REVUE DE STATISTIQUE APPLIQUÉE

A. LECLERC
A. CHEVALIER
D. LUCE
M. BLANC

Analyses des correspondances et modèle logistique : possibilités et intérêt d'approches complémentaires

Revue de statistique appliquée, tome 33, n° 1 (1985), p. 25-40

<http://www.numdam.org/item?id=RSA_1985__33_1_25_0>

ANALYSES DES CORRESPONDANCES ET MODELE LOGISTIQUE : POSSIBILITES ET INTERET D'APPROCHES COMPLEMENTAIRES

A. LECLERC*, A. CHEVALIER, D. LUCE, M. BLANC

*INSERM U. 88, 91, Bd de l'hôpital, 75634 Paris Cedex 13

RESUME

On envisage ici, à partir d'un exemple, une utilisation complémentaire de l'analyse des correspondances (A.F.C.) et du modèle logistique : l'A.F.C. est utilisé dans une première étape comme une méthode descriptive, fournissant un cadre à l'application du modèle logistique et aidant au choix des variables à retenir, dans une situation où les données se préteraient mal à utiliser d'emblée le modèle logistique.

SUMMARY

"correspondence analysis and logistic model : possibility and interest of a complementary use"

A complementary use of two statistical methods is considered here, with an exemple : correspondence analysis as a first step, and logistic model for a subgroup of variables. Some results of correspondence analysis are used in order to reduce the number of explanatory variables, or to choice variables to be explained. Some links between the two method are stressed.

Combining an exploratory method (correspondence analysis) and a confirmatory method (logistic model) is suggested as being an adapted solution in the analysis of some survey data, like health survey data.

I. INTRODUCTION

Le modèle logistique permet d'"expliquer" une variable à deux modalités par plusieurs variables qualitatives ou quantitatives, prenant en compte le rôle propre de chacune, et l'effet éventuel des interactions ; le modèle est bien adapté à des situations où on peut facilement isoler une variable à expliquer, et un petit nombre de variables explicatives pertinentes ; ceci est le cas, bien souvent, en épidémiologie où ce modèle est largement utilisé [19, 6].

Si on s'éloigne de ces conditions idéales, on peut être intéressé à utiliser aussi le modèle logistique, mais quelques problèmes peuvent alors se présenter :

Si les variables "à expliquer" sont nombreuses, toutes ne méritent peut-être pas une analyse détaillée, et on peut apprécier de disposer d'éléments facilitant le choix des variables à retenir.

Un manque de connaissance sur le sujet étudié a en général comme conséquence un nombre assez important de variables explicatives dont toutes ne sont peut-être pas à garder : certaines sont redondantes, d'autres ont un pouvoir explicatif faible.

Mots-clés : **Modèle logistique, Analyse des correspondances.**

Dans le cas d'une situation ainsi partiellement exploratoire [22] utiliser directement le modèle logistique n'est pas une bonne solution, même si c'est matériellement possible ; le grand nombre de modèles à essayer amène à multiplier les tests statistiques, avec le risque de trouver des associations significatives sans aucun caractère général ; les effectifs faibles de certaines associations de variables peuvent aussi limiter largement à la fois les possibilités de mise en œuvres et la portée des conclusions ; une solution adaptée peut être alors de procéder à une première étape de description des données de façon à réduire le nombre de variables et à disposer d'informations synthétiques qui serviront de cadre aux modèles ; c'est cette démarche qui est présentée ici sur un exemple.

Bien que le lien soit étroit entre modèle log-linéaire et modèle logistique, notre approche diffère de celle de J.J. DAUDIN ou C. LAURO [8. 15] qui mettent "en concurrence" modèle log-linéaire et analyse des correspondances ; nous n'envisageons pas qu'une méthode remplace l'autre, mais qu'elles se complètent en étant utilisées successivement ; nous tenons aussi à nous placer le plus près possible d'une situation réelle où des contraintes d'effectif limitent la possibilité de certaines investigations, et où le grand nombre de variables justifie pleinement le recours à l'analyse de données.

L'exemple que nous présentons comporte relativement peu de variables, mais des fichiers de taille beaucoup plus importante se prêtent à la même approche [20].

Au-delà de cet exemple nous nous demandons quelles informations utiles peut apporter l'analyse des correspondances vis-à-vis de l'application de modèles logistiques.

II. LES METHODES UTILISEES

Nous avons appliqué l'analyse des correspondances multiples [16], en traitant en variables actives toutes les variables potentiellement à expliquer (ici, un ensemble de variables de santé déclarée), et en variables supplémentaires les variables explicatives (variables socio-démographiques et professionnelles) ; nous discuterons plus loin du choix de cette approche.

Nous allons développer ce qui concerne le modèle logistique, car cette méthode est, en France, moins bien connue :

On considère une variable à expliquer, Z, à deux modalités, (oui-non, Z = 1 pour le oui, O pour le non), et des variables explicatives $X_1^*, \ldots\ldots X_q^*$, quantitatives ou qualitatives, dont on connaît les réalisations sur un échantillon de taille n.

On se restreint ici au cas où les X_j^* sont qualitatives.

On note $\pi = P(Z = 1)$; $X_1 \ldots\ldots X_p$ sont les variables en 0 ou 1 associées à toutes les modalités des variables X_j^* moins une (choisie arbitrairement) par variable. X_0 prend la valeur 1 pour toute observation.

Dans le modèle logistique, π s'exprime en fonction de $X_1 \ldots . X_p$ par une relation de la forme :

$$\text{Log}\ \frac{\pi}{1-\pi} = \sum_{k=0}^{p} \alpha_K X_K \qquad (1)$$

(modèle sans terme d'interaction).

Dans un modèle avec interaction s'ajoutent aux X_K certains de leurs produits.

La quantité de gauche dans (1) est appelée "logit" ; c'est une transformation couramment utilisée pour les pourcentages, qui a l'avantage d'opérer une forme de normalisation, et de fournir un modèle représentant bien un certain nombre de situations observées de relation dose-effet [7].

Certains développements en analyse discriminante amènent aussi à proposer le modèle logistique comme une méthode de discrimination [9, 11, 21]. Malgré cette ressemblance formelle, cependant, les objectifs sont différents : dans le modèle logistique, on étudie les événements rares ; l'objectif de prévision est le plus souvent absent et le modèle n'est pas utilisé comme aide à la décision [6].

Dans le cas où les X_j^* sont qualitatives, le modèle peut être considéré comme une application du modèle log-linéaire à l'ensemble des variables y compris la variable à expliquer [4, 10, 11] ; les deux approches diffèrent en général par leur écriture : il est habituel dans le modèle log-linéaire d'associer un coefficient à chaque modalité, avec des contraintes sur les coefficients comme en analyse de variance ; dans le modèle logistique, la contrainte intervient en fixant arbitrairement à 0 le coefficient d'une modalité par variable. Cette écriture est bien adaptée au cas où une modalité par variable représente le "niveau plancher" d'un risque (par exemple, le risque de cancer du poumon peut être un risque minimum, augmenté par la présence d'un ou plusieurs facteurs de risque).

Il serait erroné de croire que le modèle logistique se réduit à un cas particulier de modèle linéaire ; en effet, les réalisations des probabilités π concernent des groupes et non des sujets, et leur estimation est de ce fait très dépendante des variables qualitatives introduites.

Pratiquement, les coefficients α_K sont estimés par itération, le critère d'ajustement étant le plus souvent celui du maximum de vraisemblance [4, 12, 14]. Comme en modèle linéaire, il paraît difficile de se borner à utiliser la méthode pour estimer ou tester des coefficients et la pratique la plus courante consiste à comparer plusieurs modèles, ce qui pose des problèmes de stratégie auxquels une réponse possible est une procédure de pas à pas [3, 5]. La stratégie adoptée ici est basée sur les éléments suivants :

Dans le modèle logistique il est possible de tester une hypothèse "modèle restreint" H0 comportant un sous ensemble de variables explicatives, contre une hypothèse "modèle large" H1 comportant ces variables plus d'autres, de la façon suivante :

Soit L0 (resp. L1) la vraisemblance sous l'hypothèse H0 (resp. H1). Sous H0 la quantité $-2 \log L0 + 2 \log L1$ suit approximativement un X^2 dont le nombre de degrés de liberté est le nombre de modalités explicatives (indépendantes) présentes sous H1 et non sous H0. A chaque modèle est donc associé, en vue de comparaisons la quantité $-2 \log L$ et le nombre de degrés de liberté (dd1), nombre de modalités explicatives indépendantes dans le modèle.

Les tests de comparaison entre modèles emboités interviennent ici comme indicateurs dans une procédure, par étapes, de choix de modèle de la façon suivante :

Le premier modèle essayé comporte toutes les variables explicatives (sans interaction). Un modèle comportant une variable explicative V de moins lui est comparé et conservé si la différence entre les quantités $-2 \log L$ ne conclut pas à la nécessité de conserver V. La variable V est choisie parmi celles dont on peut penser qu'elles jouent un rôle mineur, au vu des intervalles de confiance des coefficients du modèle complet. (Ceux ci sont donnés par le programme, comme en modèle linéaire).

La procédure se poursuit par essai d'exclusion d'une nouvelle variable, jusqu'à obtenir un modèle qui ne peut être réduit. A partir de celui-ci on envisage enfin l'introduction de termes d'interaction complétant les variables retenues.

La procédure choisie n'assure pas que le modèle final ait des qualités optimales.

III. UN EXEMPLE :

1. Les données analysées

Les données proviennent d'une enquête en milieu de travail, auprès de 444 personnes réparties sur 3 sites professionnels ; une partie des personnes interrogées a un horaire de travail normal ; d'autres travaillent en 2 x 8 (équipes alternantes de jour) ou en 3 x 8 (par roulement le matin, l'après midi, ou la nuit) ; l'intérêt principal de l'étude porte sur les effets du travail en horaires alternants sur la santé au sens large, y compris plaintes ou déclarations de symptomes ; les résultats ont été publiés[13] ; l'analyse qui suit est un complément, dont l'intérêt épidémiologique est un peu limité par l'existence de certains biais que nous n'avons pas essayé de corriger. (La façon dont les questions ont été posées n'est pas entièrement homogène, et il subsiste un "effet enquêteur").

L'intérêt de l'exemple est principalement qu'il illustre une situation fréquemment rencontrée dans l'analyse des questionnaires sur la santé et de façon plus générale, de certains questionnaires centrés sur un thème particulier décrit par des variables multiples, les variables explicatives étant de nature socio- démographique.

Les variables de santé, qui ont été traitées en variables actives dans l'analyse des correspondances multiples, sont l'existence ou non de :

TEN Tension élevée ;
TRE Tremblements ;
ALT Altération de l'état général ;
PAL Palpitations ;
PRE Précordialgie ;
DYS Dyspnée d'effort ;
GAS Gastralgie ;
ULC Ulcère ;
FMU Fatigue musculaire ;
FIN Fatigue intellectuelle ;
TAP Troubles de l'appétit ;
SPP Somnolence après les repas ;
TRA Troubles du transit ;

SOM Troubles du sommeil ;
SEX Troubles sexuels ;
CEP Symptomes céphalées ;
ASO Angoisse somatique ;
APS Anxiété psychique ;
CAR Troubles du caractère ;
MEX Troubles de la sensibilité au monde extérieur ;
DEP Dépression ;
HYP Hypocondrie ;
VIE Fait médicalement vieux pour son âge ;
ALC Consommation d'alcool ;
TAB Consommation de tabac.

d'après un interrogatoire par le médecin du travail. Par la suite, l'identification des modalités de réponse est : 1 pour "non", 2 pour "oui".

Les variables socio-démographiques et professionnelles suivantes ont été placées en variables supplémentaires :

HOR Horaires de travail			HOR 1 horaire normal HOR 2 2 x 8 HOR 3 3 x 8
SIT Site industriel			SIT 1 SIT 2 SIT 3
ANC Ancienneté dans le poste			ANC 1 $\leqslant$ 3 ans ANC 2 4 à 13 ans ANC 3 $\geqslant$ 14 ans.
AGE	Age	AGE 1	$\leqslant$ 25 ans
		AGE 2	26-35 ans
		AGE 3	$\geqslant$ 36 ans
CAT Catégorie hiérarchique			CAT 1 ouvrier CAT 2 Contremaître.

2. Résultats de l'analyse des correspondances multiples

On s'intéresse principalement à la représentation graphique sur le plan 1-2 (Fig. 1) où on a souligné les modalités des variables supplémentaires dont le carré de la distance à l'origine, sur le plan, est supérieure à $5.99/n_j$, où n_j est l'effectif associé à la modalité j ; ceci correspond à une position significativement différente de l'origine au risque 5 % [16].

L'axe 1 oppose les "oui" aux "non", pour toutes les variables de santé. Ceci est une configuration classique [1] ; l'axe 2 sépare deux groupes de symptômes ou de troubles déclarés ; les uns semblent plus liés au site 3, les autres plus liés à l'horaire en 3 x 8 et à l'âge élevé. Toutes les modalités supplémentaires sont éloignées de l'origine (distance^2 $>$ $5.99/n_j$) à l'exception d'ancienneté moyenne, âge moyen et catégorie ouvrier, qui sont des modalités soit médianes soit de poids très important.

Au-delà du plan 1-2, les modalités supplémentaires sont exceptionnellement excentrées, et les modalités de santé qui jouent un rôle important concernent peu de sujets : dyspnée d'effort (17 personnes) et "fait médicalement vieux pour

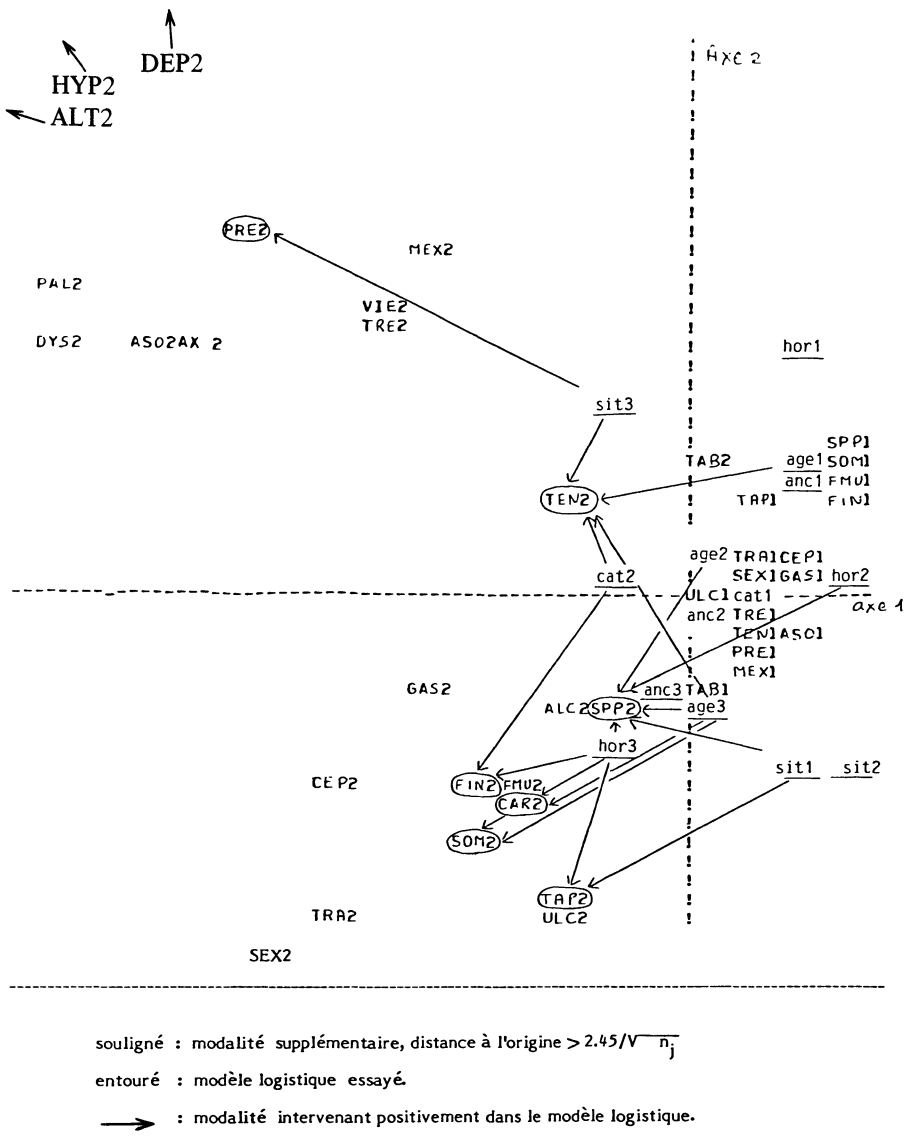

Figure 1. – Plan 1-2 de l'A.F.C.

son âge" (23 personnes) sur l'axe 3, alcoolisme (12 personnes), troubles sexuels (28 personnes) sur l'axe 4.

Les axes 1 et 2 paraissent par ailleurs avoir un pouvoir explicatif nettement plus élevé ; les pourcentages d'inertie $\lambda/\Sigma\lambda$ issus du tableau disjonctif analysé, sont respectivement égaux à 13 %, 9 %, puis 6 %, 5 %, 5 %, les quantités $\lambda^2/\Sigma\lambda^2$, qui sont les pourcentages d'inertie du tableau de Burt correspondant et qui peuvent être considérées comme des mesures mieux adaptées, valent : 31 %, 15 %, puis 7 %, 5 %, 5 % .

Nous avons utilisé la représentation sur le plan 1-2, ainsi que les descriptions des variables seules ou deux à deux (tableaux croisés) pour choisir les

variables à retenir pour le modèle logistique (variables à expliquer et variables explicatives).

Pour les variables à expliquer, le nombre de variables à retenir était limité par des raisons matérielles (lourdeur de la mise en œuvre des modèles logistiques). 7 variables ont été choisies à partir des critères suivants :

– Exclusion des variables dont une analyse un peu plus détaillée est impossible ou sans intérêt, donc soit trop rares, soit dont la signification peut apparaître comme trop subjective : "angoisse" par exemple.

– Inclusion de variables intéressantes du point de vue des liens avec le travail en 3 x 8, pour lesquelles les résultats de l'analyse des correspondances sont insuffisamment précis : fatigue intellectuelle, troubles de l'appétit, somnolence post-prandiale, troubles du sommeil et du caractère.

– Inclusion de variables pour lesquelles l'analyse des correspondances suggère l'existence de liens privilégiés avec certaines variables explicatives : tension élevée et précordialgie ont été ainsi choisis pour confirmer par un modèle ce qui apparaît sur le graphique.

Les variables choisies sont de plus fortement liées aux variables explicatives du point de vue des Khi – 2 des tableaux de contingence (voir tableau 1). La variable

TABLEAU 1

Khi-2 des tableaux de contingence
Modèle logistique retenu

	Effectif Modalité +	Khi - 2					Modèle logistique; variables retenues
		HOR	SIT	ANC	AGE	CAT	
TEN	127	4.0	32.8 *	2.2	13.3 *	18.9 *	SIT, AGE, CAT,(SIT x AGE)
TRE	37	5.7	25.1 *	3.4	2.6	0.2	-
ALT	11	1.7	8.9 *	6.2 *	3.1	6.1 *	-
PAL	29	1.5	8.1 *	1.6	1.8	0.3	-
*PRE	46	2.9	48.3 *	3.7	5.8	4.5 *	SIT
DYS	17	1.9	15.7 *	5.8	2.4	7.0 *	-
GAS	93	5.9	1.4	5.0	7.4 *	0.3	-
ULC	13	0.1	3.2	3.4	4.5	0.0	-
FMU	169	14.1 *	4.8	4.4	3.4	0.1	-
FIN	157	40.9 *	3.3	3.9	2.2	5.7 *	HOR, CAT
TAP	47	57.5 *	9.8 *	0.3	1.1	1.1	HOR, SIT,(HORxSIT)
SPP	267	47.0 *	26.2 *	18.9 *	24.2 *	2.9	HOR, SIT, AGE
TRA	37	6.6 *	0.5	7.5 *	3.5	0.1	-
SOM	160	70.9 *	5.5	13.5 *	5.6	0.2	HOR, AGE
SEX	28	22.3 *	4.4	2.5	8.2 *	2.4	-
CEP	61	10.8 *	11.5 *	5.3	13.3 *	3.0	-
ASO	48	9.9 *	22.1 *	1.0	0.8	6.5 *	-
APS	47	18.4 *	14.8 *	4.7	1.3	4.0	-
CAR	189	89.3 *	0.3	12.9 *	10.3 *	3.9 *	HOR, AGE
MEX	75	9.8 *	100.6 *	11.4 *	26.1 *	1.0	-
DEP	23	0.7	22.9 *	0.2	1.5	1.0	-
HYP	15	0.4	17.3 *	0.1	4.7	0.7	-
VIE	23	0.7	5.9	0.5	3.5	0.0	-
ALC	12	0.1	2.7	2.2	4.1	2.7	-
TAB	192	1.9	31.9 *	30.5 *	32.9 *	4.0 *	-

— Pas d'essai de modèle logistique
* Khi-2 significatif à 5%.

"précordialgie" est un exemple extrême, car le lien avec le site 3 est très important (ce qui peut être du à un effet enquêteur).

Pour les variables explicatives l'analyse permet de réduire le nombre de variables ou modalités :

– Les sites 1 et 2 ont été agrégés car ils ont même projection. Il s'agit d'ailleurs de la même ville et du même médecin du travail ;

– il n'apparaît pas nécessaire de conserver à la fois l'âge et l'ancienneté ; l'importance du lien entre les deux variables est bien vérifiée sur les données ; en conséquence l'ancienneté dans le poste est exclue de l'analyse.

3. Modèles logistiques

Le modèle logistique a été appliqué aux 7 variables retenues, avec les variables explicatives :

Horaire (Jour, 2 x 8, 3 x 8)
Site (1 et 2, ou 3)
Age (1, 2, 3)
Catégorie (ouvrier, contremaître).

La stratégie utilisée pour retenir un modèle est celle qui a été présentée en II ; par exemple, pour "fatigue intellectuelle", les modèles suivants ont été essayés :

Modèle	ddl	-2logL
1. FIN = f(HOR, SIT, AGE, CAT)	6	521.75
2. FIN = f(HOR, SIT, CAT)	4	523.76
3. FIN = f(HOR, CAT)	3	526.56
4. FIN = f(HOR)	2	531.37
5. FIN = f(HOR, CAT, HORxCAT)	5	525.83
6. FIN = Cte	0	574.80

Les modèles 1, 2, 3, par comparaisons successives, amènent à enlever du modèle l'âge et le site ; par contre le modèle 4 apparaît trop réducteur (écart de 4.81 entre les quantités – 2 log L, à comparer à un Khi–2 à 1 ddl).

La comparaison des modèles 3 et 5 permet de juger de l'amélioration apportée par un terme d'interaction ; ici on retient en définitive le modèle 3 :

$$\text{Logit P(FIN)} = -1.76 + \begin{cases} 0 & \text{si HOR 1} \\ 0.72 & \text{si HOR 2} \\ 1.54 & \text{si HOR 3} \end{cases} + \begin{cases} 0 & \text{si CAT 1} \\ 0.50 & \text{si CAT 2} \end{cases}$$

dont la qualité se mesure par comparaison au modèle 6 (écart de 48.24 entre les quantités – 2 log L, à comparer à un X^2 à 3 ddl). Aucun modèle excluant l'horaire (catégorie seule, par exemple) n'a été essayé, le coefficient associé à "horaire 3 x 8" apparaissant partout comme hautement significatif.

TABLEAU 2

Modèles logistiques retenus

VARIABLE	MODELE	-2log L. Modèle (1)	-2log L. Cte seule (2)	Diff. (3)	ddl (4)
TEN Tension élevée	-1.65 + 1.44(SIT3) - 2.09 (AGE2) + 0.30(AGE3) +0.80(CAT2) + 1.30(SIT3 x AGE2) - 0.70(SIT3 x AGE3)	461.09	530.85	69.76	6
PRE Précordialgie	-4.75 + 3.40(SIT3)	235.48	291.09	55.60	1
FIN Fatigue intell.	-1.76 + 0.72(HOR2) + 1.54(HOR3) + 0.50(CAT2)	526.56	574.80	48.24	3
TAP Troubles de l'appétit.	-4.11 + 1.78(HOR2) + 4.06(HOR3) + 1.53(SIT3) -0.51(HOR2 x SIT3) - 2.71(HOR3 x SIT 3)	389.06	475.64	86.58	5
SPP Somnolence après les repas.	-0.81 + 1.60(HOR2) + 1.27(HOR3) - 0.64(SIT3) + 0.76(AGE2) + 0.99(AGE3)	517.43	592.44	75.02	5
SOM Troubles du sommeil.	-2.44 + 0.70(HOR2) + 2.12(HOR3) + 0.43(AGE2) +0.76(AGE3)	494.92	578.37	83.46	4
CAR Troubles du caractère.	-2.26 - 0.03(HOR2) + 2.09(HOR3) + 0.75(AGE2) +1.16(AGE3)	493.53	605.14	109.61	4

(1) et (2) : Quantités -2Log L pour le modèle présenté et pour le modèle "constante seule"

(3) : Différence entre les quantités précédentes.

(4) : Nombre de modalités explicatives indépendantes dans le modèle présenté.

Le même modèle peut s'écrire avec des coefficients liés par des contraintes de centrage comme en analyse de variance :

$$\text{Logit P(FIN)} = -0.70 + \begin{cases} -0.91 & \text{SI HOR 1} \\ -0.19 & \text{SI HOR 2} \\ 0.62 & \text{SI HOR 3} \end{cases} + \begin{cases} -0.15 & \text{si CAT 1} \\ 0.35 & \text{si CAT 2} \end{cases}$$

Pour HOR 3, CAT 2, par exemple, la probabilité estimée est la plus élevée ; elle vaut :

$$\text{P(FIN)} = \frac{\exp(0.27)}{1 + \exp(0.27)} \simeq 0.57$$

On trouvera dans les tableau 1 et 2 tous les modèles retenus ; par rapport aux informations apportées par les tableaux croisés (variables de santé X variables explicatives), il n'est pas étonnant de trouver des ressemblances importantes. Le fait de tenir compte des associations entre variables (ce que fait le modèle logistique) modifie cependant certains effets : l'existence d'une précordialgie déclarée n'est liée qu'au site 3 (en effet sur les 46 observées, 44 sont du site 3 !) ; par contre l'effet de l'âge sur les troubles du sommeil apparaît plus nettement dans le modèle logistique que dans les tableaux croisés.

La représentation graphique de l'analyse des correspondances (figure 1) permet de visualiser partiellement les résultats des modèles logistiques ; sur le plan 1-2, on a relié par une flèche une modalité "explicative" à une variable expliquée, chaque fois que la présence de la modalité, d'après le modèle retenu, augmente la probabilité de présenter le problème de santé en question, donc chaque fois que le coefficient associé à la modalité est positif (avec l'écriture du modèle type "analyse de variance").

Cette représentation graphique semble assez satisfaisante : en général, si une modalité de variable supplémentaire est placée dans la direction d'une variable de santé, elle influe bien (positivement) sur la fréquence de la déclaration : par exemple, travailler au site 3 augmente la fréquence de "précordialgie" (PRE 2) et de "tension élevée" (TEN 2) ; l'horaire 3 x 8 (HOR 3) augmente la fréquence des déclarations de fatigue intellectuelle, troubles du sommeil, de l'appétit, du caractère, et somnolence après les repas.

Parallèlement, les variables de santé sont influencées positivement par les modalités "explicatives" qui se projettent dans les mêmes directions, à quelques exceptions près.

Le graphique ne permet pas cependant de prévoir de façon détaillée les variables explicatives qui jouent un rôle important pour le modèle logistique ; par exemple, il n'apparaît pas graphiquement que, pour la fatigue intellectuelle, l'âge soit moins important dans le modèle logistique que la catégorie hiérarchique, alors qu'il l'est plus pour les troubles du sommeil et du caractère.

IV. COMPARAISON DE DIFFERENTES APPROCHES DESCRIPTIVES, CHOIX DES VARIABLES A EXPLIQUER PAR UN MODELE

L'analyse des correspondances aurait pu, dans cet exemple, être utilisée différemment et on peut envisager au moins trois façons simples de l'utiliser dans cette optique de complémentarité au modèle logistique :

A Variables à expliquer actives (notre exemple) ;
B Variables explicatives actives
C Analyse du tableau de contingence juxtaposé croisant toutes les variables à expliquer avec toutes les variables explicatives.

A priori, l'approche B, qui consiste à projeter les variables à expliquer sur un plan explicatif est la démarche la plus proche de la régression logistique. Ceci pourrait être une raison déterminante de la préférer dans le cas où les variables explicatives sont nombreuses. Cependant, essayée sur nos données, l'analyse B est très pauvre : les variables explicatives ne sont pas si nombreuses, et les liens qu'elles présentent entre elles sont faciles à connaître par ailleurs ; la description des relations entre variables de santé est aussi moins informative, puisque tout est médiatisé par un petit nombre de variables explicatives.

D'autre part, avec un objectif d'explication de type modèle logistique, l'intérêt ne se porte pas principalement sur les relations entre variables explicatives, qui sont plutôt considérées comme une gêne à la compréhension des phénomènes, certaine variables (l'âge par exemple) intervenant comme "tiers facteur", perturbant les relations que l'on veut analyser.

L'approche A est plus complémentaire, car elle apporte des informations sur les relations entre variables à expliquer, au niveau des individus.

L'approche C pourrait apparaître aussi comme une bonne étape préalable : on analyse les relations entre variables à expliquer et variables explicatives, sans tenir compte des inter-relations entre variables explicatives, ce que le modèle logistique fait ensuite. Essayé sur nos données, le résultat est proche de ce qui a été retenu ici [18]. La lecture des graphiques est agréable du fait de la bonne

qualité de représentation des points (on ne traite plus un tableau disjonctif complet) ; sur nos données par exemple le plan 1-2 explique 80 % de l'inertie ; par contre, le fait que l'analyse porte précisément sur les relations entre variables explicatives et variables à expliquer complique l'interprétation : il n'existe aucune notion de distance entre une variable à expliquer et une variable explicative, les deux n'appartenant pas au même espace. Quelque soit l'approche, la mise en œuvre de modèles logistiques apparaît comme complémentaire en répondant à deux objectifs non contradictoires :

– Validation d'un graphique représentant les variables à expliquer, les variables soumises à un modèle logistique étant alors choisies comme formant un sous ensemble couvrant les différentes parties du graphique.

– Recherche de précisions supplémentaires, par exemple différence entre deux variables à expliquer que l'analyse des correspondances distingue mal.

V. QUELQUES LIENS ENTRE ANALYSE DES CORRESPONDANCES MULTIPLES ET MODELE LOGISTIQUE. CHOIX DES VARIABLES EXPLICATIVES EN MODELE LOGISTIQUE

De façon générale, les approches descriptives envisagées ne permettent pas de prévoir l'ensemble des résultats du modèle logistique ; les liens existants entre les deux méthodes permettent cependant une prévision partielle et expliquent la cohérence observée entre les deux approches.

Dans tout ce qui suit, on supposera que les observations sont assez nombreuses pour que les fluctuations aléatoires puissent être négligées.

1. Mesures d'association entre variables

On considère ici la mesure de la relation entre une variable à expliquer Z et une variable explicative X_j toute deux à deux modalités, en modèle logistique ou en analyse des correspondances multiples ; la relation entre Z et X_j est calculée à partir du tableau de contingence :

		X_j	
		1	0
Z	1	a	b
	0	c	d

n = a+b+c+d
= nombre total d'observations

Un résultat classique en modèle logistique, présenté par exemple dans [23] est le suivant :

Dans le modèle logistique exprimant Z ou plus exactement son logit, en fonction de X_j le coefficient associé à X_j est log (ad/cd) ; la quantité ad/cd est appelée odds-ratio.

En analyse des correspondances multiples, la mesure la plus comparable à ce coefficient est le cosinus de l'angle formé par les axes joignant le centre de gravité d'une part à la modalité positive de Z, d'autre part à la modalité positive de X_j (ceci dans R^n muni de la métrique euclidienne usuelle), cosinus égal à $r(Z, X_j)$, coefficient de corrélation entre Z et X_j ; la quantité $n\, r^2\, (Z, X_j)$ est aussi égale au X^2 de contingence associé au tableau précédent.

Les deux quantités Log(ad/cb) et $r(Z, X_j)$ ont toujours même signe ; en effet elles sont nulles pour le même type de tableau (indépendance), et de même signe que $a - (a + c)(a + b)/n$, quantité encore égale à $(ad - bc)/n$. (Voir exemple en Fig. 2).

TABLEAU 2X2 : COSINUS ET LOG DU ODDSRATIO

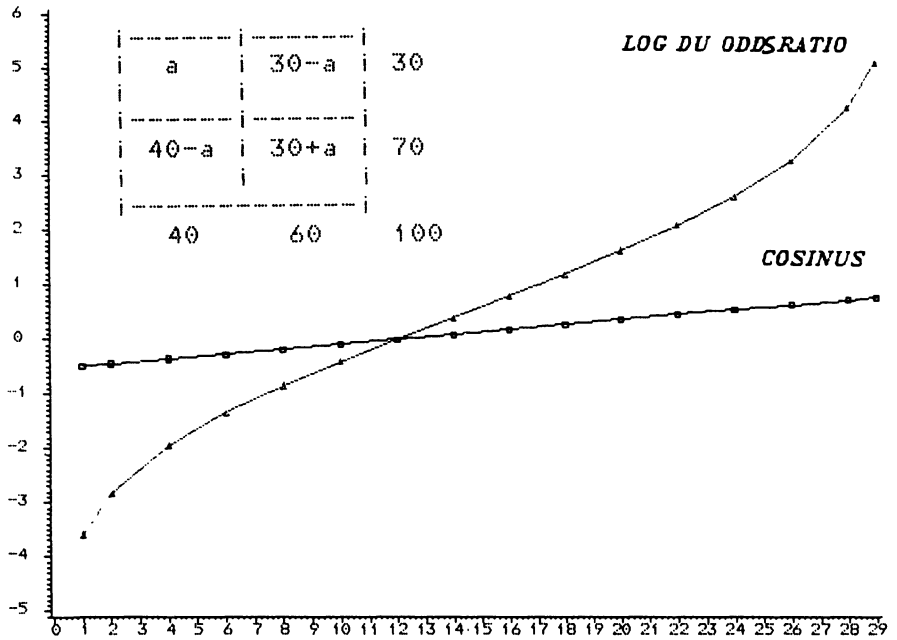

Figure 2

Conséquence pour la lecture des résultats en analyse des correspondances multiples

Si l'axe joignant l'origine à X_j et l'axe joignant l'origine à Z font un angle inférieur à $\pi/2$ (dans R^n, ou sur un plan factoriel avec une bonne qualité de représentation), alors, dans un modèle logistique exprimant Z en fonction de X_j, cette dernière variable intervient avec un coefficient positif.

Pour les variables explicatives à plus de deux modalités, le fait qu'une modalité explicative se projette dans la même direction que la variable à expliquer, avec une bonne qualité de représentation, signifie que cette modalité interviendrait avec un coefficient positif dans un modèle logistique où la seule variable explicative serait la présence ou l'absence de la modalité.

Ce qui précède concerne uniquement un modèle logistique à une variable explicative ; dans le cas de plusieurs variables explicatives, on peut démonter qu'il en est encore de même au moins dans un cas particulier : absence de termes d'interaction et indépendance entre les variables explicatives dans les deux sous groupes ($Z = 1$ et $Z = 0$) [23]. Dans le cas général, les coefficients, comme en modèle linéaire, dépendent des variables présentes dans le modèle ; les effets les plus importants, cependant, sont peu sensibles à la présence ou à l'absence d'autres variables, si on prend soin de ne pas avoir de variables explicatives trop liées.

2. Choix de bonnes variables prédictives

L'analyse des correspondances multiples des variables explicatives (variables actives) permet de réduire celles-ci, et leurs modalités, de façon à améliorer la stabilité et la précision des coefficients du modèle.

Si les variables explicatives ont été traitées en variables supplémentaires, l'information sur leurs inter-relations est moins directe, mais, jointe aux tableaux croisés, peut être suffisante pour en sélectionner un sous ensemble.

Par ailleurs, cette dernière analyse donne des indications sur les variables les plus liées, globalement, à un ensemble de variables à expliquer bien représentées sur un sous-espace de faible dimension.

Pour une variable explicative X_j la mesure à prendre peut être la dispersion des diverses modalités de X_j sur le sous-espace considéré, complétée par la mesure de la qualité globale, sur ce sous-espace, de la représentation des modalités de X_j, telle que cette mesure est proposée dans [2] ou [17], comme somme pondérée par les inerties, des qualités de représentation des modalités.

L'une et l'autre de ces mesures peuvent intervenir comme élément de choix de variables explicatives si les variables à expliquer sont nombreuses, et que l'on veut se restreindre aux mêmes variables explicatives pour tous les modèles.

3. Détection des termes d'interaction

Le modèle liant Z à deux variables explicatives à deux modalités peut être :

$$\text{Logit } P(Z) = \alpha_\square + \alpha_1 X_1 + \alpha_2 X_2 \qquad (2)$$

(modèle sans terme d'interaction)

ou :

$$\text{Logit } P(Z) = \alpha_\square + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_{12} X_1 X_2 \qquad (3)$$

l'absence d'interaction se définit aussi par :

$$\text{odds } R(Z, X_1 X_2) = \text{odds } R(Z, X_1) \text{ odds } R(Z, X_2) \quad (4)$$

odds R étant le odds ratio du tableau croisé correspondant.

La notion d'interaction, dans le modèle logistique, est complexe ; il peut y avoir interaction ou non, qu'il y ait indépendance ou non entre les variables X_1 et X_2. Pour qui est familier du modèle log-linéaire, un terme d'interaction (d'ordre 2) en modèle logistique est un terme d'interaction d'ordre 3 en modèle log-linéaire.

L'analyse des correspondances détecte des interactions d'ordre 2 en modèle Log-linéaire [15], mais il paraît difficile d'imaginer que soient détectées des interactions d'ordre 2 en modèle logistique. Les seules interactions que l'on peut espérer mettre en évidence (par exemple en projetant les associations de modalités explicatives) sont tellement extrêmes qu'elles amèneraient à remettre en cause le choix du modèle : par exemple, α_1 et α_2 de même signe, α_{12} de signe opposé, et : $|\alpha_{12}| > |\alpha_1 + \alpha_2|$.

CONCLUSION

Aucune méthode ne permet une analyse complète des données, dans le cas ou les questions que l'on se pose sont à la limite de "l'exploration" et de la "validation". Ceci justifie l'intérêt qu'il y aurait à mieux préciser en quoi les approches très différentes peuvent se compléter, et à tester des procédures incluant diverses méthodes, pour des données issues de domaines d'application variés. Ce qui est présenté ici ne peut être formalisé comme une démarche précise et rigoureuse, susceptible d'être appliquée de façon systématique. Cela tient en partie à ce que les liens entre méthodes sont limités, l'application de l'une ne permettant de prédire que très partiellement les résultats de l'autre.

Un point également important est que, dans une démarche de recherche, dans le choix des investigations et des approches, interviennent des éléments propres au domaine d'application, à l'intérêt porté à différents aspects qui ne peuvent se traduire en termes statistiques.

REMERCIEMENTS

Nous remercions messieurs M. GUIGNARD et M. CARRE pour les données analysées.

REFERENCES

[1] P. AIAICH, A. LECLERC, A. PHILIPPE. – Facteurs de différenciation dans la déclaration de symptômes *Rev. Epidemiol. Santé Pub.*, 29, 1, 27-44, 1981.

[2] C. BASTIN, J.P. BENZECRI, C. BOURGARIT, P. CAZES. – *Pratique de l'analyse des données,* Vol. 2, DUNOD, 1980, 466 p.

[3] J.K. BENEDETTI, M.B. BROWN. – Stratégies for the selection of Log-linear models. *Biometrics,* 34 : 680-686, 1978.

[4] Y.M.M. BISHOP, S.E. FIENBERG, P.W. HOLLAND. – Discrete multivariate analysis : theory and practice. *The M.I.T. Press,* 1975.

[5] B.M.D.P. – 79. – University of California Press.

[6] N.E. BRESLOW, N.E. DAY. – Statistical methods in cancer research Vol. I : The Analysis of Case Control Studies. *IARC Scientific Publication n° 32 : WHO,* Lyon 1980.

[7] D.R. COX. – *Analysis of binary data.* Methuen London, 1970.

[8] J.J. DAUDIN, P. TRECOURT. – Analyse factorielle des correspondances et modèle Log-linéaire : comparaison de deux méthodes sur un exemple. *Rev. Stat. Appl.*, 28 : 5-24, 1980.

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

[10] J.M. DEVAUD. – *Discrimination et description sur variables qualitatives Application à des données d'accident de travail.* Thèse de 3^{e} cycle, Paris-Dauphine, 1982.

[11] S.E. FIENBERG. – The Analysis of cross-classified categorical data. *The M.I.T. Press,* 1980.

[12] S. GREENLAND. – Tests for interaction in epidemiologic studies : a review and a study of power. *Statistics in medicine,* Vol. 2, n° 2, 1983.

[13] M. GUIGNARD, M. CARRE. – Vécu et santé du travailleur en service continu. Thèse de doctorat d'Université, *Ergonomie et Ecologie,* Paris I, 1983.

[14] S.J. HABERMAN. – *The Analysis of frequency data.* Chicago, Univ. of Chicago Press, 1974.

[15] N.C. LAURO, A. DECARLI. – Correspondence analysis and Log-linear models in multiway contingency tables study. *Some remarks on experimental data.* Metron, Vol. XL, N° 1-2, 213-234, 1982.

[16] L. LEBART, A. MORINEAU, N. TABARD. – *Techniques de la description statistique.* DUNOD, 1977.

[17] A. LECLERC. – *Aides à l'interprétation et procédures de validation en analyse de données.* Thèse d'état, Paris 6, 1982.

[18] A. LECLERC. – Modelos logísticos.Journées d'analyse de données et informatique CARACAS, 1983.

[19] J . LELLOUCH. – Le risque : définitions et procédés de calcul *Rev. Epidém. et santé publ.*, 1976, 24, 201-210.

[20] P. LOGEAY, J .F. CHASTANG, F. LERT. – *Etude épidémiologique de l'incidence médicale et sociale des horaires de travail sur le personnel infirmier* (en préparation).

[21] J .P. NAKACHE. – *Méthodes de discrimination sur variables de nature quelconque.* Théorie et pratique. Thèse d'état, Paris 6, 1980.

[22] J.W. TUKEY. – *Styles of data analysis and their implications for statistical computing.* COMPSTAT 1980, Proceedings in computational statistics Physica-Verlag, 21-31.

[23] Méthodes quantitatives en épidémiologie. *Séminaire technologique INSERM,* Document multigraphié, 1983.