REVUE DE STATISTIQUE APPLIQUÉE

M. TENENHAUS
Y. LE ROUX
C. GUIMART
P.-L. GONZALEZ

Modèle linéaire généralisé et analyse des correspondances

Revue de statistique appliquée, tome 41, n° 2 (1993), p. 59-86
<http://www.numdam.org/item?id=RSA_1993__41_2_59_0>

© Société française de statistique, 1993, 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/

Rev. Statistique Appliquée, 1993, **XLI** (2), 59-86

MODÈLE LINÉAIRE GÉNÉRALISÉ ET ANALYSE DES CORRESPONDANCES

M. Tenenhaus (1), Y. Le Roux (2), C. Guimart (2), P.-L. Gonzalez (3)

(1) Groupe HEC, (2) Rhône-Poulenc Rorer, (3) CNAM-IIE

RÉSUMÉ

Nous étudions dans cet article le problème de la régression qualitative : expliquer une variable Y qualitative à l'aide d'un ensemble de variables explicatives $X_1, \ldots, X_k$ qualitatives. Nous présentons deux approches complémentaires : le modèle linéaire généralisé et l'analyse des correspondances. Le modèle linéaire généralisé permet de modéliser la loi de probabilité de Y, ou une fonction de cette loi de probabilité, en fonction des variables $X_1, ..., X_k$. Notre présentation du modèle linéaire généralisé correspond à la PROC CATMOD du logiciel SAS. L'analyse des correspondances du tableau de contingence croisant $(X_1 * X_2 * ... * X_k)$ avec Y permet de visualiser les profils-lignes, mais aussi de tester approximativement l'influence des variables X_j sur la dispersion de ces profils. Dans l'exemple traité (le problème des goitres au Mali) les deux approches concordent.

Mots-clés : *Modèle linéaire généralisé, analyse des correspondances, analyse discriminante qualitative, régression qualitative, régression logistique.*

ABSTRACT

In this paper we study the qualitative regression problem : to explain a categorical variable Y with the help of a set of categorical explanatory variables $X_1, ..., X_k$. We present two complementary approaches : generalized linear model and correspondence analysis. The generalized linear model allows to model the probability distribution of Y, or a function of this probability distribution, as a function of the variables $X_1, ..., X_k$. Our presentation of the generalized linear model follows the PROC CATMOD of the SAS software. Correspondence analysis of the contingency table which crosses $(X_1 * X_2 * ... * X_k)$ by Y allows visualisation of row-profils, and also approximate tests on the influence of the variables X_j on the dispersion of these profils. In the treated example (the goitre problem in Mali) the two approaches agree.

Key-words : *Generalized linear model, correspondence analysis, qualitative discriminant analysis, qualitative regression, logistic regression.*

Introduction

L'analyse des données qualitatives s'est considérablement développée ces dernières années sur deux voies parallèles : les méthodes de codage optimal (la géométrie) et le modèle linéaire généralisé (la statistique). Selon la nature des données à analyser et le type de problème posé, la géométrie et/ou la statistique apporteront les réponses les plus appropriées. Nous allons étudier dans cet article le cas particulier des méthodes explicatives : il s'agit d'étudier les méthodes permettant d'analyser la liaison entre une variable qualitative Y et un ensemble de variables quantitatives ou qualitatives $X_1, ..., X_k$. Dans une première partie nous présenterons le modèle linéaire généralisé qui permet de généraliser tous les concepts de la régression multiple (estimation, test, analyse des résidus, prévision) au cas où les variables Y, $X_1, ..., X_k$ sont qualitatives. Nous montrerons dans la deuxième partie l'intérêt complémentaire de l'analyse des correspondances qui, tout en illustrant graphiquement la première approche, permet de mettre en évidence d'autres aspects des données.

Notre présentation du modèle linéaire généralisé correspond à la PROC CATMOD de SAS et s'appuie sur les références suivantes : Agresti (1990), Hosmer et Lemeshow (1989), McCullagh et Nelder (1989), les notes de cours de SAS Institute (1988), et bien sûr la documentation de la PROC CATMOD du SAS/STAT User's guide (1990).

I. LE MODÈLE LINÉAIRE GÉNÉRALISÉ

1. Présentation du problème

Le modèle linéaire généralisé permet d'étudier la liaison entre une variable qualitative Y et un ensemble de variables explicatives $X_1, ..., X_k$ qualitatives ou quantitatives. La variable dépendante Y peut elle-même être formée à partir du croisement de p variables qualitatives $Y_1, ..., Y_p$. Les s croisements disponibles $(x_{i1}, ..., x_{ik})$ des variables $X_1, ..., X_k$ définissent s populations. Notons π_i la loi de probabilité de Y sur la population i. On cherche à relier linéairement q fonctions de réponse $F_h(\pi_i)$, $h = 1, ..., q$, aux caractéristiques de la population i :

$$F_h(\pi_i) = x_i \beta_h \tag{1}$$

où x_i est un vecteur-ligne caractérisant la population i et β_h un vecteur-colonne de paramètres. Les fonctions de réponse F_h sont des transformations logistiques, ou toute autre fonction choisie par l'utilisateur. Lorsque les effectifs des cases du tableau de contingence croisant $(X_1 * X_2 * ... * X_k)$ par Y sont faibles, alors seule la méthode du maximum de vraisemblance, et par voie de conséquence la transformation logistique, sont utilisables. Ceci est en particulier le cas lorsque certaines variables explicatives sont quantitatives.

Posant :

$$F(\pi) = \begin{pmatrix} F_1(\pi_1) \\ \vdots \\ F_q(\pi_1) \\ \vdots \\ F_1(\pi_s) \\ \vdots \\ F_q(\pi_s) \end{pmatrix}, \quad X = \begin{pmatrix} x_1 & 0 & \dots & 0 \\ 0 & x_1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & x_1 \\ \vdots & \vdots & & \vdots \\ x_s & 0 & \dots & 0 \\ 0 & x_s & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & x_s \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_q \end{pmatrix}$$

le modèle (1) s'écrit plus globalement

$$F(\pi) = X\beta$$

Disposant d'observations de la variable à expliquer Y et des variables explicatives $X_1, ..., X_k$ nous allons étudier dans cet article les méthodes permettant

(1) d'estimer le vecteur des paramètres β,

(2) de tester une hypothèse de la forme $H_0 : L\beta = 0$, où L est une matrice,

(3) de valider le modèle étudié.

2. Les données

On observe sur n individus les variables Y, $X_1, ..., X_k$. La variable dépendante Y est qualitative à r modalités, éventuellement ordonnées. Les variables explicatives $X_1, ..., X_k$ peuvent être qualitatives ou quantitatives. Une variable qualitative X_j prend p_j modalités. Les croisements disponibles entre les niveaux des variables $X_1, ..., X_k$ définissent les s populations. Les données disponibles se présentent donc sous la forme d'un tableau de contingence :

	Réponse Y					
Echantillon	1	...	j	...	r	
1	n_{11}	...	n_{1j}	...	n_{1r}	n_1
⋮	⋮		⋮		⋮	⋮
i	n_{i1}	...	n_{ij}	...	n_{ir}	n_i
⋮	⋮		⋮		⋮	⋮
s	n_{s1}	...	n_{sj}	...	n_{sr}	n_s

On extrait de chaque population i un échantillon de taille n_i. Ces s échantillons sont supposés indépendants les uns des autres. Chaque échantillon i se répartit en r classes d'effectifs $n_{i1}, ..., n_{ij}, ..., n_{ir}$ en fonction des réponses à la variable Y_i :

n_{ij} représente le nombre de personnes de l'échantillon i prenant la modalité j de la variable Y.

Nous allons illustrer cette note à l'aide des données de Rhône-Poulenc Rorer reproduites dans le tableau 1. Il s'agit de tester l'efficacité d'un diffuseur d'iode pour lutter contre le goitre. A la suite d'un concept transmis par les Docteurs Alain Fisch et Thierry Prasuck, Rhône-Poulenc a mis au point et testé, en collaboration avec ces derniers, et l'Institut Santé Développement (Dr. R. Sebbag, G. Cyprien) ainsi que l'Ecole nationale de médecine de Bamako (Pr. Ag Rhaly, Dr. Eric Pichard...), le Rhodifuse^R iode breveté par Rhône-Poulenc (brevets européens N° EP 283.407.285.525.283.408 et 284.521) dans le but de lutter contre ces carences. La technique utilisée, basée sur la diffusion continue d'un principe actif dans l'eau des puits et des forages, a conduit à des résultats intéressants et pourrait être mise en œuvre pour prévenir d'autres maladies endémiques dans le tiers-monde.

Un prototype, élaboré à partir d'une matrice unitaire de 70 % de silicone et de 30 % d'iodure de sodium, a été mis en œuvre dans un pays africain, sur le territoire de la république du Mali. L'étude statistique, dont nous ne présentons ici qu'une partie à titre illustratif, a été réalisée après la mise en œuvre d'un tel système pilote, afin de prouver l'efficacité du procédé. Il y a trois villages, le premier servant de village témoin. Au jour 0, on regarde la répartition des goitres par niveau de gravité (codé de 1 : pas de goitre à 5 : goitre irréversible). On sait que les goitres sont dus à une carence alimentaire en iode, carence que devrait pallier la diffusion continue d'iode dans l'eau. Dans les villages 2 et 3, on installe donc ces diffuseurs d'iode dans les puits et on observe les résultats six mois après (jour 180), en mesurant les taux d'iode urinaire. On sait d'autre part que la répartition des goitres dépend du sexe. On cherche donc à relier le niveau de goitre (variable Y) aux variables X_1 = Village, X_2 = Sexe (H = 1, F = 2), X_3 = Présence d'iode dans l'urine (Absence = 1, Présence = 2) et X_4 = Jour.

Les données (en pourcentage en ligne) du tableau 2 montrent clairement une aggravation des goitres dans le village témoin (1) et une amélioration de la situation dans les deux autres villages.

Le modèle linéaire généralisé va nous permettre de quantifier et de valider cette impression.

3. Le modèle linéaire généralisé

Il s'agit de modéliser les probabilités π_{ij} = Prob($Y = j$ dans la population i) ou, plus généralement, des fonctions $F_h(\pi_i)$ du vecteur des probabilités $\pi_i = (\pi_{i1}, ..., \pi_{ir})'$ en fonction des caractéristiques de la population i :

$$F_h(\pi_i) = x_i \beta_h \tag{1}$$

où x_i est le vecteur-ligne des caractéristiques de la population i, et β_h un vecteur-colonne de paramètres. Les q transformations $F_1, ..., F_q$ sont choisies *a priori*. Donnons les fonctions de réponse disponibles en standard dans la PROC CATMOD de SAS, version 6.

TABLEAU 1
Les données

				Niveau de goitre					
Village	Sexe	Iode	Jour	1	2	3	4	5	Total
1	1	1	0	106	12	46	11	0	175
1	1	1	180	60	31	46	15	0	152
1	2	1	0	77	21	71	65	11	245
1	2	1	180	46	28	63	65	11	213
2	1	1	0	127	27	45	12	1	212
2	1	2	180	145	28	19	1	1	194
2	2	1	0	69	21	65	50	2	207
2	2	2	180	76	40	41	13	2	172
3	1	1	0	91	8	14	6	0	119
3	1	2	180	94	14	10	0	0	118
3	2	1	0	42	18	45	34	4	143
3	2	2	180	50	29	38	13	3	133

TABLEAU 2
Les données (en % en ligne)

				Niveau de goitre				
Village	Sexe	Iode	Jour	1	2	3	4	5
1	1	1	0	60.57	6.86	26.29	6.29	0
1	1	1	180	39.47	20.39	30.26	9.87	0
1	2	1	0	31.43	8.57	28.98	26.53	4.49
1	2	1	180	21.60	13.15	29.58	30.52	5.16
2	1	1	0	59.91	12.74	21.23	5.66	0.47
2	1	2	180	74.74	14.43	9.79	0.52	0.52
2	2	1	0	33.33	14.43	31.40	24.15	0.97
2	2	2	180	44,19	23.26	23.84	7.56	1.16
3	1	1	0	76.47	6.72	11.76	5.04	0
3	1	2	180	79.66	11.86	8.47	0	0
3	2	1	0	29.37	12.59	31.47	23.78	2.80
3	2	2	180	37.59	21.80	28.57	9.77	2.26

(1) *Identité*

$F_h(\pi_i) = \pi_{ih}$, $h = 1, \ldots, r-1$

(2) *Logit généralisé*

$F_h(\pi_i) = \text{Log}(\pi_{ih}/\pi_{ir})$, $h = 1, ..., r-1$

(3) *Logit adjacent*

$F_h(\pi_i) = \text{Log}(\pi_{i(h+1)}/\pi_{ih})$, $h = 1, ..., r-1$

(4) *Logit cumulé*

$F_h(\pi_i) = \text{Log}(\text{Prob}(Y > h)/\text{Prob}(Y \leqslant h))$, $h = 1, ..., r-1$

(5) *Moyenne*

$$F(\pi_i) = \sum_{j=1}^{r} j\pi_{ij}.$$

Les trois dernières transformations sont adaptées au cas d'une variable dépendante Y ordinale.

Disposant d'observations de la variable à expliquer Y et des variables explicatives $X_1, ..., X_k$ il est possible

(a) d'estimer le vecteur des paramètres $\beta = (\beta_1', ..., \beta_q')'$ en utilisant la méthode des moindres carrés généralisée, ou, mais seulement pour la fonction de réponse *Logit généralisé*, le maximum de vraisemblance.

(b) de tester une hypothèse de la forme $H_o : L\beta = 0$, où L est une matrice, en utilisant la statistique de Wald ou, lorsque la fonction de réponse est le *Logit généralisé*, le logarithme du rapport des vraisemblances.

(c) de valider le modèle étudié en comparant les estimations $F_h(p_i)$ des $F_h(\pi_i)$ à celles obtenues en utilisant le modèle estimé $\widehat{F}_h(\pi_i) = x_i\widehat{\beta}_h$, p_i étant le vecteur des fréquences relatives n_{ij}/n_i observées dans l'échantillon i.

Nous allons étudier dans les sections suivantes la résolution de ces trois problèmes et montrer leur intérêt pratique sur l'exemple Mali.

4. Etude des fonctions de réponse

On considère que, dans chaque échantillon i, la répartition des effectifs observés suit une loi multinomiale :

$$\text{Prob}(n_{i1}, \ldots, n_{ij}, \ldots, n_{ir}) = n_i! \, \frac{\pi_{i1}^{ni1} \ldots \pi_{ij}^{n_{ij}} \ldots \pi_{ir}^{nir}}{n_{i1}! \ldots n_{ij}! \ldots n_{ir}!}$$

On construit ensuite le tableau des proportions observées $p_{ij} = n_{ij}/n_i$:

	Proportions					
Echantillons	1	...	j	...	r	
1	p_{11}	...	p_{1j}	...	p_{1r}	
⋮	⋮		⋮		⋮	
i	p_{i1}	...	p_{ij}	...	p_{ir}	p_i'
⋮	⋮		⋮		⋮	
s	p_{s1}	...	p_{sj}	...	p_{sr}	

Le vecteur $p_i = (p_{i1}, ..., p_{ij}, ..., p_{ir})'$ représente la répartition des fréquences relatives des différentes modalités de Y dans l'échantillon i. On peut calculer moyenne et variance de ce vecteur p_i. On obtient :

$$\begin{aligned} E(p_i) &= \pi_i \\ V(p_i) &= (\text{Diag}(\pi_i) - \pi_i \pi_i')/n_i \end{aligned} \tag{2}$$

où $\text{Diag}(\pi_i)$ représente la matrice diagonale formée des termes $\pi_{i1}, ..., \pi_{ir}$. La matrice $V(p_i)$ est estimée en remplaçant π_i par p_i dans (2) :

$$V_i = \widehat{V}(p_i) = (\text{Diag}(p_i) - p_i p_i')/n_i.$$

On note $p = (p_1', ..., p_s')'$ le vecteur formé de toutes les proportions observées et $\pi = (\pi_1', ..., \pi_s')'$ le vecteur de toutes les probabilités. Le vecteur p est une estimation de π avec

$$E(p) = \pi$$

$$V(p) = \begin{bmatrix} V(p_1) & & 0 \\ & \ddots & \\ 0 & & V(p_s) \end{bmatrix}$$

La matrice $V(p)$ est estimée en remplaçant chaque bloc $V(p_i)$ par son estimation V_i. D'où

$$V = \widehat{V}(p) = \begin{bmatrix} V_1 & & 0 \\ & \ddots & \\ 0 & & V_s \end{bmatrix}$$

On construit ensuite le vecteur F formé de tous les $F_h(p_i)$:

$$F_i = F(p_i) = \begin{bmatrix} F_1(p_i) \\ \vdots \\ F_q(p_i) \end{bmatrix}, \quad F = F(p) = \begin{bmatrix} F_1 \\ \vdots \\ F_s \end{bmatrix}$$

On peut calculer approximativement la moyenne et la variance de chaque vecteur F_i, et par conséquent de F.

On suppose tout d'abord, qu'asymptotiquement,

$$E(F_i) = E(F(p_i)) = F(\pi_i) = \begin{bmatrix} F_1(\pi_i) \\ \vdots \\ F_q(\pi_i) \end{bmatrix} = \begin{bmatrix} x_i \beta_1 \\ \vdots \\ x_i \beta_q \end{bmatrix} = X_i \beta$$

où X_i est la matrice diagonale par blocs formés du même vecteur x_i.

On calcule ensuite une approximation de la variance de F_i en développant $F(p_i)$ au voisinage de $F(\pi_i)$:

$$F(p_i) \approx F(\pi_i) + \frac{\partial F}{\partial p_i}(\pi_i) \cdot (p_i - \pi_i)$$

D'où :

$$V(F_i) \approx \frac{\partial F}{\partial p_i}(\pi_i) \cdot V(p_i) \cdot \frac{\partial F'}{\partial p_i}(\pi_i)$$

Posant $H_i = \frac{\partial F}{\partial p_i}(p_i)$ estimation de $\frac{\partial F}{\partial p_i}(\pi_i)$ et utilisant V_i, estimation de $V(p_i)$, on obtient finalement une estimation S_i de $V(F_i)$:

$$S_i = \widehat{V}(F_i) = H_i V_i H_i'.$$

Globalement, sur l'ensemble des populations :

$$E(F) = X\beta, \quad \text{où} \quad X = \begin{bmatrix} X_1 \\ \vdots \\ X_s \end{bmatrix}$$

et $V_F = V(F)$ est estimée par $\widehat{V}(F) = S = \begin{bmatrix} S_1 & & 0 \\ & \ddots & \\ 0 & & S_s \end{bmatrix}$.

5. Estimation et test

On utilise la méthode des moindres carrés généralisée pour estimer les paramètres du modèle. Lorsque la fonction de réponse est le logit généralisé il est préférable (et parfois nécessaire) d'utiliser la méthode du maximum de vraisemblance.

5.1 La méthode des moindres carrés généralisés

Notre modèle s'écrit

$$F(p) = X\beta + \varepsilon \tag{3}$$

où ε est un terme d'erreur de moyenne nulle et de matrice variances/covariances $V_F = \text{Var}(F(p))$ estimée par la matrice S. Pour se ramener à un modèle de régression usuelle on prémultiplie les deux termes de (3) par $V_F^{-1/2}$:

$$V_F^{-1/2} F(p) = V_F^{-1/2} X\beta + V_F^{-1/2}\varepsilon \tag{4}$$

Cette fois le terme d'erreur $V_F^{-1/2}\varepsilon$ est de moyenne nulle et de matrice variances/covariance *l'identité*. Utilisant les formules habituelles de la régression multiple on obtient donc :

a) Estimation de β

$$\widehat{\beta} = (X'V_F^{-1}X)^{-1}X'V_F^{-1}F(p) \approx (X'S^{-1}X)^{-1}X'S^{-1}F(p)$$

b) Estimation de $V(\widehat{\beta})$

$$V(\widehat{\beta}) = (X'V_F^{-1}X)^{-1} \approx (X'S^{-1}X)^{-1}$$

c) Valeur prédite de F

$$\widehat{F} = X\widehat{\beta}$$

d) Variance de $\widehat{F}$

$$\text{Var}(\widehat{F}) = X \cdot V(\widehat{\beta}) \cdot X' \approx X(X'S^{-1}X)^{-1}X'$$

e) Somme des carrés résiduelles

$$\begin{aligned}\text{SCR} &= \|V_F^{-1/2} \cdot F(p) - V_F^{-1/2}X\widehat{\beta}\|^2 \\ &= (F(p) - \widehat{F})'V_F^{-1}(F(p) - \widehat{F}) \\ &\approx (F(p) - \widehat{F})'S^{-1}(F(p) - \widehat{F})\end{aligned}$$

Sous l'hypothèse du modèle étudié, la somme des carrés résiduelles SCR suit une loi du Khi-deux à sq-d degrés de liberté où d est le nombre de paramètres du modèle. On peut donc tester la validité du modèle étudié.

f) Test d'une hypothèse linéaire

L'hypothèse $H_0 : L\beta = 0$ est testée à l'aide de la statistique de Wald $Q = (L\widehat{\beta})'(\widehat{\text{Var}}(L\widehat{\beta}))^{-1}L(\widehat{\beta})$ où $\widehat{\text{Var}}(L\widehat{\beta}) = L\widehat{\text{Var}}(\widehat{\beta})L' = L(X'S^{-1}X)^{-1}L'$.

On retrouve un résultat de la régression multiple : la statistique Q correspond exactement à l'augmentation de la somme des carrés résiduelles du modèle (3) lorsqu'on impose à $\widehat{\beta}$ de vérifier la contrainte $L\widehat{\beta} = 0$. Sous l'hypothèse H_0 vraie, la statistique Q suit une loi du Khi-deux dont le nombre de degrés de liberté est égal au rang de la matrice L.

5.2 La méthode du maximum de vraisemblance

Cette méthode n'est disponible dans la PROC CATMOD de SAS que pour le modèle logistique. Il est possible dans ce modèle de calculer les π_{ij} en fonction du vecteur β des paramètres. Le modèle étudié s'écrit, pour $i = 1, ..., s$ et $j = 1, ..., r-1$,

$$\text{Log}(\pi_{ij}/\pi_{ir}) = x_{ij}\beta \tag{5}$$

où x_{ij} est la j-ième ligne de la matrice X_i. Les fonctions de réponse $F_h(\pi_i) = \text{Log}(\pi_{ih}/\pi_{ir})$, $h = 1, ..., r-1$ s'appellent les logits généralisés.

De la formule (5) et de la condition $\sum_{j=1}^{r} \pi_{ij} = 1$, nous déduisons :

$$\pi_{ij} = f_{ij}(\beta) = \begin{cases} \dfrac{e^{x_{ij}\beta}}{1 + \sum_{j=1}^{r-1} e^{x_{ij}\beta}}, & j = 1, \ldots, r-1 \\ \dfrac{1}{1 + \sum_{j=1}^{r-1} e^{x_{ij}\beta}}, & j = r \end{cases}$$

La vraisemblance du modèle pour les données observées s'écrit donc

$$\varphi(\text{Modèle}) = \prod_{i=1}^{s} n_i! \frac{\pi_{i1}^{n_{i1}} \ldots \pi_{ir}^{n_{ir}}}{n_{i1}! \ldots n_{ir}!}$$
$$= \prod_{i=1}^{s} n_i! \frac{f_{i1}^{n_{i1}}(\beta) \ldots f_{ir}^{n_{ir}}(\beta)}{n_{i1}! \ldots n_{ir}!}.$$

On recherche β maximisant la vraisemblance φ du modèle. Pratiquement on cherche à minimiser $-2\text{Log}\,\varphi(\text{Modèle})$ en annulant les dérivées partielles. L'hypothèse $H_0 : L\beta = 0$ peut être testée en utilisant la statistique

$$D = -2\text{Log}\left[\frac{\varphi\ (\text{Modèle sous } H_0)}{\varphi\ (\text{Modèle sans contrainte})}\right]$$

qui, sous H_0, suit une loi du Khi-deux dont le nombre de degrés de liberté est égal au rang de la matrice L.

L'adéquation du modèle aux données est testée en utilisant un test du rapport des vraisemblances. On compare la vraisemblance du modèle étudié à celle d'un modèle saturé. Précisons qu'un modèle saturé est un modèle reconstruisant parfaitement les proportions p_{ij}. Le nombre de paramètres indépendants d'un modèle saturé est égal au nombre de proportions p_{ij} indépendantes. Dans ces conditions $\hat{\pi}_{ij} = p_{ij}$ et la vraisemblance d'un modèle saturé vaut

$$\varphi(\text{Modèle saturé}) = \prod_{i=1}^{s} n_i! \frac{p_{i1}^{n_{i1}} \ldots p_{ir}^{n_{ir}}}{n_{i1}! \ldots n_{ir}!}.$$

Si le modèle étudié est exact, alors la statistique

$$D = -2 \log\left[\frac{\varphi\ (\text{Modèle étudié})}{\varphi(\text{Modèle saturé})}\right]$$

suit une loi du Khi-deux à $N - d$ degrés de liberté où N est le nombre de p_{ij} indépendants et d le nombre de paramètres du modèle étudié.

6. Etude de l'exemple «Mali»

Le plus simple est de modéliser directement les probabilités π_{ij} = Prob(goitre de niveau j dans la population i) en fonction des facteurs Village, Sexe, Iode et Jour :

$$\pi_{ij} = \beta_{oj} + \underset{\text{Village}}{\begin{matrix}1\\2\\3\end{matrix}\begin{bmatrix}\beta_{1j}\\ \beta_{2j}\\ -\beta_{1j}-\beta_{2j}\end{bmatrix}} + \underset{\text{Sexe}}{\begin{matrix}\text{H}\\ \text{F}\end{matrix}\begin{bmatrix}\beta_{3j}\\ -\beta_{3j}\end{bmatrix}} + \underset{\text{Iode}}{\begin{matrix}\text{Abs}\\ \text{Prés}\end{matrix}\begin{bmatrix}\beta_{4j}\\ -\beta_{4j}\end{bmatrix}} + \underset{\text{Jour}}{\begin{matrix}0\\ 180\end{matrix}\begin{bmatrix}\beta_{5j}\\ -\beta_{5j}\end{bmatrix}}$$

pour $j = 1$ à 4. La contrainte «Somme des coefficients associés aux modalités d'un facteur nulle» est celle adoptée dans la PROC CATMOD. Nous étudierons également la modélisation du logit généralisé $\text{Log}(\pi_{ij}/\pi_{is})$, du logit cumulé $\text{Log}(\text{Prob}(Y > j \mid i)/\text{Prob}(Y \leqslant j \mid i))$ et de la moyenne $\sum_{j=1}^{s} j\pi_{ij}$ en fonction des mêmes facteurs. Il est possible d'introduire des termes d'interaction entre les facteurs.

Lorsqu'on utilise la méthode des moindres carrés généralisée, il y a calcul explicite des fonctions de réponse. Dans ce cas la présence de p_{ij} nul provoque des difficultés de calcul. Pour éviter ces difficultés nous avons alors ajouté une unité à chaque n_{ij} du tableau des données. Lorsqu'on utilise la méthode du maximum de vraisemblance pour le modèle logistique, les logits généralisés ne sont pas calculés au niveau de l'estimation et on peut donc utiliser les données d'origine.

Précisons les vecteurs x_i caractérisant les 12 populations disponibles :

				Vecteurs x_i					
Village	Sexe	Iode	Jour	Constante	Village1	Village2	SexeH	IodeAbs	Jour0
1	1	1	0	1	1	0	1	1	1
1	1	1	180	1	1	0	1	1	−1
1	2	1	0	1	1	0	−1	1	1
1	2	1	180	1	1	0	−1	1	−1
2	1	1	0	1	0	1	1	1	1
2	1	2	180	1	0	1	1	−1	−1
2	2	1	0	1	0	1	−1	1	1
2	2	2	180	1	0	1	−1	−1	−1
3	1	1	0	1	−1	−1	1	1	1
3	1	2	180	1	−1	−1	1	−1	−1
3	2	1	0	1	−1	−1	−1	1	1
3	2	2	180	1	−1	−1	−1	−1	−1

Ainsi le vecteur $\pi_j = (\pi_{1j}, ..., \pi_{ij}, ..., \pi_{12j})'$ des différentes probabilités d'avoir un goitre de niveau j dans les populations i s'écrit :

$$\pi_j = \begin{bmatrix} \pi_{1j} \\ \vdots \\ \pi_{ij} \\ \vdots \\ \pi_{12j} \end{bmatrix} = \begin{array}{c} \begin{matrix} \text{Constante} & \text{Village} & & \text{Sexe} & \text{Iode} & \text{Jour} \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & -1 \\ 1 & 1 & 0 & -1 & 1 & 1 \\ 1 & 1 & 0 & -1 & 1 & -1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & -1 & -1 \\ 1 & 0 & 1 & -1 & 1 & 1 \\ 1 & 0 & 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix} \\ \text{"Design Matrix"} \end{array} \begin{bmatrix} \beta_{0j} \\ \beta_{1j} \\ \beta_{2j} \\ \beta_{3j} \\ \beta_{4j} \\ \beta_{5j} \end{bmatrix} \quad \beta_j$$

Résultats

a) Fonction de réponse «Identité»

En utilisant la méthode des moindres carrés généralisée on obtient les estimations $\widehat{\beta}_{hj}$ des paramètres du modèle avec leurs écarts-types $s(\widehat{\beta}_{hj})$, la statistique de Wald $(\widehat{\beta}_{hj}/s(\widehat{\beta}_{hj})^2$ et le niveau de signification. Les résultats sont présentés dans le tableau 3. On peut voir qu'aucun des paramètres décrivant l'influence du facteur Village n'est significatif. Ainsi on peut modéliser les probabilités Prob(Goitre de Niveau j | Village,Sexe,Iode,Jour). On a par exemple pour $j = 1$:

$$\widehat{\text{Prob}}(\text{Goitre de Niveau 1} \mid \text{Village,Sexe,Iode,Jour})$$
$$= 0.52 + \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} \begin{bmatrix} -0.015 \\ -0.001 \\ 0.016 \end{bmatrix} + \begin{matrix} \text{H} \\ \text{F} \end{matrix} \begin{bmatrix} 0.14 \\ -0.14 \end{bmatrix} + \begin{matrix} \text{Abs} \\ \text{Près} \end{matrix} \begin{bmatrix} -0.11 \\ 0.11 \end{bmatrix} + \begin{matrix} 0 \\ 180 \end{matrix} \begin{bmatrix} 0.07 \\ -0.07 \end{bmatrix}$$
$$\text{Village} \qquad \text{Sexe} \qquad \text{Iode} \qquad \text{Jour}$$

On trouve pour la première population (Village = 1, Sexe = H, Iode = Abs, Jour = 0) $\widehat{\pi}_{11} = 0.52 - 0.015 + 0.14 - 0.11 + 0.07 = 0.6$ à comparer à la proportion observée $p_{11} = \frac{107}{180} = 0.594$ (on vérifie ici qu'une unité a bien été ajoutée à chaque case du tableau des données).

Dans le tableau 4 on trouve le test de Wald sur les paramètres associés à chaque facteur. Décrivons par exemple le test de l'influence du facteur Village. Il s'agit du test

$$H_0 : \beta_{1j} = \beta_{2j} = 0, \quad j = 1, \ldots, 4$$

La statistique de Wald s'écrit

$$Q = \widehat{\beta}'_{\text{village}}(\widehat{\text{Var}}(\widehat{\beta}_{\text{Village}}))^{-1}\widehat{\beta}_{\text{Village}}$$

TABLEAU 3
Estimation et Test des paramètres
[Fonction de Réponse : Identité]

Effet	Numéro de la fonction de réponse	Estimation des paramètres	Ecart-type	Statistique de Wald	Niveau de signification
Constante	1	0.52	0.012	1849.1	0.0001
	2	0.13	0.009	223.8	0.0001
	3	0.22	0.010	438.5	0.0001
	4	0.11	0.007	234.1	0.0001
Village 1	1	−0.015	0.019	0.6	0.44
	2	−0.019	0.012	2.8	0.10
	3	0.023	0.017	1.7	0.19
	4	0.01	0.013	0.6	0.45
2	1	−0.001	0.015	0.0	0.92
	2	0.015	0.010	2.0	0.16
	3	−0.004	0.013	0.1	0.76
	4	−0.006	0.008	0.6	0.43
Sexe H	1	0.14	0.010	209.5	0.0001
	2	−0.01	0.007	2.0	0.15
	3	−0.06	0.009	43.0	0.0001
	4	−0.07	0.006	111.9	0.0001
Iode Abs	1	−0.11	0.020	29.1	0.0001
	2	0.006	0.015	0.2	0.68
	3	0.05	0.019	6.4	0.01
	4	0.06	0.014	16.1	0.0001
Jour 0	1	0.07	0.016	17.6	0.0001
	2	−0.04	0.011	10.6	0.001
	3	−0.009	0.016	0.3	0.61
	4	−0.02	0.012	2.8	0.09

où $\widehat{\beta}_{\text{Village}} = (\widehat{\beta}_{11}, \ldots, \widehat{\beta}_{14}, \widehat{\beta}_{21}, \ldots, \widehat{\beta}_{24})'$ est le vecteur des paramètres estimés associés au facteur Village. Ici $Q = 6.40$. Si H_0 est vraie la statistique Q suit une loi du Khi-deux à 8 degrés de liberté. Le niveau de signification de $Q = 6.40$ valant $P = \text{Prob}(\chi^2_{(8)} \geqslant 6.40) = 0.6$, on ne rejette pas l'hypothèse H_0. L'apport marginal du facteur Village n'est pas significatif. On peut remarquer que tous les autres facteurs sont significatifs.

La somme des carrés résiduelle pondérée $\text{SCR} = (F - \widehat{F})'\widehat{V}(F)^{-1}(F - \widehat{F})$ vaut 63.67. Ici F représente le vecteur

$$F = (p_{11}, ..., p_{14}, ..., p_{i1}, ..., p_{i4}, ..., p_{12.1}, ..., p_{12.4})'$$

formé des $N = 12 \times 4 = 48$ proportions p_{ij} indépendantes. Si le modèle étudié est exact, alors SCR suit une loi du Khi-deux à $48 - 24$ degrés de liberté, puisque le modèle étudié contient $d = 24$ paramètres. Le niveau de signification du résidu $P = \text{Prob}(\chi^2_{(24)} > 63.67)$ valant 0.0001 on doit rejeter le modèle étudié : les résidus sont trop importants.

TABLEAU 4
Tableau d'analyse de la variance

Source de variation	Degrés de liberté	Statistique de Wald	Niveau de signification
Constante	4	91175.72	0.0001
Village	8	6.40	0.6030
Sexe	4	266.65	0.0001
Iode	4	36.37	0.0001
Jour	4	23.22	0.0001
Résidu	24	63.6	0.0001

Vérifions que la statistique Q mesure bien un apport marginal. Le tableau d'analyse de la variance pour le modèle sans Village est donné dans le tableau 5. On peut donc vérifier que

$$Q(\text{Village}) = \text{SCR(Sexe,Iode,Jour)} - \text{SCR(Village,Sexe,Iode,Jour)}$$
$$= 70 - 63.60 = 6.40.$$

Ainsi la statistique de Wald Q(Village) représente la diminution de la somme des carrés résiduelle lorsqu'on passe du modèle à trois variables (Sexe,Iode,Jour) au modèle à quatre variables (Village,Sexe,Iode,Jour).

TABLEAU 5
Tableau d'analyse de la variance pour le modèle sans Village

Source de variation	Degrés de liberté	Statistique de Wald	Niveau de signification
Constante	4	108315	0
Sexe	4	270	0
Iode	4	96	0
Jour	4	34	0
Résidu	32	70	0

b) *Le modèle logistique*

On étudie maintenant le modèle

$$\mathrm{Log}(\pi_{ij}/\pi_{i5}) = \beta_{oj} + \underset{\text{Village}}{\begin{matrix}1\\2\\3\end{matrix}\begin{bmatrix}\beta_{1j}\\ \beta_{2j}\\ -\beta_{1j}-\beta_{2j}\end{bmatrix}} + \underset{\text{Sexe}}{\begin{matrix}\mathrm{H}\\ \mathrm{F}\end{matrix}\begin{bmatrix}\beta_{3j}\\ -\beta_{3j}\end{bmatrix}} + \underset{\text{Iode}}{\begin{matrix}\mathrm{Abs}\\ \mathrm{Prés}\end{matrix}\begin{bmatrix}\beta_{4j}\\ -\beta_{4j}\end{bmatrix}} + \underset{\text{Jour}}{\begin{matrix}0\\ 180\end{matrix}\begin{bmatrix}\beta_{5j}\\ -\beta_{5j}\end{bmatrix}}$$

On peut estimer les paramètres de ce modèle en utilisant la méthode des moindres carrés généralisée ou bien celle du maximum de vraisemblance. Le tableau d'analyse de la variance correspondant à l'utilisation des moindres carrés généralisés est donné dans le tableau 6 et celui associé au maximum de vraisemblance dans le tableau 7. Le tableau 6 montre que seul le facteur Village est non significatif. Le test d'adéquation du modèle étudié montre que le modèle logistique est acceptable : le niveau de signification du résidu vaut 0.1768. L'utilisation du maximum de vraisemblance conduit à un résultat plus ambigu. Le test du rapport des vraisemblances donne un niveau de signification de 0.0262.

TABLEAU 6
Tableau d'analyse de la variance
[Moindres carrés généralisée]

Source de variation	Degrés de liberté	Statistique de Wald	Niveau de signification
Constante	4	638.78	0.0001
Village	8	7.39	0.4952
Sexe	4	211.33	0.0001
Iode	4	50.53	0.0001
Jour	4	24.96	0.0001
Résidu	24	30.24	0.1768

Précisons les calculs effectués en utilisant le maximum de vraisemblance :

(1) Il n'est pas nécessaire de corriger le tableau des données car les logits généralisés $\mathrm{Log}(p_{ij}/p_{i5})$ ne sont pas calculés explicitement dans la phase estimation. Les calculs ont donc été réalisés sur les données d'origine.

(2) On a obtenu $-2\mathrm{Log}\,\varphi(\text{Modèle étudié}) = 4985.91$. Pour calculer $-2\mathrm{Log}\,\varphi(\text{Modèle saturé})$ on a modélisé le logit généralisé $\mathrm{Log}(\pi_{ij}/\pi_{i5})$ en fonction des effets principaux Village,Sexe, Iode et Jour et de toutes les interactions d'ordre 2, 3 et 4. On a obtenu : $-2\mathrm{Log}\,\varphi(\text{Modèle saturé}) = 4946.74$. Le test «H0 : Modèle étudié exact» s'effectue à l'aide de la statistique

$$D = -2\mathrm{Log}\left[\frac{\varphi(\text{Modèle étudié})}{\varphi(\text{Modèle saturé})}\right] = 4985.91 - 4946.74 = 39.17.$$

TABLEAU 7
Tableau d'analyse de la variance
[Maximum de vraisemblance]

Source de variation	Degrés de liberté	Statistique de Wald	Niveau de signification	−2Log (Rapport des vraisemblances)	Niveau de signification
Constante	4	596.25	0.0001	1058	0.0001
Village	8	8.85	0.3548	9.23	0.3
Sexe	4	226.82	0.0001	278	0.0001
Iode	4	54.41	0.0001	57	0.0001
Jour	4	25.54	0.0001	26	0.0001
Résidu	24			39.17	0.0262

Le modèle étudié contenant 24 paramètres et le modèle saturé 48, on en conclut que la statistique D suit une loi du Khi-deux à 24 degrés de liberté sous l'hypothèse H_0 vraie. D'où son niveau de signification $P = \text{Prob}(\chi^2_{(24)} \geqslant 39.17) = 0.0262$.

(3) Les tests du rapport des vraisemblances pour chaque facteur sont réalisés comme suit :

Pour l'effet Village par exemple on calcule

$$D_{\text{Village}} = -2\text{Log}\left[\frac{\varphi(\text{Sexe,Iode,Jour})}{\varphi(\text{Village,Sexe,Iode,Jour})}\right]$$
$$= 4995.14 - 4985.91$$
$$= 9.23.$$

Sous H_0 : $\beta_{1j} = \beta_{2j} = 0$, $j = 1, ..., 4$ vraie, la statistique D_{Village} suit une loi du Khi-deux à 8 degrés de liberté. D'où son niveau de signification $P = \text{Prob}(\chi^2_{(8)} \geqslant 9.23) = 0.3$.

L'effet Village n'est pas significatif.

(4) L'estimation des paramètres du modèle logistique et les tests correspondants sont donnés dans le tableau 8. On en déduit les estimations des logits généralisés.

Par exemple :

$$\widehat{\text{Log}}(\pi_{11}/\pi_{15}) = 4.39 - 0.60 + 1.68 - 0.52 + 0.33 = 5.28.$$

Et de même :

$$\widehat{\text{Log}}(\pi_{12}/\pi_{15}) = 3.16, \quad \widehat{\text{Log}}(\pi_{13}/\pi_{15}) = 4.22, \text{ et } \widehat{\text{Log}}(\pi_{14}/\pi_{15}) = 2.93.$$

On peut ensuite estimer les probabilités π_{ij}. Sur notre exemple il vient :

$$\widehat{\pi}_{11} = \widehat{\pi}_{15}e^{5.28},\ \widehat{\pi}_{12} = \widehat{\pi}_{15}e^{3.16},\ \widehat{\pi}_{13} = \widehat{\pi}_{15}e^{4.22} \text{ et } \widehat{\pi}_{14} = \widehat{\pi}_{15}e^{2.93}.$$

TABLEAU 8
Estimation et Test des paramètres du modèle logistique
[Maximum de vraisemblance]

Effet	Numéro de la fonction de réponse	Estimation des paramètres	Ecart-type	Statistique de Wald	Niveau de signification
Constante	1	4.39	0.40	117.85	0.0001
	2	2.98	0.41	52.86	0.0001
	3	3.49	0.41	73.71	0.0001
	4	2.29	0.42	29.66	0.0001
Village 1	1	−0.60	0.33	3.34	0.0678
	2	−0.77	0.35	4.72	0.0297
	3	−0.49	0.33	2.17	0.1411
	4	−0.47	0.34	1.97	0.1607
2	1	0.52	0.35	2.46	0.1165
	2	0.68	0.34	4.03	0.0447
	3	0.52	0.33	2.38	0.1226
	4	0.47	0.34	1.89	0.1690
Sexe H	1	1.68	0.37	21.04	0.0001
	2	1.26	0.37	11.66	0.0006
	3	1.10	0.37	8.98	0.0027
	4	0.55	0.37	2.16	0.1412
Iode Abs	1	−0.52	0.36	2.04	0.1531
	2	−0.07	0.38	0.03	0.8552
	3	0.047	0.36	0.02	0.8966
	4	0.59	0.38	2.45	0.1177
Jour 0	1	0.33	0.22	2.16	0.1415
	2	−0.25	0.24	1.09	0.2966
	3	0.07	0.22	0.09	0.7687
	4	−0.02	0.23	0.00	0.9454

On en déduit

$$\widehat{\pi}_{15} = (1 + e^{5.28} + e^{3.16} + e^{4.22} + e^{2.93})^{-1} = 0.00325$$

puis

$$\widehat{\pi}_{11} = 0.00325 \times e^{5.28} = 0.638$$
$$\widehat{\pi}_{12} = 0.00325 \times e^{3.16} = 0.076$$
$$\widehat{\pi}_{13} = 0.00325 \times e^{4.22} = 0.221$$
$$\widehat{\pi}_{14} = 0.00325 \times e^{2.93} = 0.061$$

Ces probabilités peuvent être comparées aux proportions observées :

$$p_{11} = 0.61$$
$$p_{12} = 0.07$$
$$p_{13} = 0.26$$
$$p_{14} = 0.06$$
$$p_{15} = 0.$$

L'adéquation est ici excellente.

c) Modélisation des logits cumulés

La variable dépendante Y = Niveau de goitre étant ordinale nous avons essayé de modéliser le logit cumulé :

$$\text{Log}\left(\frac{P(Y > j)}{P(Y \leqslant j)}\right) = \beta_{oj} + \underset{\text{Village}}{\begin{matrix}1\\2\\3\end{matrix}\begin{bmatrix}\beta_{1j}\\ \beta_{2j}\\ -\beta_{1j}-\beta_{2j}\end{bmatrix}} + \underset{\text{Sexe}}{\begin{matrix}\text{H}\\ \text{F}\end{matrix}\begin{bmatrix}\beta_{3j}\\ -\beta_{3j}\end{bmatrix}} + \underset{\text{Iode}}{\begin{matrix}\text{Abs}\\ \text{Prés}\end{matrix}\begin{bmatrix}\beta_{4j}\\ -\beta_{4j}\end{bmatrix}} + \underset{\text{Jour}}{\begin{matrix}0\\ 180\end{matrix}\begin{bmatrix}\beta_{5j}\\ -\beta_{5j}\end{bmatrix}}$$

Les paramètres sont estimés en utilisant la méthode des moindres carrés généralisée et on aboutit au tableau d'analyse de la variance donné dans le tableau 9.

TABLEAU 9
Tableau d'analyse de la variance
(Fonction de réponse : logit cumulé)

Source de variation	Degrés de liberté	Statistique de Wald	Niveau de signification
Constante	4	683.31	0.0001
Village	8	8.20	0.4145
Sexe	4	240.82	0.0001
Iode	4	51.28	0.0001
Jour	4	24.67	0.0001
Résidu	24	23.82	0.4718

On retrouve que le Village n'est pas significatif. Le résidu n'est pas du tout significatif. Son niveau de signification vaut 0.4718. La modélisation des logits cumulés est donc acceptable.

d) Modélisation de la moyenne

La modélisation de la liaison entre une variable ordinale et des facteurs explicatifs est particulièrement facile à analyser lorsqu'on utilise la fonction de réponse «Moyenne» :

$$F(\pi_i) = \sum_{j=1}^{5} j\pi_{ij} = \beta_{oj}$$

$$+ \underset{\text{Village}}{\begin{matrix}1\\2\\3\end{matrix}\begin{bmatrix}\beta_1\\ \beta_2\\ -\beta_1-\beta_2\end{bmatrix}} + \underset{\text{Sexe}}{\begin{matrix}\text{H}\\ \text{F}\end{matrix}\begin{bmatrix}\beta_3\\ -\beta_3\end{bmatrix}} + \underset{\text{Iode}}{\begin{matrix}\text{Abs}\\ \text{Prés}\end{matrix}\begin{bmatrix}\beta_4\\ -\beta_4\end{bmatrix}} + \underset{\text{Jour}}{\begin{matrix}0\\ 180\end{matrix}\begin{bmatrix}\beta_5\\ -\beta_5\end{bmatrix}}$$

La fonction de réponse $F(\pi_i) = \sum_{j=1}^{5} j\pi_{ij}$ représente $E(Y/i)$ moyenne de Y dans la population i, lorsqu'on considère que la variable Y prend les valeurs numériques 1 à 5. Ce modèle est donc tout-à-fait analogue à une analyse de la variance de Y sur les facteurs Village,Sexe,Iode et Jour. Mais on ne suppose pas que la loi de probabilité de Y dans la population i est normale avec une variance constante d'une population à l'autre. On suppose simplement que les données disponibles suivent une loi multinomiale propre à chaque population i. Ainsi le modèle linéaire généralisé semble nettement préférable à l'analyse de la variance dans la modélisation d'une variable ordinale Y à l'aide de facteurs explicatifs.

En utilisant la méthode des moindres carrés généralisée pour estimer les paramètres du modèle «Moyenne» on obtient le tableau d'analyse de la variance donné dans le tableau 10.

TABLEAU 10
Tableau d'analyse de la variance
(Fonction de réponse : «Moyenne»)

Source de variation	Degrés de liberté	Statistique de Wald	Niveau de signification
Constante	1	5137.84	0.0001
Village	2	3.96	0.1381
Sexe	1	279.85	0.0001
Iode	1	36.96	0.0001
Jour	1	9.90	0.0017
Résidu	6	11.02	0.0877

Le score moyen est ainsi convenablement reconstitué à l'aide des effets principaux. Le village reste non significatif.

On peut écrire le modèle estimé

$$\widehat{F}(\pi_i) = \widehat{E}(Y/i) = 1.97 + \underset{\text{Village}}{\begin{matrix}1\\2\\3\end{matrix}\begin{bmatrix}0.09\\-0.03\\-0.06\end{bmatrix}} + \underset{\text{Sexe}}{\begin{matrix}\text{H}\\\text{F}\end{matrix}\begin{bmatrix}-0.39\\0.39\end{bmatrix}} + \underset{\text{Iode}}{\begin{matrix}\text{Abs}\\\text{Prés}\end{matrix}\begin{bmatrix}0.30\\-0.30\end{bmatrix}} + \underset{\text{Jour}}{\begin{matrix}0\\180\end{matrix}\begin{bmatrix}-0.13\\0.13\end{bmatrix}}$$

analogue à une équation de régression, mais les coefficients ont été estimés sous l'hypothèse de loi multinomiale beaucoup moins restrictive que celle de loi normale.

II. UTILISATION DE L'ANALYSE DES CORRESPONDANCES

1. Analyse des correspondances des données du Mali et Biplot

Il est tout-à-fait naturel de réaliser une analyse des correspondances du tableau des effectifs n_{ij} observés, nombre de personnes ayant la réponse $Y = j$ dans l'échantillon tiré de la population i. Rappelons les notations usuelles de l'analyse des correspondances :

Effectif : $k_{ij} = n_{ij}$

Total des effectifs : $k = \sum_{i=1}^{s}\sum_{j=1}^{r} k_{ij}$

Fréquence : $f_{ij} = k_{ij}/k$

Effectif marginal en ligne : $k_{i.} = \sum_{j=1}^{r} k_{ij} = n_i$

Effectif marginal en colonne : $k_{.j} = \sum_{i=1}^{s} k_{ij}$

Fréquence marginale en ligne : $f_{i.} = k_{i.}/k$

Fréquence marginale en colonne : $f_{.j} = k_{.j}/k = \widehat{\text{Prob}}(Y = j)$

Fréquence conditionnelle en ligne : $f_{ij}/f_{i.} = p_{ij} = \widehat{\text{Prob}}(Y = j/i)$

Profil-ligne : $f_J^i = (\ldots, p_{ij}, \ldots) = p_i$

Profil-colonne : $f_I^j = (\ldots, f_{ij}/f_{.j}, \ldots)$

F_h : composante principale du nuage des profils-lignes associée à la valeur propre λ_h.

G_h : composante principale du nuage des profils-colonnes associée à la valeur propre λ_h.

Nous avons réalisé une analyse des correspondances du tableau des données du Mali formé des colonnes du tableau 1 correspondant aux différents niveaux de goitre. Les résultats sont donnés dans le tableau 11 et le premier plan principal dans la figure 1. La signification des identificateurs des lignes est claire :

V1S1I1J1 = Village 1/Sexe1/Iode1/Jour0. Le premier plan principal explique 94,26 % de l'inertie totale. Toutes les lignes et les colonnes du tableau 1 sont bien représentées dans ce plan. Le premier axe traduit exactement le niveau de goitre puisque $G_1(j)$ est une fonction décroissante du niveau de goitre j. Le deuxième axe met en évidence le rôle particulier des goitres de niveau 2. Dans le tableau 12 on a construit les différences de profils entre les jours 180 et 0, pour un même village et un même sexe.

TABLEAU 11

Résultats de l'analyse des correspondances des données du Mali

EDITION DES VALEURS PROPRES

APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION 0.2247
SOMME DES VALEURS PROPRES 0.2247

HISTOGRAMME DES 4 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	0.1835	81.66	81.66	**********
2	0.0283	12.58	94.24	*******
3	0.0107	4.76	99.00	***
4	0.0023	1.00	100.00	*

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES INDIVIDUS SUR LES AXES 1 A 4

INDIVIDUS ACTIFS

INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0
V1S1I1J1	8.40	0.13	0.27	-0.10	0.19	0.08	0.00	3.4	3.1	20.4	26.8	0.0	0.58	0.08	0.28	0.06	0.00
V1S1I1J2	7.30	0.09	-0.01	0.29	0.10	-0.03	0.00	0.0	21.1	7.4	2.3	0.0	0.00	0.88	0.12	0.01	0.00
V1S2I1J1	11.76	0.25	-0.46	-0.16	-0.05	0.05	0.00	13.8	11.3	3.0	10.7	0.0	0.87	0.11	0.01	0.01	0.00
V1S2I1J2	10.23	0.43	-0.64	-0.06	-0.13	0.00	0.00	22.8	1.3	16.1	0.1	0.0	0.95	0.01	0.04	0.00	0.00
V2S1I1J1	10.18	0.09	0.30	0.01	0.04	0.03	0.00	5.1	0.0	1.6	3.2	0.0	0.97	0.00	0.02	0.01	0.00
V2S1I2J2	9.31	0.38	0.60	-0.04	-0.12	-0.02	0.00	18.5	0.5	13.5	1.8	0.0	0.95	0.00	0.04	0.00	0.00
V2S2I1J1	9.94	0.15	-0.35	-0.06	0.15	-0.09	0.00	6.5	1.3	20.4	36.3	0.0	0.78	0.02	0.14	0.05	0.00
V2S2I2J2	8.26	0.11	0.10	0.30	-0.07	-0.02	0.00	0.4	26.7	3.4	1.0	0.0	0.09	0.87	0.04	0.00	0.00
V3S1I1J1	5.71	0.35	0.53	-0.25	0.00	-0.04	0.00	8.9	12.5	0.0	3.1	0.0	0.82	0.18	0.00	0.00	0.00
V3S1I2J2	5.46	0.48	0.68	-0.12	-0.09	-0.03	0.00	14.1	2.7	4.3	2.9	0.0	0.95	0.03	0.02	0.00	0.00
V3S2I1J1	6.87	0.17	-0.41	0.00	0.03	-0.01	0.00	6.4	0.0	0.7	0.2	0.0	0.99	0.00	0.01	0.00	0.00
V3S2I2J2	6.39	0.10	-0.06	0.29	-0.05	0.06	0.00	0.1	19.3	1.3	11.6	0.0	0.03	0.90	0.02	0.04	0.00

FREQUENCES			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDEN - LIBELLE COURT	P.REL	DISTO	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0
FREQUENCES ACTIVES																	
GOI1 - GOI1 : goitre1	47.19	0.16	0.39	-0.09	-0.01	0.00	0.00	39.4	13.0	0.4	0.0	0.0	0.95	0.05	0.00	0.00	0.00
GOI2 - GOI2 : goitre2	13.30	0.15	0.01	0.37	-0.11	-0.04	0.00	0.0	63.6	15.4	7.7	0.0	0.00	0.91	0.08	0.01	0.00
GOI3 - GOI3 : goitre3	24.15	0.11	-0.28	0.08	0.13	0.04	0.00	10.5	6.0	38.7	20.7	0.0	0.75	0.07	0.16	0.02	0.00
GOI4 - GOI4 : goitre4	13.68	0.61	-0.76	-0.18	-0.03	-0.07	0.00	42.5	15.1	0.9	27.8	0.0	0.94	0.05	0.00	0.01	0.00
GOI5 - GOI5 : goitre5	1.68	1.20	-0.90	-0.20	-0.53	0.24	0.00	7.5	2.3	44.7	43.8	0.0	0.68	0.03	0.24	0.05	0.00

Il y a aggravation de la situation dans le village témoin 1 où le diffuseur d'iode n'a pas été installé dans les puits, et amélioration systématique dans les villages 2 et 3 où les diffuseurs d'iode ont été installés. Le niveau de goitre 2 joue un rôle particulier puisque dans les 3 villages il y a augmentation des proportions des goitres de niveau 2 des jours 0 à 180. Dans le village 1 il y a des goitres de niveau 1 qui passent en 2. Dans les autres villages il y a des 3 ou 4 qui passent en 2.

L'analyse des positions des profils-lignes dans le premier plan principal est très claire :

(1) Au jour 0 (J1) les trois villages se regroupent à sexe fixé : les hommes (S1) à droite (du côté des petits niveaux de goitre), les femmes (S2) à gauche (du côté des plus forts niveaux de goitre). Les profils des niveaux de goitre diffèrent beaucoup d'un sexe à l'autre, et peu entre village à sexe fixé.

TABLEAU 12
Evolution des profils du jour 0 au jour 180

		Niveau de goitre				
Village	Sexe	1	2	3	4	5
1	1	−0.21	0.14	0.04	0.04	0
1	2	−0.10	0.05	0.00	0.04	0.01
2	1	0.15	0.02	−0.11	−0.06	0
2	2	0.11	0.13	−0.08	−0.16	0.00
3	1	0.03	0.05	−0.03	−0.05	0
3	2	0.08	0.09	−0.03	−0.14	−0.01

(2) Au jour 180 (J2) on observe une nette typologie des profils en trois groupes.

Premier groupe : Les hommes des villages 2 et 3 ont des profils qui se sont améliorés et rapprochés. Ils témoignent de l'effet positif du diffuseur d'iode chez les hommes de ces villages.

Deuxième groupe : Il rassemble les hommes du village 1 (non traité) et les femmes des villages 2 et 3. Ainsi on observe que le diffuseur d'iode ramène le profil des femmes traitées à celui des hommes sans traitement.

Troisième groupe : Les femmes du village témoin 1 subissent une aggravation de leurs niveaux de goitre.

Ainsi l'effet du diffuseur d'iode est clairement visualisé : il y a amélioration systématique dans les villages traités (flèches reliant les villages 2 et 3, sexes 1 et 2, aux jours 0 et 180 allant vers la droite) et dégradation systématique dans le village non traité (flèche reliant le village 1, Sexes 1 et 2 aux jours 0 et 180 allant vers la gauche).

Les données ici ne sont pas du tout symétriques. Il est possible de prendre en compte cette situation au niveau graphique. La formule de décomposition à

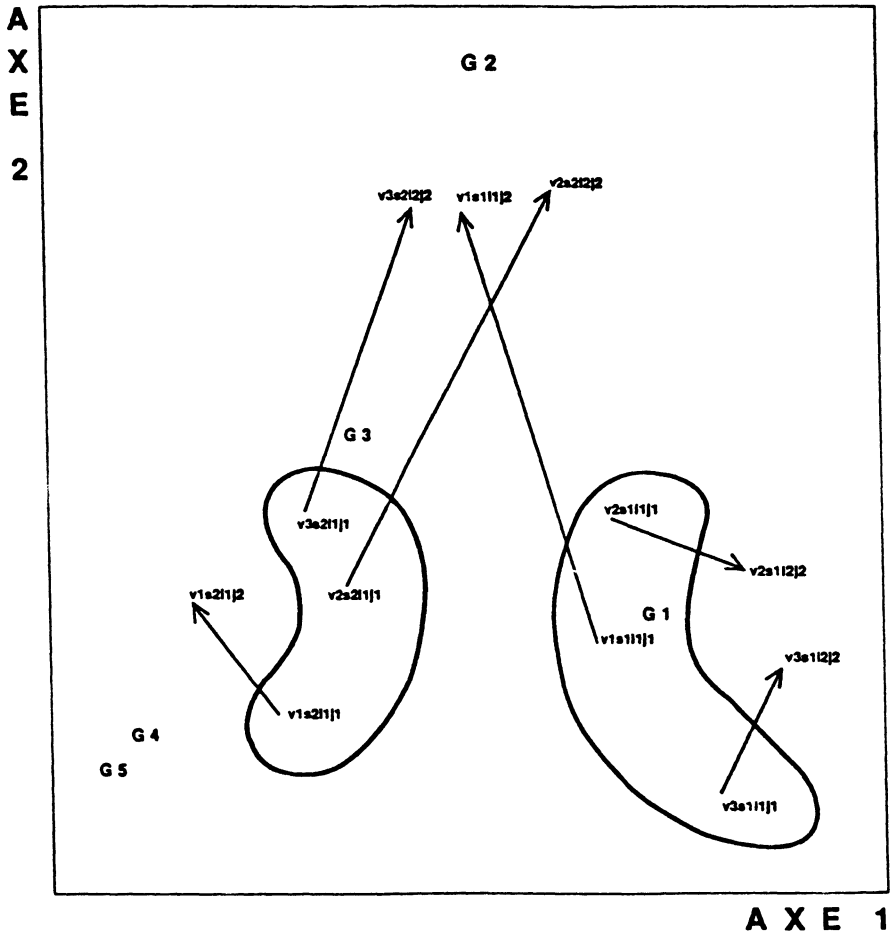


FIGURE 1
Le premier plan principal

l'ordre 2 s'écrit :

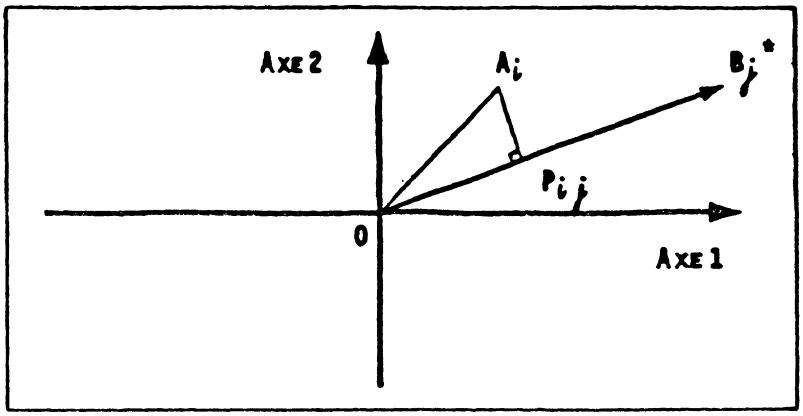
$$\frac{f_{ij}}{f_{i.}} \approx f_{.j}\Big(1 + \sum_{h=1}^{2} \frac{1}{\sqrt{\lambda_h}} F_h(i) G_h(j)\Big) \tag{6}$$

Sur notre exemple l'approximation est excellente pour tous les couples (i, j).

Posons $G_h^*(j) = f_{.j} G_h(j)/\sqrt{\lambda_h}$. La formule (6) s'écrit donc :

$$\frac{f_{ij}}{f_{i.}} - f_{.j} \approx \sum_{h=1}^{2} F_h(i) G_h^*(j) \tag{7}$$

Suivant les idées de Gabriel (1982), il est alors intéressant de construire la carte «biplot» superposant $(F1, F2)$ et (G_1^*, G_2^*) :



On pose $A_i = (F_1(i), F_2(i))$ et $B_j^* = (G_1^*(j), G_2^*(j))$.

La formule (7) donne :

$$\frac{f_{ij}}{f_{i.}} - f_{.j} \approx \langle A_i, B_j^* \rangle$$

produit scalaire entre les vecteurs A_i et B_j^*.

On a donc approximativement :

$$\text{angle } A_i O B_j^* \text{aigu} \rightarrow \widehat{\text{Prob}}(Y = j/i) > \widehat{\text{Prob}}(Y = j)$$
$$\text{angle } A_i O B_j^* \text{obtu} \rightarrow \widehat{\text{Prob}}(Y = j/i) < \widehat{\text{Prob}}(Y = j)$$

Notons P_{ij} la projection du point A_i sur l'axe (engendré par le vecteur) B_j^*.

De $\overline{OP_{ij}} = \frac{1}{\|B_j^*\|}\langle A_i, B_j^* \rangle$ on déduit $\widehat{\text{Prob}}(Y = j/i) - \widehat{\text{Prob}}(Y = j) \approx \|B_j^*\|\overline{OP_{ij}}$.

La répartition des P_{ij} sur l'axe B_j^* reflète donc approximativement les écarts entre les probabilités $\widehat{\text{Prob}}(Y = j/i)$ et $\widehat{\text{Prob}}(Y = j)$.

On a calculé G_1^* et G_2^* :

Niveau de goitre	G_1^*	G_2^*
1	0.43	−0.253
2	0.003	0.293
3	−0.16	0.115
4	−0.24	−0.147
5	−0.035	−0.02

D'où le biplot de la figure 2, où nous avons représenté (F_1, F_2) et $2(G_1^*, G_2^*)$ pour des raisons de lisibilité. Seuls les angles et les projections entre les A_i et les B_j^* étant interprétés, cette dilatation de la représentation des niveaux de goitre est sans influence sur l'interprétation des résultats.

Interprétons le biplot :

(1) Pour les villages 2 et 3 on visualise l'augmentation des probabilités des goitres de niveau 1, pour les hommes comme pour les femmes, lorsqu'on passe du jour 0 au jour 180. On visualise également l'augmentation des probabilités des goitres de niveau 2 pour tous les villages et les deux sexes, à l'exception du village 2, sexe 1, mais ici le graphique ne correspond pas à la réalité.

(2) Dans le village 1 on observe l'augmentation pour les deux sexes des probabilités des goitres de niveau 2, 3 et 4, et la diminution des goitres de niveau 1.

(3) La position de G5 (goitre de niveau 5) est due à la faible fréquence de ce type de goitre. Sur les six mois observés il y a stabilité des probabilités des goitres de niveau 5. Il faut signaler que ce type de goitre est irréversible.

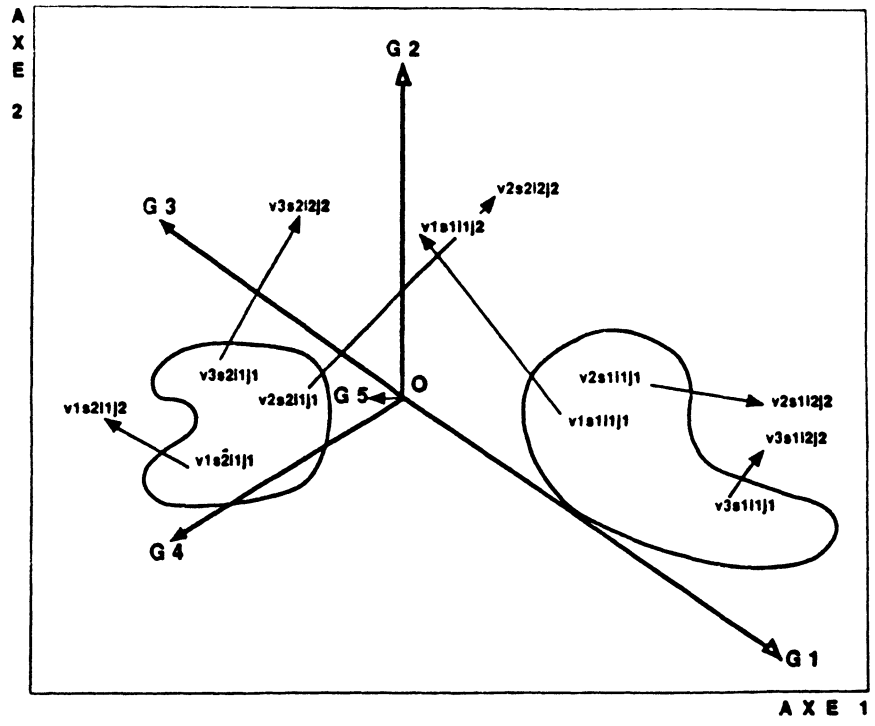


FIGURE 2
Biplot des données du Mali : (F_1, F_2) *et* $2(G_1^*, G_2^*)$

2. Analyse de la variance et analyse des correspondances

Les profils-lignes sont complètement décrits par les composantes principales (F_1, F_2, F_3, F_4). Ces profils-lignes dépendent-ils des variables Village,Sexe,Iode et Jour? Une analyse de la variance sur chaque composante F_h ou globale sur l'ensemble des composantes permet de répondre à cette question. Nous donnons dans le tableau 13 les analyses de la variance univariées et dans le tableau 14 l'analyse de la variance multivariée. Ces analyses de la variance ont été effectuées en pondérant chaque profil-ligne par sa fréquence.

Les composantes principales F_h sont obtenues par analyse canonique entre les variables indicatrices des lignes et celles des colonnes. Il ne faut donc pas interpréter les F des tableaux 13 et 14 comme provenant d'une loi de Fisher-Snedecor. Mais ils gardent une valeur descriptive que nous pouvons exploiter. Décidons qu'un F est «significatif» dès qu'il est supérieur à 4. On retrouve dans le tableau 14 l'essentiel de l'approche plus rigoureuse par le modèle linéaire généralisé. Dans le tableau 13 on constate que seule la première composante principale F_1 est reliée aux variables explicatives des variations entre profils-lignes. Les résultats sont plus marqués que dans l'analyse globale qui intègre les composantes principales F_2, F_3, F_4 peu liées aux variables explicatives.

TABLEAU 13

Analyses de la variance des composantes principales F_1, F_2, F_3, F_4

Source de variation	Valeurs de F pour les composantes principales			
	F_1	F_2	F_3	F_4
Village	0.94	0.23	0.05	2.99
Sexe	159	0.26	0.58	0.15
Iode	24.20	0.05	0.26	3.02
Jour	6.35	1.60	0.70	2.91

TABLEAU 14

Analyse de la variance multivariée des composantes principales F_1, F_2, F_3, F_4

Source de variation	Λ de Wilks	F de Rao
Village	0.27	0.68
Sexe	0.03	23.37
Iode	0.14	4.6
Jour	0.14	4.6

On peut résumer les variations entre profils en utilisant la formule de décomposition des $\frac{f_{ij}}{f_{i.}}$ à l'ordre 1 et la décomposition de F_1 par l'analyse de la variance. On a obtenu la décomposition suivante de F_1 :

$$F_1(i) \approx 0.003 + \begin{matrix}1\\2\\3\end{matrix}\begin{bmatrix}-0.11\\-0.02\\0\end{bmatrix} + \frac{1}{2}\begin{bmatrix}0.68\\0\end{bmatrix} + \frac{1}{2}\begin{bmatrix}-0.55\\0\end{bmatrix} + \frac{1}{2}\begin{bmatrix}0.22\\0\end{bmatrix}$$

Village Sexe Iode Jour

Le R^2 de l'ajustement vaut 0.976 et l'écart-type $\widehat{\sigma}$ est égal à 0.0269. La constante peut être négligée.

La formule de reconstitution à l'ordre 1 s'écrit

$$\frac{f_{ij}}{f_{i.}} = f_{.j}\left(1 + \frac{1}{\sqrt{\lambda_1}} F_1(i) G_1(j)\right)$$

D'où :

$$\begin{bmatrix}\widehat{\text{Prob}}(Y=1/i)\\ \widehat{\text{Prob}}(Y=2/i)\\ \widehat{\text{Prob}}(Y=3/i)\\ \widehat{\text{Prob}}(Y=4/i)\\ \widehat{\text{Prob}}(Y=5/i)\end{bmatrix} = \begin{bmatrix}\widehat{\text{Prob}}(Y=1)\\ \widehat{\text{Prob}}(Y=2)\\ \widehat{\text{Prob}}(Y=3)\\ \widehat{\text{Prob}}(Y=4)\\ \widehat{\text{Prob}}(Y=5)\end{bmatrix}\left(1 + (1/\sqrt{\lambda_1})F_1(i)\begin{bmatrix}G_1(1)\\G_1(2)\\G_1(3)\\G_1(4)\\G_1(5)\end{bmatrix}\right) \qquad (8)$$

$$\approx \begin{bmatrix}0.47\\0.13\\0.24\\0.14\\0.02\end{bmatrix}\left(1 + \left(\left(\begin{matrix}1\\2\\3\end{matrix}\begin{bmatrix}-0.11\\0.02\\0\end{bmatrix} + \frac{1}{2}\begin{bmatrix}0.68\\0\end{bmatrix} + \frac{1}{2}\begin{bmatrix}-0.55\\0\end{bmatrix} + \frac{1}{2}\begin{bmatrix}-0.22\\0\end{bmatrix}\right)\begin{bmatrix}0.91\\0.02\\-0.65\\-1.77\\-2.1\end{bmatrix}\right)\right)$$

Village Sexe Iode Jour

En comparant les profils observés et les profils reconstitués on peut constater que les ordres de grandeur sont respectés. Nous donnons trois exemples dans le tableau 15. On vérifie que la qualité de la reconstitution dépend bien sûr de la valeur du cosinus carré du profil avec l'axe 1.

TABLEAU 15
Profils observés et reconstitués

	1	2	3	4	5	
V1S1I1J1 $\cos^2$ avec l'axe 1 = 0.58	0.61	0.07	0.26	0.06	0	observé
	0.57	0.13	0.20	0.08	0.00	reconstitué
V3S2I2J2 $\cos^2$ avec l'axe 1 = 0.03	0.38	0.22	0.29	0.09	0.00	observé
	0.47	0.13	0.24	0.14	0.02	reconstitué
V3S2I1J1 $\cos^2$ avec l'axe 1 = 0.99	0.29	0.13	0.31	0.24	0.03	observé
	0.33	0.13	0.29	0.22	0.03	reconstitué

Concluons en disant que la formule (8) fournit globalement un bon résumé de la variation des profils des niveaux de goitre en fonction des variables Village, Sexe, Iode et Jour.

Conclusion

Lorsque les effectifs des cases du tableau croisant $(X_1 * X_2 * ... * X_k)$ par Y sont suffisamment importants nous avons montré qu'il était intéressant d'utiliser simultanément le modèle linéaire généralisé et l'analyse des correspondances. Le modèle linéaire généralisé permet de décrire de manière précise l'influence de chaque variable explicative sur la loi de probabilité de Y. Les résultats de l'analyse se présentent d'une manière comparable à ceux d'une régression multiple ou d'une analyse de la variance. L'analyse des correspondances fournit une carte des profils-lignes et des profils-colonnes donnant une vision globale de la dispersion des profils-lignes. Une analyse de la variance multivariée des composantes principales définies sur les profils-lignes sur les facteurs $X_1, ..., X_k$ permet de restituer les résultats du modèle linéaire généralisé, d'une manière sans doute plus descriptive, les niveaux de signification des tests ne pouvant être évalués. Le modèle linéaire généralisé permet de valider statistiquement des résultats que l'analyse des correspondances ne fait que suggérer. L'analyse des correspondances donne une information synthétique mettant en valeur les résultats du modèle linéaire généralisé. L'analyse des correspondances apparaît ainsi comme le soubassement descriptif naturel du modèle linéaire généralisé.

Références

AGRESTI A., (1990). *Categorical Data Analysis*, Wiley.

GABRIEL K.R., (1982). *Biplot*, in S. Kotz, N.L. Johnson & C. B. Read (Eds) : Encyclopedia of Statistical Sciences, Vol. 1, Wiley.

HOSMER D.W. & LEMESHOW S., (1989). *Applied Logistic Regression*, Wiley.

McCULLAGH P. & NELDER J.A., (1989). *Generalized Linear Models*, Chapman and Hall, 2^e ed.

SAS Institute (1988). *Categorical Data Analysis*, Course Notes, SAS Institute.

SAS Institute (1990). *SAS/STAT User's guide*, Version 6, SAS Institute.