Bull. Soc. roy. belge Anthrop. Préhist. 99 : 35-51, 1988

Les fluctuations des coefficients de corrélation en paléontologie humaine

par

André LEGUEBE

INTRODUCTION

Depuis les premiers travaux de Galton, la notion d'association entre mesures joue en anthropologie un rôle fondamental :

« It is easy to see that co-relation must be the consequence of the two organs being partly due to common causes. If they were wholly due to common causes, the co-relation would be perfect ... If they were in no respect due to common causes, the co-relation would be nil. Between these two extremes are an endless number of intermediate cases, and it will be shown how the closeness of co-relation in any particular case admits of being expressed by a single number » (Galton, 1888).

Ce nombre connu d'abord sous l'appellation de « réversion » et de « fonction de Galton » sera désigné par Edgeworth sous le nom de coefficient de corrélation (Pearson, 1920). Dès 1888, Galton avait remarqué que, quand chaque déviation est mesurée en écarts-type, le coefficient de régression s'identifie au coefficient de corrélation.

Dans la suite, Pearson, professeur de mathématiques appliquées de University College à Londres, a largement contribué à préciser divers aspects théoriques et pratiques de ce coefficient : dès 1894, il fait acheter une machine à calculer pour réaliser les calculs indispensables qui étaient relativement lourds. Il a été le promoteur de nombreux travaux visant à approfondir la signification des associations entre mensurations et leur portée pour l'interprétation des données craniométriques en particulier (Pearson E.S., 1938).

C'est en 1895 que Pearson publie les premiers coefficients de corrélation entre la longueur et la largeur du crâne pour trois séries d'origines géographiques différentes (Bavarois, Français et Naqada).

En 1896, il aborde le problème des corrélations factices (spurious) qui peuvent découler de l'utilisation d'indices, en se servant pour cela des mesures de longueur, de largeur et de hauteur et des indices qui en découlent pour une série de 100 crânes de Bavarois. Déjà, dans une note infrapaginale, il envisage de considérer les différences qui peuvent être observées en fonction de la race et du sexe notamment.

En annexe de ce travail, Weldon publie une note mentionnant les résultats d'une épreuve de tirages d'échantillons aléatoires, la première de son espèce probablement, et illustrant l'effet que peut avoir, sur la valeur du coefficient de corrélation, une opération de standardisation des mensurations par une même mesure les transformant ainsi en indices.

Lee, en 1899, publie des coefficients plus nombreux et elle considère les variations en fonction de la race. La même année, Boas apporte quelques résultats complémentaires relatifs à des Amérindiens, des Eskimo et des Indiens du Bengale en tentant, par l'intermédiaire des coefficients de corrélation des caractères entre eux, de réduire la corrélation apparente entre la longueur et la largeur de la tête par élimination des effets des autres facteurs qui pourraient être considérés comme des « causes ». Il conclut que : « while the cephalic index is a convenient practical expression of the form of the head, it does not express any important anatomic relation » (p. 461) et il évoque la possibilité d'estimer, à partir des mensurations céphaliques, la capacité crânienne qu'il n'est pas possible de mesurer sur le vivant.

En 1902, Fawcett sur la base d'une série « homogène » de 400 squelettes Naqada récoltés en Egypte par Flinders Petrie, analyse de manière détaillée les relations entre une vingtaine de variables du crâne : elle souligne qu'il n'est probablement pas justifié d'étendre à d'autres séries les résultats obtenus pour une race :

« ...to classify a few individuals into different races by means of two or three measurements, such as the cephalic index, the length or the facial angle, — before the correlation and the variation of these characters have been determined for even a single race — is a very dangerous proceeding, and calculated to bring craniometry into discredit » (Fawcett, 1902, p. 409).

En 1924, en collaboration avec Davin, Pearson publie le résultat de quatorze années de recherches du *Biometric Laboratory,* consacrées à une série masculine de 700 à 900 crânes et une série féminine de 400 à 600 crânes provenant d'Egypte. L'objectif est de dégager à partir de l'étude d'un seul groupe homogène et d'effectif important, ce que pourraient être les caractères structuraux du crâne chez l'*Homo sapiens*. Une attention particulière est accordée aux coefficients de corrélation pour lesquels il établit un classement en fonction de la nature des variables couplées et de l'importance de la valeur atteinte par les coefficients : Pearson souligne que les résultats se rapportent uniquement à la variabilité intraraciale. L'existence d'une variabilité interraciale au niveau du degré d'association entre les variables a été maintes fois mise en évidence et elle se trouve confirmée par les résultats malheureusement très partiels publiés par Howells (1973, table 50) concernant les crânes masculins et féminins de 17 populations.

SIGNIFICATION DES ASSOCIATIONS

En paléontologie humaine, aux variations que l'on rencontre naturellement, s'ajoutent des fluctuations qui résultent de différents facteurs comme la nature des

échantillons eux-mêmes, la proportion relativement importante de valeurs manquantes et la présence de sujets marginaux par certaines de leurs mensurations; tous ces facteurs sont susceptibles de cumuler leurs effets au niveau du calcul des coefficients de corrélation (Leguebe, 1989).

Ces fluctuations des valeurs des coefficients de corrélation peuvent avoir des répercussions considérables au niveau de l'analyse multivariée des données puisque c'est de la matrice de corrélation que sont extraits les vecteurs propres qui servent à définir le nouvel espace, et les valeurs propres correspondantes qui partitionnent la variabilité globale en parties indépendantes.

Je me propose donc, dans cet article, d'essayer de cerner l'importance des fluctuations susceptibles d'affecter les coefficients de corrélation, et de tenter de préciser leur origine. A titre d'exemple, j'utilise l'échantillon de crânes du Paléolithique supérieur mesurés par Riquet (1970) et Billy (1972) sous le rapport de douze variables qui sont :

1. capacité crânienne,
2. longueur maximum,
3. largeur maximum,
4. largeur frontale minimum,
5. largeur frontale maximum,
6. hauteur basion-bregma,
7. largeur bizygomatique,
8. hauteur faciale supérieure,
9. largeur orbitaire,
10. hauteur orbitaire,
11. largeur nasale,
12. hauteur nasale.

Pour aucune de ces variables, l'hypothèse de normalité de la distribution, éprouvée au moyen d'un test de Filliben, n'est rejetée (Leguebe et Albert, 1989).

Le calcul de la matrice de corrélation peut se réaliser de deux manières différentes soit qu'on ne tienne compte, dans le calcul des coefficients, que des objets pour lesquels on dispose des valeurs pour toutes les variables, soit qu'on utilise pour chaque couple de variables le nombre maximum d'objets disponibles. Chacune de ces façons de procéder a des avantages et présente des inconvénients mais, en principe, aucune n'est a priori supérieure à l'autre. Dans la situation qui nous occupe, où les effectifs dont nous disposons sont généralement faibles, les tests classiques et les estimations d'intervalles de confiance ne sont pratiquement d'aucune utilité.

Les deux matrices sont données dans le tableau 1. Il est possible au moyen d'une procédure fishérienne non paramétrique (Lebart et al., 1979, p. 116) d'éprouver l'hypothèse nulle de l'indépendance entre deux variables : on compare pour une statistique choisie, dans ce cas la variable aléatoire «produit scalaire», le résultat

	1	2	3	4	5	6	7	8	9	10	11	12
1	—	.760	.420	.649	.554	.687	.650	.568	.255	-.021	.180	.528
2	.796	—	.446	.589	.597	.587	.701	.728	.577	-.056	.305	.589
3	.572	.588	—	.505	.580	.007	.734	.232	.194	-.202	-.005	.307
4	.355	.477	.676	—	.635	.310	.728	.497	.423	-.170	.225	.522
5	.659	.713	.683	.694	—	-.063	.467	.325	.157	-.347	.367	.329
6	.454	.394	-.174	-.258	.160	—	.437	.555	.162	.181	.131	.524
7	.460	.634	.625	.566	.440	.156	—	.619	.525	-.077	.236	.472
8	.446	.634	.119	.350	.391	.413	.510	—	.480	.409	.374	.872
9	-.188	.211	.044	.096	.000	-.153	.326	.294	—	.165	.431	.348
10	-.253	-.172	-.414	-.373	-.430	.195	-.232	.393	.089	—	.057	.406
11	.070	.165	-.260	.097	.166	.187	-.002	.372	.397	.165	—	.273
12	.494	.596	.229	.489	.500	.466	.441	.832	.074	.280	.365	—

Tableau 1. Matrice de corrélation de Pearson : au-dessous de la diagonale, les valeurs correspondent aux vecteurs de données complets (N = 18); au-dessus de la diagonale, les valeurs correspondent au nombre maximum d'objets pour chaque couple de variables (N = max.).

observé pour l'échantillon initial à ceux obtenus pour toutes ou une partie des configurations résultant de l'appariement des données au hasard. Si la probabilité de l'hypothèse nulle est très faible, on acceptera l'hypothèse alternative de l'existence d'une dépendance monotone croissante entre les deux variables.

Pour 300 échantillons tirés au hasard, les probabilités d'obtenir un échantillon présentant une valeur aussi extrême de la statistique que celle observée, sont données dans le tableau 2 pour les divers couples de variables. Toutes les valeurs marquées par * (27 au total) sont proches de zéro, ce qui signifie que, pour les paires considérées, l'association des variables dans l'échantillon de crânes du Paléolithique supérieur, est réelle; dans quinze autres cas, cette probabilité est inférieure à 0.05, ce qui conduit également à rejeter l'hypothèse d'indépendance. Pour les 24 dernières associations au contraire ($p > 0.05$), on peut considérer que le coefficient de corrélation obtenu est dépourvu de signification.

LES COEFFICIENTS DE CORRELATION DE RANGS

Une autre possibilité de mesurer le degré d'association entre deux séries de mensurations nous est offerte par les coefficients de corrélation de rangs : ils présentent cette propriété de ne pas être influencés par la présence, au sein d'un échantillon d'effectif faible, d'un objet caractérisé par des valeurs marginales (Kendall, 1938). En fait, on observe fréquemment, pour une série de mensurations, un coefficient de corrélation élevé entre la mesure elle-même et son rang (Kendall, 1970, p. 126). Pour des échantillons tirés d'une population normale, on a la relation suivante :

$$E(t) = 2 \sin^{-1} \rho/\pi \text{ (cn radians)}$$

et il est donc possible de procéder à une estimation du coefficient de corrélation à partir de la valeur du *tau* de Kendall, soit :

$$r = \sin 0.5 \pi t$$

Le tableau 3 donne le résultat des valeurs transformées des coefficients de Kendall correspondants dans le cas où on n'a considéré que les objets pour lesquels on a toutes les mensurations (matrice triangulaire inférieure) et dans le cas où on a utilisé le nombre maximum de paires pour lesquelles on dispose des mesures (matrice triangulaire supérieure).

Il est évident qu'en utilisant les rangs plutôt que les valeurs primitives, on néglige une partie de l'information initiale mais on peut supposer avec une certaine vraisemblance qu'il s'agit de cette partie de l'information qui est susceptible de conduire à des résultats plus labiles.

ROLE DES VARIABLES

Dès que le nombre de variables utilisées est quelque peu important, la comparaison visuelle des différentes matrices de corrélation devient impossible. On peut

	1	2	3	4	5	6	7	8	9	10	11	12
1	—											
2	*	—										
3	*	*	—									
4	0.01	*	*	—								
5	*	*	*	*	—							
6	*	*	0.48	0.03	0.43	—						
7	*	*	*	*	*	0.02	—					
8	*	*	0.15	*	0.07	0.02	*	—				
9	0.14	0.01	0.12	0.01	0.22	0.21	*	0.03	—			
10	0.45	0.35	0.14	0.16	0.04	0.16	0.35	0.01	0.17	—		
11	0.16	0.06	0.46	0.10	0.02	0.24	0.12	0.04	*	0.41	—	
12	*	*	0.05	*	0.06	0.01	0.05	*	0.02	*	0.08	—

Tableau 2. Résultats des tests de dépendance monotone par la méthode de Fisher : les tests ont été réalisés sur le nombre maximum d'objets pour chaque couple de variables (nombre de tirages au hasard : 300).

	1	2	3	4	5	6	7	8	9	10	11	12
1	—	.805	.456	.530	.574	.589	.743	.560	.296	-.131	.249	.598
2	.791	—	.435	.503	.607	.593	.728	.779	.586	-.135	.311	.588
3	.620	.643	—	.605	.597	-.067	.808	.310	.247	-.173	.028	.327
4	.419	.483	.799	—	.604	.143	.777	.574	.463	-.139	.301	.587
5	.735	.740	.746	.749	—	-.099	.495	.313	.082	-.351	.485	.304
6	.310	.339	-.255	-.253	.233	—	.422	.558	.233	.246	.087	.623
7	.627	.740	.717	.698	.576	.254	—	.703	.567	-.117	.270	.616
8	.435	.719	.217	.498	.442	.380	.676	—	.555	.300	.297	.839
9	-.178	.156	.128	.085	-.181	-.179	.325	.360	—	.143	.482	.389
10	-.394	-.242	-.489	-.339	-.299	.307	-.288	.110	.075	—	-.009	.394
11	.107	.096	-.185	.109	.227	.076	.098	.165	.350	.088	—	.282
12	.509	.636	.315	.533	.544	.523	.638	.760	.063	.296	.214	—

Tableau 3. Matrice de corrélation de Kendall : au-dessous de la diagonale, les valeurs correspondent aux vecteurs de données complets (N = 18) ; au-dessus de la diagonale, les valeurs correspondent au nombre maximum d'objets pour chaque couple de variables.

déterminer dans quelle mesure les coefficients de corrélation de chacune des variables avec l'ensemble des autres, sont en moyenne plus ou moins élevés. Le calcul des moyennes s'effectue en utilisant la transformation z de Fisher des valeurs absolues des coefficients : pour les quatre matrices de corrélation, les moyennes des coefficients incluant les diverses variables, accompagnées de leur rang, sont données dans le tableau 4. Les figures 1 et 2 illustrent, dans le cas des coefficients de corrélation pearsoniens et dans le cas des valeurs transformées des *tau* de Kendall, la relation qui existe entre les moyennes de chacune des variables pour les matrices obtenues à partir de l'effectif N = 18 et de l'effectif N = max.

Variables	Coefficient de Pearson				Coefficient de Kendall transformé			
	N = max		N = 18		N = max		N = 18	
1	.508	4	.457	5	.530	4	.495	5
2	.563	1	.519	1	.580	2	.548	1
3	.353	9	.425	6	.399	8	.504	4
4	.496	6	.423	7	.497	6	.487	6
5	.417	7	.469	2	.426	7	.532	3
6	.353	8	.279	9	.353	10	.287	9
7	.538	3	.415	8	.602	1	.542	2
8	.545	2	.457	4	.558	3	.461	8
9	.347	10	.173	12	.380	9	.192	11
10	.194	12	.276	10	.197	12	.271	10
11	.239	11	.208	11	.261	11	.157	12
12	.498	5	.459	3	.529	5	.481	7
Moyenne	.428		.385		.451		.421	

Tableau 4. Moyennes et rangs des coefficients de corrélation (après application de la transformation de Fisher) pour chacune des variables dans les quatre matrices de corrélation.

Dans chacune des situations, il y a deux groupes de variables, les unes, les variables 10 (hauteur orbitaire), 3 (largeur maximum) et 5 (largeur frontale maximum), pour lesquelles la moyenne est plus élevée avec N = 18, et toutes les autres pour lesquelles la moyenne avec N = 18 est au contraire plus basse. D'un autre point de vue, les variables 6 (hauteur basion-bregma), 9 (largeur orbitaire), 10 (hauteur orbitaire) et 11 (largeur nasale) sont caractérisées par les moyennes les plus faibles, les variables 3 et 5 par des moyennes intermédiaires et les autres ont les moyennes les

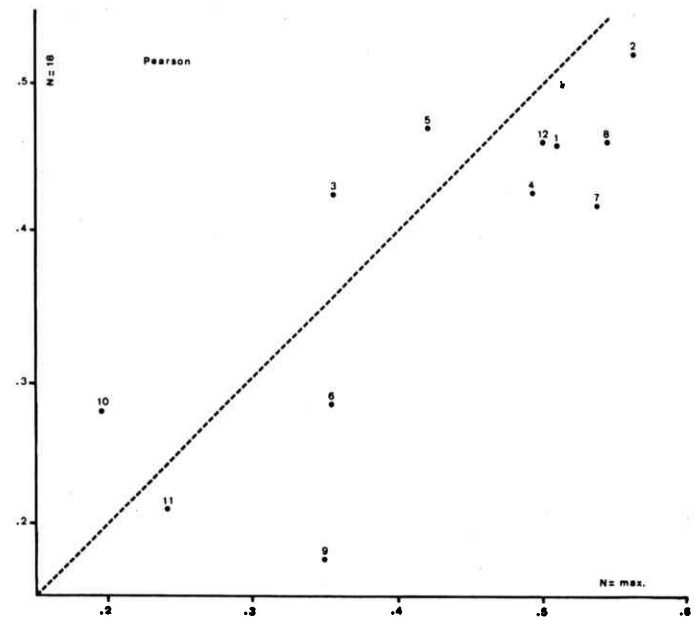

Figure 1. Relation entre les moyennes des coefficients de corrélation avec N = max. et N = 18 pour la matrice de corrélation de Pearson.

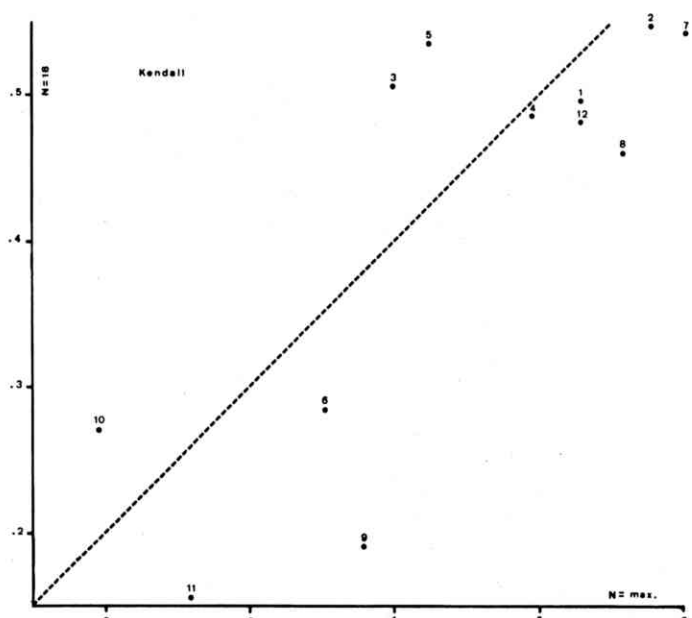

Figure 2. Relation entre les moyennes des coefficients de corrélation avec N = max. et N = 18 pour la matrice de corrélation du *tau* de Kendall transformé.

plus fortes, tout particulièrement la variable 2 (longueur maximum). La plus ou moins grande discordance entre les moyennes d'une même variable pour les deux matrices d'effectifs différents est donnée par la distance par rapport à la diagonale.

D'autre part, le coefficient de corrélation entre deux variables exprime dans quelle mesure l'information fournie par l'une des variables est semblable à celle contenue dans l'autre variable : il peut donc être assimilé à un coefficient de similarité.

En soumettant une matrice de corrélation à une analyse en coordonnées principales, on définit les axes orthogonaux d'un nouvel espace dans lequel on peut situer les variables : ces axes constituent les nouvelles variables qui sont des fonctions linéaires des variables primitives et qui sont indépendantes les unes des autres (Gower, 1966; Hills, 1969).

Les valeurs propres et les pourcentages de la variabilité totale auxquels elles correspondent, figurent dans le tableau 5 : les valeurs obtenues pour la matrice des valeurs transformées du *tau* de Kendall ne sont pratiquement pas différentes de celles obtenues pour la matrice de corrélation de Pearson.

	r (Pearson)				tau (Kendall)			
Vecteur	N = max		N = 18		N = max		N = 18	
	E. val	%	E. val	%	E. val	%	E. val	%
1	2.13	31.0	2.87	36.0	1.81	22.4	2.34	26.6
2	1.31	19.0	1.64	20.5	1.27	15.7	1.45	16.4
3	0.95	13.8	0.94	11.7	1.05	13.0	1.02	11.6
4	0.80	11.6	0.85	10.7	0.91	11.3	0.93	10.5
5	0.59	8.5	0.66	8.2	0.72	8.9	0.77	8.7
Trace	6.90		7.99		8.11		8.79	

Tableau 5. Valeurs propres des matrices de corrélation de Pearson et de Kendall transformées pour les douze variables.

Au contraire, pour la matrice de corrélation de Kendall, les traces sont plus grandes et les deux premiers vecteurs expliquent une part relativement plus faible de la variabilité totale. On constate également que les traces des matrices et les pourcentages des deux premiers vecteurs sont plus élevés pour les matrices obtenues à partir des vecteurs de données complets.

Les figures 3 et 4 représentent la position des extrémités des vecteurs « variables » pour les deux premiers axes du nouvel espace défini par l'analyse en coordonnées principales des matrices de corrélation de Pearson et de Kendall, respectivement. Sur chacune des figures, sont données les valeurs correspondant aux résultats de l'analyse faite sur les coefficients de corrélation obtenus pour le nombre maximum de paires (● surmonté du numéro de la variable) et sur les coefficients correspondant aux vecteurs complets (□ avec le numéro de la variable souscrit).

Les résultats présentent, pour les matrices de Pearson et de Kendall, une similitude générale de l'ensemble : toutefois les discordances peuvent être, en fonction de l'effectif utilisé, plus ou moins prononcées selon les variables. Le cas le plus frappant est celui de la largeur frontale maximum (5) et de la largeur bizygomatique (7). Pour N = max, la largeur bizygomatique (7) est associée à la capacité crânienne (1) et à la longueur maximum (2) alors que la largeur frontale maximum (5) est associée à la largeur maximum (3) et à la largeur frontale minimum (4); on observe l'inverse pour N = 18. La largeur orbitaire (9) et la largeur nasale (11) présentent également une inversion de l'importance de leur rôle selon l'effectif utilisé.

Pour l'échantillon de crânes du Paléolithique supérieur, les différences entre les deux types de coefficients de corrélation paraissent négligeables. Le recours au *tau* de Kendall transformé à la place du coefficient de Pearson reste toutefois une possibilité intéressante dans le cas où l'hypothèse de normalité est rejetée : la distribution de r peut en effet être très sensible à la non-normalité (Kowalski, 1972).

On pourrait vérifier si le schéma obtenu pour cette série de crânes du Paléolithique supérieur répond à un modèle relativement général en soumettant au même type d'analyse des matrices de corrélation relatives à d'autres échantillons; toutefois pour les échantillons dont on dispose en paléontologie, il semble plus important d'essayer de détecter dans quelle mesure certains sujets peuvent être responsables, en fonction de leurs mensurations, des fluctuations du coefficient de corrélation.

INFLUENCE DES SUJETS DEVIANTS

Etant donné le caractère souvent limité de l'effectif, il est important de repérer ceux parmi les objets qui sont susceptibles d'influencer la valeur du coefficient de corrélation : le fait que l'on ne tienne pas compte d'un objet dans le calcul des coefficients de corrélation n'exclut d'ailleurs pas la possibilité de le positionner dans l'espace nouvellement défini si on dispose de toutes les données pour cet objet. La démarche la plus classique consistant à ne pas inclure dans les calculs les sujets qui sont distants de la moyenne d'un nombre choisi d'écarts-type, est évidemment à éviter parce qu'elle néglige les effets éventuels des associations entre les variables : elle

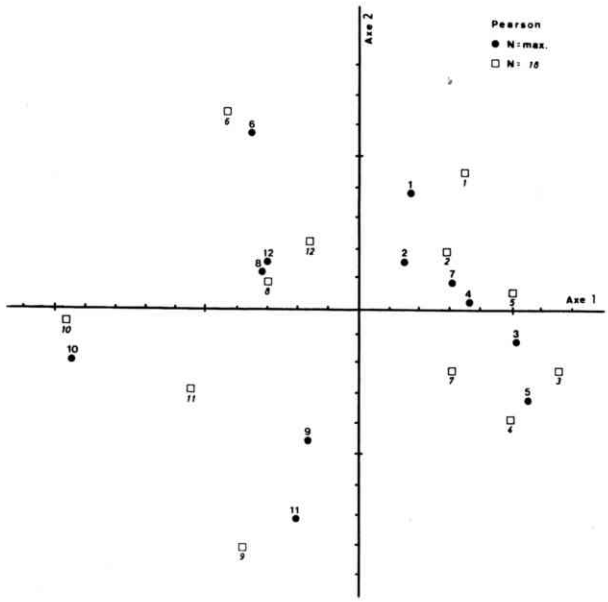

Figure 3. Distribution des vecteurs «variables» dans l'espace des deux premiers axes de l'analyse en coordonnées principales de la matrice de corrélation pearsonienne (● : matrice avec N = max; □ : matrice avec N = 18).

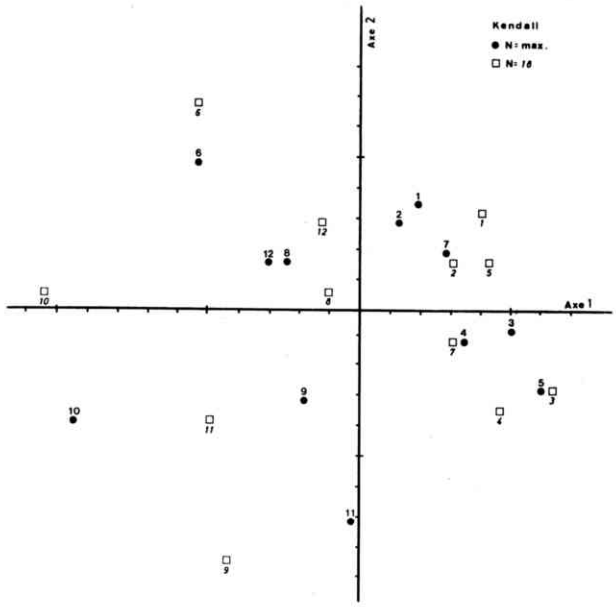

Figure 4. Distribution des vecteurs «variables» dans l'espace des deux premiers axes de l'analyse en coordonnées principales d'une matrice de corrélation de Kendall (● : matrice avec N = max; □ : matrice avec N = 18).

est susceptible d'engendrer, dans le cas de petits échantillons, des erreurs importantes (Leguebe, 1989).

Divers procédés pour identifier, dans la situation bivariée, la présence d'individus marginaux ont été proposés, certains d'entre eux pouvant être étendus à une situation multivariée (Devlin et al., 1975). Vu que l'intérêt se porte ici sur le coefficient de corrélation, nous avons adopté une technique développée par Quenouille et Tukey, visant à réduire ou à supprimer le biais d'un estimateur et à fournir une estimation approchée mais robuste d'un intervalle de confiance autour de l'estimation (Lebart et al., 1979, p. 71).

Pour chaque couple de variables d'une matrice de données, ces estimations utilisent les coefficients de corrélation d'une matrice de données pour tous les ensembles d'objets moins un, qu'il est possible de constituer : après avoir classé leurs valeurs par ordre croissant, on repère ceux de ces coefficients qui sont le plus marginaux et les sujets dont l'élimination a entraîné soit une diminution marquée, soit une augmentation prononcée de sa valeur. Deux exemples suffiront à illustrer l'importance que peut avoir sur la valeur du coefficient de corrélation l'inclusion ou l'exclusion de certains objets.

Considérons le couple de variables largeur frontale maximum (5) / largeur bizygomatique (7), variables pour lesquelles on avait constaté dans l'analyse en coordonnées principales, une inversion de leur position par rapport à l'axe 1. Le coefficient de corrélation vaut 0.467 et l'estimation de Quenouille-Tukey basée sur le calcul de 26 coefficients, est égale à 0.451. L'élimination soit de Brno III (26), soit de la Grotte des Enfants (5), dont la position dans l'espace bivarié est donnée dans la figure 5, ramène le coefficient respectivement à 0.374 et 0.387; au contraire sa valeur atteint 0.589 ou 0.502 si on élimine ou Obercassel I (35) ou Lautsch I (14).

Pour le couple de variables largeur orbitaire (9) / largeur nasale (11), c'est l'importance de leur intervention dans l'espace des deux premiers axes de l'analyse en coordonnées principales qui se trouve inversée. Le coefficient de corrélation vaut 0.431 et son estimation 0.426.

La position des crânes dont l'élimination de l'échantillon a pour effet d'engendrer des variations importantes du coefficient de corrélation, est donnée dans la figure 6. L'omission de Combe-Capelle (3) fait tomber le coefficient à 0.328; au contraire l'omission soit de Cro-Magnon I (1), d'Obercassel (35) ou de Grimaldi (25) a pour conséquence de porter la valeur du coefficient respectivement à 0.509, 0.506 et 0.500.

On observe que les estimations de Quenouille-Tukey sont en général assez proches de la valeur du coefficient calculé sur l'ensemble de l'échantillon : en effet, pour un échantillon d'effectif N, une paire de valeurs marginales n'est en fait éliminée qu'une fois du calcul alors qu'elle est incluse N - 1 fois.

D'un point de vue pratique, la valeur du coefficient de corrélation la meilleure à adopter pour les analyses multivariées, sera celle qui évite, dans la mesure du possi-

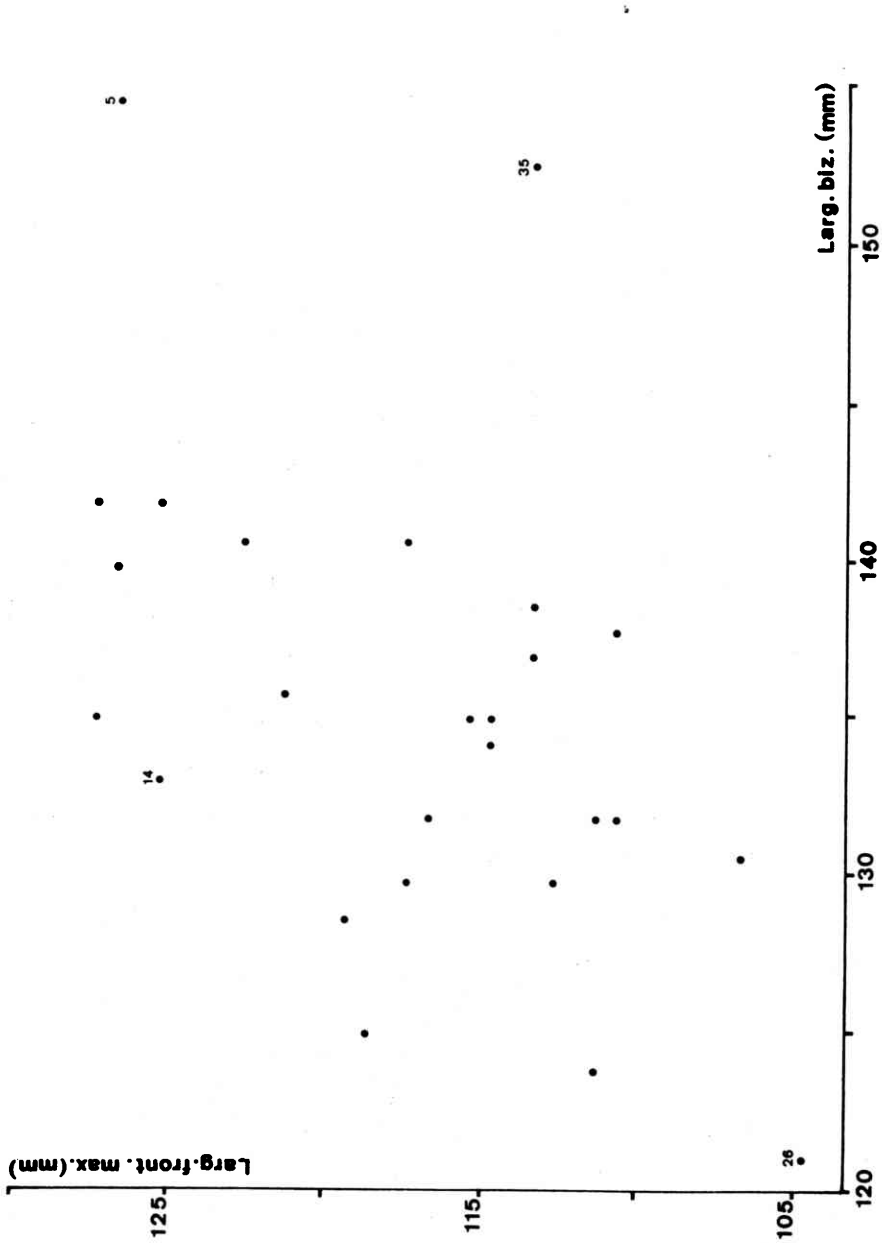

Figure 5. Distribution des crânes du Paléolithique supérieur (N = 26) en fonction de la largeur bizygomatique et de la largeur maximum.

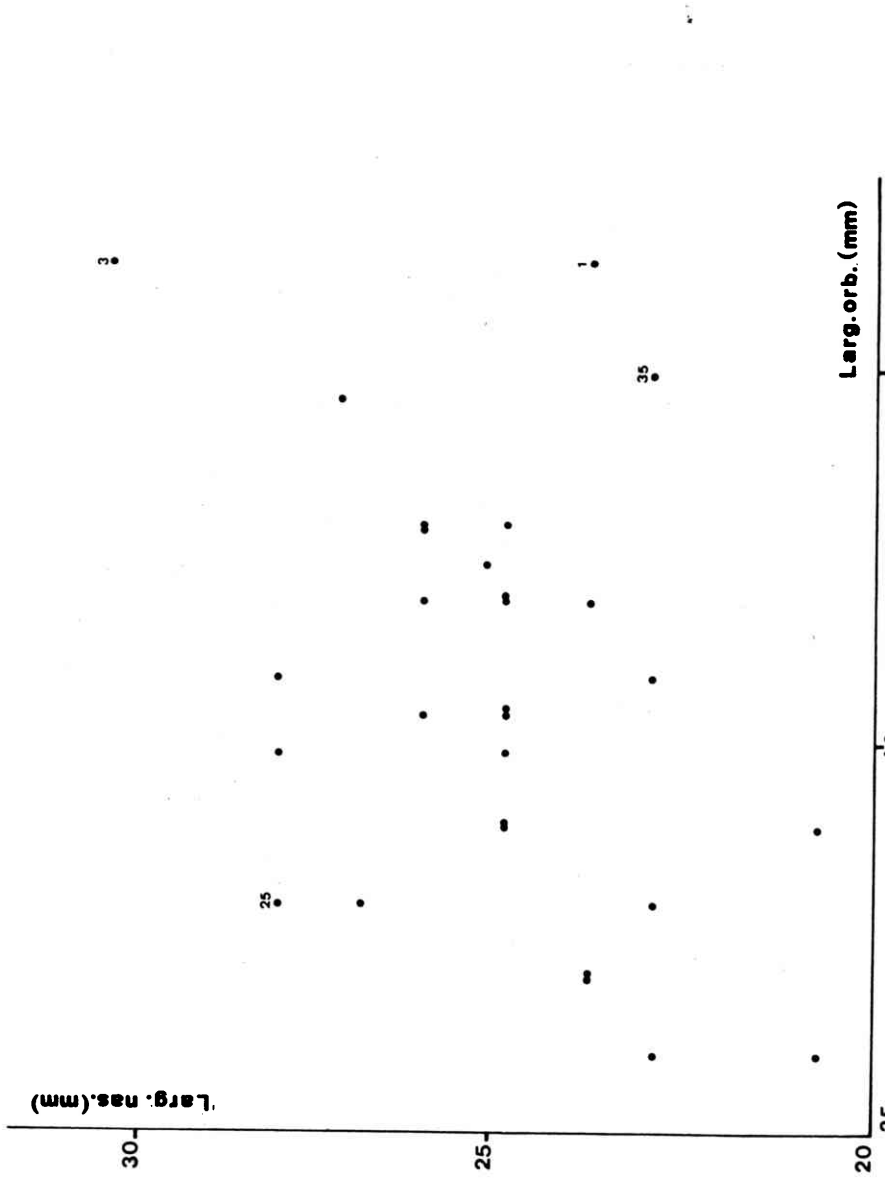

Figure 6. Distribution des crânes du Paléolithique supérieur (N = 29) en fonction de la largeur orbitaire et de la largeur nasale.

ble, d'inclure des objets très atypiques. Il est possible de repérer ces objets au moyen des indices d'atypisme dont le calcul n'est d'ailleurs pas limité au cas de deux variables (Albert et Leguebe, 1989). Il est bien évident que si on ne veut pas réduire de manière importante l'effectif de l'échantillon, il s'imposera de n'éliminer que ceux des objets qui se révèlent être atypiques pour un large ensemble de combinaisons de mensurations.

CONCLUSION

Certains auteurs considèrent que la non-normalité des distributions n'invalide pas le recours à l'analyse multivariée si les résultats ne sont pas prolongés par des épreuves de signification. L'ensemble des observations précédentes montre que les coefficients de corrélation occupent une position clé dans la détermination de la structure multivariée; or, leur valeur dépend dans une large mesure des distributions des variables.

Il n'est pas évident que le calcul des coefficients de corrélation sur le plus grand nombre de paires de valeurs disponibles conduise à des résultats plus fiables. L'impossibilité de prendre certaines mesures du fait que le matériel est fragmentaire, peut affecter de manière plus marquée certaines pièces plutôt que d'autres : une distorsion de la matrice pourrait résulter de la différence de composition des séries de données ayant servi de base au calcul des coefficients de corrélation.

En augmentant le nombre de variables prises en considération, on risque à la fois de réduire le nombre de vecteurs pour lesquels on dispose de toutes les valeurs et d'accroître la quantité d'information qui est redondante, plus rapidement que celle susceptible de nous apporter des précisions complémentaires sur la forme.

Lorsque les matrices de corrélation comportent des valeurs moyennes ou relativement fortes, il paraît donc plus opportun de sélectionner les variables de façon à disposer d'un effectif aussi large que possible de vecteurs complets. Comme nous le montrerons ultérieurement, cette sélection pourrait se faire en tenant compte à la fois des particularités de l'échantillon et des hypothèses que l'on se propose de vérifier.

BIBLIOGRAPHIE

ALBERT, A. et A. LEGUEBE.
1989 Indices d'atypisme en paléontologie humaine. *Z. Morph. Anthrop.* **77** (3) : 273-286.

BILLY, G.
1972 L'évolution humaine au Paléolithique supérieur. *Homo,* **23** (1/2) : 2-12.

BOAS, F.
1899 The cephalic index.
Am. Anthropologist, NS **1** : 448-461.

DEVLIN, S.J., R. GNANADESIKAN et J.R. KETTENRING.
1975 Robust estimation and outlier detection with correlation coefficients.
Biometrika, **62** : 531-545.

FAWCETT, C.D.
1902 A second study of the variation and correlation of the human skull, with special reference to the Naqada crania.
Biometrika, **1** : 408-467.

GALTON, Fr.
1888 Co-relations and their measurement, chiefly from anthropometric data.
Proc. Roy. Soc., London, **45** : 135-145.

GNANADESIKAN, R.
1973 Graphical methods for informal inference in multivariate data analysis.
Bull. Inst. internat. Statist., **45** : 195-206.

GOWER, J.C.
1966 Some distance properties of latent root and vector methods used in multivariate analysis.
Biometrika, **53** : 325-338.

HILLS, M.
1969 On looking at large correlation matrices.
Biometrika, **56** (2) : 249-253.

HOWELLS, W.W.
1973 Cranial variation in man.
Papers Peabody Mus. Archaeol. Ethnol., **67** : 259.

KENDALL, M.G.
1938 A new measure of rank correlation.
Biometrika, **30** : 81-93.
1970 *Rank correlation methods.*
London, Griffin; 4th ed., 202 p.

KOWALSKI, Ch. J.
1972 On the effects of non-normality on the distribution of the sample product-moment correlation coefficient.
Applied Statistics, **21** : 1-12.

LEBART, L., A. MORINEAU et J.-P. FENELON.
1979 *Traitement des données statistiques.*
Paris, Dunod; 511 p.

LEE, A.
1899 A first study of the correlation of the human skull.
Philos. Trans., London, **196A** : 225-264.

LEGUEBE, A.
1989 Analyse biométrique des données en paléontologie humaine.
L'Anthropologie, Paris (sous presse).

LEGUEBE, A. et A. ALBERT.
1989 Test de normalité graphique en paléontologie humaine.
Z. Morph. Anthrop., **77** (3) : 259-271.

PEARSON, E.S.

1938 *Karl Pearson : an appreciation of some aspects of his life and work.* Cambridge Univ. Press, VIII + 170 p.

PEARSON, K.

1895 Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia.
Philos. Trans., London, **187A** : 253-318.

1896 Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurements of organs.
Proc. Roy. Soc., London, **60** : 489-498.

1898 On the probable errors of frequency constants and on the influence of random selection on variation and correlation.
Philos. Trans., London, **191A** : 229-311.

1920 Notes on the history of correlation.
Biometrika, **13** : 25-45.

PEARSON, K et A.G. DAVIN.

1924 On the biometric constants of the human skull.
Biometrika, **16** : 328-363.

RIQUET, R.

1970 La race de Cro-Magnon : abus de langage ou réalité objective?
In G. Camps et G. Olivier : *L'homme de Cro-Magnon. Anthropologie et Archéologie.* Paris, Arts et Métiers graphiques : 37-58.

Adresse de l'auteur : A. LEGUEBE
Institut royal des Sciences naturelles de Belgique
rue Vautier, 29
B 1040 BRUXELLES