LES CAHIERS DE L'ANALYSE DES DONNÉES

L. LEBART

Exemple d'analyse d'un tableau dont l'une des colonnes a un poids prédominant

Les cahiers de l'analyse des données, tome 4, n° 4 (1979), p. 417-422

<http://www.numdam.org/item?id=CAD_1979__4_4_417_0>

Les Cahiers de l'Analyse des Données
Vol IV 1979 n° 4 p 417 422

EXEMPLE D'ANALYSE D'UN TABLEAU DONT L'UNE DES COLONNES A UN POIDS PRÉDOMINANT [COL. PRED. EX.]

par L Lebart (1)

0 Données

Tableau 13 x 88 des décès départementaux selon 13 causes, pour les deux sexes, et pour la classe *d'âge 45-64 ans (année 1962).*

L'analyse de ce tableau donnerait une représentation de la forme des mortalités départementales (importance de chacune des rubriques dans le total des décès), sans nous renseigner sur le niveau de cette mortalité (importance des causes de décès par rapport à la population départementale).

Nous ferons donc une deuxième analyse, en adjoignant aux 13 colonnes du tableau précédant une 14-ème colonne où figurera la population de la classe d'âge étudiée.

Cette colonne est d'un ordre de grandeur tout à fait différent :

Symboles et poids de chacune des variables (effectifs des décédés)

TU	Tuberculose pulmonaire	3639	(1)
CA	Cancer	30678	(2)
CB	Cancer bronco-pulmonaire	4044	(3)
LE	Leucémies	888	(4)
DI	Diabète sucré	1417	(5)
LV	Lésions vasculaires intra-crâniennes	9847	(6)
CO	Coeur	18223	(7)
GR	Grippe	552	(8)
PN	Pneumonie	1109	(9)
AL	Alcoolisme	2812	(10)
CI	Cirrhose du foie	7659	(11)
SU	Suicide	3184	(12)
AC	Accident	5975	(13)
PO	Population de la classe d'âge	10.479.000	(14)

(1) *Maître de recherches C.N.R.S.*

1 *Analyse du tableau homogène (Sans la colonne population)*

La représentation simultanée, figure 1, dans le plan des deux premiers facteurs, nous donne les proximités entre variables et entre départements repérés par leurs numéros minéralogiques.

Les six premières valeurs propres sont respectivement :

0.017432 , 0.013149 , 0.005307 , 0.004044 , 0.003280, 0.002888.

Ces valeurs propres sont déjà assez faibles, car il existe de grandes disparités entre les poids des diverses causes de décès, ainsi qu'entre les poids des divers départements.

En effet, leur somme, inertie totale du nuage, est voisine, au facteur 2Log2 près, de l'information mutuelle $H(P_{IJ},P_IP_J)$ des ensembles I et J mis en correspondance. Cette quantité est égale à $(H(P_I)+H(P_J)-H(P_{IJ})$, et est positive (CF Inf. Tab.). Comme d'autre part la quantité $H(P_J) - H(P_{IJ})$ est négative, on voit qu'il suffit que $H(P_I)$ soit très petit pour que l'inertie totale soit très faible. C'est en particulier le cas lorsque les disparités de poids sont grandes, puisque le degré d'indétermination est très faible.

A la limite,si une colonne a un poids prédominant, l'inertie totale, et par conséquent les valeurs propres, sont infiniment petites.

Interprétation : Le premier facteur, défini par la variable "alcoolisme" et par quelques variables paraissant lui être rattachées plus ou moins directement (Tuberculoses, Suicides, Cirrhoses, Accidents), prend en compte les principales causes de sur-mortalité de cette classe d'âge, en les opposant simultanément aux deux groupes d'affections qui représentent la grosse masse des décès : Tumeurs d'une part, avec le cancer, le cancer bronco-pulmonaire, la leucémie ; et maladies cardio-artérielles.Ces deux groupes sont séparés par le second facteur.

Il est intéressant de voir comment se modifie cette configuration lorsque l'on ajoute la colonne de poids quasi-infini des populations des départements.

2 *Analyse du tableau avec la 14° colonne*

La représentation simultanée n'est plus possible , puisque les points-départements sont des barycentres des points-variables, avec comme poids les valeurs des variables. Ils vont donc tous s'abîmer sur le point-population ! La figure 2 représente les points-variables sur les premiers facteurs, et la figure 3, à une autre échelle , les points-départements. Pour fixer les idées, le format réel de la figure 3 a été tracé sur la figure 2, autour de l'origine.

Les valeurs propres sont pour les 6 premières :

0.00002386, 0.00001330, 0.00000835, 0.00000368, 0.00000341, 0.00000277

Il existe bien un facteur 10^3 entre ces valeurs propres et les précédentes qui est également le facteur d'échelle de la colonne rajoutée.

Le point-population, joint à l'origine donne l'axe de mortalité générale (J.P. B.).

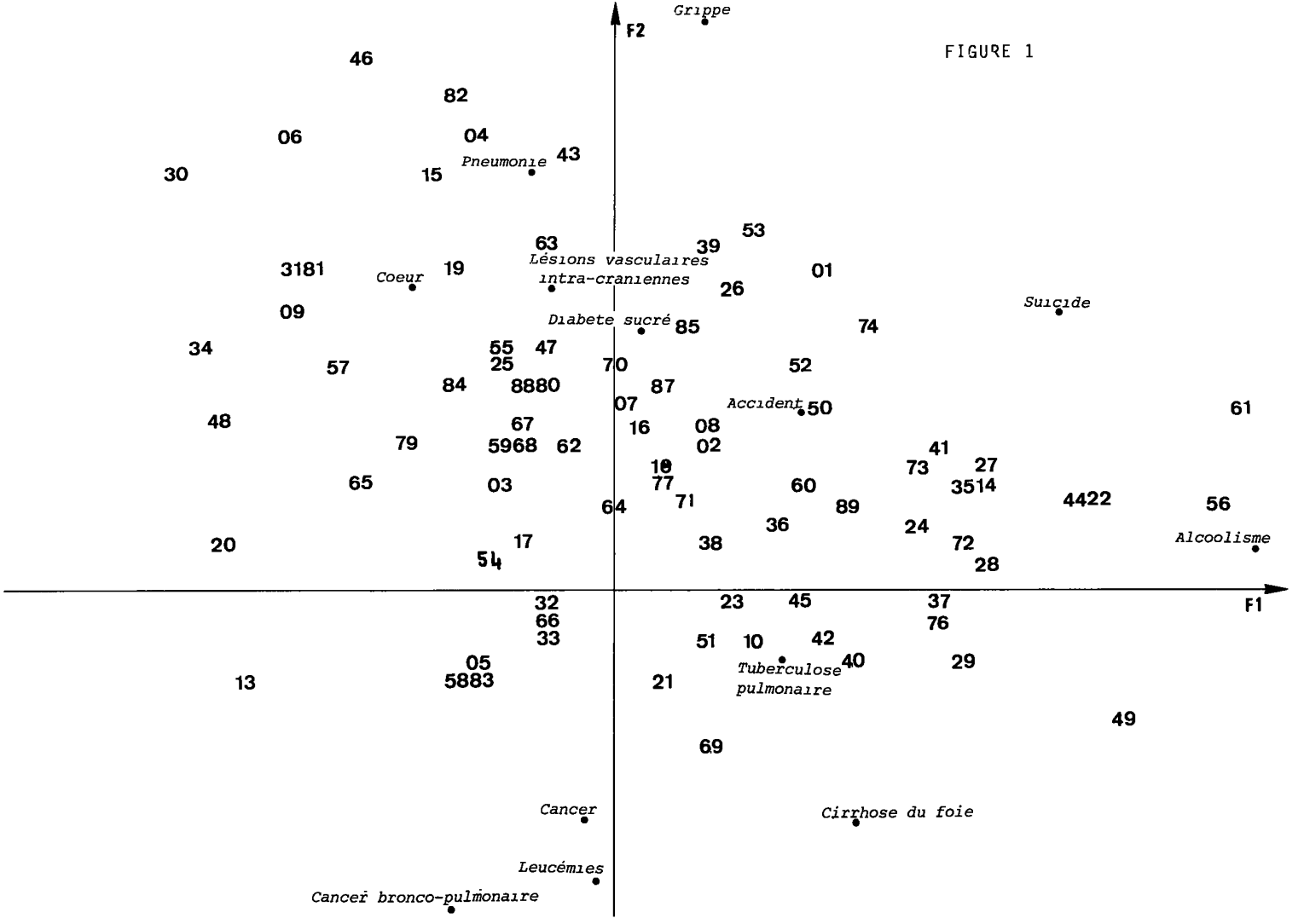

Grippe
F2
F1
46
82
06
04
Pneumonie
43
30
15
53
63
39
Lésions vasculaires
intra-crantennes
3181
Coeur
19
01
26
Suicide
09
Diabete sucré
85
74
34
55
47
25
70
52
57
84
8880
87
07
Accident
50
61
48
67
16
08
79
5968
62
02
41
10
73
27
65
03
77
60
3514
64
71
89
4422
56
36
24
20
17
38
72
Alcoolisme
54
28
32
23
45
37
66
76
33
51
10
42
05
40
29
Tuberculose
pulmonaire
13
5883
21
49
69
Cancer
Cirrhose du foie
Leucémies
Cancer bronco-pulmonaire

FIGURE 1

FIGURE 2

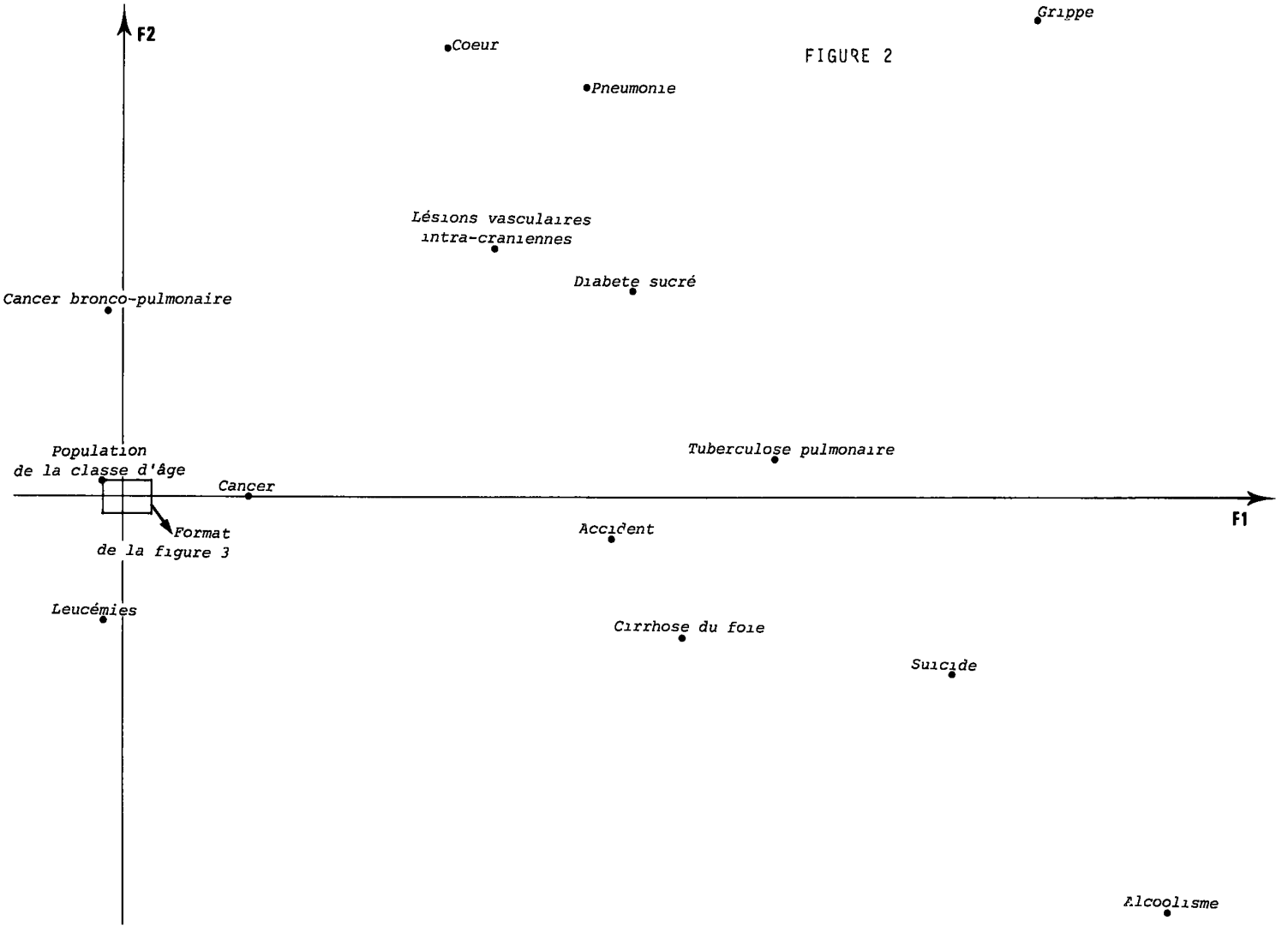
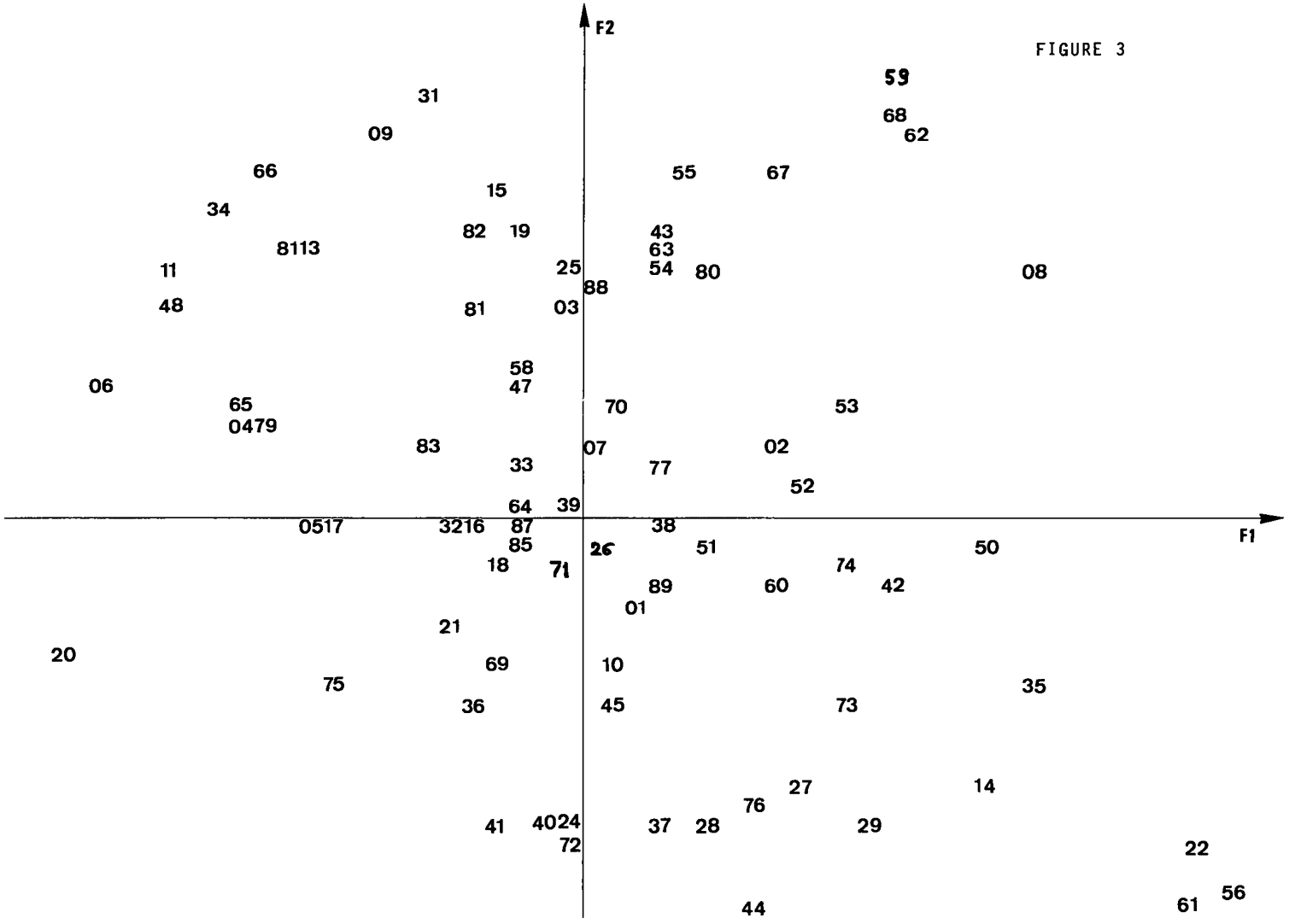

F2
F1
Grippe
Coeur
Pneumonie
Lésions vasculaires intra-craniennes
Diabete sucré
Cancer bronco-pulmonaire
Population de la classe d'âge
Cancer
Tuberculose pulmonaire
Format de la figure 3
Accident
Leucémies
Cirrhose du foie
Suicide
Alcoolisme

FIGURE 3


F2
F1
59
31
68
09
62
66
55
67
15
34
82
19
43
8113
63
11
25
54
80
08
88
48
81
03
58
06
47
65
70
53
0479
83
07
02
33
77
52
64
39
0517
3216
87
38
85
26
51
50
18
71
74
89
60
42
01
21
20
69
10
75
35
36
45
73
27
14
76
41
4024
37
28
29
72
22
44
61
56

Le point-population occupe le centre de gravité, au voisinage immédiat duquel se trouvent maintenant les points "Cancer", "Leucémie" , et "Cancer broncho-pulmonaire". Le point "Leucémie", dont le poids est faible, peut être considéré comme n'ayant pas participé à la détermination du plan des deux premiers facteurs. La distance de ce point à l'origine est donc en projection sur ce plan à comparer à la quantité $\sqrt{6/888}$,(6 est la valeur qu'un χ^2 à 2 D. L. ne dépasse que dans 5% des cas, et 888 est l'effectif total des décédés par leucémie). ($\sqrt{6/888} = 0.08$).

Il apparaît que le point "Leucémie" n'est pas significativement différent de la population, ce qui indique une répartition homogène de cette affection (ou plus précisément du diagnostic de cette affection).

Cette procédure de contrôle ne s'applique pas au point Cancer, qui, vu son poids prédominant, n'a pas manqué d'attirer à lui le plan des deux premiers facteurs ; en fait, sa distance à l'origine dans R^{88} est de l'ordre de 0.10, alors que pour cet espace, la sphère de garde a pour rayon 0.08.

En définitive, toutes les causes de décès ont des répartitions géographiques spécifiques, à l'exception de la leucémie, simplement distribuée au prorata des effectifs de la classe d'âge.

Il existe vraisemblablement des disparités régionales de la morbidité, indiscernables ici des disparités régionales dans la nature et la qualité des diagnostics.

Les proximités entre les différents points ne sont pas fondamentalement modifiées.

On a obligé les axes factoriels à passer par le point "population", qui n'était pas loin du plan des deux premiers facteurs de la première analyse, car la figure 2 se déduit de la figure 1, avec une concordance assez bonne, par une rotation de 45°.

La présence de ce point Population a donc ici pour principal effet d'enrichir l'interprétation.

La figure 3 nous donne les proximités entre les points - départements. Le taux de mortalité de ces départements est d'autant plus élevé que ceux-ci sont situés vers la droite.

On pouvait peut-être s'attendre à des zones concentriques d'égale mortalité autour de l'origine, mais le point population n'est pas strictement à l'origine , et a pour abscisse -0.00011.

C'est surtout le second facteur qui nous renseigne sur la forme de la mortalité. (Le premier aussi, puisque dans le cas précis de la mortalité, le niveau n'est pas indépendant de la forme..)