REPRESENTATIONS GRAPHIQUES D'UN TABLEAU DE CONTINGENCE

Georges LAPLACE

1. ETAPES DE LA CONSTRUCTION GRAPHIQUE

1.1 Représentation graphique non pondérée

1.1.1 Le tableau de contingence

Nous référant à notre article "Le Lien" comme mesure de l'information dans un tableau de contingence (LAPLACE 1979-1980), nous reprendrons l'exemple des trois ensembles aurignacoïdes de la Grotte Gatzarria (LAPLACD 1966), Cjn2 et Cjn1 du Protoaurignacien et Cbf de l'Aurignacien ancien, considérés au niveau des Ordres typologiques A, B, E, P, S et SE, pour dresser un tableau croisé ou tableau de contingence, les ensembles industriels étant classés selon la diachronie et les catégories typologiques étant rangées selon l'ordre dégressif de leurs effectifs globaux :

	S	A	SE	E	B	P	
Cjn2	132	93	19	0	26	0	270
Cjn1	70	25	21	3	6	0	125
Cbf	216	42	76	52	8	0	394
	418	160	116	55	40	0	789

1.1.2 Le tableau de départ

Comme nous nous proposons d'étudier les variations des catégories typologiques au niveau de chaque ensemble industriel dans la totalité des tableaux des données, nous calculons les fréquences conditionnelles pour chacune des trois modalités du caractère ensemble lorsque l'on se trouve dans chacune des cinq modalités du caractère catégorie ainsi que les fréquences des effectifs marginaux des lignes et des colonnes, ces dernières étant figurées entre parenthèses :

	S	A	SE	E	B	P	
Cjn2	.316	.581	.164	....	.650	....	.342
Cjn1	.167	.156	.181	.055	.150	....	.158
Cbf	.517	.263	.655	.955	.200	....	.499
	1	1	1	1	1	....	
	(.530)	(.203)	(.147)	(.070)	(.051)	(....)	1

1.1.3 La construction graphique non pondérée

a - La structure du tableau de contingence suggère que l'ensemble des distributions étudiées en lignes et en colonnes puisse être représenté par un carré de côté égal à la somme des fréquences marginales, c'est-à-dire à l'unité. Nous fixerons à dix centimètres la longueur de ce côté afin d'obtenir une précision suffisante, soit un individu pour environ 12 millimètre carré.

b - Une première division de la surface globale du carré sera effectuée selon les valeurs des fréquences marginales des catégories typologiques portées entre parenthèses (Figure 1). On notera la disparition de la catégorie P dont l'effectif est nul.

c - Quant à la division de la surface globale selon les ensembles industriels, il est évidemment impossible de procéder pour l'obtenir comme précédemment. C'est donc colonne par colonne que l'on représentera la répartition des ensembles (Figure 2). Ainsi chaque surface du graphique correspont-elle à une case du tableau de contingence dont l'effectif lui est proportionnel.

d - La Figure 2 étant peu lisible, nous isolerons les surfaces correspondant à chacun des ensembles industriels. On obtient ainsi, pour chaque ensemble, un profil constitué par une série de fréquences de catégories typologiques (Figure 3). De ce fait, nous pouvons mettre en rapport chaque fréquence catégorielle de chaque ensemble avec la fréquence marginale correspondante, c'est-à-dire avec la fréquence moyenne toutes catégories confondues. S'il n'existait pas de lien entre ensembles et catégories, c'est-à-dire si ces deux variables étaient indépendantes, toutes les hauteurs des fréquences catégorielles de chaque ensemble seraient identiques et chaque profil se réduirait à la juxtaposition de surfaces de même hauteur figurée par un pointillé, formant un rectangle élevé pour l'ensemble Cbf (.499), moyen pour l'ensemble Cjn2 (.342) ou bas pour l'ensemble Cjn1 (.158). Les oscillations, positives ou négatives, au-dessus ou au-dessous de ces valeurs moyennes représentent en abscisse les écarts à la moyenne et, en surface, les écarts à l'indépendance.

e - Pour visualiser les écarts à l'indépendance, il suffit de donner comme base au graphique non pas zéro mais la valeur de la moyenne de chaque ensemble toutes catégories confondues (Figure 4). Au-dessus de la moyenne se trouveront les surfaces proportionnelles aux effectifs catégoriels au-dessus de l'indépendance. Inversement, au-dessous de la moyenne sont représentées les surfaces proportionnelles aux déficits des effectifs catégoriels relativement à l'indépendance.

Ainsi, pour chaque ligne, la différence entre les fréquences catégorielles en colonnes et la fréquence marginale toutes catégories confondues donne les écarts à la moyenne, c'est-à-dire les hauteurs orientées des rectangles ayant pour base les fréquences marginales de chacune des catégories :

	S	A	SE	E	B
Cjn2	-.026	+.239	-.178	-.342	+.308
Cjn1	+.009	-.002	+.023	-.104	-.008
Cbf	+.017	-.237	+.156	+.446	-.299
	0	0	0	0	0

On en déduit les écarts à l'indépendance en faisant le produit des écarts à la moyenne, ou écarts conditionnels, par les fréquences marginales catégorielles correspondantes :

	S	A	SE	E	B	
Cjn2	-.01399	+.04848	-.02623	-.02385	+.01560	0
Cjn1	+.00479	-.00044	+.00332	-.00724	-.00043	0
Cbf	+.00921	-.04803	+.02291	+.03110	-.01518	0
	0	0	0	0	0	

On constate que pour chaque profil la somme des écarts à l'indépendance est nulle et que, pour chaque colonne catégorielle, la somme des surfaces au-dessus de la moyenne est égale à la somme des surfaces situées en-dessous.

1.1.4 Interprétation des données

Elle s'effectue à un double point de vue, celui des associations spécifiques entre lignes et colonnes et celui des similitudes entre profils.

a - Une association spécifique est une liaison privilégiée entre une catégorie typologique et un ensemble industriel. Elle nous est signalée par une surface au-dessus de la moyenne, c'est-à-dire par un écart à l'indépendance positif, dans le cas d'une association positive, ou par une surface au-dessous de la moyenne, c'est-à-dire par un écart à l'indépendance négatif, dans le cas d'une association négative. Si son importance numérique est indiquée par l'ampleur de la surface du rectangle, une hauteur élevée suffit à marquer une association spécifique même si elle ne concerne qu'un effectif de faible importance, c'est-à-dire un rectangle de base relativement réduite. Tel est le cas de la catégorie B dans les ensembles Cjn2 et Cbf. Dans les surfaces considérées, la hauteur, c'est-à-dire l'écart à la moyenne, est un indicateur de force de l'association spécifique pondéré par l'importance de l'effectif marginal de la catégorie concernée.

Ainsi pour l'ensemble Cjn2, on note l'importance de l'association positive entre cet ensemble et les catégories A et B, compensée par une association négative avec les catégories SE, E et S, cette dernière se situant près de l'indépendance. Pour l'ensemble Cjn1, les deux catégories en association positive S et SE, voisines de l'indépendance, s'opposent aux catégories en association négative E, B et SE, ces deux dernières étant à proximité de l'indépendance. Enfin, pour l'ensemble Cbf, on remarque une opposition entre les catégories en association positive E, SE et S, cette dernière étant proche de l'indépendance, et les catégories en association négative A et B.

b - Quant à la similitude entre profils, si les ensembles Cjn2 et Cjn1 n'ont en commun que l'écart à l'indépendance négatif de la catégorie E, il n'en est pas de même pour les ensembles Cjn1 et Cbf qui sont similaires à l'exclusion de la différence d'orientation des écarts à l'indépendance de la catégorie E, négative dans l'ensemble Cjn1 et positive dans l'ensemble Cbf.

1.1.5 Remarques

a - Notons immédiatement que l'on peut construire à partir du tableau de contingence un autre graphique où, les catégories typologiques étant considérées comme profils, on calculerait les écarts entre fréquences conditionnelles en ligne et les fréquences marginales correspondantes toutes lignes confondues.

b - Notons encore que, pour faciliter l'interprétation des données, on peut changer dans les graphiques aussi bien l'ordre des profils que l'ordre de la série des profils.

c - Notons enfin que, grâce à l'écart à l'indépendance, concept statistique de valeur théorique, on dispose d'une méthode de représentation des données

qui met en lumière les traits pertinents du point de vue statistique. Cette notion joue un rôle capital dans l'élaboration de la contingence quadratique ou indicateur d'écart KHI-DEUX, somme des écarts quadratiques réduits, c'est-à-dire d'effectifs en écart à l'indépendance pondérés dans un tableau de contingence.

1.2 Représentation graphique pondérée

1.2.1 Le tableau de départ de la pondération

Les profils en lignes ont été établis en calculant les écarts entre les fréquences de cases en colonnes et les fréquences marginales toutes colonnes confondues correspondantes. De même, nous pourrions construire les profils en colonnes en calculant les écarts entre les fréquences de cases en lignes et les fréquences marginales toutes lignes confondues correspondantes. Ce sont ces fréquences que nous figurons sur le tableau suivant, les fréquences marginales toutes colonnes confondues étant portées pour mémoire entre parenthèses

	S	A	SE	E	B		
Cjn2	.489	.344	.070	....	.098	1	(.342)
Cjn1	.560	.200	.168	.024	.048	1	(.158)
Cbf	.107	.107	.193	.132	.020	1	(.499)
	.530	.203	.147	.070	.051		1

1.2.2 La construction graphique pondérée

En effectuant les différences entre les fréquences de cases pour chaque ligne aux fréquences marginales correspondantes toutes lignes confondues on obtient les écarts à la moyenne :

	S	A	SE	E	B	
Cjn2	-.041	+.142	-.077	-.070	+.046	0
Cjn1	+.030	-.003	+.021	-.046	-.003	0
Cbf	+.018	-.096	+.046	+.062	-.030	0

La valeur absolue de chacun de ces écarts représentant la valeur pondérée de la base de chacun des rectangles composants des trois profils en lignes, nous pouvons construire les graphiques pondérés en substituant simplement les bases pondérées aux bases non pondérées (Figure 5).

Notons que la pondération peut être amplifiante, réductrice ou égale à l'unité. Dans notre exemple, toutes les pondérations sont réductrices à l'exception de celle de la catégorie E dans l'ensemble Cjn2 où elle est égale à l'unité car l'effectif de cette case est nul.

1.2.3 Les écarts à l'indépendance

En faisant le produit des hauteurs de chaque rectangle par la longueur de sa base pondérée on obtient les écarts à l'indépendance pondérés. Dans le tableau suivant nous présentons ces écarts ainsi que les sommes marginales de leurs valeurs absolues :

	S	A	ES	E	B	
Cjn2	- 0,00108	+ 0,03386	- 0,01368	- 0,02385	+ 0,01404	0,08651
Cjn1	+ 0,00027	- 0,00001	+ 0,00047	- 0,00475	- 0,00002	0,00552
Cbf	+ 0,00032	- 0,02278	+ 0,00715	+ 0,02778	- 0,00910	0,06713
	0,00167	0,05665	0,02130	0,05638	0,02316	0,15916

Si nous comparons ce tableau au tableau du "Lien" (LAPLACE 1979-1980), calculé pour le même tableau de contingence, nous constatons leur parfaite identité : mêmes contributions relatives signées de cases, mêmes contributions relatives en lignes et en colonnes et, bien entendu, mêmes sommes globales.

2. FORMALISATION DE LA DEMARCHE

2.1 Notations

Soit un tableau de contingence dont les lignes sont indicées par i et les colonnes par j. On adoptera les notations suivantes (LAPLACE 1979-1980) :

2.1.1 Effectifs

n_{ij} ... effectif d'une case quelconque

n_i ... effectif marginal d'une ligne ..... $n_i = \sum_i n_{ij}$

n_j ... effectif marginal d'une colonne ... $n_j = \sum_j n_{ij}$

n ... effectif global ..... $n = \sum_{ij} n_{ij}$

2.1.2 Fréquences

$$f_{ij} = \frac{n_{ij}}{n} \quad \text{avec} \quad \sum_{ij} f_{ij} = 1$$

$$f_i = \frac{n_i}{n} \quad \text{avec} \quad \sum_j f_i = 1$$

$$f_j = \frac{n_j}{n} \quad \text{avec} \quad \sum_i f_j = 1$$

2.1.3 Fréquences conditionnelles en ligne ou en colonne

$$f_j^i = \frac{f_{ij}}{f_i} = \frac{n_{ij}}{n_i} \quad \text{avec} \quad \sum_j f_j^i = 1$$

$$f_i^j = \frac{f_{ij}}{f_j} = \frac{n_{ij}}{n_j} \quad \text{avec} \quad \sum_i f_i^j = 1$$

2.2 Ecarts à la moyenne

On note e_i^j l'écart à la moyenne en ligne et e_j^i l'écart à la moyenne en colonne tels que :

$$e_i^j = f_i^j - f_i \qquad \text{et} \qquad e_j^i = f_j^i - f_j$$

On vérifie que :

$$\sum_i e_i^j = \sum_j e_j^i = 0$$

en effet $\sum_i e_i^j = \sum_i \frac{f_{ij}}{f_j} - \sum_i f_i = 1 - 1 = 0$

$$\sum_j e_j^i = \sum_j \frac{f_{ij}}{f_i} - \sum_j f_j = 1 - 1 = 0$$

Notons que pour construire les profils en lignes de la Figure 3 nous avons utilisé la différence entre chaque fréquence de case en colonne et la fréquence marginale toutes colonnes confondues correspondante, c'est-à-dire e_i^j, correspondant à la hauteur du rectangle.

2.3 Ecarts à l'indépendance

La base de chacun des rectangles étant proportionnelle à la fréquence de chacune des colonnes, c'est-à-dire à f_j, et sa hauteur étant proportionnelle à e_i^j, sa surface est proportionnelle au produit $f_j e_i^j$. Ce produit représente donc la proportion d'individus qui sont en écart à l'indépendance pour une case donnée :

$$f_j e_i^j = f_j (f_i^j - f_i) = f_j f_i^j - f_i f_j$$

or comme $f_j f_i^j = f_j \frac{f_{ij}}{f_j} = f_{ij}$ on a $f_j e_i^j = f_{ij} - f_i f_j$

La proportion d'individus $f_{ij} - f_i f_j$ en écart à l'indépendance sera notée e_{ij} :

$$e_{ij} = f_{ij} - f_i f_j$$

Pour un profil donné représentant une ligne on a $\sum_j e_{ij} = 0$

en effet $\sum_j e_{ij} = \sum_j f_{ij} - \sum_j f_i f_j = f_i - f_i = 0$

De même, si l'on considère la superposition des rectangles appartenant à différents profils mais situés sur la même colonne, on a $\sum_i e_{ij} = 0$

en effet $\sum_i e_{ij} = \sum_i f_{ij} - \sum_i f_i f_j = f_j - f_j = 0$

2.4 Ecarts à l'indépendance et indicateur d'écart

Les profils de la Figure 3 représentent les catégories typologiques en écart à l'indépendance dans chaque ensemble industriel. Par une simple modification de la base de chacun des rectangles nous allons pondérer chacune de leurs surfaces de manière à la rendre proportionnelle au KHI-DEUX de chacune des cases. En effet, l'écart à l'indépendance ayant été exprimé en construisant des profils en lignes, il serait aussi possible de l'exprimer en construisant des profils en colonnes et l'on obtiendrait alors pour la surface de chaque rectangle :

$$e_{ij} = f_i e^i_j$$

Si l'on exprime les écarts à la moyenne en fonction de l'écart à l'indépendance on a :

$$e^j_i = \frac{e_{ij}}{f_j} = \frac{f_{ij} - f_i f_j}{f_j} \quad \text{et} \quad e^i_j = \frac{e_{ij}}{f_i} = \frac{f_{ij} - f_i f_j}{f_i}$$

Le produit des deux écarts à la moyenne est égal au PHI-DEUX de chaque case, c'est-à-dire au quotient du carré de l'écart à l'indépendance par le produit des fréquences marginales, car nous travaillons sur des fréquences :

$$e^i_j e^j_i = \frac{(e_{ij})^2}{f_i f_j} = \frac{(f_{ij} - f_i f_j)^2}{f_i f_j} = \phi^2_{ij}$$

ϕ^2_{ij} étant la "contribution relative signée de case à l'inertie du nuage", si l'on se réfère à notre article relatif au "Lien" (LAPLACE 1979-1980), on constate qu'il est identique à c_{ij} défini comme "contribution de la case (i,j) à l'information apportée par le tableau f_{IJ}" :

$$\phi^2_{ij} = c_{ij} = \frac{(e_{ij})^2}{f_i f_j} = \frac{(f_{ij} - f_i f_j)^2}{f_i f_j} = \frac{X^2_{ij}}{n}$$

Ainsi, le PHI-DEUX global ϕ^2_{IJ} est-il identique au Lien (I,J) :

$$\phi^2_{IJ} = \sum_{ij} \phi^2_{ij} = \text{Lien } (I,J) = \sum_{ij} c_{ij} = \sum_{ij} \frac{X^2_{ij}}{n} = \frac{X^2}{n}$$

2.5 Conclusion

Nous disposerons désormais de deux représentations graphiques pour visualiser l'information apportée par une case, une ligne ou une colonne d'un tableau de contingence : la représentation graphique pondérée des écarts à l'indépendance et la représentation graphique du "spectre" de ligne ou de colonne (LAPLACE 1979-1980).

En conséquence, il nous semble utile de normaliser le vocabulaire en utilisant les notations suivantes :

ϕ^2_{ij} ... pour la contribution relative signée de case à l'information ;

$\sum_i \phi^2_{ij}$ ... pour la contribution relative de ligne à l'information ;

$\sum_j \phi^2_{ij}$ ... pour la contribution relative de colonne à l'information ;

$\sum_{ij} \phi^2_{ij}$ ... pour la mesure de l'information globale.

3. DISTANCE DU KHI-DEUX ET PHI-DEUX GLOBAL

Si l'on désigne par i et i' deux lignes quelconques d'un tableau de contingence, par g la ligne moyenne toutes lignes confondues (point moyen ou centre de gravité), la distance du KHI-DEUX entre i et i' est définie par les formules :

$$d^2(i,i') = \sum_j \frac{1}{f_j} (f^i_j - f^{i'}_j)^2 \quad \text{et} \quad \sum_i f_i \, d^2(i,g)^2 = \phi^2$$

avec $$d^2(i,g) = \sum_j \frac{1}{f_j} (f^i_j - f_j)^2$$

Cette distance pondérée est appelée distance du KHI-DEUX car on démontre que la somme, pondérée par la fréquence marginale de chacune des lignes, des distances de chaque ligne au profil moyen, c'est-à-dire aux fréquences marginales toutes lignes confondues, est égale au ϕ^2 global.

En effet, si pour la case c_{ij} nous portons en $f_i \, d^2(i,g)$, tiré de la seconde formule, la valeur de $d^2(i,g)$, donnée par la troisième formule, on obtient successivement :

$$c_{ij} = f_i \frac{1}{f_j} (f^i_j - f_j)^2 = f_i (f^i_j - f_j) \frac{1}{f_j} (f^i_j - f_j)$$

$$= f_i \left(\frac{f_{ij}}{f_i} - f_j\right) \frac{1}{f_j} \left(\frac{f_{ij}}{f_i} - f_j\right) = (f_{ij} - f_i f_j) \left(\frac{f_{ij}}{f_i f_j} - \frac{f_j}{f_j}\right)$$

$$= (f_{ij} - f_i f_j) \left(\frac{f_{ij}}{f_i f_j} - 1\right) = (f_{ij} - f_i f_j) \left(\frac{f_{ij} - f_i f_j}{f_i f_j}\right)$$

$$= \frac{(f_{ij} - f_i f_j)^2}{f_i f_j} = \frac{(e_{ij})^2}{f_i f_j} = \phi^2_{ij}$$

ainsi on a bien :

$$\sum_i f_i \, d^2(i,g)^2 = \sum_i c_{ij} = \sum_i \phi^2_{ij} = \phi^2 = \frac{X^2}{n} \quad \text{c.q.f.d.}$$

BIBLIOGRAPHIE

CIBOIS P. 1980 - La représentation factorielle des tableaux croisés et des données d'enquête. Laboratoire d'informatique pour les Sciences de l'homme. C.N.R.S., 451 p.

LAPLACE G. 1966 - Les niveaux castelperroniens, protoaurignaciens et aurignacien de la Grotte Gatzarria à Suhare en Pays Basque. Quartär, 17, 117-140, 4 fig., 5 tabl.

LAPLACE G. 1979-1980 - Le "lien" comme mesure de l'information dans un tableau de contingence. Dialektikê. Cahiers de Typologie analytique 1979-1980, 1-15.

VOLLE M. 1981 - Analyse des données. Economiac, Paris, 320 p.

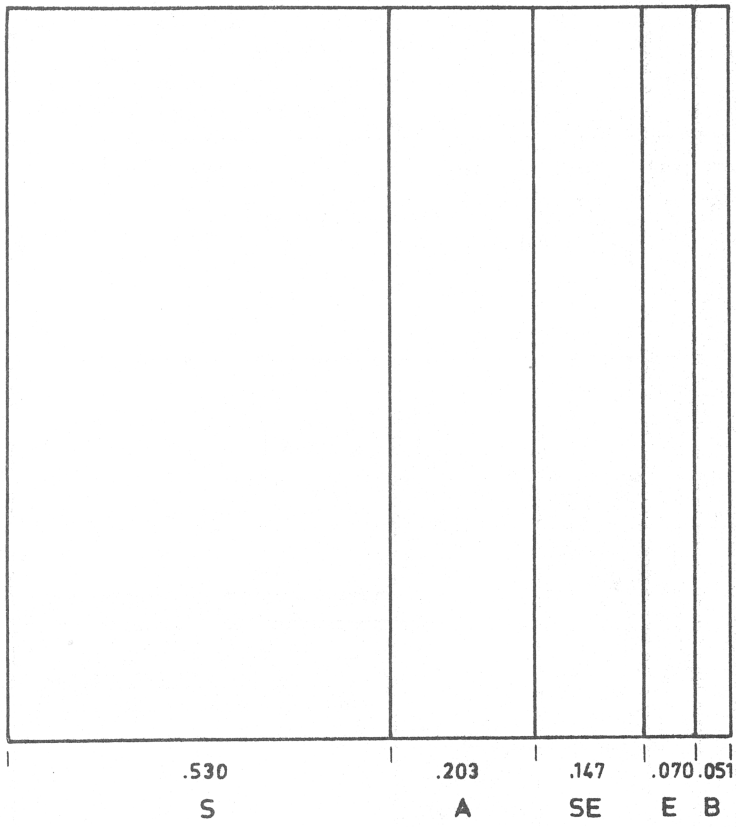


FIGURE 1

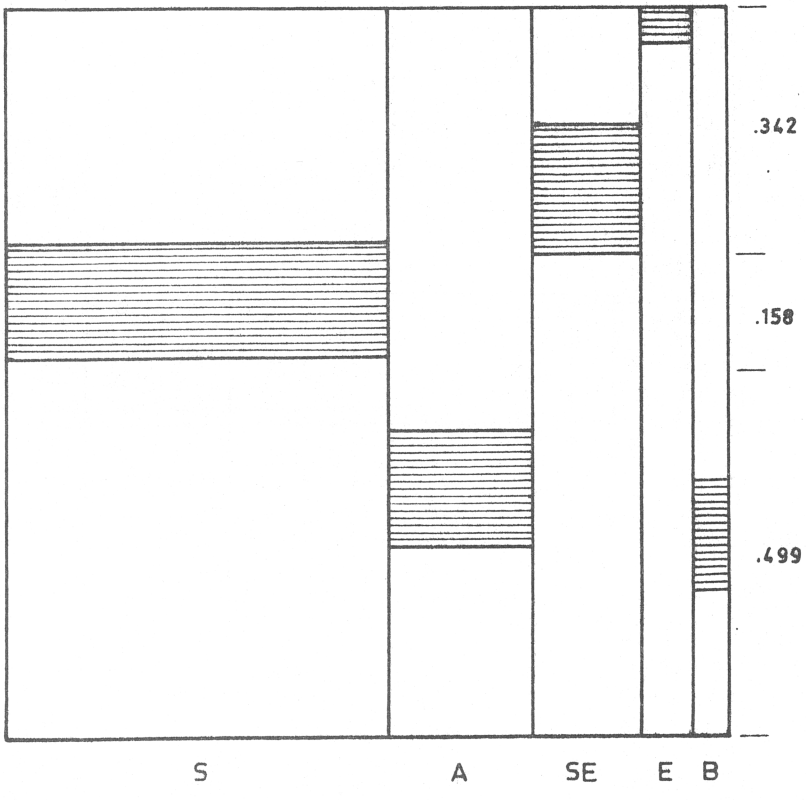


FIGURE 2

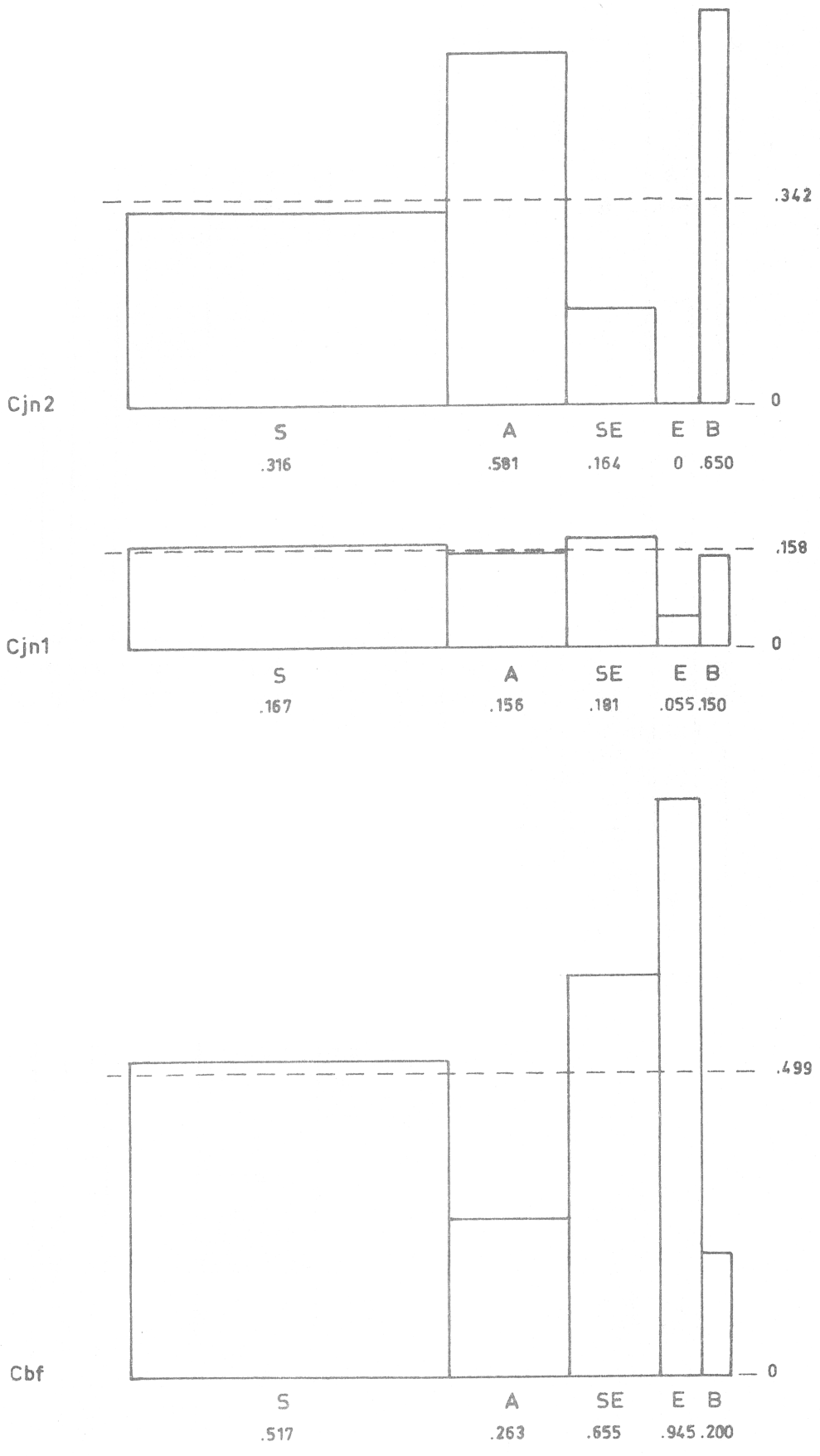

.342
Cjn2
0
S
A
SE
E
B
.316
.581
.164
0
.650
.158
Cjn1
0
S
A
SE
E
B
.167
.156
.181
.055
.150
.499
Cbf
0
S
A
SE
E
B
.517
.263
.655
.945
.200

FIGURE 3

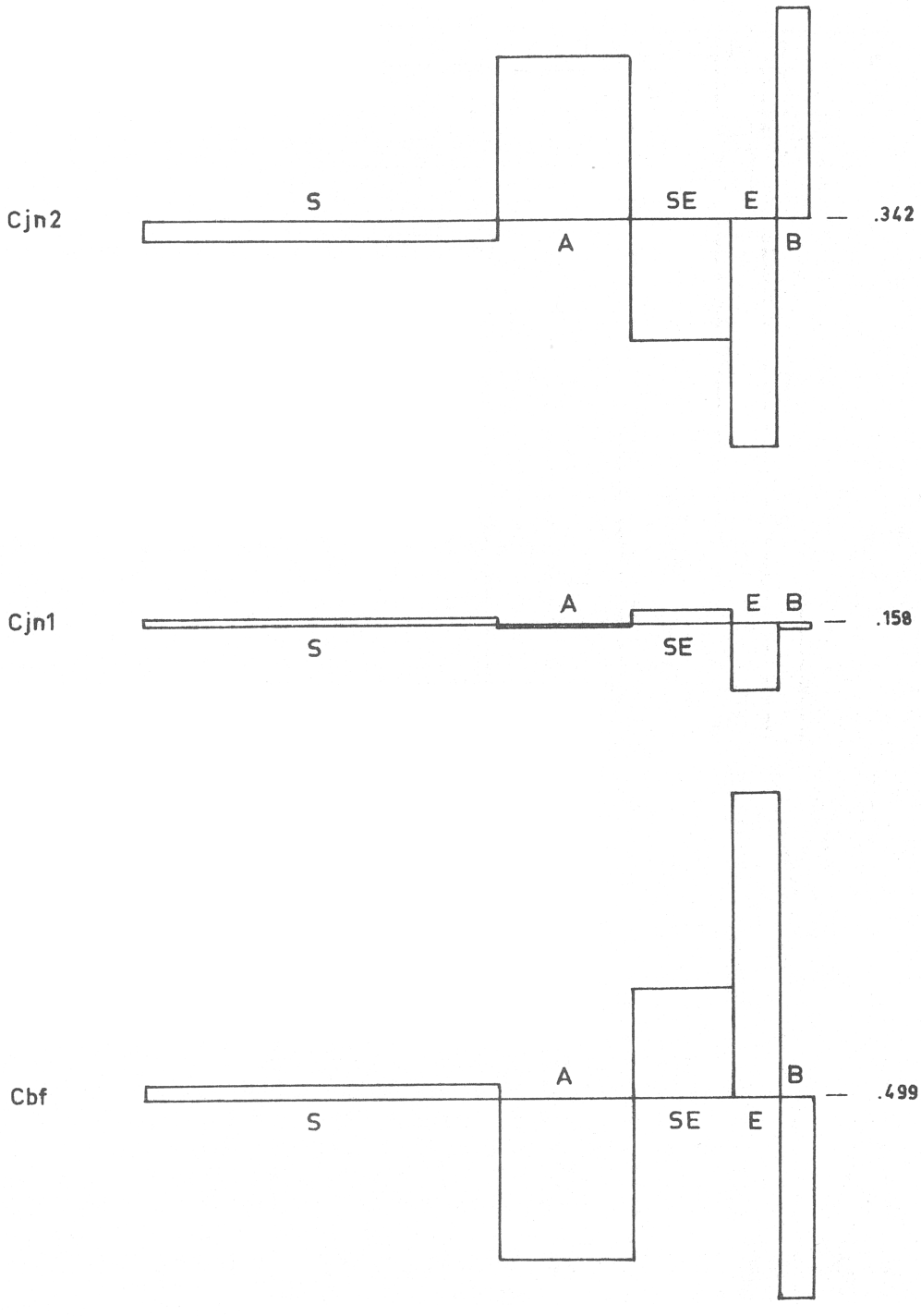
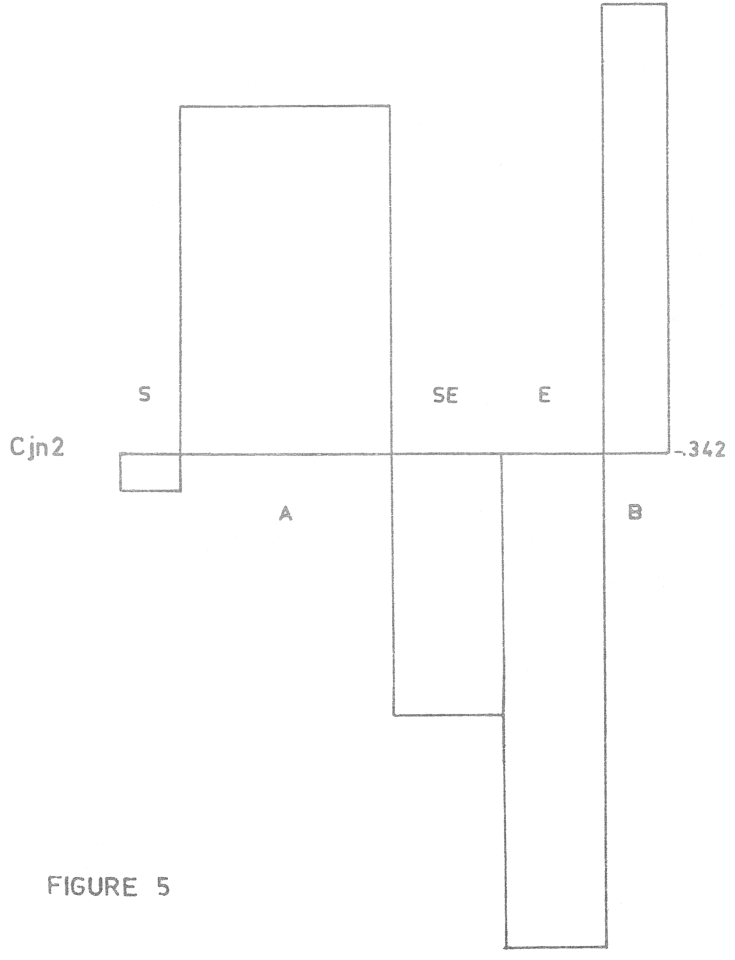
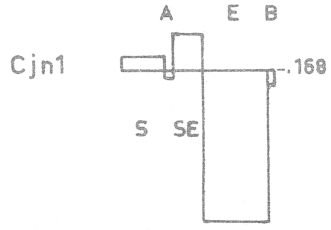


FIGURE 4





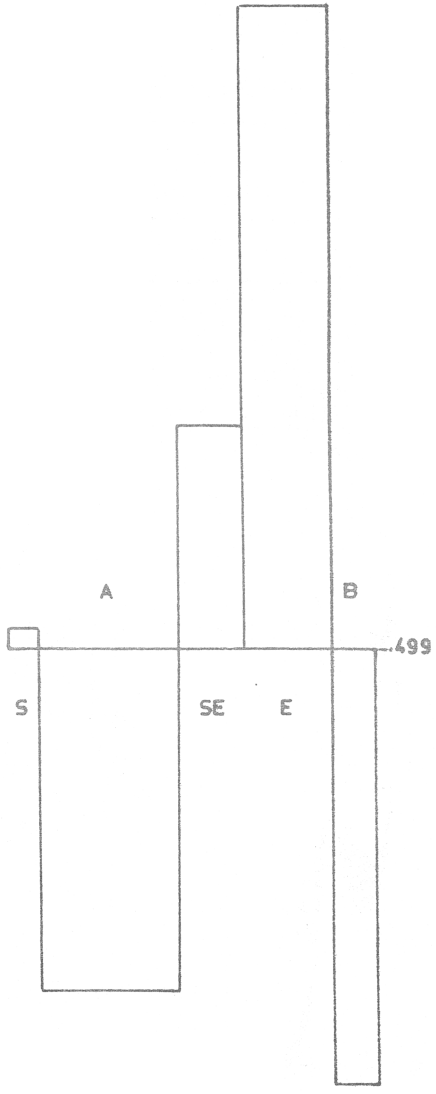


FIGURE 5

Echelle : 2/1