REVUE DE STATISTIQUE APPLIQUÉE

A. MORIN

Comparaison de plusieurs méthodes de classification sur un exemple de lexicométrie

Revue de statistique appliquée, tome 32, n° 4 (1984), p. 37-49

<http://www.numdam.org/item?id=RSA_1984__32_4_37_0>

COMPARAISON DE PLUSIEURS METHODES DE CLASSIFICATION SUR UN EXEMPLE DE LEXICOMETRIE

A. MORIN I.R.I.S.A.
Université de Rennes 1

Résumé

L'objectif est le classifier n individus caractérisés par leur courbe de fréquence de mots prononcés au cours d'une entrevue et par quelques variables sociologiques. Nous utilisons trois méthodes de classification, classification hiérarchique (I.C. LERMAN), méthode des nuées dynamiques (E. DIDAY), classification sur tableau de contingence (G. GOVAERT). La comparaison des résultats obtenus se fait par analyse factorielle et par calcul d'un indice mis au point par I.C. LERMAN.

I. INTRODUCTION

Il existe peu d'outils statistiques pour comparer et classifier des courbes de fréquences de mots. En effet, le corpus des données est en général grand et la présence de mots rares de fréquence 1 et de mots fréquents, jusqu'à 25 000 occurences dans certains cas, compliquent encore le problème. Notre objectif est de classifier n individus lorsque nous disposons d'une courbe de fréquence de mots par individu. Pour cela, nous avons utilisé trois méthodes : la première est une variante optimale suivant un certain critère que nous préciserons, de la méthode des nuées dynamiques, la seconde utilise l'algorithme de classification d'un tableau de contingence de G. GOVAERT, la troisième est la méthode de classification hiérarchique de I.C. LERMAN. Pour comparer les résultats, nous utilisons un indice permettant de découvrir les liens éventuels des modalités des caractères des classifications et une analyse du tableau disjonctif complet nxk où n est le nombre d'individus et k le nombre total de groupes obtenus par les trois méthodes de classification. Nous mettons en évidence l'intérêt de la première méthode de classification adoptée ainsi que la cohérence des résultats obtenus par les trois méthodes.

II. DONNEES DU PROBLEME ET OBJECTIF

Les données concernent 120 individus francophones de Montréal. Le corpus a été fourni par D. SANKOFF [9]. Notre objectif est de choisir une pondération optimale pour la comparaison et la classification des individus, pondération inter-

Mots-clés : **Classification hiérarchique, Méthode des nuées dynamiques, Analyse des correspondances.**

venant dans la variante des nuées dynamiques que nous avons utilisée [6]. Le but du projet global est l'étude de la structure de la variation linguistique à l'intérieur de la communauté francophone de Montréal. Les interviews des 120 individus sont des conversations informelles sur des sujets divers concernant la vie à Montréal. Elles permettent d'obtenir en moyenne 8 000 mots par personne interrogée. Une description complète de la méthode de sondage est publiée dans [9] [10] et [11]. Chaque individu est caractérisé par des variables supplémentaires telles que son âge, sa profession, son sexe, son nombre d'années de scolarité (SCO) et un indice sociolinguistique (CSL) (Annexe 1).

Notons m le nombre de mots (on dira aussi formes) différents dans tout le corpus, n le nombre d'individus interviewés (ici n = 120).

Précisons la forme de notre tableau des données.

Soit f(j) la fréquence du j^{è} mot dans le corpus entier, avec les mots ordonnés de la façon suivante :

$$0 < f(m) \leqslant f(m-1) \leqslant \qquad \leqslant f(1)$$

et

$$\sum_{j=1}^{m} f(j) = 1.$$

Nous disposons d'un tableau de nombres compris entre 0 et 1, à m lignes et 120 colonnes. Nous noterons (x_i^p) la fréquence relative du p^{è} mot pour le i^{è} individu.

III. METHODES DE CLASSIFICATION UTILISEES

Nous avons considéré 2 catégories de techniques. Les deux méthodes du premier groupe sont basées sur l'algorithme des nuées dynamiques. La première est une variante des nuées dynamiques de Diday [1] dans laquelle on pondère les variables et classifie les individus. On recherche donc d'une part des groupes d'individus A_k et d'autre part des pondérations w_p. La fonction à minimiser s'écrit :

$$F = \sum_{k=1}^{G} \frac{1}{n_k (n_k - 1)} \sum_{i,j \in A_k} \sum_{p=1}^{m} w_p^2 (x_i^p - x_j^p)^2$$

sous la contrainte, $\sum_{p=1}^{m} w_p = 1 \quad w_p > 0 \; \forall p.$

avec G = le nombre de groupes
A_k = k^{è} groupe
n_k = la taille du k^{è} groupe
x_i^p = l'observation de la p^{è} variable sur le i^{è} individu
n = nombre d'individus
m = nombre de variables

SEBESTYEN [12] avait déjà résolu un problème analogue. Mais nous introduisons une simplification supplémentaire. Au lieu de conserver autant de valeurs

de w qu'il y a de valeurs différentes des fréquences, nous divisons l'intervalle des fréquences en un nombre de régions inférieur au nombre de fréquences et nous supposons que w est constant sur chaque région.

Dans ce cas, lorsque nous minimisons F par rapport à w, nous obtenons :

$$w(r) = \frac{C}{\sum_{k=1}^{G} \frac{1}{n_k (n_k - 1)} \sum_{i,\, j \in A_k} \sum_{p,\, f(p) \in \text{Région } (r)} (x_i^p - x_j^p)^2}$$

où C est une constante déterminée de telle sorte que $\sum_r m_r\, w(r) = 1$, m_r désignant le nombre de fréquences tombant dans la région r.

Les A_k ne sont pas connus a priori mais redéterminés à chaque itération. On reporte l'expression de w(r) dans F qu'on minimise.

L'algorithme de calcul est décrit et détaillé dans [6]. Notons que LUMESKY [5] retient aussi cette méthode pour obtenir des groupes plus compacts.

La seconde méthode proposée par G. GOVAERT [2] recherche une suite de couples de partitions (P, Q) sur les lignes et sur les colonnes d'un tableau de contingence (n_{ij}) $i \in I, j \in J$.

Dans ce cas, on applique successivement sur I et sur J une variante de la méthode des nuées dynamiques (mnd). Le critère utilisé est sur les lignes

$$F(P) = \sum_{k=1}^{G} \sum_{i \in p_k} f_{i\cdot}\, d^2\,(i, G(p_k))$$

où $f_{i\cdot} = \sum_j \frac{n_{ij}}{\sum_i \sum_j n_{ij}}$ et $f_{\cdot j} = \sum_i \frac{n_{ij}}{\sum_i \sum_j n_{ij}}$, et où p_k est le $k^{\text{ème}}$ groupe de la partition P, $G(p_k)$ le centre de gravité de p_k et $d(i, G(p_k))$ la distance entre i et $G(p_k)$. La métrique utilisée est définie par la matrice diagonale des $(1/f_{\cdot j})$. On a une quantité analogue sur les colonnes. Govaert démontre que minimiser F revient à maximiser χ^2 (P, Q) où χ^2 est le χ^2 du tableau de contingence obtenu lorsqu'on a la partition P sur les lignes et Q sur les colonnes. Enfin, pour utiliser l'algorithme de I.C. LERMAN [4] sur un tableau de contingence, il fallait définir un indice de proximité entre les lignes du tableau. Cet indice a été défini en associant à chaque ligne une variable numérique définie de telle sorte que la distance entre deux variables soit la distance du χ^2. Cet indice, conforme à la classe des indices de proximité de I.C. LERMAN, n'est rien d'autre que le coefficient de corrélation entre deux variables.

IV. RESULTATS ET COMPARAISON DES RESULTATS

Algorithme combiné de pondération des variables et de classification des individus appelé algorithme MORIN-SANKOFF

Nous avons choisi G = 4 groupes. Il n'y avait pas de différenciation prononcée dans les données qui aurait pu justifier un plus grand nombre de groupes.

La taille du noyau de chaque groupe a été fixée à 6 éléments. Nous voulions des noyaux suffisamment grands pour bien représenter un groupe. D'autre part, nous devions tenir compte des temps de calculs. Après quelques essais, nous avons donc pris 6 éléments par noyau.

Une dernière décision concernait le choix des régions. Soit L le nombre total de mots utilisés ; ici L est voisin de 10^6. Nous avons pensé que les mots qui n'apparaissent qu'une seule fois, il y en avait environ 12 000, ne nous aidaient pas à classifier les individus ; par conséquent, ils n'ont pas été considérés. Les régions sont les suivantes :

R_1 = mots dont la fréquence est $\frac{2}{L}$

R_2 = $\frac{3}{L}$

R_3 = $\frac{4}{L}$

R_4 = $\frac{5}{L}$

R_5 = mots dont la fréquence est comprise entre $\frac{6}{L}$ et $\frac{10}{L}$

R_6 = $\frac{11}{L}$ $\frac{20}{L}$

R_7 $\frac{21}{L}$ $\frac{50}{L}$

R_8 $\frac{51}{L}$ $\frac{100}{L}$

R_9 = $\frac{101}{L}$ $\frac{200}{L}$

R_{10} = $\frac{201}{L}$ $\frac{500}{L}$

R_{11} = $\frac{501}{L}$ $\frac{1000}{L}$

R_{12} = $\frac{1001}{L}$ $\frac{2000}{L}$

R_{13} = $\frac{2001}{L}$ $\frac{5000}{L}$

$R_{14} = \dfrac{5001}{L} \quad \dfrac{8000}{L}$

R_{15} = mots dont la fréquence est supérieure à $\dfrac{8000}{L}$. Nous avons donc 15 régions.

L'algorithme converge en 5 ou 10 itérations vers un minimum local. D'un passage à l'autre, les groupes sont parfois différents dépendant des conditions initiales choisies. Bien que la majorité des membres des noyaux réapparaisse dans des configurations assez voisines d'un passage à l'autre, les solutions diffèrent. Ce n'est pas très surprenant du fait de la nature des données mais cela rend l'interprétation des résultats difficile. Après avoir classifié nos individus plusieurs fois en modifiant les conditions initiales, nous avons tenté d'identifier les formes fortes.

Pour rendre cette notion de formes fortes opérationnelle, nous avons utilisé l'algorithme de classification hiérarchique suivant la distance moyenne ("l'average-linkage"), l'indice de proximité entre individus i_1 et i_2 utilisé étant

$$I(i_1, i_2) = \frac{1}{p} \sum_{i=1}^{p} l_i(i_1, i_2).$$

où p est le nombre de classifications réalisées (ici 11)

$l_i(i_1, i_2) = 1$ si pour la classification i, les individus i_1 et i_2 se trouvent dans la même classe. $l_i(i_1, i_2) = 0$ sinon.

Notre algorithme (nuées dynamiques suivi de classification hiérarchique) décompose notre échantillon de 120 individus en 11 groupes. Il a tendance à différencier les locuteurs de la façon suivante : un groupe de personnes plus jeunes et un groupe de personnes plus âgées. A l'intérieur de la majorité restante, il y a des regroupements consistants mais non corrélés de façon évidente à des caractéristiques socio-démographiques ou à un comportement socio-linguistique particulier [7].

Quant à w, d'un passage à l'autre, il reste assez constant et nous obtenons :

$$w\left(\frac{2}{L}\right) = 0.15$$

$$w\left(\frac{3}{L}\right) = 0.20$$

$$w\left(\frac{4}{L}\right) = 0.23$$

$$w\left(\frac{5}{L}\right) = 0.23$$

$$w(f) = 0{,}25\, f^{-2/3} \text{ sauf pour } f > \frac{8000}{L}.$$

Pour cette valeur en effet, w(f) est très inférieur à la valeur prévue par la fonction définie ci-dessus. Cette chute importante est probablement une des conséquences du choix du nombre de groupes et du découpage en régions.

TABLEAU 1

Résumé des caractéristiques des 11 groupes
obtenus par l'algorithme MORIN-SANKOFF

Groupe	Taille	Age moyen	% de femmes	CSL moyen	SCO moyenne	Professions dominantes	Remarques
1	8	55	75	0,29	8 ans	ménagères	-
2	19	28	63	0,60	11 ans	étudiants	CSL très dispersé - 50% d'étudiants de moins de 20 ans.
3	9	37	22	0,76	15 ans	cadres sup.	très homogène.
4	30	40	57	0,23	8 ans	ménagères ouvriers	-
5	6	53	66	0,03	6 ans	ménagères chômeurs	-
6	8	25	12	0,20	10 ans	étudiants	-
7	8	19	75	0,39	10 ans	étudiants chômeurs	-
8	5	22	100	0,71	14 ans	étudiants chômeurs	groupe très homogène
9	6	29	0	0,21	11 ans	ouvriers	homogène / sexe et SCO
10	10	37	40	0,65	15 ans	-	- 50% d'éléments ont un CSL > 0,74. Dispersé / âge.
11	11	59	66	0,32	9 ans	retraités	homogène / âge. dispersé / autres variables.

CSL : coefficient socio-linguistique déterminé par les linguistes [8].
SCO : nombre d'années de scolarité

Classification d'un tableau de contingence

Bien que nous ayons demandé 10 groupes, l'algorithme de Govaert nous fournit 9 classes assez homogènes par rapport aux variables linguistiques (voir tableau 2).

Algorithme de vraisemblance du lien (A.V.L.)

Nous avons considéré l'arbre de classification en examinant les niveaux et noeuds significatifs. Nous obtenons 12 classes (voir tableau 3).

Comparaison des résultats

Les groupes obtenus par les trois méthodes de classification ne sont pas identiques ; leur nombre diffère : 9, 11, 12. Pour comparer les résultats, nous voulons en quelque sorte mesurer la proximité entre les classifications d'une part, et ensuite visualiser globalement ces proximités. Pour évaluer la distance entre 2 classifications, nous avons repris une idée exposée par I.C. LERMAN quand il utilise l'AVL pour la juxtaposition de tableaux de contingence :

Lorsque le tableau de contingence représente le croisement de classifications, jusqu'à présent on calculait la valeur observée du χ^2 et on acceptait ou refusait

TABLEAU 2

Synthèse des classes obtenues par la MND appliqués aux tableaux de contingence

Groupe	Taille	Age moyen	% des femmes	CSL moyen	SCO moyen	Professions dominantes	Remarques
1	8	53	75	0.5	10	-	CSL dispersé
2	11	41	64	0.15	6	ménagère	très homogène CSL très faible
3	9	49	23	0.81	16	Cadre sup.	homogène
4	11	55	64	0.13	6	retraités	groupe homogène
5	28	31	57	0.34	10	ouvriers et chômeurs	assez homogène
6	14	32	14	0.18	9	ouvriers	homogène
7	14	27	35	0.64	14	étudiants	homogène
8	15	19	66	0.67	12	étudiants	homogène jeune
9	10	60	70	0.21	8	ménagère	assez homogène

CSL : coefficient socio-linguistique déterminé par les linguistes [8].
SCO : nombre d'années de scolarité

TABLEAU 3

Synthèse des classes obtenues par l'AVL.

Groupe	Taille	Age moyen	% des femmes	CSL moyen	SCO moyenne	Professions Dominantes	Remarques
1	15	53	86	0.27	7	ménagère	homogène / âge
2	5	56	60	0.20	7	retraité	très homogène
3	12	26	50	0.41	12	étudiant	CSL jeune ou moyen
4	6	40	66	0.69	13	-	CSL assez élevé et homogène
5	4	51	25	0.22	8	-	SCO étendue
6	16	49	62.5	0.17	7	ménagère	assez homogène / CSL
7	13	40	47	0.75	14	-	CSL homogène
8	13	25	46	0.28	11	étudiant	CSL bas
9	3	37	0	0.30	7	-	-
10	23	24	26	0.38	11	-	jeunes
11	6	22	83	0.76	15	-	très homogène
12	2	-	50	-	-	-	-

CSL : coefficient socio-linguistique déterminé par les linguistes [8].
SCO : nombre d'années de scolarité

l'hypothèse d'indépendance. Mais l'importance du χ^2 observé peut n'être dûe qu'à quelques cases du tableau I.C. LERMAN a donc suggéré d'associer les liens éventuels entre les modalités des caractères dans les deux classifications.

$$q_{rs} = \frac{(\text{résultat observé})_{rs} - (\text{résultat espéré})_{rs}}{\sqrt{(\text{résultat espéré})_{rs}}}$$

r (resp. s) étant ici le r^e (resp. le s^e) groupe de la 1^{ere} classification (resp. la 2^e).

Ceci en fait n'est rien d'autre que la racine carrée de la contribution au χ^2 calculée pour une case du tableau. La somme des q_{rs}^2 est donc égale au χ^2 de contingence.

Nous nous sommes intéressés aux croisements de classes pour lesquels l'indice q_{rs} était supérieur à 2, indiquant un lien important entre les deux classes. Dans les tableaux 4, 5, 6, nous donnons non pas la valeur de cet indice mais les pourcentages d'éléments d'une classe j appartenant aussi à la classe i et réciproquement. Lorsqu'un des deux nombres ne figure pas, c'est que ce pourcentage était inférieur à 15 %, ou ne signifiait pas grand chose étant donné le nombre d'éléments concernés.

TABLEAU 4

Croisement des classifications observées par l'AVL et l'algorithme MORIN-SANKOFF en pourcentage

le χ^2 calculé par le tableau complet est 321,58, et on a $Pr[\chi^2_{110} \leq 135{,}48] = 0{,}95$

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6	Classe 7	Classe 8	Classe 9	Classe 10	Classe 11	Classe 12
Classe 1	37,5 / 20,0				25 / 50	37,5						
Classe 2				33,3			31,5 / 40					
Classe 3			22,2				26,5	22,2				
Classe 4	20 / 40				50	26,5 / 50						
Classe 5	50,0 / 20	33,3 / 40										
Classe 6								62,5 / 38,8	33,3			
Classe 7										87,5 / 30,4		
Classe 8											100 / 83,3	
Classe 9										100 / 26,8		
Classe 10				30 / 50								20 / 100
Classe 11		40				36,3 / 25						

Dans le croisement des classifications Morin-Lerman, si nous nous intéressons aux intersections pour lesquelles l'indice q_{rs} est supérieur à 2, cela concerne 56 individus sur les 120 à classifier. Ce nombre est de 51 pour le croisement des classifications MORIN-GOVAERT et de 58 pour LERMAN-GOVAERT. Ce nombre est en quelque sorte l'effectif des formes fortes pour les deux classifications impliquées.

TABLEAU 5

Croisement des classifications observées par les algorithmes MORIN-SANKOFF et GOVAERT en pourcentage
le χ^2 calculé pour le tableau complet est 186,01 et Pr $(\chi^2_{80} \leqslant 101{,}87) = 0{,}95$

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6	Classe 7	Classe 8	Classe 9	Classe 10	Classe 11
Classe 1	37,5 / 37,5									25 / 20	
Classe 2				72,7 / 26,5							
Classe 3			44,4 / 44,4							33,3 / 30	
Classe 4					27,3 / 50	25					45,5 / 45,5
Classe 5					33,3	21,4 / 75	62,5				
Classe 6				57,1 / 26,5					28,5 / 66,7		
Classe 7		28,5 / 21	21,4 / 33,3						33,3		
Classe 8		40 / 31,5					20 / 37,5	26,6 / 80			
Classe 9	30 / 37,5										

TABLEAU 6

Croisement des classifications observées par l'AVL et l'algorithme de GOVAERT en pourcentage
le χ^2 calculé pour le tableau complet est 199 et Pr $[\chi^2_{89} \leqslant 110{,}89] = 0.95$

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6	Classe 7	Classe 8	Classe 9	Classe 10	Classe 11	Classe 12
Classe 1	37,5 / 20				25							
Classe 2			36,4 / 33,3		50	27,3						
Classe 3							77,8 / 46,6					
Classe 4						63,5 / 43,8						
Classe 5				50				28,5 / 61,5	100			
Classe 6										71,4 / 43,4		100
Classe 7								21,4 / 23		35,7 / 21,8		
Classe 8			26,6 / 33,3				26,6 / 26,6				26,6 / 66,7	
Classe 9	30 / 20					50 / 31,3						

Pour les tableaux 4, 5, 6, seules sont considérées les cases où l'indice de Lerman est supérieur à 2 et le pourcentage est supérieur ou égal à 20.

i \ j · pourcentage d'éléments de la classe j se trouvant dans la classe i

j \ i pourcentage d'éléments de la classe i se trouvant dans la classe j

Comme il nous paraissait important de visualiser globalement les résultats, nous avons réalisé une analyse des correspondances sur le tableau disjonctif complet 120 x 32 où les 32 variables sont ici, les 32 groupes obtenus dans les trois classifications. Les variables supplémentaires introduites sont le sexe, l'âge, le coefficient socio-linguistique (CSL), la catégorie socio-professionnelle et le nombre d'années de scolarité.

La figure 1 donne la représentation des individus et des variables dans le plan factoriel 1-2.

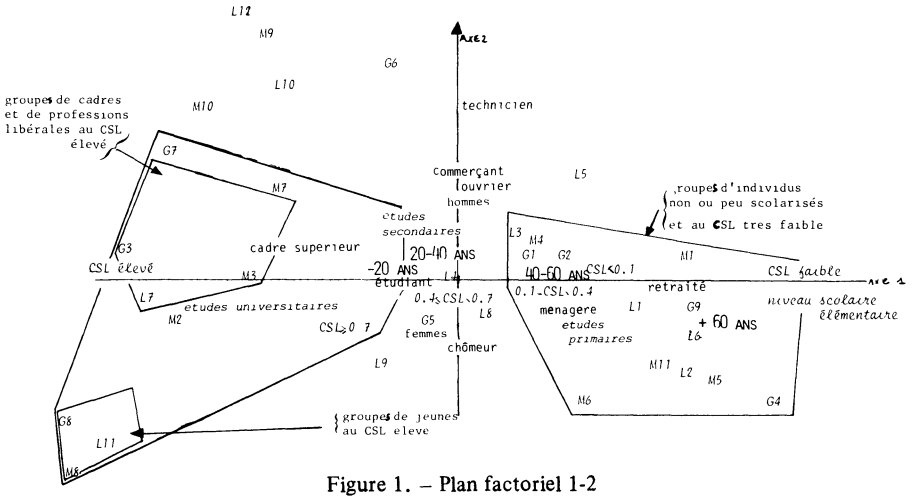

Figure 1. – Plan factoriel 1-2

- le $i^{ème}$ groupe obtenu par l'algorithme MORIN-SANKOFF est noté Mi, $1 \leqslant i \leqslant 11$.
- le $i^{ème}$ groupe obtenu par l'algorithme de LERMAN est noté Li, $1 \leqslant i \leqslant 12$.
- le $i^{ème}$ groupe obtenu par l'algorithme de GOVAERT est noté Gi, $1 \leqslant i \leqslant 9$.

Les dix premiers axes expriment 61 % de la variance et dans [6] nous avons considéré les résultats dans les plans engendrés par les 4 premiers facteurs.

Le second axe est moins clairement défini par les variables sociologiques. On distingue cependant trois groupes importants : un premier groupe constitué d'individus cadres supérieurs, formation universitaire, un second groupe de jeunes et étudiants au CSL élevé, et un troisième groupe formé d'individus plutôt agés au CSL faible, en majorité retraités et ménagères.

Nous avons quelques cas d'associations bizarres, inexplicables en particulier celle de ce policier de 40 ans au CSL moyen et d'une ménagère de 54 ans au CSL nul ! quelle que soit la méthode, ces deux individus sont toujours dans le même groupe ou bien même ils peuvent constituer un groupe à eux deux (AVL).

V. CONCLUSION

Notre objectif n'est pas de privilégier une méthode parmi les trois adoptées. Les distances utilisées sont différentes. Signalons cependant une grande cohérence des résultats (cf. les tableaux 4, 5, 6) même si les classes ne sont pas toujours sociologiquement déterminées. Quelle que soit la méthode, le groupe d'individus à l'in-

dice linguistique fort est bien déterminé, le groupe d'individus peu scolarisés l'est aussi. Ce sont les deux extrêmes de notre échantillon ! reste à savoir si le vocabulaire est déterminant sociologiquement pour la classe moyenne. Nous devons ici signaler la différence de notre démarche avec celle de H. LAZARE et STEINBERG [3] qui utilisent seulement un sous-ensemble du vocabulaire pour faire une analyse des correspondances. Les résultats que nous obtenons sont moins "spectaculaires" qu'attendus en ce sens que nous ne sommes pas arrivés à une classification sociale rigoureuse. Mais pouvions-nous espérer mieux avec un corpus de 8000 mots pour 120 individus et environ 10^6 occurences au total ?

ANNEXE

Description des individus interviewés selon leur niveau d'études achevées (SCO) et leur coefficient socio-linguistique (CSL)

FEMMES

	SCO / CLS	< primaire	primaire	secondaire	supérieur	Total
Professions libérales Cadres supérieurs 4	faible					0
	moyen			4		4
	bon					0
Chômeurs 5	faible	0	1	2		3
	moyen			1		1
	bon				1	1
Employés Ouvrières 1	faible					0
	moyen			1		1
	bon					0
Ménagères 33	faible	3	6	10		19
	moyen	1		7	2	10
	bon			4		4
Retraitées 3	faible		1	1		2
	moyen			1		1
	bon					0
Etudiantes 15	faible			4		4
	moyen			4		4
	bon			2	5	7
	Total	4	8	41	8	61

HOMMES

	SCO / CSL	< primaire	primaire	secondaire	supérieur	Total
Professions libérales Cadres Supérieurs 23	faible	1	2	4	2	9
	moyen			3	2	5
	bon		1		8	9
Chômeurs 9	faible		3	5		8
	moyen			1		1
	bon					0
Employés - Ouvriers 5	faible			1		1
	moyen			2		2
	bon				2	2
Retraités 9	faible	1	2	4		7
	moyen			1	1	2
	bon					0
Etudiants 13	faible					0
	moyen			8	1	9
	bon			1	3	4
	Total	2	8	30	19	59

AGE DES INDIVIDUS INTERVIEWES SELON LE SEXE

	- de 20 ans	20-40 ans	40-60 ans	+ de 60 ans
Femmes 61	19	19	14	9
Hommes 59	15	17	21	6

Note : Dans le texte précédent, la variable SCO représente le nombre d'années de scolarité. Pour simplifier la description des individus, nous avons transformé SCO en variable qualitative à 4 classes : < primaire pour 3 années d'études ou moins, primaire pour les individus ayant de 4 à 6 années d'études, secondaire de 7 à 12 années d'études, supérieur sinon.

BIBLIOGRAPHIE

[1] E. DIDAY et J.C. SIMON. – 1976, *Clustering Analysis*, Springer-Verlag, Ed. K.S. Fu.

[2] G. GOVAERT. – 1980 *Algorithme de Classification d'un tableau de contingence*, Doc. Int. INRIA Rocquencourt.

[3] H. LAZARE et H. STEINBERG. – 1973, *Les feuilletons du Petit Journal de 1890 à 1894*, dans la Pratique de l'Analyse des Données, T3, J.B. BENZECRI *et coll.* (paru en 1981).

[4] I.C. LERMAN. – 1981, *Classification et Analyse Ordinale des Données*, Dunod.

[5] V.J. LUMESKY. – 1982, *A combined algorithm for weighting the variables and clustering in the clustering problems,* Pattern Recognition, vol 15, n° 2, p. 53-60.

[6] A.M. MORIN. – 1981, *Recherche d'une métrique optimale en Analyse de données. Application en lexicométrie*, Thèse de 3e cycle, Université de Rennes 1.

[7] A.M. MORIN et D. SANKOFF. – 1978, *A weighting function for the comparison of word-frequency diestribution*, Doc. Int. C.R.M.A. Université de Montréal, n° 838.

[8] D. SANKOFF. – 1978, *Linguistic variation. Models and methods*, Ac. Press.

[9] D. SANKOFF, G. SANKOFF. – 1973, *Sample survey methods and computer assisted analysis in the study of grammatical variation*, in Canadian Languages in their social context édité par R. DARNELL, p. 7-64. Edmonton Linguistic Research.

[10] D. SANKOFF, G. SANKOFF, S. LABERGE et M. TOPHER. – 1976, *Méthodes d'échantillonnage et utilisation de l'ordinateur dans l'étude de la variation grammaticale*, Cahiers de linguistique de l'Université du Québec, Vol. 6, p. 85-125.

[11] D. SANKOFF, R. LESSARD, N.B. TRUONG. – 1978, *Computational linguistics and statistics in the analysis of the Montréal French Corpus*, Computers and the humanities, Vol. 11, p. 185-191.

[12] G.S. SEBESTYEN. – 1962, *Decision making process in pattern recognition*, Mac MILLAN Company.