

REVUE DE STATISTIQUE APPLIQUÉE

J. J. DAUDIN

P. TRECOURT

Analyse factorielle des correspondances et modèle log-linéaire : comparaison des deux méthodes sur un exemple

Revue de statistique appliquée, tome 28, n° 1 (1980), p. 5-24

http://www.numdam.org/item?id=RSA_1980__28_1_5_0

© Société française de statistique, 1980, 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/>

ANALYSE FACTORIELLE DES CORRESPONDANCES ET MODELE LOG-LINEAIRE : COMPARAISON DES DEUX METHODES SUR UN EXEMPLE

J.J. DAUDIN et P. TRECOURT

Institut National Agronomique Paris — Grignon

Service de Mathématiques

16 rue Claude Bernard — 75005 PARIS

RESUME. — Quand dans une table de contingence, l'une des deux variables est de structure factorielle, il est intéressant de comparer les résultats de l'analyse factorielle des correspondances qui ne tient pas compte de cette structure à ceux donnés par le modèle log-linéaire qui la prend en compte. C'est ce que nous avons fait à partir de la table des répartitions des exploitations agricoles par région de programme et type de main d'œuvre. Si chaque méthode a son intérêt pour l'étude de la liaison entre type de main d'œuvre et région de programme, le modèle log-linéaire est mieux adapté à l'étude des liaisons globales entre toutes les variables concernées.

1. INTRODUCTION

Les données sont issues de l'enquête structure CEE de 1969 ([2] p. 300-301). Un échantillon de 19921 exploitations agricoles est réparti selon les 21 régions de programme françaises d'une part et 32 variables caractérisant le type de main d'œuvre d'autre part. Ces dernières ont une structure factorielle : elles sont le résultat du croisement de cinq variables binaires :

Variable 1 : Chef d'exploitation (à temps partiel-à temps complet) ;

Variable 2 : Aide familiale à temps complet (absence-présence) ;

Variable 3 : Aide familiale à temps partiel (absence-présence) ;

Variable 4 : Salariés à temps complet (absence-présence) ;

Variable 5 : Salariés saisonniers (absence-présence).

Les données sont résumées par la table de contingence à deux entrées (Table 1).

Nous voulons étudier les liaisons entre toutes les variables et notamment la liaison entre le groupe de variables caractérisant le type de main d'œuvre et la variable région de programme. Le modèle log-linéaire étant d'emploi moins courant en France que l'analyse factorielle des correspondances, nous en développerons davantage les résultats.

2. ANALYSE FACTORIELLE DES CORRESPONDANCES

L'analyse de la table 1 a donné les résultats suivants rapidement résumés ici :

<i>Valeurs propres</i>	<i>pourcentage d'inertie</i>	<i>pourcentage cumulé</i>
0,135	49,3	49,3
0,046	16,7	66,0
0,022	08,2	74,2

Les 2 premiers axes, qui représentent 66 % de la liaison type de main d'œuvre x Région, donnent lieu à une interprétation très nette comme on va le voir sur les figures 1 à 8 qui représentent toutes, le plan principal. La figure 1 est la représentation graphique brute du plan principal, les figures 2 à 7 en détaillent des aspects particuliers. Précisons les legendes :

Dans la figure 1, et les suivantes chaque type de main-d'œuvre est décrit par 5 chiffres (0 ou 1), chacun caractérisant une variable :

- 1^{er} chiffre : chef d'exploitation 0 à temps partiel
1 à temps complet
- 2^e chiffre : main d'œuvre familiale à temps complet 0 absente
1 présente
- 3^e chiffre : main d'œuvre familiale à temps partiel 0 absente
1 présente
- 4^e chiffre : salariés permanents 0 absents
1 présents
- 5^e chiffre : salariés saisonniers 0 absents
1 présents

On a représenté sur la figure 2 les modalités supplémentaires suivantes :

- 0 . . . chef d'exploitation à temps partiel (barycentre des points dont le 1^{er} chiffre est 0)
- 1 . . . chef d'exploitation à temps complet (barycentre des points dont le 1^{er} chiffre est 1)
- .01. . . présence de main d'œuvre familiale à temps partiel seulement (barycentre des points ont les 2^e et 3^e chiffres sont respectivement 0 et 1)
- .10. . . présence de main d'œuvre familiale à temps complet seulement
- .00. . . pas de main d'œuvre familiale
- .11. . . présence de main d'œuvre familiale à la fois à temps partiel et complet
- . . .00 pas de salarié
- . . .01 présence de salarié(s) saisonnier seulement
- . . .10 présence de salarié(s) à temps complet seulement
- . . .11 présence de salarié(s) à temps complet + salarié(s) saisonnier

La figure 1 étant assez rébarbative, il est utile de la compléter par les figures 2, 3, 4, 5, 6 et 7 qui vont nous permettre d'en faciliter la lecture.

Dans la figure 2 nous avons gardé seulement les modalités supplémentaires.

TABLE 1

Région de progr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	Ensemble	
Région Par.	9	7	9	6	2	4	2	8	—	9	2	5	—	5	2	6	13	19	5	9	1	3	2	10	—	10	2	5	1	7	2	3	168	
Champagne	29	9	2	3	22	4	5	5	6	19	1	7	—	14	1	4	48	56	2	10	40	62	2	22	6	—	6	35	3	—	28	5	9	465
Picardie	31	12	7	4	8	6	5	10	1	21	2	4	5	16	1	12	38	47	2	18	19	44	8	27	8	32	7	4	13	—	5	5	417	
Hte Normandie	41	14	10	6	10	4	5	4	3	26	3	10	2	13	3	6	93	67	1	24	16	27	—	19	2	—	2	—	5	1	6	—	2	441
Centre	70	21	16	12	74	12	10	9	58	16	7	48	—	16	7	9	117	124	3	45	68	116	9	41	24	10	74	14	6	41	2	7	1048	
Basse Norm.	80	36	27	26	25	15	10	8	17	134	3	51	5	35	1	9	128	137	10	51	32	37	4	24	19	8	43	2	24	4	5	1	5	999
Bourgogne	44	23	8	4	29	18	4	5	4	38	—	3	—	25	—	4	93	142	5	31	60	80	5	19	8	44	2	7	7	46	2	10	630	
Nord	34	13	7	5	11	10	3	6	11	54	—	6	2	33	5	16	59	84	6	22	10	76	3	28	8	44	2	7	7	46	2	10	630	
Lorraine	41	4	1	—	22	8	4	2	6	37	2	1	3	26	—	3	133	62	4	7	54	42	—	4	10	50	1	6	4	20	—	2	559	
Alsace	11	3	2	—	22	2	2	1	—	16	—	2	3	13	—	3	91	33	1	7	98	39	2	3	11	35	—	2	6	26	—	1	435	
Franche-Comté	26	12	1	2	11	5	1	—	8	32	—	4	1	10	—	70	70	—	6	24	26	—	3	10	62	1	2	1	8	—	—	—	396	
Pays-Loire	92	33	20	14	109	26	12	11	12	77	—	16	4	28	1	10	194	376	14	79	117	194	8	29	23	188	2	19	7	41	2	4	1762	
Bretagne	150	37	12	8	36	25	4	7	40	183	2	19	15	81	—	4	257	411	10	43	45	133	4	29	41	199	4	18	10	92	—	12	1931	
Poit. Char.	61	23	13	9	40	18	9	13	9	51	1	4	3	36	2	6	147	205	8	51	49	114	10	34	8	124	3	13	4	30	—	12	1110	
Aquitaine	58	45	16	2	76	31	29	21	26	140	3	6	17	90	1	12	195	233	10	15	115	111	14	31	32	127	2	5	9	49	4	8	1533	
Midi-Pyrénéen	118	62	13	9	44	23	14	8	38	121	1	7	8	38	—	3	258	345	14	19	74	101	11	13	59	178	3	11	9	54	5	8	1669	
Limousin	24	26	7	9	14	10	5	4	7	66	—	8	3	16	—	—	67	135	1	12	28	49	1	3	13	85	1	13	5	20	—	3	635	
Rhône-Alpes	134	45	11	7	122	38	9	8	21	84	1	5	11	24	1	4	326	305	4	16	168	181	5	28	35	122	—	8	12	52	1	12	1800	
Auvergne	38	44	6	2	27	36	4	3	13	48	—	2	8	27	—	1	123	199	10	27	30	93	2	13	21	126	2	14	12	37	1	4	163	
Languedoc	51	9	5	2	271	33	72	26	—	8	1	3	12	12	3	5	158	69	4	3	319	111	40	33	5	10	1	2	15	28	1	5	1317	
Prov-Côte-A Corse	73	20	6	5	127	32	17	11	8	24	—	5	5	20	3	7	105	61	7	8	110	82	7	17	8	32	—	3	10	28	5	9	855	
France	1295	498	199	135	1102	360	228	171	239	1246	22	184	114	600	24	124	2713	3180	121	503	1477	1721	137	430	325	1670	34	192	132	660	35	130	19921	

Code n° temps complet 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32

temps partiel 1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31

temps complet 9 ————— 16

temps partiel 1 ————— 16

temps complet 34 78 1112 1516 1920 2324 2728 3132

saisonnier 5 — 8 13 16 21 24 29 32

Aide Familiale

Salariés

temps complet

temps partiel

temps complet

temps partiel

temps complet

saisonnier

Les traits hachurés relient les variables concernant les salariés, les traits simples relient les variables de main d'œuvre familiales, et le trait double relie les 2 modalités du chef d'exploitation. Dans tous les cas le sens de la flèche va du 0 vers le 1.

Il est possible d'interpréter le plan principal de la façon suivante :

L'axe 1 est celui des salariés saisonniers, du chef d'exploitation et de la main d'œuvre familiale à temps complet : les régions se trouvant dans la partie positive de l'axe 1 sont celles qui font peu appel aux salariés saisonniers, davantage à la main d'œuvre familiale à temps complet et où les chefs d'exploitation travaillent plutôt à temps complet.

L'axe 2 est celui des salariés permanents et de la main d'œuvre familiale à temps partiel : les régions se trouvant du côté positif de l'axe 2 font appel aux salariés permanents et à la main d'œuvre familiale à temps partiel.

Les figures 3, 4, 5, 6 et 7 permettent de confirmer cette interprétation : nous y avons relié respectivement les 2 modalités des variables chef d'exploitation (temps partiel — temps complet) main d'œuvre familiale à temps complet (absence — présence), à temps partiel (absence — présence), salarié permanent (absence — présence) et saisonnier (absence — présence).

En effet, il est remarquable de constater que toutes les flèches des figures 3 et 4 vont de la gauche vers la droite, toutes celles des figures 5 et 6 sont orientées du haut vers le bas, celles de la figure 7 étant dirigées vers la gauche.

La seule exception à cette propriété de l'analyse est due aux modalités 00011 et 00111 dans la figure 5 : la flèche qui les relie est très légèrement en sens contraire de la tendance générale. Il faut remarquer que ces 2 modalités ont un poids faible : 228 et 430 exploitations ont respectivement la configuration de main d'œuvre 00011 et 00111 sur les 19921 exploitations de l'analyse.

L'analyse factorielle permet également de mettre en évidence un autre phénomène : la main d'œuvre salariée est plus variable d'une région à l'autre que la main d'œuvre familiale et le type de chef d'exploitation : ceci se traduit dans la figure 2 par le fait que les modalités supplémentaires caractéristiques des salariés sont plus éloignées de l'origine. Ce phénomène se traduit également par un allongement plus grand des flèches des figures 6 et 7 par rapport à celles des figures 3, 4 et 5.

Ces résultats obtenus par l'analyse factorielle sont remarquables et substantiels. Il reste cependant des questions auxquelles il semble difficile de répondre.

Les axes suivants ne permettent aucune interprétation valable : il y a donc 34 % de la liaison Région x Type de main d'œuvre qui reste dans l'ombre. D'autre part nous avons analysé cette liaison en considérant une seule variable de type de main d'œuvre à la fois.

N'y a-t-il pas des phénomènes plus compliqués faisant intervenir 2 ou plusieurs de ces variables ensemble ?

Autrement dit quelle est la part des interactions entre ces variables et comment décrire ces dernières ?

Cette question est d'ailleurs à relier à celle concernant l'interprétabilité des axes 3, 4 et suivants. Nous n'avons pas pu les interpréter probablement parce qu'y interviennent justement des interactions entre les variables de main d'œuvre.

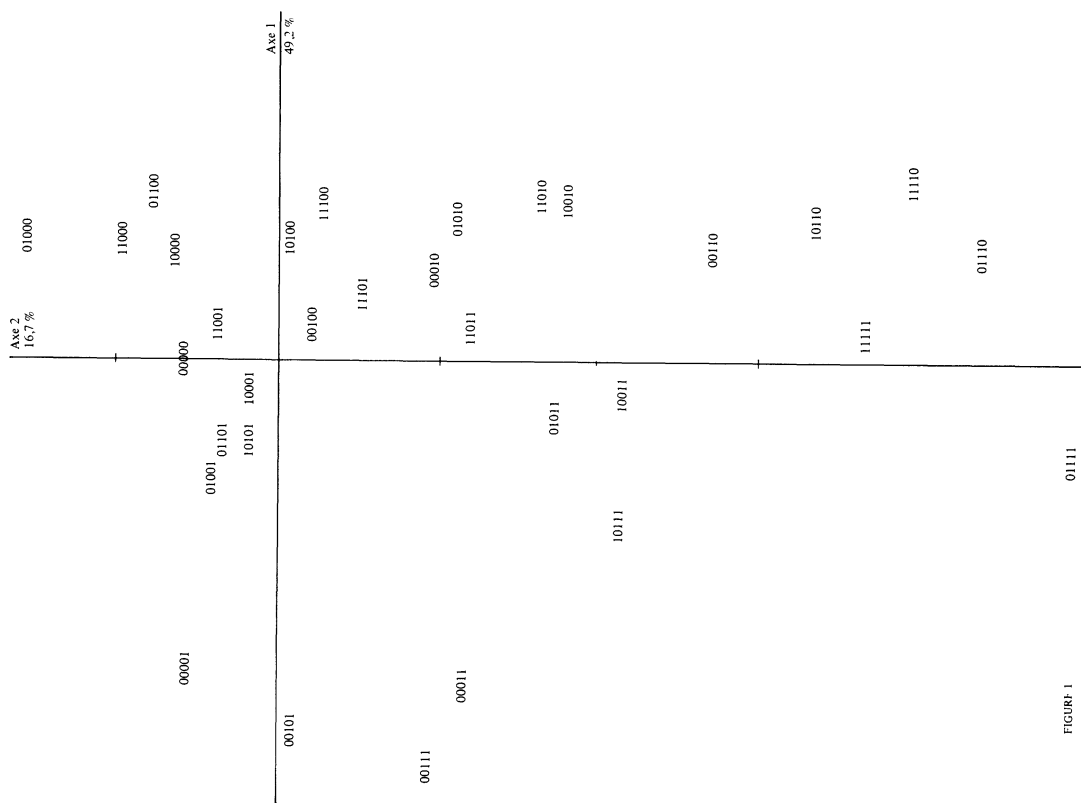


FIGURE 1

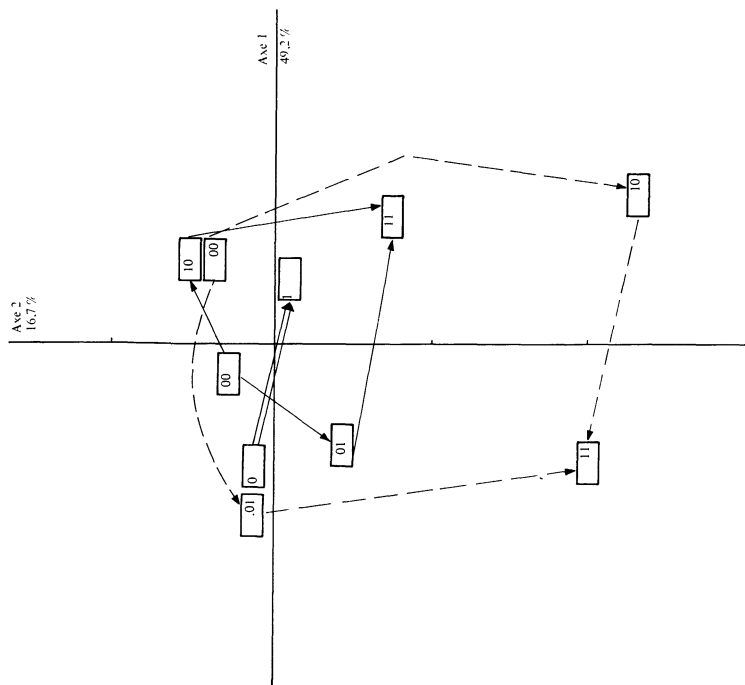
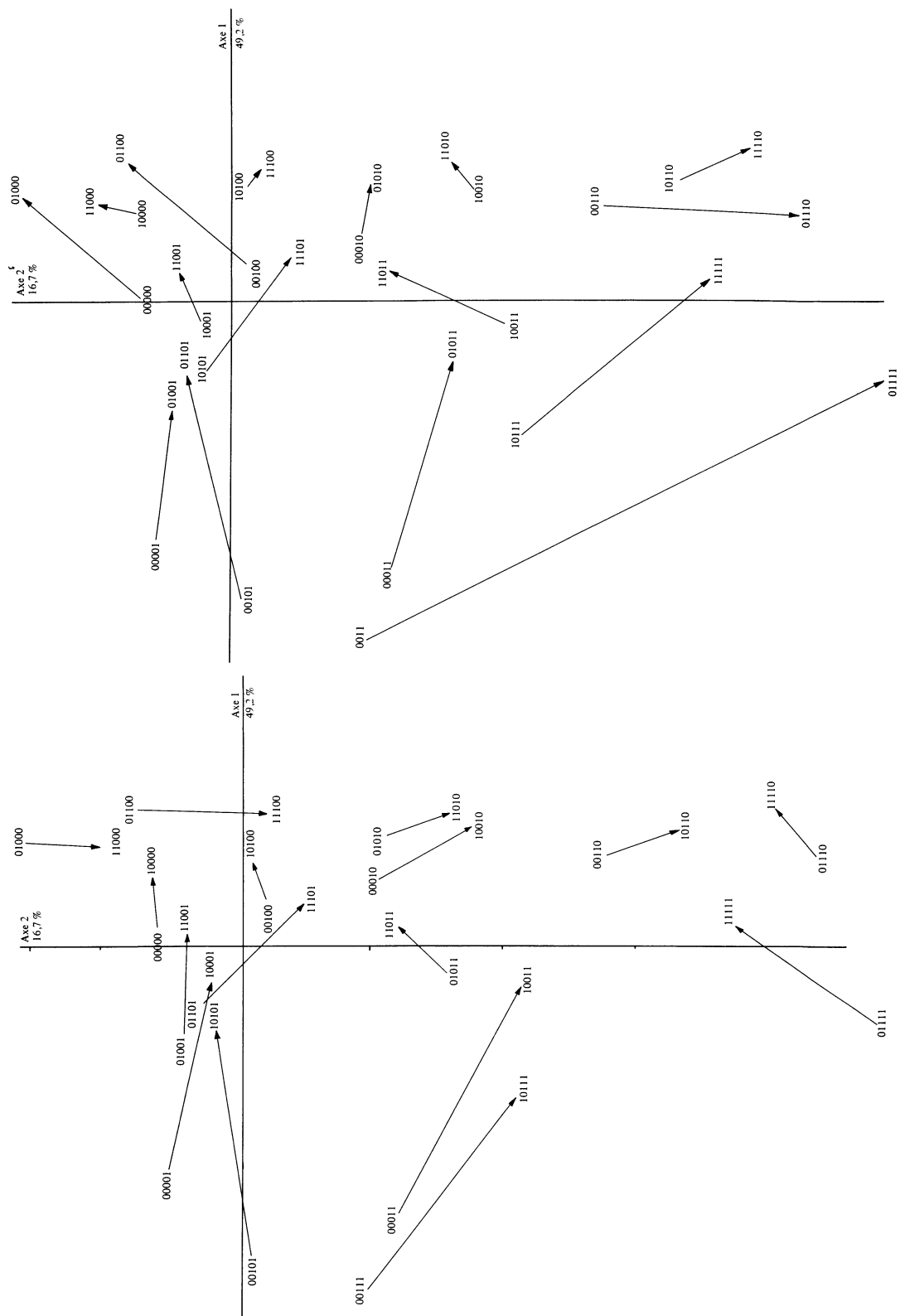
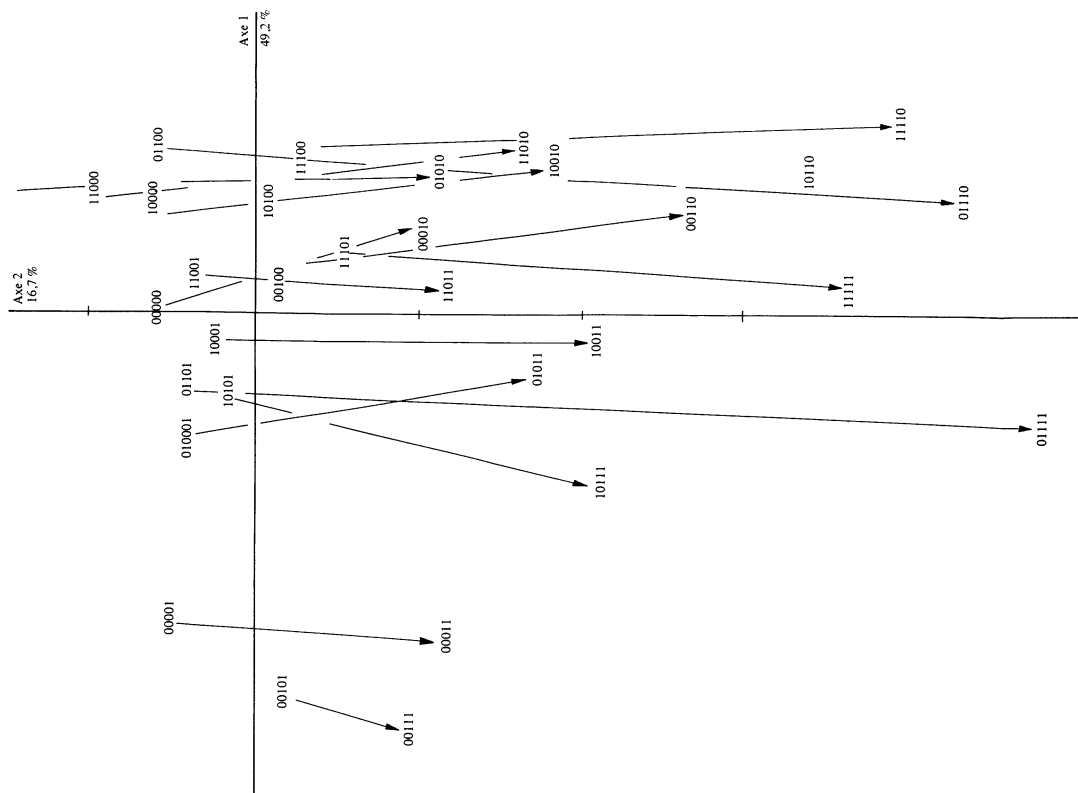
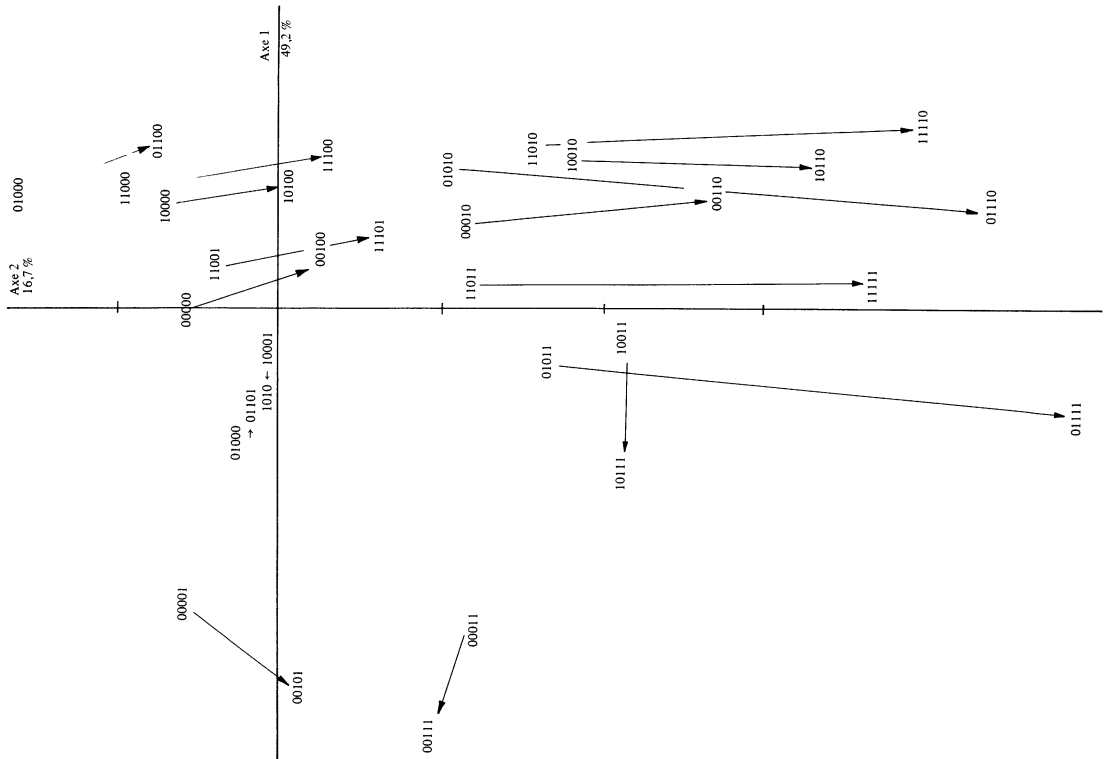


FIGURE 2 Representation des modalités supplémentaires dans le plan principal





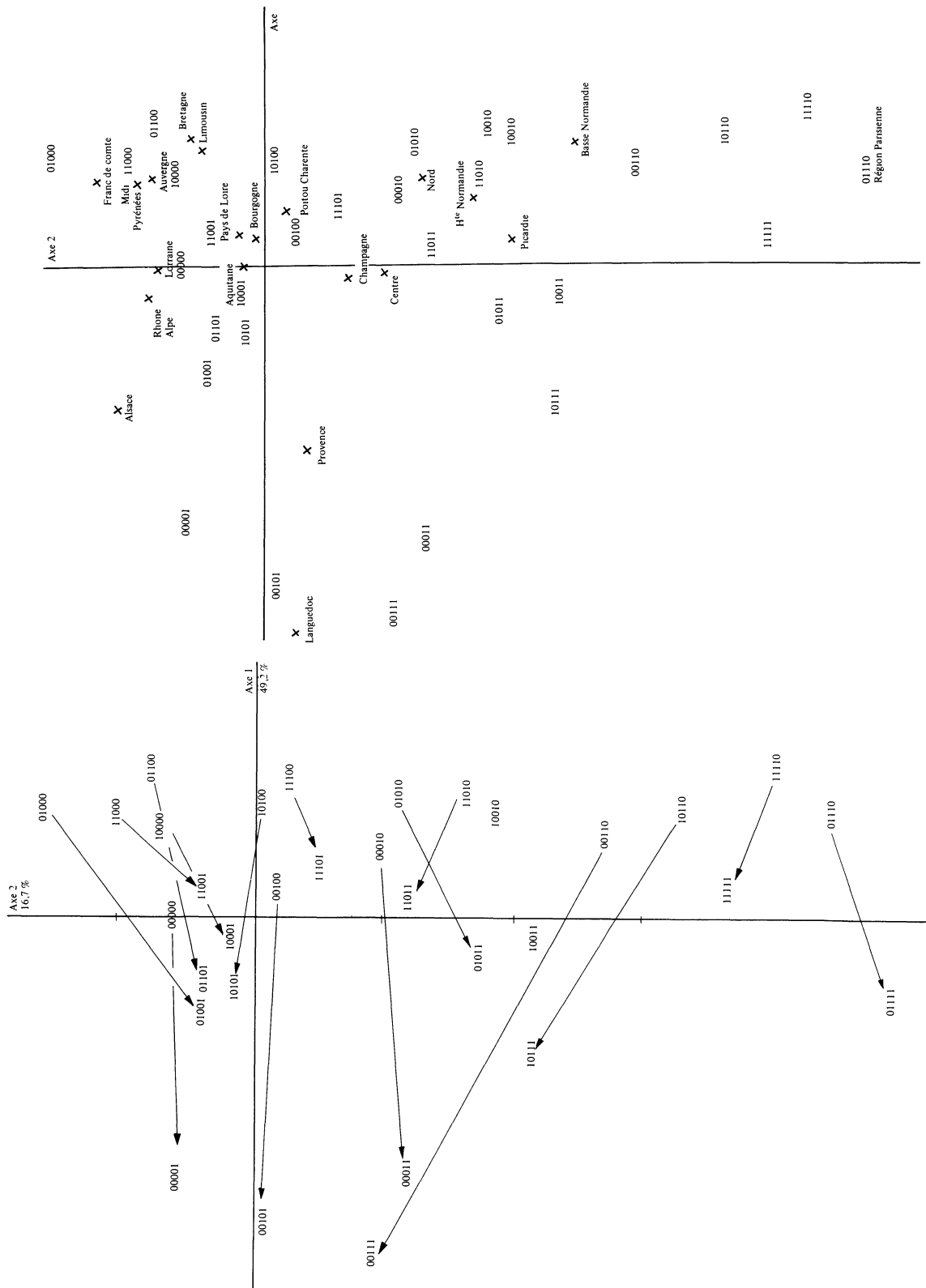


FIGURE 8

FIGURE 7 Salaires saisonniers (absence - presence)

En ce qui concerne les régions, figure 8, elles s'ordonnent sur l'axe 1 selon l'emploi de main d'œuvre saisonnière. Il faut noter que le Languedoc et la Provence Côte d'Azur contribuent beaucoup à cet axe : leurs contributions absolues sont respectivement de 0,65 et 0,09. Les régions s'ordonnent sur l'axe 2 selon l'emploi de main d'œuvre salariée à temps complet, la Région Parisienne (contribution de 0,18) et la Basse Normandie (0,27) étant celles qui y font le plus appel.

3. ANALYSE PAR MODELE LOG-LINEAIRE

3.1. Introduction

Il n'est pas possible dans le cadre de cet article d'exposer les propriétés du modèle log-linéaire. Nous donnons ici très brièvement celles qui sont nécessaires à la compréhension de l'analyse de l'exemple.

On considère les données sous la forme d'une table de contingence multidimensionnelle à 6 entrées, ayant $2^5 \times 21$ cases.

Soit n_{ijklmp} l'effectif de la case i, j, k, l, m, p , l'indice i correspondant à la variable 1, l'indice j à la variable 2, ... l'indice p à la variable 6 (région).

Un modèle log linéaire sur cette table s'écrit

$$\log (E(n_{ijklmp})) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \theta_m + \xi_p + (\alpha\beta)_{ij} + \dots \text{etc.}$$

Nous considérons un modèle où $E(n_{ijklmp}) > 0$ pour tout i, j, k, l, m, p . C'est à dire qu'aucune case de la table de contingence n'est structurellement vide ; comme c'est le cas dans notre exemple. Parmi les modèles possibles on s'intéressera particulièrement aux modèles suivants :

$$\log (E(n_{ijklmp})) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \theta_m + \epsilon_p \quad (1)$$

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_k \gamma_k = \sum_l \delta_l = \sum_m \theta_m = \sum_p \epsilon_p = 0$$

$$\sum_{ijklm} \exp (\mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \theta_m + \xi_p) = n$$

ou n est l'effectif total de la table de contingence. Cette dernière contrainte est nécessitée par le fait que l'effectif total est fixe à n .

$$\begin{aligned} \log (E(n_{ijklmp})) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \theta_m + \xi_p + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} \\ + \dots + (\theta\xi)_{mp} \end{aligned} \quad (2)$$

avec des contraintes analogues au cas précédent.

Il est facile de vérifier que le modèle (1) est celui de l'indépendance mutuelle : en effet (1) implique

$$E(n_{ijklmp}) = e^\mu \cdot e^{\alpha_i} \cdot e^{\beta_j} \cdot e^{\gamma_k} \cdot e^{\delta_l} \cdot e^{\theta_m} \cdot e^{\xi_p}$$

Dans le modèle (2) on autorise des dépendances directes 2 à 2 entre les variables. Si dans le modèle (2) on supprime le paramètre $(\alpha\beta)$ on obtient un nouveau modèle qui implique l'indépendance des variables 1 et 2 conditionnellement aux variables 3, 4, 5 et 6 ; autrement dit supprimer $(\alpha\beta)$ revient à poser un modèle où la liaison entre ces 2 variables provient des autres variables.

On considère également des modèles comportant les interactions d'ordre 3, comme par exemple $(\alpha\beta\gamma)_{ijk}$, ou les dépendances 3 à 3 entre les variables sont admises et des modèles comportant des interactions d'ordre 4.

Estimation des paramètres

Les paramètres sont estimés par la méthode du maximum de vraisemblance. Le programme ECTA que nous avons utilisé est basé sur une méthode itérative de maximisation appelé Itérative scaling et qui procède de la façon suivante :

On considère le modèle (2). On définit les marginales d'ordre 2 de la table de contingence, $n_{ij...}$, $n_{ik...}$, $n_{...mp}$ par les formules :

$$n_{ij...} = \sum_{klmp} n_{ijklmp}, n_{ik...} = \sum_{jlm p} n_{ijklmp} \text{ etc.}$$

$n_{ij...}$ est donc le nombre d'individus de l'échantillon appartenant à la modalité i de la 1^{ère} variable et j de la 2^{ème}.

Les marginales d'ordre 2 sont les seuls éléments de la table de contingence dont nous ayons besoin pour estimer les paramètres du modèle (2).

Soit $m_{ijklmp}^{(s)}$, la valeur obtenue à la s ème itération. On obtient l'itération suivante par la formule.

$$m_{ijklmp}^{(s+1)} = m_{ijklmp}^{(s)} \cdot \frac{n_{ij...}}{m_{ij...}^{(s)}} \quad \text{où} \quad m_{ij...}^{(s)} = \sum_{klmp} m_{ijklmp}^{(s)}$$

et l'on applique successivement le même type de formule aux 14 autres marginales d'ordre 2.

Ce qui précède se généralise à n'importe quel modèle autre que le modèle (2).

En particulier pour estimer les paramètres d'un modèle comportant des interactions d'ordre quelconque on a besoin des tables marginales du même ordre. Par exemple si le modèle comporte l'interaction $(\alpha\beta\gamma\delta)$, on a besoin de la table $(n_{ijkl...})$ où

$$n_{ijkl...} = \sum_{np} n_{ijklmp}$$

Les estimateurs ainsi obtenus sont asymptotiquement efficaces et de loi normale. Le programme ECTA calcule leur variance asymptotique. Si $\hat{\alpha}_i$ est l'estimateur du paramètre α_i le rapport $\hat{\alpha}_i/\sqrt{v_i}$ où v_i est l'estimateur de la variance asymptotique de $\hat{\alpha}_i$, est un indicateur de la signification du paramètre α_i . Ce rapport est à comparer aux valeurs de la loi normale 1.96 ou 3 par exemple ce qui permet de juger grossièrement l'importance de ce paramètre.

Un test rigoureux de l'hypothèse $\alpha_i = 0$ pour tout i est basé sur la statistique de χ^2 décrite au paragraphe suivant.

A chaque modèle correspond la statistique d'ajustement aux données qui, sous l'hypothèse que le modèle choisi est le bon suit asymptotiquement une loi de χ^2 . On peut mesurer l'importance d'un effet ou d'une interaction particulière en considérant la différence des deux statistiques d'ajustement suivantes : celle obtenue par un modèle contenant l'effet ou l'interaction en question, et celle du même modèle d'où on a supprimé cet effet ou cette interaction.

Si ce dernier modèle est le vrai modèle, la statistique mesurant l'importance de l'effet ou de l'interaction suit asymptotiquement une loi de χ^2 avec le nombre de degrés de libertés correspondant au nombre de niveaux de l'effet ou au produit des nombres de niveaux (moins un) des indices de l'interaction.

Etant donné la taille importante de l'échantillon la plupart des interactions sont significatives. Plutôt que de chercher un modèle particulier s'ajustant bien aux données et d'estimer les paramètres, nous nous sommes intéressés davantage à la décomposition des statistiques d'ajustement, aux ordres de grandeurs des effets et des interactions, de façon à obtenir une vue globale des liaisons entre les six variables. Notre optique est donc ici proche de l'analyse de l'information menée par Kullback [4]. Pour un exposé détaillé sur le modèle log-linéaire, on se reportera à Bishop, Fienberg et Holland [1].

Nous donnons ici quelques éléments de la méthode d'analyse de l'information proposée par S Kullback. Pour plus de détails on se reportera à Ku et Kullback [3].

On considère deux systèmes de probabilités Π et p d'une table de contingence multidimensionnelle. Kullback a défini une mesure de "l'écart" de Π à p :

$$I(\Pi : p) = \sum_i \Pi_i \log \frac{\Pi_i}{p_i}$$

où i est ici considéré comme un multi index : $i = (i_1, i_2, \dots, i_p)$.

Soit H_1 et H_2 2 hypothèses concernant les probabilités de la table de contingence et $\hat{\Pi}$ et $\hat{p}$ les estimateurs de ces probabilités sous respectivement l'hypothèse H_1 et H_2 . On calcule la statistique de discrimination de H_1 contre H_2 :

$$2nI(\hat{\Pi} : \hat{p}) = n \sum_i \hat{\Pi}_i \log \frac{\hat{\Pi}_i}{\hat{p}_i}$$

où n est la taille de l'échantillon observé, i.e. la somme des termes de la table de contingence étudiée.

Par exemple si H_2 est l'hypothèse relative au modèle le plus général et H_1 est l'hypothèse d'indépendance mutuelle des variables, $2nI(\hat{\Pi} : \hat{p})$ est la statistique qui permet de mesurer l'écart des observations à l'indépendance.

Remarquons à ce propos que : Modèle log linéaire et mesure de l'information pour des variables qualitatives sont étroitement liés, voir [4].

Enfin notons que si H_1 est vraie, $2nI(\hat{\Pi} : \hat{p})$ est distribué asymptotiquement selon une loi de χ^2 pourvu que $\hat{\Pi}$ et $\hat{p}$ soient estimés par la méthode du maximum de vraisemblance. Dans ce dernier cas la quantité $2nI(\hat{\Pi} : \hat{p})$ est appelée mdis (minimum discriminant information statistics).

La table de contingence comporte un certain nombre de cases vides (aléatoires) notamment pour les régions de faible effectif (Alsace, Franche-Comté).

Pour ces deux régions des marginales d'ordre 3 sont nulles. On ne peut donc pas ajuster certains modèles comportant des interactions d'ordre 4 (voir [1]). Nous avons donc ajouté 1/2 à chaque case pour remédier à cette situation. Cette modification contribue à diminuer quelque peu les statistiques d'ajustement, notamment pour les interactions d'ordre élevé.

3.2. Ajustement de 4 modèles

La table 2 donne les résultats obtenus pour 4 modèles : celui contenant toutes les interactions d'ordre 4, celui contenant toutes les interactions d'ordre 3, celui contenant toutes les interactions d'ordre 2, celui contenant tous les effets principaux.

TABLE 2

<i>Modèle</i>	<i>Statistique d'Ajustement Information discriminante</i>	<i>dl</i>	<i>Probabilité⁽¹⁾</i>
Interactions d'ordre 4	91	121	> 1/2
Interactions d'ordre 3	366	326	0,06
Interactions d'ordre 2	1 463	536	0,00
Effets principaux	9 037	646	0,00

Il faut noter que le dernier modèle est celui de l'indépendance globale entre les 6 variables. D'autre part le modèle complet contenant toutes les interactions jusqu'à l'ordre 6 a une statistique d'ajustement nulle.

En effet la statistique d'ajustement d'un modèle donné vaut $2nI(\hat{\Pi}, f)$, où $\hat{\Pi}$ est l'estimation des probabilités sous ce modèle et f le vecteur des fréquences observées, c'est-à-dire que :

$$f_i = \frac{n_i}{n}, \quad i = (i_1, i_2, \dots, i_p) \in I_1 \times I_2 \dots I_p$$

Or il est facile de voir que si le modèle est complet i.e. si le vecteur Π n'est tenu de satisfaire qu'aux seules restrictions :

$$\Pi_i \in]0,1[\quad \text{et} \quad \sum_i \Pi_i = 1$$

alors $\hat{\Pi} = f$ et donc $2nI(\hat{\Pi}, f) = 0$

Plus on admet d'interactions dans le modèle, plus celui-ci devient acceptable. On observe que le modèle contenant les interactions d'ordre 3 est acceptable au niveau 5 %. En revanche les interactions d'ordre 2 permettent de diminuer considérablement la statistique d'ajustement, beaucoup plus que les interactions d'ordre 3, ces dernières étant à leur tour plus importantes que les statistiques d'ordre 4.

(1) Il s'agit de la probabilité qu'une loi de $X^2(dl)$ dépasse la valeur calculée de la statistique d'ajustement.

Cette particularité des données est confirmée par l'analyse des interactions une à une. Elle est très importante, car pratiquement elle permet de s'en tenir à l'étude des interactions d'ordre 2 qui prennent compte $(9\ 037-1\ 463)/9\ 037 = 84\ %$ de la mesure de la liaison entre les six variables.

3.3. Etude des interactions d'ordre 4

Si on considère le modèle contenant les interactions d'ordre 3 comme modèle de référence et que l'on ajoute à tour de rôle chaque interaction d'ordre 4, on obtient par différence, l'importance de l'apport de chaque interaction ; les résultats sont résumés par la table 3 :

TABLE 3

<i>Interaction</i>	<i>dl</i>	<i>mdis</i>	<i>Probabilité</i>
1 246	20	38	0,01
2 456	20	36	0,01
1 256	20	32	0,02
1 346	20	32	0,02
2 356	20	32	0,02
2 346	20	29	0,03
1 236	20	28	0,04
1 356	20	21	0,20
3 456	20	21	0,20
1 456	20	16	> 1/2

Les interactions d'ordre 4 ne contenant pas le facteur 6 ne sont pas significatives.

3.4. Etude des interactions d'ordre 3

Prenant toujours le modèle contenant les interactions d'ordre 3 comme référence, et considérant les modèles obtenus en retirant tour à tour une interaction d'ordre 3, on obtient la table 4.

TABLE 4

<i>Interactions à 20 dl⁽¹⁾</i>		<i>Interactions à 1 dl</i>	
<i>Interaction</i>	<i>mdis</i>	<i>Interaction</i>	<i>mdis</i>
126	89	123	204
136	81	124	70
146	64	125	9
156	82	134	6
236	121	135	2
246	35	145	2
256	54	234	44
346	36	235	4
356	45	245	1
456	78	345	5

(1) Les degrés de liberté se calculent comme en analyse de variance : $(21 - 1)(2 - 1)(2 - 1) = 20$ pour les interactions contenant la région $(2 - 1)(2 - 1)(2 - 1) = 1$ pour les autres interactions. Les seuils au niveau 1 % sont les suivants :

$$X^2(20) = 45,32 \quad \text{et} \quad X^2(1) = 10,83$$

Toutes ces interactions sont très significatives sauf 145, 135 et 245. Il faut noter l'importance de 123, 234, 124, 236. Les autres sont du même ordre de grandeur.

3.5. Etude des interactions d'ordre 2

On prend le modèle contenant toutes les interactions d'ordre 2 comme référence. On retire à tour de rôle chaque interaction d'ordre 2. On obtient la table 5.

TABLE 5

<i>Interactions à 20 dl⁽¹⁾</i>		<i>Interactions à 1 dl⁽¹⁾</i>	
Interaction	mdis	Interaction	mdis
16	683	12	2 205
26	298	13	618
36	280	14	161
46	541	15	7
56	1 346	23	948
		24	24
		25	47
		34	99
		35	93
		45	122

(1) Voir la note 1 relative à la table 4.

Toutes ces interactions sont très largement significatives. De plus en comparant les ordres de grandeurs des "mdis" des tables 3, 4 et 5 on constate que les interactions d'ordre 4 sont négligeables devant les interactions d'ordre 3 et que celles-ci sont à leur tour peu importantes (sauf 123, 124 et 236) devant les interactions d'ordre 2. Nous allons donc nous limiter à étudier ces dernières.

Il est important de caractériser la nullité d'une interaction d'ordre 2 : On considère par exemple le modèle contenant toutes les interactions d'ordre 2 sauf une, par exemple l'interaction 12 (Main d'œuvre familiale à temps complet chef d'exploitation). On peut montrer (voir [1]) que ce modèle implique l'indépendance entre les variables 1 et 2 conditionnellement aux autres variables (3, 4, 5 et 6) ; Par suite la mesure de l'apport de l'interaction 12 (mdis de 2 205) est une mesure de la liaison partielle entre les variables 1 et 2 conditionnellement aux autres variables. On observe donc une liaison partielle très forte entre les variables 1 et 2 : toutes choses égales par ailleurs chef d'exploitation à temps partiel est très fortement lié à la présence de main d'œuvre familiale à temps complet.

On peut résumer la table 5 par le graphique suivant : (fig. 9).

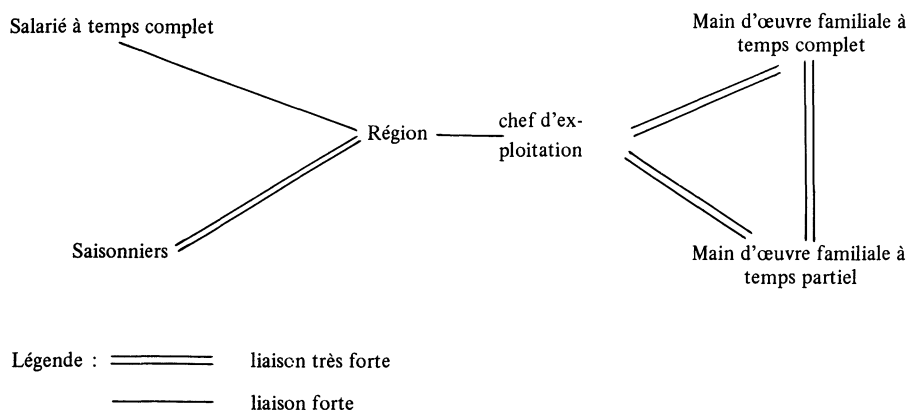


FIGURE 9. — Schéma représentant globalement les liaisons entre les différentes variables et leur intensité⁽¹⁾.

3.6. Analyse des paramètres

On se limite aux interactions simples. Leurs estimations sont données dans les tables 6 et 7.

Dans la Table 6 sont indiquées les interactions régions x type de main d'œuvre. Dans la table 7 sont inscrites les interactions entre les variables de type de main d'œuvre. Les effets principaux sont sur la diagonale de la table 7. Cette dernière est triangulaire car chaque interaction n'est indiquée qu'une fois. D'autre part les variables de main d'œuvre étant binaires nous n'avons indiqué pour ces dernières qu'un seul paramètre, l'autre lui étant égal et de signe opposé.

Par exemple l'interaction $(\beta\zeta)_{11}$ entre la région parisienne et l'absence de main d'œuvre familiale à temps complet vaut $-0,90$. Par suite l'interaction $(\beta\zeta)_{21}$ entre la région parisienne et la présence de main d'œuvre familiale à temps complet vaut $0,90$. Ces deux interactions sont largement significative puisque le rapport du paramètre à son écart type vaut $8,93$.

L'estimation de l'interaction chef d'exploitation à temps partiel absence de main d'œuvre familiale à temps complet vaut $0,36$:

$$\begin{aligned}
 &(\hat{\alpha\beta})_{11} = 0,36 \\
 \text{d'où} &(\hat{\alpha\beta})_{12} = -0,36 \\
 &(\hat{\alpha\beta})_{21} = -0,36 \\
 \text{et} &(\hat{\alpha\beta})_{22} = -0,36
 \end{aligned}$$

Cette interaction est très importante comme on peut le mesurer sur le rapport du paramètre à son écart-type : $19,45$.

(1) Les variables qui ne sont pas reliées par un trait simple ou double ne sont pas pour autant indépendantes. Cependant leur liaison est faible en comparaison avec les liaisons mises en évidence dans la figure 9.

TABLE 6
Interactions Régions x Type de main d'œuvre

Région	Chef d'exploitation temps partiel $\mu/\sigma^{(1)}$		Main d'œuvre salariée				Main d'œuvre familiale			
			Temps complet (absence) $\mu/\sigma^{(1)}$		saisonnier (absence) $\mu/\sigma^{(1)}$		temps complet (absence) $\mu/\sigma^{(1)}$		temps partiel (absence) $\mu/\sigma^{(1)}$	
Région parisienne	0,102	-1,017	-0,90	-8,93	0,12	1,18	-0,16	-1,56	-0,16	1,60
Champagne	0,010	0,128	-0,20	-2,61	-0,17	-2,27	-0,15	-1,96	0,11	1,47
Picardie	0,075	-0,999	-0,34	-4,53	-0,13	-1,76	-0,06	-0,76	-0,06	-0,82
Hte Normandie	-0,081	-0,857	-0,32	-3,42	0,19	2,00	0,10	1,03	-0,25	-2,62
Centre	-0,181	-2,427	-0,04	- .61	-0,08	-1,05	0,14	1,91	-0,12	-1,56
B. Normandie	-0,101	-1,601	-0,34	-5,39	0,41	6,54	0,01	0,15	-0,34	-5,41
Bourgogne	-0,079	-1,044	-0,02	- .27	-0,01	0,18	0,08	1,00	0,24	3,23
Nord	-0,148	-2,273	-0,22	-3,34	-0,04	0,68	-0,28	-4,24	-0,04	-0,61
Lorraine	0,148	1,482	0,28	2,77	0,04	0,40	-0,18	-1,77	0,09	0,86
Alsace	0,126	0,984	0,27	2,08	-0,20	-1,54	-0,09	-0,69	0,23	1,82
Franche Comté	-0,034	-0,233	0,412	2,83	0,37	2,51	0,09	-0,59	0,01	0,05
Pays de Loire	-0,092	-1,436	0,07	1,15	0,17	2,66	0,22	3,46	0,03	0,49
Bretagne	-0,144	-2,394	0,28	4,60	0,28	4,64	-0,21	-3,51	0,00	0,00
Poitou - Charentes	0,163	-2,348	-0,07	1,07	0,04	.65	0,16	2,33	0,09	1,26
Aquitaine	0,137	2,567	0,14	2,71	-0,23	4,30	-0,03	-0,52	-0,13	-2,43
Midi Pyrénées	0,206	3,872	0,21	3,93	0,16	3,07	-0,08	-1,51	0,06	1,18
Limousin	-0,122	-1,418	0,12	1,36	0,30	3,43	-0,12	-1,42	-0,04	0,44
Rhône-Alpes	0,056	0,862	0,38	5,89	-0,07	-1,16	0,10	1,57	0,04	0,62
Auvergne	-0,029	-0,344	0,28	3,36	0,19	2,29	0,01	-0,10	0,29	3,44
Languedoc	0,448	4,833	0,00	0,02	-0,92	-9,95	0,52	5,56	0,03	0,33
Provence - Côte d'Azur	0,221	3,353	0,01	0,15	-0,41	-6,17	0,10	1,51	-0,09	-1,35

(1) Dans cette colonne est indiqué le rapport du paramètre estimé à sa variance asymptotique. Cette valeur est à comparer aux valeurs classiques de la loi normale 1,96 ou 3.

TABLE 7
Interactions et effets principaux des types de main d'œuvre

Variable Variable	1 (Temps partiel) $\mu/\sigma^{(1)}$		2 (absence) $\mu/\sigma^{(1)}$		3 (absence) $\mu/\sigma^{(1)}$		4 (absence) $\mu/\sigma^{(1)}$		5 (absence) $\mu/\sigma^{(1)}$	
1 (tp)	-0,41	-21,72								
2 (abs)	0,36	19,45	0,50	26,74						
3 (abs)	0,16	8,39	0,14	7,63	23	12,25				
4 (abs)	0,09	4,57	0,01	0,56	11	6,13	0,83	44,54		
5 (abs)	0,02	0,83	0,04	2,37	0,05	2,71	0,11	6,16	0,18	9,87

(1) Voir note 1, table 6

A l'aide des paramètres, on obtient la représentation graphique des régions (fig. 10 et fig. 11).

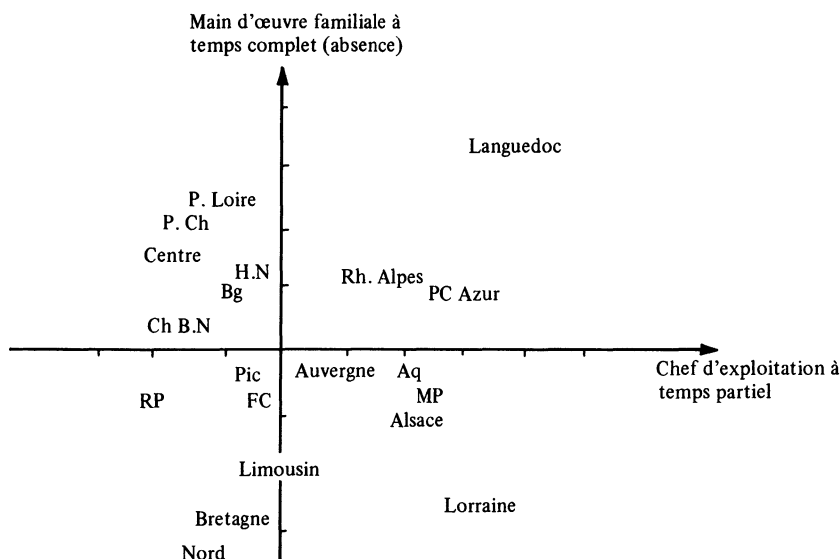
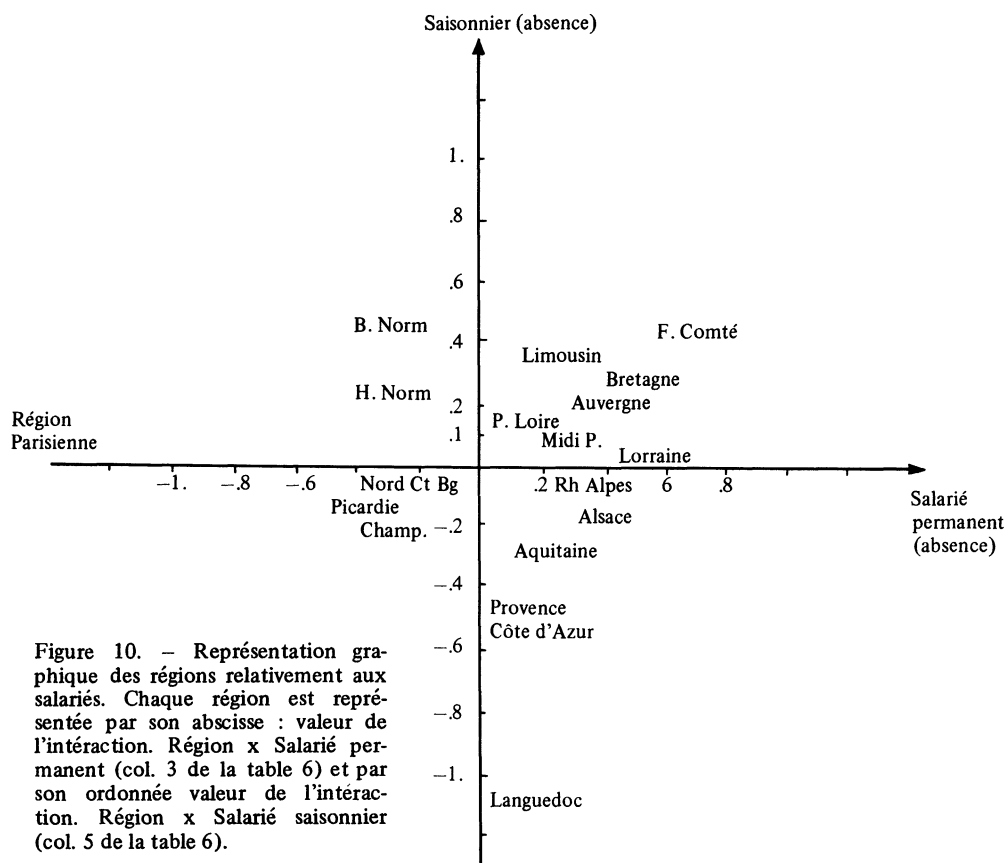


Figure 11. — Représentation graphique des régions relativement au statut du chef d'exploitation et à l'importance de la main d'œuvre familiale à temps complet.

En abscisse : interaction Région x chef d'exploitation (col. 1 de la table 6)

En ordonnée : interaction Région x main d'œuvre familiale (col. 7 table 6).

La figure 10 est presque identique au plan principal donné par l'analyse factorielle des correspondances. Cependant le premier axe de la figure 10 est uniquement celui des salariés saisonniers à la différence de celui de l'analyse factorielle qui est un axe composite.

Remarquons que comme en analyse factorielle, il est possible de donner le pourcentage de la liaison totale représentée par un graphique comme la figure 10 :

La première idée qui vient à l'esprit est de considérer les statistiques associées à chaque interaction dans la table 5 et de les additionner. Par exemple si on ajoute les statistiques relatives aux interactions Région x saisonnier et Région x salarié à temps complet, on obtient $1\,346 + 541 = 1\,887$. D'autre part si on considère le modèle d'indépendance entre les régions et type de main d'œuvre, on obtient une statistique d'ajustement qui vaut 4 711. Par suite on pourrait utiliser le rapport $1\,887/4\,711$ pour indiquer le pourcentage de liaison représentée sur la figure 10.

Cette façon de procéder suppose implicitement que les statistiques sont additives ce qui n'est pas le cas, comme en analyse de variance non orthogonale où les sommes de carrés ajustées ne sont pas additives, sauf si on considère une suite de modèles emboîtés.

On considère alors les modèles de la table 8 où un modèle est caractérisé par des marginales ajustées : par exemple le modèle 1 qui s'ajuste aux marginales $(n_{ijklm..})$ et $(n_{...p})$ s'écrit :

$$\log E(n_{ijklmp}) = \alpha_{ijklm-} + \beta_p$$

c'est le modèle d'indépendance Région x Type de main d'œuvre

TABLE 8
Table d'analyse de modèles emboîtés

	Modèle	Statistique d'ajustement	dl
1	12345,6	4 711	620
2	12345,56	3 247	600
3	12345,56,46	2 634	580
4	12345,56,46,16	1 668	560
5	12345,56,46,16,26	1 352	540
6	12345,56,46,26,36	1 081	520
7	123456	0	0

On voit sur la table que la variable emploi de saisonnier permet de prendre en compte $(4\,711 - 3\,247)/4\,711 = 31\%$ de la liaison Région x type de main d'œuvre, que les 2 variables d'emploi de salariés permettent de prendre en compte $(4\,711 - 2\,634)/4\,711 = 44\%$ de la liaison. On peut donc dire que la figure 10 prend en compte 44 % de la liaison. Si on cumule les informations des figures 10 et 11 on obtient $(4\,711 - 1\,352)/4\,711 = 71\%$ de la liaison. Enfin si on ne tient compte que des liaisons simples (sans considérer les interactions d'ordre supérieur) on obtient $(4\,711 - 1\,081)/4\,711 = 77\%$ de la liaison globale. Les liaisons entre 2 variables de main d'œuvre (ou plus) de la variable Région représentent donc 23 %

de la liaison totale quand on a déjà pris en compte les liaisons simples. La table 9 résume ces réflexions, en attribuant à chaque variable de main d'œuvre la part de liaison qui lui revient :

TABLE 9
Décomposition de la liaison Région x Type de main d'œuvre

Variable	% de liaison
Saisonnier	31 %
Salarié	13 %
Statut du chef d'exploitation	21 %
Main d'œuvre familiale à temps complet	6 %
Main d'œuvre familiale à temps partiel	6 %
Intéractions	23 %
Total	100 %

Remarque : l'ordre d'introduction des variables dans les tables 8 et 9 est important.

L'apport de chaque variable est mesuré en tenant compte des variables déjà mentionnées dans la table. Un changement de l'ordre entraîne donc un changement des résultats dans les tables 8 et 9.

Pour traiter cet exemple, nous avons utilisé le programme ECTA de l'université de Chicago. Ce programme fonctionne actuellement sur l'IRIS 80 du Ministère de l'Agriculture. A titre d'exemple l'ajustement d'un modèle comportant les interactions d'ordre 4 sur l'exemple étudié prend environ 10 secondes CPU.

Pour obtenir des précisions sur ce programme le lecteur intéressé pourra s'adresser aux auteurs de l'article. Signalons que Collombier (Université P. Sabatier-Toulouse) dispose d'une version modifiée d'ECTA et que le laboratoire de Biométrie du CNRA Versailles utilise un programme appelé ZAN.

4. COMPARAISON DES DEUX METHODES

4.1. Etude des liaisons globales

Si on cherche une vue globale des liaisons entre les 6 variables concernées il est indéniable que le modèle log-linéaire est la méthode la mieux adaptée. Il permet d'obtenir la figure 9 qui donne un résumé agréable de l'ensemble des liaisons concernées. Il est clair que l'analyse factorielle des correspondances sur la table 1 ne peut donner d'information sur les liaisons entre les seules variables de main d'œuvre. Il est frappant de constater que les très fortes associations entre les variables 1, 2, 3 sont indécélables sur la figure 1. Pourtant leur intensité, me-

surée par le mdis, est supérieure à celle des liaisons entre variables de main d'œuvre et région, pour un nombre de degrés de liberté 20 fois plus faible ! Ce phénomène a des raisons évidentes : l'analyse factorielle sur la table 1 ne permet d'analyser que la liaison entre variables de main d'œuvre d'une part et régions d'autre part. Les proximités entre variables de main d'œuvre sur la figure 1 sont calculées sur les régions : une faible distance entre 2 modalités de type de main d'œuvre signifie seulement que ces deux modalités varient de la même façon d'une région à l'autre, et non qu'elles sont liées marginalement. Le fait de considérer les profils, et donc de diviser par les marges de la table 1, élimine les liaisons marginales. On peut évidemment compléter l'analyse factorielle des correspondances de la table 1 par des analyses des liaisons marginales entre les variables de main d'œuvre, mais on ne voit pas bien comment regrouper les résultats de ces différentes analyses en un résumé synthétique semblable à celui donné par le modèle log-linéaire.

4.2. Etude de la liaison type de main d'œuvre x région

Les deux méthodes donnent des résultats assez semblables : l'importance du facteur saisonnier, du type de chef d'exploitation et de la main d'œuvre salariée à temps complet sont mis en évidence par les 2 analyses, à ceci près que l'analyse factorielle confond les 2 premiers éléments sur son premier axe, soulignant par là le fait qu'ils varient ensemble d'une région à une autre. Il nous semble cependant que le fait que le modèle log linéaire respecte la structure factorielle des données est utile pour l'interprétation : il est agréable en effet de considérer les effets principaux, puis les interactions de différents ordres au lieu de considérer l'ensemble des 32 modalités de main d'œuvre à la fois. Malgré la présence des modalités supplémentaires, l'interprétation du graphique 1 est un exercice assez délicat et le praticien habitué à l'analyse de variance aura moins de difficultés avec le modèle log-linéaire, qui se révèle ici être une bonne méthode de description des données.

BIBLIOGRAPHIE

- [1] Y.M. BISCHOP, S.E. FIENBERG, R.W. HOLLAND. — "Discrete Multivariate Analysis". MIT. Press (1975).
- [2] Statistique Agricole. — Supplément "Série Etudes" n° 116 (1973). Ministère de l'Agriculture.
- [3] H.H. KU, S. KULLBACK. — "Interaction in multidimensional contingency table. An information theoretic approach". *Journal of Research of the National Bureau of Standards*, 72 B, 3, 159 198, (1968).
- [4] S. KULLBACK. — Information Theory and Statistics. Dover Publications Inc. (1968).