

VARIATION LINGUISTIQUE  
DANS L'ESPACE :  
DIALECTOLOGIE ET ONOMASTIQUE

ACTES DU XVIIÈME CONGRÈS INTERNATIONAL  
DE LINGUISTIQUE ET PHILOGIE ROMANES  
(AIX-EN-PROVENCE, 29 AOÛT - 3 SEPTEMBRE 1983)

VOL. N° 6

1986

publications

UNIVERSITE DE PROVENCE  
29, av. Robert-Schuman  
13621 Aix-en-Provence Cedex

diffusion

JEANNE LAFFITTE  
25, crs d'Estienne-d'Orves  
13001 Marseille

L'AUTOMATISATION DES ATLAS LINGUISTIQUES

DENIS PHILPS

#### Remarque préliminaire

La première génération des grands atlas linguistiques européens, de type manuel, remonte au tournant du siècle (a) et s'échelonne jusqu'aux oeuvres les plus récentes : l'Atlas linguistique de l'Angleterre (b), ou les Nouveaux atlas de la France par régions (c).

Il peut sembler étonnant, près d'un siècle plus tard, que les bases théoriques et méthodologiques de ces atlas précurseurs n'aient guère évolué par rapport à celles proposées par G. PARIS en 1888, lorsqu'il opina qu'"il faudrait que chaque commune d'une part, chaque mot de l'autre, eût sa monographie, purement descriptive, faite de première main et tracée avec toute la rigueur d'observation qu'exigent les sciences naturelles" (d). Ces concepts, à savoir l'exhaustivité géolinguistique, la description, l'enquête directe, la cartographie, et l'empirisme scientifique respectivement, caractérisent encore aujourd'hui la plupart des atlas linguistiques manuels, le principe d'exhaustivité étant cependant jugé impraticable ou superflu (e).

L'un des seuls atlas européens de grande envergure à faire évoluer cette méthodologie de façon significative était le sixième volume de l'Atlas Linguistique de la Gascogne du regretté Jean SEGUY (f). Ce tome a donné naissance à une nouvelle discipline, appelée 'dialectométrie' par son fondateur, qui a pour objet l'analyse quantitative de l'espace linguistique. Enfin un grand atlas dialectal comprenait, outre les cartes géolinguistiques désormais classiques, une analyse mathématique de son propre objet, la variation linguistique.

Or, l'inconvénient majeur d'un atlas linguistique national ou régional réside notamment dans son exploitation, car son accès n'est pas toujours aisé pour le chercheur, enseignant ou étudiant, en raison de l'énorme quantité de données éparpillées au travers de chaque ouvrage, et de l'inadéquation des instruments d'analyse traditionnels.

L'apport de l'automatisation dans notre domaine est tel qu'il nous semble justifié de parler d'atlas de deuxième génération, engendrés par leurs prédécesseurs (les NALF en France). En effet, ce sont eux qui fournissent les données et qui bénéficieront les premiers (cf. l'ALG 6 et peut-être bientôt l'ALLOc) des traitements ultrapuissants rendus possibles par les nouveaux logiciels informatiques : classification hiérarchique, analyse des correspondances, cartographie automatisée (g).

Compte tenu de l'étendue de l'espace linguistique mondial restant à explorer et à cartographier - aucun atlas national de l'arabe n'a encore vu le jour par exemple - ainsi que de l'accumulation de connaissances dialectales non traitées, l'utilisation de l'instrument d'analyse le plus puissant qui soit paraît logique.

Nous allons donc tenter de montrer ici qu'une approche quantitative comprenant le traitement, l'analyse et la cartographie automatiques des données dialectales fournit cette puissance analytique, et inaugure une nouvelle génération d'atlas linguistiques.

- (a) Cf. Atlas linguistique de la France, J. GILLIERON et E. EDMONT, Paris, 1902-10.
- (b) Linguistic Atlas of England, H. ORTON, S. SANDERSON, J. WIDDOWSON, Croom Helm, London, 1978.
- (c) E.g. Atlas linguistique du Languedoc occidental, vol.2, X. RAVIER, CNRS, Paris, 1982.
- (d) In Les parlers de France, Revue des patois gallo-romans, II, pp.161f, Paris, 1888.
- (e) Mais cf. Atlas linguistique des Pyrénées orientales, H. GUITER, CNRS, Paris, 1966, entre autre.
- (f) Atlas linguistique de la Gascogne, vol.6, J. SEGUY, CNRS, Paris, 1973.
- (g) Cf. Dialectology : An Introduction, W.N. FRANCIS, Longman Linguistics Library (29), pp.164-5, Harlow, 1983.

\*  
\* \*

L'on admet (1) que la dialectométrie, c'est-à-dire l'analyse quantitative de la variation dialectale, est née en France, et plus particulièrement dans le grand Sud-Ouest, au début des années 1970.

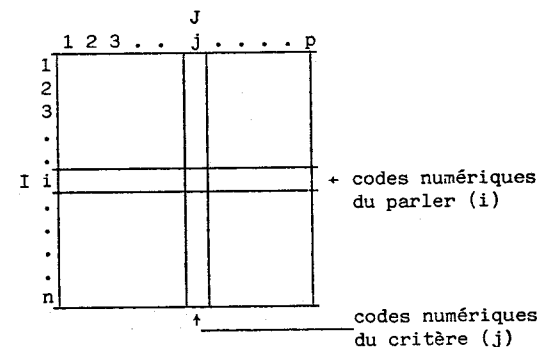
Les esprits scientifiques que l'on associe à cette naissance sont ceux de Henri GUITER (Université de Perpignan), et surtout, du moins en ce qui concerne l'école toulousaine, de Jean SEGUY, auteur aujourd'hui disparu de l'Atlas linguistique de la Gascogne (ALG), dont le volume 6 constitue la pierre angulaire de la dialectométrie.

Le but de ce volume, accompagné de matrices codées et d'une notice explicative, était d'opérer une synthèse cartographique de l'ensemble des données linguistiques éparpillées au travers des cinq premiers tomes de l'ALG, et ainsi de rendre compte de la variation dialectale gasconne. La collecte des données s'était faite par questionnaire (méthode directe et indirecte) dans 154 localités gasconnes essentiellement rurales pendant

les années 1950 et 1960, les points 790N et 790SO ayant été enquêtés par nous-mêmes en 1973-75. Les réponses fournies par les informateurs - des ruraux sédentaires originaires du point en question pour la plupart - étaient transcrites 'sur le tas' ou enregistrées sur bande magnétique en vue d'une transcription ultérieure.

L'approche variationniste adoptée par Jean SEGUY consistait à regrouper les données ainsi recueillies en cinq paramètres globalement représentatifs du fait linguistique : phonétique diachronique, phonologie, morphosyntaxe du pronom, morphologie verbale, et lexique. Chaque paramètre comprend une sélection pseudo-aléatoire de critères (phonèmes pour la phonologie, lexèmes pour le lexique, etc.), codés numériquement et disposés en forme de tableaux rectangulaires IxJ. L'axe horizontal de chaque tableau aligne les parlers enquêtés, alors que l'axe vertical comporte les critères :

Fig. 1 Schéma d'un tableau dialectométrique (matrices de l'ALG 6)



Les principes régissant la collecte, la sélection pseudo-aléatoire, et la composition paramétrique des données appellent sans doute un certain nombre de critiques, mais nous nous intéresserons ici aux seuls tableaux rectangulaires élaborés tels quels par Jean SEGUY.

La mesure de la variation spatiale du gascon s'effectue grâce à une technique de codage des erreurs évaluée au cours des années 50 par R. HAMMING (3). Ce procédé, adapté à l'ALG, consiste à attribuer un code discret à chaque donnée réunie sous la même variable (j), ceci pour l'ensemble des parlers en question. A partir d'un tableau paramétrique ainsi constitué (voir Fig.1), l'on relève pour chaque interpoint ( $i_1, i_2 ; i_1, i_3 ; i_2, i_3$  etc.), la somme des identités (0,0; 1,1; etc.) et des non-identités (0,1; 1,0; etc.).

Fig 2 Exemple de comptage

		J					
		i <sub>1</sub>	i <sub>2</sub>	i <sub>3</sub>	i <sub>4</sub>	i <sub>5</sub>	
I	i <sub>1</sub>	1	0	1	0	0	(i <sub>1</sub> , i <sub>2</sub> ) : 2 différences, 3 identités)
	i <sub>2</sub>	1	1	0	0	0	(i <sub>1</sub> , i <sub>3</sub> ) : 3 différences, 2 identités)
	i <sub>3</sub>	1	1	0	0	1	(i <sub>2</sub> , i <sub>3</sub> ) : 1 différence, 4 identités)
							Total ; 6 différences, 9 identités

La distance de HAMMING (dH) se calcule grâce à la formule suivante :

$$dH(i_1, i_2) = \frac{\sum d}{j} \text{ où } d = \text{somme des différences interponctuelles}$$

et j = somme des variables

La dH étant généralement exprimée en pourcentage, l'on constate les résultats suivants pour la Fig.2 ci-dessus :

$$dH(i_1, i_2) = 40\% (2/5)$$

$$dH(i_1, i_3) = 60\% (3/5)$$

$$dH(i_2, i_3) = 20\% (1/5)$$

La distance linguistique globale se calcule par l'addition des dH paramétriques selon la même formule.

Pourvu qu'il s'agisse d'un tableau logique, c'est-à-dire quand les valeurs du tableau sont 0 ou 1 ayant valeur booléenne ( $K_{IJ} \cdot (i, j) = 0$  ou 1), l'on obtient des résultats analogues par l'application de la formule euclidienne usuelle :

$$d^2(i_1, i_2) = \sum_{j \in J} (x(i_1, j) - x(i_2, j))^2$$

Cependant, la méthode de classification critérielle adoptée par Jean SEGUY fait que plusieurs numéros de code peuvent être réunis sous un même critère. C'est ainsi qu'en lexique, chaque critère constitue un référent capable d'engendrer de 0 à n lexèmes, tous codés différemment. L'on comprendra qu'il s'agit donc d'un codage pseudo-quantitatif, les chiffres n'ayant en réalité aucune valeur numérique. Que la comparaison interponctuelle donne (0,1) ou (6,17), l'on compte toujours un point de distance. En effet, seule la matrice phonologique constitue un véritable tableau logique de présence (1) - absence (0).

Depuis la disparition de Jean SEGUY en mars 1973, la recherche dialectométrique se poursuit à l'Université de Toulouse-Le-Mirail sous l'impulsion de J-L FOSSAT, Professeur d'Etudes Romanes et responsable du Centre de linguistique et de dialectologie (UTMII). Grâce à lui est né le projet d'informatisation de l'ALG (projet DELTA), en collaboration avec N. BEAUCHEMIN, Professeur à l'Université de Sherbrooke (Canada). C'est dans ce cadre que se situe notre recherche sur le gascon centre-pyrénéen (4), sous la direction de J-L FOSSAT.

Le but de cette intervention, compte tenu de l'état actuel de nos travaux, est plutôt modeste. Il s'agit de présenter la démarche méthodologique que nous avons suivie en vue de l'étude de la variation spatiale du gascon centre-pyrénéen, grâce aux techniques de la classification automatique et de l'analyse des correspondances qui nous ont été exposées par Bernard BURTSCHY, maître-assistant à l'Ecole Nationale Supérieure des Télécommunications (ENST, Paris), lui-même élève de J-P BENZECRI, Professeur de statistique à l'Université Pierre et Marie Curie (Paris).

D'un point de vue épistémologique, l'approche dialectométrique de Jean SEGUY se compare à la récente évolution survenue dans un certain nombre de disciplines scientifiques, dont les télécommunications, où la numérisation joue un rôle de plus en plus prépondérant. L'étude qualitative de la surface dialectale n'ayant fourni ni théorie ni heuristique permettant une analyse adéquate de la variation géolinguistique, notre point de départ a été de considérer, dès 1975 (5), que seul le traitement quantitatif des données mettrait en évidence la structure des aires en question. Ce traitement a pu être effectué à l'ENST au cours de 1981-82 grâce à l'inestimable concours de J-P BACH et surtout de B. BURTSCHY. Les programmes suivants, écrits en langage FORTRAN sur IRIS 80, ont été employés :

(1) J-P BACH : Distance de HAMMING (dH) : calcul des identités et non-identités de codage en chiffres bruts et en pourcentage concernant les trente points de notre zone (28 ALG + 2 BBB (5)). Ce programme, fourni en annexe, calcule toutes les dH paramétriques et cumulées, et assure donc la compatibilité et la comparabilité de nos résultats avec ceux de l'ALG6. Il permet également de vérifier les mesures séguyennes, et d'en augmenter le nombre, car seuls les calculs relatifs aux points contigus (et à quelques interpoints limitrophes) apparaissent dans les cartes dialectométriques de l'ALG6.

(2) M. JAMBU(6)/B. BURTSCHY : Classification ascendente hiérarchique (CAH)

Ce programme fournit les analyses suivantes (distribution des objets pour (i)=32) :

- fréquences conditionnelles PR(I|J), PR(J|I), PR(I|Q)
- contribution de la paire (i, j) à l'inertie du nuage
- tableaux des distances interponctuelles et ultramétriques (critère d'agrégation : maximisation du moment centré d'ordre 2 d'une partition)
- contribution des individus à la distance des classes de la hiérarchie A(J) au centre de gravité de l'ensemble J
- contribution relative des classes Q de la hiérarchie à la distance de la classe au centre du nuage.

Deux problèmes méthodologiques s'étaient posés comme préalable à la CAH :

- le choix d'une métrique

- la programmation des tableaux paramétriques à codage non-binaire.

Il a été décidé d'adopter comme métrique la distance distributionnelle de J-P BENZECRI (7), également appelée distance du CHI-2, soit :

$$d^2(f_{Ij}, f_{Ij'}) = \sum (1/f_i)(f_{ij} - f_{ij'})^2 \quad | \quad i \in I$$

sur l'espace des profils sur I, soit :

$$d^2(f_{ji}, f_{ji'}) = \sum (1/f_j)(f_{ji} - f_{ji'})^2 \quad | \quad j \in J$$

sur l'espace des profils sur J.

Compte tenu de la présence de la notion de masse (ou poids) dans ces formules ( $f, f'$ ), il a été nécessaire d'écrire un sous-programme permettant de transformer les tableaux non-binaires de l'ALG6 en tableaux de compatibilité logique, ce qui fait ressortir non seulement la classification des présences (1), mais aussi celle des absences (0).

Fig.3 Extrait du tableau des données phonologiques (profil sur

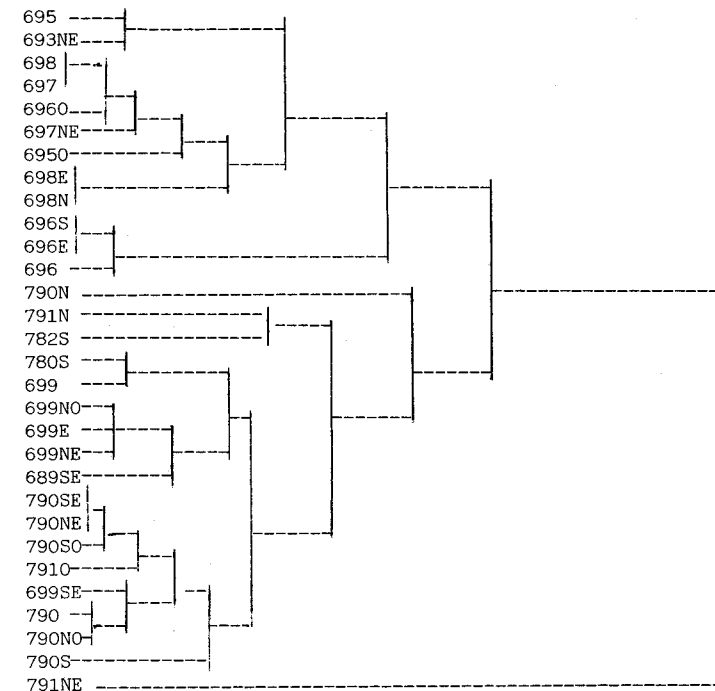
	J											
	6 0	6 1	7 0	7 1	8 0	8 1	10 0	10 1	16 0	16 1	21 0	21 1
689SF	0	1.0	1.0	0	0	1.0	0	1.0	1.0	0	0	1.0
693NE	1.0	0	0	1.0	1.0	0	0	1.0	1.0	0	0	1.0
695	1.0	0	0	1.0	1.0	0	0	1.0	1.0	0	0	1.0
6950	1.0	0	0	1.0	1.0	0	0	1.0	1.0	0	0	1.0
696	1.0	0	1.0	0	1.0	0	0	1.0	1.0	0	0	1.0
696E	1.0	0	1.0	0	1.0	0	0	1.0	1.0	0	0	1.0
696S	1.0	0	1.0	0	1.0	0	0	1.0	1.0	0	0	1.0
6960	0	1.0	0	1.0	1.0	0	0	1.0	1.0	0	0	1.0
697	1.0	0	0	1.0	1.0	0	0	1.0	1.0	0	0	1.0
697NE	1.0	0	0	1.0	1.0	0	0	1.0	1.0	0	0	1.0
698	1.0	0	0	1.0	1.0	0	0	1.0	1.0	0	0	1.0
698N	1.0	0	1.0	0	1.0	0	0	1.0	1.0	0	0	1.0
698E	1.0	0	1.0	0	1.0	0	0	1.0	1.0	0	0	1.0
699	1.0	0	0	1.0	0	1.0	0	1.0	1.0	0	0	1.0
699NE	1.0	0	1.0	0	0	1.0	0	1.0	1.0	0	0	1.0
699E	1.0	0	1.0	0	1.0	0	0	1.0	1.0	0	0	1.0
699SE	0	1.0	1.0	0	0	1.0	0	1.0	1.0	0	0	1.0
699NO	1.0	0	0	1.0	0	1.0	0	1.0	1.0	0	0	1.0
780S	0	1.0	0	1.0	0	1.0	0	1.0	1.0	0	0	1.0
782S	0	1.0	0	1.0	1.0	0	0	1.0	1.0	0	0	1.0
790	0	1.0	1.0	0	0	1.0	0	1.0	1.0	0	0	1.0

Pour un exposé statistique de la CAH, l'on consultera l'ouvrage de M. JAMBU et de J-P BENZECRI (op.cit.), car ici ne seront commentées que les possibilités d'interprétation variationniste. Le point d'entrée le plus transparent dans un tableau logique est sans doute la représentation arborescente des données, qui permet une saisie visuelle de leur classification.

L'on s'aperçoit (Fig.4) que notre zone comprend un parler (791 NE) dont le système phonologique le différencie de tout autre

parler - il s'agit en effet d'un point se trouvant en zone interférentielle gascon-languedocien (8). Le poids structurant de ce parler justifie sa mise en élément supplémentaire d'un point de vue statistique, mais ce n'est pas son cas dans les autres paramètres. Outre cette observation, nous constatons que la zone en question se scinde en deux aires phonologiques primaires (9), dont le degré individuel d'homogénéité est plus ou moins comparable (distance ultramétrique 0,075 et 0,088 respectivement).

Fig.4 CAH : représentation arborescente des parlers (phonologie)



Les cartes phonologiques fournies en annexe, dérivées de la Fig.4, (1,2,3,A), montrent la décomposition de notre zone en aires secondaires et tertiaires territorialement intactes, à deux nuances près, sans qu'aucune information à caractère géographique ait été intégrée au programme informatique.

Cette observation nous paraît d'une importance capitale, car elle démontre que la surface dialectale est effectivement dotée d'une structure géolinguistique cohérente. Nous verrons

plus loin (p.14) que les frontières circonscrivant les aires de cette surface constituent des interfaces synthétiques à valeurs multiples.

La visualisation de la représentation arborescente ne suffit cependant pas à tout expliquer - l'on s'interrogera par exemple sur les éventuelles incohérences territoriales, les séparations entre classes, les niveaux d'homogénéité, les phénomènes structurants, etc., auquel cas l'on consultera les tableaux des distances simples ou ultramétriques entre chaque sous-ensemble de parlers (Q).

Les tableaux des fréquences conditionnelles  $PR(I|J)$  et  $PR(I|Q)$  soit la contribution de la présence ou absence de chaque critère à chaque parler (I|J) ou sous-ensemble de parlers (I|Q) revêtent peu d'intérêt dans la mesure où la distribution des objets est un constante, car  $0 = 0$  et  $1 = 1/\Sigma$ . Par contre, le tableau  $PR(J|I)$ , soit la contribution de chaque parler à la présence ou absence de chaque critère est plus révélateur, la distribution des objets étant une variable ; l'on voit bien y apparaître le poids structurant des critères, ainsi que leur répartition spatiale.

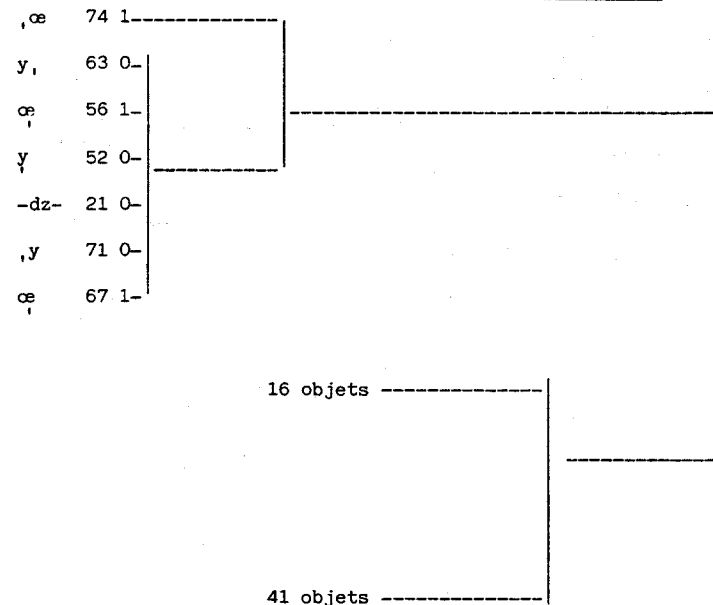
Le tableau des contributions de la paire (i,j) à l'inertie du nuage rapporte la dispersion spatiale des parlers et celle des critères à leur barycentre linguistique. Les deux autres sorties permettent d'étudier respectivement les contributions relatives des sous-ensembles de parlers (Q) de la hiérarchie à la distance de chaque classe au centre du nuage, et les contributions des critères à la distance des classes de la hiérarchie A(J) au barycentre de l'ensemble J.

Si notre tableau logique  $I \times J$  permet d'analyser la variation spatiale des parlers, son retournement en tableau  $J \times I$  met en évidence la notion de distance intercritérielle, notion que l'on ne retrouve ni en dialectométrie séguyenne, où il est uniquement question de distance entre parlers, ni en dialectologie classique, d'où son intérêt fondamental. Ici, la distribution des objets est nécessairement une variable, les valeurs additionnées de chaque paire (présence + absence) pouvant varier entre (1) et  $(\Sigma_i - 1)$ , mais devant obligatoirement totaliser  $\Sigma_i(I)$ , soit les 30 parlers enquêtés dans notre zone.

L'on s'aperçoit (Fig.5) que la surface phonologique centre-pyrénéenne se divise en deux sous-ensembles de dimensions inégales, l'un composé de 57 objets (distance ultramétrique 0,113), l'autre, relativement homogène, des sept objets restants (distance ultramétrique 0,023). Ce dernier sous-ensemble comprend la présence ou absence de six phénomènes vocaliques, le septième étant consonantique, caractérisant exclusivement le système phonologique de 791NE, le parler interférentiel désigné plus haut. Sans vouloir approfondir ici l'étude qualitative de nos données, nous soulignons son utilité - l'on s'interroge par exemple sur la possibilité d'une éventuelle corrél-

lation distributionnelle entre la présence ou absence de certains traits distinctifs gascons.

Fig.5 Représentation arborescente des phonèmes (extrait)



Laissant de côté le sous-ensemble de sept objets, nous constatons que notre zone se divise phonologiquement en deux aires secondaires elles aussi inégales - un sous-ensemble composé de seize variables (distance ultramétrique 0,046), l'autre des quarante-et-un objets restants (distance ultramétrique 0,073), que nous n'analyserons pas ici.

(3) E.R. YAGOL-NITZER/J-P BENZECRI/B. BURTSCHY : analyse des correspondances

Ce programme fournit les sorties suivantes :

- moyenne, écart type, coefficient de variation, biais, aplatissement (linguistiquement non-pertinents pour une distribution (0,1)
- analyse factorielle des correspondances
- valeurs et vecteurs propres (listing, histogramme)
- facteurs de correspondance
- graphiques des valeurs propres (V1/V2, V1/V3, V2/V3).

La matrice phonologique de l'ALG relative à notre zone constitue un tableau de correspondance où l'ensemble I comprend les

trente parlars en question, et l'ensemble J les trente-deux phonèmes différentiels (sur soixante-dix-sept au total dans l'ALG), chacun réparti en deux colonnes (présence et absence), soit soixante-quatre variables.

L'analyse factorielle des correspondances donne lieu à un tableau de valeurs exprimant les corrélations relatives entre parlars (profil sur I), ou entre variables (profil sur J). C'est ainsi qu'une valeur maximale identique est attribuée (à une exception près) aux variables comprenant le sous-ensemble primaire mis en évidence par la CAH (cf. Fig.5), car ce sont les seules dont la distribution spatiale, identique, caractérise un parler unique (791NE).

La deuxième analyse effectuée concerne les valeurs propres ( $\lambda$ ), c'est-à-dire les moments principaux d'inertie d'un nuage. Alors que l'analyse factorielle sert à définir les directions de dispersion du nuage autour de son barycentre, la détermination des valeurs propres permet de mesurer cette même dispersion dans chacune des directions principales. En outre, le fait qu'il y ait égalité des valeurs propres de (I) et de (J), et donc le même nombre d'axes principaux d'inertie, permet leur représentation simultanée sur un même graphique, ce qui facilite le rapprochement des éléments de (J) dus aux éléments de (I), et vice-versa.

Le tableau des valeurs comprend quatre colonnes contenant :

- l'énumération des  $\lambda$  à partir de 2 (la première étant un calcul théorique en principe égal à 1)
- le pourcentage d'inertie représenté par la  $\lambda$  afférente à cet axe
- le cumul des pourcentages
- l'histogramme des  $\lambda$  en longueurs proportionnelles.

Fig.6 Extrait du tableau des valeurs propres (phonologie, profil sur J)

N° Iter	$\lambda$	%	cumul	Histog. des valeurs propres
2 0	0,24541849	24,541	24,541	*****
3 1	0,17882180	17,882	42,423	*****
4 0	0,09237272	9,237	51,660	*****
5 1	0,08482128	8,482	60,142	*****
6 1	0,07115906	7,116	67,258	*****
7 2	0,06183777	6,184	73,442	*****
8 1	0,04544279	4,544	77,986	*****

Pour notre zone, les axes concernant les deux premières  $\lambda$  sont relativement individualisés, et représentent plus de 40% de l'inertie totale du nuage.

Les deux tableaux regroupant les facteurs de correspondance (ou coordonnées situées sur les axes principaux d'inertie) se lisent en millièmes sur I et sur J. Chaque listing comprend les données suivantes :

QLT = qualité de la représentation des (i) : plus le chiffre est près de 1.000, mieux est la représentation  
 POID = la masse (ou poids)  $f_i$   
 INR = inertie des (i) par rapport au centre du nuage  
 1#F = première, etc., valeur propre  
 COR = contribution relative du facteur au carré de la distance à l'origine  
 CTR = valeur relative du point à l'inertie le long de l'axe sur lequel est pris le facteur.

Fig.7 Extrait du tableau des facteurs de correspondance (phonologie)

(A1) Valeurs propres sur I					
1#F		2#F		3#F	
-2257	791NE	-1162	791NE	986	790N
-----		-656	696E	470	695
528	696E	-619	696S	455	697NE
471	696S	-561	696	421	6950
465	696	-----		(129 791NE)	
418	695	782	790N	-----	
		432	782S	-392	696E
		432	791N	-353	696S

(A2) Valeurs propres sur J (0=absence, 1=présence)					
1#F		2#F		3#F	
-4456	-dz- (0)	-2747	-dz- (0)	3245	-h- (0)
-4556	y (0)	-2747	y (0)	3245	-r- (0)
-4556	œ (1)	-2747	œ (1)	2316	-h (0)
-4556	y, (0)	-2747	y, (0)	1524	-r (0)
-4556	œ, (1)	-2747	œ, (1)	1440	-r- (0)
-4556	y (0)	-2747	y (0)	-----	
-2430	œ (1)	-----		-1360	ø (1)
		1849	-h- (0)	-1099	^ (1)
986	^ (1)	1849	-r- (0)	-1099	,^ (1)
986	,^ (1)	1022	-g- (0)		
365	-j (0)	1022	-j (0)		

(B1) Facteurs de correspondance sur I					
CTR1		CTR2		CTR3	
693	791NE	252	791NE	351	790N
38	698E	114	790N	80	695
30	696S	80	696E	75	697NE
29	696	72	696S	64	6950
24	695	59	696	62	790S
				( 6 791NE)	



(B2) Facteurs de correspondance sur J (0=absence, 1=présence)

CTR1	CTR2	CTR3
88 -dz- (0)	70 j- (1)	121 -λ (0)
88 y (0)	54 j- (0)	129 -λ- (0)
88 œ (1)	46 - (0)	119 -r- (0)
88 y <sub>i</sub> (0)	44 -dz-(0)	117 r- (0)
88 œ <sub>i</sub> (1)	44 y (0)	63 -t- (0)
88 y (0)	44 œ (1)	

L'étude de la première valeur propre du nuage (I) permet d'apprécier l'écrasant poids structurant du point interférentiel 791NE, et, à l'autre extrémité de l'axe, le sous-ensemble 696, 696E, 696S (cf. Fig.4), doté d'un poids beaucoup moins structurant. Ces extrémités correspondent parfaitement à celles de l'axe géographique SE-NW de notre zone (cf. cartes phonologiques en annexe).

Passant à la deuxième  $\lambda$ , l'on note que les extrémités de son axe opposent 791NE à 790N, 782S et 791N, points géographiques rapprochés, dont les systèmes phonologiques comportent un petit nombre d'éléments, souvent endémiques, très structurants : articulation dorso-uvulaire et yodisée de /r/ et /λ/ respectivement à 790N, absence de /-s/ à 782S, et de /-d-/ à 791N.

La contribution massive de 791NE aux deux premiers axes principaux d'inertie confirme la nécessité de la mise en élément supplémentaire de ce point, du moins en ce qui concerne le paramètre phonologique, ce que tend à faire la troisième  $\lambda$ , où la valeur de 791NE est de 129 (cf. Fig.7, A1, 3#F). Chacune des  $\lambda$  permet en outre d'étudier les rapports d'opposition entre parlers, grâce à l'attribution de valeurs positives et négatives.

Etant donné qu'il y a identification des axes principaux d'inertie des deux nuages (I,J), et qu'une formule de transition (voir J-P BENZECRI op.cit. vol.3, p.100) permet de dériver les facteurs sur (J) à partir des facteurs sur (I), il y aura égalité des valeurs propres concernant les deux nuages. C'est ainsi que la première  $\lambda$  de (J) oppose la présence ou absence des phonèmes caractérisant 791NE, soit /-dz-, y, œ, y<sub>i</sub>, œ<sub>i</sub>, y, œ/, aux éléments structurants du sous-ensemble 696, 696E, 696S, soit /λ, λ, -r, -n/. Les valeurs (positives) de cette classe sont d'ailleurs beaucoup plus faibles que celles de 791NE, démontrant ainsi une spécificité gasconne plus prononcée.

Les extrémités de l'axe de la deuxième  $\lambda$  correspondent aux éléments structurants propres à 791NE d'un côté, et ceux de 790N (/λ-, -r-), 782S (/s/), et 791N (/d-/) de l'autre, alors que la troisième  $\lambda$  oppose les éléments caractérisant 790N à ceux de 790S (/θ/) et de 696E/696S.

Cet aspect de l'analyse démontre clairement l'intérêt du concept d'"espace phonologique" : si le premier axe principal du nuage (I) correspond à un axe de dispersion géographique des parlers, celui du nuage (J) correspond tout autant à un axe de dispersion de traits phonologiques. L'étude de cet espace nous permet de découvrir les oppositions ou complémentarités d'éléments qui existent en gasco centre-pyrénéen. L'on constate par exemple que la présence de /y/ tonique (/y/) et de /œ/ tonique y est mutuellement exclusive, et que cette opposition s'étend également aux positions prétonique et posttonique.

Nous nous sommes déjà interrogé quant à l'éventuelle existence d'une corrélation distributionnelle de certains phonèmes ou traits distinctifs. L'étude de l'espace phonologique gascon nous permettrait de déterminer par exemple si l'agrégation de l'articulation dorso-uvulaire de /r/ et celle, yodisée, de /λ/ est aléatoire ou non.

Les listings des contributions relatives de chaque parler ou variable à la dispersion des trois axes de la Fig.7 (B1, B2), constituent des tableaux hiérarchiques (en millièmes), dont l'intérêt est de rendre compte de l'importance relative des éléments écartés de l'analyse, par rapport aux éléments principaux.

Les dernières sorties concernent les représentations graphiques des nuages N(I) et N(J), dont les principaux axes factoriels sont imprimés individuellement ou en superposition (dans notre cas :  $\lambda_1/\lambda_2$ ,  $\lambda_1/\lambda_3$ ,  $\lambda_2/\lambda_3$ ). L'avantage de ces graphiques est de faciliter une saisie visuelle instantanée de la dispersion des facteurs.

Fig.8 Représentation graphique des  $\lambda_1/\lambda_2$  du nuage (I) - phonologie

(voir page suivante)



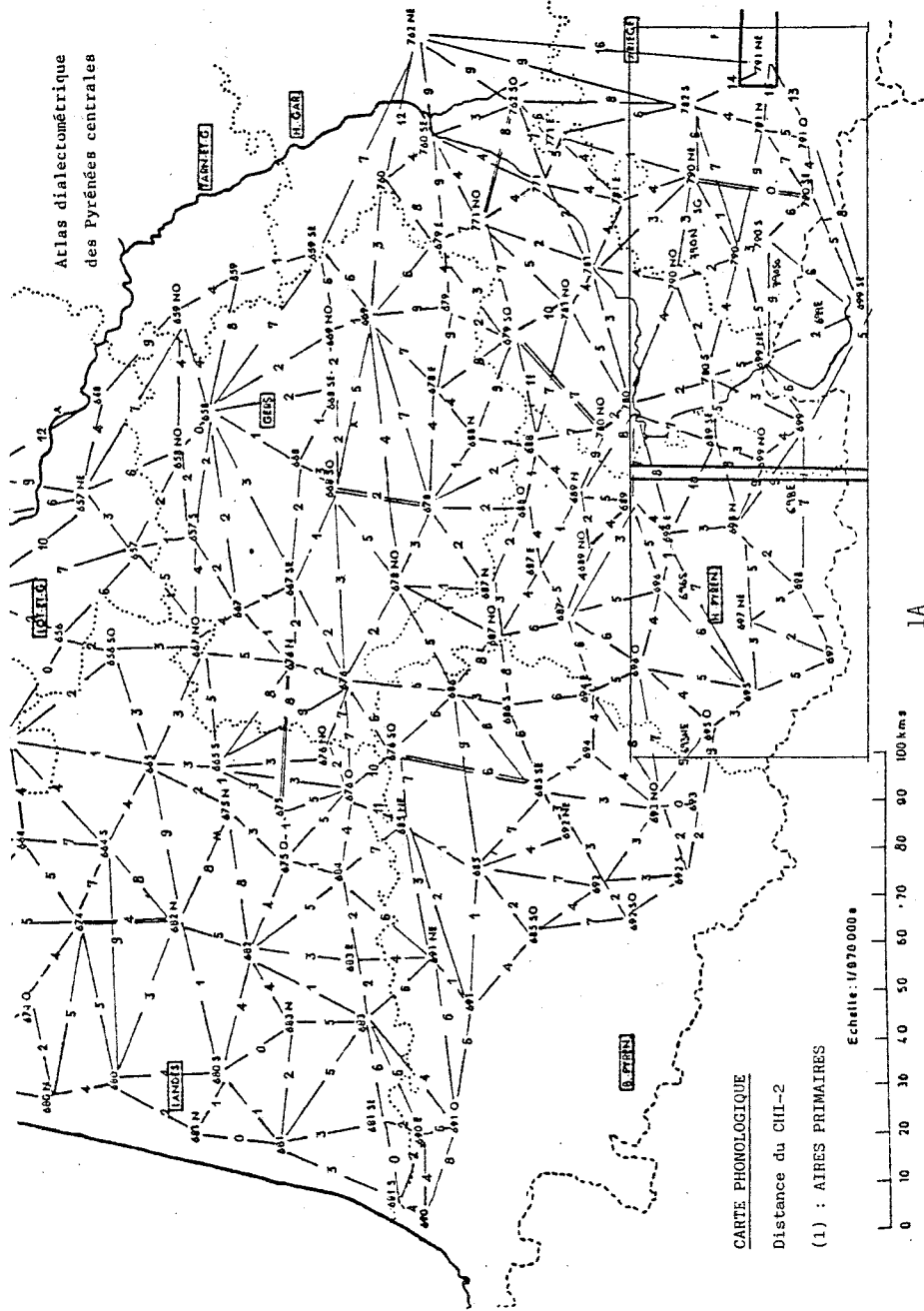
En conclusion, nous rappelons que nos résultats sont exemplifiés ici par rapport au seul paramètre phonologique. Cependant, l'ensemble de nos observations vaut également pour les autres sorties paramétriques ou cumulées : phonétique diachronique, morpho-syntaxe du pronom, morphologie verbale et lexicale, compte tenu de leurs aires divergentes. Notons enfin l'intérêt de l'application des techniques automatiques exposées ci-avant à la dialectologie historique, sociale ou ethnographique, où la caractérisation d'un espace humain requiert souvent une connaissance des aires sociolinguistiques déterminées depuis la préhistoire par l'effet variationniste.

NOTES ET BIBLIOGRAPHIE

- (1) Voir par exemple Dialectology, CHAMBERS & TRUDGILL, Cambridge University Press, 1980, p.112.
- (2) Atlas linguistique (et ethnographique) de la Gascogne, vols. 1-6, CNRS, Paris, 1956-73.
- (3) Cf. Error Detecting and Error Correcting Codes, R. HAMMING, in Key Papers in the Development of Coding Theory, E. BERLEKAMP, IEEE Press, New York, 1974.
- (4) Atlas linguistique automatisé des Pyrénées centrales, thèse d'Etat en cours, CLD-Université de Toulouse II, D. PHILIPS.
- (5) Voir Balaguères-Bethmale-Biros : Etude dialectométrique, D. PHILIPS, thèse de 3e cycle, CNRS-UTMII, 1976.
- (6) Classification automatique pour l'analyse des données, M. JAMBU, 1 - Méthodes et algorithmes, 2 - Logiciels, Bordas-Dunod, Paris, 1978.
- (7) Pratique de l'analyse des données 1 - Analyse des correspondances, J-P et F. BENZECRI, Bordas-Dunod, 1980 ; 3 - Linguistique et lexicologie, J-P BENZECRI et collaborateurs, Bordas-Dunod, 1981.
- (8) Voir Les interférences linguistiques entre gascon et languedocien dans les parlers du Comminges et du Couserans, P. BEC, Presses Universitaires de France, Paris, 1968.
- (9) Voir Carte phonologique 1A : aires primaires, en annexe.

\*\*\*\*\*

NOTA : Les logiciels (données ALG6 et programmes) désignés sont implantés sur DPS8/Multics au CICT - Université de Toulouse III, en Fortran, et sont accessibles depuis les locaux du CLD-URL8 (UTMII-CNRS).

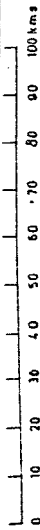


CARTE PHONOLOGIQUE

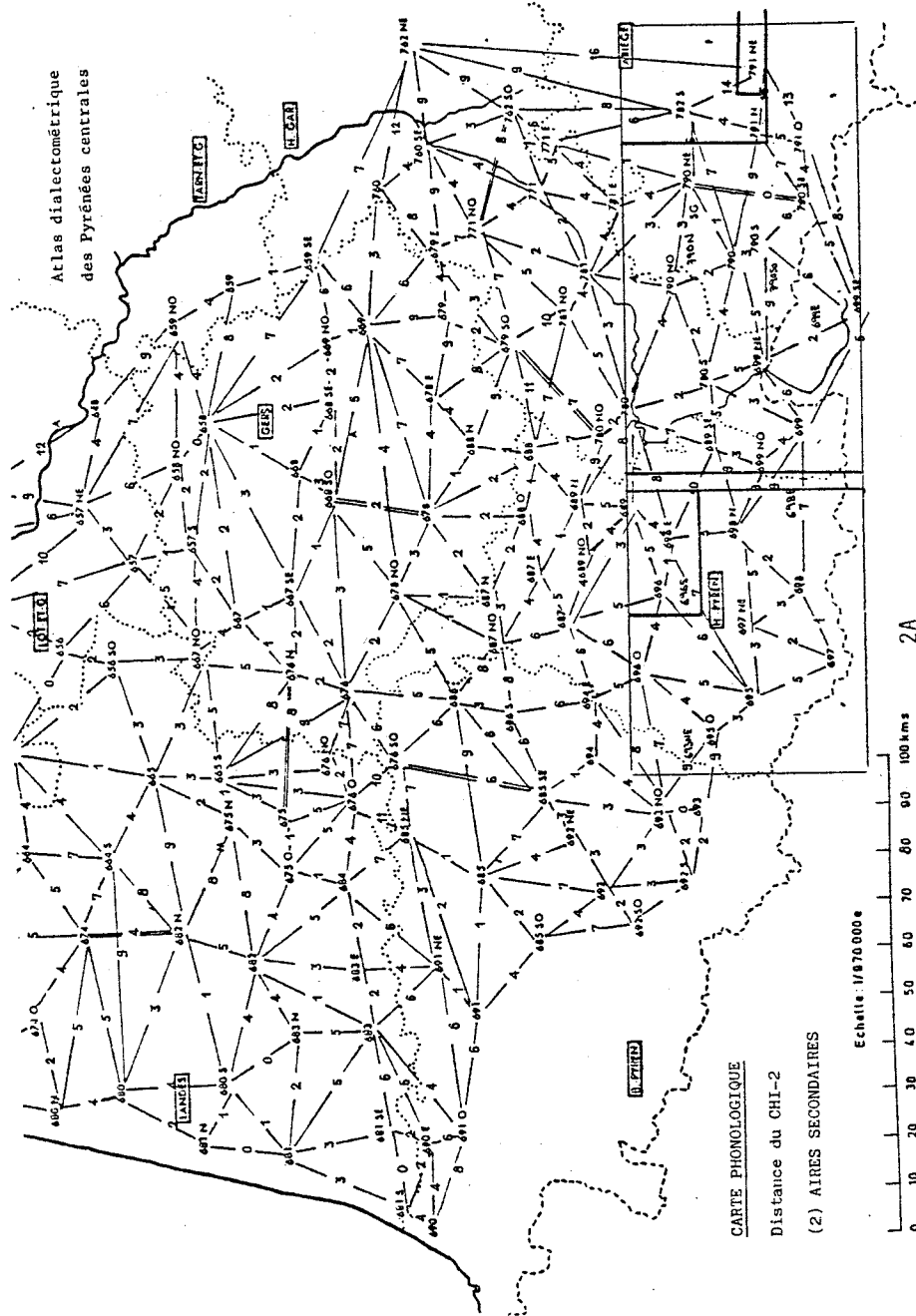
Distance du CHI-2

(1) : AIRES PRIMAIRES

Echelle: 1/870 000 e



IA

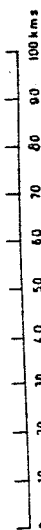


CARTE PHONOLOGIQUE

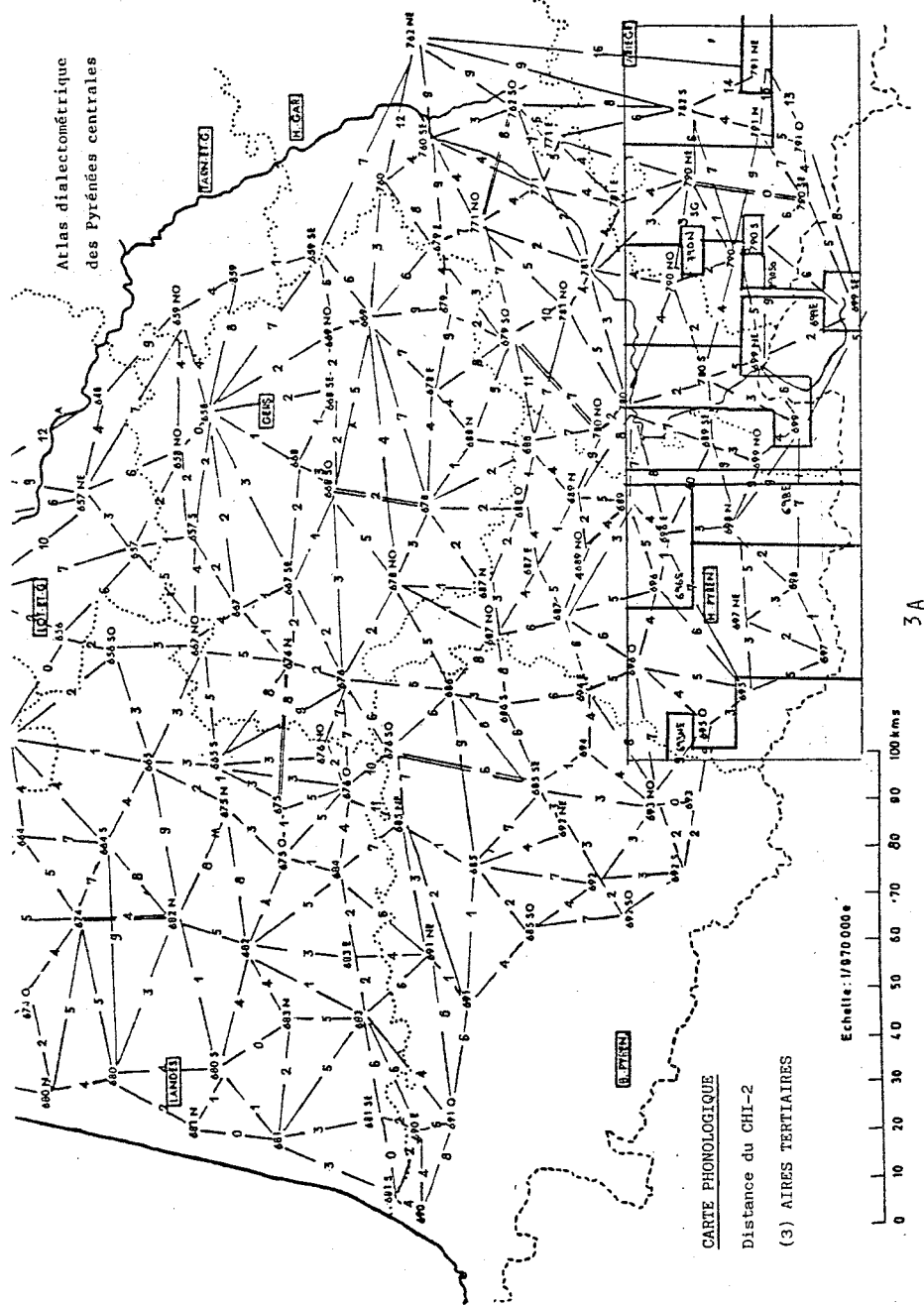
Distance du CHI-2

(2) AIRES SECONDAIRES

Echelle: 1/870 000 e



2A



## PROGRAMME DE J-P BACH : DISTANCE DE HAMMING

## PHONOLOGIE

```

0001      IMPLICIT INTEGER (A-Z)
0002      INTEGER B(100)
0003      INTEGER NOM (50), CODE (50), VAL(50,100), MAT (50,50)
0004      REAL PC1,PC2
0005      C
0006      C      NVAL = NOMBRE DE PARAMETRES
0007      NVAL = 77
0008      C      LECTURE DE DONNEES
0009      I = 0
0010      1      READ (1,100,END=9) NOMB, CODEB, (B(J), J=1,NVAL)
0011      C100      FORMAT (I3,R2,68Z1)
0012      100      FORMAT (I3,R2,77RI)
0013      DO 11 J = 1,NVAL
0014      IF (B(J).EQ.64) B(J) = -1; GO TO 11
0015      B (J) = B(J) - 240
0016      11      CONTINUE
0017      I=I+1
0018      CODE (I) =CODEB
0019      NOM(I) = NOMB
0020      DO 2 J = 1,NVAL
0021      VAL (I,J) = B(J)
0022      IF (I.LT.50) GO TO 1
0023      9      CONTINUE
0024      C      FIN DE LECTURE : I = NOMBRE DE TERMES LUS
0025      MAX = I; OUTPUT MAX
0026      C
0027      C
0028      C      CALCUL DE LA MATRICE DES DIFFERENCES
0029      C      MAT (I1,I2) = SOMME DES DIFFERENCES SI I1<I2
0030      C      MAT (I2,I1) = SOMME DES DIFFERENCES SI I1<I2
0031      C
0032      MAXM1 = MAX-1
0033      DO 10 I1=1, MAXM1
0034      I1P1 = I1+1
0035      DO 20 I2 = I1P1,MAX
0036      MAT (I1,I2) = 0
0037      MAT (I2,I1) = 0
0038      DO 30 J=1, NVAL
0039      IF (VAL (I1,J) .NE.VAL (I2,J)) GO TO 35
0040      C      ICI, LES DEUX VALEURS SONT EGALES
0041      MAT (I2,I1) = MAT (I2,I1) + 1
0042      GO TO 30
0043      35      CONTINUE
0044      C      ICI, LES DEUX VALEURS SONT DIFFERENTES
0045      MAT (I1,I2) = MAT (I1,I2) + 1
0046      30      CONTINUE
0047      IF (MAT (I1,I2) + MAT (I2,I1) .NE.NVAL) OUTPUT 'YAUNOS'

```



