ANALYSE MULTIVARIEE DES DONNEES DISCRETES: UNE APPLICATION A L'ETUDE DE LA DISTRIBUTION REGIONALE DES VIRUS DU SIDA DANS L' ANGOLA

H. Bacelar-Nicolau

CEAUL / INIC - UNIV. LISBOA
Laboratório de Estatística e Análise de Dados - Faculdade de Psicologia e de Ciências da Educação da Universidade de Lisboa - Alameda da Universidade, 1600 Lisboa, Portugal

RESUME: En Analyse Classificatoire Discrète on doit souvent traiter des tableaux de fréquences décrivant un ensemble de Catégories / Unités de Données par un ensemble de variables discrètes. Le traitement par des méthodes de classification hiérarchique ascendante de ces tableaux, de nature généralement descriptive, peut alors être complété par l'étude inférentielle de plusieurs tableaux de contingence uni / bi ou multidimensionnelles. On associe donc à l'analyse exploratoire du tableau global, des analyses partielles confirmatoires, dont les hypothèses à tester sont suggérées par les résultats issus de la première analyse.
Cette méthodologie est appliquée dans le travail présent à l'étude de la distribution régionale des virus HIV1 et HIV2 du SIDA dans la population de l' Angola. On utilise des approches classique aussi bien que probabiliste pour l'Analyse Classificatoire Hiérarchique du tableau initial. On fait des analyses classiques et loglinéaires pour les Tableaux de Contingence partiels.

Mots clés: analyse classificatoire, tableaux de contingence, classification hiérarchique ascendante, modèles loglinéaires

---

Ce travail a été partiellement supporté par le Project d'Analyse des Données entre le Laboratoire de Statistique et Analyse des Données (LEAD) de la Faculté de Psichologie et Sciences de l'Education, Université de Lisbonne et le Project CLOREC, de l' INRIA, dans le cadre du Programme de Coopération Scientifique et Technique Luso-Française entre l' Instituto Nacional de Investigação Cientifica (INIC) et l' Ambassade de France au Portugal.

1. INTRODUCTION

L' étude statistique développé dans ce travail concerne une recherche épidémiologique sur les virus de l'immunodéficience humaine du type I (HIV1) et du type II (HIV2) dont le but est d'analyser comparativement la distribution du HIV sur six régions de la République de l'Angola. Il s'agit d'une partie d'un travail plus vaste qui a été realisé au Laboratoire de Statistique et Analyse des Données (LEAD) de la Faculté de Psychologie et des Sciences de l'Education, à l'Université de Lisbonne.

Les données recueillies ont été rassemblées dans des tableaux de fréquences, faisant le croisement des régions et/ou des sous-groupes spécifiques d'individus dans ces régions, avec les deux types de virus auxquelles a été ajouté le type HIV1+2 où sont simultanément présents le HIV1 et le HIV2 (Santos-Ferreira et al, 1990). L'information contenue dans ces tableaux est la donnée de base de notre étude.

Les échantillons observés étant de taille importante, le plan d'échantillonage n'a pourtant pas respecté des critères statistiques de représentativité concernant la distribution de la population par région. Le traitement statistique a donc du tenir compte de ce fait. Nous l'avons divisé en deux étapes complémentaires. Dans la première partie de l'étude nous avons utilisé des méthodes d'analyse multivariée exploratoire, sans caractère inférenciel, cherchant à trouver les aspects les plus robustes, caractéristiques de la structure sousjacente aux données. Dans la seconde on a fait des analyses confirmatoires où sont étudiées du point de vue de la statistique classique des hypothèses particulières suggérées dans l'étape initialle.

La méthodologie statistique adoptée peut donc être décrite shématiquement en deux étapes, de la façon suivante:
1 - Analyse de données multivariées.
Développée à partir de modèles d'analyse classificatoire hiérarchique ascendante basés sur le coefficient d'affinité (Bacelar-Nicolau, 1980, 1985, 1988, 1990). Des critères d'aggrégation classiques (Lerman, 1981) ont été utilisés, ainsi que des critères probabilistiques de la famille VL de la Vraisemblance du Lien (Bacelar Nicolau, 1972, 1980, 1985; Lerman, 1970, 1981; Nicolau, 1983, 1985). L'étude a été complétée par l'application d'analyses factorielles des correspondances (Benzecri et al, 1973; Nicolau, 1973), dont les résultats ne sont pas présentés ici: ils ont simplement renforcé ceux qui ont été obtenus par les analyses classificatoires.
2 - Analyse des tableaux de contingence bi et tridimensionnels.
Pour tester la signification de certains des résultats que l'analyse multivariée exploratoire met en évidence, des analyses des tableaux de contingence particuliers ont été accomplies. Nous avons considéré l' approche classique et celle des modèles loglinéaires (Bishop et al, 1975; Everitt, 1980).

Les calculs ont été effectués à l'aide des programmes CLASSIF pour la classification hiérarchique ascendante basée sur le

coefficient d'affinité (Bacelar-Nicolau et al,1980) et BIDIM et TRIDIM pour l' analyse des tableaux de contingence bi et tridimensionnels (Mendes Leal,1986).

Dans les deux sections suivantes nous décrivons brièvement les méthodes d'analyse classificatoire et l'approche loglinéaire employées. La dernière section du travail présente une partie des résultats issus de l'application des méthodes statistiques aux données concernant la distribution régionale des virus du SIDA.

2. ANALYSE EXPLORATOIRE : CLASSIFICATION HIERARCHIQUE ASCENDANTE

Cette analyse a été fondée sur le coefficient d'affinité défini entre paires d'éléments de l'ensemble des régions et/ou de certains sous-groupes d'individus choisis dans ces régions. Les régions et les sous-groupes sont décrits par leurs fréquences en individus présentant les virus HIV1, HIV2, ou HIV1+2 et le nombre de séronégatifs.

étant donné le tableau de contingence $N=[n_{ij}]$, d' élément générique n_{ij} , $i=1,\ldots,k$; $j=1,\ldots,r$; soit $n_{.j}$ et $n_{.j'}$ les fréquences marginales des colonnes j et j', respectivement.

L'affinité a(j,j') entre chaque paire d'éléments (j,j') de l'ensemble à classifier est définie par le produit interne des racines carrées des profils associés à j et j'. Alors, on a:

$$a(j,j')=\Sigma\sqrt{(n_{ij}/n_{.j})(n_{ij'}/n_{.j'})}=\Sigma\sqrt{n_{i/j}\ n_{i/j'}}$$

où la sommation est indicée par i=1,k.

Il est facile de démontrer que l'affinité est un coefficient de similarité prenant ses valeurs dans l'intervalle [0,1]. Des propriétés du coefficient de similarité a(j,j') ont été étudiées par exemple en Bacelar-Nicolau (1980). Nous avons alors étendu au domaine de l'analyse classificatoire la notion d'affinité entre lois de probabilité proposée par Matusita (1955,1967).

Le coefficient simple a(j,j') peut être associé à des critères d'aggrégation classiques comme ceux du lien simple, du lien complet ou du lien moyen , engendrant donc des modèles d'analyse classificatoire hiérarchique ascendente, de nature purement descriptive.

D'autre part, la considération d'hypothèses de référence appropriées permet d'introduire un nouveau coefficient de similarité $\alpha(j,j')$ associé à la notion de base a(j,j') d'affinité, à savoir :

$$\alpha(j,j') = P(A(j,j') \leq a(j,j'))$$

où A(j,j') représente la variable aléatoire dont a(j,j') est une réalisation. Il s'agit donc de la fonction de répartition de A(j,j') au point a(j,j').

La probabilité $\alpha(j,j')$ mesure la validité du lien défini par l'affinité $a(j,j')$, sous l'hypothèse de référence considérée: nous l'avons appellé le coefficient de validité de l'affinité . Des propriétés de ce coefficient ont été présentées par Bacelar-Nicolau (1980,1988,1990).

Dans l'analyse classificatoire des tableaux de contingence nous nous référons généralement à deux sortes d' hypothèses (Bacelar-Nicolau,1988). La première, utilisant un modèle multinomial basé sur le théorème limite de la méthode δ (Tiago de Oliveira,1982) ne peut pas être directement apliquée dans le cas des tableaux comprennant des zéros; on peut cependant ajouter une constante positive à toutes les cases du tableau, à l'image du procédé employé dans l'analyse loglinéaire des tableaux de contingence. La seconde hypothèse de référence, s'appuyant sur le thèorème limite de Wald et Wolfovitz (Fraser,1957), utilise un modèle permutationnel; elle suppose une distribution uniformément définie sur le sous-espace des vecteurs engendrés par les permutations d'indices du vecteur associé à chaque élément j de l'ensemble à classifier.

Dans chacun des cas on fera le calcul respectivement de la valeur moyenne et de la variance asymptotiques et on sera amené à une loi limite normale réduite. Nous avons alors, le résultat commun suivant:

$$\alpha(j,j') \approx P(A^*(j,j') \leq a^*(j,j')) \approx \Phi(a^*(j,j'))$$

où A^* représente la variable aléatoire asymptotiquement centrée et réduite dont a^* est une réalisation et Φ la fonction de répartition de la loi normale réduite.

La définition du coefficient probabiliste α entraine naturellement l'utilisation de critères d'aggrégation de nature probabiliste, du type des algorithmes appartenant à la famille VL (Nicolau,1983,1989).

Le choix des modèles d'analyse hiérarchique ascendante mieux ajustés aux données aussi bien que celui des niveaux les plus significatifs dans chaque hiérarchie, est fondé sur la comparaison des valeurs des statistiques "des niveaux", locale et globale (Bacelar-Nicolau,1970,Lerman,1970).

3. ANALYSE CONFIRMATOIRE: MODELES LOGLINEAIRES

Dans l'approche loglineaire au traitement statistique des tableaux de contingence on étudie des modèles où les logarithmes des fréquences espérées (ou des probabilités correspondantes) sous chaque hypothèse de référence, sont des fonctions linéaires d'un certain nombre de paramètres. Ces paramètres ont une interprétation connue: ils représentent les effets des variables ou des interactions entre variables dans la détermination des valeurs des observations.

L'approche loglinéaire aux tableaux de contingence permet de calculer les estimations de la grandeur des effets représentés par les paramètres, aussi bien que de tester la qualité de l'ajustement du modèle aux données. En plus, la formalisation mathématique de l'analyse des données qualitatives par les modèles loglinéaires devient tout à fait semblable à celle de l'analyse de la variance pour des données quantitatives.

Prenons, par exemple, le cas de l'indépendance entre variables, dans un tableau bidimensionnel de dimension r*s . Il s'agit donc d'un tableau où les r lignes représentent les catégories ou modalités d'une variable et les s colonnes les modalités de l'autre (Everitt,1980). Il est bien connu dans l'approche classique, que l'hypothèse statistique correspondante peut s'écrire de la façon suivante:

$$H_0: p_{ij} = p_{i.} * p_{.j} \qquad i=1,\ldots,r \; ; \; j=1,\ldots,s$$

Il s'agit là d'un modèle multiplicatif: la probabilité qu'une observation de la population tombe dans la case (i,j) est égale au produit des probabilités marginales.

On peut démontrer qu' un modèle aditif équivalent au précédent est donné par:

$$\log f_{ij} = u + u_{1(i)} + u_{2(j)} \qquad i=1,\ldots,r \; ; \; j=1,\ldots,s$$

où les f_{ij} représentent les fréquences espérées sous l'hypothèse H_0 et les paramètres du modèle, à savoir:

$$u = \Sigma\Sigma \log f_{ij} / (r*s) \qquad i=1,\ldots,r \; ; \; j=1,\ldots,s$$

$$u_{1(i)} = \Sigma \log f_{ij} / s - u \qquad j=1,\ldots,s$$

$$u_{2(j)} = \Sigma \log f_{ij} / r - u \qquad i=1,\ldots,r$$

représentent respectivement, l'effet de la moyenne totale, l'effet principal de la i-ème modalité de la variable en ligne et l'effet principal de la j-ème modalité de la variable en colonne. On a:

$$u_{1(.)} = \Sigma u_{1(i)} = 0 \qquad i=1,\ldots,r$$

$$u_{2(.)} = \Sigma u_{2(j)} = 0 \qquad j=1,\ldots,s$$

ce qui revient à dire que les effets "se compensent".

Dans la notation adopté on utilise donc des indices numériques pour représenter les variables et des indices alphabétiques pour représenter les catégories dans chaque variable.

Le modèle loglinéaire d'indépendance introduit ci-dessus est un modèle théorique, bien evidemment. Dans la pratique on doit estimer soit les fréquences espérées soit les paramètres du modèle.

Pour généraliser le modèle précédent au cas de non indépendance on introduira un nouveau terme qui représente l'effet de l'interaction entre les catégories i et j des deux variables, dans l'expression des fréquences logarithmiques, soit:

$$\log f_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

où on a aussi:

$$u_{12(i.)} = \Sigma\, u_{12(ij)} = 0 \qquad j=1,\ldots,s$$

$$u_{12(.j)} = \Sigma\, u_{12(ij)} = 0 \qquad i=1,\ldots,r$$

Ce dernier modèle est dit saturé, car le nombre de paramètres est précisément égal à celui des fréquences espérées (le nombre de degrés de liberté est égal à zero). Dans le modèle saturé, tester l' indépendance est équivalent à tester si tous les paramètres d'interaction de 1er ordre sont nuls, soit l'hypothèse suivante:

$$H_0 : u_{12(ij)} = 0 \qquad i=1,\ldots,r \;;\; j=1,\ldots,s$$

Dans l' étude des tableaux tridimensionnels de dimension r*s*t le modèle saturé apparait comme une extension tout à fait naturelle du modèle saturé défini dans le cas bidimensionnel. On a:

$$\log f_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(ij)} + u_{123(ijk)} \qquad i=1,\ldots,r \;;\; j=1,\ldots,s \;;\; k=1,\ldots,t$$

où, en plus des effets principaux pour chaque variable, et des effets d' intersection de 1er ordre, on trouve maintenant des effets d'intersection de 2d ordre, expliquant l'association simultané entre les trois variables.

L'anulation de certains termes dans le modèle saturé définit de nouveaux modèles, non saturés, plus intéressants du point de vue statistique. Le modèle correspondant à l'hypothèse de non existence d'interaction de 2d ordre, par example, s'écrit en faisant sortir les derniers i*j*k termes du modèle saturé. Nous avons:

$$H_0 : u_{123(ijk)} = 0 \qquad i=1,\ldots,r \;;\; j=1,\ldots,s \;;\; k=1,\ldots,t$$

dans l'approche loglinéaire, tandis que dans l'approche classique le modèle multiplicatif équivalent peut s'écrire comme suit:

$$H_0: p_{rst} * p_{ijt} / p_{ist} * p_{rjt} = p_{rsk} * p_{ijk} / p_{isk} * p_{rjk}$$

$$i=1,\ldots,r-1 \;;\; j=1,\ldots,s-1 \;;\; k=1,\ldots,t-1$$

L'hypothèse d'indépendence mutuelle et le modèle loglinéaire associé peuvent à son tour s'écrire:

$$H_o : u_{12} = u_{13} = u_{23} = u_{123} = 0$$

$$\log f_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)}$$
$$i=1,\ldots,r \; ; \; j=1,\ldots,s \; ; \; k=1,\ldots,t$$

correspondant au modèle multiplicatif classique:

$$H_o : p_{ijk} = p_{i..} * p_{.j.} * p_{..k}$$

$$i=1,\ldots,r \; ; \; j=1,\ldots,s \; ; \; k=1,\ldots,t$$

D'autres types d'indépendance à étudier dans les tableaux tridimensionnels se situant entre les deux hypothèses précédentes, sont les indépendances partielles et conditionnelles, dont on ne s'occupera pas dans le texte présent. Toutes ces hypothèses sont associées à des modèles hierarchiques, soit des modèles où la présence d'un effet entraine la présence de tous les effets d'ordre inférieure.

La généralisation des modèles loglinéaires tridimensionnels à ceux multidimensionnels apparait généralement beaucoup plus simple et intuitive que celle des hypothèses correspondantes dans l'approche classique aux tableaux de contingence.

Dans cette étude nous n'avons utilisé que des modèles bi et tridimensionnels.

4. DESCRIPTION ET ANALYSE DES DONNEES

4.1. INTRODUCTION

L'echantillon total, de dimension 1695, est composé par deux groupes à savoir: le groupe A comprenant 968 individus qui ne présentaient aucune évidence clinique de la maladie et le groupe B constitué par 727 personnes qui étaient soignés à l' hopital, soit les individus malades. Les uns et les autres ont été recueillis dans six provinces de l'Angola à savoir: Zaire (ZAI: 13), Lunda-Nord (LUN: 749), Luanda (LUA: 556), Huambo (HUA: 154), Kuando-Kubango (KUA: 49) et Namibe (NAM: 119).

Le groupe A est formé par les sous-groupes suivants: Donneurs de sang (112), Migrants (250), Femmes enceintes (51), Militaires (105) et Autres individus urbains sains (420).

Le groupe B comprend les sous-groupes: individus portant des maladies transmises sexuellement (204), Malades internés à l'hopital (151), Malades non internés (251) et Tuberculeux (121).

Dans chaque cas on dispose aussi de l'information sur la variable sexe.

Nous avons tout d'abord étudié les deux groupes A et B séparément. Pour chaque analyse l'ensemble d'éléments à classifier a été obtenu à partir de la sous-division des provinces par sous-groupes d'individus (ceux décrits précédemment), auxquels on a ajouté les deux sous-totaux réunissant respectivement les provinces se situant à la frontière et celles de l'intérieur.

En tenant compte des résultats issus des analyses classificatoires aussi bien que des analyses loglinéaires des groupes A et B nous avons aussi fait une analyse conjointe de A et B, soit du groupe G.

Dans les paragraphes suivants nous présentons un résumé des résultats obtenus sur le groupe A et nous faisons une allusion brève à l'analyse du groupe G, où il est question de l'application des modèles loglinéaires tridimensionnels.

4.2. ANALYSE DU GROUPE A

4.2.1. CLASSIFICATION HYERARCHIQUE

Ce sont les hiérarchies de classifications basées respectivement, sur le coefficient d'affinité simple et le critère d'aggregation du lien simple, parmi les modèles classiques, et sur le coefficient de validité de l'affinité et l'algorithme de validité du lien / Bacelar (AVLB), parmi les modèles probabilistes, celles qui ont aboutit aux meilleurs résultats.

Dans l'annexe 1 on donne la représentation graphique de l'arbre correspondant à la hiérarchie des classifications associée au modèle défini par la pair (Affinité, Lien Unique). Le modèle probabiliste (Validité du Lien, AVLB) a produit dans ce cas une structure classificatoire très semblable. La stabilité des résultats est aussi confirmé par d'autres modèles, bien ajustés aux données, d'où ressort la même structure générale.

En annexe 2 on trouve un ensemble d' "aides à l'interprétation" à savoir: des statistiques caractéristiques de la distribution des similarités / affinités; la représentation polonnaise de l'arbre (Lerman, 1970); les valeurs de l'affinité et du coefficient de stratification (Bacelar-Nicolau, 1972) responsables par l'aggrégation entre classes, à chaque niveau de l'arbre; les valeurs des statistiques des niveaux.

Une analyse globale de l'arbre associé à la hiérarchie rend compte qu'il se développe en deux grandes classes. Au niveau 9 sont formés les noyaux des classes, le premier étant à son tour le résultat de la réunion de deux petites classes très bien définies. Les grandes classes s'achèvent respectivement aux niveaux 11 et 12 et vont se réunir au niveau 14, le plus significatif de l'arbre. Tous ces niveaux sont associés à des maxima d'une ou plusieurs statistiques des niveaux. Du 14ème au dernier niveau de l'arbre on observe l'effet de chaine généralement

présent pour le critère d'aggrégation du lien simple et aussi présent à l'intérieur de chaque classe.

En lisant maintenant l'arbre du haut vers le bas, on trouve les classes suivantes et leur interprétation par rapport à l'information contenue dans le tableau des données:

Classe I = {1,6,18} : classe des individus séronégatifs

Classe II = {5,22,13} : il s'agit des sous-groupes d'individus où les fréquences du HIV1 et du HIV2 sont à peu près semblables, avec une légère tendance à avoir des valeurs supérieures de HIV2.

Aux deux classes précédentes vont s'ajouter les éléments 14 et 19, présentant des valeurs de HIV2 plus grandes ou égales à celles de HIV1 et qui n'ont pas de HIV1+2.

Classe III = {3,7,11,19,20,12,17} : ce sont des sous-groupes qui n'ont pas de HIV1+2, à l'exception du 7; ils présentent des valeurs de HIV1 et HIV2 généralement élevées, avec HIV2 nettement supérieure à HIV1 dans le noyau et une tendance inverse au fur et à mesure qu'on s'éloigne vers la périphérie de la classe.

Tous les sous-groupes en dessous de la classe III sont caractérisés par la présence du HIV2, absence du HIV1+2 et faible dimension (inférieure ou égale à 9). La classe {2, 16}, en particulier, ne présente que le virus HIV2.

Du point de vue de la séroprévalence globale, on remarque que les deux sexes ne semblent pas être significativement différents. En effet, à l'exception des cas particuliers (cf. tableau) de la région Intérieure et des Donneurs de Sang (à LUN et à LUA), les sexes masculin (M) et féminin (F) sont toujours dans les mêmes classes.

Par contre, les sous-groupes des Migrants (à LUA, KUA et HUA) et des Urbains sains (à LUN et LUA) se trouvent bien séparés ainsi que les Donneurs de sang, en classes distinctes, ce qui suggère des différences significatives par rapport à la séroprévalence globale. Pareillement les régions dans la frontière (FRN) et celles à l'intérieur (INT) sont bien séparés entre elles.

En ce qui concerne la comparaison entre le HIV1 et le HIV2, c'est dans les régions et les sous-groupes de la Classe III, que l'on cherchera tout d'abord à tester les différences significatives. On constate par rapport à la Région, que la province de Lunda-Nord est celle la plus présente dans cette classe; seul le sous-groupe 4 des Donneurs de sang féminins sort de la classe, pour être aggrégé, juste au dernier niveau de l'arbre, à la grande classe formée par tous les autres sousgroupes (il est composé d'une seule femme qui ne présente que le HIV2). D'autre part ce sont les sous-groupes des Donneurs de sang, des Migrants, des Urbains sains, et des Femmes Enceintes, qu'il convient d'étudier et de comparer dans cette classe.

A l'exception de Lunda-Norte, concentrée dans la Classe III de l'arbre, toutes les autres provinces sont dispersées dans les différentes classes de la hiérarchie. Luanda, par exemple, est présente dans la première classe des séronégatifs (Donneurs de sang et Migrants masculins), dans la seconde (Militaires et Urbains sains masculins et féminins), où les valeurs du HIV1 et du HIV2 sont peu importants et dans la petite classe (2,16) (Donneurs de sang féminins), dont les éléments n'ont que le HIV2.

4.2.2. TABLEAUX DE CONTINGENCE

Pour tester la signification des aspects les plus pertinents que l'analyse classificatoire du groupe A a mis en évidence, on a étudié des tableaux de contingence, en utilisant les approches classique et loglinéaire. On en donne quelques exemples dans la suite.

Les variables et leurs modalités respectives considérées sont les suivantes: la Séroprévalence (HIV1, HIV2, HIV1+2, séronégatifs); les Groupes (A, B); les Sous-groupes de A; les Sous-groupes de B; les Provinces (Luanda, Lunda-Nord, Huambo); les grandes Régions (intérieur, frontière).

L'analyse des tableaux de contingence bidimensionnels montre que la séroprévalence se trouve significativement associée à toutes les autres variables.

En ce qui concerne les sous-groupes de A, cet association est expliquée par les individus Migrants et les Urbains sains. Les effets les plus significatifs sont les interactions de 1er ordre (Migrants, HIV1), (Migrants, Séronégatifs) et (Urbains sains, Séronégatifs), dont les valeurs réduites des paramètres du modèle bidimensionnel ajusté (modèle saturé) sont estimés respectivement par: 2.11929, -5.78112 et 3.66775. Ces valeurs sont largement supérieures aux limites asymptotiques correspondants à 95% de confiance. Nous pouvons donc affirmer qu'il existe une association significative positive entre les Migrants et le HIV1 et entre les Urbains sains et les Séronégatifs, tandis qu'une association négative très forte est verifiée entre les Migrants et les Séronégatifs.

En analysant les tableaux partiels où seules les modalités HIV1 et HIV2 de la Séroprévalence sont prises en considération, on obtient l'indépendence entre les variables, ce qui revient à dire que les sous-groupes étudiés de A ne diffèrent pas significativement par rapport à la distribution du HIV1 et du HIV2.

Aussi on demontre que la séroprévalence globale chez les Migrants est significativement plus élevée dans la province de Huambo que dans celle de Kuando-Kubango, alors que Lunda ne présente pas de Migrants infectés. D'autre part la séparation entre les régions de l'intérieur et celles à la frontière, suggèrée par les résultats de l'analyse classificatoire, s'est révélée significative d'après l'analyse du tableau de contingence associé.

Par contre, les différences observées entre les HIV1 et HIV2 dans les sous-groupes appartennant à la Classe III de la hiérarchie de classifications ne sont pas significatives.

D'autres aspects particuliers des données ont également été traités à l'aide de l'analyse des tableaux de contingence, mais on arrête ici le rapport des expériences effectuées.

4.3. ANALYSE COMPARATIVE ENTRE LES GROUPES A ET B - GROUPE G

L'analyse classificatoire conjointe des groupes A et B - groupe G, ainsi que l'analyse du groupe B, suit le même schéma général que l'étude du groupe A présentée dans la section précédente. Dans ce travail nous ne retiendrons de ces deux analyses que quelques références concernant les résultats obtenues par l'étude du groupe G.

Le modèle classificatoire qui a produit le meilleur ajustement aux données a été celui basé sur le coefficient d'affinité asymptotiquement centré et réduit selon la méthode-δ et le critère d'aggrégation AVLB.

L'interprétation des résultats obtenus par l'analyse classificatoire du groupe G, a suggéré d'entreprendre des analyses particulières de certains tableaux de contingence, de même qu'il a été fait dans le cas des analyses séparées des deux groupes A et B. Il s'agit maintenant d'étudier des tableaux bi et tridimensionnels.

En ce qui concerne le traitement des tableaux tridimensionnels, le modèle le mieux ajusté aux données a été celui de non existence d'interaction du second ordre. En effet, le non reject de ce modèle est suivi par le reject des trois hypothèses possibles d'indépendence conditionnelle.

On peut observer dans l'annexe 3 une partie des résultats fournis par le programme TRIDIM appliqué au tableau tridimensionnel (Séroprévalence, Provinces, Groupes), de dimension 4*3*2.

L'analyse de la matrice des résidus, ainsi que celle des valeurs estimées des paramètres du modèle et de leurs valeurs réduites, permet de savoir quels sont les associations positives ou négatives les plus significatives. En utilisant maintenant la notation employée dans les sorties du programme, dont la correspondance avec celle proposée dans la section 3 devient très claire, on remarque, par exemple, qu'il y a une association positive significative entre les individus séronégatifs et la province de Luanda ($u12(4,1) = 0.58523$; $u^*12(4,1) = 2.51668$), à l'inverse de l'association négative significative entre les séronégatifs et le Huambo ($u12(4,3) = -0.41930$; $u^*12(4,3) = -2.57917$). Par ailleurs, sont aussi significatives les associations positive entre séronégatifs et le groupe A ($u13(4,1) = 0.29160$; $u^*13(4,1) = 2.22232$) et négative entre séronégatifs et le groupe B ($u13(4,2) = -0.29160$; $u^*13(4,2) = -2.22232$). D'au-

tres effets aussi bien que leur signification peuvent être trouvés de façon analogue.

3.5. CONCLUSIONS

Nous avons présenté une étude appliquée où des méthodes robustes d'analyse des données multivariées s'articulent et sont complètés par des méthodes d'analyse inférentielle, plus restrictives en ce qui concerne les contraintes d'utilisation. Cette procédure s'est avérée fort productive, les méthodes d'analyse exploratoire fournissant alors le support à l'application des méthodes d'analyse confirmatoire. Dans le cas présent il s'agit d'une étude des tableaux de fréquence par des modèles de classifition hiérarchique ascendante, en particulier des modèles résultants d'un approche probabiliste, basés sur le coefficient d'affinité. D'après les résultats obtenus, nous procédons à la sélection des tableaux de contingence bi et tridimensionnels à analyser par des modèles classiques et loglinéaires.

REFERENCES

BACELAR-NICOLAU, H. (1972): Analyse d'un Algorithme de Classification Automatique Rapport MSH: Maison des Sciences de l'Homme, Paris.

BACELAR-NICOLAU, H. (1980): Contribuições ao Estudo dos Coeficientes de Comparação em Análise Classificatória. Thèse de Doctorat. Université de Lisbonne.

BACELAR-NICOLAU, H. et al (1980): Programa CLASSIF de Classificação Hierárquica Ascendente. Notas e Comunicações CEAUL/INIC.

BACELAR-NICOLAU, H. (1985): The Affinity Coefficient in Cluster Analysis. Meth. Oper. Res., vol. 53, Verlag Anton Hain, p. 507-512.

BACELAR-NICOLAU, H. (1988): Two Probalistic Models for Classification of Variables in Frequency Tables. in: Classification and Related Methods of Data Analysis, H. H. Bock (ed.). North Holland, p. 181-186.

BACELAR-NICOLAU, H. (1990): Classifying Integer Scale Data by the Affinity Coefficient. Meth. Oper. Res., vol. 60, Verlag Anton Hain, p. 587-595.

BACELAR-NICOLAU, H.; MENDES-LEAL, M. (1990): Análise Multivariada de Dados Discretos. Uma Aplicação à Distribuição Regional do Virus da SIDA: Actas da I Conferência em Estatistica e Optimização, Tróia, p. 405-417.

BENZECRI, J. P. et al (1973): L'Analyse des Données, Tome 2: L'Analyse des Correspondances: Dunod, Paris.

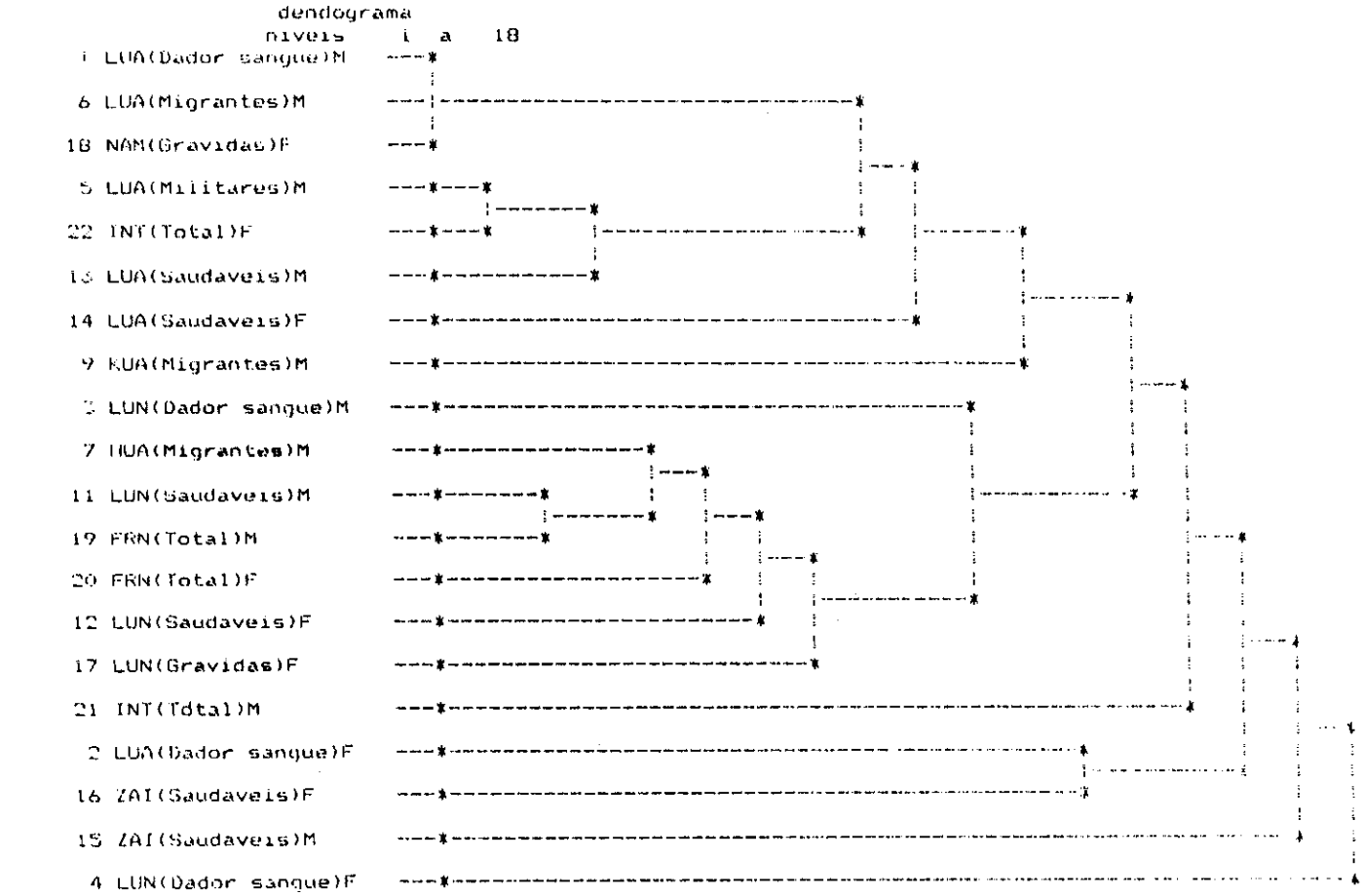
BISHOP, Y. M. M.; Fienberg, S. E.; Holland, F. W. (1975): Discrete Multivariate Analysis, Massachussets Institute of Technology Press.

EVERITT, B. S. (1980): The Analysis of Contingency Tables, Chapman and Hall, London.

FRASER, D. A. S. (1975): Non Parametric Methods in Statistics. Chapman and Hall, p. 235-239.

LERMAN, I.C. (1970): Sur l'Analyse des Données Préalable à une Classification Automatique. Rev.Math. et Sc.Hum., vol.32, 8.ème année, p.5-15.
LERMAN, I.C. (1981):Classification et Analyse Ordinale des Données. Dunod, Paris.
MATUSITA, K. (1955): Decision Rules Based on Distance for Problems of Fit, Two samples and Estimation. Ann.Math.Stat., vol.26, 4, p.631-640.
MATUSITA, K. (1967): On the Notion of Affinity of Several Distributions and Some of its Applications. Ann.Math.Stat., vol.19, 2, p.181-192.
MENDES LEAL, M. (1986): Análise de Tabelas de Contingência. Thèse de 3ème Cycle, Faculté des Sciences de l'Université de Lisbonne.
NICOLAU, F.C. (1973): Programme CORRESPO, in Benzecri, J.P. et Col., 1973 L'Analyse des Données, Tome 2: L'Analyse des Correspondances: Dunod, Paris.
NICOLAU, F.C. (1980): Critérios de Análise Classificatória Hierárquica Baseados na Função de Distribuição. Thèse de doctorat, Université de Lisbonne.
NICOLAU, F.C. (1983): Cluster Analysis and Distribution Function. Meth.Oper.Res., vol.45, Verlag Anton Hain, p.431-433.
NICOLAU, F.C. (1985): Analysis of a Non-Hierarchic Clustering Method Based on VL-Similarity. Meth.Oper.Res., vol.53, Verlag Anton Hain, p.603-610.
NICOLAU, F.C.; BRITO, M.P. (1989): Improvements in NHMEAN Method, in Data Analysis, Learning Symbolic and Numeric Knowledge, E.Diday (ed.): New Science Publ., p.109-115.
SANTOS-FERREIRA, M.O. et al (1980): A Study of Seroprevalence of HIV1 and HIV2 in Six Provinces of the People's Republic of Angola: Clues to the Spread of HIV Infection. Journal of Acquired Immune Deficiency Syndromes, 3, Raven Press, New York, p.780-786.
TIAGO DE OLIVEIRA, J. (1982): The δ-Method for Obtention of Asymptotic Distributions: Applications. Publ.Inst.Stat.Univ.Paris, vol.XXVII, p.49-70.

** Análise Classificatoria - Coeficiente de afinidade - Sida/Grupo A, Angola **
H. Bacelar Nicolau, Ulília Dias, Ana Cristina, Catarina Isabel, David Diogo



Parametros da distribuicao das semelhancas:

media = 1.40942200

variance = .33282160 sigma = .57716810

skewness= -2.50407900

kurtosis= 5.23855300

representacao polaca da arvore

-18 -17 -16 -15 -14 -12 -10 -9 -6 -1 1 6 16 -4 -2 5 20 11 12 8
-11 3 -8 -7 -5 -3 7 -3 9 17 18 10 15 19 -13 2 14 13 4

indice de semelhanca

niv	semelhanca
1	.1732E+01
2	.1731E+01
3	.1731E+01
4	.1731E+01
5	.1731E+01
6	.1730E+01
7	.1730E+01
8	.1730E+01
9	.1729E+01
10	.1727E+01
11	.1725E+01
12	.1719E+01
13	.1708E+01
14	.1707E+01
15	.1688E+01
16	.1671E+01
17	.1612E+01
18	.0000E+00

indice de estratificacao

niv	strat
1	.4160E-01
2	.4168E-01
3	.4169E-01
4	.4177E-01
5	.4178E-01
6	.4180E-01
7	.4183E-01
8	.4185E-01
9	.4187E-01
10	.4209E-01
11	.4248E-01
12	.4281E-01
13	.4363E-01
14	.4792E-01
15	.4567E-01
16	.5147E-01
17	.5287E-01
18	.1000E+01

estatistica local

niv	stat
1	2.9684
2	1.7228
3	1.7228
4	2.4164
5	2.0489
6	2.0557
7	2.9050
8	2.9772
9	4.8730
10	4.1494
11	3.9870
12	4.2953
13	1.7198
14	9.0968
15	3.4045
16	5.4869
17	5.1962

estatistica global

niv	stat	dif
1	2.9684	.0000
2	3.4185	.4500
3	3.8117	.3932
4	4.4786	.6669
5	5.0210	.5425
6	5.6132	.5922
7	6.2279	.6147
8	6.7803	.5524
9	8.2777	1.4974
10	9.1969	.9192
11	9.7602	.5633
12	10.3267	.5665
13	10.4107	.0839
14	11.0292	.7185
15	10.6724	-.3568
16	9.7812	-1.8912
17	7.9436	-1.8376

valores de stat por ordem decrescente

nos	stat
14	11.0292
15	10.6724
13	10.4107
12	10.3267
11	9.7602
10	9.1969
16	9.7812
9	8.2777
17	7.1436
8	6.7803
7	6.2279
6	5.6132
5	5.0210
4	4.4786
3	3.8117
2	3.4185
1	2.9684

ANNEXE 3

********** Matriz dos Residuos **********

-0.6174	0.4825	0.0937	2.4128	-0.2117	-0.1158
-0.6107	0.4060	0.0753	3.1840	-0.2378	-0.1239
-0.0064	-0.7237	0.6605	0.0167	0.2186	-0.5539
0.1751	-0.1928	-0.1309	-0.9793	0.1210	0.2302

*** Modelo: vijk = u + u1 + u2 + u3 + u12 + u13 + u23 ***

*** parametro	** estimador	** d.padrao	** v.reduzidos ***
u12(1 , 1)	-0.33420	0.38939	-0.85827
u12(1 , 2)	0.17731	0.23383	0.75829
u12(1 , 3)	0.15689	0.25090	0.62529
u12(2 , 1)	-0.56777	0.43521	-1.30458
u12(2 , 2)	0.27590	0.24841	1.11068
u12(2 , 3)	0.29187	0.26408	1.10521
u12(3 , 1)	0.31674	0.45308	0.69908
u12(3 , 2)	-0.28728	0.33630	-0.85426
u12(3 , 3)	-0.02945	0.35310	-0.08341
u12(4 , 1)	0.58523	0.23254	2.51668
u12(4 , 2)	-0.16593	0.15099	-1.09892
u12(4 , 3)	-0.41930	0.16257	-2.57917
u13(1 , 1)	-0.06670	0.21184	-0.31487
u13(1 , 2)	0.06670	0.21184	0.31487
u13(2 , 1)	0.22062	0.23125	0.95405
u13(2 , 2)	-0.22062	0.23125	-0.95405
u13(3 , 1)	-0.44552	0.27174	-1.63950
u13(3 , 2)	0.44552	0.27174	1.63950
u13(4 , 1)	0.29160	0.13121	2.22232
u13(4 , 2)	-0.29160	0.13121	-2.22232
u23(1 , 1)	1.11558	0.22368	4.98742
u23(1 , 2)	-1.11558	0.22368	-4.98742
u23(2 , 1)	-1.06848	0.14499	-7.36913
u23(2 , 2)	1.06848	0.14499	7.36913
u23(3 , 1)	-0.04711	0.15379	-0.30631
u23(3 , 2)	0.04711	0.15379	0.30631
u1(1)	-0.40779	0.21184	-1.92500
u1(2)	-0.25230	0.23125	-1.09103
u1(3)	-1.66357	0.27174	-6.12191
u1(4)	2.32366	0.13121	17.70888
u2(1)	-0.85875	0.22368	-3.83922
u2(2)	0.96330	0.14499	6.64375
u2(3)	-0.10455	0.15379	-0.67982
u3(1)	0.31643	0.12564	2.51861
u3(2)	-0.31643	0.12564	-2.51861
u	2.47530		