REVUE DE STATISTIQUE APPLIQUÉE

P. CAZES

Régression par boule et par l'analyse des correspondances

Revue de statistique appliquée, tome 24, n° 4 (1976), p. 5-22

<http://www.numdam.org/item?id=RSA_1976__24_4_5_0>

RÉGRESSION PAR BOULE ET PAR L'ANALYSE DES CORRESPONDANCES

P. CAZES

Laboratoire de Statistique,

Université Pierre et Marie Curie (Paris VI)

I – INTRODUCTION

En présence de variables hétérogènes (quantitatives, semi-quantitatives, qualitatives) il est possible pour prendre en compte l'information apportée par toutes ces variables, soit de coder les variables qualitatives, codage qui dépend du but poursuivi, soit de rendre qualitatives toutes les variables par découpage en classes des variables qui ne sont pas déjà qualitatives. Dans ce dernier cas, on peut si l'on veut décrire ces variables faire l'analyse factorielle du tableau disjonctif complet ainsi construit ou du tableau de contingence associé (tableau de BURT).

Pour mettre en évidence les liaisons d'une des variables y avec les autres variables, on peut faire l'analyse des correspondances du tableau obtenu en croisant les modalités de y avec celles des variables explicatives, tableau que nous appellerons tableau de régression.

Pour prévoir les valeurs de y, on peut dans l'espace des premiers facteurs de l'analyse précédente, où l'on a projeté chaque observation, soit faire une régression usuelle, soit faire une régression par boule.

C'est dans une optique de régression que nous nous plaçons ici. Nous définissons en II les notations utilisées, tandis qu'en III nous donnons quelques propriétés des facteurs du tableau de BURT et du tableau de régression. Nous montrons en particulier que les facteurs sur les individus rajoutés en supplémentaire au tableau de régression sont centrés, et que dans le cas où les variables explicatives sont indépendantes 2 à 2, ils sont non corrélés. Après avoir rappelé en IV les principes de la régression par boule, un exemple d'application est présenté en V. Cet exemple concerne un problème géologique où il s'agissait d'expliquer la teneur en matière organique (ou kérogène) d'un certain nombre de roches de l'Oxfordien du bassin de Paris en fonction d'un certain nombre de caractéristiques quantitatives comme la teneur en calcaire, ou la granulométrie, ou qualitatives comme la couleur de la roche par exemple.

Dans ce paragraphe, nous comparons les résultats fournis par différentes régressions par boule, en fonction de l'espace où s'effectue cette régression.

Manuscrit remis le 5-4-76, révisé en septembre 1976.

Mots-clés : **Analyse factorielle des correspondances, corrélation, facteur, liaisons, régression.**

II – LES NOTATIONS

Soient $y = x_0, x_1, x_2, \ldots, x_r$, l'ensemble des variables, y étant à expliquer en fonction de $x_1, x_2, \ldots, x_r$.

Nous supposerons $y, x_1, \ldots, x_r$ qualitatives, ce qui revient à supposer que les variables quantitatives ont été rendues qualitatives par un découpage préalable en classes, et nous désignerons par J_q l'ensemble des modalités de la $q^{ième}$ variable $(0 \leqslant q \leqslant r)$

On posera :

$Q = \{0, 1, 2, \ldots, r\}$	ensemble des variables
$Q_e = \{1, 2, \ldots, r\}$	ensemble des variables explicatives
$J = U\{J_q \mid q \in Q\}$	ensemble des modalités de toutes les variables
$J_e = U\{J_q \mid q \in Q_e\}$	ensemble des modalités des variables explicatives
$I = \{1, 2, \ldots, n\}$	ensemble des n individus (ou observations) pour lesquels on a mesuré $y, x_1, \ldots, x_r$.

II.1 – Les tableaux (cf. figure 1)

a) k_{IJ} : tableau initial des données, qui est un tableau disjonctif complet :

$\forall i \in I, \forall j \in J_q \subset J : k_{IJ}(i, j) = 1$ si i a pris la modalité j de J_q, 0 sinon.

b) k_{JJ} : tableau de BURT associé à k_{IJ}, obtenu en croisant l'ensemble des modalités de toutes les variables avec lui-même :

$$\forall j, j' \in J, j \in J_q, j' \in J_{q'} :$$

$$k_{JJ}(j, j') = \Sigma \{k_{IJ}(i, j)\, k_{IJ}(i, j') \mid i \in I\} \qquad (0)$$

désigne le nombre d'individus (ou d'observations) possédant les modalités j et j' de J_q et $J_{q'}$. Si $q = q'$, $k_{JJ}(j, j') = o$ sauf si $j = j'$ auquel cas $k_{JJ}(j, j)$ désigne le nombre d'individus possédant la modalité j de J_q.

A ce tableau on peut adjoindre en supplémentaire le tableau k_{IJ}

c) $k_{J_0 J_e}$: sous tableau de k_{JJ} croisant J_0 ensemble des modalités de la variable à expliquer avec Je ensemble de toutes les modalités des variables explicatives. C'est le tableau de régression ou encore de dépendance (ou de liaison) de y avec $x_1, x_2, \ldots, x_r$.

A ce tableau on peut adjoindre en supplémentaire le tableau k_{IJ_e} (sous tableau de k_{IJ})

C'est en effectuant l'analyse factorielle des correspondances de ce tableau $k_{J_0 J_e}$ qu'on étudiera la liaison de y avec $x_1, x_2, \ldots, x_r$, la mise en supplémentaire du tableau k_{IJ_e} permettant de faire sur les facteurs associés aux n individus supplémentaires une régression usuelle ou une régression par boule.

d) $k_{J_q J_{q'}}$ sous tableau de k_{JJ} croisant J_q avec $J_{q'}$. La somme des éléments de ce tableau valant n, nous poserons également :

$$p_{J_q J_{q'}} = k_{J_q J_{q'}}/n$$

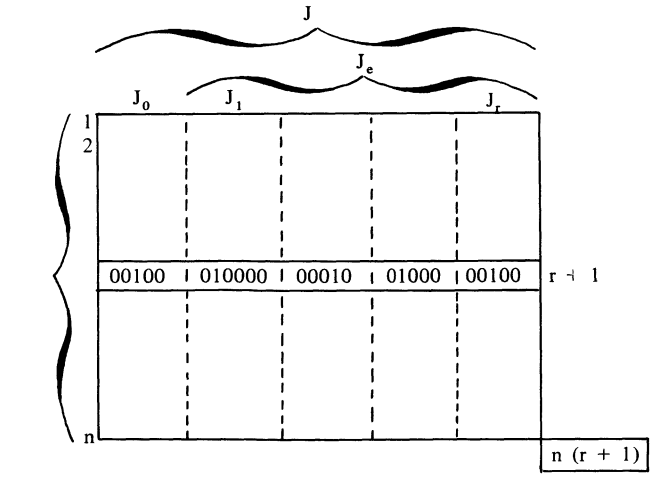

tableau disjonctif complet initial k_{IJ}

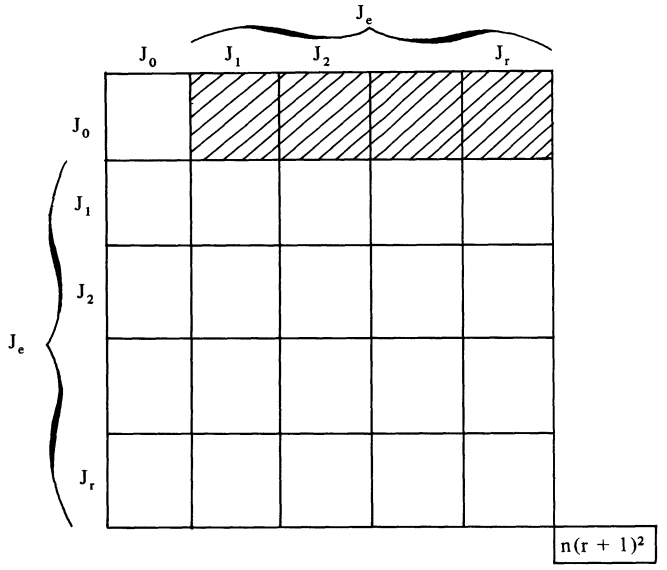

tableau de BURT k_{JJ} associé à k_{IJ}

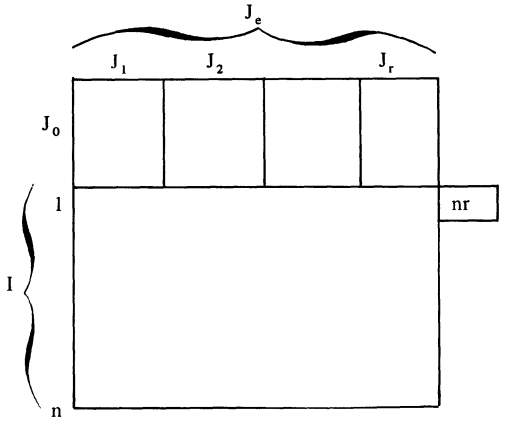

tableau "de régression" $k_{J_oJ_e}$ (sous tableau de k_{JJ}) avec en supplétaire le sous tableau k_{IJ_e} de k_{IJ}

Figure 1

II.2 – Caractéristiques associées aux tableaux précédents.

II.2.1. Lois marginales

Nous désignerons par p_{J_q} , $p_{J_{q'}}$ les lois marginales associées à $p_{J_q J_{q'}}$:

$$\forall\, j \in J_q : p_{J_q}(j) = \Sigma\{p_{J_q J_{q'}}(j, j') \mid j' \in J_{q'}\}$$

désigne la proportion des individus ayant pris la modalité j de J_q.

Les lois marginales associés aux différents tableaux introduits en II.1 (tableaux ramenés à avoir la somme de leurs éléments égale à 1 par division par une constante) se calculent aisément en fonction des p_{J_q} On obtient ainsi pour $k_{J_o J_e}$ les lois p_{J_o} sur J_o et $\frac{1}{r}(p_{J_1}, \ldots, p_{J_r})$ sur J_e, tandis que pour k_{JJ} et k_{IJ} la loi sur J est $\frac{1}{r+1}(p_{J_o}, \ldots, p_{J_r})$, la loi marginale p_I sur I de k_{IJ} étant la loi uniforme.

II.2.2. Lois conditionnelles

Nous désignerons par $p_{J_q}^{J_{q'}}$ et $p_{J_{q'}}^{J_q}$ les tableaux des lois conditionnelles associées à $p_{J_q J_{q'}}$. On a :

$$p_{J_q}^{J_{q'}} = \{p_{j'}^{j} \mid j \in J_q , j' \in J_{q'}\}$$

avec $\forall\, j \in J_q$, $\forall\, j' \in J_{q'}$:

$$p_{j'}^{j} = p_{J_q J_{q'}}(j, j')/p_{J_q}(j) = k_{JJ}(j, j')/k_{JJ}(j, j)$$

Notons que si $q = q'$, $p_{J_q J_q}$ est le tableau diagonal des $\{p_{J_q}(j) \mid j \in J_q\}$ tandis que $p_{J_q}^{J_q}$ est la matrice unité d'ordre card J_q.

Les lois conditionnelles associées aux tableaux k_{JJ} , $k_{J_o J_e}$ se calculent alors aisément d'après II.2.1 en fonction des $p_{J_{q'}}^{J_q}$.

III – ETUDE DES FACTEURS

III.1 – Propriétés et équations des facteurs

Rappelons quelques résultats classiques (cf. (1), (4), (6), (7)), relatifs aux facteurs issus de l'analyse des correspondances des tableaux k_{IJ} , k_{JJ} , $k_{J_o J_e}$.

a) k_{IJ} et k_{JJ} ont mêmes facteurs de variance 1 sur J, ce qui permet de déduire les résultats de l'analyse des correspondances de k_{IJ} à partir de ceux de k_{JJ}.

b) Tous les facteurs du tableau symétrique k_{JJ} sont directs (i. e. ont même valeur sur J que J soit considéré comme premier ensemble ou comme second ensemble).

c) Tout facteur sur J (resp. J_e) de k_{JJ} (resp. $k_{J_oJ_e}$) non trivial (i.e. non constant, et relatif à une valeur propre non nulle) est centré (pour p_{J_q}) sur chacun des sous ensembles J_q de J (resp. J_e).

d) Equation des facteurs :

Soient :

$$\varphi^J = \{\varphi^{J_q} \mid q \in Q\}$$

un facteur de k_{JJ} associé à la valeur propre λ,

$$(\psi^{J_o}, \psi^{J_e}) \text{ où } \psi^{J_e} = \{\psi^{J_q} \mid q \in Q_e\}$$

un couple de facteurs associés de $k_{J_oJ_e}$ correspondant à la valeur propre μ.

Ces facteurs vérifient les équations :

$$\forall\, q \in Q : \Sigma\, \{\varphi^{J_{q'}} \circ p^{J_q}_{J_{q'}} \mid q' \in Q - \{q\}\} = ((r+1)\sqrt{\lambda} - 1)\,\varphi^{J_q} \quad (1)$$

$$\left.\begin{aligned} \forall\, q \in Q_e : \quad \psi^{J_o} \circ p^{J_q}_{J_o} &= \sqrt{\mu}\,\psi^{J_q} \quad 2a \\ \Sigma\, \{\psi^{J_q} \circ p^{J_o}_{J_q} \mid q \in Q_e\} &= r\sqrt{\mu}\,\psi^{J_o} \quad 2b \end{aligned}\right\} \quad (2)$$

où une équation condensée telle que 2a par exemple signifie que :

$$\forall\, q \in Q_e, \forall\, j \in J_q : \Sigma\, \{\psi^{j_o} p^j_{j_o} \mid j_o \in J_o\} = \sqrt{\mu}\,\psi^j$$

De (2), l'on déduit :

$$\Sigma\, \{\psi^{J_o} \circ p^{J_q}_{J_o} \circ p^{J_o}_{J_q} \mid q \in Q_e\} = r\mu\,\psi^{J_o} \quad (3)$$

formule qui nous servira en III.2.

e) Dans le cas où les variables explicatives sont deux à deux indépendantes, on peut montrer à partir des équations précédentes (cf. annexe 1) que l'analyse du tableau k_{JJ} se ramène du moins en ce qui concerne les facteurs non triviaux à l'analyse du tableau $k_{J_oJ_e}$.

On montre de la même façon (cf. annexe 2) que si les variables explicatives sont divisées en deux groupes Q_1 et Q_2, et si y est indépendante des variables du groupe Q_2, l'analyse du tableau $k_{J_oJ_e}$ est équivalente à celle du tableau $k_{J_oJ^1_e}$ où

$$J^1_e = \cup\, \{J_q \mid q \in Q_1\}$$

désigne l'ensemble des modalités des variables de Q_1.

III.2 – Etude du cas particulier où il existe un facteur commun φ^{J_o} à toutes les correspondances $p_{J_oJ_q}$ $(q \in Q_e)$

Si $(\varphi^{J_o}, \varphi^{J_q})$ est un couple de facteurs de variance 1 de $p_{J_oJ_q}$, associé à la valeur propre λ_q, on a :

$$\forall\, q \in Q_e : \varphi^{J_o} \circ p^{J_q}_{J_o} \circ p^{J_o}_{J_q} = \lambda_q\,\varphi^{J_o}$$

d'où l'on déduit que φ^{J_o} vérifie (3) avec :

$$\mu = \Sigma\, \{\lambda_q \mid q \in Q_e\}/r \quad (4)$$

$\psi^{J_o} = \varphi^{J_o}$ est donc facteur de $k_{J_oJ_e}$ relatif à la valeur propre

$$(1/r)\ \Sigma\ \{\lambda_q \mid q \in Q_e\}$$

De (2), et de la relation :

$$\psi^{J_o} \circ p_{j_o}^{J_q} = \sqrt{\lambda_q}\ \varphi^{J_q}$$

l'on déduit que le facteur $\psi^{J_e} = \{\psi^{J_q} \mid q \in Q_e\}$ associé à ψ^{J_o} est tel que :

$$\forall\ q \in Q_e : \psi^{J_q} = \sqrt{\lambda_q/\mu}\ \varphi^{J_q}$$

Application au cas de l'analyse discriminante où il y a deux groupes.

Dans ce cas la variable à expliquer y est qualitative et comporte deux modalités j_0 et j_0'. Toutes les correspondances $p_{J_oJ_q}$ $(q \in Q_e)$, ainsi que $k_{J_oJ_e}$ possèdent alors le même (et seul) facteur non trivial φ^{J_o} qui est le facteur de moyenne nulle (pour p_{J_o}).

Posant :

$$p_{J_o}(j_o) = a \quad ; \quad p_{J_o}(j_o') = 1 - a$$

on a :

$$\varphi^{J_o} = (1/\sqrt{a\,(1-a)}) \begin{pmatrix} 1-a \\ -a \end{pmatrix}$$

Le facteur associé ψ^{J_e} pour $k_{J_oJ_e}$ est alors tel que :

$$\forall\ q \in Q_e : \psi^{J_q} = \left(\frac{1}{\mu\ a\ (1-a)}\right)^{1/2} ((1-a)\ p_{j_o}^{J_q} - a\ p_{j_o'}^{J_q})$$

μ étant donné par la formule (4), où λ_q $(q \in Q_e)$ désigne la valeur propre non triviale de la correspondance $p_{J_oJ_q}$.

III.3 – Etude des facteurs sur I obtenus en adjoignant en supplémentaire k_{IJ_e} à $k_{J_oJ_e}$.

Soit (ψ^{J_o}, ψ^{J_e}) un couple de facteurs non triviaux de variance 1 de $k_{J_oJ_e}$ et G (i) la valeur de ce facteur pour l'individu supplémentaire i ; on a :

$$G(i) = (1/r)\ \Sigma\ \{\psi^j\ k_{IJ_e}(i, j) \mid j \in J_e\} \qquad (5)$$

Donnant même poids $p_i = 1/n$ à chaque individu i, et compte tenu de de que :

$$\forall\ j \in J_q \subset J_e : \Sigma\ \{k_{IJ_e}(i, j) \mid i \in I\} = n\, p_{J_q}(j)$$

et de ce que les facteurs non triviaux ψ^{J_e} de $k_{J_oJ_e}$ sont centrés sur chaque J_q $(q \in Q_e)$, on a :

$$\Sigma\ \{p_i\ G(i) \mid i \in I\} = 0$$

G est donc centré. Cette propriété est intéressante lorsqu'on effectue des régressions sur ces facteurs.

Nous allons maintenant calculer la covariance entre les facteurs sur I, G_α et G_β, G_α (resp. G_β) étant associé au facteur $\psi_\alpha^{J_e}$ (resp. $\psi_\beta^{J_e}$) de $k_{J_oJ_e}$.

On a :

$$\text{Cov}\,(G_\alpha\ ,\ G_\beta) = (1/n)\ \Sigma\ \{G_\alpha\ (i)\ G_\beta(i) | i \in I\}$$

Cette expression peut se mettre d'après (0) et (5) sous la forme suivante :

$$\text{Cov}\,(G_\alpha\ G_\beta) = (1/n\ r^2)\ \Sigma\ \{k_{JJ}\ (j\ ,\ j')\ \psi_\alpha^j\ \psi_\beta^{j'} \mid j \in J_e\ ,\ j' \in J_e\}$$

Compte tenu de ce que :

a) si $j \in J_q$, $j' \in J_{q'}$, $q \in Q_e$, $q' \in Q_e$: $k_{JJ}\ (j\ ,\ j') = n\ p_{J_q J_{q'}}\ (j\ ,\ j')$

b) si $j\ ,\ j' \in J_q \subset J_e$: $k_{JJ}\ (j\ ,\ j') = n\, p_{J_q}\ (j)\ \delta_{j'}^{j} = n\ p_{J_q}\ (j)$ si $j = j'$, o sinon.

c) $(1/r)\ \Sigma\ \{p_{J_q}\ (j)\ \psi_\alpha^j\ \psi_\beta^j \mid j \in J_e\} = \delta_\alpha^\beta = 1$ si $\alpha = \beta$, o sinon.

Cov $(G_\alpha\ ,\ G_\beta)$ s'écrit :

$$\delta_\alpha^\beta / r + (1/r^2) \sum_{q \in Q_e} \sum_{q' \in Q_e - \{q\}} \sum_{j \in J_q} \sum_{j' \in J_{q'}} p_{J_q J_{q'}}(j\ ,\ j')\ \psi_\alpha^j\ \psi_\beta^{j'} \qquad (6)$$

On voit sur cette expression assez compliquée que les facteurs G_α et G_β $(\alpha \neq \beta)$ sont en général corrélés.

Par contre, si les variables explicatives sont deux à deux indépendantes

$$(p_{J_q J_{q'}} = p_{J_q} \otimes p_{J_{q'}},\ \forall\ q\ ,\ q' \in Q_e\ ,\ q' \neq q),$$

compte tenu de ce que les facteurs ψ_α^{Je} et ψ_β^{Je} sont centrés sur chaque J_q $(q \in Q_e)$, le dernier terme de (6) est nul, et l'on a :

$$\text{Cov}\,(G_\alpha\ ,\ G_\beta) = \delta_\alpha^\beta / r = 1/r \text{ si } \alpha = \beta, \text{ o sinon,}$$

ce qui signifie que les facteurs sur I sont non corrélés et de variance $1/r$.

IV – REGRESSION PAR BOULE

On possède un n échantillon des variables y, x_1, x_2, . . . , x_r. Pour prédire la valeur de y en un point M de l'espace supposé métrique des variables explicatives x_1, x_2, . . . ,x_r, on effectue une régression par boule (cf. (5)) en recherchant les points de l'échantillon en nombre s, avec s fixé, les plus proches de M. La moyenne des valeurs de y en ces s points voisins fournit l'estimation cherchée.

Cette procédure est simple. De plus, alors que la régression usuelle fournit une précision uniforme (le même écart type) pour toute prévision de y, la régression par boule donne une précision qui dépend du point M où l'on se place pour prédire y ; cette précision est mesurée par l'écart type des valeurs de y associées aux s points voisins de M. La régression par boule est donc une technique fiable puisqu'elle permet de contrôler la qualité de chaque prévision.

Pour se rendre compte de la qualité globale de la régression par boule, on peut calculer sur l'échantillon le coefficient de corrélation entre y et sa valeur prédite y*, coefficient qui est l'analogue du coefficient de corrélation multiple dans la régression usuelle.

N.B. – Pour reconstituer y en un point M de l'échantillon, on recherche les s points voisins de M parmi les points de l'échantillon différents de M.

V – UN EXEMPLE D'APPLICATION

V.1. – Les données et les analyses factorielles effectuées

Les données analysées ici ont été fournies par Y. REYRE, géologue au Muséum. Ces données sont décrites en détail dans la référence (3) où figure également l'interprétation géologique des résultats que nous exposons ici. Elles concernent un ensemble I de 277 roches caractérisées par 11 variables :

– cinq variables lithologiques correspondant à la sédimentation "en grand" : la longitude LN, la latitude LA, le type de formation géologique FM, la disposition structurale en couches ST, l'épaisseur de la strate ES.

– cinq variables pétrologiques correspondant à la sédimentation ponctuelle : la couleur CL, la texture en surface polie SS, le diamètre du grain maximum MX, le diamètre du grain le plus fréquent FR, le pourcentage de carbonate de calcium (ou calcaire) CC.

– le kérogène (ou matière organique) KE, variable que l'on désire expliquer en fonction des précédentes.

Toutes les variables quantitatives ont été découpées en classes. On obtient ainsi 70 variables logiques, 8 caractérisant le kérogène et notées KE1, KE2, ... KE8 et 62 caractérisant les variables explicatives (on a éliminé les modalités SS4 et SS7 correspondant à des effectifs nuls).

Les découpages adoptés ainsi que les sigles que nous avons choisis pour ces 70 variables sont résumés sur le tableau n°1.

Nous désignerons par I l'ensemble des 277 roches et par J l'ensemble des modalités de toutes les variables :

$$J = J_o \cup J_e$$
$$J_e = J_t \cup J_p \cup J_c$$

avec J_o : ensemble des modalités du kérogène (Card $J_o = 8$)

J_t : ensemble des modalités des variables lithologiques (Card $J_t = 32$)

J_p : ensemble des modalités des variables pétrologiques (sans le calcaire) (Card $J_p = 24$)

J_c : ensemble des modalités du calcaire (Card $J_c = 6$)

Les analyses de correspondance suivantes ont été effectuées :

1) Analyse du tableau disjonctif complet 277 x 70 : k_{IJ}

2) Analyse du tableau de régression 8 x 62 : $k_{J_o J_e}$

croisant les modalités du kérogène avec l'ensemble des modalités de toutes les variables explicatives.

Tableau 1

Désignation des variables et modalités. Effectifs

	Variables		*Classes ou modalités*	*Effectifs*
géographiques	Longitude	LN1	$1 \leqslant 8092$	90
		LN2	$8092 < 1 \leqslant 8174$	46
		LN3	$8174 < 1 \leqslant 8244$	54
		LN4	$8244 < 1 \leqslant 8398$	44
		LN5	$8398 < 1$	43
	Latitude	LA1	$L \leqslant 0726$	74
		LA2	$0726 < L \leqslant 1267$	48
		LA3	$1267 < L \leqslant 1490$	20
		LA4	$1490 < L \leqslant 1843$	82
		LA5	$1843 < L$	53
sédimentaire	Formation J_t	FM1	de type 1	12
		FM2	de type 2	39
		FM3	de type 4 ou 7	68
		FM4	de type 5 ou 6	6
		FM5	de type 8 ou 10	13
		FM6	de type 9	10
		FM7	de type 11 ou 13	89
		FM8	de type 12 ou 14 ou 15	31
		FM9	de type 16	9
lithologiques	Disposition structurale des couches	ST1	plane	116
		ST2	en biseau	4
		ST3	plane oblique	1
		ST4	entrecroisée	13
		ST5	ondulée	59
		ST6	enchevêtrée	28
		ST7	lenticulaire	16
		ST8	biohermale	40
	Epaisseur des strates	ES1	millimétrique	83
		ES2	centimétrique	47
		ES3	décimétrique	58
		ES4	métrique	66
		ES5	décamétrique	23

	Variables		*Classes ou modalités*	*Effectifs*
pétrologiques	Couleur	CL1	Beige	99
		CL2	Blanc	51
		CL3	Gris	86
		CL4	Brun	21
		CL5	Brique	20
	Texture en surface polie J_p	SS1	laminaire plane	51
		SS2	laminaire ondulée	39
		SS3	éléments orientés	6
		SS4	laminaire oblique	0
		SS5	lenticulaire	34
		SS6	enchevêtrée	2
		SS7	en poches	0
		SS8	vacuolaire	1
		SS9	confuse	111
		SS0	isotrope	33
	Diamètre du grain maximum	MX1	moins de 20 μ	15
		MX2	de 20 μ à 125 μ	126
		MX3	de 125 μ à 2 mm	59
		MX4	de 2 mm à 5 mm	39
		MX5	plus de 5 mm	38
	Diamètre du grain le plus fréquent	FR1	moins de 1 μ	5
		FR2	de 1 μ à 20 μ	149
		FR3	de 20 μ à 125 μ	13
		FR4	de 125 μ à 0,5 mm	50
		FR5	de 0,5 mm à 2 mm	51
		FR6	plus de 2 mm	9
	J_c % de $CaCO_3$	CC1	$p = 99$	73
		CC2	$97 \leqslant p < 99$	47
		CC3	$94 \leqslant p < 97$	52
		CC4	$80 \leqslant p < 94$	43
		CC5	$50 \leqslant p < 80$	37
		CC6	$p < 50$	25
principale	J_o Kérogène	KE1	$m < 50$	40
		KE2	$50 \leqslant m < 100$	36
		KE3	$100 \leqslant m < 200$	60
		KE4	$200 \leqslant m < 300$	28
		KE5	$300 \leqslant m < 500$	23
		KE6	$500 \leqslant m < 1000$	32
		KE7	$1000 \leqslant m < 2000$	31
		KE8	$2000 \leqslant m$	27

3) Analyse du sous tableau 8 x 56 : $k_{J_o J_t \cup J_p}$ où l'on ne tient plus compte du calcaire

4) Analyse du sous tableau 8 x 32 : $k_{J_o J_t}$ pour voir l'influence des variables lithologiques seules sur le kérogène.

5) Analyse du sous tableau 8 x 30 : $k_{J_o J_p \cup J_c}$ pour voir l'influence des variables pétrologiques

6) Analyse du sous tableau 8 x 24 : $k_{J_o J_p}$ pour voir l'influence des variables pétrologiques sans le calcaire

Dans toutes les analyses effectuées sur un sous tableau de $k_{J_o J_e}$, le sous tableau restant, non rentré dans l'analyse a été placé en supplémentaire.

On a été amené à mettre le calcaire en supplémentaire (cf. analyses 3) et 6)) car étant très lié au kérogène (le coefficient de corrélation entre KE et CC vaut – 0,747) il avait une importance prépondérante dans les analyses 2) et 5).

Toutes ces analyses ayant donné des résultats très semblables, une seule sera présentée, à savoir la seconde, celle du tableau $k_{J_o J_e}$, qui résume finalement toutes les autres.

V.2. – Résultats de l'analyse des correspondances du tableau $k_{J_o J_e}$

Le plan 1-2 résume 67 % de l'inertie, le premier facteur rendant compte à lui tout seul de 51 % de cette inertie, tandis que les quatre premiers facteurs correspondent à 89 % de cette inertie (cf. tableau 2)

Tableau n°2

Résultats de l'analyse du tableau $k_{J_o J_e}$

Axe	1	2	3	4
Valeur propre	0,279	0,088	0,072	0,046
% d'inertie	51,05	16,10	13,19	8,33
% d'inertie cumulée	51,05	67,15	80,34	88,67

Dans le plan 1-2 (cf. figure 2) les modalités du kérogène décrivent une courbe parabolique correspondant à l'effet GUTTMAN usuel. Sur l'axe 1, ces modalités se suivent dans l'ordre à part une légère interversion entre les deux premières modalités KE1 et KE2. Les modalités KE8, KE7, KE6, correspondant à un poids élevé en kérogène sont situées du côté négatif de cet axe 1 tandis que celles correspondant à un poids faible sont situées du côté positif de l'axe. Le premier facteur est donc lié négativement au kérogène.

Toutes les variables explicatives non qualitatives (à part la longitude) se projettent sensiblement en bon ordre sur l'axe 1, dans le même sens que le kérogène pour la latitude, en sens inverse pour les autres variables, CC – FR – MX – ES, avec une exception notable : les strates décamétriques (ES 5) qui sont du même côté sur cet axe 1 que les strates millimétriques (ES 1).

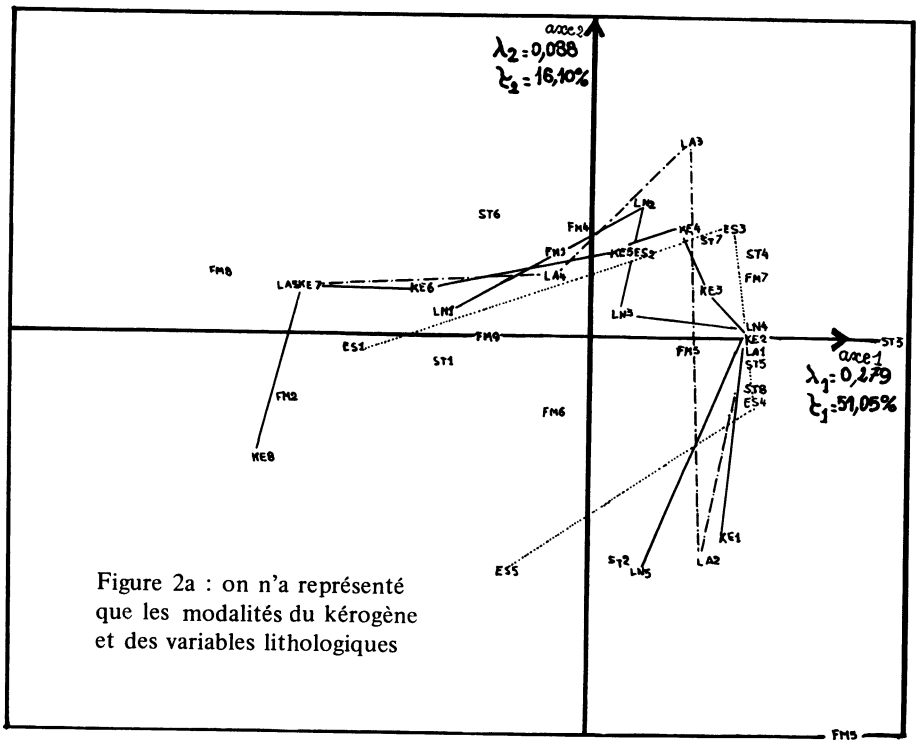

Figure 2a : on n'a représenté que les modalités du kérogène et des variables lithologiques

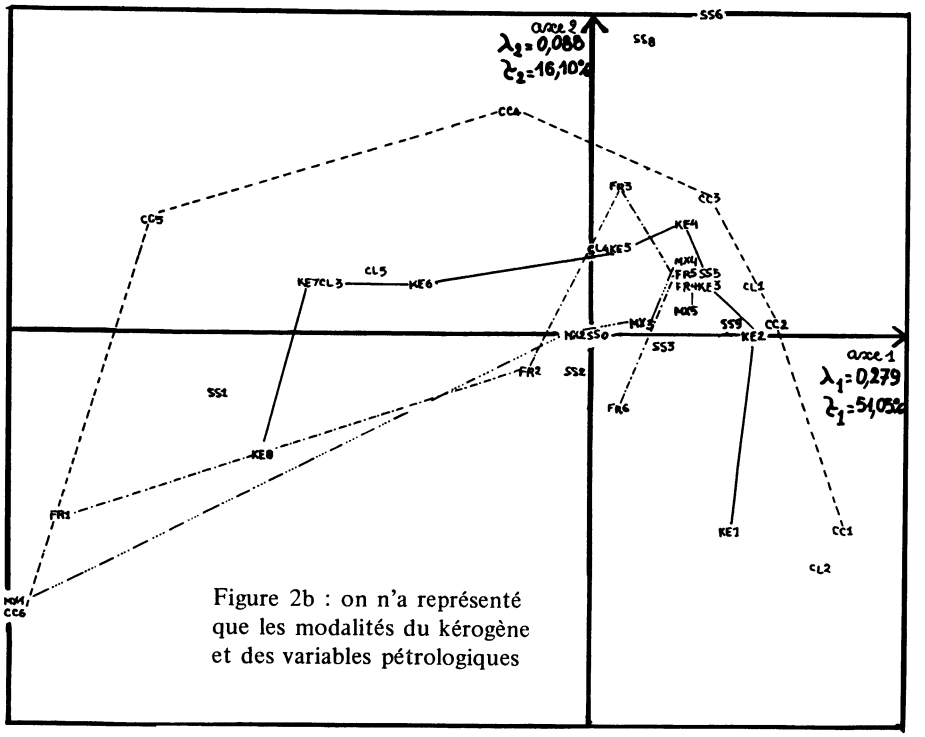

Figure 2b : on n'a représenté que les modalités du kérogène et des variables pétrologiques

Analyse Factorielle de $k_{J_o J_e}$ (plan 1-2)
Figure 2

La teneur en kérogène est donc d'autant plus forte que la granulométrie est fine, qu'il y a peu de calcaire (ou ce qui est équivalent beaucoup d'argile) et qu'il se trouve dans une roche relevée à une latitude élevée.

En ce qui concerne les variables qualitatives, une teneur élevée en kérogène semble liée :

- à une couleur grise (CL 3) ou brique (CL 5)
- aux formations de type 2 (FM 2), 12, 14 ou 15 (FM 8)
- à une disposition structurale plane (ST 1) ou enchevêtrée (ST 6)
- à une texture laminaire plane (SS 1)

De même une teneur moyenne ou faible en kérogène semble liée :

- à une couleur blanche (CL 2) et à un degré moindre beige (CL 1)
- aux formations de type 8 ou 10 (FM 5), 11 ou 13 (FM 7)
- à une disposition structurale ondulée (ST 5)
- à une structure confuse (SS 9)

Notons que le calcaire contribue respectivement pour 25 %, 29 %, 38 %, et 51 % à la formation des quatre premiers axes factoriels. C'est la raison pour laquelle, on a refait cette analyse en mettant cette variable, qui est la variable explicative la plus importante, en supplémentaire , les résultats obtenus étant proches comme on l'a déjà dit de ceux obtenus ici, nous ne les présenterons pas.

V.3. – Prévision de la teneur en kérogène

Les caractéristiques des variables quantitatives initiales (moyenne, écart-type, matrice de corrélation) sont résumées sur les tableaux 3 et 4*

Tableau n° 3

Caractéristiques des variables quantitatives

variable	LN	LA	CC	MX	FR	ES	KE
moyenne : m	8 211	1 372	85,4	4,68	3,19	2,64	840
écart-type : s	165	544	22,1	2,10	1,59	1,35	1 757
m/s	49,9	2,52	3,87	2,23	2,01	1,96	0,48

(*) Nous avons adopté une échelle logarithmique pour la variable ES, en donnant respectivement les valeurs 1, 2, 3, 4, 5 selon que l'on a une épaisseur de strate millimétrique, centimétrique, décimétrique, métrique ou décamétrique.

Tableau n° 4

Matrice de corrélation des variables quantitatives

	LN	LA	CC	MX	FR	ES	KE
LN	1						
LA	− 0,4	1					
CC	0,35	− 0,38	1				
MX	0,14	− 0,21	0,31	1			
FR	0,09	− 0,12	0,26	0,92	1		
ES	0,19	− 0,375	0,37	0,23	0,16	1	
KE	− 0,22	0,25	− 0,747	− 0,22	− 0,21	− 0,22	1

On voit que la variable à expliquer, le kérogène, est très dispersée (son écart-type est plus de deux fois plus grand que sa moyenne) et très liée, comme on l'a déjà vu au calcaire (r (KE, CC) = − 0,747).

On retrouve également d'une part la liaison positive du kérogène avec la latitude, (r (KE, LA) = 0,251), d'autre part les liaisons négatives du kérogène avec les diamètres du grain maximum MX, et du grain le plus fréquent FR (r (KE, MX) = − 0,219 ; r (KE, FR) = − 0,206) qui sont très corrélés entre eux (r (MX, FR) = 0,916).

Néanmoins, les liaisons linéaires du kérogène avec les variables autres que le calcaire sont faibles, la valeur absolue du coefficient de corrélation de KE avec une variable autre que CC étant inférieure ou égale à 0,251.

L'analyse des correspondances, grâce au découpage en classes (qui correspond à une procédure non linéaire) a permis de mieux apprécier les liaisons entre le kérogène et ces variables quantitatives, tout en y incluant les variables qualitatives ou semi-quantitatives.

Nous avons effectué les régression suivantes :

1) Régression usuelle sur les variables explicatives quantitatives, avec et sans calcaire

2) Régression usuelle sur les facteurs issus des différentes analyses factorielles effectuées

3) Régression par boule*

a) dans l'espace des quatre premiers facteurs issus des différentes analyses factorielles

(*) Dans chacun des espaces considérés, nous avons adopté la métrique usuelle pour rechercher les voisins d'un point donné, et nous avons retenu 20 voisins pour estimer la teneur en kérogène.

b) dans l'espace R_{J_e} ou $R_{J_t \cup J_p}$ associé au tableau disjonctif complet $k_{IJ_e \cup J_p}$, qui sont des sous tableaux de k_{IJ}

c) dans l'espace des variables explicatives quantitatives (avant découpage en classes), avec et sans calcaire, chaque variable étant ramenée à avoir une variance égale à 1.

Les résultats obtenus sont schématisés sur les tableaux n° 5 et 6. On voit sur le tableau 5 que la régression par boule sur les facteurs donne des résultats équivalents ou meilleurs que la régression usuelle sur les mêmes facteurs, en ce qui concerne la corrélation entre y la variable à expliquer (i.e. le kérogène) et y* sa valeur approchée.

Il est par ailleurs intéressant de noter que dans la régression par boule, la corrélation entre y et sa valeur approchée y* est de 0,59 (resp. 0,47) dans l'espace initial à 62 (resp. 56) dimensions R_{J_e} (resp. $R_{J_t \cup J_p}$) alors que cette corrélation est de 0,71 (resp. 0,57) dans l'espace des quatre premiers facteurs de l'analyse de $k_{J_o J_e}$ (resp. $k_{J_o J_p \cup J_t}$) et 0,68 (resp. 0,54) dans l'espace des deux premiers facteurs.

Pour voir l'influence du nombre de voisins, nous avons fait varier ce nombre entre 5 et 50, en nous plaçant dans l'espace des quatre premiers facteurs de $k_{J_o J_e}$. La corrélation entre y et y* a varié entre 0,67 et 0,71 (cf. tableau 7).

En ce qui concerne le calcaire, son influence est très nette dans la régression usuelle, où le fait de rajouter les variables quantitatives LN, LA, ES, FR, MX, augmente de façon négligeable la qualité de l'ajustement, le coefficient de corrélation multiple du kérogène en fonction de ces six variables étant de 0,752, à peine supérieur à la valeur absolue de la corrélation entre calcaire et kérogène qui est de 0,747.

On peut noter que les corrélations obtenues dans les différentes régressions par boule où figure le calcaire (soit directement, soit indirectement par les facteurs d'une analyse factorielle où il se trouve en variable de base) sont inférieures à la corrélation (en valeur absolue) entre kérogène et calcaire. Ceci semble dû d'une part à l'importance du calcaire, et d'autre part au fait que l'objet même de la régression usuelle est la maximisation (parmi les combinaisons linéaires des variables explicatives) de la corrélation entre la variable à expliquer y et sa valeur approchée y*.

L'intérêt de la régression par boule par rapport à la régression usuelle réside, comme on l'a déjà mentionné dans le fait qu'elle fournit une précision différente pour chaque estimation de y. D'un point de vue pratique, cette propriété est fort intéressante, surtout quand la variable à expliquer est très dispersée, ce qui est le cas ici du kérogène.

Si l'on retire l'influence du calcaire, la régression usuelle sur LA, LN, ES, FR et MX ne permet d'expliquer le kérogène qu'avec un coefficient de corrélation multiple égal à 0,346, tandis qu'avec la régression par boule, on n'obtient qu'une corrélation de 0,315[1]; par contre, si l'on se place dans

(1) Cette corrélation égale à 0,315 pour 20 voisins, vaut 0,337 avec 10 voisins, 0,364 avec 30 voisins et 0,358 avec 40 voisins.

Tableau n°5

Corrélation entre le kérogène et sa valeur approchée dans la régression par boule (20 voisins) et la régression usuelle, effectuées dans l'espace des quatre premiers facteurs des analyses factorielles 1 à 6 (cf. V.1)

Analyse n°	1	2	3	4	5	6
tableau analysé	k_{IJ}	$k_{J_o J_e}$	$k_{J_o J_t \cup J_p}$	$k_{J_o J_t}$	$k_{J_o J_p \cup J_c}$	$k_{J_o J_p}$
Régression usuelle	0,585	0,680	0,575	0,482	0,660	0,535
Régression par boule (20 voisins)	0,684	0,71*	0,57**	0,46	0,695	0,557

* Corrélation égale à 0,68 avec deux facteurs et à 0,58 avec un facteur.
** Corrélation égale à 0,54 avec deux facteurs.

Tableau n°6

Corrélation entre le kérogène et sa valeur approchée dans la régression par boule (20 voisins) et la régression usuelle, effectuées sur les variables quantitatives initiales et sur les variables logiques associées à J_e et $J_t \cup J_p$ (cf. V.1 et tableau n°1)

Variables	LN, LA, MX FR, ES, CC	LN, LA, MX FR, ES.	variables logiques associées à	
			J_e	$J_t \cup J_p$
Régression usuelle	0,752	0,346		
Régression par boule (20 voisins)	0,701	0,315	0,587	0,474

Tableau n°7

Influence du nombre de voisins sur la corrélation R entre y et sa valeur approchée y* dans la régression par boule faite dans l'espace des quatre premiers facteurs de $k_{J_o J_e}$

Nombre de voisins	5	8	10	15	20	30	40	50
R	0,691	0,666	0,679	0,703	0,709	0,707	0,696	0,687

l'espace des quatre premiers facteurs de $k_{J_o J_t \cup J_p}$ (où l'influence du calcaire a été enlevée) la régression par boule (resp. usuelle) fournit une corrélation de 0,57 (resp. 0,575), on voit ainsi l'intérêt de se placer dans cet espace.

On peut par ailleurs noter que l'influence des variables pétrologiques autres que le calcaire semble être un peu plus importante que l'influence des variables lithologiques, les corrélations fournies par la régression par boule (resp. usuelle) dans l'espace des quatre premiers facteurs de $k_{J_o J_p}$ et $k_{J_o J_t}$ étant respectivement de 0,557 et 0,46 (resp. 0,535 et 0,482).

N.B. – Pour se rendre compte de la significativité des corrélations calculées dans les différentes régressions par boule effectuées, nous avons tiré au hasard les voisins. Avec 10 voisins, on a obtenu une corrélation de – 0,054, et avec 20 voisins une corrélation de 0,037, ce qui montre que les résultats que l'on a obtenus par ce type de régression sont très significatifs.

V.4 – Conclusion

L'analyse des correspondances du tableau croisé $k_{J_o J_e}$ des modalités des variables explicatives avec celles de la variable à expliquer y, ici le kérogène, a permis de bien voir les liaisons entre cette variable et les différentes variables explicatives.

La régression par boule permet d'estimer la valeur de y, connaissant les variables explicatives, ainsi que la précision de cette estimation. Il semble qu'il y ait intérêt à faire cette régression (en même temps que la régression usuelle) sur les facteurs issus du tableau disjonctif complet k_{IJ_e} associé aux variables explicatives, mis en supplémentaire sur le tableau croisé $k_{J_o J_e}$.

ANNEXE N°1

Supposons les variables explicatives 2 à 2 indépendantes ; on a alors :

$$\forall\ q, q' \in Q_e^{*}:\ p_{J_q J_{q'}} = p_{J_q} \otimes p_{J_{q'}}$$

Cette relation est équivalente, si δ^{J_q} désigne la fonction constante et égale à 1 sur J_q, à :

$$\forall\ q, q' \in Q_e^{*}:\ p^{J_q}_{J_{q'}} = p_{J_{q'}} \otimes \delta^{J_q}$$

Pour tout facteur non trivial φ^J de k_{JJ}, i.e. centré sur chaque J_q ($q \in Q$), on a :

$$\forall\ q, q' \in Q_e^{*}:\ \varphi^{J_{q'}} \circ p^{J_q}_{J_{q'}} = (\varphi^{J_{q'}} \circ p_{J_{q'}})\ \delta^{J_q} = 0$$

Les équations des facteurs non triviaux de k_{JJ} s'écrivent alors (cf. équation (1) du III.1) d)) :

$$\left.\begin{aligned} \Sigma\ \{\varphi^{J_q} \circ p^{J_o}_{J_q} \mid q \in Q_e\} &= ((r+1)\sqrt{\lambda} - 1)\ \varphi^{J_o} \\ \forall\ q \in Q_e :\ \varphi^{J_o} \circ p^{J_q}_{J_o} &= ((r+1)\sqrt{\lambda} - 1)\ \varphi^{J_q} \end{aligned}\right\} \qquad (A1)$$

(*) $q \neq q'$

Posant

$$\left.\begin{aligned} \forall\, q \in Q_e : \psi^{Jq} &= a\, \varphi^{Jq} \\ \psi^{Jo} &= b\, \varphi^{Jo} \end{aligned}\right\} \qquad \text{(A2)}$$

On peut calculer a et b de façon à ce que

$$(\psi^{Jo}\ ,\ \psi^{Je} = \{\, \psi^{Jq} \mid q \in Q_e \,\})$$

soit un couple de facteurs normés de $k_{J_o J_e}$, associé à la valeur propre μ (i.e. vérifie les équations (2) du III 1) d))

Il vient tout calcul fait, si φ^J est de variance 1 :

$$\left.\begin{aligned} a &= (2r/(r+1))^{1/2} \\ b &= (2/(r+1))^{1/2} \times \text{signe}\,((r+1)\sqrt{\lambda} - 1) \\ \mu &= ((r+1)\sqrt{\lambda} - 1)^2/r \end{aligned}\right\} \qquad \text{(A3)}$$

A tout facteur non trivial de variance 1 φ^J de k_{JJ} relatif à une valeur propre λ différente de $1/(r+1)^2$ correspond donc avec les formules (A2) et (A3) un couple de facteurs normés non triviaux (ψ^{Jo}, ψ^{Je}) de $k_{J_o J_e}$, associé à la valeur propre μ. Réciproquement à tout couple de facteurs normés non triviaux (ψ^{Jo}, ψ^{Je}) de $k_{J_o J_e}$ associé à la valeur propre μ correspond deux facteurs normés non triviaux de k_{JJ} que nous désignerons par φ^J_ϵ, ϵ pouvant prendre les valeurs 1 ou -1 . φ^J_ϵ se déduit de $(\psi^{Jo}\ ,\ \psi^{Je})$ à partir des formules (A2) où :

$$a = (2r/(r+1))^{1/2}$$

$$b = (2/(r+1))^{1/2}\ \epsilon$$

$$\lambda_\epsilon = ((1 + \epsilon\sqrt{r\mu})/(r+1))^2$$

λ_ϵ désignant la valeur propre associée à φ^J_ϵ.

Remarque : ψ^{Jo} et ψ^{Je} étant de variance unité, les contributions de φ^{Jo}_ϵ et φ^{Je}_ϵ à la variance de φ^J_ϵ qui vaut 1, sont égales à 1/2.

ANNEXE N°2

Supposons les variables explicatives divisées en deux groupes Q_1 et Q_2, y étant indépendante des variables du groupe Q_2. On a donc :

$$\left.\begin{aligned} \forall\, q \in Q_2 : p^{J_o}_{J_q} &= p_{J_q} \otimes \delta^{Jo} \\ p^{J_q}_{J_o} &= p_{J_o} \otimes \delta^{Jq} \end{aligned}\right.$$

Les équations (2) (cf. III.1) d)) des facteurs non triviaux de $k_{J_o J_e}$ se simplifient alors et s'écrivent :

$$\Sigma \{ \psi^{J_q} \circ p^{J_o}_{J_q} \mid q \in Q_1 \} = r\sqrt{\mu}\, \psi^{J_o}$$

$$\forall\, q \in Q_1 : \psi^{J_o} \circ p^{J_q}_{J_o} = \sqrt{\mu}\, \psi^{J_q}$$

$$\forall\, q \in Q_2 : \sqrt{\mu}\, \psi^{J_q} = 0$$

Si l'on pose $J^1 = \cup \{ J_q \mid q \in Q_1 \}$, on voit que l'analyse du tableau $k_{J_o J_e}$ est équivalente à l'analyse du sous tableau $k_{J_o J^1}$ croisant les modalités de y avec les modalités des variables explicatives non indépendantes de y.

De façon précise, si r_1 désigne le cardinal de Q_1 (r_1 supposé non nul) et si J^2 désigne la réunion des J_q pour $q \in Q_2$, au couple de facteurs normés non triviaux $(\psi^{J_o}, \psi^{J_e} = (\psi^{J^1}, 0^{J^2}))$ relatif à la valeur propre μ, correspond le couple de facteurs normés non triviaux $(\psi^{J_o}, (r_1/r)^{1/2}\, \psi^{J^1})$ de $k_{J_o J^1}$ relatif à la valeur propre $(r/r_1)\mu$ et réciproquement.

Dans le cas où $r_1 = 0$, i.e. dans le cas où y est indépendante de toutes les variables explicatives, alors $k_{J_o J_e}$ n'admet que des facteurs triviaux. Dans ce dernier cas l'analyse du tableau de BURT k_{JJ} est équivalente à celle du sous tableau $k_{J_e J_e}$.

BIBLIOGRAPHIE

(1) BENZECRI J.P. – Sur l'analyse des tableaux binaires associés à une correspondance multiple ([Bin. Mult.]) : publication du Laboratoire de Statistique (1 972).

(2) BRENOT J., CAZES P., LACOURLY N. – Pratique de la régression : qualité et protection. *Cahiers du B.U.R.O.*, N°23, (1975).

(3) CAZES P., REYRE Y. – La fossilisation du kérogène en milieu argilo-carbonaté. Etude statistique de ses liaisons avec les propriétés lithologiques et petrologiques dans l'Oxfordien du Bassin de Paris (Partie Orientale) – *Bulletin du B.R.G.M.*, 2ème série, section IV, n° 2, pp. 85-102 (1976).

(4) CAZES P. – Etude de quelques propriétés extrémales des facteurs issus d'un sous tableau d'un tableau de BURT ([Extr. – Fac.]), publication du Laboratoire de Statistique, (Novembre 1975).

(5) LEBEAUX M.O. – Programmes de régression et de classification utilisant la notion de voisinage. Thèse de 3ème cycle, Paris (1974).

(6) LECLERC A. – Etude de certains types de tableaux par l'analyse des correspondances. Thèse de 3ème cycle, Paris (1973)

(7) LECLERC A. – L'analyse des correspondances sur juxtaposition de tableaux de contingence *R.S.A.* vol. XXIII N°3, pp. 5-16 (1975).