LES CAHIERS DE L'ANALYSE DES DONNÉES

F. TEKAÏA
PH. SANSONETTI
J.-M. CLAVERIE

Estimation du stade de l'infection par le VIH chez les sujets séro-positifs

Les cahiers de l'analyse des données, tome 15, n° 3 (1990), p. 261-278

<http://www.numdam.org/item?id=CAD_1990__15_3_261_0>

Les Cahiers de l'Analyse des Données
Vol XV 1990 - n°3 - pp. 261 278.

ESTIMATION DU STADE DE L'INFECTION PAR LE VIH CHEZ LES SUJETS SÉRO-POSITIFS

[STADES VIH]

*F. TEKAÏA**
*Ph. SANSONETTI***
*J.-M. CLAVERIE**

1 Introduction: le problème, les données, les analyses

1.1 Typologie biologique et stades de l'infection

Les facteurs pronostiques de l'évolution de l'infection par le Virus de l'Immunodéficience Humaine (VIH) sont encore mal connus; on ne dispose pas d'une formule permettant de calculer, en fonction de données cliniques et biologiques, un indicateur de gravité, dont la régression éventuelle permettrait d'affirmer l'efficacité d'un traitement.

À partir des données recueillies au service des consultations de l'Hôpital de l'Institut Pasteur nous souhaitons:

1) établir une typologie des états des sujets en fonction des paramètres cliniques biologiques sensibles à la gravité de l'infection,

2) estimer le stade d'infection du malade; ou, mieux encore, concevoir un nouvel indicateur précis de son état.

1.2 Les données de l'Hôpital de l'Institut Pasteur

La Base de Données Relationnelle "CONSULT" regroupe les données relatives aux sujets infectés par le Virus de l'Immunodéficience Humaine (VIH), suivis par le service des consultations à l'Hôpital de l'Institut Pasteur. Cette base

(*) Institut Pasteur, Unité d'Informatique Scientifique, 25 Rue du Dr. Roux, 75724, Paris Cedex 15.

(**) Institut Pasteur, Hôpital service des consultations, 28 Rue du Dr. Roux, 75724, Paris Cedex 15.

est constituée essentiellement des données cliniques et immunologiques recueillies à chaque consultation (cf. 1), chaque sujet étant examiné, en moyenne deux fois par an.

Cette base de données évolutives a été conçue, d'une part, pour suivre l'état de chaque malade; et, d'autre part, pour rechercher des indices biologiques ayant valeur pronostique quant à l'évolution de l'infection, exprimée, présentement, selon une classification par stade tenant compte des paramètres biologiques et cliniques.

Dans une précédente étude (cf. 2) nous avons analysé les données recueillies pendant la première année. La présente étude est fondée sur 1990 observations relatives à 749 patients infectés par le VIH. Chaque observation, correspondant à une consultation d'un malade, est définie par 14 paramètres biologiques et par 3 paramètres cliniques.

Les paramètres biologiques sont: le nombre total de lymphocytes (LYMP), le nombre de cellules T4 (T4), le nombre de cellules T8 (T8), l'hypersensibilité cutanée à la candidine (IDRC), l'hypersensibilité cutanée à la tuberculine (IDRT), la présence de colonies de *Candida albicans* (MYCO), la présence ou l'absence d'antigénémie-p25 (AG), le taux de β2-microglobuline (ß), l'antigène HBs (AGH), l'anticytomégalovirus (AC), les taux d'immunoglobuline IgG (IGG), IgM (IGM) et IgA (IGA) et le nombre de plaquettes (PLQ).

Les paramètres cliniques sont: les 2 types de classifications de l'infection selon les critères utilisés à l'Hôpital de l'Institut Pasteur (HIP) ou au CDC (Center for Diseases Control) d'Atlanta; et l'apparition de signes fonctionnels, neurologiques ou d'infections caractérisées ne répondant pas aux critères du SIDA (Syndrome d'Immuno-Déficience Acquise). Les critères de classification selon HIP ou CDC sont indiqués dans un tableau.

La classification du CDC comporte 3 classes {II, III, IV}. Selon la classification du HIP le stade de l'évolution de l'infection est codé de 2 à 14, allant du Porteur Asymptomatique jusqu'au SIDA; les classes du CDC étant subdivisées respectivement, selon HIP, en {2, 3, 4}, {5, 6, 7} et {8,...,14}.

1.3 Enchaînement des analyses

Afin de répondre aux objectifs de l'étude, indiqués au §1.1, nous avons utilisé trois méthodes: l'Analyse Factorielle des Correspondances (AFC), la Régression Linéaire Multiple (RLM) et l'Analyse Discriminante Barycentrique.

Une première étude (§2) prend en compte la totalité des 1990 observations, même si elles comportent des données manquantes. Les 13 variables biologiques sont découpées en classes (cf. §2.1) et codées en (0,1) avec la variable SYMP (symptomes), en prenant soin de prévoir, pour chaque variable une modalité 'manque'; après élimination de ces modalités, on a un tableau

Classification CDC	Classe	Classification Pasteur HIP	Stade
Primo-infection	I		1
Porteur Asymptomatique	II	-sans signe biologique	2
		-avec signe biologique	
		. T4 ≥ 500/µl	3
		. T4 < 500/µl	4
Lymphadénopathie Chronique	III	-sans signe biologique	5
		-avec signe biologique	
		. T4 ≥ 500/µl	6
		. T4 < 500/µl	7
États Apparentés au SIDA (ARC)	IV	-avec ou sans lymphadénopathie chronique, marqué par des signes fonctionnels	
		. T4 ≥ 500/µl	8
		. T4 < 500/µl	9
		-avec ou sans lymphadénopathie marqué par des signes neurologiques	
		. T4 ≥ 500/µl	10
		. T4 < 500/µl	11
		-avec ou sans lymphadénopathie chronique marqué par des infections caractérisées ne répondant pas aux critères d'inclusion dans le SIDA	
		. T4 ≥ 500/µl	12
		. T4 < 500/µl	13
SIDA (*stricto sensu*)			14

Tableau de définition des stades de l'infection par le virus VIH

(1990 × 44), en (0,1), mais non sous forme disjonctive complète, du fait des données manquantes.

L'analyse de ce tableau, auquel on a adjoint en lignes supplémentaires les centres de gravité des classes d'examen par stade, montre clairement la pertinence des variables retenues. Le but étant de mettre en rapport les données biologiques avec le stade, on a également procédé au cumul des lignes (observations) suivant le stade HIP; d'où un tableau (13 × 44) qui est analysé

comme tableau principal et auquel les observations sont adjointes en élements supplémetaires. Nous avons tenté, avec la RLM, de quantifier directement le stade considéré comme une fonction des paramètres biologiques et cliniques, ou des facteurs obtenus par l'AFC.

Une deuxième étude (§3) est fondée sur un tableau (867 × 10) sans données manquantes. Le nombre des observations reste élevé, mais on a dû écarter 5 variables d'un réel intérêt: {IDRC, IDRT, MYCO, AGHBs, ACMV}. Le tableau des données retenues est transformé en utilisant le codage linéaire par morceau (dit encore "codage flou", cf. §3) afin de préserver au mieux l'information; d'où un tableau à 32 colonnes.

Ce tableau n'est pas analysé tel quel: on s'est borné à analyser le tableau des cumuls (cf. *supra*). En considérant, dans le plan (1,2), les sous-nuages d'observations par classe, on constate, à la fois, la pertinence des informations retenues et leur relative imprécision. L'analyse discriminante barycentrique (qui rattache chaque observation individuelle au centre ou 'stade' dont elle est le plus proche) permet de faire un bilan de ces constatations.

2 Analyses fondées sur la totalité des observations

2.1 Codage disjonctif complet: bornes, classes et effectifs

Afin d'homogénéiser le tableau initial (1990 × 17) précédemment présenté, on a utilisé un codage disjonctif complet. Il s'agit pour chaque variable de fixer des classes de variation (ou modalités) et de remplacer la valeur initiale de l'observation pour cette variable par une suite de 0 et de 1; les valeurs 1 et 0 indiquant, respectivement, l'appartenance et la non-appartenance de la valeur initiale à une classe. Pour les variables continues, on a choisi des bornes de façon à obtenir des classes d'effectifs équivalents.

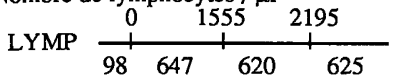
Pour toutes les variables la borne 0 définit la modalité des valeurs manquantes, modalité qui sera éliminée du tableau à analyser.

Ce codage founit un tableau disjonctif complet; après élimination des modalités correspondant aux valeurs manquantes, il reste un tableau (1990×44), (qui n'est plus sous forme disjonctive complète). Ce tableau a été soumis à l'Analyse des Correspondances.

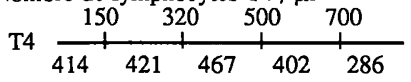
2.2 Analyse des Correspondances

Le tableau des valeurs propres et taux d'inertie afférents aux 7 premiers axes signale l'importance du 1-er facteur, nettement séparé des suivants; c'est pourquoi les résultats de l'analyse des correspondances seront simplement présentés dans le plan (1,2).

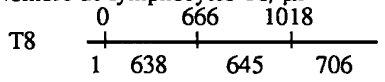
Nombre de lymphocytes / µl



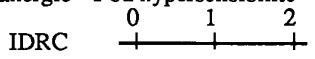
Nombre de lymphocytes T4 / µl



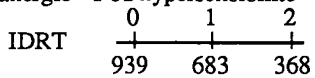
Nombre de lymphocytes T8/ µl



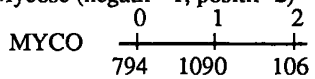
Réaction cutanée à la Candidine
(anergie =1 ou hypersensibilité =2)



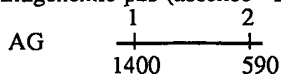
Réaction cutanée à la Tuberculine
(anergie =1 ou hypersensibilité =2)



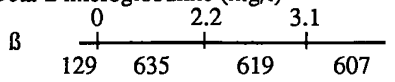
Mycose (négatif =1, positif=2)



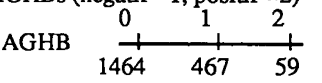
Antigénémie-p25 (absence =1, présence =2)



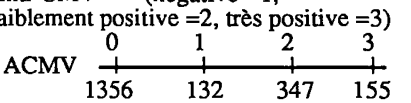
Béta-2 microglobuline (mg/l)



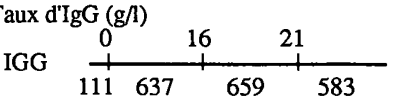
AGHBs (négatif =1, positif =2)



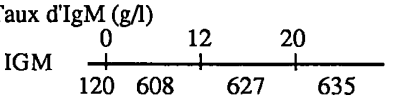
Anti-CMV (négative =1,
faiblement positive =2, très positive =3)



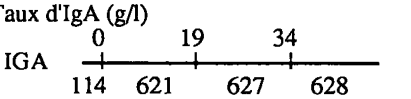
Taux d'IgG (g/l)



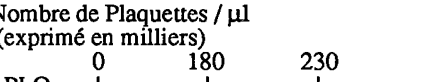
Taux d'IgM (g/l)



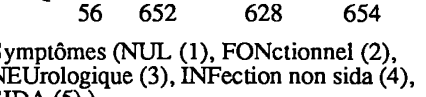
Taux d'IgA (g/l)



Nombre de Plaquettes / µl
(exprimé en milliers)



Symptômes (NUL (1), FONctionnel (2),
NEUrologique (3), INFection non sida (4),
SIDA (5))

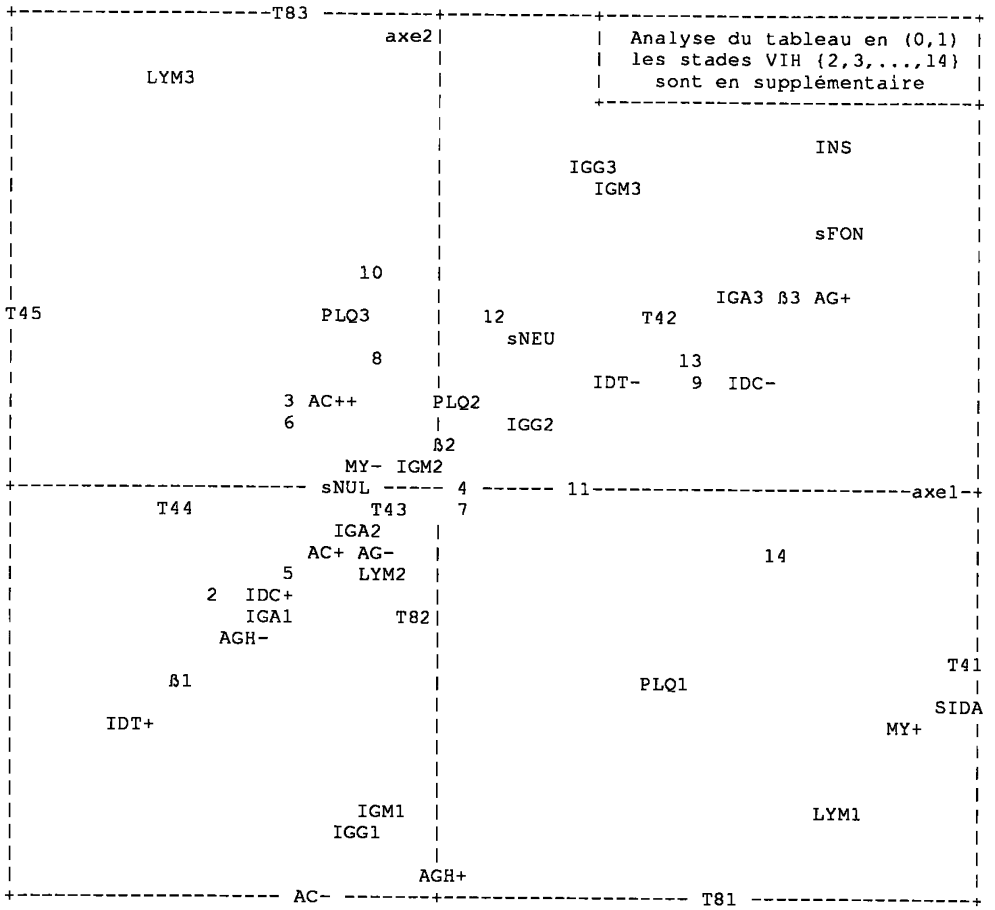


ci-dessus:***Tableau des modalités des variables retenues***

ci-dessous: *Tableau des valeurs propres et taux d'inertie*

Rang	1	2	3	4	5	6	7
Lambda	.3012	.1472	.1323	.111	.1	.1	.09
Taux (%)	11.2	5.5	4.95	4.18	3.8	3.7	3.6
Cumul (%)	11.2	16.77	21.72	25.9	29.74	33.51	37.1

L'examen conjoint, sur le plan (1, 2) des modalités des variables et des stades des observations montre une succession, selon le premier axe, allant du Porteur Asymptomatique jusqu'au SIDA. On constate que selon cet axe les variables se divisent en deux groupes selon l'ordonnance de leurs modalités.



Pour un premier groupe de variables, les modalités fortes sont associées aux manifestations bénignes ou inexistantes (Porteur Asymptomatique) et les modalités faibles au SIDA; ces variables sont:

nombre total de lymphocytes; nombres de cellules T4 et T8; nombre de Plaquettes: il y a donc Pancytopénie dans le SIDA. Il faut toutefois **prendre garde** au fait que, la modalité forte de T8 étant nettement supérieure à la normale, les sujets asymptomatiques, mais déjà infectés, n'ont généralement des numérations quasi normales que pour les cellules *autres que* les T8.

Pour un second groupe de variables biologiques, au contraire, les faibles modalités sont associées aux manifestations bénignes ou inexistantes (Porteurs Asymptomatiques) et les modalités fortes au SIDA; ces variables sont:

taux d'immunoglobulines IgG, IgA, IgM; taux de β2-microglobuline; présence d'antigénémie et de *Candida albicans*. À propos des 4 premières variables, on parlera d'hyper-γ-globulinémie dans le SIDA.

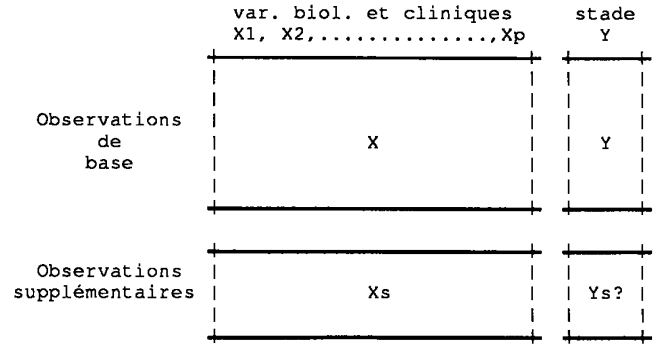
D'autre part, réaction cutanée à la Candidine et à la Tuberculine et forte présence de *Candida albicans* sur la langue sont des séquelles d'infections opportunistes. On trouve aussi avec les états intermédiaires avancés du mal la présence de signes fonctionnels et neurologiques ainsi que les infections de type non SIDA.

Cette analyse confirme les résultats obtenus dans (2) et surtout l'intérêt des paramètres: nombres de cellules T4, de plaquettes; taux de β2- microglobulines, d'IgG, d'IgA et d'IgM ainsi que l'apparition de l'antigénémie et l'hypersensibilité à la candidine et à la tuberculine.

Au-delà de cette description, on se propose d'exprimer l'état de l'infection (stade) en fonction des paramètres biologiques et cliniques. À cet effet, on a utilisé la Régression Linéaire Multiple.

2.3 Régression Linéaire Multiple

Nous avons exploré plusieurs voies et rendrons compte brièvement des résultats obtenus. Le numéro du stade est ici considéré par nous comme une



variable numérique, que l'on cherche à exprimer en combinaison linéaire des données, éventuellement transformées ou remplacées par les facteurs issus de l'analyse du tableau en (0,1) ou du tableau des cumuls par stade.

La Régression Linéaire Multiple s'applique à des données de la forme indiquée sur le tableau: Y est la "variable à expliquer"; X1, X2,....., Xp sont les "variables explicatives". Dans notre étude la variable à expliquer Y est le stade de l'infection; elle prend des valeurs discrètes, comprises respectivement entre 2 et 14 ou entre 1 et 3 selon qu'on considère la classification HIP ou CDC. Les Xi sont soit les variables biologiques et cliniques considérées plus haut soit les facteurs obtenus par l'analyse du §2.2 ou celle du tableau des cumuls (§2.4).

On cherche à approcher Y par une combinaison linéaire Y' des variables explicatives choisies.

Il faut rappeler qu'en toute rigueur, l'application de la RLM requiert qu'il n'y ait pas de valeurs manquantes; ici, on a repris 2084 observations sans données manquantes: qui ont été les seules utilisées pour les régressions sur variables et sur facteurs.

Le critère d'évaluation de la régression est le coefficient de corrélation multiple R entre la variable à expliquer et les variables explicatives: R n'est autre que le coefficient de corrélation usuel entre la variable à expliquer Y et son estimation Y'; plus R est proche de 1, meilleure est la qualité de la régression.

On remarque néanmoins que la valeur de R augmente automatiquement avec le nombre des variables explicatives. Lorsque les variables explicatives sont corrélées, la formule obtenue est instable: une petite variation des variables explicatives peut induire une variation importante des coefficients de la régression. Pour ces raisons il convient, pour une bonne application de la RLM, d'avoir un petit nombre de variables explicatives indépendantes entre elles.

Vu l'hétérogénéité des variables et les variations d'ordre de grandeur de plusieurs d'entre elles, on ne peut appliquer directement la RLM. On a donc, dans un premier temps, considéré les logarithmes à base 10 des variables explicatives et de la variable à expliquer.

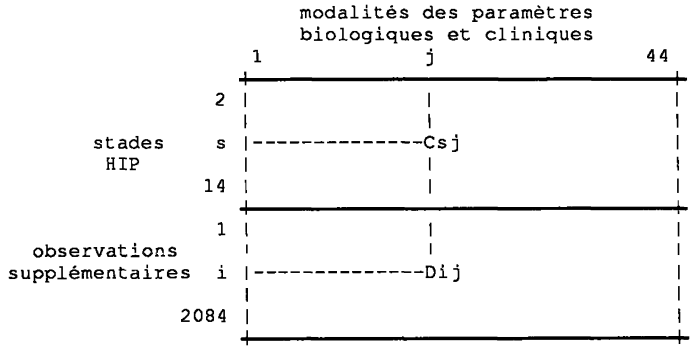
Dans un deuxième temps, et afin de prendre en compte les remarques précédentes, on a pris comme variables explicatives les facteurs obtenus par l'AFC du tableau X. Ces facteurs sont par construction orthogonaux entre eux. D'autre part, nous en avons limité le nombre. Un tableau donne les coefficients de corrélation multiple, R, obtenus.

Variable à expliquer	variables explicatives	R
log(stade HIP)	log(par. biologiques et cliniques)	.886
log(stade CDC)	log(par. biologiques et cliniques)	.834
stade HIP	7 premiers facteurs de l'AFC	.753
stade HIP	1 premier facteur de l'AFC	.693
log(stade HIP)	7 premiers facteurs de l'AFC	.723
stade CDC	7 premiers facteurs de l'AFC	.660
log(stade CDC)	7 premiers facteurs de l'AFC	.625

On voit que, pour les mêmes variables explicatives, le coefficient de corrélation multiple est meilleur pour le stade HIP que pour la classe CDC: ceci est dû principalement à ce que HIP constitue une interpolation de CDC. Pour une même classification, la corrélation diminue peu lorsqu'on considère les facteurs); pour la classification selon le HIP, lorsqu'on se limite au premier facteur, le coefficient R passe de .75 à .69.

2.4 Analyse du tableau cumulé et régression

Dans une étude de prévision du pH en fonction du pourcentage de présence d'espèces de diatomées (5), nous avons proposé une méthode de régression sur les facteurs obtenus par l'AFC de différents tableaux construits en croisant la variable à expliquer (pH) subdivisée en classes avec l'ensemble des espèces de diatomées. Cette méthode a fourni le meilleur coefficient de corrélation multiple. Ceci a été récemment confirmé dans une étude comparative de différentes méthodes de prévision (6). Nous avons appliqué la même méthode (5) pour construire, par cumul à partir du tableau D des données codées en (0,1), un tableau C croisant les stades d'infection (exceptés, 8, 10, 12, de faible effectif), avec l'ensemble des modalités des paramètres biologiques et cliniques.



De façon précise, C croise les 10 stades d'infection selon le HIP avec les modalités des paramètres biologiques et cliniques: Csj est le nombre d'observations au stade s rentrant dans la modalité j; la ligne s de C est le cumul des lignes i de D pour lesquelles i est au stade s. Un tel tableau exprime le mieux les liaisons entre modalités du stade et les modalités des paramètres biologiques. À l'analyse de C, on a adjoint, en supplémentaire, le tableau D.

rang	1	2	3	4	5	6	7
Lambda	.215	.095	.09	.087	.069	.016	.0025
Taux	36.9	16.3	15.47	15.08	11.9	2.87	.445
Cumul	36.9	53.22	68.7	83.78	95.68	98.55	99.

Dans cette analyse, la prédominance du 1-er facteur est encore plus nette que dans celle du §2.2.

A partir des facteurs obtenus sur les observations i placées en éléments supplémentaires on a effectué des régressions, comme au §2.3., en considérant dans chaque cas le stade ou son logarithme à base 10 comme variable à expliquer. Les principaux résultats sont indiqués dans un tableau.

Variable à expliquer	variables explicatives	R
stade HIP	7 premiers facteurs de l'AFC	.619
stade HIP	le premier facteur de l'AFC	.574
log(stade HIP)	7 premiers facteurs de l'AFC	.615
stade CDC	2 premiers facteurs de l'AFC	.753
log(stade CDC)	2 premiers facteurs de l'AFC	.711

On note une nette amélioration des coefficients de corrélation multiple par rapport à ceux du §2.3; et, d'autre part, comme au §2.3, le coefficient de corrélation est meilleur lorsqu'on considère les stades selon le HIP que lorsqu'on considère les stades selon le CDC; et, pour les stades selon le HIP, R diminue peu si l'on ne considère qu'une seule variable explicative, le premier facteur (R passe de .85 à .77).

3 Nouvelle analyse fondée sur un tableau sans donnée manquante

3.1 Choix du tableau et codage

Après avoir, au §2, confirmé le lien de toutes les variables retenues avec la gravité de l'atteinte HIV, exprimée en terme de stade, nous concentrerons désormais notre attention sur un tableau sans lacune. Plusieurs analyses ont été effectuées sur les variables décrivant, en termes généraux, l'état du système immunitaire, et mettant en jeu des observations de plus en plus nombreuses; l'analyse présentée ici est la première d'une série dont il est rendu compte ailleurs.

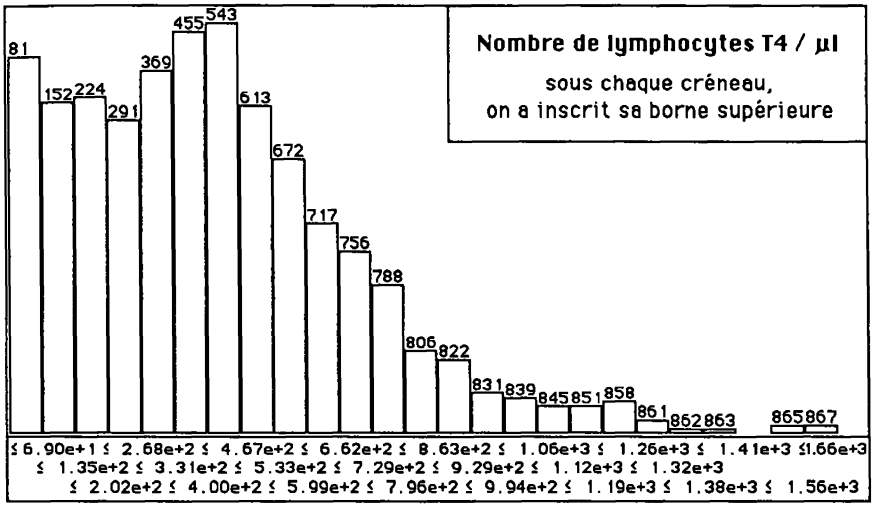
Après consultation d'un listage de données manquantes, on a retenu les 10 variables:

{LYMP, T4, T8, AG, ß, IGG, IGM, IGA, PLQT, SYMP};

et, corrélativement, 867 observations où sont présentes ces 10 variables. Pour plus de rigueur, la dernière de ces variables, SYMP, n'est elle-même retenue que comme élément supplémentaire, car elle comporte une modalité 'SIDA', dont l'inclusion semble constituer une pétition de principe.

Quant au codage, les 8 variables continues n'ont pas été découpées en classes (pour être codées en 0,1) mais soumises à un codage linéaire par morceau (ou codage flou: cf. 3). En bref, à chaque variable correspond une suite de modalités (généralement 3, mais 4 pour T4) allant de faible à fort ({-,=,+} ou {<<,<=,>=,>>} et définies par des "valeurs pivot".

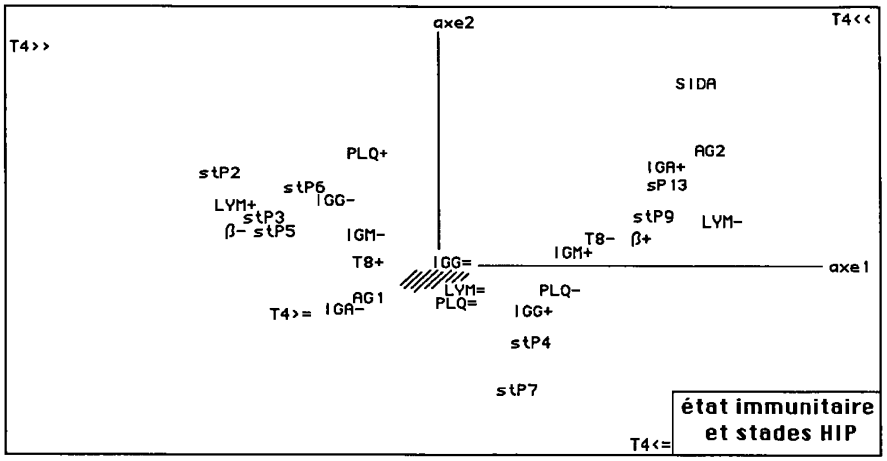
Une valeur coïncidant avec un pivot est codée 1 dans la modalité correspondante et zéro ailleurs; mais, pour une valeur tombant entre deux pivots, la note est partagée entre deux modalités, (suivant une formule linéaire, d'où le nom du codage), avec au total 1.



Pour choisir les pivots, on se base sur des histogrammes, tel que celui présenté pour 'T4'.

S'il y a, à gauche ou à droite, un étalement de la distribution, le pivot n'est pas placé à l'extrémité correspondante de l'intervalle de variation, mais en retrait. De ce fait, les valeurs situées au-delà du pivot de droite sont codées avec 1 dans la modalité correspondante, comme si elles étaient égales à ce pivot; on fait de même en deçà du pivot de gauche. On perd ainsi en précision; mais,

```
Données Pasteur relatives aux séropositifs
B:tek:Ctk**Dcodx: bornes pour le découpage des variables
le nombre des variables est  10
LYMP a  3 modalités dont les sigles et valeurs pivot (./µl) sont
    LYM-    LYM=    LYM+          216  1739  3654
  T4 a  4 modalités dont les sigles et valeurs pivot (./µl) sont
  T4<<   T4<=   T4>=   T4>>      4   274   524   882
  T8 a  3 modalités dont les sigles et valeurs pivot (./µl) sont
     T8-     T8=     T8+           75   789  1642
  AC a  2 modalités dont les sigles et valeurs pivot (N--O) sont
         AG1    AG2                  1     2
   ß a  3 modalités dont les sigles et valeurs pivot (µg/l) sont
      ß-      ß=      ß+           900  2600  4400
 IGG a  3 modalités dont les sigles et valeurs pivot (cg/l) sont
    IGG-    IGG=    IGG+         1000  1700  3000
 IGM a  3 modalités dont les sigles et valeurs pivot (cg/l) sont
    IGM-    IGM=    IGM+           40   130   380
 IGA a  3 modalités dont les sigles et valeurs pivot (cg/l) sont
    IGA-    IGA=    IGA+           60   239   720
 PLQ a  3 modalités dont les sigles et valeurs pivot (k/µl) sont
    PLQ-    PLQ=    PLQ+           80   200   320
 SYM a  5 modalités {nul, fonct, neuro, infection non SIDA, SIDA}
  sNON  sFON  sNEU  sINF  sSID      1  2  3  4  5
```



d'une part, le codage conserve sa simplicité; et, d'autre part, il importe à la stabilité de l'analyse qu'il n'y ait pas de modalités extrêmes de poids très faible, donc sensibles aux fluctuations d'échantillonnage.

Au total, on a un tableau (867 × 32). Parmi les 32 modalités, seules sont en (0,1) les deux modalités {AG1, AG2} (absence ou présence d'antigénémie-p25) et les 5 modalités de la variable supplémentaire SYMP.

Les 25 modalités suivant lesquelles on a codé les 8 variables continues sont des nombres positifs variant entre 0 et 1; mais, comme dans le codage disjonctif *complet*, le total de toute ligne afférente à une observation est égal au nombre des variables de base retenues.

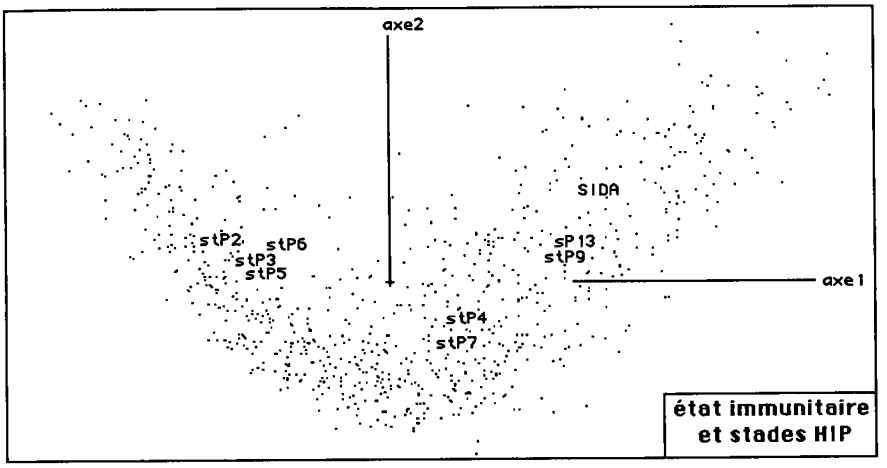
Pour l'analyse factorielle, le tableau principal est, comme au §2.4, construit par cumul; mais parce que certains stades sont très peu représentés, on n'en a retenu que 9, énumérés ci-dessous avec leurs fréquences:

stP2(100); stP3(109); stP4(299); stP5(43); stP6(27); stP7(85); stP9(70); sP13(54); SIDA(65).

De plus, les stades 5 et 6 étant eux-mêmes peu attestés, on en a introduit le cumul, noté sP56; et, comme de règle, on a considéré les classes du CDC, dont ls stades de l'HIP sont des subdivisions.

3.2 Analyse factorielle

Compte tenu des valeurs propres et taux, on s'est borné à figurer le plan (1,2); même si les axes suivants recèlent, sans doute, des faits intéressant la distinction fine des stades. On a, d'une part, un plan où figurent les modalités des variables et les stades; et, d'autre part, un plan à échelle réduite, avec un



trace : 1.143e-1									
rang	1	2	3	4	5	6	7	8	
lambda	910	131	47	31	13	7	3	1	e-4
taux	7963	1151	410	271	112	59	24	10	e-4
cumul	7963	9113	9523	9794	9907	9966	9990	10000	e-4

Tableau des valeurs propres et taux d'inertie

nuage dont chaque point figure l'une des 867 observations individuelles (non cumulées par stade; éléments supplémentaires).

Il apparaît que les stades forment 4 îlots: {2, 3, 5 ,6}, {4,7}, {9,13} et SIDA; ces îlots s'échelonnent sur le nuage en croissant des observations. (L'image présentée au §2.2 suggère d'ailleurs une répartition analogue.) Nos données représentant essentiellement l'état du système immunitaire (la variable AG étant seule spécifique du SIDA), il n'y a pas à s'étonner que contrairement à l'ordre des nombres, 4 aille avec 7, laissant en arrière 5 avec 6.

L'analyse discriminante barycentrique aidera à préciser les relations entre stades; mais il faut d'emblée noter l'ambiguïté de la notion de stade qui, pour le CDC, caractérise globalement la gravité de l'état pathologique; et dont l'HIP entend faire un instrument de mesure biologique précis, pouvant servir à mesurer la réponse des patients à un traitement immuno-stimulant.

Quant aux modalités, on trouve confirmées, pour les variables conservées, les conclusions du §2: avec la gravité de l'atteinte, {LYM, T4, T8, PLQ} décroissent; tandis que {ß, IGG, IGM, IGA} augmentent: autrement dit, l'atteinte est pancytopénique et hyper-ɣ-globulinémique.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	0	66	22	6	17	5	3	1	0	0	0	0	0	1
3	0	17	58	22	10	10	3	3	3	0	0	0	0	3
4	0	1	2	104	4	2	18	0	12	0	1	0	6	3
5-6	0	14	19	11	9	7	5	0	0	0	0	0	0	0
7	0	2	6	72	2	2	40	0	12	0	6	1	10	3
9	0	0	0	19	0	1	3	0	11	0	1	0	4	4
13	0	0	2	34	0	0	10	0	12	0	0	0	17	6
14	1	0	0	31	1	0	3	0	20	0	0	1	17	45
2	0	56	19	3	12	4	0	1	0	0	0	0	0	0
3	0	14	53	19	7	6	2	3	1	0	0	0	0	3
4	0	1	0	99	4	2	17	0	12	0	1	0	6	2
5	0	25	21	21	15	7	10	0	0	0	0	0	0	2
6	0	2	11	4	2	6	3	0	2	0	0	0	1	0
7	0	2	4	70	2	1	37	0	12	0	6	1	9	3
9	0	0	0	18	0	1	3	0	11	0	1	0	4	4
13	0	0	1	34	0	0	10	0	12	0	0	0	17	6
14	1	0	0	31	1	0	3	0	20	0	0	1	17	45

en colonne: stade réel HIP; en ligne: affectation

3.3 Analyse discriminante barycentrique

Le programme 'discri' permet d'affecter un ensemble d'individus à un ensemble de centres, pourvu que l'on possède, pour chacun des deux ensembles, un fichier de coordonnées sur le même système d'axes factoriels. Ici, on a pris pour individus les 867 observations retenues; et on a considéré deux systèmes de centres: d'une part, le système de 9 stades:

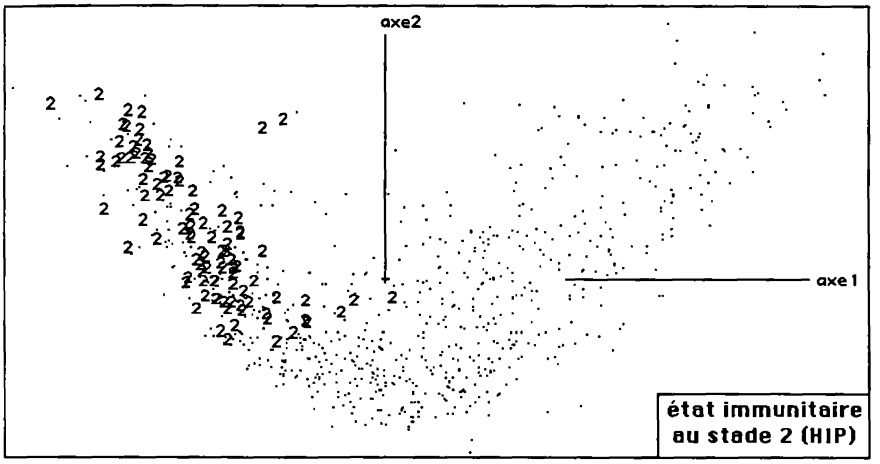
{stP2, stP3, stP4, stP5, stP6, stP7, stP9, sP13, SIDA};

et, d'autre part, un système réduit, où stP5 et stP6 sont cumulés en sP56:

{stP2, stP3, stP4, sP56, stP7, stP9, sP13, SIDA};

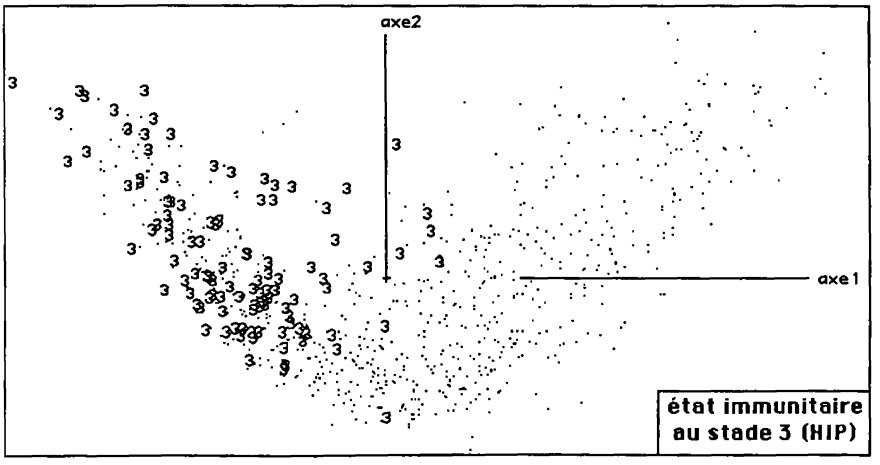
Dans l'un et l'autre cas, chaque observation est simplement affectée au centre dont elle est le plus proche: nous rappelons que ces centres ne sont autres, chacun, que le centre de gravité, ou *barycentre*, de l'ensemble des observations relevant d'un même stade; d'où le terme de *barycentrique*. On a ainsi, pour chaque observation, un stade réel et un stade estimé par affectation.

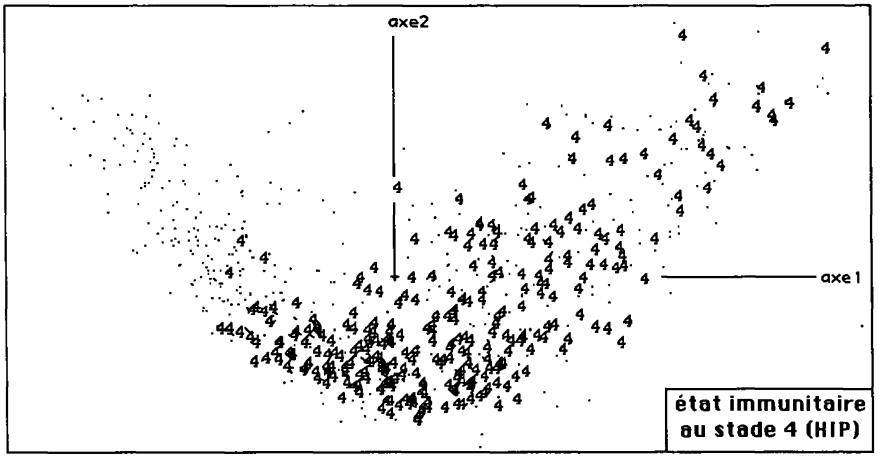
L'idéal serait la coïncidence entre stade réel et stade estimé. Le double tableau de contingence publié ici permet d'apprécier dans quelle mesure cet idéal est atteint pour chacun des deux systèmes de centres: la tendance générale seule est satisfaisante. On retrouve la proximité déjà notée entre stP4 et stP7: 70 (ou 72) observations du stade 4 sont affectées à stP7: on voit même, sur la ligne 7, que, parmi les observations affectées à stP7, la majorité ont pour stade réel 4; mais il faut se souvenir que les observations de ce stade sont les plus nombreuses (299 au stP4 contre 85 au stP7). On note encore que le cumul sP56 ne parvient pas à rallier plus des 2/3 des individus affectés à stP5 et stP6 pris séparément: ce qui suggère l'importance des facteurs de rang supérieur à 2.



Pour rendre visible le succès de l'affectation barycentrique et ses limites, on a construit des graphiques du plan (1,2) où figurent d'une part toutes les 867 observations, marquées d'un simple point, et d'autres part les observations d'un stade déterminé marquées du chiffre de ce stade (ou d'une lettre: 'D' pour 13, ou 'E' pour 14=SIDA). (On construit rapidement de tels graphiques à l'aide des programmes 'soustab' et 'planF'.)

Tandis que la dispersion de stP3 s'écarte peu de celle de stP2, on trouve dans stP4 des profils immunitaires très perturbés, s'étendant jusqu'à l'extrémité {F1>0, F2>0} (SIDA) du croissant des observations. On doit souligner

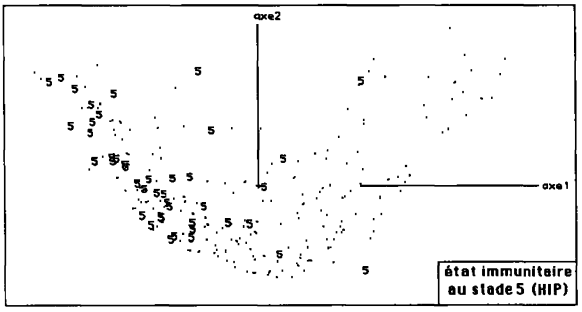


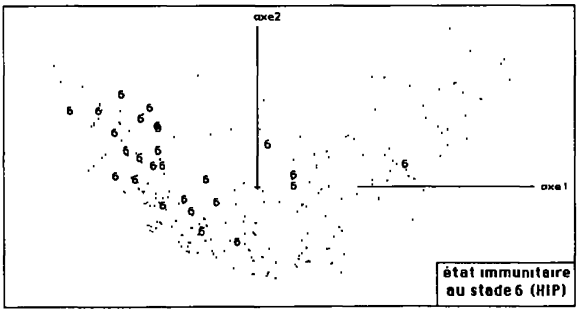


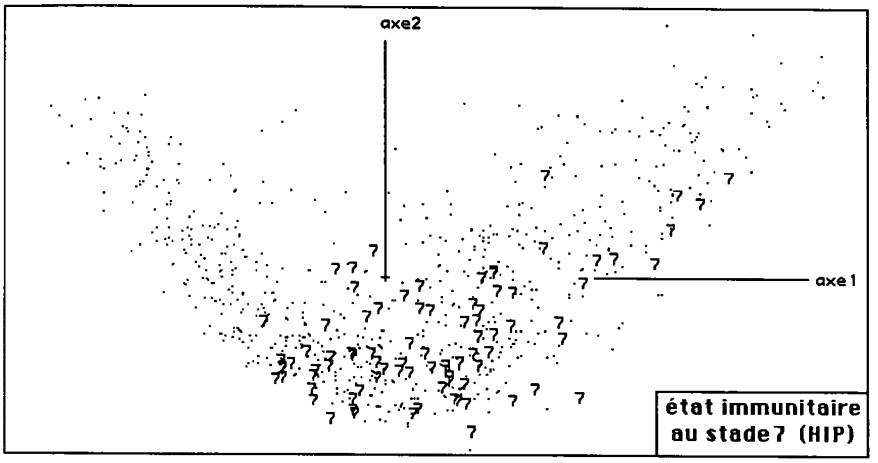
l'hétérogénéité du stade stP4, défini suivant des critères mixtes; et qu'on n'est sans doute pas capable, présentement, de subdiviser convenablement.

On ne s'étonnera pas de voir que, quant à l'état du système immunitaire, les stades stP5 et stP6 sont moins perturbés que stP4: car stP5 et stP6 sont précisément définis par des taux de T4 normaux ou > 500 / µl.

Dans le plan (1,2), les sous-nuages de stP5 et stP6 se recouvrent à peu près; et l'on n'y trouve que par exception des points ayant un facteur F1 positif.



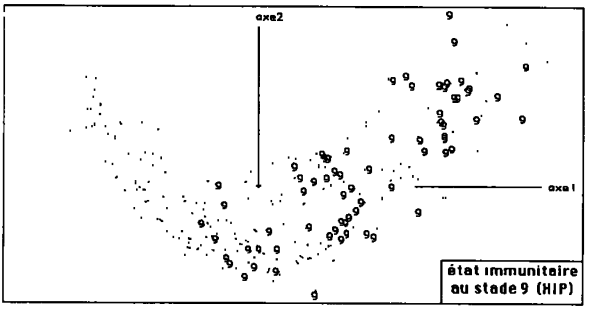


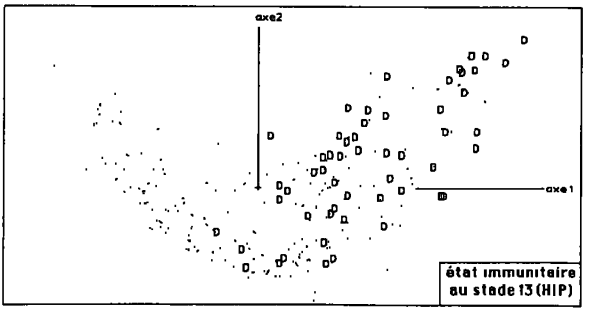


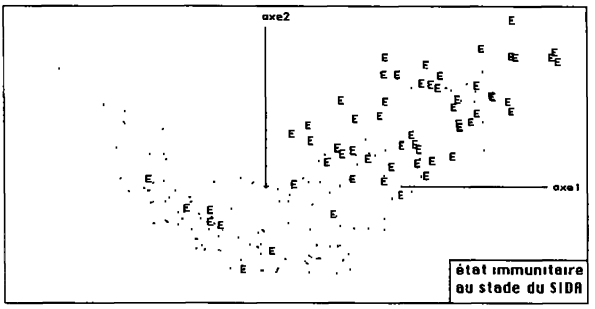
Dans le plan (1,2), le domaine de stP7 semble compris dans celui de stP4, étant plus concentré que celui-ci au voisinage du demi-axe {F2<0}; mais le petit nombre des observations relevant de stP7 ne permet pas d'être affirmatif.

Avec stP9 et sP13, on trouve des états immunitaires différant peu de ceux qui se rencontrent dans le SIDA proprement dit.

En effet, le tableau des affectations montre que des observations des stades stP9 et sP13 sont fréquemment affectées au SIDA; mais il est plus rare que des cas de SIDA soient affectés à stP9 ou stP13; ce qu'il convient de rapprocher du fait que les centres des stades stP9 et sP13 ne sont pas si écartés dans le plan (1,2) que celui du SIDA (stade 14) proprement dit.







NB Les observations classées SIDA du côté (F1<0), sont celles de sujets ayant des tumeurs de Kaposi; ces sujets ont un profil immunologique similaire à celui des porteurs asymptomatiques.

4 Conclusions et perspectives

L'analyse des correspondances a montré l'intérêt des paramètres biologiques (nombre de cellules T4, T8, plaquettes, lymphocytes, les taux d'IgG, d'IgA, d'IgM et de β2-microglobuline ainsi que l'apparition de l'antigénémie et l'hypersensibilité à la Candidine et à la Tuberculine) qui sont liés à l'évolution de l'infection par le VIH. Et les cliniciens s'intéressent à l'aspect pronostique de chacun de ces paramètres.

Le présent travail a pris la forme sous laquelle il paraît, après de longues discussions, notamment avec deux Médecins en stage à l'Institut Pasteur, Cruz AYERBE et José ALCAMI. Le Pr. J.-P. BENZÉCRI, quant à lui, nous a prodigué les encouragements,... et les critiques! Il rend compte lui-même, dans ce cahier (cf. 7), d'analyses ultérieures faites sur des tableaux construits à sa requête.

Références Bibliographiques

1. F. Tekaïa, F. Bimet, Ph. Sansonetti, J. Riou, I. Sauvaget, J.-M. Claverie, 1989: Relational Biomedical Data Bases pp. 483-497, 7th European ORACLE Users Group Conference BRUSSELS - BELGIUM.

2. F. Tekaïa, J. de Saint Martin, Ph. Sansonetti, F. Vuillier, J.-M. Claverie, 1988: Intérêt de l'Antigénémie dans la définition des stades de l'infection par le Virus VIH. *CAD* Vol XIII - n° 4, pp. 407-424.

3. J.-P & F. Benzécri, 1989: Codage linéaire par morceaux et équation personnelle. *CAD*, Vol. XIV, n° 3, pp. 331-336.

4. M. Roux, 1979: Estimation des paléoclimats d'après l'écologie des foraminifères. *CAD* Vol. IV n° 1 pp. 61-79.

5. F. Gasse, F. Tekaïa, 1983: Transfer function for estimating paleoecological conditions (pH) from East African diatoms. *Hydrobiologia* 103, pp.85-90.

6. C. J. F. Ter Braak, H. Van Dam, 1989: Interferring pH from diatoms: a comparison of old and new calibration methods. *Hydrobiologia* 178, pp. 209-223.

7. J.-P. Benzécri, 1990: Analyse des données biologiques et pathologie clinique. *CAD*, Vol XV, n°3; et, *ibid.*: État du système immunitaire et histoire clinique chez les patients infectés par le VIH.