

H. BACELAR-NICOLAU

F. C. NICOLAU

**Analyse classificatoire de l'identité nationale chez les portugais.
Une étude exploratoire fondée sur le coefficient d'affinité**

Mathématiques et sciences humaines, tome 109 (1990), p. 55-64

http://www.numdam.org/item?id=MSH_1990__109__55_0

© Centre d'analyse et de mathématiques sociales de l'EHESS, 1990, tous droits réservés.

L'accès aux archives de la revue « Mathématiques et sciences humaines » (<http://msh.revues.org/>) 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 CLASSIFICATOIRE DE L'IDENTITÉ NATIONALE
CHEZ LES PORTUGAIS.
UNE ÉTUDE EXPLORATOIRE FONDÉE SUR
LE COEFFICIENT D'AFFINITÉ

H. BACELAR-NICOLAU¹ , F. C. NICOLAU²

RÉSUMÉ - On cherche des typologies d'un ensemble de pays à partir des notes moyennes de défiance qui leurs ont été accordées par six régions différentes du Portugal, d'après un sondage mené en 1988 sur six groupes de citoyens portugais résidant dans ces régions. Trois méthodes de classification hiérarchique ascendante ont été utilisées, toutes basées sur le coefficient d'affinité et associées à des hypothèses de référence de plus en plus spécifiques. On observe que les résultats des analyses deviennent progressivement meilleurs au sens de l'ajustement aux structures des similarités de base, évalué par une statistique appropriée.

ABSTRACT - *Classifying analysis of national identity among Portugueses.*

In this paper we search for typologies on a set of sixteen countries from the knowledge of distrust mean scores given by people of six different regions of Portugal. Three hierarchical clustering procedures based on the notion of affinity coefficient were used. These procedures are supported by some basic probabilistic hypothesis of more and more specific constraints. The clustering results progressively become better in the sense of well fitting the initial structures of data, as evaluated by an appropriate statistic.

1. INTRODUCTION

Dans ce travail nous étudions un ensemble de données psychosociologiques recueillies par sondage réalisé en 1988 dans six régions portugaises différentes (Oliveira and Marques, 1988). Le but du sondage était de situer le Portugal et d'autres pays les uns par rapport aux autres d'après les attitudes de défiance exprimées par les citoyens portugais de ces régions envers eux. Nous voulons chercher des typologies des pays associées à des modèles de classification hiérarchique ascendante (C.H.A.).

Dans chaque région le même effectif d'individus a été pris pour l'échantillon. Chaque élément a donné une note de défiance entre 0 et 5 à chacun des pays P_j , $j=1,2,\dots,16$.

¹ Lab. Statistique et Analyse des Données (LEAD), Fac. Psychologie et Sciences de l'Éducation de l'Université de Lisbonne, 17 Rua Pinheiro Chagas, 1200 Lisboa, Portugal .

² Lab. Analyse des Données, Statistique et Reconnaissance des Formes (LADERF)/Mathématique, Université de Aveiro, Campus de Santiago, 3800 Aveiro, Portugal .

Le psychologue, ne voulant pas retenir de chaque région que le comportement global, a calculé l'individu moyen par région; et le tableau de données auquel on a eu accès finalement a été celui de dimensions 6×16 où la région R_k , $k=1,2,\dots,6$, est représentée par son individu moyen. Le traitement de ce tableau est justifié du fait que les observations sont peu dispersées autour de leur moyenne régionale.

La première étape de notre étude comprend l'analyse classificatoire du tableau $R \times P$, où $R=\{R_k\}$ et $P=\{P_j\}$, basée sur le coefficient d'affinité (Matusita, 1955, 1967 ; Bacelar-Nicolau, 1980, 1985, 1988).

Dans la deuxième étape du travail nous utilisons le coefficient d'affinité centré et réduit sous des hypothèses de référence appropriées. Le centrage et la réduction sont asymptotiques par rapport au nombre d'unités de données ce qui pourrait ne pas être justifié pour un tableau de six lignes. On montre que l'analyse de ce tableau est équivalente à celle d'un tableau fictif de $6N \times 16$ où tous les N individus dans la même région sont égaux à l'individu moyen. Ceci justifie parfaitement l'utilisation des méthodes asymptotiques lorsque N est suffisamment grand. Nous avons là en réalité une conséquence du fait que l'affinité est un des coefficients vérifiant le principe d'équivalence distributionnelle énoncé par Benzécri (1973, p.31) dans le cas de la distance du χ^2 .

L'affinité réduite est en elle-même un coefficient de similarité, lequel donne fréquemment des résultats mieux adaptés aux données que l'affinité de base quand on utilise les mêmes algorithmes d'agrégation.

Dans la troisième étape de notre étude nous introduisons le coefficient de validité de l'affinité, un coefficient probabiliste auquel on associe généralement des critères d'agrégation appartenant à la famille d'Algorithmes de Validité du Lien (AVL) développée par Nicolau (1980,1983) à partir de l'Algorithme de la Vraisemblance du Lien (Lerman, 1970, 1981 ; Bacelar-Nicolau, 1972a, 1972b, 1980).

Le paragraphe suivant contient la description des trois coefficients d'affinité, tout d'abord celui de base, ensuite l'affinité centrée et réduite et enfin la validité de l'affinité.

Au troisième paragraphe nous présentons les résultats des trois classifications hiérarchiques obtenues avec ces coefficients, en utilisant les critères d'agrégation qui ont donné les meilleurs résultats, au sens de l'ajustement aux structures des similarités : celui du lien maximal (ultramétrique inférieure maximale) dans les deux premiers cas et un critère de la famille AVL, l'Algorithme de la Validité Moyenne (AVM) dans le cas du coefficient probabiliste.

Le dernier paragraphe fait le point des résultats principaux de ce travail, en mettant l'accent sur la structure de l'identité nationale portugaise et sur la méthodologie de classification utilisée.

2. LES COEFFICIENTS D'AFFINITÉ

2.1. Coefficient d'affinité simple

Soit U_1, U_2, \dots, U_n un ensemble de n individus ou, plus généralement unités de données: sujets, sous-ensembles, sous-échantillons, etc. et X_1, X_2, \dots, X_r , r variables ayant des valeurs positives : scores, mesures, etc... Nous pouvons alors construire un tableau de données dont le terme général x_{ij} , $i=1,2,\dots, n$, $j=1,2,\dots,r$, représente la valeur de la j -ème variable sur l' i -ème

unité de données . La note totale de X_j relative à l'ensemble des n unités est égale à $x_j = \sum\{x_{ij} \mid i=1,2,\dots,n\}$.

Associons maintenant à chaque X_j le vecteur de proportions définies par $x_{i/j} = x_{ij} / x_j$, $i=1,2,\dots,n$, soit le profil des contributions des unités de données à la j -ème variable. La ressemblance de chaque paire (j,j') sera déterminée par le coefficient d'affinité $a_{jj'}$ entre les vecteurs de proportions associés à j et j' , soit le produit des racines carrés des profils :

$$a_{jj'} = \sum_{i=1}^n \sqrt{\frac{x_{ij}}{x_{.j}} \frac{x_{ij'}}{x_{.j'}}} = \sum_{i=1}^n \sqrt{x_{i/j} x_{i/j'}}$$

On remarquera que $a_{jj'}$ est un coefficient de dépendance monotone entre j et j' , et aussi un coefficient de similarité en analyse classificatoire car :

- (i) il prend des valeurs réelles (positives)
- (ii) il est symétrique : $a_{jj'} = a_{j'j}$, $j \neq j'$
- (iii) et $a_{jj} \geq a_{jj'}$, $j \neq j'$

Le maximum de $a_{jj'}$ est égal à 1, et est atteint pour $j=j'$; il correspond donc à deux vecteurs identiques des proportions. Le minimum de $a_{jj'}$ est égal à 0 et intervient lorsque les vecteurs de proportions sont orthogonaux (au sens de la métrique identité usuelle).

2.2 Coefficient d'affinité réduit

Supposons que tous les $x_{i/j}$ sont de même ordre de grandeur n^{-1} lorsque n va vers l'infini. La condition W de Wald et Wolfowitz (Fraser, 1957, p.235-239) est vérifiée par les suites de nombres réels :

$$E_n = (\sqrt{x_{i/j}} \quad , \quad i=1, 2, \dots, n)$$

et

$$E'_n = (\sqrt{x_{i/j'}} \quad , \quad i=1, 2, \dots, n)$$

On montre en effet (Bacelar-Nicolau 1985, 1988), qu'à des cas triviaux près en classification, la suite E_n satisfait la condition générale: $m_r(j) / m_2(j)^{r/2} = O(1)$ où $m_r(j)$ est le moment centré d'ordre r et $O(1)$ représente une limite finie ; et de même pour la suite E'_n .

Associons à E_n (E'_n) une variable aléatoire (v.a.) Y_n (Y'_n) définie sur l'ensemble des permutations d'indices muni de la loi uniforme, où chaque suite permutée de E_n (E'_n) a la même probabilité égale à $1/n!$. Le coefficient d'affinité $a_{jj'}$ est une réalisation soit de la (v.a.) $A_{jj'} = \langle Y_n, E'_n \rangle$ soit de la (v.a.) $B_{jj'} = \langle E_n, Y'_n \rangle$; ces deux variables ont la même loi de probabilité et en particulier les valeurs moyennes et les variances sont les mêmes. D'après un théorème limite de Wald et Wolfowitz (Fraser, 1956) on démontre alors que :

$$A^*_{jj'} = [A_{jj'} - E(A_{jj'})] / s(A_{jj'})$$

est asymptotiquement normale centrée réduite, $E(A_{jj'})$ et $s^2(A_{jj'})$ étant donnés par :

$$E(A_{jj'}) = n \left(\sum_{i=1}^n \frac{\sqrt{x_{ij}}}{n} \right) \left(\sum_{i=1}^n \frac{\sqrt{x_{i'j'}}}{n} \right)$$

et

$$\sigma(A_{jj'}) = \frac{1}{n-1} \left[\sum_{i=1}^n \left(\sqrt{x_{ij}} - \sum_{i=1}^n \frac{\sqrt{x_{ij}}}{n} \right)^2 \right] \left[\sum_{i=1}^n \left(\sqrt{x_{i'j'}} - \sum_{i=1}^n \frac{\sqrt{x_{i'j'}}}{n} \right)^2 \right]$$

En considérant le fait que les $(x_{ij})^{1/2}$ sont des coordonnées sphériques, car on a $\sum (x_{ij})^{1/2} = 1$, la réalisation $a_{jj'}$ de $A_{jj'}$ devient, d'après le théorème cité ci-dessus :

$$a_{jj'} = \sqrt{n-1} \frac{\sum_{i=1}^n \sqrt{x_{ij} x_{i'j'}} - \frac{1}{n} \sum_{i=1}^n \sqrt{x_{ij}} \sum_{i=1}^n \sqrt{x_{i'j'}}}{\sqrt{\left[x_{.j} - \frac{1}{n} \left(\sum_{i=1}^n \sqrt{x_{ij}} \right)^2 \right] \left[x_{.j'} - \frac{1}{n} \left(\sum_{i=1}^n \sqrt{x_{i'j'}} \right)^2 \right]}}$$

Ici $a_{jj'}$ est un nouveau coefficient de similarité entre j et j' car :

- (i) il prend des valeurs réelles
- (ii) il est symétrique : $a_{jj'} = a_{j'j}$, $j \neq j'$
- (iii) il prend sa valeur maximum pour $j=j'$: $a_{jj} = (n-1)^{1/2} \geq a_{jj'}$, $j \neq j'$
(coefficient normalisé à la valeur $(n-1)^{1/2}$)

2.3. Coefficient de validité de l'affinité

La (v.a.) $A_{jj'}$ nous amène directement à l'introduction d'un coefficient de similarité probabiliste, le coefficient de validité de l'affinité, à savoir :

$$a_{jj'} = \text{Prob} (A_{jj'} \leq a_{jj'})$$

Le coefficient $a_{jj'}$:

- (i) prend des valeurs réelles dans l'intervalle $[0,1]$
- (ii) est symétrique : $a_{jj'} = a_{j'j}$, $j \neq j'$
- (iii) est normalisé à l'unité.

Plus petite est la probabilité d'avoir des valeurs supérieures à $a_{jj'}$, plus grande est, d'après la définition, la validité $a_{jj'}$ de la similarité $a_{jj'}$.

En rappelant la notion de Vraisemblance du Lien (VL) introduite en Classification par Lerman (1970), qu'on préfère appeler généralement Validité du Lien (Bacelar-Nicolau 1980,1985,1988 ; Nicolau 1980,1983,1985)) où l'idée de base est précisément celle de prendre la fonction de répartition comme coefficient de similarité, on voit tout de suite que $a_{jj'}$ est un coefficient de type VL.

Les résultats asymptotiques du paragraphe précédent justifient le calcul approché de $a_{jj'}$ à travers la fonction de répartition Φ de la loi normale réduite, d'après la relation $a_{jj'} \approx \Phi(a^*_{jj'})$.

L'utilisation du coefficient $a_{jj'}$ rend comparables du point de vue probabiliste les affinités des différents paires de variables. Son extension à des critères d'agrégation entre classes nous amène naturellement aux méthodes de la famille AVL développée par F.Nicolau à partir de l'algorithme de même nom. Dans les méthodes d'agrégation AVL l'hypothèse probabiliste de référence de base consiste à supposer une distribution uniforme sur les valeurs des similarités initiales. Pour une paire (X,Y) de classes, chaque critère d'agrégation dans la famille AVL tiendra toujours compte et des similarités des couples appartenant au produit cartésien $X \times Y$ et de l'hypothèse probabiliste de base.

3. L'IDENTITÉ NATIONALE CHEZ LES PORTUGAIS

3.1. Les données

Dans chacune des six régions choisies R_k ($k=1,2,\dots,6$) du Portugal les $N=60$ individus interrogés devaient accorder une note de défiance entre 0 et 5 à chacun des $r=16$ pays suivants :

1 - Portugal (POR)	9 - Angola (ANG)
2 - Espagne (ESP)	10 - Chine (CHI)
3 - Italie (ITA)	11 - Japon (JAP)
4 - Brésil (BRA)	12 - Mozambique(MOC)
5 - Grande-Bretagne (ING)	13 - URSS (RUS)
6 - Allemagne (RFA)	14 - Etats Unis(USA)
7 - Vénézuéla (VEN)	15 - France (FRA)
8 - S. Tomé (STO)	16 - Cap-Vert (CVE)

Nous avons vu que le tableau des données $R \times P$ analysé décrit en ligne les régions R_k ($k=1,2,\dots,6$), qui sont les unités de données, et en colonne les pays P_j ($j=1,2,\dots,16$), correspondant aux variables à classifier. Chaque région est comme nous l'avons dit auparavant, représentée dans ce tableau par les scores moyens de défiance conférés à chaque pays.

Montrons maintenant que l'affinité entre j et j' calculée sur ce tableau est égale à l'affinité entre j et j' calculée sur le tableau de dimension $n \times 16$, avec $n=6N$. En effet le coefficient $a_{jj'}$ devient alors :

$$a_{jj'} = \sum_{k=1}^6 \frac{\sum_{l=(k-1)N+1}^{kN} \sqrt{\frac{x_{lj} x_{lj'}}{x_{.j} x_{.j'}}}}{6} = \sum_{k=1}^6 \sqrt{\frac{x'_{kj} x_{kj'}}{x'_{.j} x_{.j'}}$$

où x'_{kj} représente le score moyen des N individus dans chaque région R_k , et $x'_{.j} = x_{.j}/N$. On retrouve donc la valeur de l'affinité entre j et j' définie à partir du tableau $R \times P$.

Le remplacement du tableau de dimension $6N \times 16$ par celui de dimension 6×16 dans l'analyse classificatoire peut être comparé à la procédure de groupement des données en statistique descriptive consistant à rassembler les données en intervalles de classe et considérer que toutes les observations appartenant à une même classe sont égales au point moyen de l'intervalle de classe (hypothèse d'homogénéité relative au groupement des données). Dans notre

cas, nous l'avons dit plus haut, cela est justifié par la concentration des observations autour de leur moyenne régionale.

Si les cardinaux n_k des R_k ne sont pas tous égaux, nous avons alors

$$a_{jj'} = \frac{\sum_k n_k \sqrt{\frac{x'_{kj} x'_{kj'}}{\sum_k n_k x'_{kj} \sum_k n_k x'_{kj'}}}}{\sum_k n_k}$$

où $x_{.j} = \sum_k n_k x'_{kj}$, et on peut encore se ramener à analyser le tableau le plus petit en affectant les x'_{kj} des masses n_k .

Comme nous l'avons déjà dit ces résultats justifient dans notre étude l'utilisation des méthodes asymptotiques.

Passons maintenant à l'analyse abrégée des arbres associés aux trois modèles de c.h.a. fondés respectivement sur l'affinité $a_{jj'}$, l'affinité centrée et réduite $a^*_{jj'}$ et la validité $a_{jj'}$. Les critères d'agrégation ont été, dans les deux premiers cas, celui du lien maximal et, dans le troisième cas, celui de la validité moyenne des liens, AVM.

Les suites de valeurs numériques qui se trouvent sous chacun des arbres sont les réalisations d'une "statistique des niveaux" et mesurent le degré d'association entre la partition donnée à chaque niveau de l'arbre et la structure initiale de similarités entre paires de l'ensemble P à classifier. Dans cette étude, il s'agit de la statistique globale STAT. Cette statistique est une fonction du cardinal de l'intersection entre le graphe associé au préordre défini par la partition et le graphe associé au préordre défini par les similarités entre paires d'éléments différents de l'ensemble à classifier (Lerman, 1970 ; Bacelar-Nicolau, 1972a, 1972b).

Les niveaux les plus significatifs de l'arbre, d'après la statistique STAT, sont tout d'abord les niveaux correspondants aux maximums locaux et ensuite ceux identifiés par des accroissements locaux maximaux de la statistique. Dans l'un et l'autre cas, il s'agit de "bonnes partitions". Par contre les derniers niveaux de l'arbre sont généralement associés à des minimums locaux de la statistique ou de la suite d'accroissements et correspondent aux réunions des classes forcées par l'algorithme d'agrégation.

3.2. Coefficient d'affinité. Agrégation par le lien maximal.

Les niveaux les plus significatifs sont, par ordre croissant de la statistique globale STAT, les 4-ème, 9-ème, 11-ème et 13-ème, tel que l'on peut le voir sur la figure 1.

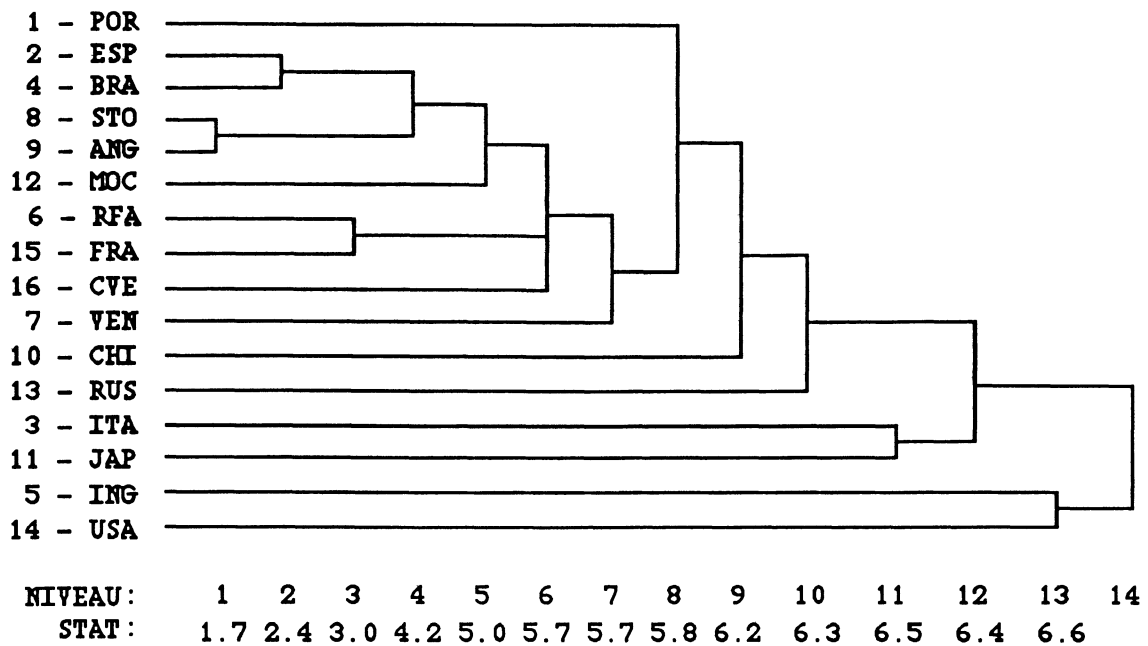


Figure 1. Coefficient d'affinité. Agrégation par le lien maximal.

Au 4-ème niveau, se constitue le groupe (ESP, BRA, STO, ANG) de profils de défiance très similaires. Le 9-ème niveau correspond à une étape dans la formation de la grande classe où sont déjà inclus tous les pays d'expression portugaise ou espagnole et aussi le couple (FRA, RFA) et CHI.

Au 11-ème niveau, on trouve une partition avec la grande classe précédente enrichie par RUS, la paire (ITA, JAP) et les singletons ING et USA.

Enfin le 13-ème niveau, le plus significatif, correspond à la formation de la classe (ING, USA), c'est-à-dire les pays envers lesquels il y a la moindre défiance, tous les autres pays étant déjà réunis dans une même classe.

3.3. Coefficient d'affinité réduit. Agrégation par le lien maximal.

La figure 2 représente l'arbre fourni par le binôme : affinité réduite, lien maximal.

On observe tout de suite qu'il y a un accroissement global des valeurs de la statistique des niveaux vis-à-vis du cas antérieur, mettant en évidence un meilleur ajustement entre les partitions de la c.h.a. et la structure des similarités initiales.

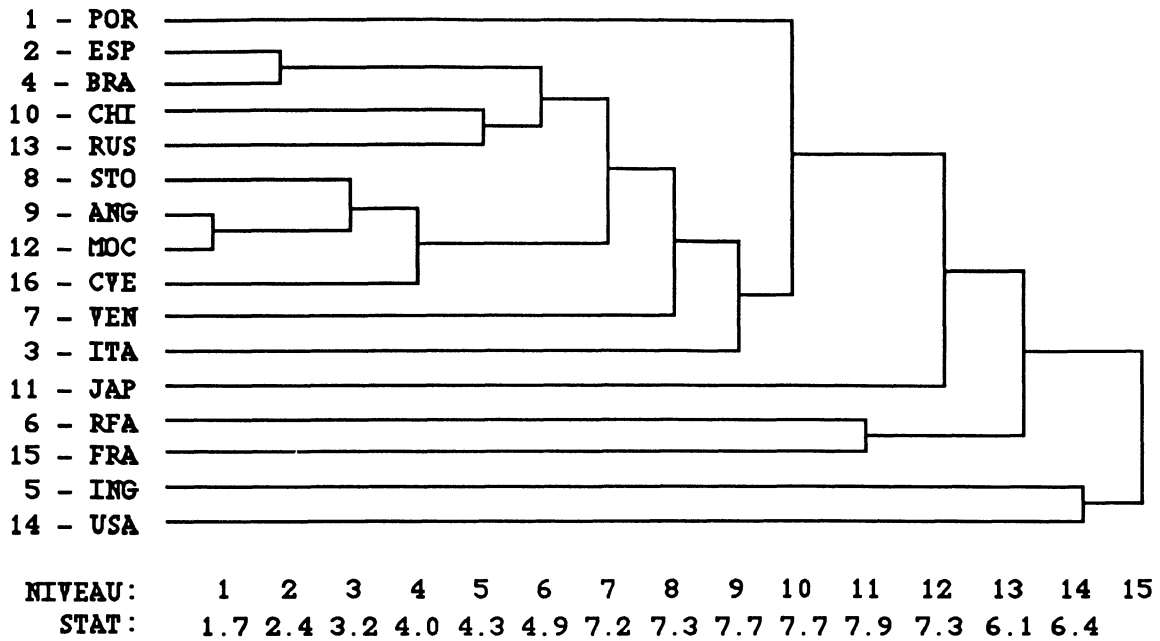


Figure 2. Coefficient d'affinité centré réduit. Agrégation par le lien maximal.

Les niveaux les plus significatifs sont maintenant les 3-ème, 7-ème, 9-ème, 11-ème et 14-ème.

Au 3-ème niveau, est constituée la classe (STO, ANG, MOC) de pays d'expression portugaise.

Au 7-ème niveau, sont réunies les deux classes (ESP, BRA, CHI, RUS) et (STO, ANG, MOC, CVE), fournies aux 6-ème et 4-ème niveaux, respectivement.

Au 9-ème niveau ITA rejointe la classe qui s'était constituée au 7-ème niveau, déjà accrue de VEN. Les autres pays, y inclus le Portugal, sont encore isolés.

Le 11-ème niveau est le plus significatif de cet arbre. La partition trouvée est formé par: la grande classe du 9-ème niveau à laquelle POR est déjà agrégé; le singleton JAP; la paire (RFA, FRA); et ING et USA encore isolés.

Au 14-ème niveau ING et USA sont mis ensemble, et se détachent à nouveau de tous les autres pays.

3.4. Coefficient de validité de l'affinité. Agrégation par la validité moyenne.

Dans la figure 3 on voit nettement que la statistique des niveaux atteint des valeurs globalement supérieures à celles obtenues dans les deux cas précédents.

Les niveaux les plus significatifs sont les 4-ème, 7-ème, 9-ème et 11-ème.

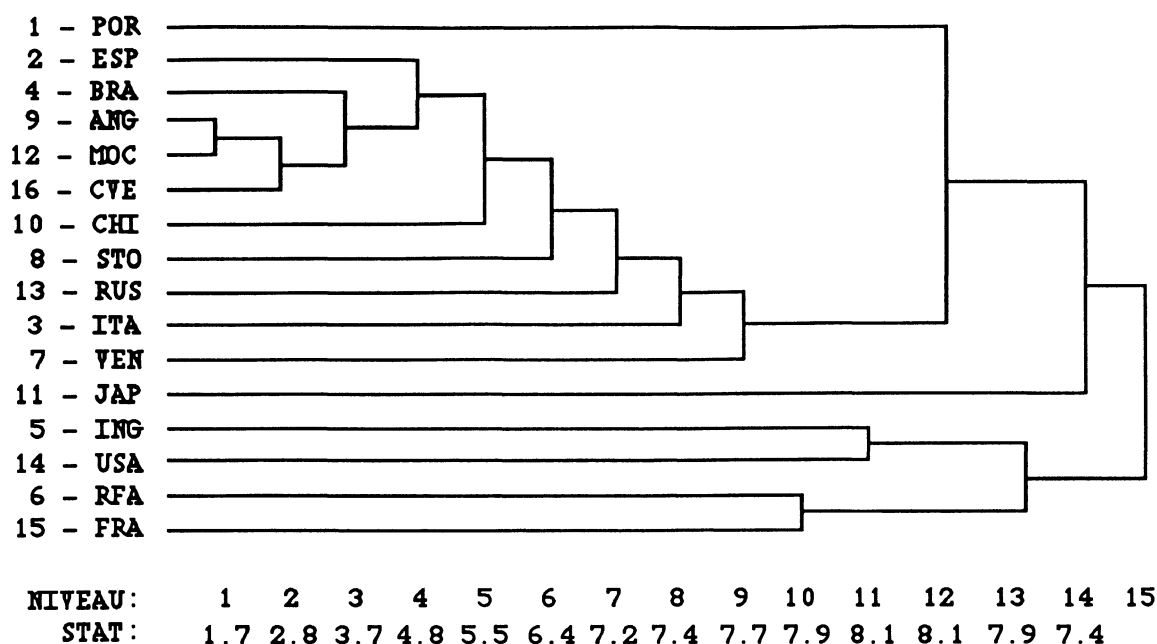


Figure 3. Coefficient de validité de l'affinité. Agrégation par la validité moyenne.

Au 4-ème niveau est formé le groupe le plus cohérent, rassemblant l'Espagne avec des pays d'expression portugaise: (ANG, MOC, CVE, BRA, ESP) ; les autres pays sont encore tous isolés.

Au 7-ème niveau la classe précédente est augmentée avec CHI, STO et RUS.

Le 9-ème niveau est encore en rapport avec la construction de cette même classe, à laquelle se sont maintenant associés ITA et VEN. Le Portugal et les autres pays sont encore isolés.

Le 11-ème niveau présente la partition la plus significative; nous avons: tout d'abord la grande classe achevée au 9-ème niveau; POR et JAP sont encore des singletons et séparent cette classe des couples (ING, USA) et (RFA, FRA).

4. CONCLUSIONS

L'analyse comparative des trois arbres de classification rend compte qu'ils traduisent, avec des nuances, une même structure globale d'attitudes de défiance envers les seize pays.

Le Portugal remplit toujours la position intermédiaire entre l'ensemble des pays envers lesquels est manifesté une plus grande défiance et ceux auxquels est accordé la moindre défiance: la description des arbres présentés plus haut montre très bien qui sont les uns et les autres.

D'autre part, il y a une évolution positive des résultats quand on va de la première à la seconde et puis à la troisième hiérarchie. En effet, en passant du coefficient d'affinité simple à celui centré et réduit et ensuite au coefficient probabiliste, on observe un accroissement global des valeurs de la statistique des niveaux, qui traduit un meilleur ajustement aux structures de similarité de base et qui a son correspondant dans une plus claire et plus satisfaisante interprétation des résultats. Ce fait, déjà constaté dans d'autres exemples préalablement traités,

met en valeur le choix des mesures probabilistes de ressemblance associées à l'affinité et, de façon plus générale, l'approche globale de la Validité du Lien en Classification.

BIBLIOGRAPHIE

BACELAR-NICOLAU, H., 1972a., Algorithme de construction de la représentation polonaise d'un arbre de classifications. *Rapport MSH : M.S.H.*, Paris, 1972.

BACELAR-NICOLAU, H., 1972b, Analyse d'un algorithme de classification automatique. *Rapport MSH :M.S.H.*, Paris, 1972.

BACELAR-NICOLAU, H., 1980, *Contributions à l'Etude des Coefficients de Comparaison en Analyse Classificatoire*. Thèse de Doctorat : Univ. Lisboa (en portugais), 1980.

BACELAR-NICOLAU, H. 1985, The affinity coefficient in cluster analysis. *Meth. Oper. Res.*, vol. 53, 1985, p. 507-512.

BACELAR-NICOLAU, H., 1988, Two probabilistic models for classification of variables in frequency tables. In *Classification and Related Methods of Data Analysis*, H. H. Bock (ed.): North Holland, 1988, p. 181-186.

BENZECRI, J.P. et Col., 1973, *L'Analyse des Données, Tome 2: L'Analyse des Correspondances* : Dunod, Paris, 1973.

FRASER, D. A. S. , 1957. *Non Parametric Methods in Statistics* : Chapman and Hall, 1957, 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, 1970, p. 5-15.

LERMAN, I. C., 1981, *Classification et Analyse Ordinale des Données* : Paris, Dunod, 1981.

MATUSITA, K., 1955, Decision rules based on distance for problems of fit, two samples and estimation. *Ann. Math. Stat.*, vol. 26, 4, 1955, p.631-640.

MATUSITA, K., 1967, On the notion of affinity of several distributions and some of its applications. *Ann. Inst. Stat. Math.*, vol.19, 2, 1967, p. 181-192.

NICOLAU, F. C., 1980, *Critères d'Analyse Classificatoire Hierarchique Basées sur la Fonction de Répartition*. Thèse de Doctorat: Univ. Lisboa (en portugais), 1980.

NICOLAU, F. C., 1983, Cluster analysis and distribution function. *Meth. Oper. Res.*, 45, 1093, p. 431-433.

NICOLAU, F. C., 1985, Analysis of a non-hierarchic clustering method based on VL-similarity. *Meth. Oper. Res.*, 53, 1985, p. 603-610.

OLIVEIRA, J. P. , MARQUES, L. G., 1988, A Identidade Nacional, Uma Investigação Psicossocial, *Rapport Inst. Damião de Gois*, Lisboa, 1988.