LES CAHIERS DE L'ANALYSE DES DONNÉES

A. EL OUADRANI

Généralisation du tableau de Burt et de l'analyse de ses sous-tableaux dans le cas d'un codage barycentrique

Les cahiers de l'analyse des données, tome 19, n° 2 (1994), p. 229-246

<http://www.numdam.org/item?id=CAD_1994__19_2_229_0>

Les Cahiers de l'Analyse des Données
Vol. XIX - 1994 - n°2 - pp.229-246.

GÉNÉRALISATION DU TABLEAU DE BURT ET DE L'ANALYSE DE SES SOUS-TABLEAUX DANS LE CAS D'UN CODAGE BARYCENTRIQUE

[BURT COD. BARY.]

A. El OUADRANI

Le présent article comprend deux parties. Au §1, après avoir rappelé les propriétés du tableau de BURT dans le cas usuel de variables discrètes (ou de variables continues découpées en classes), on généralise ces propriétés pour le cas de variables continues soumises à un codage barycentrique. Au §2, des données, relatives aux propriétés mécaniques de 1511 échantillons de bois d'épicéa, sont traitées en recourant à un tableau de BURT généralisé.

1 Définition et calcul du tableau de BURT: cas usuel et généralisation

1.1 Le tableau de description logique des individus

Rappelons brièvement la situation classique à laquelle se réfère la notion de tableau de BURT.

Un ensemble I, comprenant cari individus, est décrit par un ensemble J de carj variables:

$$J = \{j \mid j = 1, .., carj\} .$$

Chaque variable j est découpée en un ensemble de modalités, qu'on supposera d'abord numérotées de 1 à nmj[j]; soit, au total, pour l'ensemble J des variables, un ensemble MJ de modalités, avec:

$$card(MJ) = carm = \Sigma\{nmj[j] \mid j = 1, .., carj\} .$$

Il est commode d'attribuer à chaque modalité un numéro de 1 à carm; à cette fin on note dj[j] le nombre total des modalités des variables de rang strictement inférieur à j:

$$dj[j] = \Sigma\{nmj[jp] \mid jp = 1, .., j-1\} ;$$

en particulier, dj[1]=0; ceci fait, la μ-ème modalité de la variable j reçoit, dans la suite MJ de toutes les modalités, le numéro $m = dj[j] + \mu$.

Désormais, nous utiliserons systématiquement la lettre grecque μ pour

désigner le numéro d'une modalité parmi celles d'une variable j particulière ($1 \leq \mu \leq nmj[j]$); et *m* pour le numéro d'une modalité considérée comme élément de MJ ($1 \leq m \leq carm$).

L'ensemble I des individus peut être décrit par un tableau k_{IM}, en (0,1), ayant cari lignes et carm colonnes; avec:

$k_{IM}[i, m] = 1$ *si* i *rentre dans la modalité* m, *et* zéro *sinon* ;

ainsi, dans chacun des blocs {dj[j]+1, .., dj[j]+nmj[j]} afférents à une variable j, la ligne i, afférente à un individu, comprend, une suite de zéros avec un seul chiffre 1; celui-ci étant dans la colonne dj[j]+μ, pour la modalité μ, de la variable j, dans laquelle rentre i.

Sous une forme plus compacte, la même information peut être disposée dans un tableau k_{IJ}, ayant cari lignes et carj colonnes; avec:

$k_{IJ}[i,j] = m = dj[j]+\mu$, *si* i *rentre dans la* μ-ème *modalité de la variable* j ;

ainsi, dans la ligne décrivant l'individu i, le j-ème nombre est un entier, dont la valeur doit être comprise entre dj[j]+1 et dj[j]+nmj[j].

Il va sans dire qu'on peut inscrire, dans la colonne j, le nombre μ lui-même, compris entre 1 et nmj[j], au lieu du nombre m. En inscrivant μ plutôt que m, on a même l'avantage d'éviter toute ambiguïté dans la cas où des variables seraient supprimées ou introduites; et c'est pourquoi l'on procède généralement ainsi dans les programmes. Mais, dans le présent exposé, nous préférons supposer qu'on inscrit m, afin de simplifier, au maximum, la formule de calcul du tableau de BURT; ou sa généralisation au cas d'un codage barycentrique.

Il est d'usage d'appeler: tableau de description logique, le tableau k_{IM}, en (0,1); le tableau k_{IJ} étant, naturellement, appelé: tableau de numéro des modalités.

1.2 Tableau de description logique et tableau de BURT

On sait que, dans de nombreuses études, notamment quand les variables j sont des questions fermées, dont chacune admet un nombre déterminé de réponses, mais aussi avec des variables continues découpées en classes, l'analyse du tableau de description logique k_{IM} a fourni des résultats très satisfaisants.

Avec un tel tableau, où la densité d'information est très faible, un pareil succès était, *a priori*, inattendu. Il s'explique parce que l'analyse de k_{IM}, se ramène à celle d'un tableau carré B_{MM}, ou tableau de BURT, qui est un véritable tableau de contingence; et renferme, dans un format carré carm ×

carm, une information d'autant plus dense que le nombre cari des individus est plus élevé.

De façon précise, le tableau de BURT ne contient rien d'autre que le dénombrement des cooccurrences entre modalités:

$B_{MM}[m, mp] =$

nombre des individus i *tels que* $k_{IM}[i, m] = k_{IM}[i, mp] = 1$.

De cette définition, il résulte, en particulier, que $B_{MM}[m, m]$ est le nombre des individus rentrant dans la modalité m; et, également, que $B_{MM}[m,mp] = 0$, si m et mp sont deux modalités différentes d'une même question; etc.

Il se trouve que les facteurs *normalisés* (de variance 1) sur M, issus de l'analyse de B_{MM} sont les mêmes que ceux issus de l'analyse du tableau de description logique k_{IM}; et la seule différence entre les facteurs F(m) de même rang issus des deux tableaux, consiste en ce que les valeurs propres (ou variances de ces facteurs) sont, pour B_{MM}, le carré de ce qu'elles sont pour k_{IM}. Quant aux facteurs sur I, que l'analyse de k_{IM} donne directement, on peut les obtenir également en adjoignant ce même tableau k_{IM} en supplément au tableau de BURT, B_{MM}.

Il n'y a pas lieu de reprendre ici un exposé théorique complet, maintes fois donné ailleurs, mais il importe d'en rappeler ce qui fonde la généralisation au codage barycentrique.

À un coefficient de proportionnalité près, le tableau de BURT, B_{MM}, correspond, en bref, à la transition composée de $k_J{}^J$, de J vers J, sur laquelle se fonde l'analyse de tout tableau de correspondance k_{IJ}. On sait que les facteurs sur J ne sont autres que les vecteurs propres de la transition composée $k_J{}^I \circ k_I{}^J$ de J vers J (de J vers I; puis de I vers J). De cela, il résulte que les facteurs sur J peuvent encore s'interpréter comme issus d'un tableau de correspondance symétrique s_{JJ} dont la définition ne diffère de celle de la transition composée que par ce qu'il faut pour établir la symétrie. Tandis que:

$$k_J{}^I \circ k_I{}^J(j, jp) = \Sigma\{(k(i, j) / k(i)).(k(i, jp) / k(jp)) \mid i \in I\} ;$$

on a pour s_{JJ}:

$$s_{JJ}(j,jp) = \Sigma\{k(i, j).k(i, jp) / k(i) \mid i \in I\} .$$

En sommant par rapport à jp et à i, on voit que le tableau symétrique, s_{JJ}, a même marge sur J que le tableau de départ k_{IJ}; d'où il résulte que la transition $s_J{}^J$, de J vers J, n'est autre que la transition composée $k_J{}^I \circ k_I{}^J(j, jp)$, associée à k_{IJ}.

L'analyse de s_{JJ} repose donc sur la diagonalisation d'une transition qui est le carré de celle construite pour analyser k_{IJ}; et c'est pourquoi, l'analyse de s_{JJ} fournit, sur J, les mêmes facteurs normalisés que l'analyse du tableau de base k_{IJ}; les valeurs propres (ou variances des F) étant pour s_{JJ}, le carré de ce qu'elles sont pour k_{IJ}; les facteurs sur I pouvant s'obtenir en adjoignant k_{IJ} en supplément à s_{JJ}.

Pour revenir au tableau de BURT, il faut considérer que dans le cas du tableau k_{IM}, en (0,1), toutes les lignes i ont la même masse: carj, le nombre des variables. La division par k(i) est donc superflue. Reste l'expression:

$$s_{MM}(m, mp) = \Sigma\{k(i, m).k(i, mp) \mid i \in I\} ;$$

laquelle, puisque les k(i, m) valent tous 1 ou 0, n'est autre que le nombre, $B_{MM}(m, mp)$, des individus rentrant à la fois dans les modalités m et mp.

1.3 Codage barycentrique et modalités fractionnaires

Nous avons rappelé que l'analyse des tableaux k_{IM} et B_{MM} s'applique non seulement à des variables j données comme discrètes; mais à des variables continues découpées en classes. Dans ce dernier cas, il est clair que le codage en classes fait perdre quelque information. Si, par exemple on définit, parmi les employés d'une entreprise, 4 classes d'âge: 18-25ans, 25-35ans, 35-50ans, >50ans; la distinction est perdue entre sujet de 26 ans et de 34 ans...

Cette perte d'information compte vraisemblablement peu si l'on a un grand nombre d'individus, et qu'on s'intéresse à la structure globale donnée par les facteurs sur J. Au niveau même des individus, avec un grand nombre de variables, la redondance de celles-ci, permet que l'imprécision des notes individuelles soit corrigée par la conjonction de plusieurs notes. Mais si le nombre des individus est restreint; ou qu'il y a peu de variables et qu'on s'intéresse aux cas individuels; on s'appliquera à ne rien perdre de l'information recueillie. Ce que permet le codage barycentrique.

En bref, au lieu de découper l'intervalle X(j) où peut être compris j, en une suite de nmj[j] segments succesifs, dont chacun définit une modalité μ, on place sur X(j) une suite de nmj[j] valeurs repère successives, communément appelées pivots; et que nous noterons $x(\mu)$.

Si, pour l'individu i, la valeur xi de la variable j est exactement $x(\mu)$, on dit que i est exactement dans la modalité μ. Mais si x_i tombe entre $x(\mu)$ et $x(\mu+1)$, on attribue à i un numéro de modalité qui est fractionnaire, compris entre μ et $\mu+1$; et, comme il se doit, d'autant plus près de μ que, sur l'intervalle $[x(\mu), x(\mu+1)]$, la valeur x_i est plus proche de l'extrémité gauche, $x(\mu)$. Si x_i tombe à gauche du 1-er pivot, on pose $\mu=1$, comme si xi était sur le pivot; et si

x_i est au-delà du dernièr pivot, on pose, de même, μ=nmj[j].

Ceci posé, on peut reprendre les notations du §1.1. Le tableau k_{IJ} de numéros des modalités, est simplement défini en posant: k(i,j) = dj[j]+μ; la seule particularité étant que k(i,j) n'est pas, en général, un nombre entier; mais un nombre réel compris entre dj[j]+1 et dj[j]+nmj[j].

Quant au tableau k_{IM}, ce ne sera pas un tableau en (0,1); mais, dans chacun des blocs {dj[j]+1, .., dj[j]+nmj[j]} afférents à une variable j, la ligne i, afférente à un individu, comprend, une suite de zéros avec deux nombres réels non nuls dont la somme est 1; ceux-ci étant dans les colonnes afférentes aux pivots qui encadrent la valeur x_i.

De façon précise, si tμ=trunc(μ) désigne la partie entière du numéro de la modalité de j attribuée à i, on aura:

$$k_{IM}(i, dj[j]+t\mu) = \mu+1-t\mu \quad ; \quad k_{IM}(i, dj[j]+t\mu+1) = \mu-t\mu \quad ;$$

de la sorte, si x_i est proche du pivot tμ, la masse la plus faible, μ-tμ, tombe sur le pivot suivant; et c'est le contraire si x_i est plus proche de celui-ci.

On remarquera que, si les seules valeurs prises par une variable j sont celles choisies pour les pivots, les numéros de modalités ne peuvent être que des entiers. On retrouve le cas du codage logique. Éventuellement, codage logique et codage barycentrique proprement dit se conjuguent dans une même étude.

1.4 Généralisation du tableau de BURT

Jusqu'ici, quand on applique le codage barycentrique, on analyse le tableau k_{IM}. Cependant, dans le cas de données en (0,1), le tableau de BURT n'a pas pour seuls mérites de rendre raison du succès de l'analyse tout en donnant pour les taux d'inertie afférents aux facteurs succesifs une décroissance plus rapide que celle obtenue avec le tableau logique, il permet de fonder sur l'analyse des correspondances la solution de problèmes de régression, compris dans le sens le plus général.

De façon précise, pour voir les liens réciproques entre deux groupes de variable, J1 et J2, on analyse le sous-tableau de BURT croisant les blocs M1 et M2 des modalités respectives des variables de ces deux groupes. L'étude peut considérer les individus eux-mêmes introduits en supplément: d'une part, comme décrits par J1, avec le tableau k_{IM1}; et, d'autre part, avec k_{IM2}, comme décrits par J2. Ainsi, on a, de chaque individu i, deux images i1 et 12, sur les mêmes axes factoriels; et la corrélation entre les coordonnées de ces deux images indique dans quelle mesure chacune de celle-ci peut être estimée d'après l'autre; ce qui est le problème fondamental de la régression.

Pour étendre ces considérations à des données codées barycentriquement, il faut, outre le tableau k_{IM}, construire un analogue B_{MM} du tableau de BURT. Avec les notations que nous avons posées, et compte tenu de la manière dont, au §1.2, on a compris le tableau de BURT comme une forme, particulièrement simple, du tableau carré symétrique s_{JJ}, associé à toute correspondance k_{IJ}, la généralisation est immédiate.

En effet, comme au §1.2, toute ligne de k_{IM} a même total carj. On posera donc:

$$s_{MM}(m, mp) \approx B_{MM}(m, mp) = \Sigma\{k(i, m).k(i, mp) \mid i \in I\} .$$

Relativement au cas du codage en (0,1), le nombre des termes non nuls est seulement multiplié par 4; car, en bref, pour le croisement de deux variables j et jp interviennent deux couples de modalités consécutives; et, de plus, les termes ne valent pas 1, mais sont des produits quelconques de nombres compris entre 0 et 1.

Quant au calcul du tableau de BURT, un programme complet doit comporter les déclarations appropriées; la procédure d'entrée du tableau k_{IJ}, de numéros des modalités, avec le dénombrement de celles-ci par variable (i.e. les fonctions introduites ci-dessus, au §1.1: nmj[j], nombre de modalités de la variable j; et dj[j], numéro précédant immédiatement le bloc j, dans la suite des modalités de toutes les variables); et la procédure de sortie du tableau de BURT, B_{MM}, créé.

Nous nous bornerons à publier des procédures qui montrent comment passer de k_{IJ} à B_{MM}; en évitant, d'une part, de parcourir une grande boucle comportant le calcul de termes dont la plupart sont nuls; et tenant compte, d'autre part, de la symétrie de B_{MM}.

Le tableau B_{MM} étant mis à zéro, la prodédure BURTer doit être exécutée dans une boucle, pour tous les individus, i allant de 1 à cari. [De façon précise, le tableau k_{IJ} est lu ligne par ligne; et les lignes sont traitées au fur et à mesure qu'elles sont lues.] On distingue deux éventualités: codage logique, en (0,1), rpw='L'; et codage barycentrique, ou continu, 'C'.

Pour chaque individu i, BURTer utilise bien l'information compacte apportée par le tableau k_{IJ} de numéros des modalités afin de n'effectuer que les produits afférents aux termes non nuls. De plus, sont seuls remplis les blocs de B_{MM} croisant les modalités de deux variables {j, jp} pour lesquelles $j \leq jp$; toutefois, on n'a pas cru utile de prendre en compte explicitement la symétrie des blocs diagonaux.

Après l'exécution de BURTer, la procédure COMPLer calcule les blocs {j, jp} manquants.

```
procedure BURTer;
  var wjj,wjp:integer;aoo,aou,auo,auu,rjj,rjp:single;
begin
  if (rpw='L') then for j:=1 to carj do for jp:=j to carj do
    BMM[kIJ[i,j],kIJ[i,jp]]:=BMM[kIJ[i,j],kIJ[i,jp]]+1;
  if (rpw='C') then for j:=1 to carj do
    for jp:=j to carj do begin
      wjj:=trunc(kIJ[i,j]);wjp:=trunc(kIJ[i,jp]);
      rjj:=kIJ[i,j]-wjj;rjp:=kIJ[i,jp]-wjp;
      auu:=rjj*rjp;aou:=rjp-auu;auo:=rjj-auu;aoo:=1+auu-(rjj+rjp);
      if (0<aoo) then BMM[wjj,wjp]     :=BMM[wjj,wjp]     + aoo;
      if (0<auo) then BMM[wjj+1,wjp]   :=BMM[wjj+1,wjp]   + auo;
      if (0<aou) then BMM[wjj,wjp+1]   :=BMM[wjj,wjp+1]   + aou;
      if (0<auu) then BMM[wjj+1,wjp+1]:=BMM[wjj+1,wjp+1]+ auu;end;
end;

procedure COMPLer;begin
 for j:=1 to carj-1 do for m:=dj[j]+1 to dj[j]+nmj[j] do
   for mp:=dj[j]+nmj[j]+1 to carm do
     BMM[mp,m]:=BMM[m,mp];
end;
```

2 Exemple d'application à des propriétés mécaniques du bois

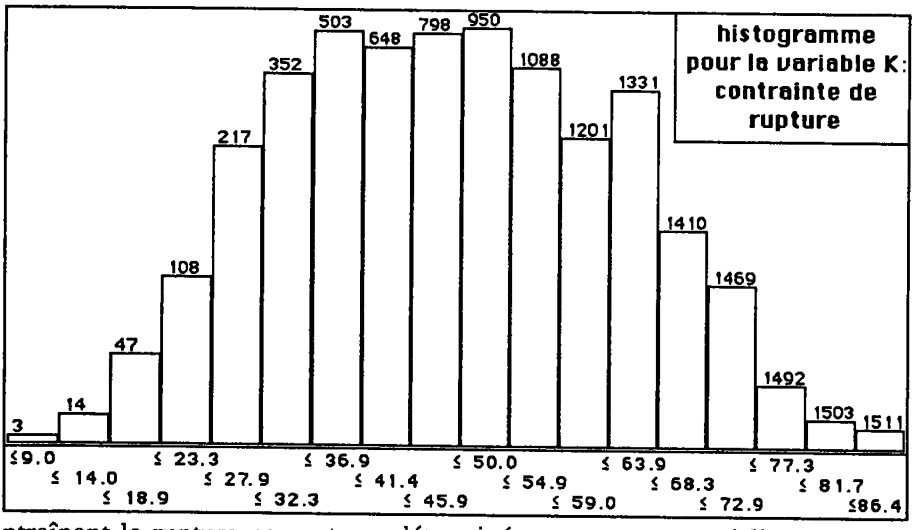
2.1 Le problème et les données

Les données considérées ici sont extraites d'un vaste corpus relatif aux propriétés physiques et mécaniques du bois. Nous avons, par ailleurs, consacré à ce corpus de multiples analyses dont nous nous réservons de rendre compte ultérieurement. Dans le présent article notre but est seulement de suivre les étapes d'une analyse fondée sur le tableau de BURT associé à un codage barycentrique.

L'ensemble I des individus comprend 1511 échantillons de bois d'épicéa issus de la forêt française. Ces échantillons ont été minutieusement décrits; et l'étude mécanique en a été poursuivie jusqu'à une épreuve destructrice afin de mesurer la *Contrainte de rupture en flexion*, K.

On souhaiterait, dans l'avenir, estimer K en se bornant à des mesures et à des essais non destructifs.

Certes, c'est là poser un problème insobluble; car, tandis que les propriétés élastiques se manifestent dans des processus à peu près réversibles et déterministes; la rupture est un phénomène essentiellement aléatoire. Même si une même pièce ne peut être brisée deux fois, on admet que les contraintes



entraînant la rupture ne sont pas déterminées; en tout cas, qu'elles ne le sont pas par les propriétés macroscopiques, mais résultent de particularités de la microstructure. Il est donc certain, *a priori*, que la *Contrainte de rupture* ne pourra être exprimée, par une formule de régression, sans un fort résidu d'erreur.

Dans la présente étude, on se propose d'analyser les variations conjointes de K et deux variables physiques {L J} qui, sur notre échantillon, lui sont notablement corrélées:

L : *Module d'élasticité en flexion* : corr(K, L) = .755 ;

J : *Masse volumique à* 12° : corr(K, J) = .468 ;

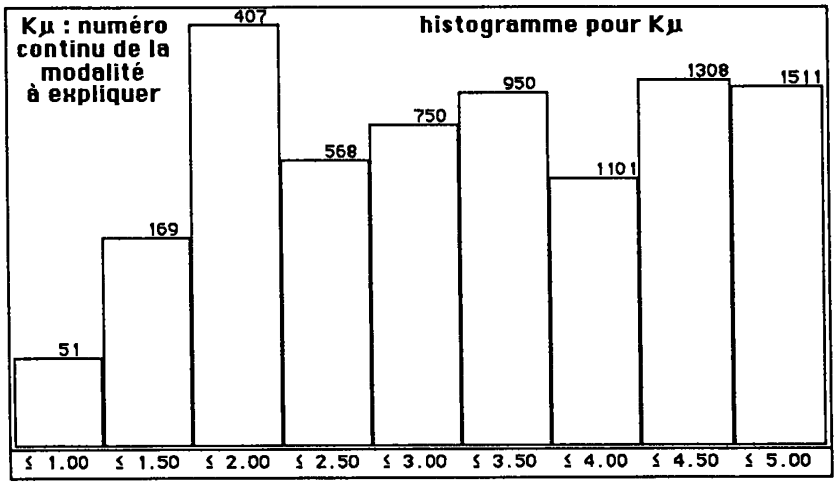
et sont corrélées entre elles: corr(J , L) = .590 .

Dans la suite, nous supposerons que les données constituent un tableau 1511 × 3, rangé dans un fichier appelé 'Disq:bois'.

2.2 Codage des variables et création de tableaux

2.2.1 La variable à expliquer

En observant les données de base par le programme 'zrang', on voit que l'histogramme de la variable à expliquer K est une courbe en cloche aplatie au sommet: peut-être cet aplatissement s'explique-t-il par le caractère aléatoire de la rupture; laquelle, pour un échantillon donné, se produirait équiprobablement sur un intervalle assez large relativement à la dispersion de ses bornes sur l'ensemble I des 1511 échantillons étudiés.



Quand on utilise un codage logique, en (0,1), on a coutume de découper la variable à expliquer en un grand nombre de classes, e.g. 10, afin de ne pas en compromettre - *a priori* - l'estimation.

Avec le codage barycentrique, qui conserve l'information, il suffit de choisir un nombre de pivots assez élevé pour déceler le glissement des associations de K avec les modalités explicatives.

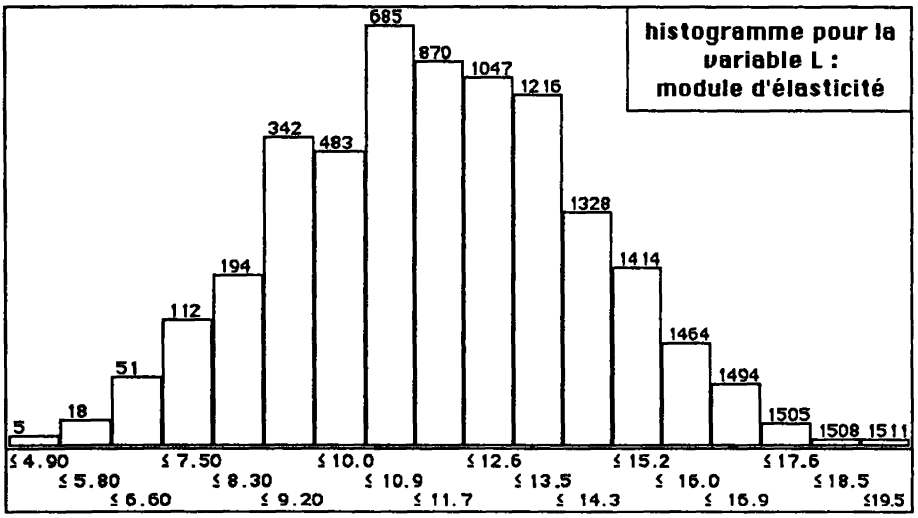
Ici, au vu de l'histogramme, on a placé 5 pivots; définis, dans le dialogue du programme de découpage 'zrang', par leur **rang** rK (lequel varie de 1 à 1511) au sein de l'ensemble I; soit:

rK< (50) ; rK≤ (400) ; rK≈ (750) ; rK≥ (1100) ; rK> (1450) ;

(les valeurs **numériques** de K se lisent sous les créneaux l'histogramme des mesures brutes; ainsi que dans le tableau de découpage des variables, §2.2.3).

Ainsi, le numéro continu de la modalité, Kμ, est un nombre variant de 1 à 5. Les pivots extrêmes ont été choisis de telle sorte que, les abords gauche et droits de la courbe en cloche (de l'histogramme de la variable donnée K) étant codés respectivement Kμ=1 et Kμ=5, l'étalement (de Kμ) de 1 à 5 corresponde (sur ce même histogramme de K) à un massif dont la hauteur ne varie pas plus que dans un rapport de 1 à 3; et les cinq pivots sont équidistants quant au rang.

L'histogramme du numéro continu Kμ est lui-même plat (si l'on met à part notre 1-er créneau; qui ne comprend que les 50 individus codés Kμ=1). Et, dans la suite, l'analyse confirmera la régularité de la distribution continue obtenue (cf. §2.3.2, croisement de Kμ avec le facteur F1).



2.2.2 Les variables explicatives

Pour la variable explicative L, l'histogramme, sans évoquer de près le modèle normal, n'a pas, comme celui de K, un sommet très plat.

On a placé 4 pivots; définis par leurs **rangs**, rL, en suivant les mêmes principes que pour ceux de K; soit:

$$rL< (50)\ ;\ rL\leq (500)\ ;\ rL\geq (1000)\ ;\ rL> (1450)\ ;$$

Le découpage de J est tout analogue à celui de L. Les valeurs numériques des pivots sont dans le listage, en tête du §2.2.3.

2.2.3 Les tableaux créés

```
données des épicéa
Disq:boisDcodx: bornes pour le découpage des variables
le nombre des variables est 3
J a  4 modalités dont les sigles et valeurs pivot sont
   J<   J≤   J≥   J>            372   426   460   519
K a  5 modalités dont les sigles et valeurs pivot sont
   K<   K≤   K≈   K≥   K>        19    34    44.5  55.8  70.9
L a  4 modalités dont les sigles et valeurs pivot sont
   L<   L≤   L≥   L>             6.6  10.2  12.4  15.7
```

Pour définir le codage adopté, le programme de découpage 'zrang' a créé un listage auxiliaire 'Disq:boisDcodx' où les bornes ou pivots sont spécifiées par des valeurs (et non des rangs au sein de I). D'une part, ce fichier permet de découper, sans reprendre le dialogue à l'écran, tout fichier complémentaire, comportant des individus décrits par les mêmes variables. D'autre part il sert au programme 'zBurt' pour créer le tableau de BURT.

Le tableau 1511 × 3 des numéros continus de modalités est rangé, par 'zrang', dans un fichier 'Disq:boisS'; et le tableau éclaté k_{IM}, 1513 × 13 (13 est le nombre total des modalités de {J, K, L}), dans 'Disq:boisQ'. À partir du fichier 'Disq:boisS', et compte tenu des informations qu'apporte 'Disq:boisDcodx' relativement à l'ensemble des modalités, le programme, 'zBurt', crée le tableau de BURT (comme on l'a expliqué au §1.3); et le range dans un fichier 'Disq:boisB'.

```
données des épicéa: tableau de BURT
 13013
```

	J<	J≤	J≥	J>	K<	K≤	K≈	K≥	K>	L<	L≤	L≥	L>
J<	140	74	0	0	61	78	49	24	2	87	95	29	2
J≤	74	369	86	0	88	165	139	103	33	103	240	155	30
J≥	0	86	389	72	37	105	133	167	105	35	155	234	123
J>	0	0	72	150	8	30	44	66	74	7	36	88	92
K<	61	88	37	8	134	60	0	0	0	107	75	11	1
K≤	78	165	105	30	60	259	59	0	0	86	201	82	10
K≈	49	139	133	44	0	59	245	60	0	27	158	148	32
K≥	24	103	167	66	0	0	60	241	59	11	79	177	93
K>	2	33	105	74	0	0	0	59	156	1	14	88	112
L<	87	103	35	7	107	86	27	11	1	153	78	0	0
L≤	95	240	155	36	75	201	158	79	14	78	365	84	0
L≥	29	155	234	88	11	82	148	177	88	0	84	348	75
L>	2	30	123	92	1	10	32	93	112	0	0	75	172

Dans le tableau de BURT, 13 × 13, B_{MM}, publié ici, les valeurs ont été arrondies à l'entier le plus proche.

On remarquera que les 3 blocs diagonaux de B_{MM}, sans être réduits à leur diagonale comme dans le cas usuel du codage en (0,1), sont chacun des tableaux tridiagonaux: en effet, pour une variable donnée, il ne peut y avoir de cooccurrence qu'entre deux modalités consécutives.

Les blocs extradiagonaux, rendant compte des cooccurrences entre modalités de variables distinctes {j ≠ jp}, n'ont pas de case nulle; mais du fait des corrélations entre variables, les coins supérieur droit et inférieur gauche sont peu chargés.

2.3 Analyse du tableau de BURT et régression

2.3.1 Les ensembles représentés

Dans l'analyse du tableau de BURT, B_{MM}, figurent en principal, d'une part, le bloc de 5 colonnes {K< K≤ K≈ K≥ K>}, des modalités de la variable à expliquer; d'autre part, l'ensemble de 8 lignes {J< J≤ J≥ J> L< L≤ L≥ L>} des modalités des variables explicatives.

À l'analyse du tableau de BURT, le tableau éclaté k_{IM} est adjoint en supplément de deux manières différentes: d'une part, comme un ensemble Ia de lignes supplémentaires; d'autre part, comme un ensemble Ib de colonnes supplémentaires. Ainsi chaque individu i (échantillon de bois d'épicéa) est projeté deux fois: d'une part, comme ia, d'après le profil de la variable à expliquer, K; d'autre part, comme ib, d'après le profil des variables explicatives {J L}.

Dans une application ultérieure, à des échantillons de bois ix soumis à un essai non destructif, permettant de mesurer {J L}, mais non K, on disposera seulement du point ixb; d'après lequel on devra estimer K. Dans la présente étude, disposant à la fois de ia et ib, on cherche, en bref, à apprécier dans quelle mesure la connaissance de ib indique la place de ia.

Voici une manière de procéder. On crée un tableau Ka × M (à 13 colonnes) dont chaque ligne correspond à une valeur déterminée de Kμ, depuis 1 jusqu'à 5, par pas successifs de 0,2: soit, en dixièmes, {10 12 14 16 ... 40 42 44 46 48 50}:

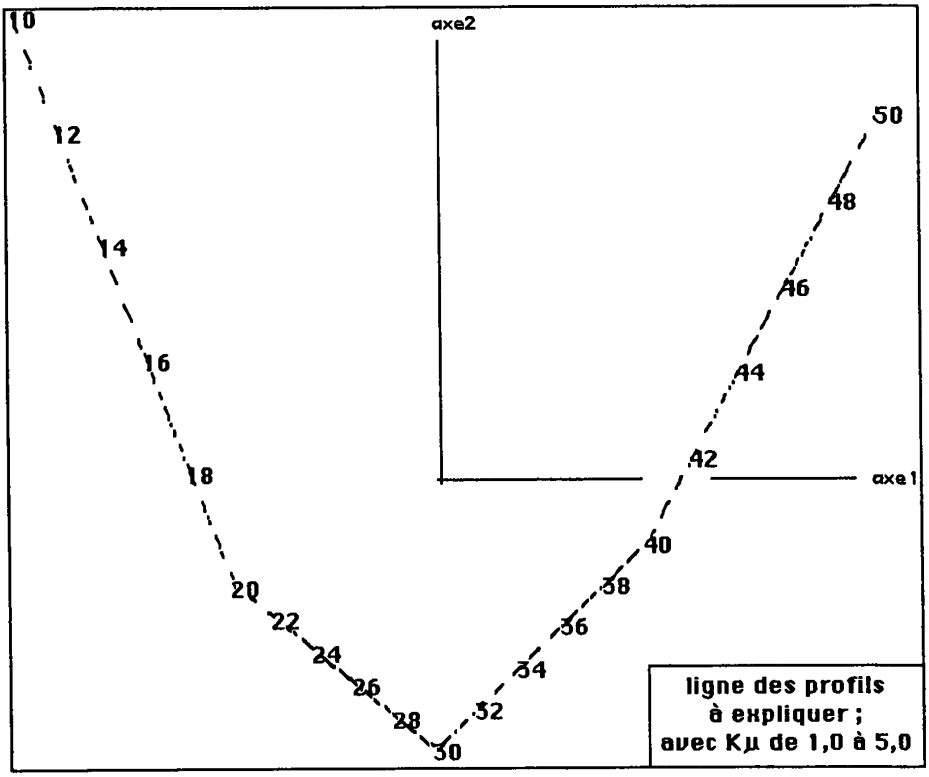
```
données des épicéa (variable K découpée)
  13  J<  J≤  J≥  J>  K<  K≤  K≈  K≥  K>  L<  L≤  L≥  L>
10    0   0   0   0 100   0   0   0   0   0   0   0   0
12    0   0   0   0  80  20   0   0   0   0   0   0   0
.....................................................................
44    0   0   0   0   0   0   0  60  40   0   0   0   0
46    0   0   0   0   0   0   0  40  60   0   0   0   0
48    0   0   0   0   0   0   0  20  80   0   0   0   0
50    0   0   0   0   0   0   0   0 100   0   0   0   0
```

Dans ce tableau, destiné à être adjoint à l'analyse de B_{MM} comme un ensemble Ka de lignes supplémentaires, seules comptent les valeurs du bloc K, le reste a été simplement mis à zéro.

Supposons l'ensemble Ka ainsi représenté: on a, pour chaque individu i, une estimation de Kμ(i) d'après {Jμ(i) Lμ(i)}, i.e. d'après ib, en cherchant, dans le chapelet des points de Ka, de 10 à 50, celui qui est le plus proche de ib: ce qui est un problème usuel de discrimination traitable par le programme 'discri' (cf. *infra*).

2.3.2 Résultats de l'analyse

```
données des épicéa
trace  :   3.468e-1
rang   :      1     2     3     4
lambda :   2876   531    60     1   e-4
taux   :   8292  1530   173     4   e-4
cumul  :   8292  9823  9996 10000   e-4
```

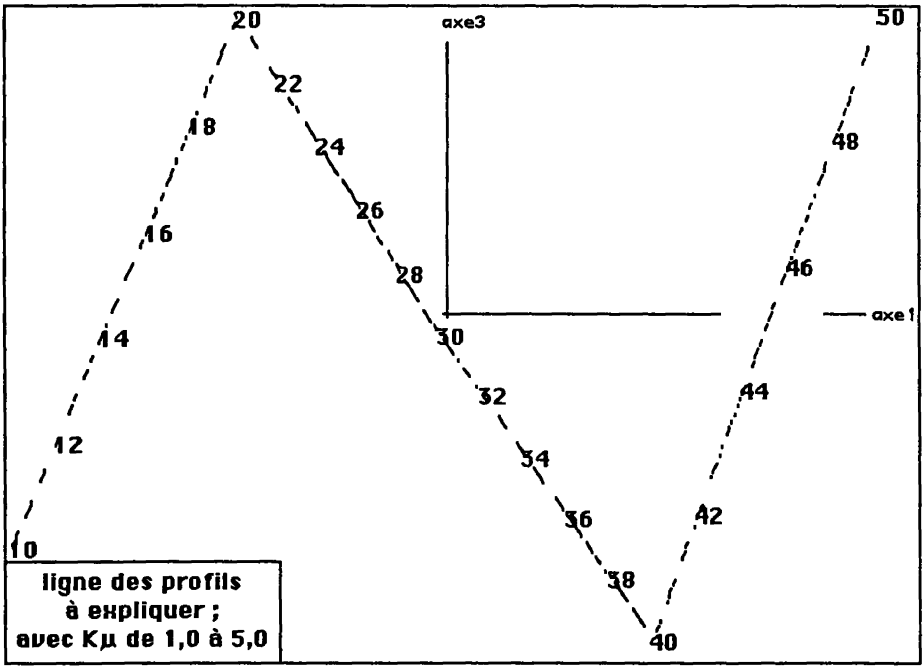


Avec 5 colonnes principales, on a seulement 4 facteurs non triviaux; dont le 1-er se détache nettement; sans que toutefois l'étalement de F2 soit négligeable sur les graphiques.

Considérons d'abord, dans les plans (1,2) et (1,3), la représentation des seuls ensembles Ia (profils de la variable à expliquer, K, pour les individus réels); et Ka (profils fictifs).

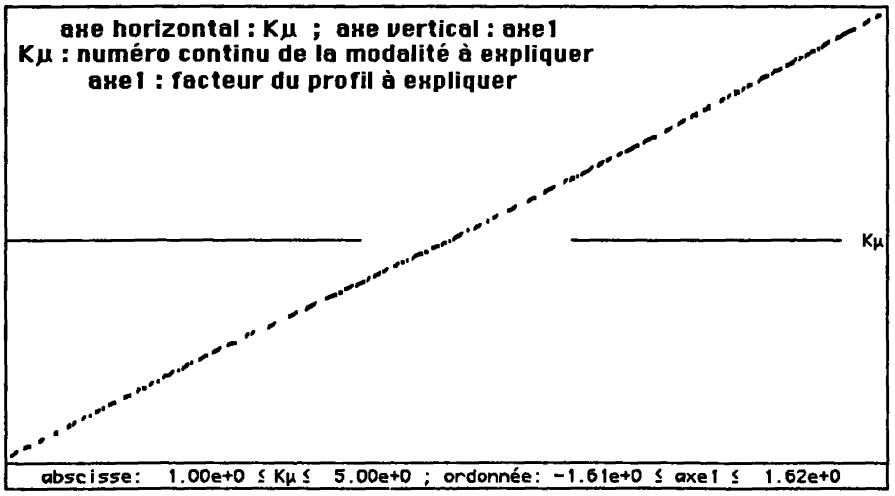
Chacun des éléments ia de Ia étant représenté par un simple tiret, tandis que ceux de Ka figurent par leur sigle (nombre de 10 à 50), on a, pour image des nuages, une ligne brisée, graduée par des nombres.

En effet, d'après le principe barycentrique, un point ia, dont la ligne (dans le tableau k_{IM}) ne comporte, dans son profil sur K, que deux modalités non nulles (disons, par exemple, la 3-ème, K≈ ; et la 4-ème, K≥), se projette (au coefficient $1/\sqrt{}$(lambda) près) comme un barycentre de ces modalités; donc aussi, en un point du segment joignant les projections des points correspondants de Ka (points 30 et 40, dans l'exemple choisi).



Dans les plans (1,2) et (1,3), l'ensemble Ia dessine une ligne brisée, jalonnée par les valeurs rondes de Ka.

En croisant la coordonnée F1(ia), notée axe1, avec la valeur de la modalité continue Kμ(i), on obtient une ligne qui diffère à peine d'un segment

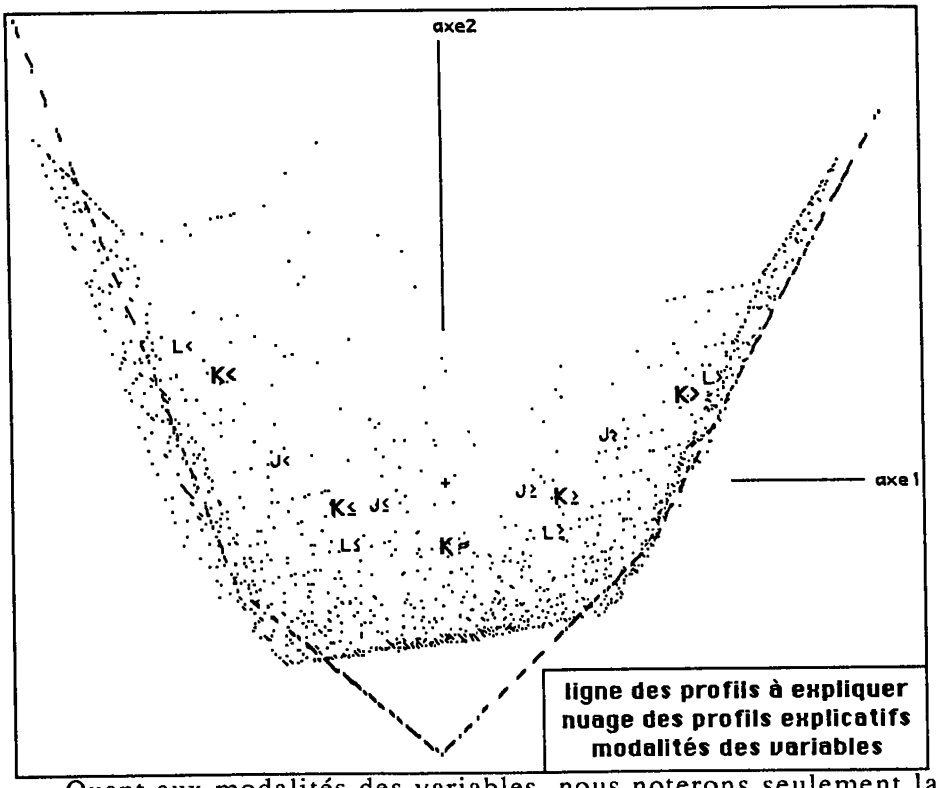


de droite; impression que confirme un calcul de corrélation:

corr(Kμ, axe1) = .9998 .

Ainsi qu'on l'annonçait au §2.2.1, cette linéarité, résulte du choix des bornes.

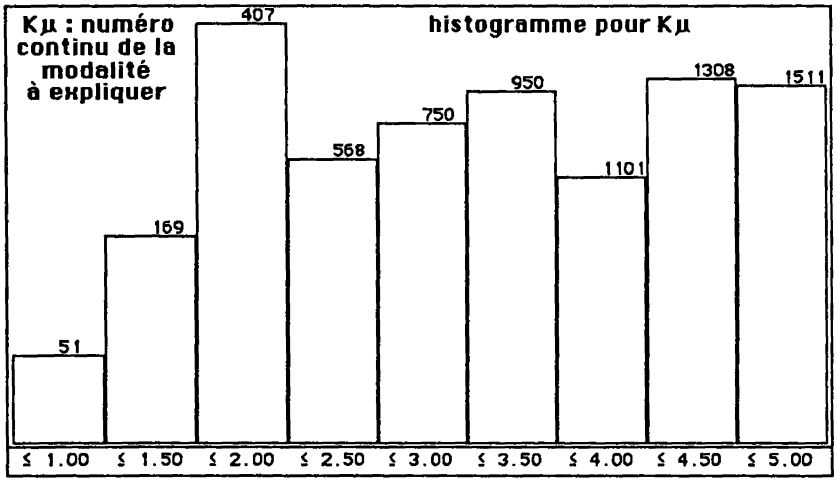
Considérons maintenant un graphique du plan (1,2) où figurent, d'une part, les deux ensembles principaux (colonnes, i.e. modalités de la variable à expliquer, marquées en gras; et lignes, modalités des variables explicatives); et, d'autre part, les deux projections de l'ensemble I mis en supplément: Ia, ligne dessinée par des tirets, comme sur les graphiques déjà vus; et Ib, véritable nuage de points, où l'image ib, de chaque individu est déterminé par les valeurs des variables explicatives.

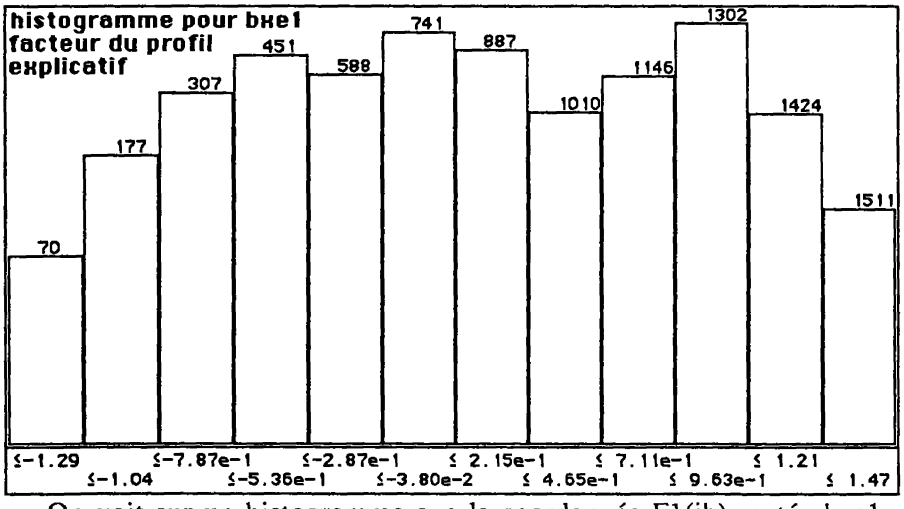


Quant aux modalités des variables, nous noterons seulement la disposition régulière des trois chapelets afférents à {J, K, L}; L étant plus étalé que J, ce qui correspond au fait que corr(K, Lμ) est supérieur à corr(K;Jμ); cf. supra, §2.1.

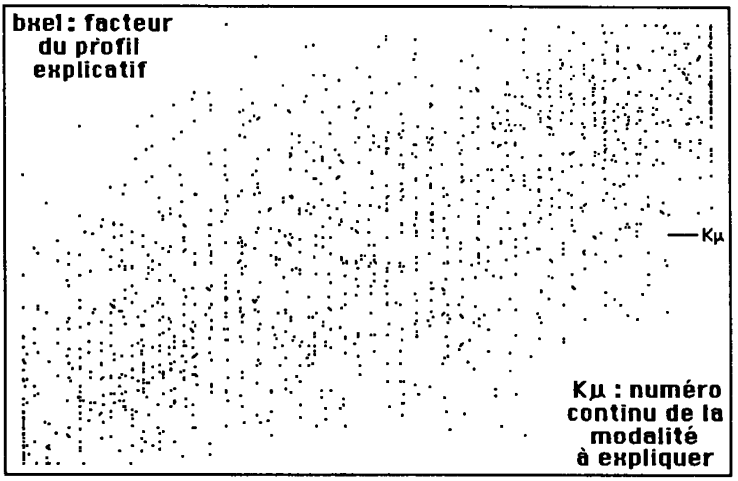
Si l'on applique à la représentation de Ib le même principe barycentrique qu'à celle de Ia, on trouve non plus une suite de segments, mais un dallage de quadrilatères, dont certains se devinent sur le graphique. Expliquons cette configuration sur un exemple: soit Jμ(i) = 1,7; Lμ(i) = 4,1 ; alors le point ib se projettera dans un quadrilatère dont les sommets correspondent aux 4 points ib pour lesquels on a:

{Jμ=1 ; Lμ=4} {Jμ=1 ; Lμ=5} {Jμ=2 ; Lμ=4} {Jμ=2 ; Lμ=5}





On voit sur un histogramme que la coordonnée F1(ib), notée bxe1, a, comme la coordonnée F1(ia), notée axe1, une distribution régulière.



abscisse: 1.0 ≤ Kμ ≤ 5.0 ; ordonnée: -1.54 ≤ bxe1 ≤ 1.47
plan croisant Kμ (axe horizontal) et bxe1 (axe vertical)

ValSup	1.00	1.50	2.00	2.50	3.00	3.50	4.00	4.50	5.00
-1.29e+0	24	23	18	3		1	1		
-1.04e+0	12	25	31	14	12	8	4		
-7.87e-1	7	23	55	15	15	10	5		
-5.36e-1	4	21	35	20	27	19	12	5	1
-2.87e-1	1	9	29	30	17	26	10	11	4
-3.80e-2	1	9	27	21	27	29	20	15	4
2.15e-1		6	20	20	24	30	17	17	12
4.65e-1	1	1	11	15	17	25	19	22	12
7.11e-1			5	8	22	27	19	29	26
9.63e-1		1	6	11	12	13	23	48	42
1.21e+0			1	2	7	6	14	41	51
1.47e+0				2	2	6	7	19	51

tri croisant Kμ (colonnes) et bxe1 (lignes)

2.3.3 Diverses formes de régression

L'interprétation des facteurs suggère d'assimiler la variable à estimer, Kμ(i) au 1-er facteur, F1(ib)=bxe1, calculé d'après les variables explicatives. On a sur le nuage des points une corrélation manifeste; mais un calcul précis de coefficient de corrélation donne:

corr(Kμ, bxe1) = .708 .

On n'a donc pas amélioré l'estimation immédiatement fournie par la variable explicative L, prise seule (cf. §2.1).

Afin de prendre en compte non seulement bxe1=F1(ib), mais aussi bxe2=F2(ib), on a cherché, dans le plan (1,2), le point du chapelet Ka qui est

le plus proche de la projection de ib. Le programme 'discri' donne, sous la forme d'un tableau à une colonne, le numéro du point de Ka ainsi obtenu pour chaque i: par exemple, s'il s'agit du point désigné par le sigle '10' le numéro est 1; pour '12', c'est 2;... ; pour 50, c'est 21. Donc, en calculant la corrélation entre ce numéro, $n\mu$, et $K\mu(i)$, on appréciera l'efficacité de ce nouveau procédé de régression; il vient:

$$corr(K\mu, n\mu) = .701 \ ;$$

il n'y a donc pas d'amélioration. En fait, bien qu'avec $n\mu$ deux facteurs soient pris en compte, le résultat ne diffère pas notablement de celui obtenu avec un seul; car, de façon précise, on a:

$$corr(bxe1, n\mu) = .976 \ ;$$

3 Conclusions et perspectives

Il n'y a pas lieu de conclure à l'inefficacité des méthodes de régression: car, avec une variable explicative prédominante, L, les données ne s'y prêtaient pas. Nous avons toutefois atteint notre but qui était de montrer, sur un exemple aussi simple que possible, l'enchaînement des calculs avec un tableau de BURT généralisé; ainsi que la régularité parfaite des estimations obtenues en conjuguant la précision numérique, caractéristique d'une régression, avec l'élaboration des données, propre à l'analyse des tableaux de BURT.

Il reste à publier des applications de la méthode à des données plus complexes; pour lesquelles la meilleure régression possible doive prendre en compte plusieurs variables.