PRESENTATION GENERALE DES METHODES D'ANALYSE MULTIDIMENSIONNELLE

(Aspect pragmatique et heuristique)

par R. TOMASSONE
Laboratoire de Biométrie C.N.R.Z. - 78350 Jouy-en-Josas

RESUME: Texte présentant à l'aide d'un exemple les différentes méthodes de l'analyse des données multidimentionnelles: quels types de questions peut-on se poser devant un tableau de données? On tente aussi de montrer comment une première analyse peut-être source d'hypothèses nouvelles permettant de mieux comprendre et de mieux expliquer une structure de données.

DESCRIPTEURS:

02 - Analyse multivariante.
12 - Discrimination.
04 - Classification.
06 - Analyse factorielle.

Classement.

PLAN

1. INTRODUCTION.

Nous voulons dans cet exposé volontairement très simple présenter l'ensemble des méthodes d'analyse multidimensionnelle; il n'existe malheureusement pas une façon unique de les présenter, façon qui satisferait à la fois les théoriciens et ceux qui souhaitent simplement appliquer à leur domaine de travail un ensemble de techniques dont ils ressentent plus ou moins confusément qu'elles pourraient leur être d'un grand secours. Les voies les plus axiomatiques consisteraient, à partir de quelques principes admis par tous, à construire un bel édifice où chacun selon ses besoins trouverait le « logement » de ses rêves. Malheureusement l'esprit humain est si casanier qu'ayant enfin trouvé sa tanière il n'en veut plus sortir! Nous ne voudrions pas qu'ayant enfin une méthode très satisfaisante il s'y enferme; il serait sans nul doute bien prétentieux d'affirmer qu'elle s'adaptera à tous ses problèmes. Pour cette raison, nous allons partir de deux exemples concrets qui nous ont été donnés par des biologistes, pour essayer de bâtir notre présentation générale. Naturellement les « vrais » statisticiens, c'est-à-dire ceux formés à l'école de la théorie de l'estimation et de celle des tests, ceux pour qui la préparation d'un plan d'expérience est essentielle, vont opposer un certain nombre d'objections. La principale sera sans doute qu'un expérimentateur construit une expérience et collecte des données pour vérifier une hypothèse, et qu'il ne peut pas avec les mêmes données répondre à plusieurs questions. Dans un « état idéal » cette attitude pure et dure est sans nul doute concevable; la réalité physique est souvent fort éloignée, et l'on essaie presque toujours de retirer le maximum d'information d'une expérience toujours limitée par la dure réalité: l'échantillon à partir duquel on travaille est « petit » parce qu'on ne peut pas faire autant de mesures qu'on le souhaite, le temps entre en jeu quand il s'agit d'étudier un animal et on ne peut guere accélérer sa croissance pour avoir davantage de mesures.

A l'opposé certains nous diront: « amassez le plus d'observations possible, ne faites aucune hypothèse, puis à l'aide d'une technique quasiparfaite découvrez dans le magma informe de vos chiffres une structure. Cette structure vous permettra de mettre en évidence des facteurs jusque là inconnus, et à partir d'eux de découvrir des lois ». C'est sans doute faire fi de la connaissance a priori que nous avons; et même si cette connaissance doit être améliorée il n'est peut

TABLEAU 1. - *Matrice des données de base (a), c. texte; et premières statistiques élémentaires (b).*

	X_1	X_2	X_3	X_4	X_5	X_6		
(a)	395	224	35.1	79.1	6.0	14.9	1*A*	1
	410	232	31.9	73.4	9.7	16.4	1*B*	2
	405	233	30.7	76.5	7.0	16.5	1*C*	3
	405	240	30.4	75.3	8.7	16.0	1*D*	4
	390	217	31.9	76.5	7.8	15.7	1*E*	5
	415	243	32.1	77.4	7.1	15.5	1*F*	6
	390	229	32.1	78.4	4.6	17.0	1*G*	7
	405	240	31.1	76.5	8.2	15.3	1*H*	8
	420	234	32.4	76.0	7.2	16.8	1*I*	9
	390	223	33.8	77.0	6.2	16.8	1*J*	10
	415	247	30.7	75.5	8.4	16.1	1*K*	11
	400	234	31.7	77.6	6.7	18.7	1*L*	12
	400	224	28.2	73.5	11.0	15.5	2*M*	13
	395	229	29.4	74.5	9.3	16.1	2*N*	14
	395	219	29.7	72.8	8.7	18.5	2*O*	15
	395	224	28.6	73.7	8.7	17.3	2*P*	16
	400	223	28.5	73.1	9.1	17.7	2*Q*	17
	400	224	27.8	73.2	12.2	14.5	2*R*	18
	400	221	26.5	72.5	13.2	14.5	2*S*	19
	410	233	25.9	72.3	11.1	16.6	2*T*	20
	402	234	27.1	72.1	10.4	17.5	2*U*	21
	400	223	26.8	70.3	13.5	16.2	2*V*	22
	400	213	25.8	70.4	12.1	17.5	2*W*	23

	Variable	Minimum	Maximum	Moyenne	*E*-type	*C*-varia	Deary	Rac(B1)
(b)	1	390.000	420.000	401.609	8.173	2.035	0.789	0.549
	2	213.000	247.000	228.826	8.658	3.784	0.843	0.297
	3	25.800	35.100	29.922	2.571	8.593	0.859	0.024
	4	70.300	79.100	74.670	2.479	3.319	0.877	0.013
	5	4.600	13.500	3.952	2.424	27.080	0.810	0.239
	6	14.500	18.700	16.422	1.131	6.887	0.807	0.200

	1	2	3	4	5	6
1	1000					
2	691	1000				
3	— 33	283	1000			
4	— 58	390	895	1000		
5	101	—328	—861	—908	1000	
6	—110	—100	— 63	—128	—263	1000

matrice des coefficients de corrélation ×1000

être pas inutile d'en tenir compte, quitte, si l'observation nous contredit et la met en défaut, à la modifier pour fournir un nouvel a priori qu'une expérimentation future nous contraindra peut-être à corriger. Mais quelquefois l'expérimentateur est beaucoup plus ambitieux; il veut par exemple à partir de résultats actuels prédire un futur qu'il ne connait pas encore; il veut aussi affirmer, en prenant un minimum de risques, que cet échantillon qu'il ne connait pas appartient à une des populations qu'il connait déjà ou bien qu'il est contraint par la réalité expérimentale de créer une population nouvelle. Dans ce cas la connaissance théorique actuelle que nous avons, imposera de faire quelques hypothèses supplémentaires; si elles sont vérifiées, nous pourrons affirmer avec plus de certitude que cet échantillon nouveau n'était en réalité qu'un élément d'un ensemble déjà connu.

Sans vouloir à tout prix concilier deux écoles, celle des statisticiens et celle des analystes de données, il nous semble qu'il y a suffisamment de richesse dans chacune d'elle pour utiliser leur façon de penser à bon escient. Enfin il serait absurde de vouloir ignorer que toutes ces techniques, connues depuis le début du siècle pour certaines, et depuis plus de vingt ans pour la totalité, ne sont sorties de l'anonymat mathématique que grâce à ce que nous pourrions appeler la démocratisation de l'emploi des ordinateurs. Ces outils ne sont pas indispensables au niveau conceptuel, mais sans eux aucun calcul ne serait raisonnablement faisable. De plus l'aspect algorithmique qu'il a été nécessaire de développer pour « programmer » ces méthodes, c'est-à-dire pour que les ordinateurs puissent résoudre automatiquement nos problèmes, a permis de rapprocher des méthodes qui n'avaient pour les seuls théoriciens aucun lien fondamental. Ce rapprochement est aussi une source non négligeable pour faciliter la compréhension de résultats, pour fournir un guide dans des interprétations qui ne sont jamais très simples. De façon plus matérielle de rapprochement purement numérique, peut permettre d'utiliser un nombre limité de programmes pour ordinateur, à condition de savoir les utiliser de façon convenable.

2. PREMIER EXEMPLE: MENSURATION DE BOVINS.

2.1. Les donnees et les questions qu'on peut se poser a leur sujet.

2.1.1.

Cet exemple est de faible dimension, il nous permettra de mieux faire le passage entre les données brutes et les différentes analyses que nous leur ferons subir. Il s'agit d'un lot de 23 bovins élevés à Cuba, donc dans un pays tropical, sur lesquels six types de mesures ont été faits. Ce tableau que nous présenterons sous la forme donnée à la fig. 1 est-ce que nous appellerons la *matrice des données de base*, nous la noterons X; un élément de cette matrice x_{ij} représente la valeur

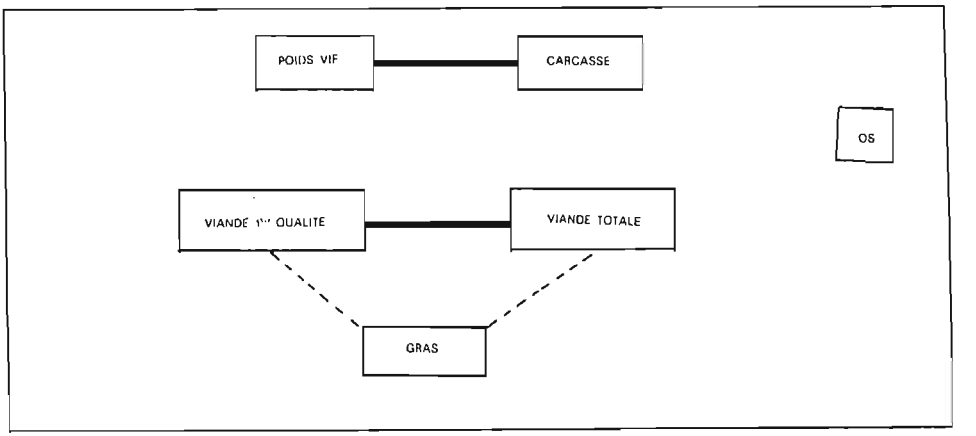

Figure 1. – Graphe de la matrice des coefficients de corrélation. ▬▬▬: coefficient > 0 et significatif à 1% (.526); – – – –: coefficient < 0 et significatif à 1% (— .526).

de la variable j pour le bovin correspondant à la ligne i de notre tableau ; ainsi x_{34} représente la valeur de la quatrième variable pour le troisième bovin. Jusqu'ici l'ordre des bovins et celui des variables est une simple affaire de convention à partir du moment où nous aurons fixé l'un et l'autre il deviendra naturellement capital pour l'interprétation.

2.1.2. *Les variables.*

Les six variables que nous noterons X_1, X_2, X_3, X_4, X_5 et X_6 ou plus brièvement X_i, l'indice i pouvant prendre les valeurs allant de 1 à 6 ont la signification suivante :

X_1 : Poids vif,

X_2 : Poids de la carcasse,

X_3 : Poids de la viande de première qualité,

X_4 : Poids de la viande totale,

X_5 : Poids du gras,

X_6 : Poids des os.

2.1.3. *Les observations.*

Elles sont au nombre de 23 ; chaque observation est parfaitement définie à l'aide des six valeurs des variables X_i. La connaissance que nous aurons d'un bovin dans cette expérience ne sera faite qu'au travers de ce *vecteur observation* et de *rien d'autre*. Naturellement nous voulons avoir davantage d'information

sur la façon dont ont été collectées ces observations; nous savons qu'elles proviennent toutes de troupeaux soigneusement controlés que ce ne sont ni des bovins particulièrement rachitiques ni des bovins particulièrement sélectionnés, ils donnent donc une image relativement représentative de leur espèce. Néanmoins ils appartiennent à deux races parfaitement bien déterminées:

— l'une est la race *charolaise* importée au début du siècle dans la perle des Antilles, les descendants sur lesquels les mesures ont été faites sont parfaitement acclimatés,

— l'autre est la race *zèbu* autochtone donc parfaitement adaptée aux conditions locales.

Pour des raisons de pure commodité de présentation nous avons placé les 12 charolais en tête et les 11 zébus à la suite.

2.1.4. *Les questions possibles.*

Aucune variable ne joue ici un rôle particulier, ou tout au moins l'expérimentateur les place sur le même plan. Par contre nous pouvons tenir compte ou ne pas tenir compte du fait que nous savons a priori qu'il existe deux races. Nous allons voir que selon que nous prenons en compte ou non l'existence de ces deux races, les techniques sont différentes, elles fournissent en partie des renseignements du même ordre mais peuvent aussi se compléter.

Dans les deux cas nous souhaitons étudier la *variabilité* de notre échantillon de 23 bovins en utilisant au mieux les mensurations qui ont été faites. En particulier nous pensons qu'il y a sûrement une certaine redondance entre les variables: sans mesurer le même facteur, il est certain qu'il doit exister une corrélation entre le poids de la carcasse (X_2) et celui de la viande totale (X_4). Il est donc logique de se demander si un *indice-unique* ne peut synthétiser la variabilité totale, et la résumer à lui seul.

2.2. Les premiers elements synthetiques.

Avant de rechercher cet indice hypothétique, il est naturel d'étudier les variables, de voir les liens qu'elles peuvent avoir entre elles c'est ce qui est donné au tableau 2. Pour X_1 nous voyons, dans la première partie du tableau 2, que si la valeur moyenne est 401,6 le poids vif le plus faible est 390,0 et le plus grand 420,0 avec un écart-type de 8,17, donc un coefficient de variation de 2.04%. De plus les coefficients de Geary (0.789) et le coefficient $\sqrt{b_1}$ de Pearson (0.549) permettent de vérifier si la distribution est plus ou moins normale (*).

(*) Nous employons ici les termes indépendance et non corrélation dans le même sens puisque le premier est plus restrictif: dans notre contexte il revient à supposer que les distributions sous jacentes sont gaussiennes.

TABLEAU 2. - *Composantes principales « bovins ».*

Vecteur i propre	1	2	3	4	5	6
λ_i	2.9544	1.6566	1.0455	0.24865	0.08390	0.01090
$100^* \lambda_i/6$	49.24	27.61	17.43	4.14	1.40	.18
Somme cumulée	49.24	76.85	94.28	98.42	99.82	100.00
x_1	.05	.72	.20	—.59	.30	.00
x_2	.30	.60	.17	.63	—.35	.00
x_3	.54	—.12	—.13	—.47	—.67	.07
x_4	.56	.08	—.17	.13	.48	.63
x_5	—.55	.19	—.19	—.05	—.33	.72
x_6	.00	—.26	.92	—.04	—.08	.28

Si l'on pousse encore plus loin l'investigation, en étudiant les liaisons entre les 6 variables, nous voyons que l'examen de la matrice des coefficients de corrélation (deuxième partie du tableau 2) est déjà très instructif, puisque si on élimine les liaisons que la référence au seuil de signification sous hypothèse normale nous permet de faire, on peut tracer le graphe de la figure 1 et voir déjà apparaître trois groupes de variables sans lien très important entre eux bien que le poids de carcasse ne soit pas totalement indépendant (*) de la viande de première qualité (0.283) de la viande totale (0.390) et du gras (— 0.328). Il peut être interessant d'utiliser un moyen plus objectif d'analyser ces données et de s'en servir pour effectuer une classification ou une ordination des observations à l'aide de variables plus synthétiques.

2.3. PREMIERE ANALYSE: ANALYSE EN COMPOSANTES PRINCIPALES.

2.3.1. *L'idée de départ est très simple.*

Si au lieu des six variables de départ qui décrivent nos observations nous n'en avions qu'une seule que nous noterons Y_1 nous pourrions décrire nos observations à l'aide d'un seul indice, c'est-à-dire caractériser un bovin par *un* nombre au lieu de *six*. Cet indice doit correspondre à une fonction des variables de départ X_i ($i = 1, 6$) et à un certain nombre de paramètres $a_1, \ldots, a_m$. Si nous notons

$$Y_1 = f(\boldsymbol{X}, \boldsymbol{a}) \tag{1}$$

ou $\boldsymbol{X}$ est le vecteur dont les 6 composantes sont les X_i ($i = 1, 6$) et $\boldsymbol{a}$, le vecteur

(*) Les valeurs pour Geary sont 0.730 — 0.877, et $|\sqrt{b_1}|$ .711 d'après les tables de Biometrika (p. 207), donc ici rien ne s'oppose à admettre la normalité des six distributions univariates.

dont les 6 composantes sont les a_i ($i = 1, \ldots, 6$). Des considérations pratiques — déterminer une fonction f simple qui fasse passer des X_i à Y_i et employer une fonction qui soit manipulable avec nos outils mathématiques — font que le choix de la *transformation* (c'est-à-dire de la fonction f) est limité à la *transformation linéaire* suivante:

$$(2) \qquad Y_1 = a_1 X_1 + a_2 X_2 + \ldots + a_6 X_6 = \sum_{i=1}^{i=6} a_i X_i .$$

Une fois connue la forme analytique de la transformation, nous nous trouvons devant un problème *d'éstimation*: comment choisir les a_i? Pour cela il faut donner une condition qui nous parait raisonnable pour Y_1; la plus simple consiste à dire que nous voulons que, parmi toutes les combinaisons linéaires de X_i, Y_1 soit celle qui ait la plus grande variance. Le problème est alors très bien posé de façon mathématique, il admet une solution: c'est le premier *vecteur propre* de la matrice des covariances des variables X_i. Pour des raisons que nous developperons plus tard on préfère utiliser la matrice des coefficients de corrélation des X_i (ce qui revient à travailler sur des variables X_i qui ont la même échelle).

TABLEAU 3. - *Coefficients de corrélation entre variables de départ et composantes principales.*

Composante i	y_1	y_2	y_3	y_4	y_5	y_6
x_1	.09	.92	.21	−.30	.09	.0
x_2	.52	.77	.17	.32	−.10	.0
x_3	.93	−.15	−.13	−.24	−.19	.01
x_4	.96	−.11	−.17	.06	.14	.07
x_5	−.94	.25	−.20	−.02	−.09	.08
x_6	.00	−.34	.94	−.02	−.02	.03

Le calcul du premier vecteur propre nous est donné au tableau 3, nous voyons que y_1 peut s'écrire en fonction des x_i (c'est-à-dire des variables divisées par leur ecart type: $x_1 = X_1/8.173$, $x_2 \neq X_2/8.658$, ..., $x_6 = X_6/1.131$)

$$(3) \qquad y_1 = 0.05x_1 + 0.30x_2 + 0.54x_3 + 0.56x_4 - 0.55x_5 + 0.00x_6$$

on voit que les coefficients a_i sont assez différents, les uns importants ($a_3 = 0.54$, $a_4 = 0.56$, $a_5 = -0.55$), les autres très faibles ($a_1 = 0.05$, $a_6 = 0.00$) et le dernier prend une valeur intermédiaire ($a_2 = 0.30$); comme les variables x_i ont la même échelle, nous pouvons admettre qu'ils pondèrent les variables de départ de façon différente: la viande de première qualité, la viande totale et le gras parti-

cipent pour la plus grande part à y_1, alors que le poids vif et le poids des os n'interviennent pratiquement pas, et que la carcasse, sans être négligeable, n'y a pas une grande importance. Nous avons donc notre indice, c'est la *première composante principale.* Il faut naturellement avoir le moyen de juger de son importance c'est ce qui nous est donné par la valeur propre associée à cette composante; ici $\lambda_1 = 2.9544$. On démontre que λ_1 est la variance de y_1

$$\lambda_1 = \text{variance } (y_1) \tag{4}$$

comme chaque variable x_i a par définition une variance égale à 1, nous avons à l'aide de la seule composante y_1 une variable de variance maximum parmi toutes les combinaisons linéaires des x_i.

De plus nous savons que cette variance ne peut pas être supérieure au nombre total de variables, ici 6, il est commode de l'exprimer en pourcentage de ce maximum, c'est-à-dire 49.24%. Nous dirons que y_1 traduit presque la moitié de la variabilité totale de l'ensemble des six variables.

2.3.2. *Interprétation de la première composante.*

A ce stade nous pouvons interpréter les composantes en comparant les valeurs respectives des poids. Plus précisément, nous avions besoin pour déterminer de façon unique les coefficients a_i de nous imposer la condition supplémentaire de normalisation

$$a_1^2 + a_2^2 + \ldots + a_6^2 = \sum_{i=1}^{i=6} a_i^2 = 1 \tag{5}$$

cette condition nous permet de dire que les poids interviennent par leur carré, donc que x_3, x_4, x_5 avec des coefficients sensiblement égaux sont pratiquement quatre fois plus importants que $x_2((0.55/0.30)^2 \simeq 4)$ ce sont donc ces variables qui doivent servir à interpréter y_1.

De façon plus précise si un bovin a des mesures x_3 et x_4 élevées (poids de viande élevé) et x_5 (gras) faible il aura une valeur y_1 grande; par contre avec des valeurs inversées il aura une valeur y_1 faible. On peut donc affirmer que l'indice y_1 oppose les bovins avec un poids de viande élevé et un poids de gras faible à ceux qui ont moins de viande et plus de gras.

Statistiquement, on peut aller plus loin encore, si chaque a_i est multiplié par $\sqrt{\lambda_1} = \sqrt{2.9544}$ on obtient les coefficients du tableau 4 qui représentent les coefficients de corrélation entre y_1 et les six variables x_i, où nous voyons bien que x_3, x_4 et x_5 sont très correlées avec y_1, et les autres x_1, x_5 pratiquement pas et x_2 moyennement.

TABLEAU 4. - *Valeurs numériques des 6 composantes principales pour les 23 bovins.*

Composante principale	1	2	3	4	5	6
Observation						
CA	2.55	−1.19	−1.83	− .49	− .02	.03
CB	.12	.97	.18	− .82	− .68	− .05
CC	1.19	.32	.22	.05	.37	− .07
CD	.71	1.10	− .09	.53	− .27	− .01
CE	.61	−1.92	−1.24	− .25	.10	.01
CF	2.07	2.03	− .29	− .13	.19	− .02
CG	2.22	−1.71	.16	.71	.20	− .13
CH	1.24	1.15	− .74	.50	− .10	.00
CI	1.51	1.59	.49	−1.33	.29	− .02
CJ	1.69	−1.99	− .24	− .14	− .39	− .02
CK	1.19	2.40	.37	.28	− .19	− .02
CL	1.94	− .75	1.88	.31	.13	.40
ZM	−1.27	.02	− .88	.01	.14	.03
ZN	− .26	− .44	− .41	.59	− .17	− .03
ZO	− .80	−1.69	1.50	− .36	− .25	− .04
ZP	− .65	−1.04	.01	.30	.08	− .14
ZQ	− .90	− .71	1.06	− .17	.13	− .08
ZR	−1.69	.35	−1.66	.08	.08	.08
ZS	−2.50	.34	−1.75	.04	.24	− .09
ZT	−1.67	1.42	.64	.27	.41	− .04
ZU	−1.32	.48	1.19	.68	− .23	.04
ZV	−2.89	.16	− .25	− .04	− .47	− .09
ZW	−3.10	− .90	.79	− .60	.32	.00

2.3.3. *Recherche d'autres composantes.*

Nous avons bien sûr 50% de la variabilité exprimée par y_1, mais il reste encore 50% que nous n'avons pas encore maitrisé aussi est-il logique de partir à la recherche d'une deuxième composante y_2 qui ait des propriétés analogues à y_1. Nous pouvons effectuer les mêmes calculs mais il nous faut imposer une condition supplémentaire, (sinon nous retrouverions encore y_1!), cette condition de bon sens consiste à chercher y_2 *non corrélée* avec y_1; nous obtenons le deuxième vecteur propre:

$$(6) \qquad y_2 = 0.72x_1 + 0.60x_2 - 0.12x_3 - 0.08x_4 + 0.19x_5 - 0.26x_6$$

avec une variance $\lambda_2 = 1.6566$ elle traduit 27.61% de la variabilité. Nous l'interprétons naturellement en fonction du poids vif et de celui de la carcasse.

2.3.4. *L'Ordination des observations.*

Pour l'instant nous ne nous sommes interessés qu'aux variables; nous allons revenir aux observations en nous servant des résultats précédents. Pour un bovin donné, c'est-à-dire pour six valeurs mesurées $X_1, X_2, \ldots, X_6$ — donc pour six valeurs réduites $x_1, x_2, \ldots, x_6$ — nous pouvons calculer la valeur numérique des deux premières composantes y_1 et y_2 par les équations (3) et (6), et même de toutes à l'aide des coefficients du tableau 4. Ces valeurs sont données au tableau 5. Nous voyons pour la première composante que les douze premièreres

TABLEAU 5. - *Analyse discriminante entre charolais et zébu.*

Variable	P_1 charolais	P_2 zébu	Ecart-type intra-race	Test univarsate $F(1, 21)$
1	403.3	399.7	8.1	1.12
2	233.0	224.3	7.6	7.57
2	32.0	27.7	1.3	59.51
4	76.6	72.6	1.4	47.15
5	7.2	10.8	1.6	29.54
6	16.3	16.5	1.2	.24

MATRICE DES COEFFICIENTS DE CORRELATION INTRA-RACE

$$S = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{pmatrix} 1.000 & & & & & \\ .689 & 1.000 & & & & \\ -.458 & -.364 & 1.000 & & & \\ -.454 & -.079 & .634 & 1.000 & & \\ .435 & .117 & -.621 & -.763 & 1.000 & \\ -.089 & -.053 & .055 & -.072 & -.539 & 1.000 \end{pmatrix}$$

ETUDE DE LA STRUCTURE INTERNE DE S (COMPOSANTES PRINCIPALES)

λ	2.943	1.311	1.112	.432	.176	.025
λ%	49.05	21.85	18.53	7.37	2.93	.37
1	.45	.40	.17	.50	-.60	.01
2	.30	.62	.37	-.18	.60	.01
3	-.48	-.02	.22	.77	.35	.09
4	-.46	.17	.49	-.32	-.36	.54
5	.50	-.43	.04	.11	.20	.72
6	-.16	.49	-.74	.06	.05	.43

FONCTION DISCRIMINANTE D_6^2 (charolais, zébu) = 18.69

Composantes principales des		.458	-.444	-.586	-.459	-.039	-.191
variables	réduites	-.437	-.308	-.919	-.850	-.679	-.387
de dèpart	non réduites	-.0105	-.0077	-.1309	-.1105	-.0740	-.0584

observations (les charolais) ont une valeur toujours positive, et les onze dernières ont une valeur toujours négative. Nous avons réalisé une *ordination* des observations à l'aide de cette première composante; grâce à elle nous avons séparé les charolais des zébus, et nous savons pourquoi: les premiers ont un poids de viande plus élevé et les seconds beaucoup plus de gras.

Naturellement une représentation graphique va être d'un grand secours pour mieux voir où se situent les différentes observations, pour mieux analyser leurs positions respectives; c'est ce qui est fait à la figure 1. Les charolais sont à droite les zébus à gauche; mais une fois ce facteur isolé on peut trouver des zébus et des charolais qui ont la même valeur pour la deuxième composante par exemple CC (.32) et ZS (.34).

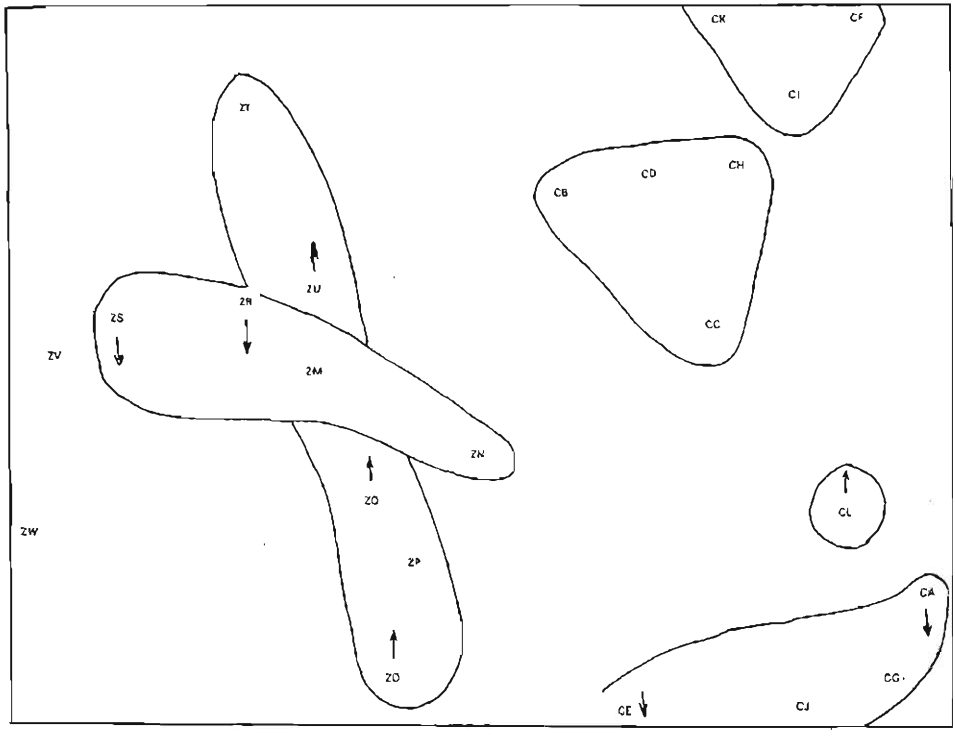

Figure 2. – Projection des 23 bovins dans le plan des 2 premières composantes principales.

Notons au passage que l'étendue de la première composante (la différence entre la plus grande valeur de y_1 2.55 pour CA et la plus faible — 3.10 pour ZW) est 5.65, et les cinq suivantes sont 4.39, 3.71, 2.04, 1.10, et 0.54; donc un résultat concordant avec la nature des composantes; ce résultat n'est pas général dans la mesure où une observation particulière peut introduire une distorsion.

2.4. Notion de distance et classification.

2.4.1. *Calcul des distances à partir des données de base.*

Cette simple ordination nous permet de dire que certains bovins sont plus voisins que d'autres ainsi CB et CD sont très proches, mais ZV et ZS le sont aussi; mais notre graphique est un plan et nous ne pouvons pas affirmer que deux points voisins dans le plan formé par y_1 et y_2 le sont aussi dans l'espace. Il nous faut donc une quantité supplémentaire; nous allons la définir en prenant tout simplement comme distance entre deux points k et l:

$$d^2(k, l) = \sum_{j=1}^{j=p} (x_{kj} - x_{lj})^2 \tag{7}$$

c'est-à-dire en appliquant le théorème de Pythagore. De façon plus synthétique si nous appelons $\boldsymbol{x}$ la matrice, à n lignes et p colonnes, formée à partir des éléments de la matrice des données de base $\boldsymbol{X}$, par centrage et réduction (c'est-à-dire en prenant l'écart à la moyenne générale de chaque variable $\overline{X}_j$, et en divisant par l'écart type de cette variable s_j)

$$\mathfrak{x} = \left(x_{ij} = \frac{X_{ij} - \overline{X}_j}{s_j}\right) \tag{8}$$

on voit que la matrice des distances entre toutes les observations prises deux à deux se déduit facilement de la matrice $\boldsymbol{\delta}$ définie par:

$$\underset{(nxn)}{\boldsymbol{\delta}} = \underset{(nxp)}{\boldsymbol{x}} \; \underset{(pxn)}{\boldsymbol{x}'} \tag{9}$$

(où $\boldsymbol{x}'$ est la matrice transposée de $\boldsymbol{x}$).

$$d_{ij} = \delta_{ii} + \delta_{jj} - 2\delta_{ij} \,. \tag{10}$$

2.4.2. *Utilisation des distances pour une classification.*

L'utilisation des distances permet donc de corriger ce qu'une vue plane efface; ainsi en reprenant les deux couples d'éléments déjà vus plus haut:

$$d^2(\text{CB, CD}) = 0.19$$

$$d^2(\text{ZV, ZS}) = 1.29 \,.$$

Mais cette simple correction peut être encore mieux utilisée; la matrice des

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
2	1490	0																				
3	993	1686	0																			
4	1230	983	668	0																		
5	240	1609	691	1089	0																	
6	991	1331	299	550	1121	0																
7	1338	2704	507	1080	870	1273	0															
8	916	1254	517	83	831	354	954	0														
9	1571	1358	923	1868	1677	719	2505	1758	0													
10	536	930	1076	812	437	384	989	891	1986	0												
11	1594	951	688	163	1604	320	1533	242	1366	1321	0											
12	2855	3695	2444	2331	2571	2601	3002	2424	3108	2467	2503	0										
13	803	1381	488	585	392	695	1174	469	1447	944	877	2340	0									
14	1024	1271	618	208	607	917	680	256	2212	548	657	2503	356	0								
15	1568	926	1066	1117	966	1679	1362	1441	1606	548	1434	2416	1013	819	0							
16	1447	1561	405	718	750	1113	441	816	1828	797	1070	3025	609	341	502	0						
17	1591	1251	362	850	742	958	835	971	1119	850	1020	2491	525	584	283	154	0					
18	922	1643	926	766	595	957	1719	586	1944	1279	1119	2656	96	554	1577	1114	1055	0				
19	1226	2094	1079	1132	766	1205	1985	906	2105	1714	1486	2920	166	828	1853	1292	1187	52	0			
20	2100	2056	398	838	1427	703	1306	820	1394	1999	784	2704	508	827	1371	626	497	850	843	0		
21	2197	1328	901	340	1537	1226	1209	653	2299	1171	573	2460	817	343	730	453	553	1183	1455	651	0	
22	1707	856	1659	764	1239	1684	2471	1019	2330	1103	1136	2575	597	664	896	1228	1043	611	818	1338	719	0
23	1894	1840	956	1731	1026	1644	1886	1807	1312	1635	1916	2785	628	1280	695	774	342	1045	959	722	1265	1107

Figure 3. – Matrice des distances entre les 23 bovins (distance cartésienne et variables centrées réduites) $\times 100$.

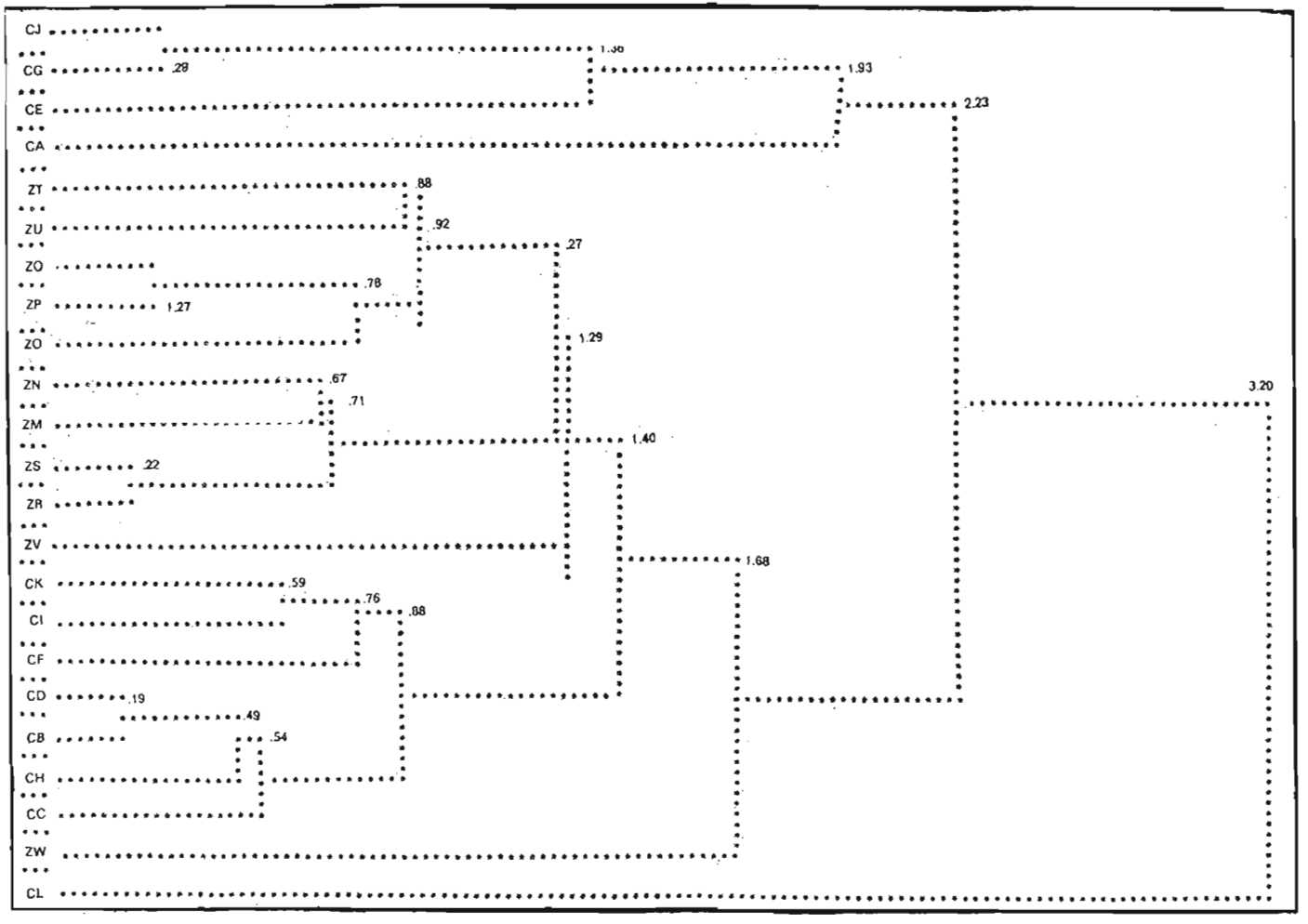

Figure 4. – Dendrogramme des 23 bovins: distance cartesienne et ultramétrique inférieure minima.

distances peut servir à réaliser une *classification*, c'est-à-dire à *former des classes*. Sans entrer dans les détails de la théorie de la classification automatique, ou *taxinomie numérique*, on peut d'un point de vue algorithmique envisager quatre étapes :

a) comment créer les premiers groupes? Par exemple en prenant les deux observations les plus voisines, ici CD et CB qui formeront les premiers groupes élémentaires ;

b) comment fusionner des éléments nouveaux à des groupes déjà créés? Ainsi CH à la distance 0.49 de CD sera fusionné avec le premier groupe déjà créé, alors que ZS et ZR formeront un groupe à part ;

c) comment interdire une nouvelle fusion? Si on impose, pour créer un groupe, qu'une observation quelconque soit à une distance inférieure à 1.0 de toute observation d'un groupe déjà créé. Le groupe — ou la classe — formé par CC, CH, CB, CD, CF, CI, CK sera considéré comme un tout homogène ;

d) comment échanger des élements entre groupes? Cette dernière étape constitue un excellent moyen de vérifier la stabilité des groupes déjà établis.

Le résultat est généralement condensé dans un *dendrogramme*, ou arbre de classification. Cet arbre, coupé à un niveau donné permet de voir que tout ce qui est « au dessus » de la coupure, appartient à une même classe.

Naturellement les méthodes de classification ne sont pas aussi simples que la description rapide que nons venons de donner peut le laisser croire.

En particulier, il existe un certain nombre de problèmes que nous avons laissé de côté :

— comment le *choix de la distance* — sa nature analytique ici définie par une distance cartésienne — peut influencer le résultat? Une réponse possible consiste à remplacer les distances par des *relations d'ordre* sur les distances. On peut penser que ces relations d'ordre seront respectées quelle que soit la forme analytique donnée à la distance ;

— comment le critère du *choix de la fusion* peut avoir une influence sur la classification? Va-t-on regrouper un élément nouveau en fonction de sa distance au point le plus voisin, au point le plus éloigné ou a un point « virtuel » qui serait le barycentre des points déjà regroupés? On voit donc qu'avec tous ces a priori, les techniques de classification ne donnent pas des critères absolument objectifs pour classer des observations. Néanmoins elles peuvent fournir une aide qui n'est pas négligeable dans une étude comme celle que nous avons vue ; il faudrait toutefois se garder de l'employer seule sans se réserver des moyens de la contrôler. Pourtant elle peut se révéler un outil parfaitement adapté quand les données de base sont fournies directement à partir de distances, ou d'*indices de ressemblance*.

Dans l'exemple de notre étude on voit très bien se dessiner trois groupes si on coupe le dendrogramme aux environs de 1.36:

a) l'un formé par 10 zébus (tous sauf ZW);

b) l'autre formé par les 7 charolais CK, CI, CF; CD; CB, CH, CC;

c) le troisième par les 3 charolais CE, CG, CJ.

Restent trois éléments CA, ZW, et CL, pour lesquels un examen plus détaillé doit être fait *avec un éventuel retour aux données de base*:

1) ZW se situe nettement à gauche dans le premier plan de l'analyse en composantes principales; c'est ce que nous pourrions baptiser un « superzébu ». Sa place dans le dendrogramme, sous un groupe de charolais, n'a aucun sens particulier vis à vis de ce groupe, il est simplement rattaché au groupe formé par la réunion de *a* et *b*. Nous voyons apparaître sur cet exemple ce qui est appelé un effet de *chaînage*: la technique de regroupement utilisée (ici ultramétrique inférieure minima) entraîne une transitivité néfaste.

2) CA se rattache sans nul doute au groupe *c*) bien que sa composante y_3 soit faible, parce que le poids d'os est faible, en effet sa valeur dans les données de base est $X_6(\mathrm{CA}) = 14.9$.

3) CL était intermédiaire entre les deux groupes de charolais assez voisin du groupe *c*), mais la valeur élevée de y_3 ne permet pas de l'intégrer dans ce groupe. C'est au fond un charolais proche de CA dont il est le plus voisin, il en diffère uniquement par un poids d'os très élévé. ($X_6(\mathrm{CL}) = 18.7$).

2.4.3. *Analyse factorielle d'une matrice de distance.*

Nous avons vu q'une matrice des données de base après un certain nombre de transformations élémentaires — centrage, réduction — pouvait être analysée de façon très simple par une analyse en composantes principales. C'est-à-dire que les variables servaient à définir des facteurs sous-jacents, qui, interprétés en fonction des variables de départ, permettaient de réaliser une ordination des observations et de connaître les clés de cette ordination.

Cette même matrice permettait aussi de définir des ressemblances entre observations, à l'aide d'une mesure de distance, et de réaliser avec elle une classification.

Nous avons dit que des observations pouvaient être faites directement sur des distances ou sur des ressemblances, on peut directement faire une classification, mais on peut aussi analyser cette matrice par une *analyse des proximités*; cette analyse consiste tout simplement à faire une recherche des valeurs propres et des vecteurs propres de la matrice des d_{ij} définie en (10), c'est-à-dire une analyse factorielle sur une matrice de distances au lieu de la faire sur une matrice

de coefficients de corrélation. Le principe consiste à projeter le solide géométrique défini par les points des observations définis par les d_{ij} dans un espace à $n-1$ dimensions, en essayant d'obtenir une déformation aussi réduite que possible du solide initial. Nous ne présenterons pas des résultats numériques ici; mais on peut résumer l'ensemble des techniques d'analyse de matrices non structurées comme indiqué sur la figure 5.

2.5. Le role symetrique des observations et des variables: la dualite.

Nous voyons bien grâce à la figure 5 qu'on peut suivre plusieurs chemins pour arriver à une réalité diffuse: l'analyse de données obtenues par l'expérience ou simplement par l'observation. Mais de plus, on voit poindre une notion fondamentale la *dualité*: par une analyse en composantes principales, grâce à des variables on arrive à faire une ordination des observations. Mais naturellement les variables, et leur regroupement qui apparait dans les composantes principales, peuvent être caractérisées par les observations les plus typiques. On devine donc à travers cette remarque que le rôle entre observations et variables est symétrique; mathématiquement, on peut simplement dire que la diagonalisation de la matrice $\boldsymbol{x}'\boldsymbol{x}$ fournit le même résultat que celle de $\boldsymbol{x}\boldsymbol{x}'$ à $n-p$ valeurs propres nulles près (si $n \geqslant p$).

Cette dualité qui permet d'enterpréter des variables en fonction d'observations, ou des observations à l'aide de variables pour lesquelles elle présentent des valeurs particulières n'est pas traduite par une parfaite symétrie des analyses. Ceci tient à la nature de la métrique choisie, ici métrique cartésienne et corrélativement celle de la covariance.

2.6. Retour a une symetrie parfaite: l'analyse factorielle des correspondances.

Cette symétrie peut être retrouvée en grande partie par un choix convenable de la métrique, il suffit au lieu de travailler sur les variables de départ d'utiliser leurs valeurs relatives. C'est ce que permet de faire l'analyse factorielle des correspondances primitivement utilisée pour l'analyse des tableaux de contingence, c'est-à-dire des tableaux de dénombrement n_{ij} représentant le nombre d'observation correspondant au niveau i d'un premier facteur et au niveau j d'un second facteur (le premier facteur pouvant représenter la couleur des yeux, ses niveaux étant: bleu, vert, marron et noir, et le second celle des cheveux).

Ces tableaux peuvent se ramener simplement à des tableaux de pourcentage en divisant n_{ij} par le nombre total d'observations $n..(n.. = \sum n_{ij})$ l'analyse des correspondances puisse mieux se justifier théoriquement on peut dire que dans un premier temps on l'a appliquée à des tableaux de données qui étaient des résultats de mesures sans se préoccuper beaucoup de la justifier dans cette situa-

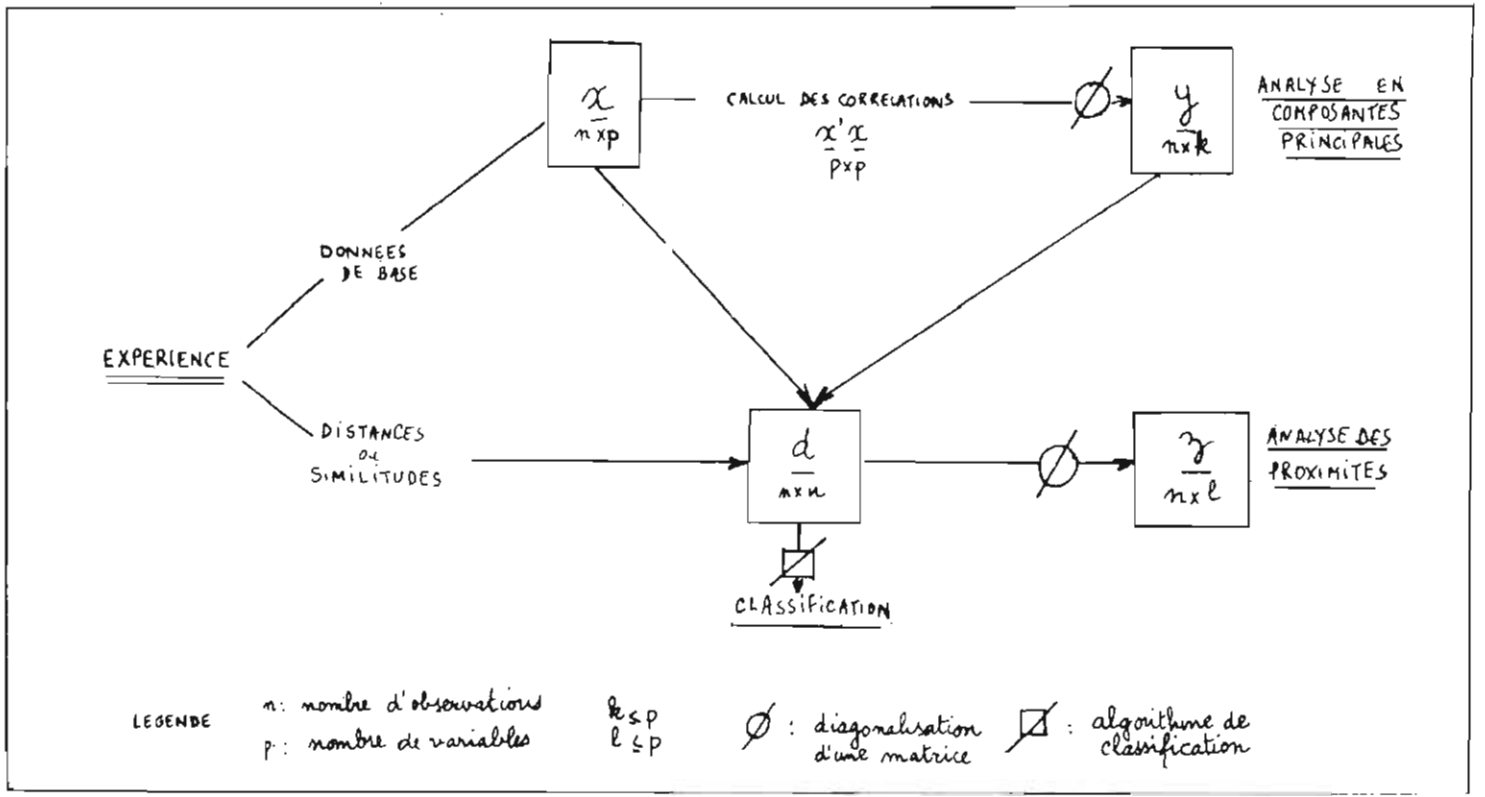

Figure 5. – Schéma simplifié des voix possibles de traitement à partir de données expérimentales non structurées (les chiffres sous les matrices indiquent leur dimension.)

tion. Si toutes les observations ont la même unité on voit qu'elle va permettre d'étudier de façon parfaite les « *profils de variables* » pour une observation ou les « *profils d'observations* » pour une variable. Ce profil sous entend donc que les valeurs relatives des variables vont être mises en évidence plus que les valeurs absolues. C'est pour cette raison qu'on peut l'appliquer de façon très fructueuse dans des *études de morphologie*; l'analyse en correspondance fera beaucoup mieux disparaître un facteur taille; en effet, lorsque celui-ci est présent et important il apparait presque toujours comme première composante principale sans intérêt.

Techniquement, on introduit une métrique différente; celle à laquelle les statisticiens sont habitués dans l'analyse des tables de contingence; *la métrique du* χ^2; c'est en fait elle qui permet de respecter la symétrie variable-observation et permet sous certaine réserve-une représentation simultanée des deux ensembles, comme on peut le voir sur la figure 6. Le gras s'oppose à la viande de première qualité pour le premier facteur mis en évidence, alors que le second oppose os à carcasse. Le premier facteur permet toujours de séparer charolais des zébus mais avec un certain recouvrement quatre zébus sont à gauche du charolais CB; mais surtout on voit disparaitre le poids vif comme caractère marquant dans la mesure où il révélait un facteur de taille qui est ici éliminé.

2.7. Premieres conclusions: analyses factorielles ou classification?

Les analyses en composantes principales ou en correspondance sont généralement regroupées sous la vocable d'*analyse factorielle*. Il existe un autre type d'analyse factorielle plus ambitieuse et relevant davantage d'un *modèle* qui n'existe pas ici, il est beaucoup plus complexe, et pour cette raison nous ne l'aborderons pas.

Les analyses factorielles permettent donc de faire surgir d'un tableau complexe un certain nombre de facteurs dont on peut espérer qu'il sont significatifs. Naturellement la technique mathématique ne fait que révéler des axes de variabilité maximum, elle ne garantit pas qu'ils sont facilement interprétables; elle fournit seulement un moyen objectif de dégager une structure d'une masse a priori informe. C'est un *excellent outil de réduction de données*; mais elle ne prend pas de décision; elle ne permet pas de dire qu'une observation est un charolais ou est un zébu, elle n'établit pas de frontières strictes. La classification établit, elle, des coupures franches; mais nous avons vu que ces coupures réalisées à partir de règles algorithmiques sont fondées sur des règles quelquefois bien arbitraires: la nature se prête rarement à un cloisonnement.

Alors, que conseiller à celui qui doit analyser un tableau de données? Les analyses factorielles ont l'avantage de ne rien imposer et par là doivent être conseillées; si toutefois des décisions doivent être prises des classes faites, la classification employée avec grande prudence n'est pas à rejeter. Mais surtout c'est par un retour aux sources, c'est-à-dire aux données observées qu'on évi-

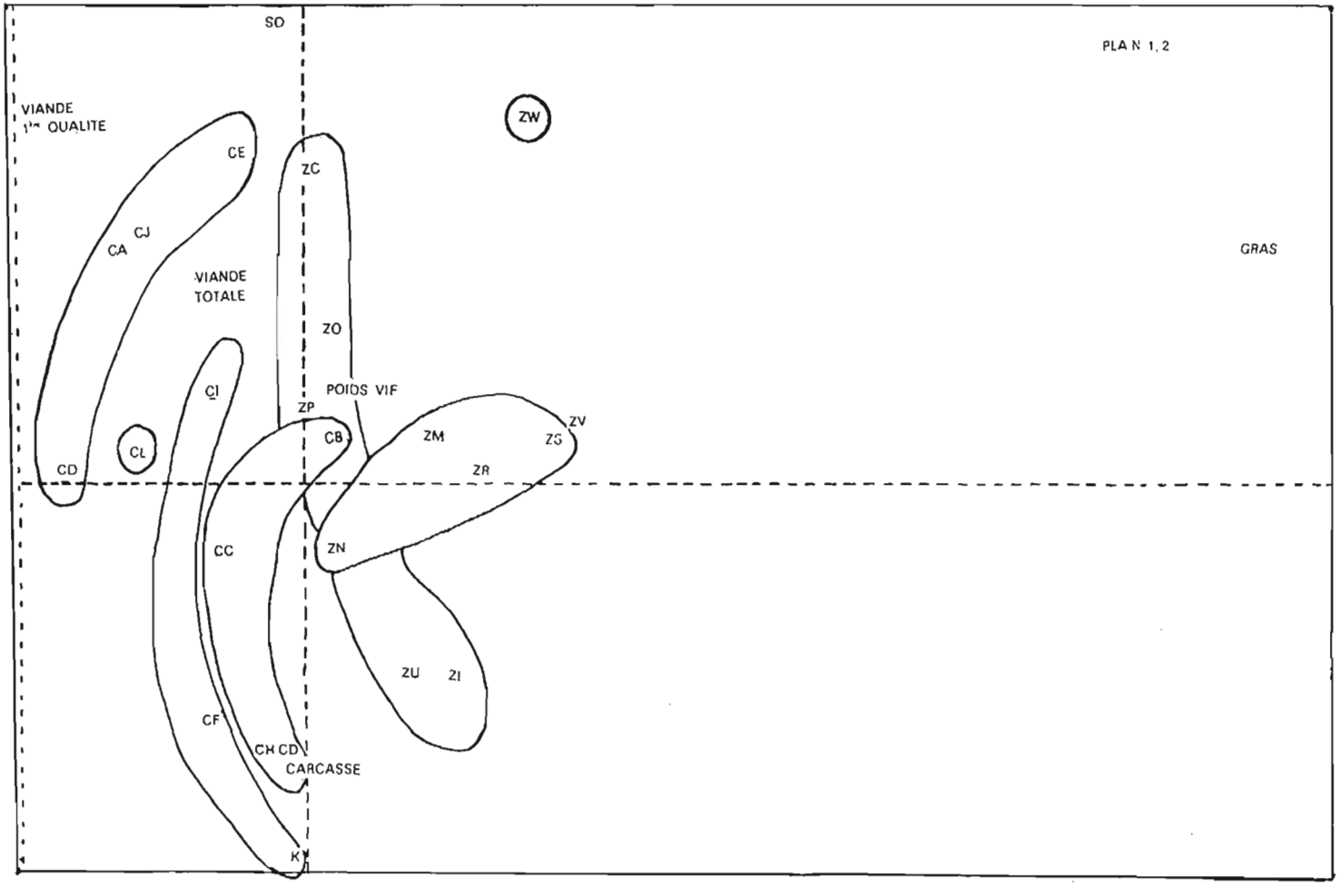

Figure 6. – Projection des 6 variables et des 23 bovins dans le plan des deux premiers facteur de l'analyse des correspondances.

tera tout risque important d'erreur; au delà, un emploi judicieux de chaque technique permet d'éclairer sous un jour différent des données ce qui n'étaient qu'une série de techniques d'interprétation peut devenir un art.

3. PREMIER EXEMPLE: MENSURATIONS DE BOVINS: STRUCTURATION A PRIORI.

3.1. L'ANALYSE DISCRIMINANTE A DEUX POPULATIONS.

3.1.1.

Les différents types d'analyse factorielle précédents nous ont offert dejà de très bons moyens d'ordonner nos observations; très souvent, et lorsqu'elles sont employées avec rigueur les analyses factorielles sont largement suffisantes. Mais on peut vouloir tenir compte d'informations supplémentaires connues a priori sur les données; ici par exemple nous savons que ces bovins appartiennent à deux races bien connues: les charolais et les zébus. Nous n'avons pas encore introduit cette information dans l'analyse, nous l'avons observée « après coup ».

L'analyse discriminante va permettre de l'introduire, mais nous allons être obligés de préciser davantage ce que nous souhaitons faire: est il possible à l'aide des observations faites sur les deux races d'affirmer qu'elles sont distinctes? On sent poindre l'outil probabiliste! Il va falloir nous donner une loi de probabilité de chaque race, et établir si les recouvrements éventuels des deux races, ont une signification.

Plus précisément, on va tenter de séparer le domaine de variation de toutes nos observations, notre ensemble référentiel, en deux ensembles disjoints, l'un sera la région des charolais, l'autre celle des zébus. L'analyse discriminante va permettre de définir une frontière entre les deux ensembles et nous dirons que d'un côté se trouvent les charolais, de l'autre les zébus. La façon la plus simple de trouver la frontière est de supposer que la distribution à l'intérieur de chaque race est multivariate gaussienne, la frontière est alors une combinaison linéaire des variables d'origine.

3.1.2.

Plus précisement, nous devons *supposer* qu'il existe une variabilité moyenne à l'intérieur de chaque race, et que les deux races ne diffèrent que par la position du « centre » des nuages des observations; si nous nous reportons à la figure 7, nous devons nous placer dans les cas représentés par la partie *a*): les populations sont d'autant mieux discriminées que les zones de recouvrement

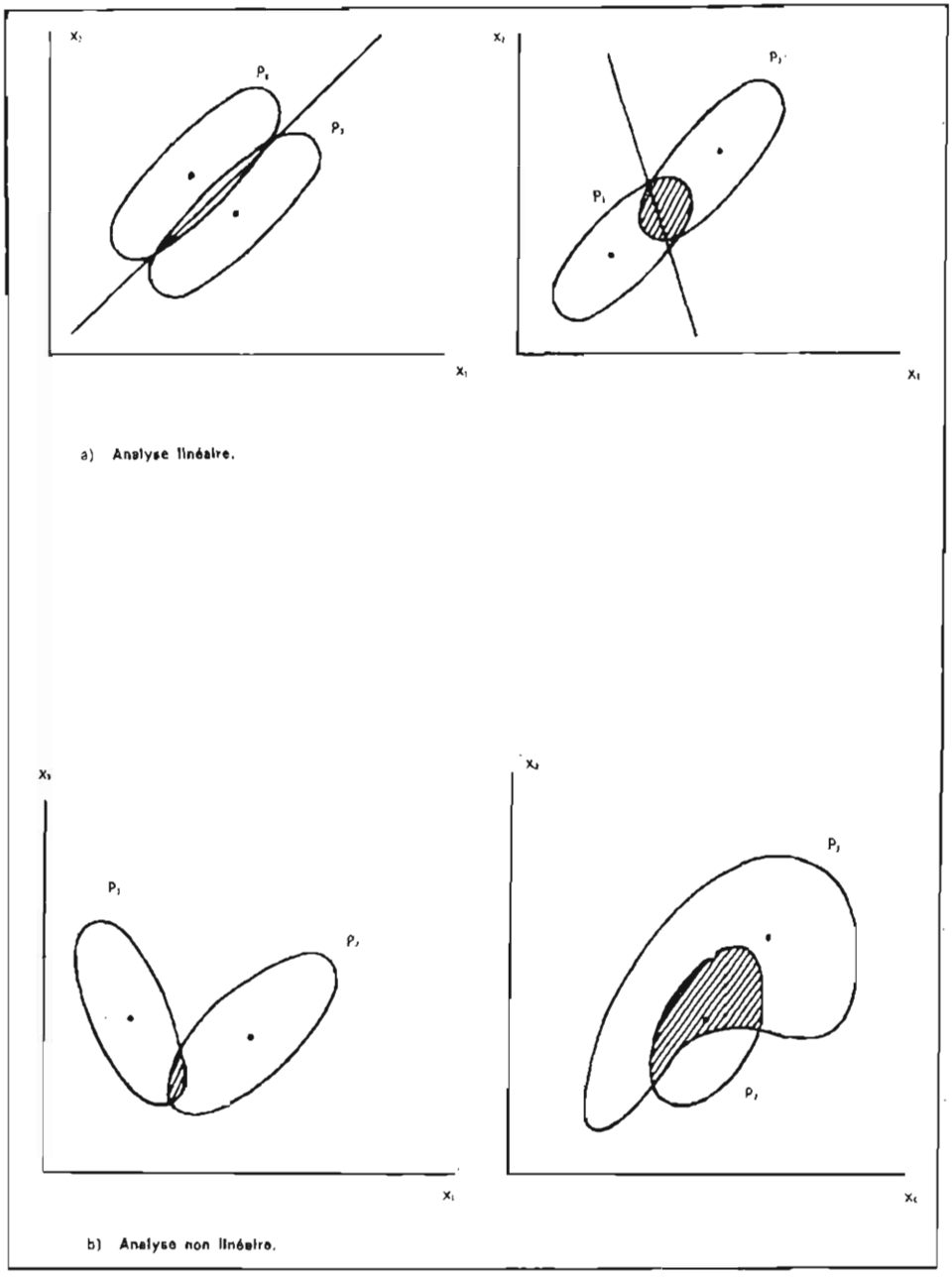

Figure 7. – Exemple de figures intervenant en analyse discriminante : *a*) analyse linéaire ; *b*) analyse non linéaire.

des observations sont plus petites. La valeur respective des points moyens a naturellement son importance, mais ce n'est pas tout: alors qu'à gauche de la figure 7*a*) les moyennes sont plus proches qu'à droite il sera plus facile néanmoins de séparer les deux populations. Ceci tient au fait que nous prenons en compte *deux niveaux de variation*:

1) l'un entre les populations, défini par les positions des deux points moyens;

2) l'autre à l'intérieur des populations défini par le nuage « moyen » de chacune d'elles.

On peut alors déterminer une fonction linéaire des variables qui représente la frontière entre les deux populations, elle nous est indiquée à la dernière ligne du tableau 5:

$$(11) \quad V(\boldsymbol{x}) = 19.8 - 0.0105x_1 - 0.0077x_2 - 0.1309x_3 - 0.1105x_4 - \\ - 0.0740x_5 - 0.0584x_6$$

$V(x)$ nous permet ensuite de faire un *classement* des observations (la constante étant ici choisie pour qu'une valeur négative corresponde à une observation classée comme zébu). Cette fonction appliquée aux éléments de notre échantillon global permet de classer a posteriori les observations de base, mais son utilisation sur ces éléments donne une vue optimiste du pouvoir discriminatoire et en particulier de la probabilité de mauvais classement fournie par la procédure. Il ne faudrait pas conclure à la vue du classement que $V(x)$ classe à coup sûr en charolais ou en zébu car il faut faire intervenir la taille de l'échantillon, le nombre de variables mais alors le problème devient nettement plus complexe.

3.1.3. *Raffinements techniques.*

Une fois la méthode bien appliquée et bien comprise, il est possible d'en améliorer l'utilisation grâce à quelques modalités techniques.

1) On peut ainsi introduire les variables l'une après l'autre, c'est ce qu'on appelle une méthode de *pas à pas*; on cherche d'abord la variable qui discrimine au mieux les deux races, dans ce cas c'est x_3, on obtient ainsi une première fonction discriminante (cf. tableau 6*a*):

$$(12) \quad V_1(x) = 5.1085 - 0.1707x_6 .$$

On cherche ensuite une deuxième variable associée à la première qui fournisse à nouveau le meilleur discriminateur

$$(13) \quad V_2(x) = 8.5963 - 0.0174x_2 - 0.1541x_3$$

TABLEAU 6. - *Analyse discriminante pas à pas entre charolais et zébu; classement a posteriori des observations:* a) *fonction discriminante pas à pas: entre parenthèses ecart type des coefficients de la fonction discriminante;* b) *classement a posteriori des observations à l'aide de la fonction* $V(x) = h + \sum a_i x_i$ *en se limitant aux deux variables* x_2 *et* x_3.

a) Estimation pas à pas de la fonction discriminante.

		1	2	3	4	5	6		
Palier	h	Poids vif	Carcasse	Chair 1ère	Chair totale	Gras	Os	F	D^2
1	5,1085			—0,1707 (221) **				59,52	10,41
2	8,5963		—0,0174 (58) **	—0,1541 (197) **				45,26	16,54
3	10,2888	—0,0065 (88)	—0,0127 (87)	—0,1593 (211) **				29,66	17,12
4	13,5293	—0,0109 (100)	—0,0073 (103)	—0,1204 (454) **	—0,0516 (532)			22,41	18,20
5	14,1672	—0,0105 (104)	—0,0076 (107)	—0,1235 (486) *	—0,0588 (633)	—0,0119 (527)		17,00	18,30
6	19,7775	—0.0105 (106)	—0,0077 (109)	—0,1309 (524) *	—0,1105 (1312)	—0,0740 (1474)	—0,0584 (1287)	13.54	18,69

b) Classement à l'aide de la fonction discriminante à 2 variables.

	N°. observation	$V(x)$	P_1	P_2	
C	1	— .73	.67	.33	$V(x)$ = fonction discriminante
	2	— .29	.59	.41	< 0 pour charolais
	3	— .23	.55	.45	> 0 pour zébu
	4	— .21	.57	.43	
	5	— .12	.52	.48	P_i $(i = 1, 2)$ probabilité
	6	— .65	.64	.36	d'appartenance à la
	7	— .29	.58	.42	population i.
	8	— .36	.59	.41	
	9	— .60	.62	.38	
	10	— .42	.62	.38	
	11	— .42	.61	.39	
	12	— .47	.59	.41	
Z	13	.31	.41	.59	
	14	.15	.48	.52	
	15	.28	.45	.55	
	16	.35	.43	.57	
	17	.34	.42	.58	
	18	.36	.40	.60	
	19	.59	.34	.66	
	20	.50	.37	.63	
	21	.44	.41	.59	
	22	.63	.36	.64	
	23	.85	.29	.71	

il est bien clair que ce ne sont pas des deux variables les plus discriminantes isolément (ce serait dans ce cas x_3 et x_4) qui donnent la meilleure fonction discriminante à deux variables; ceci est dû à la liaison qui existe entre x_3 — la meilleure variable isolée — et les deux autres pour lesquelles interviennent aussi bien la valeur absolue du coefficient que son signe (— 0.364 pour x_3 et x_2, et 0.634 pour x_3 et x_4). Nous ne nous étendrons pas davantage sur cette technique qui peut-être d'un intérêt très grand pour limiter le nombre global de variables à introduire pour discriminer deux populations (*).

2) On peut aussi analyser d'abord la structure interne de la variabilité en calculant les axes principaux des nuages de la figure 7, (c'est-à-dire les vecteurs propres et les valeurs propres de la matrice S du tableau 5). Ceci a un premier avantage qui est de bien montrer que l'analyse directe en composantes principales que nous avons déjà vue ne fournit pas les mêmes résultats puisque nous ne nous attaquons pas au même nuage de points. Le second avantage permet de bien voir que ce qui définit bien une *structure interne* n'est pas obligatoirement ce qui permet de bien discriminer deux populations; à gauche (de la figure 7*a*) ce sera sûrement la deuxième composante qui aura un poids plus grand dans la discrimination, alors qu'à droite ce sera la première composante.

Ce résultat se voit numériquement au tableau 5, puisque la fonction discriminante en fonction des composantes principales s'écrit:

$$(14) \qquad V(x) = 0.458y_1 - 0.444y_2 - 0.586y_3 - 0.459y_4 - 0.039y_5 - 0.191y_6$$

les quatres premières composantes ont un poids pratiquement égal dans la discrimination.

3.2. Notion de distance statistique.

Nous voyons qu'avec quelques hypothèses supplémentaires (même variabilité interne correspondant à des distributions gaussiennes) il est possible de tracer une frontière entre les deux populations; on peut aussi la «tester»: ce test revient à se demander si la zone de recouvrement (hachurée sur les graphiques de la figure 7) est grande par rapport aux zones où l'on trouve les populations parfaitement isolées.

Cette zone permet de définir une *probabilité de mauvais classement*, ou une quantité qui lui est fonctionnellement liée la *distance de Mahalanobis* cette distance intègre à la fois la distance géometrique entre les populations (c'est-à-dire la distance entre les points moyens) et les corrélations entre les variables. Naturellement

(*) On peut encore améliorer cette technique en supprimant des variables déjà introduites aux paliers précédents; mais naturellement ces techniques ne trouvent leur intérêt véritable qu'avec un grand nombre de variables.

cette distance est la même qu'elle soit calculée sur les variables de départ ou sur leurs composantes principales ; cette invariance est vraie pour toute transformation linéaire des variables d'origine.

3.3. Cas ou les hypotheses ne sont pas respectees.

Sans entrer dans les détails des analyses qui sont schématisées par les figures 7*b*) on peut traiter des cas plus généraux et sans nul doute plus proches de la réalité. Analytiquement les calculs sont plus complexes mais réalisables sur ordinateur :

— on peut trouver une frontière linéaire, dans le cas de gauche, bien qu'il ne s'agisse pas de la frontière qu'on aurait déterminé en faisant l'hypothèse de même forme des nuages (Anderson and Bahadur (1962)),

— on peut trouver une frontière non linéaire dans le cas de droite (Victor (1973)).

Nous ne nous étendons pas davantage sur ces analyses, puisque nous ne faisons ici qu'un exposé de présentation.

3.4. Analyse discriminante a plusieurs populations.

Ces résultats peuvent s'étendre au cas où nous savons qu'il existe plusieurs populations ; tout ce que nous avons vu dans les trois paragraphes précédents peut être appliqué. Naturellement les aspects analytiques sont plus compliqués, mais surtout l'interprétation est généralement plus délicate nous verrons une de ces analyses au paragraphe suivant. Ces analyses, qu'elles soient à deux ou plusieurs populations peuvent trouver un enrichissement profond par une *approche décisionnelle*, nous ne nous y attarderons pas ici. L'important réside en ce que nous avons deux niveaux distincts de variabilité : la première résiduelle traduisant la dispersion indépendante de tout facteur contrôlé et la seconde entre les moyennes des populations, c'est en utilisant « au mieux » ces deux niveaux que nous parviendrons à la meilleure discrimination.

3.5. L'analyse des correlations canoniques.

3.5.1.

De même qu'il peut exister une structure en ligne a priori on peut aussi imaginer une structure en colonne : les variables peuvent être groupées par exemple en deux ensembles ayant chacun une réalité physique. On peut vouloir étudier la liaison entre ces deux ensembles : sont ils fortement liés ou au contraire indépendants? il s'agit, dans l'analyse des corrélations canoniques, de rechercher deux combinaisons linéaires (une pour chaque ensemble) qui aient le plus

grand coefficient de corrélation entre elles. Dans notre exemple de bovins nous allons chercher si entre poids vif et carcasse d'un côté et les quatre dernières variables de l'autre il existe une liaison; nous pourrions songer à appliquer cette analyse pour tenter de prédire un index de conformation à deux mesures globales de poids.

3.5.2. *Corrélations canoniques sur les bovins.*

Ici encore on peut passer par une analyse de la structure interne de chaque ensemble de variables, mais naturellement le coefficient de corrélation que nous allons trouver sera le même qu'il soit calculé sur les composantes principales des groupes ou sur les variables des groupes.

Comme la technique revient à étudier des liens entre les deux ensembles, nous allons écrire la matrice globale des coefficients de corrélation:

$$R = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} \tag{15}$$

R_{11} représente les coefficients de corrélation entre les variables du premier ensemble, R_{22} entre celles du deuxième, et R_{12} (ou R_{21}) entre celles dex deux ensembles. Si les valeurs de R_{12} étaient toutes nulles, nous conclurions immédiatement à l'indépendance des deux ensembles. Ici le plus grand des coefficients est celui entre x_2 et x_4 égal à 0.390; la recherche d'un coefficient de corrélation plus important que 0.390 va montrer l'intérêt de l'étude des corrélations canoniques.

Nous arrivons ici à une valeur de 0.657 donc un gain appréciable (même s'il n'est pas significatif (*)); si nous regardons quelles sont les combinaisons linéaires qui interviennent, on a d'après la tableau 7:

1) pour le premier ensemble: $-0.620x_1 + 0.785x_2$;

2) pour le second: $-0.575x_3 + 0.744x_4 - 0.337x_5 - 0.045x_2$;

ce résultat analytique obtenu, il nous faut bien sûr l'interpréter, et cette opposition de variables peut nous suggérer qu'une variation relative de la viande première qualité (x_3) et du gras (x_5) par rapport à la viande totale (x_4) est fortement liée à une variation relative de la carcasse (x_2) par rapport au poids vif (x_1). Nous ne nous attarderons pas ici sur une analyse détaillée, nous tenons simplement à montrer la puissance d'une méthode et son potentiel, nous verrons plus tard comment extraire toute l'information contenue dans ces analyses qui représentent des *extensions des techniques de regression.*

(*) Nous notons ici qu'un test est fondé sur l'utilisation du χ^2 qui sera étudié par ailleurs.

TABLEAU 7. - *Analyse des corrélations canoniques bovines.*

	Poids vif (x_1)						
	1.000	Poids carcasse (x_2)					
	.691	1.000	Viande 1-ère qualité (x_3)				
	—.033	.283	1.000	Viande totale (x_4)			
	—.058	.390	.895	1.000	Gras (x_5)		
	.101	—.328	—.861	—.908	1.000	Os (x_6)	
	—.110	—.100	—.063	—.128	—.263	1.000	
λ	1.692	0.308	2.779	1.090	.120	.011	COMPOSANTES
λ%	84.60	15.40	69.48	27.25	3.00	.27	PRINCIPALES DE
x_1	.71	—.71	—	—	—	—	CHAQUE GROUPE
x_2	.71	.71	—	—	—	—	
x_3	—	—	.57	—.11	—.81	.07	
x_4	—	—	.58	—.16	.49	.63	
x_5	—	—	—.58	—.22	—.32	.72	
x_6	—	—	.02	.96	—.08	.28	
	ξ_1	ξ_2	η_1	η_2	η_3	η_4	

COEFFICIENTS DE CORRELATION ENTRE COMPOSANTES PRINCIPALES DES 2 GROUPES.

ξ_1	.151	—.120	.079	.022
ξ_2	.528	.004	.352	.015

CORRELATION CANONIQUE.

a) En fonction des composantes principales.

						μ	χ^2	d.l.
.266	.964	.835	—.043	.547	.031	.657	11.31 NS	8
.964	—.266	.041	—.977	—.148	.147	.120	.28 NS	3

b) En fonction des variables de depart

—.620	.785	—.575	.744	—.337	—.045	.657	11.31 NS	8
.978	.211	.322	.482	.768	—.273	.120	.28 NS	3

3.6. L'UNITE DES METHODES D'ANALYSE DE MATRICES PARTITIONNEES.

Nous avons volontairement présenté nos deux types d'analyse de façon séparée parce que c'est ainsi qu'elles se présentent physiquement. Il est possible de fournir une vue synthétique de l'analyse discriminante et de l'analyse des corrélations canoniques par l'introduction pour l'un des deux ensembles de variables indicatrices ($x_i = 1$ ou 0 selon que l'observation appartient à la population définie par x_i). Même plus, on peut montrer que l'analyse factorielle des

correspondances peut se présenter comme un cas particulier de l'analyse des corrélations canoniques, celui où les deux ensembles de variables sont chacun consititué de variables indicatrices.

Cette méthode identique pour trois aspects que nous avons exposés séparément, présente un intérêt dans la mesure où chacun des aspects « classiques » apporte son cortège technique c'est-à-dire son mode d'interprétation:

1) l'analyse discriminante apporte une optique proche de celle de l'analyse de variance: un test global (par un F de Snedecor ou un D^2 de Mahalanobis) et les comparaisons de moyenne, voire un aspect décisionnel;

2) l'analyse des corrélations canoniques apporte une technique d'interprétation voisine de la regression;

3) l'analyse factorielle des correspondances apporte des techniques d'interprétation issues de l'analyse des tables de contingence.

Nous pouvons donc affirmer que la connaissance détaillée de chaque technique fournit un enrichissement global qui peut ensuite être utilisé pour les autres. L'aspect synthètique que nous venons de présenter a un intérêt essentiellement didactique pour donner un peu de hauteur de vue à ce qui ne pourrait paraître qu'un ensemble disparate de techniques, voire dans le pire des cas de « recettes » d'analyse et d'interprétation.

4. DEUXIEME EXEMPLE: DE L'ANALYSE AUX HYPOTHESES NOUVELLES.

4.1. Pour une analyse dynamique.

Nous avons vu l'esprit général sous jacent à l'ensemble des méthodes de l'analyse des données multidimensionnelles; nous allons maintenant, sur un nouvel exemple agronomique, montrer comment on peut à partir d'hypothèses de départ élémentaires, faire une première analyse, puis à partir des résultats obtenus tenter d'aller plus loin dans la connaissance des phénomènes profonds.

4.2. Les donnees.

Il s'agit d'analyser des données permettant de comparer quatre types parentaux de pommiers à l'aide de six variables (*); ces types sont respectivement:

Delicious Spur (DS);
Delicious normal (DN);
Golden Spur (GS);
Golden Normal (GN);

(*) Ces données fournies par Monsieur Decourtye de la Station d'Arboriculture d'Angers font partie d'un ensemble plus vaste dont nous avons extrait ce qui illustre plus précisement notre présentation générale.

et les variables:

x_1 nombre d'inflorescences sur bois de deux ans;

x_2 longueur du bois de deux ans;

x_3 diamètre du bois de deux ans;

x_4 nombre d'inflorescences sur bois de un an;

x_5 longueur du bois de un an;

x_6 diamètre du bois de un an.

Les données de base et les premières statistiques élémentaires sont fournies au tableau 8.

4.3. Choix d'une methode.

Nous nous trouvons devant des données qui manifestement demandent à être traitées par une analyse discriminante à plusieurs populations (*) ou analyse factorielle discriminante. Nous avons déjà évoqué en 3.4 cette méthode, nous allons ici en dégager le principe. Les personnes habituées à l'analyse de variance, verront très rapidement qu'il s'agit d'une extension. En analyse de variance, on dispose de deux niveaux différents de variabilité:

1) l'un dû à la variabilité « naturelle » à l'intérieur de chaque population; cette variabilité est caractérisée par une variance résiduelle (on dit aussi variance d'erreur);

2) l'autre est dû à la variabilité que l'on attribue aux différences entre populations.

Tout consiste à savoir si compte tenu de la première, la seconde est suffisamment grande pour être jugée « significative ». Si nous retournons à la figure 7, ceci revient à se demander si la zone de recouvrement des points appartenant à des populations différentes est suffisamment faible.

Le test qui en découle est un test F de Fischer-Snedecor. Dans le cas où nous disposons de plusieurs variables x_i, on va chercher une combinaison linéaire de ces variables de départ telle que le rapport des deux variabilités soit maximum. On a la combinaison linéaire

$$z = a_1 x_1 + \ldots + a_6 x_6 \tag{16}$$

(*) On retrouve sous des noms différents des techniques assez voisines: analyse discriminante, analyse de dispersion, MANOVA (Multivariate Analysis of Variance), analyse des variables canoniques; depuis peu le terme *analyse factorielle discriminante* est utilisé, c'est celui que nous emploierons désormais.

TABLEAU 8. - *Etude de 4 types parentaux de pommiers données (a) et statistiques élémentaires (b et c).*

x_1	x_2	x_3	x_4	x_5	x_6	Identification		
6	485	10.0	0	335	6.5	DS	1	1
15	380	7.5	0	175	4.5		1	2
14	435	10.0	4	405	6.5		1	3
0	505	11.5	0	555	7.5		1	4
2	495	9.0	0	375	5.5		1	5
0	400	9.5	0	420	6.0		1	6
0	375	8.5	0	345	6.0		1	7
0	310	7.5	0	210	4.5		1	8
3	290	6.0	2	425	5.0	GN	2	1
2	260	8.0	3	655	6.5		2	2
5	445	10.0	15	720	8.5		2	3
9	555	9.5	14	640	7.0		2	4
1	280	6.5	3	425	5.5		2	5
9	555	9.0	13	640	6.5		2	6
1	340	7.0	2	430	5.0		2	7
10	505	8.0	1	335	4.5		2	8
4	515	8.5	0	410	5.0	DN	3	1
7	530	10.5	2	410	6.0		3	2
7	465	7.5	2	365	5.0		3	3
3	350	9.0	2	415	5.5		3	4
5	430	7.0	5	290	4.5		3	5
2	435	8.5	2	370	5.5		3	6
7	530	10.5	2	415	5.5		3	7
5	435	7.0	5	290	4.5		3	8
3	140	9.0	12	480	7.5	GS	4	1
2	190	6.5	1	140	5.0		4	2
10	330	9.5	1	240	5.5		4	3
2	230	10.5	12	610	8.5		4	4
3	145	10.5	17	555	8.0		4	5
2	225	10.0	12	615	8.0		4	6
5	285	9.0	1	115	5.5		4	7
4	285	9.0	1	210	5.5		4	8

a) Les données de base.

Moyennes	DS	4.6	423.1	9.19	0.5	352.5	5.88
	GN	5.0	403.8	8.00	6.6	533.8	6.06
	DN	5.0	461.3	8.56	2.5	370.6	5.19
	GS	3.9	228.8	9.26	7.1	370.6	6.69
Moyenne generale		4.6	379.2	8.75	4.2	406.9	5.95
Ecart type residuel		4.1	85.4	1.37	4.7	145.3	1.13
$F(3, 28)$		.13	11.67	1.47	3.74	2.74	2.37

b) Les statistiques univariates.

W_r						B_r					
1.00						1.00					
.47	1.00					.93	1.00				
.14	.54	1.00				−.78	−.49	1.00			
.10	.10	.48	1.00			−.38	−.69	−.29	1.00		
−.19	.08	.72	.74	1.00		.45	.12	−.87	.57	1.00	
−.13	.15	.79	.75	.94	1.00	−.81	−.91	.36	.67	.67	1.00

c) Matrices de correlation intrapopulations (W_r) et interpopulations (B_r).

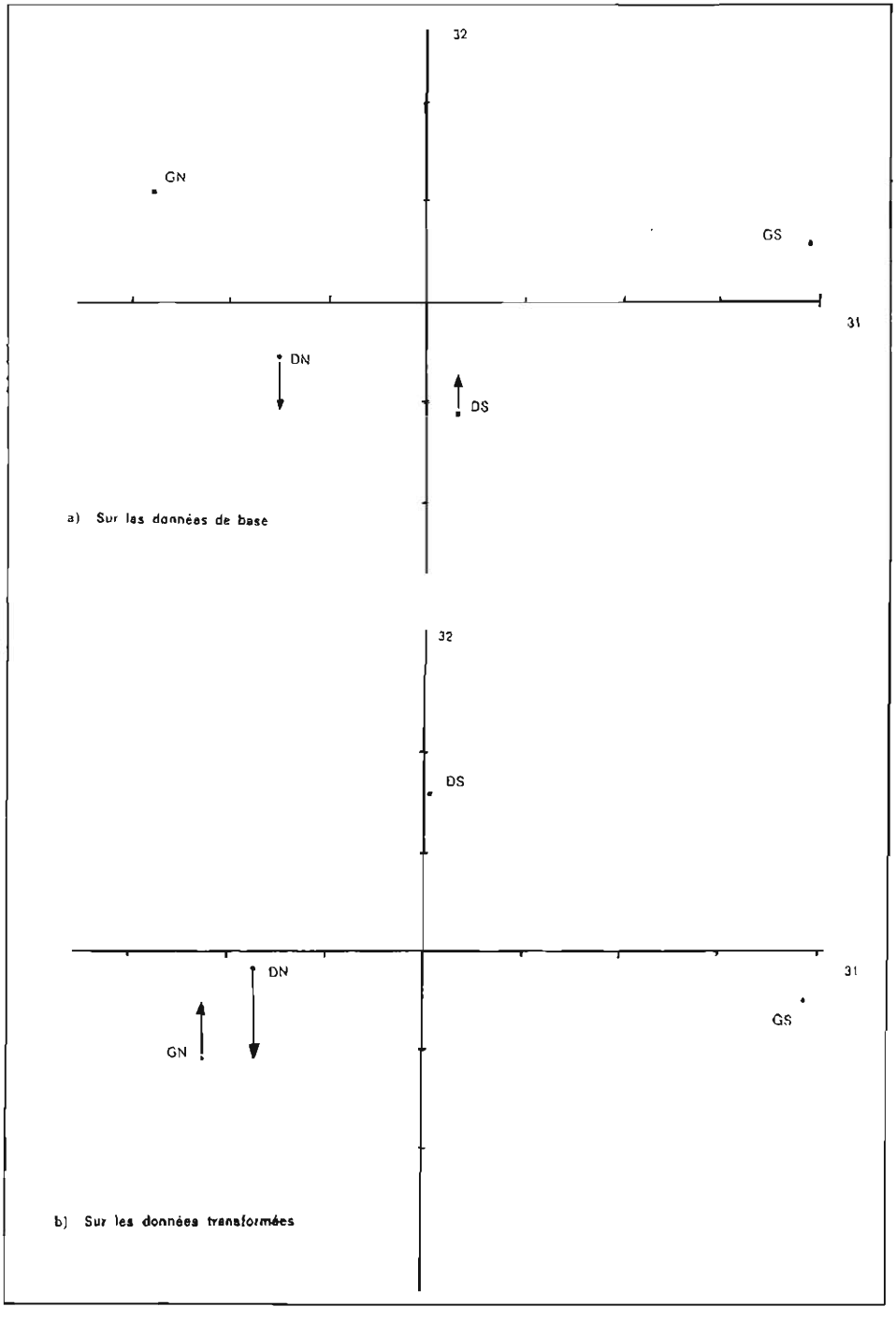
32
GN
GS
31
DN
DS
a) Sur les données de base
32
DS
DN
31
GS
GN
b) Sur les données transformées

Figure 8.

et il s'agit de déterminer les coefficients a_i; on voit qu'on a un problème très voisin de celui que nous avons analysé en 2.3.

On démontre que γ est le vecteur propre associé à la plus grande valeur de $\boldsymbol{W}^{-1}\boldsymbol{B}$ où:

— $\boldsymbol{W}$ est la matrice de variances et covariances intrapopulations; celle dont la traduction géométrique est le *nuage « moyen »* des points d'une population qui est représenté à la figure 8*a*) dans deux positions différentes.

— $\boldsymbol{B}$ est la matrice de variances et covariances interpopulations; géométriquement elle fait intervenir les seuls points moyens des populations avec un poids proportionnel au nombre d'observations de chacune d'elle (*).

On va trouver un nombre de vecteurs propres égal au plus petit des deux nombres p et $k-1$ où:

p = nombre de variables,

k = nombre de populations.

Ici avec $p = 6$ et $k = 4$, on trouvera trois vecteurs propres; ce seront les trois axes factoriels discriminants qui ont, comme en analyse en composantes principales, la propriété d'être non corrélés entre eux. On peut même faire un test qui est indiqué sur les tableau 9 mais dont nous ne donnerons ici aucune justification théorique.

4.4. Les premiers resultats et les idees nouvelles qui sont suggerees.

Dans les analyses à une variable il fallait que le F à 3 et 28 degrés de liberté soit supérieur à 2.95 (seuil 5%) ou à 4.57 (seuil 1%) pour que nous concluons à la signification d'une variable. Seules étaient significatives la longueur des bois à deux ans (x_2) au seuil 1%, et le nombre d'inflorescences à un an (x_4) au seuil: 5%; d'après le tableau 9*a*) nous pouvons dire qu'il existe deux facteurs discriminants significatifs, dont l'un, le premier, très fortement. Examinons-le, on peut écrire:

$$\gamma_1 = -0.1x_1 - 1.0x_2 + 1.3x_3 + 0.5x_4 - 2.4x_5 + 1.2x_6 . \tag{17}$$

Il est bien clair que le nombre d'inflorescences tant à deux ans qu'à un an n'ont pas un grand poids dans ce facteur; par contre, apparaissent bien nettement $x_3 - x_2$ et $x_6 - x_5$. Et ce facteur isole bien GS et GN qui apparaissent opposés: ceci est bien net car GS a toujours des valeurs de diamètre élevées pour

(*) Nous avons conservé la notation anglaise $\boldsymbol{W}$ comme Within et $\boldsymbol{B}$ comme Between.

TABLEAU 9. - *Analyse factorielle discriminante sur données de base pommiers.*

Variables canoniques	z_1	z_2	z_3
valeur propre	67.917	9.006	3.817
χ^2	81.43**	26.48**	8.92 NS
d.l.	18	10	4
x_1	— .11	.40	.56
x_2	—1.01	—.54	.33
x_3	1.29	—.60	—1.69
x_4	.46	.48	—1.54
x_5	—2.42	.79	— .29
x_6	1.17	—.19	3.11

TABLEAU 9*a*. - *Analyse factorielle discriminante*. Valeurs propres de $W^{-1}B$, approximation par un χ^2 et degrés de liberté. Coefficients des variables de départ pour les trois axes factoriels discriminants (variables canoniques)

Populations	Variables canoniques		
	z_1	z_2	z_3
DS	.36	—1.13	.73
GN	—2.77	1.13	.32
DN	—1.51	— .65	—.88
GS	3.92	.65	—.17

TABLEAU 9*b*. - *Analyse factorielle discriminante*. Coordonnées des quatre populations sur les trois nouveaux axes.

DS	0. 0			
GN	15.02	0. 0		
DN	6.28	6.17	0. 0	
GS	16.67	45.28	31.66	0.0

TABLEAU 9*c*. - *Distances de Mahalanobis entre les 4 populations.*

ce qui nous fait donc penser à un *facteur de forme des bois* l'un étant trapu (GS) l'autre plus fin (GN).

Après une première analyse de ce type on sent donc que s'il était possible de faire intervenir le rapport des variables diamètre/longueur on aurait sûrement un facteur interessant.

De même, un coup d'oeil sur les valeurs des inflorescences nous permet de vérifier que les distributions sont très dissymétriques. Ce résultat est classique avec des données provenant de comptage; généralement on conseille de *transformer ces variables* en prenant les racines carrées; comme les rapports ne sont pas faciles à manipuler directement on peut aussi prendre les logarithmes et

ses différentes hauteurs, faire la différence des logarithmes des hauteurs et des diamètres ; ce passage aux logarithmes est d'ailleurs classique lui aussi pour toutes les données correspondant à des mensurations. Tout ceci nous conduit à une seconde analyse.

4.5. Deuxieme analyse.

Nous allons donc faire une analyse factorielle discriminante sur les six nouvelles variables :

$$(18)\qquad \begin{cases} x_1' = \sqrt{x_1} & x_4' = \sqrt{x_4} \\ x_2' = \mathrm{Ln}(x_3/x_2) & x_5' = \mathrm{Ln}(x_6/x_5) \\ x_3' = \mathrm{Ln}(x_3) & x_6' = \mathrm{Ln}(x_6) \end{cases}$$

nous avons conservé x_3' et x_6' dans la mesure où le rapport seul ne résume pas toute l'information des deux variables isolées. Il est déjà bien clair que x_2' et x_5' sont très significatives, il en est d'ailleurs de même pour x_4' comme on peut le voir au tableau 10.

Tableau 10. - *Analyse factorielle discriminante sur les données pommiers transformées.*

Variables transformées		y_1	y_2	y_3	y_4	y_5	y_6
Moyennes	DS	1.43	−3.83	2.21	.25	−4.05	1.76
	GN	2.07	−3.89	2.06	2.31	−4.46	1.78
	DN	2.19	−3.99	2.14	1.44	−4.26	1.64
	GS	1.89	−3.18	2.22	2.31	−3.85	1.88
Moyenne générale		1.90	−3.72	2.16	1.58	−4.16	1.77
Ecart type résiduel		1.04	.23	.16	1.05	.28	.18
$F(3, 28)$		.82	21.33 **	1.51	6.87 **	7.43 **	2.27

Tableau 10*a*. - *Statistiques univariates sur données transformées.*

W_r						B_r					
1.00						1.00					
−.38	1.00					−.18	1.00				
.16	.21	1.00				−.66	.61	1.00			
.18	.42	.43	1.00			.71	.42	−.44	1.00		
−.10	.52	.80	−.55	1.00		−.51	.81	.96	−.16	1.00	
.18	.42	.43	.69	−.55	1.00	−.32	.86	.35	.43	.54	1.00

Tableau 10*b*. - *Matrices de correlation intrapopulations (W_r) et interpopulations (B_r) sur les données transformées.*

L'analyse factorielle discriminante résumée au tableau 11 confirme naturellement notre première idée, puisque le nouvel axe discriminant z_1 est fonction presqu'uniquement de ces deux variables de forme à un an et à deux ans: GS et à un moindre degré DS sont plus trapus que GN et DN (*).

TABLEAU 11. - *Analyse factorielle discriminante sur données pommiers transformées.*

Variables canoniques	z_1	z_2	z_3
valeurs propres	62.83	11.36	5.18
χ^2	85.35 **	32.17 **	11.48 *
d.l.	18	10	4
y_1	.03	— .18	.44
y_2	1.07	— .19	— .19
y_3	.46	.47	—1.80
y_4	.04	—1.06	— .93
y_5	1.09	.14	.06
y_6	— .05	.41	2.64

TABLEAU 11*a*. - *Analyse factorielle discriminante*. Valeurs propres de $W^{-1}B$, approximation par un χ^2 et degrés de liberté. Coefficients des variables transformées sur les trois axes factoriels discriminants (variables canoniques).

Populations	Variables canoniques		
	z_1	z_2	z_3
DS	.01	1.69	.39
GN	—2.24	—1.01	.76
DN	—1.71	— .08	—1.10
GS	3.95	— .61	— .04

TABLEAU 11*b*. - *Analyse factorielle discriminante*. Coordonnées des quatre populations sur les trois nouveaux axes.

DS	0.0			
GN	12.51	0.0		
DN	8.27	4.62	0.0	
GS	21.03	39.20	33.38	0.0

TABLEAU 11*c*. - *Distances de Mahalanobis entre les* 4 *populations*.

(*) La valeur du logarithme du rapport étant négative, les valeurs moyennes des populations sont grandes pour GS et DS que pour GN et DN.

Le deuxième facteur donne le poids le plus important à x'_4 et isole indépendamment du premier

$$\mathrm{DS}(\overline{x}'_4 = 0.25) \quad \text{et} \quad \mathrm{GN}(\overline{x}'_4 = 2.31)\,.$$

quant au troisième il fait intervenir $x'_3(-1.80)$ et $x'_6(2.64)$ c'est-à-dire une différence de croissance en diamètre entre un an et deux ans, et sépare sur ce nouveau critère GN et DN. Pour ce dernier il faut bien remarquer que les diamètres intervenaient déjà dans la forme traduite par le premier, et qu'ils apparaissent ici sous un aspect différent.

4.6. Un probleme nouveau : le classement.

Nous n'avons pas encore évoqué le problème du *classement* (qu'il ne faut pas confondre avec celui de la *classification*); il découle assez naturellement de l'analyse factorielle discriminante. Ayant, par cette dernière, discriminé des groupes, peut-on sur un nouvel échantillon déterminé par les mêmes variables le rattacher à l'un des groupes préalablement défini? On conçoit que cette opération est réalisée sans peine en calculant la position de cet échantillon dans l'espace factoriel discriminant ou en calculant sa distance aux groupes déjà définis et en l'affectant à l'un d'entre eux en s'imposant une règle de décision; par exemple l'affecter au groupe dont il est le plus proche.

Naturellement sous réserve d'hypothèses supplémentaires on peut même donner une probabilité d'appartenance à chacun des groupes initiaux.

4.7. En guise de conclusion.

Naturellement nous n'avons fait qu'une esquisse des possibilités d'une série de méthodes. Il est bien évident que nous avons passé sous silence un nombre important de problèmes qu'il va falloir aborder par la suite: le respect des hypothèses, les modèles sous-jacents à chaque méthode, les limites d'utilisation.

Nous avons voulu montrer ici le potentiel de ces méthodes et faire comprendre leur portée. L'apparente simplicité de leur utilisation ne doit pas nous masquer les difficultés; le risque le plus grand réside à notre sens dans l'utilisation inconsidérée d'un ordinateur: en effet, ces méthodes pour être appliquées imposent l'utilisation d'un ordinateur.

On a donc quelquefois tendance à oublier les données, c'est-à-dire qu'on ne trace plus d'histogrammes univariantes, qu'on n'étudie plus les distributions à l'intérieur des populations. On risque donc, par paresse, de redécouvrir par ordinateur ce qui était évident à l'oeil; sans nous attarder sur ce domaine qui déborde largement celui de l'analyse multidimensionnelle nous pouvons conclure en disant une fois encore que l'usage de l'ordinateur ne doit pas empêcher de penser et de réfléchir.

BIBLIOGRAPHIE

[1] ANDERSSON T. W., *An introduction to multivariate statistical analysis*, J. Wiley, New York, 1959, p. 374.

[2] ANDERSSON T. W. - BAHADUR R. R., *Classification into two multivariate normal distributions with different covariance matrices*, Am. Math. Stat., **33**, 420 (1962).

[3] BADIOU A., *Le concept de modèle F. Maspero*, Paris, 1972, p. 94.

[4] BENZECRI J. P., *L'analyse des données*, Tome 1: *La taxinomie*, p. 624; Tome 2: *L'analyse des correspondances*, p. 624, Dunod, Paris, 1973.

[5] CAILLEZ F. - MAILLES J. P. - NAKACHE J. P. - PAGES J. P., *Analyse des données multidimensionnelles*, Centre d'Etudes Economiques d'Entreprises, Paris, 1971, 3 tomes.

[6] DAGNELIE P., *Introduction à l'analyse statistique à plusieurs variables*, Fac. Sc. Agro. Etat, Gembloux, Belgique, 1974, à paraître.

[7] GOWER J. C., *Adding a point to vector diagrams in multivariate analysis*, Biometrika, **55**, 582-585 (1968).

[8] LAWLEY D. N. - MAXWELL A. E., *Factor analysis as a Statistical Method*, Butterworths, Londre, 1971, p. 153.

[9] VICTOR N., *A non linear discriminant analysis*, Computer Programs in Biomedecine, **2**, 36-50 (1971).

[10] PEARSON E. S. - HARTLEY H. O. ed., *Biometrika tables for statisticians*, vol. 1, Cambridge Univ. Press, 1970, p. 287.

[11] AFIFI A. A. - AZEN S. P., *Statistical Analysis: a Computer Oriented Approach*, Academic Press Inc., New York, 1972, p. 384.