

J.-P. BENZÉCRI

Analyse des données biologiques et pathologie clinique

Les cahiers de l'analyse des données, tome 15, n° 3 (1990),
p. 285-304

http://www.numdam.org/item?id=CAD_1990__15_3_285_0

© Les cahiers de l'analyse des données, Dunod, 1990, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

ANALYSE DES DONNÉES BIOLOGIQUES ET PATHOLOGIE CLINIQUE

[ANA. CLIN.]

J.-P. BENZÉCRI

Le mercredi 20 Juin 1990 se tint, dans la Salle Chamberland de l'Institut PASTEUR, une réunion de réflexion consacrée à la Recherche de facteurs pronostiques dans l'évolution du SIDA. On trouve dans le présent cahier un compte rendu de cette réunion pluri-disciplinaire. L'exposé d'analyse des données, qui ouvrit le débat, est ici publié à part, dans l'espoir de servir à des médecins et statisticiens qui participent à d'autres recherches cliniques.

0 Enchaînement des exemples présentés

Par une note intitulée "État du système immunitaire et histoire clinique chez les patients affectés par le virus VIH", nous avons proposé aux médecins et biologistes un programme dont les étapes successives n'ont jamais encore été enchaînées dans une même recherche statistique. Pour justifier ce programme, nous devons prendre argument de plusieurs exemples partiels.

Au §1, on tire, d'une seule analyse multidimensionnelle, une image globale comprenant, avec les sujets sains, un vaste domaine de la pathologie. Au §2, on montre qu'une semblable représentation géométrique peut servir de base à une démonstration. Le §3 concerne l'état du système immunitaire tel qu'il peut être décrit d'après les données recueillies, à l'Hôpital de l'Institut PASTEUR, sur les patients infectés par le Virus de l'Immunodéficience Humaine (VIH).

1 Un panorama de la pathologie clinique

Au prix d'un labeur persévérant dont il faut également louer la largeur de vue et le sens du détail, le Professeur Georges SANDOR, assisté de son épouse, a, dans le Service Chimie Physique des Protéines de ce même Institut où nous sommes réunis aujourd'hui, constitué une base de données concernant l'ensemble de la pathologie. Sans rendre compte de la totalité des travaux statistiques effectués, nous présentons l'analyse d'un tableau, d'après la thèse de L. Benyamina.

Les données de la base proviennent de plusieurs milliers d'examen individuels effectués sur des sujets sains ou malades. Nous retenons de chaque

examen cinq variables qui sont toutes des concentrations sériques, et concernent les fractions protéiques suivantes:

{ Albumine; Orosomucoïde; IgA ; IgG ; IgM }

Ces variables sont de même nature; mais, quant à l'ordre de grandeur, elles ne constituent pas un ensemble homogène: comme ce serait le cas, par exemple, si l'on ventilait par classes d'âge-sexe la population d'un grand nombre d'agglomérations (qui seraient les analogues des échantillons de sérums analysés par G. Sandor). Dans les analyses présentées aux §§2 & 3, l'hétérogénéité est encore plus grande; car les variables diffèrent non seulement quant à l'ordre de grandeur mais quant à la nature même (e.g. on considère simultanément des taux de substances chimiques et des numérations de cellules).

Avec les bilans de population par classes d'âge-sexe, la démarche de la statistique multidimensionnelle n'est qu'une généralisation des calculs classiques de pourcentages. Sont considérées comme identiques deux agglomérations où tous les taux sont les mêmes (e.g. 10% de garçons de 0 à 15 ans, ... , 8% de femmes de plus de 70 ans). Les graphiques plans produits par le calcul étalent les agglomérations et les classes pour rendre compte des gradations de taux (e.g. on aura à droite les variables '% d'adolescents' avec les agglomérations où ces % sont élevés).

En présence de variables hétérogènes, on ne peut procéder immédiatement à de tels calculs: il faut, au préalable, coder les variables de telle façon que les similitudes entre nombres soient conformes aux similitudes entre les objets décrits par ces nombres. On adopte ici le codage logique, ou découpage des variables en classes; une variante plus fine de ce codage sera présentée au §2.

Données humorales Benyamina Sandor

	17	Ab<	Ab≤	Ab≥	Ab>	Or<	Or≤	Or≥	Or>	iG<	iG≈	iG>	iM<	iM≈	iM>	iA<	iA≈	iA>
norm	0	16	123	178	45	258	14	0	240	74	3	278	37	2	271	38	8	
clgv	25	51	36	19	1	17	45	68	33	34	64	69	52	10	37	52	42	

L'intervalle de variation de chacune des mesures considérées est partagé en 3 ou 4 sous-intervalles consécutifs que l'on considère comme des *modalités* de cette mesure. Par exemple, IgG a 3 modalités notées, {iG<, iG≈, iG>}, que l'on peut appeler modalités {faible, moyenne, forte}. Dans le tableau des données codées, la colonne IgG est éclatée en 3 colonnes; la mesure propre à un individu est traduite par une suite de trois nombres qui sont respectivement {1, 0, 0}, {0, 1, 0} et {0, 0, 1} selon que le sujet rentre dans la modalité faible moyenne ou forte de IgG.

Le tableau soumis à l'analyse et dont un extrait est publié ici, est construit en cumulant les données afférentes à 71 classes de patients. La première ligne, 'norm', concerne les sujets normaux. On voit qu'aucun d'eux ne rentre dans la modalité la plus faible de l'albumine (Ab<) et que plus de la moitié (178) sont

Données humorales Benyamina Sandor

trace :	3.852e-1									
rang :	1	2	3	4	5	6	7	8	9	10
lambda :	1501	923	441	284	194	143	103	92	60	45 e-4
taux :	3896	2395	1144	737	505	371	267	238	156	118 e-4
cumul :	3896	6291	7435	8172	8677	9048	9316	9554	9709	9827 e-4

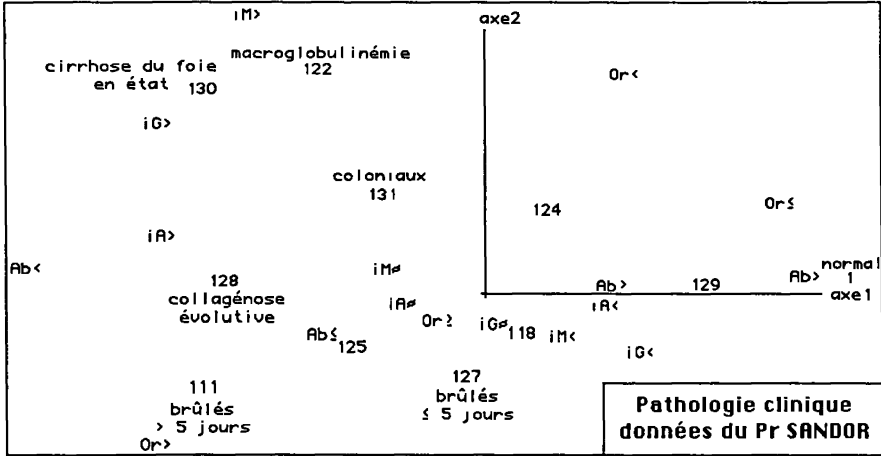
SIGJ	QLT	PDS	INR	F 1	CO2	CTR	F 2	CO2	CTR	F 3	CO2	CTR	F 4	CO2	CTR
Ab<	901	29	100	-885	584	150	106	8	3	-487	177	154	421	133	180
Ab≤	656	56	32	-332	503	41	-160	116	15	89	36	10	-3	0	0
Ab≥	721	68	23	224	383	23	36	10	1	150	170	34	-145	159	50
Ab>	817	48	56	602	800	115	72	11	3	-24	1	1	-45	5	3
Or<	759	16	58	250	44	7	854	510	124	-474	157	80	265	49	39
Or≤	905	72	92	550	614	144	356	258	99	-128	33	27	1	0	0
Or≥	511	55	36	-111	48	4	-104	43	6	227	202	64	-235	218	107
Or>	932	58	125	-645	500	160	-576	399	208	73	6	7	150	27	46
iG=	324	53	19	-6	0	0	-114	97	8	136	138	22	-109	89	22
iG>	840	44	120	-642	395	121	661	418	209	-100	10	10	-134	17	28
iM=	441	48	25	-213	224	14	96	46	5	100	49	11	-157	123	42
iM>	959	16	117	-470	78	23	1088	417	204	1030	373	382	510	91	145
iA=	385	43	23	-182	158	9	-39	7	1	102	50	10	-188	170	54
iA>	787	30	73	-633	431	81	233	58	18	-385	159	102	-359	139	138
iG<	864	103	44	279	473	53	-225	307	56	-27	5	2	114	79	47
iM<	838	136	28	129	211	15	-161	325	38	-155	302	74	-4	0	0
iA<	811	127	30	213	505	38	-43	20	3	57	37	9	150	249	100

dans la modalité supérieure (Ab>); corrélativement, plus des 3/4 rentrent dans la modalité inférieure des IgG (iG<). Le profil est tout autre avec 'clgv', collagénose évolutive: dont près de la moitié des cas sont dans iG>...

L'intérêt de tels cumuls est manifeste: on ne peut s'enquérir de la place de plusieurs milliers de profils individuels; et l'interprétation de la dispersion des résultats repose sur les diagnostics, dont chacun a un profil moyen qui n'est autre que celui de la ligne de cumul. Éventuellement (cf §3), on pourra faire l'analyse du tableau des individus et introduire dans les résultats (graphiques et tableaux de facteurs) les diagnostics (ou autres classes interprétables: pensez à la suite des % calculés pour l'ensemble des communes d'un même type: e.g. rurales de moins de 1000 habitants...) comme des centres de gravité de groupes d'individus.

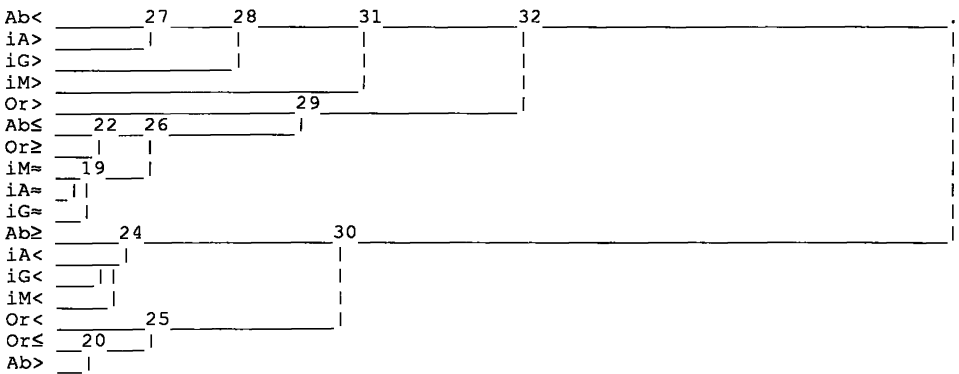
Dans le cas présent, le nombre des diagnostics considérés étant de 71, la lecture du graphique demeure difficile: et c'est pourquoi on a effectué une classification ascendante hiérarchique. On retient 11 classes de diagnostics, agrégés suivant la proximité des profils, et dont les caractères communs sont notés sur l'arbre taxinomique. Sur le graphique plan des axes (1, 2), axes qui rendent compte des 2/3 de la dispersion du nuage (exactement 62,9%), on a porté, outre les sigles de toutes les modalités des variables, les numéros des classes avec quelques diagnostics fréquents caractéristiques.

À l'extrémité positive de l'axe 1, on voit, associée aux sujets normaux, la modalité inférieure de l'albumine, suivie des modalités inférieures des diverses



immunoglobulines. La situation inverse se trouve du côté négatif de l'axe 1. Dans le demi-plan ($F1 < 0$), suivant la direction de l'axe 2, s'opposent $iM>$ et $Or>$, les modalités fortes de l'orosomucoïde et des IgM caractérisant des domaines bien distincts de la pathologie. L'association de la *cirrhose du foie en état* avec un taux élevé d'immunoglobulines est visible; quand aux diverses *macroglobulinémies* (de Waldenstrohm ou africaine), un taux élevé d'IgM rentre dans leur définition même.

Dans la classification, comme sur l'axe 1, une classe de diagnostics correspondant à des profils protéiques peu perturbés (n° 129) s'associe aux normaux (n° 1) pour s'opposer à tout le reste. Au §3, dans l'analyse des examens des sidéens aux divers stades, nous trouverons, de même, les états peu perturbé, non au centre, comme un *juste milieu*, mais à une extrémité du graphique: comme le normal s'oppose au pathologique.



Le présent travail montre qu'on peut intégrer en un profil interprétable des variables hétérogènes quant à l'ordre de grandeur; et obtenir de la pathologie infectieuse ou inflammatoire un tableau cohérent. Ce caractère global nous intéresse en ce qu'il suggère qu'on pourra rassembler les multiples formes d'une affection aussi polymorphe que le SIDA et les mettre en rapport avec la normale. Le petit nombre des variables disponibles, limite toutefois la précision du tableau clinique; de plus, une étude qui ne se fonde sur aucune information diachronique ne peut aboutir à une conclusion thérapeutique. L'étude qui suit, bien que restreinte quant au domaine de la pathologie et au nombre des cas, a le mérite de suivre avec quelque précision l'histoire de ces cas; et, après un raisonnement statistique de portée générale, permet de caractériser l'efficacité relative de plusieurs traitements.

2 Persistance comparée de l'effet de trois antihistaminiques

Il s'agit d'une étude de phase IV conduite par un laboratoire pharmaceutique. Un rapport détaillé en étant publié ailleurs, nous nous bornerons ici à suggérer des conclusions générales d'après un exposé schématique; (où nous raisonnerons, notamment, comme s'il ne manquait aucune donnée aux dossiers; bien que les essais de cette sorte soient communément affectés de telles déficiences).

Les trois produits en présence sont notés {A, T, M}. Dix sujets, atteints de rhinite allergique et initialement sans traitement, sont traités successivement pendant un mois par chacun des trois produits; avec, pour séparer les mois de traitement deux mois de repos (ou 'lavage': *wash out*). L'ordre d'application des traitements (T, A, M; ou: M, A, T, etc...) est choisi aléatoirement pour chaque sujet; et les produits sont distribués en double insu (i.e., sous présentation uniforme, à l'insu du malade et du médecin).

Chaque mois de traitement (par le produit Z) fait l'objet de 4 examens: un examen à vide 'w' au jour zéro; un examen 'z' à la fin et deux examens 'z'' et 'z"', respectivement 5 et 10 jours plus tard alors que le sujet est en repos. Il y a donc (en principe! cf. *supra*) 12 examens par sujet; soit, en tout, 120 examens.

Les examens sont tous d'un même format. On relève d'une part des signes cliniques qui sont {éternuement, rhinorrhée, obstruction nasale, prurit} appréciés à la fois par le patient et le médecin sur une échelle discrète {0, 1, 2, 3}; et, d'autre part, la réponse cutanée à quatre allergènes {histamine, deux allergènes acariens, PAF}; en notant pour chacun la positivité et, s'il y a lieu, le diamètre de la papule et de la zone d'érythème (les diamètres étant notés arbitrairement à 3mm si la réponse est nulle ou minime). En tout, 20 variables par examen.

Plus encore qu'au §1, les variables sont hétérogènes: un codage s'impose donc. Voici comment, pour plus de précision, on a codé, suivant 3 modalités, chacune des variables continues (diamètres des papules ou zones d'érythème). Au vu d'un histogramme, on choisit, pour chaque variable, 3 valeurs qu'on

appelle des pivots: un pivot inférieur, un pivot central et un pivot supérieur. Et l'on procède comme suit. Un individu situé en deçà du pivot inférieure est considéré comme très bas; et on le code (cf. §1): $\{1, 0, 0\}$; de même, au delà du pivot supérieur, on note $\{0, 0, 1\}$. Mais entre le pivot inférieur et le pivot central, la note sera partagée, par un calcul précis, entre les deux modalités correspondantes: au milieu des 2 pivots, on aura $\{1/2, 1/2, 0\}$; à $1/3$ de l'inférieur et $2/3$ du central, ce sera $\{2/3, 1/3, 0\}$. On procède de même entre le pivot central et le supérieur; e.g. $\{0, 3/4, 1/4\}$, si la mesure tombe au quart à partir du pivot central dans l'intervalle séparant celui-ci du pivot supérieur.

Étant ainsi codé, le tableau dont les 120 lignes décrivent les 120 examens peut être soumis à l'analyse des correspondances. L'analyse est dominée par le premier facteur dont l'importance est plus du double de celle du second; et, quant aux modalités, l'interprétation de l'axe 1 est très claire: à droite, du côté ($F1 > 0$), sont toutes les modalités supérieures de la souffrance clinique (rhinorrhée,...) ou de la réponse cutanée à un allergène; les modalités inférieures sont à gauche, du côté ($F1 < 0$). Ainsi, on peut assimiler l'examen à la prise d'une mesure unique, le facteur $F1$, qui exprime l'atteinte allergique; avec $F1$ fortement positif si l'atteinte est grave et $F1$ négatif si le sujet est indemne. On dira, en bref, que $F1$ mesure l'allergie. Reste à utiliser notre instrument de mesure pour chiffrer et comparer l'efficacité des traitements.

Il n'y a que 10 sujets; mais, en tout, 30 mois de traitement; le parcours de chacun étant jalonné par 4 prises de mesure. Regarder dans le détails quelques mesure est instructif; mais ne peut aboutir à une conclusion nette; parce que des accidents divers marquent le parcours de chaque sujet. On calculera donc des moyennes, non sans prendre garde aux fluctuations statistiques qui affectent toute moyenne; et on calculera des différences, avec cette réserve que notre mesure ne peut être sans précaution assimilée à une quantité ou à une distance.

Sur les mesures disponibles, on a pris les moyennes de 15 groupes de 10 nombres définis par des critères réels. Et, comme termes de comparaison, on a calculé les moyennes de 200 groupes fictifs constitués par tirage aléatoire. Voici comment.

Chaque traitement Z étant appliqué 10 fois sous le contrôle de 4 examens, on a 4 moyennes notées, $\{wZ, Z, Z', Z''\}$, calculées respectivement sur les notes à l'entrée dans le traitement, à la fin du mois, 5 et 10 jours après. De plus, on calcule 3 moyennes d'examen à vide, groupés non en fonction du traitement (comme pour les wZ : wA, wT, wM), mais de la chronologie d'ensemble: $\{W_a, W_b, W_c\}$; respectivement au début du premier mois de traitement; à la fin du premier mois de repos (et donc au début du 2-ème mois de traitement); et au début du troisième et dernier mois de traitement que comporte l'essai.

Quelle que soit la dispersion des données individuelles, les moyennes se disposent sur l'axe de façon parfaitement cohérente. Les points à vide $\{wA,$

wT, wM ou $\{W_a, W_b, W_c\}$ sont du côté ($F1 > 0$): réponse allergique élevée; en fin de mois de traitement, on trouve $\{A, T, M\}$ du côté négatif: cette première observation atteste l'efficacité des trois antihistaminiques.

Quant à la persistance de l'effet, on note d'abord que, de W_a à W_c en passant par W_b , la réponse allergique, même prise à vide, recule; quels qu'aient été les deux traitements déjà reçus (traitements attribués au hasard par les organisateurs de l'essai). Mais dès les 10 jours qui suivent l'arrêt d'un traitement, le bénéfice de celui-ci tend à céder; la réponse augmente, on a: $Z < Z' < Z''$.

Comparons les traitements. Les trois points $\{A, T, M\}$, mesures moyennes à la fin d'un mois en charge, diffèrent peu; 'M' étant le moins écarté du côté ($F1 < 0$) on conclurait que la réponse allergique est moins réprimée par ce traitement que par les deux autres; mais cette conclusion ne subsiste guère si l'on tient compte de ce qu'à l'entrée dans M, l'état moyen est plus grave qu'à l'entrée dans A ou dans T: $\{wA, wT\} < wM$.

Reste à comparer les persistances. Nous avons noté que, pour chacun des 3 traitements: $Z < Z' < Z''$; mais l'augmentation ($Z'' - Z$) de la réponse allergique, de la fin du mois de traitement jusqu'au dixième jour de repos, diffère suivant le produit. Pour T et M, on passe à l'origine au 5-ème jour: $T' \approx 0, M' \approx 0$; et au 10-ème jour la réponse est franchement positive. Pour A, la valeur algébrique augmente, mais reste négative; on a:

$$A < A' < A'' < 0; \quad T < 0 \approx T' < T''; \quad M < 0 \approx M' < M'';$$

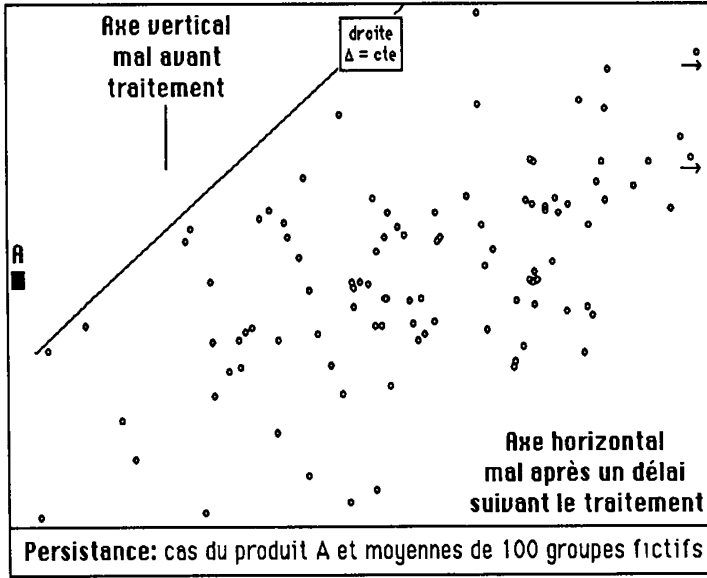
Quant à la persistance, il apparaît donc que A l'emporte sur T et M. Pour plus de précision, on tient compte de l'état des patients à l'entrée dans le mois du traitement considéré en effectuant la différence ($wZ - Z''$); il vient:

$$(wA - A'') = 608 \gg (wT - T'') = 142 \approx (wM - M'') = 133;$$

dix jours après l'arrêt, seul A laisse un bénéfice notable.

Mais les données étant peu nombreuses, et les comparaisons fondées sur des moyennes calculées d'après un instrument de mesure indirect, on s'interroge sur la validité de ce résultat. Aucun modèle probabiliste n'étant ici de mise, car il n'y a point de loi, on procède par simulation.

Du point de vue des comparaisons qu'on vient de rapporter, un traitement est caractérisé d'après les 10 mois au cours desquels il a été appliqué. Afin de prendre pour référence le hasard, on tire donc, des 30 mois (3 trait. \times 10 sujets) pour lesquels on dispose d'observations, 100 groupes aléatoires de 10 mois. Un tel groupe est considéré comme un traitement fictif f pour lequel on calcule un point wf , moyenne à l'entrée, et un point f'' , moyenne 10 jours après l'arrêt.



Sur un diagramme plan, on porte en abscisse le mal (ou mesure de la réponse allergique) 10 jours après l'arrêt; et en ordonnée, le mal à l'entrée dans le traitement. L'amélioration est mesurée par la différence Δ , constante sur toute parallèle à la première bissectrice. En comparant avec le point A, (A'' , wA), le nuage des 100 points f , (f'' , wf), on voit que la différence ($wA - A''$), mesure du bénéfice persistant, n'est égalée par aucune des différences ($wf - f''$) afférentes aux traitements fictifs: CQFD!

J'ai choisi cet exemple, pour montrer qu'en analyse multidimensionnelle, l'enchaînement des constructions de tableaux et des analyses de correspondances ou des classifications, constitue une nouvelle espèce de discours, de logique; avec une suite de questions qui peuvent aboutir à des conclusions; et, éventuellement, requièrent, pour conclure, de nouvelles expériences.

3 État du système immunitaire chez les sujets infectés par le VIH (Virus de l'Immunodéficience Humaine)

Vous savez que le thème de notre réunion a été suggéré par un travail, entrepris par trois des participants: Fr. Tékaïa, Ph. Sansonetti et J.-M. Claverie; travail intitulé: Estimation du stade de l'infection par le VIH chez les sujets séropositifs (cf. [STADES VIH], in *CAD*, Vol XV, n°3). C'est à dessein qu'ayant pour but de montrer le progrès de la pensée au fil des analyses je donne au présent § un nouveau titre. Reprenant une étude dont des résultats sont déjà publiés en détail sous une autre forme, je prends la liberté de proposer des

réflexions générales, toujours présentes à mon travail, mais qui trouvent mal leur place dans un exposé particulier. D'ailleurs, grâce à la diligence de Fr. Tekaiia, nous pouvons considérer ici un plus grand nombre d'observations que dans [STADES VIH]: il y en a 2084.

Devant un tableau de données, le statisticien peut être pris de peur! craindre que rien de lisible ne sorte d'une masse accablante d'informations. Une voie s'offre alors, pour contraindre en quelque sorte les résultats à s'ordonner. Au lieu de procéder de façon purement inductive, sans introduire aucun ordre *a priori*, on cherche à retrouver un tel ordre, un peu comme on calculerait, par régression linéaire, la meilleure approximation d'une grandeur en fonction des données; au lieu de laisser les facteurs se constituer d'eux-mêmes.

C'est ainsi que, dans l'analyse rapportée au §1, au lieu de croiser avec l'ensemble des modalités des variables l'ensemble même des cas individuels, on a cumulé ces cas par diagnostics, ce qui force les modalités à s'ordonner suivant ce qui est pertinent pour la nosographie.

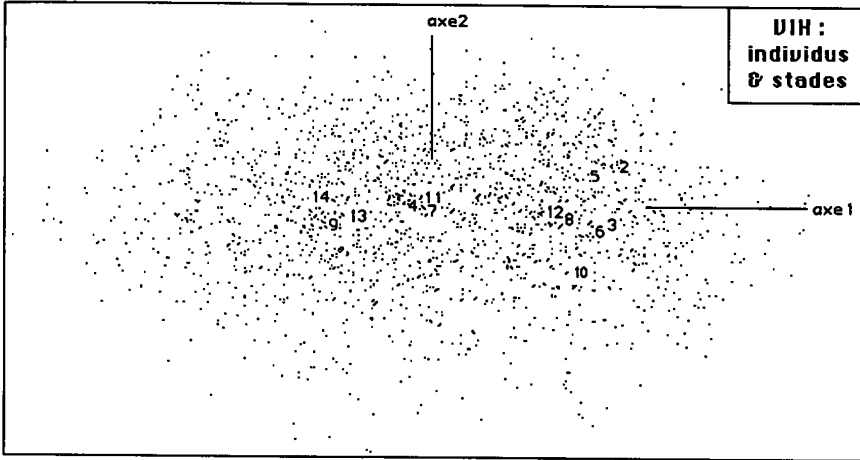
Dans l'analyse des quelque 2000 observations effectuées sur les sujets infectés par le VIH, on a suivi les deux voies: on a croisé avec les modalités des variables, d'une part, l'ensemble des 13 lignes obtenues en cumulant les observations (non par diagnostic, ici, mais par stade); et, d'autre part l'ensemble des observations individuelles. Il semble que la première voie aille directement au but qu'assigne le titre *estimation du stade*; tandis que la deuxième s'avance dans l'inconnu. En fait l'une et l'autre analyse produisent des résultats principaux équivalents, à une différence près que j'expliquerai ensuite. Hormis cette explication, l'exposé sera fondé sur des graphiques issus de l'analyse prenant directement en compte les observations.

3.1 Le premier facteur: pancytopénie et hyper- γ -globulinémie

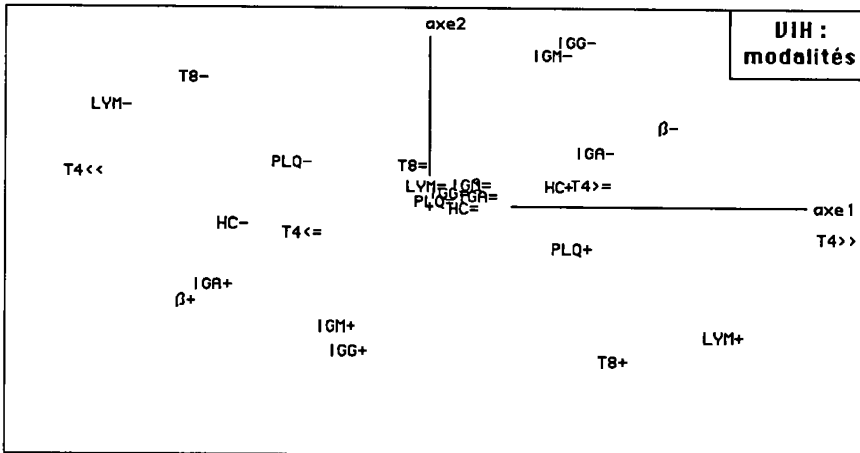
Considérez, dans le plan des axes (1, 2) ou (1, 3) cet essaim de points qui, aux fusions et superpositions près, sont plus de 2000! Seul commentaire: les nombres de 2 à 14, placés aux centres de gravité des sous-ensembles de points afférents aux divers stades... Pour comprendre, il faut considérer, à la même échelle, le nuage des modalités des variables, chacune codée avec des pivots comme on l'a expliqué au §2. Ainsi, on peut, en terme de modalités, faire, en quelque sorte, une symptomatologie de chacune des zones du graphique.

Le numérotage des stades nous a déjà montré, à gauche, du côté ($F1 < 0$), les états les plus graves (ceux des stades 9, 13, 14); et à droite ($F1 > 0$) les états des sujets asymptomatiques (stades 2, 3, 5, 6).

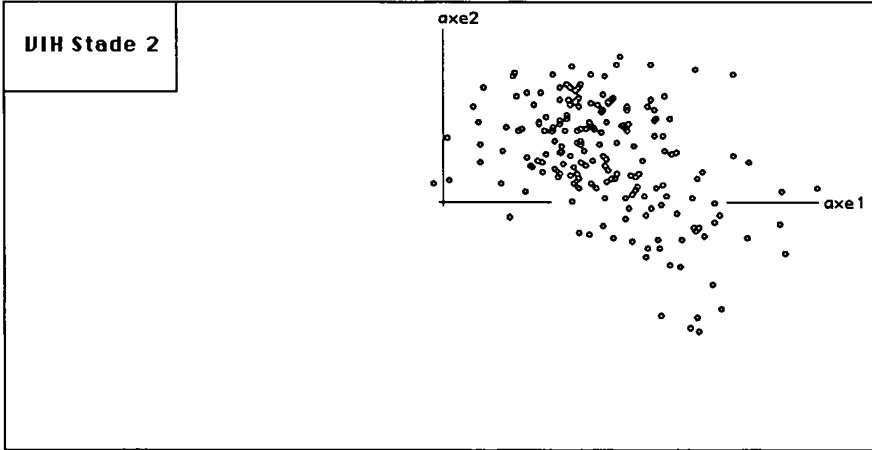
Quant aux 8 variables, elles se distribuent en deux groupes décrivant l'état du système immunitaire au niveau cellulaire et humoral. Sont dénombrés: les lymphocytes (de toute catégorie, LYM), les lymphocytes T4 et T8 (comptés



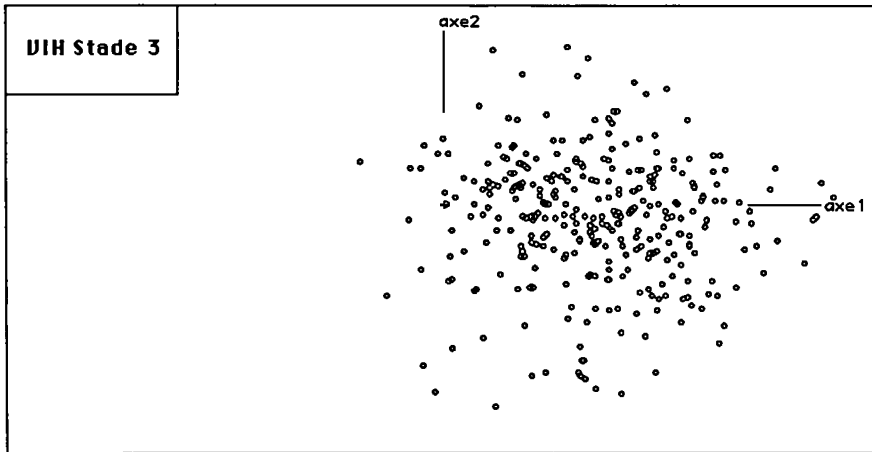
séparément), les plaquettes (PLQ); sont dosées: les IgG, IgM et IgA ainsi que la β 2-microglobuline.



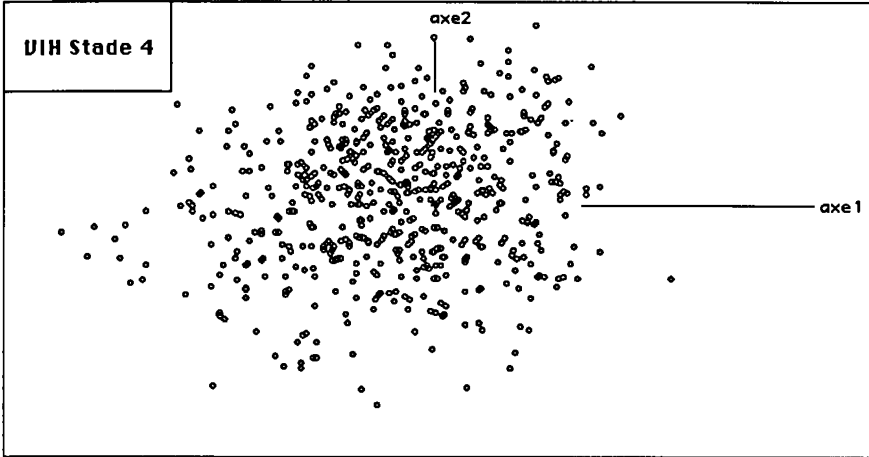
Avec ($F1 > 0$) vont les modalités supérieures de toutes les numérations cellulaires; notamment des T4, mais aussi des autres lymphocytes et des plaquettes. Corrélativement, pour les immunoglobulines IgG, IgM et IgA, ainsi que pour la β 2-microglobuline, on a les modalités inférieures. Ce qui, relativement à notre population, est appelé modalité supérieure de numération et modalité inférieures de taux de globuline, correspond, en fait, à peu près, à la normale, aux sujets sains. À une très importante exception près toutefois: La modalité T8+ des T8 est nettement supérieure à la normale, qui est proche de T8=: nous reviendrons sur ce point au §3.3.



Du côté ($F1 < 0$), avec les modalités opposées, on peut, au contraire, parler sans réserve de pancytopénie (défiance de toutes les classes de cellules dénombrées) et d'hyper- γ -globulinémie: ce syndrome n'est présent que dans les états les plus graves.



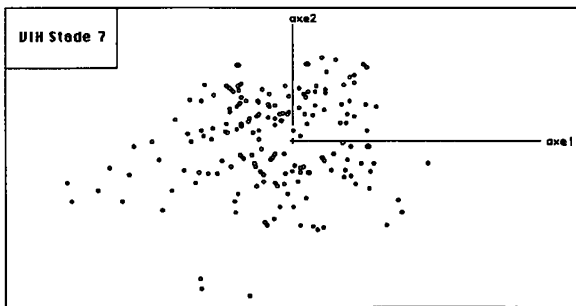
Ne disons rien présentement de l'interprétation des axes 2 et 3, mais considérons, dans le plan (1,2), les sous-nuages de points afférents à divers stades; car plusieurs d'entre eux suggèrent des réflexions critiques. Prenons d'abord le stade 2: il est conséquent qu'un ensemble de sujets infectés mais asymptomatiques ne débord pas du demi-plan droit ($F1 > 0$); plus précisément, la densité est maxima dans le quadrant ($F1 > 0, F2 > 0$). Le stade 3 est similaire: la seule différence, (pour laquelle on proposera une explication au §3.3), étant



que la densité est égale dans les deux quadrants haut et bas, ($F1 > 0, F2 > 0$) et ($F1 > 0, F2 < 0$), du demi-plan droit ($F1 > 0$).

Jusqu'ici, la dispersion observée est acceptable; conforme à ce qu'on attend de sujets dont un numéro de stade ne peut caractériser l'état avec une absolue précision.

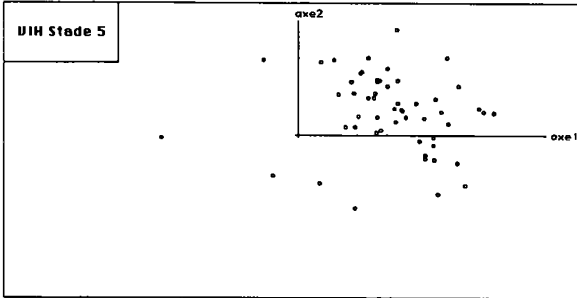
Mais voici le stade 4: celui-ci s'étale sur tout le plan (1, 2): il comprend aussi bien des observations situées à droite, pour lesquelles les variables retenues sont à peu près normales, que des observations où règnent la pancytopénie et l'hyper- γ -globulinémie ($F1 < 0$).



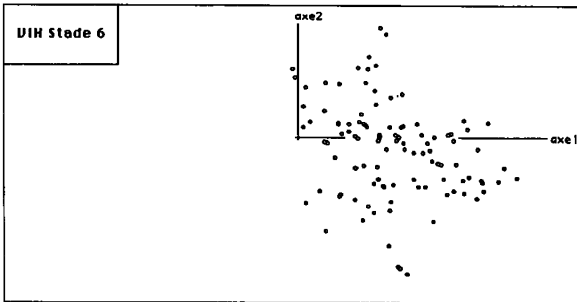
Le stade 7, quoique moins dense, recouvre le stade 4.

Bien que leur rang les place après le stade 4, les stades 5 et 6, étant presque exclusivement distribués dans le demi-plan ($F1 > 0$), comprennent des sujets dont l'état est proche de la normale (du moins au niveau des variables retenues).

Les stades {8, 10, 11, 12} sont très peu représentés.

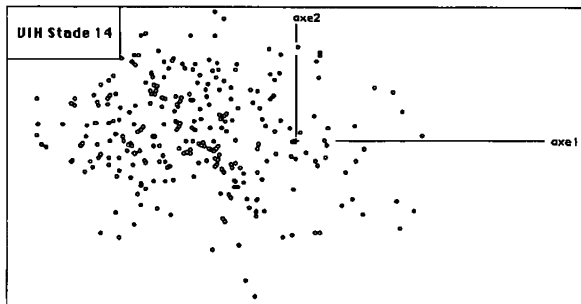
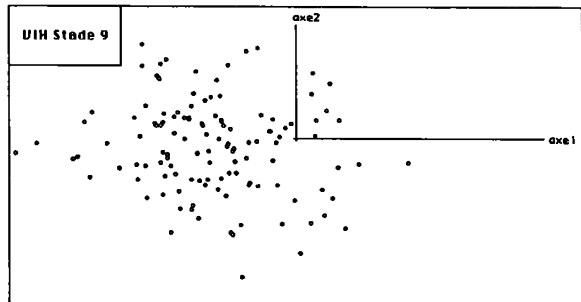


Dans le plan (1,2),
les sous-nuages des
stades 5 et 6
recouvrent,
respectivement ceux
des stades 2 et 3.



Les stades {9, 13, 14},
sont, comme il est
naturel, distribués quasi
exclusivement dans le
demi-plan ($F1 < 0$):
pancytopénie,
hyper- γ -globulinémie.

Mais, même au stade 14
du SIDA déclaré, on a
quelques sujets figurés
par un point dont
l'abscisse $F1$ est
positive. Ces sujets ont
un système immunitaire
relativement normal: le
diagnostic de SIDA a été
porté d'après la clinique,
un sarcome de Kaposi
s'étant déclaré chez eux à
un stade précoce de
l'infection par le VIH.



3.2 Les diverses sources d'information

Nous venons de recourir à l'information clinique, tandis que l'analyse présentée porte sur des données biologiques non spécifiques du SIDA, à l'exclusion des données spécifiques telles que: antigénémie, et séquelles immunitaires d'infections opportunistes (hypersensibilité à la candidine et à la tuberculine).

La classification du CDC d'Atlanta, principalement fondée sur la clinique, distingue, on le sait, quatre étapes successives de la maladie: {Porteur Asymptomatique, Lymphadénopathie Chronique, États apparentés au SIDA, SIDA *stricto sensu*}. Les stades de la nomenclature de l'Hôpital de l'Institut Pasteur subdivisent ceux du CDC en se référant, notamment, à des données biologiques précises, spécifiques ou non du SIDA.

La valeur clinique d'un tel système intégré d'informations est incontestable; et il nous a fourni la clef de l'interprétation de l'analyse factorielle. Mais, pour nous, l'intérêt des données non spécifiques vient de ce qu'on peut utilement les relever, en dehors du domaine du SIDA, sur un champ très étendu, chez des sujets sains ou malades; comme l'a fait le Pr SANDOR dans l'étude rapportée au §1.

Le raisonnement statistique a besoin de références extérieures au domaine particulier sur lequel on s'interroge; or ces références sont rarement fournies par le spécialiste. Qu'il soit médecin ou linguiste, le spécialiste consacre tous ses efforts à observer l'objet même de son étude; il n'a cure de rassembler des données ne concernant directement que des problèmes considérés par lui comme résolus.

Ainsi, pour décider de l'attribution à un auteur unique de la totalité des chapitres d'une œuvre remontant à l'antiquité grecque ou latine, le philologue est prêt à dénombrer dans ces chapitres les mots ou les tournures de phrase. Mais il ne songe pas que l'étalon de la variabilité de style compatible avec l'attribution à un auteur unique doit être cherché par le statisticien dans le dépouillement d'œuvres étendues dont l'auteur est connu et reconnu de tous (même si ces œuvres sont écrites dans une autre langue que celle du texte auquel on s'intéresse principalement). Car, s'il est vrai qu'il n'y a de science que du général, c'est dans la connaissance générale qu'on doit chercher la réponse scientifique à toute question particulière.

D'ailleurs, l'analyse des informations du bloc immunitaire générique nous paraît propre à servir de fond à la recherche, si importante aujourd'hui, de critères mesurant rapidement l'efficacité des traitements nouveaux essayés contre l'infection par le VIH. Fondamentalement la question est posée des rapports entre état immunitaire en général et maladie en particulier: nous estimons que des variables biologiques continues peuvent permettre d'apprécier l'efficacité d'un traitement à visée immuno-stimulante; sans être paralysé par les longs délais

requis pour une évaluation fondée sur les statistiques de survie ou même de temps de passage au SIDA *stricto sensu*.

De ce point de vue, les facteurs issus de l'analyse des observations pourraient offrir de plus riches perspectives que la notion même de stade à laquelle on les a confrontés par les calculs de régression et l'analyse discriminante barycentrique. Tout mouvement du sujet dans la direction positive de l'axe 1 apparaîtrait comme la marque d'une amélioration de l'état du système immunitaire à porter au crédit du traitement.

Mais, ici encore, se pose un problème d'étalonnage: car seul est significatif un mouvement dont l'amplitude est supérieure à celle des fluctuations usuelles de l'état du système immunitaire; fluctuations que l'on doit donc observer, chez le sujet séropositif ou non, sous diverses conditions; notamment lors d'incidents pathologiques bénins.

Outre l'état instantané du système immunitaire, il convient de considérer comme un individu statistique l'histoire clinique d'un cas; et de faire, de telles histoires, une typologie. Des malades rangés sous un même titre, parce que l'agent causal est reconnu pour être le même, peuvent suivre des parcours bien différents.

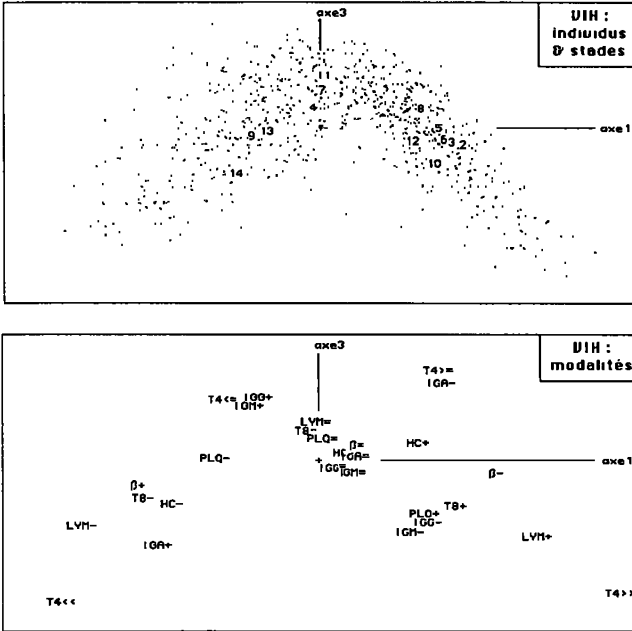
Ainsi, analysant 102 cas de Leucémie Lymphoïde Chronique suivis à l'Hôpital Saint-Louis dans le Service du Pr. Jean BERNARD, Ch. Bastin, L. Degos et coll., ont d'abord reconnu, dans le plan (1,2), quatre quadrants {A, B, C, D} qu'ils caractérisent ainsi, du point de vue de l'évolution clinique:

Dans la région A, se situent essentiellement les malades au moment du bilan initial. Les modalités de cette région correspondent à des examens anormaux mais non aux symptômes les plus alarmants. La région B rassemble les modalités normales des variables. On y trouve peu de décès: c'est une zone de rémission. Dans la région D, se concentrent la plupart des décès; associés à des symptômes très inquiétants: grosse rate, cytopénie, test de Coombs positif. Enfin la région C est une zone de transition; s'y projettent beaucoup d'individus qui sont encore vivants lors de leur dernier examen en date.

Les auteurs constatent ensuite que les malades décédés en moins de quatre ans, sont généralement passés directement de A en D. Puis, considérant l'évolution complexe des malades ayant vécu plus de quatre ans ils notent trois types d'évolution:

Des patients, à partir de A, évoluent directement vers C. D'autres passent par B avant d'aller en C. Mais certains, situés également en A lors de leur bilan initial, passent d'abord par C avant de se diriger vers B. Ces patients sont alors en rémission.

L'intérêt de semblables conclusions (dont on souhaiterait obtenir l'analogie ici) ne sera contesté par personne.



Dans le plan (1,3), nuage des individus et nuage des variables affectent la forme d'un *croissant parabolique*. C'est le classique *effet Guttman*. Le facteur 3 est, approximativement une fonction du second degré du facteur 1. Le seul intérêt éventuel du facteur 3 est qu'il signale, en les plaçant nettement à l'intérieur du croissant, les individus qui présentent une *conjonction des extrêmes*, i.e. associent certaines modalités qui appartiennent aux deux pointes du croissant.

3.3 Au-delà du premier facteur

Je rappelle qu'on a effectué deux analyses: on a croisé avec les modalités des variables, d'une part, l'ensemble des observations individuelles; et, d'autre part, l'ensemble des 13 lignes obtenues en cumulant les observations par stade. Le plan (1,2) issu de la deuxième analyse coïncide presque exactement avec le plan (1,3) issu de la première: nous ne parlerons donc que de celle-ci.

Il importe de préciser que cette coïncidence, apparente sur les graphiques, est confirmée par le calcul: entre les facteurs F1 issus des deux analyses, considérés comme fonctions sur l'ensemble des 1937 examens, la corrélation est de .992; entre la facteur F3, issu de l'analyse où les individus sont en principal, et le facteur F2, issu de celle où sont en principal les cumuls par stade, la corrélation est .92.

Avant d'examiner les facteurs 1 et 2, nous rappelons qu'en analyse des correspondances, il est commun que modifier le tableau des données, (soit en éliminant une variable ou un groupe de sujets; soit en supprimant par des cumuls une distinction), fasse disparaître un facteur, les autres restant inchangés. Nous ne citerons qu'un exemple récemment publié: J.-F. Mazet, cherchant des *critères pour l'aspect des placages de bois de chêne*, considère les opinions recueillies au cours d'une enquête. Il enlève d'une première analyse les modalités de réponse "non-choix" et les sujets "non-professionnels du bois":

l'axe 1 issu de la deuxième analyse est identique à l'axe 2 issu de la première analyse (cf. *CAD*, Vol XIV, n°3, 1989).

Dans le plan (1,3), nuage des individus et nuage des variables affectent la forme d'un *croissant parabolique*. Le facteur 3 est, approximativement une fonction du second degré du facteur 1. C'est le classique *effet Guttman*, commenté sur le graphique.

L'axe 2 mérite plus d'attention: s'opposent, sur l'axe 2, d'une part, toutes les modalités fortes et, de l'autre, toutes les modalités faibles. Puisque, à la différence de ce qui est sur l'axe 1, dénombrements cellulaires et concentrations en globulines sont solidaires, nous avons d'abord suggéré que le facteur 2 était créé par l'hémodilution, phénomène commun lié aux variations normales de la volémie au cours du nyctémère. Mais cette interprétation n'est pas confirmée par une analyse complémentaire prenant en compte l'hématocrite en plus des dénombrements et concentrations. Sur l'axe 2, les modalités {HC-, HC≈, HC+} de la variable "hématocrite" sont proches de l'origine; mais HC- et HC+ vont avec les modalités de signe contraire des autres variables.

Puisqu'il disparaît si l'on cumule les individus par stades, le facteur 2 ne semble pas d'une signification clinique majeure. On a toutefois vu que, relativement aux stades 2 et 5, les stades 3 et 6, définis par la présence de signes biologiques divers, se distinguent aussi par une extension de la distribution des cas vers le quadrant ($F1 > 0$, $F2 < 0$). Dans ce quadrant, on note des taux élevés d'IgG et d'IgM: ce qui peut concorder avec la présence de "signes biologiques".

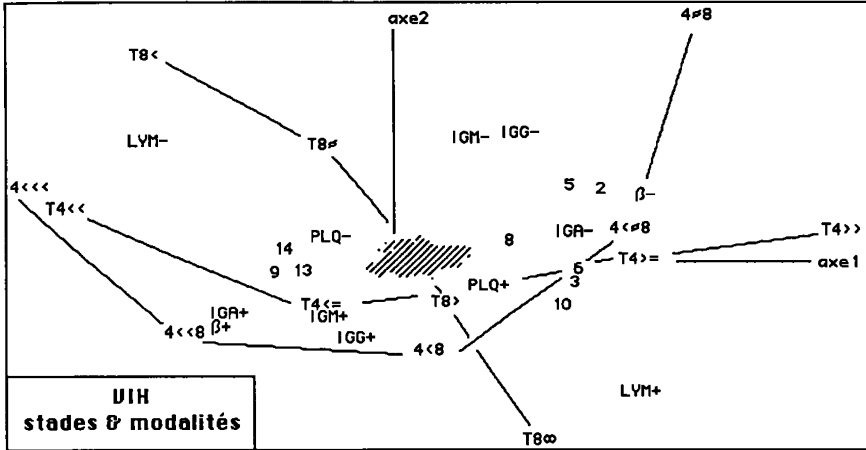
[Mais ce que nous croyons être la clef de l'interprétation du facteur 2 ne nous est apparu qu'après l'exposé dont on donne ici la rédaction différée. Nous croyons utile de proposer cette interprétation.]

C'est Fr. VUILLIER (ingénieur dans l'unité de l'Institut Pasteur que dirige le Pr. G. DIGHIÉRO), qui par sa note "Études des sous-populations lymphocitaires" (in *ADIP*, n°31, Juin 1990), a appelé notre attention sur le fait que, chez les sujets infectés par le VIH mais peu perturbés, le taux des lymphocytes T8 est très anormalement élevé; le taux des T4 ne descendant au-dessous de la normale que chez les patients dont la cytopénie revêt la plus extrême gravité. Fr. VUILLIER signale de plus que:

"le déséquilibre quantitatif des sous-populations T4 et T8 est tel que le rapport T4/T8 s'inverse et peut devenir très faible au stade ultime de la maladie; alors que ce rapport est normalement supérieur à 1."

On a donc repris les analyses en découpant la variable T8 suivant 4 modalités nouvelles, notées:

$$\{T8<, T8\approx, T8>, T8\infty\},$$



les valeurs pivot afférentes à $T8>$ et $T8\infty$ dépassant la normale. De plus, inspiré par le travail cité, on a introduit une variable quotient $T4/T8$, découpée suivant quatre modalités, notées:

$$\{4<<<, 4<<8, 4<8, 4<=8, 4=8\};$$

l'examen des histogrammes ayant montré qu'il est exceptionnel, dans nos examens, que se rencontre (conformément à la normale) une valeur de $T4/T8$ supérieure à 1.

L'analyse montre que le quadrant ($F1>0$; $F2>0$) correspond seul aux états proches de la normale, avec des valeurs peu perturbées du rapport $T4/T8$. En descendant dans la direction de l'axe 2 ($F2<0$) on se dirige vers la modalité maxima des $T8$: $T8\infty$. Il faut donc comprendre que, chez la plupart des patients, se produit une prolifération massive des $T8$ avant même que la séropositivité ne soit décelée. C'est cette phase initiale que révèle l'analyse factorielle en créant l'axe 2. Bien que la ligne des centres de stades s'écarte peu de l'axe 1, et que, dans l'analyse après cumul par stades, le facteur 2 soit perdu, le décalage entre 2 et 3 (ou entre 5 et 6) vers ($F2<0$) apparaît ainsi interprétable.

Quant au rapport $T4/T8$, il s'élève tout au long de l'observation des patients. Selon Fr. VUILLIER (loc. cit.), l'effet noté serait encore plus net si l'on distinguait la sous-population, en augmentation particulièrement forte, des $T8$ cytotoxiques.

Ceci permet d'énoncer une hypothèse, que nous suggère une note de M. ZOUALI, publiée dans le même cahier de l'ADIP. Selon cet auteur (qui cite J.-Ph. CORRE), dès le début de l'infection par le VIH, par un phénomène de double réplique, des anticorps dirigés contre ceux sécrétés contre le VIH, provoqueraient, par leur pouvoir antigénique propre, la création d'anticorps

ayant pour cibles les récepteurs CD4, c'est-à-dire, en définitive, les lymphocytes T4. Avant même que la séropositivité ne soit décelée, se serait ainsi développée une sorte de maladie auto-immune, faisant des T4 la cible d'une classe proliférante de T8 cytotoxiques. Les fluctuations mêmes, observées sur la cytopénie dont sont frappés les patients, peuvent correspondre à une amélioration momentanée du taux des T4; mais l'analyse des données atteste que le quotient (T4/T8) reste bas; et donc que le processus de destruction des T4 n'est pas enrayé. Une publication récente (cf. MIEDEMA et coll.) attire d'ailleurs l'attention sur l'importance de la phase initiale de l'infection par le VIH.

Resterait à examiner une hypothèse connexe, formulée dans la note [IMM. CLIN. VIH] distribuée aux participants de notre réunion. Puisque les lymphocytes T4 sont en faible nombre, mais que seul un pourcentage modéré d'entre eux est infecté par le VIH, il ne s'agirait pas tant de destruction des T4 que d'inhibition, de répression de leur production. Le VIH simulerait un message de saturation en T4, alors que ceux-ci sont peu nombreux; ou, avec un effet équivalent, intercepterait le message naturel stimulant la production des T4. Comme il s'agit de l'équilibre d'un système, la connaissance du comportement homéostatique normal de celui-ci pourrait révéler le processus pathogène.

Mais trop ignorant pour défendre cette hypothèse, nous laissons Baudelaire proposer pour nous, en un quatrain de ses *Correspondances*, cette image du réseau immunitaire:

*Comme de longs échos qui de loin se confondent
Dans une ténébreuse et profonde unité,
Vaste comme la nuit et comme la clarté,
Les parfums, les couleurs et les sons se répondent.*

Références bibliographiques

- Ch. Bastin, L. Degos, N. Feingold, J.-D. Rain et coll.: Les leucémies lymphoïdes chroniques: la diversité des cas et leur évolution; in *CAD*, Vol I, n°4, pp. 419-440; 1976.
- L. Benyamina: *Codage pondéré en analyse des correspondances et méthodes d'interprétation de classifications en pathologie clinique*; Thèse, Université Pierre et Marie Curie, Paris; 1982.
- J.-P. Benzécri, E. Chwetzoff, G. D. Maïti: Persistance comparée de l'effet de trois antihistaminiques chez 10 sujets souffrant de rhinite allergique; in *CAD*, Vol XV; 1990.
- J.-F. Mazet: Recherche de critères pour l'aspect (dessin et couleur) des placages en bois de chêne; in *CAD*, Vol XIV, n°3, pp. 365-376; 1989.
- Fr. Miedema, M. Tersmette, R.A.W. van Lier: AIDS pathogenesis: a dynamic interaction between HIV and the immune system; in *Immunology Today*, Vol XI, n°8; 1990.
- Fr. Vuillier: Étude des sous-populations lymphocytaires: un intérêt pour la clinique, in *ADIP*, Lettre de l'Association pour l'Institut Pasteur, n° 31, pp. 106-110; Juin 1990.
- M. Zouali: Diversité et modulation idiotypique de la réponse immunitaire contre le VIH, in *ADIP*, Lettre de l'Association pour l'Institut Pasteur, n° 31, pp. 68-72; Juin 1990.