthodes. D'une manière générale l'ACM risque donc d'occulter des caractéristiques que les modèles loglinéaires font apparaître. Par ailleurs, il est clair que ces deux approches ne répondent pas aux mêmes préoccupations, bien qu'elles s'appliquent aux mêmes données. Les modèles loglinéaires sont bien adaptés lorsque le problème est suffisamment connu pour qu'on puisse formuler des hypothèses. On a vu qu'ils permettaient alors de faire commodément les tests correspondants. En revanche, face à un problème nouveau, la construction du modèle adéquat, de proche en proche, peut s'avérer délicate et les conclusions risquent d'être difficiles à formuler. A cet égard, il faut noter que tous les cas pratiques ne connaissent par un déroulement aussi "heureux" que l'exemple présenté.

Lors de l'étude d'un phénomène mal connu, il peut donc paraître utile d'effectuer d'abord une ACM qui devrait condure à un dégrossissage efficace du problème. Ensuite, afin d'affiner l'analyse et, muni de ces premiers éléments, il sera alors sans doute pertinent de faire appel aux modèles loglinéaires. Un tel enchainement pourrait également être utile lorsque le nombre de variables est élevé.

Les méthodes explicatives, n'ont pas permis de déceler la liaison entre l'âge et la catégorie hiérarchique. Cela n'a rien d'étonnant puisque leur but est de mettre en évidence les liaisons entre la variable expliquée et les variables explicatives mais non les liaisons entre variables explicatives. A cela près, la régression logistique a apporté des conclusions identiques à celles issues des modèles loglinéaires : liaison entre le type d'A.T. et la catégorie et le sexe, mais caractère non discriminant de l'âge en plus de ces deux dernières variables. On remarquera à ce propos que, dans le cas où toutes les variables explicatives sont qualitatives, estimer un modèle loglinéaire revient à effectuer une régression logistique – cf. HABERMAN [13] –. Ici, il n'y a pas rigoureusement équivalence entre les modèles loglinéaire et logistique choisis, cependant les analogies constatées dans les résultats sont liées à la proximité des deux approches. La propension des femmes et des cadres à avoir des accidents de trajet est soulignée tant par la régression logistique que par MNA. La représentation géométrique de MNA a également permis de classer les modalités "femme" et "cadre" par ordre de propension décroissante et montré le rôle faiblement discriminant de l'âge, sans toutefois donner une réponse aussi tranchée (du moins en apparence) que celle apportée par un test. Il est d'ailleurs utile de rappeler qu'à cet égard, les statistiques utilisées dans les tests effectués en régression logistique et dans les modèles loglinéaires ne suivent qu'approximativement des lois du khi-deux et que lorsque certaines sous-populations ont des effectifs trop faibles, l'approximation est peu adéquate.

Pour finir, on peut constater que, sur le plan métthodologique, la frontière entre méthodes "explicatives" et méthodes "descriptives" apparait comme assez artificielle. On a rappelé ci-dessus les liens existant entre les modèles loglinéaires et la régression logistique lorsque toutes les variables sont qualitatives. Il n'est pas possible de les développer ici, mais ils mettent en évidence que ces deux méthodes sont finalement deux expressions d'une seule et même approche. Semblablement, on sait (cf. [20]) que la régression multiple, qui répond à une problématique d'"explication" est un cas particulier de l'analyse canonique généralisée, tout comme l'analyse en composantes principales et l'analyse des correspondances qui, elles, répondent à une problématique de description.

Tout en étant conscient de ces liens, il semble toutefois utile de classer les méthodes à partir de ces niveaux d'objectifs de la statistique que sont la "description" et l'"explication" (le troisième étant la "prévision"). En effet, ceux-ci correspondent bien à des types de problématiques rencontrées par l'homme d'études lorsqu'il analyse des situations concrètes. Et, dans le domaine des méthodes statistiques multidimensionnelles, il semble important de rester en contact avec la réalité puisque c'est essentiellement pour aider à étudier des situations réelles qu'on fait appel à ces techniques.

## BIBLIOGRAPHIE


[1] F.M. ANDREWS, R.C. MESSENGER. – Multivariate Nominal scale Analysis: a report on a new analysis technique and a computer program. Survey Research Letter, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, 1973.

[2] J.P. BENZECRI. – Histoire et préhistoire de l'analyse des données : l'analyse des correspondances, *Les cahiers de l'analyse des données,* 1977, 2, 55-71.

[3] Y.M. BISHOP, S.E. FIENBERG, P.W. HOLLAND. – *Discrete Multivariate Analysis,* MIT Press, Cambridge, 1975.

[4] BMDP – Biomedical computer programs-P series, University of California Press, Los Angeles, 1981.

[5] J.D. CARROLL. – Generalization of canonical correlation analysis to three or more sets of variables, Proceedings, 76th Annual Convention, American Psychological Association, 1968.

[6] D.R. COX. – *The Analysis of binary data,* Methuen, Londres, 1970.

[7] J.J. DAUDIN. – *Etude de la liaison entre variables aléatoires. Regression sur variables qualitatives,* Thèse de 3e Cycle, Université Paris XI, 1978.

[8] N.E. DAY, D.F. KERRIDGE. – A general maximum likelihood discriminant, *Biometrics,* 1967, 23, 313-323.

[9] J.M. DEVAUD. – *Discrimination et description sur variables qualitatives. Application à des données d'accidents du travail.* Thèse de 3e Cycle, Université Paris IX, 1982.

[10] J.M. DEVAUD. – Discrimination par l'ajustement de variables indicatrices, *Revue de Statistique Appliquée,* 1984,32, 3, 23-41.

[11] M. GOLDSTEIN, W.R. DILLON. – *Discrete Discriminant Analysis,* Wiley, New-York, 1978.

[12] L. GUTTMAN. – The quantification of a class of attributes: A theory and a method of scale construction, in HORST P. et al. ed. *The prediction of personnal adjustment,* New-York, Social Science Research Council, 1941, 319-348.

[13] S.J. HABERMAN. – *The Analysis of Frequency Data,* University of Chicago Press, Chicago, 1974, 303-321.

[14] P. HORST. – Measuring complex attitudes, *Journal of Social Psychology,* 1935, 6, 369-374.

[15] L. LEBART, A. MORINEAU, N. TABARD. – *Techniques de la description statistique,* Dunod, Paris, 1977.

[16] J. LELLOUCH. – Le risque : définition et procédés de calcul, *Revue d'Epidémiologie et de Santé Publique,* 1976, 24, 201-210.

[17] J.P. NAKACHE. – *Méthodes de discrimination sur variables de nature quelconque. Théorie et pratique,* Thèse d'Etat, Université Paris VI, 1980.

[18] C.R. RAO. – *Linear Statistical Inference and its Applications,* 2e ed., Wiley, New-York, 1973, 417-420.

[19] G. SAPORTA. – Une méthode et un programme d'analyse discriminante pas-à-pas sur variables qualitatives, *Colloques IRIA,* Analyse des données et Informatique, 1977, **1**, 201-210.

[20] G. SAPORTA. – *Théorie et méthodes de la statistique,* Technip, Paris, 1978.

[21] SAS. – Supplemental library user's guide, SAS Institute Inc., Cary, North-Corolina, 1980.

[22] M. TENENHAUS, F.W. YOUNG. – Multiple correspondence analysis and the principal components of qualitative data, soumis à *Psychometrika,* 1984.