SPAD.N

LOGICIEL POUR L'ANALYSE STATISTIQUE DES DONNÉES

Alain Morineau
C.I.S.I.A.
1 avenue Herbillon 94160 Saint-Mandé
Tél (1) 43 74 20 20 Fax (1) 43 74 17 29

Introduction

Le logiciel SPAD.N est dévolu à l'analyse statistique des données numériques. Conçu pour être autonome, il contient tout d'abord les outils traditionnels de la description statistique (comme les tris à plat, histogrammes, moyennes, écarts types, extrema), les techniques de la statistique classique (comme la régression multiple, l'analyse discriminante) et de nombreuses possibilités de recodages des variables. Il met en œuvre enfin les techniques de l'analyse exploratoire multidimensionnelle: analyse en composantes principales, analyse des correspondances simples et des correspondances multiples, combinées aux procédures de classification.

La sélection des méthodes et des algorithmes répond à des préoccupations imposées par le type et le volume des données que le logiciel traite: en général plusieurs milliers de lignes et plusieurs centaines de colonnes. Un champ d'application important est le dépouillement d'enquêtes (de type socio-économique, médical, marketing, etc). Un autre domaine est constitué par les données industrielles, souvent volumineuses.

Le logiciel, conçu avec un souci permanent de portabilité, fonctionne sur tout gros système muni d'un compilateur FORTRAN 77, sur les stations de travail SUN, APOLLO, VAX et RISK 6000. La *version 2.0* du logiciel est fournie sur Macintosh avec un environnement conversationnel complet pour le lancement du logiciel et la visualisation des résultats. Sur mico ordinateur PC la *version 2.0* est un *intégré* contenant un environnement à "menus déroulants" pour piloter le logiciel, un module graphique interactif et un éditeur autonome.

Dans la même famille de logiciels, SPAD.N est complété par le logiciel SPAD.T consacré au traitement statistique des données textuelles. La *"philosophie"* des deux logiciels est la même, les traitements et les résultats sont compatibles. Un champ d'application privilégié est l'analyse des réponses libres aux questions ouvertes d'une enquête, complétant l'étude des questions fermées (qualitatives et quantitatives).

Le logiciel est organisé de façon modulaire en procédures qui s'enchaînent pour réaliser les analyses statistiques. On dresse ici la liste de ces procédures, regroupées par type de fonction et accompagnées d'une description sommaire de leur rôle

Communication

ARDIC	*lecture et archivage du dictionnaire des variables*
ARDON	*lecture et archivage des données*
BIFOR	*exportation de fichiers de résultats*
ASCII	*exportation simplifiée*
ARMDP	*interface avec BMDP*
ARSAS	*lecture directe d'une base SAS*

Gestion

ESCAL	*recodages et archivages*
FREGA	*découpage en fréquences égales*
REDRE	*calcul de coefficients de redressement*
SELEC	*sélections préalables à une analyse*

Analyses factorielles

COPRI	*composantes principales*
CORBI	*correspondances simples*
CORMU	*correspondances multiples*
COREM	*corresp. multiples avec choix des modalités*
CORCO	*correspondances multiples conditionnelles*

Classifications

RECIP	*classification hiérarchique directe*
SEMIS	*classification mixte pour grand tableau*
PARTI	*partition par coupure d'arbre*
TYTRA	*observation et typologie de trajectoires*
ARMAC	*archivage d'une matrice de contiguïté*
CAMAC	*classification sous contrainte de contiguïté*

Description

DEMOD	*caractérisation de variables nominales*
DECLA	*caractérisation d'une partition*
DESCO	*caractérisation d'une variable continue*
DEFAC	*caractérisation des axes factoriels*
GRAPH	*représentations graphiques planes*
POSIT	*positionnement graphique a posteriori*
STATS	*statistiques usuelles des variables*
TABLE	*commandes de tabulations*

Ajustements linéaires

VAREG	*régressions multiples et analyses de variance*
DIS2G	*analyse discriminante à 2 groupes*
FUWIL	*sélection des ajustements optimaux*
SCORE	*création et étude d'une fonction score*

1. Procédures de communication

Ces procédures assurent la communication entre SPAD.N et l'extérieur : récupération de données ou d'informations provenant de l'extérieur, ou inversement transfert vers l'extérieur de résultats produits par le logiciel. Les procédures d'archivage ARDIC et ARDON lisent les libellés des variables et les données numériques, appelés fichiers-sources, pour produire leurs équivalents, mais adaptés à la suite de l'analyse : les fichiers-archives. Pour assurer la re-lecture des fichiers de résultats par tout autre logiciel, BIFOR transforme ces fichiers en fichiers de type texte (appelés aussi fichiers "ASCII"). La procédure ASCII est une version simplifiée de la procédure BIFOR. Pour les versions sur micro-ordinateur, les procédures ARMDPet ARSAS assurent la communication avec les logiciels BMDP® et SAS®.

ARDIC (archivage du dictionnaire)

La richesse et la lisibilité des tableaux de résultats et des graphiques fournis par SPAD.N sont largement fonctions de la précision des libellés qui décrivent les variables et identifient des individus. SPAD.N permet d'introduire des libellés qui ont jusqu'à 60 caractères. L'ensemble des libellés constitue le *dictionnaire* des variables. La procédure ARDIC effectue la lecture (ou la création automatique) de ce dictionnaire. Elle fournit un listage complet des libellés afin de disposer d'un document de référence sur les variables.

ARDON (archivage des données)

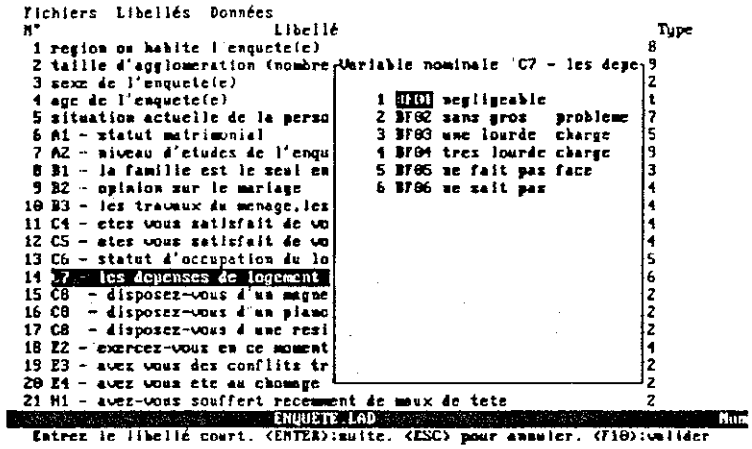
Cette procédure effectue la lecture du fichier des données numériques de l'utilisateur. Les fichiers lus sont des fichiers 'texte' (ou fichiers ASCII) pouvant avoir des formes très diverses. On peut lire en effet des fichiers écrits en *format fixe* (en spécifiant le format de lecture), aussi bien que des fichiers écrits en *format libre* (si les valeurs sont séparées par des blancs). La procédure gère automatiquement les données manquantes éventuelles et les répertorient pour les analyses statistiques ultérieures.

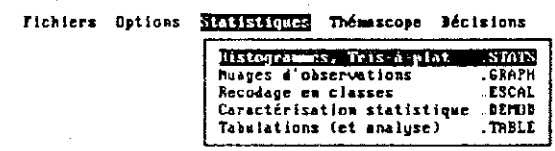
La procédure édite le maximum et le minimum de chaque variable, afin de contrôler les données. On vérifie ainsi que les variables nominales sont correctement codées, et que les variables quantitatives ne présentent pas de valeurs aberrantes. Il est possible d'obtenir le listage de tout ou partie du fichier, afin de vérifier la concordance entre les données fournies et celles qui sont enregistrées.

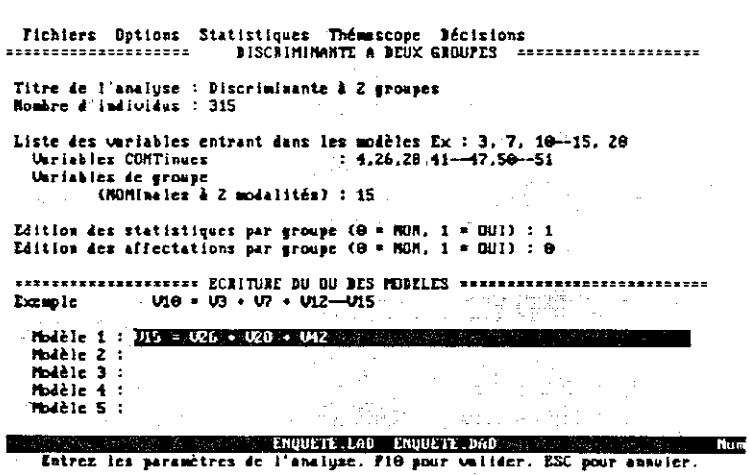
BIFOR (exportation de fichiers de résultats)

Cette procédure permet de transformer les fichiers de résultats de SPAD.N en fichiers texte ou fichiers "avec format", donc récupérables par l'utilisateur. On l'utilisera pour récupérer les fichiers des coordonnées factorielles (repérés par le mot clé NGUS), les fichiers de libellés (NDIC et NDICA) et de données (NDON et NDONA). Le format d'écriture peut être choisi par l'utilisateur ou défini automatiquement par le programme. L'utilisateur peut sélectionner tout ou partie des données à écrire sur ces fichiers.

Cette procédure est donc un outil très souple de communication entre SPAD.N et d'autres logiciels ou d'autres machines. On pourra par exemple récupérer aisément les coordonnées factorielles issues d'une analyse multidimentionnelle pour effectuer des représentations graphiques tridimentionnelles à l'aide d'un logiciel spécialisé. On pourra aussi bien récupérer les coordonnées factorielles de variables (ou de modalités de variables) sur lesquelles on souhaite réaliser (à l'aide de SPAD.N) une typologie.







Trois écrans d'utilisation de SPAD N intégré (Version 2.0 pour PC)

ASCII (création de fichiers "avec format")

La procédure ASCII est une version simplifiée de BIFOR transcrivant de façon automatique en fichiers "textes" les fichiers de résultats fournis par SPAD.N.

ARMDP (interfaçage avec BMDP®)

Pour les versions PC la procédure ARMDP permet de communiquer directement avec le logiciel BMDP: pour y entrer en sortant de SPAD.N, ou pour entrer dans SPAD N en sortant de BMDP (référence *Petitjean Roget)*.

ARSAS (lecture d'une base SAS®)

Cette procédure permet de prendre directement en entrée dans SPAD.N une base SAS existante. Elle crée, à partir de ce fichier, les fichiers de type NDICA et NDONA utilisés par SPAD.N avec sélection éventuelle des éléments à conserver. On peut également créer à partir d'une base SAS un fichier de libellés des variables au format de SPAD.N, pour l'enrichir avant d'entrer dans le logiciel. On pourra alors profiter de toutes les possibilités d'édition de SPAD.N.

2. Procédures de gestion

Les procédures de gestion permettent de manipuler et de transformer les fichiers de données, suivant les besoins de l'analyse à effectuer. Le plus souvent ce sont des procédures qui créent de nouvelles variables s'ajoutant aux variables existant dans le tableau des données. On les appelle dans ce cas procédures de recodage.

La procédure essentielle de recodage est ESCAL. Cette procédure possède un véritable langage pour la création de variables ainsi que des fonctions spécifiques pour le traitement des grands tableaux et des données d'enquête. La procédure FREGA effectue le découpage d'un ensemble de variables en classes d'effectifs égaux avec création automatique des libellés. La procédure REDRE calcule des coefficients de pondération permettant de redresser un échantillon suivant un ou plusieurs critères. On pourrait classer ici la procédure TABLE, décrite plus loin, qui calcule et archive des empilements et juxtapositions de tableaux croisés et de tableaux de moyennes. Ces tableaux sont eux-mêmes susceptibles de subir de multiples traitements statistiques au sein de SPAD.N.

La procédure SELEC est une étape de gestion particulière. Elle opère une sélection sur les fichiers-archives pour préparer les données en vue d'une analyse statistique. Suivant le type d'analyse, on sera amené à sélectionner une partie de l'échantillon des individus, à définir des groupes de variables (homogènes par exemple vis-à-vis d'un thème particulier), ou encore à choisir parmi les données une variable de pondération. A ce titre SELEC est une procédure pivot au cœur du logiciel.

ESCAL (recodages et archivages)

La procédure ESCAL a une double fonction de création de nouvelles variables. Elle permet d'une part d'archiver les résultats d'analyses (par exemple des coordonnées factorielles ou une classification) afin d'enrichir les données initiales avec ces nouvelles variables. Elle permet également de créer de nouvelles variables à partir des variables existantes, en utilisant une large gamme d'opérateurs de diverses natures.

ESCAL permet de réaliser une très grande variété de recodages, dont certains sont spécifiques aux analyses de données. L'utilisateur dispose des opérations arithmétiques, des opérateurs logiques, des parenthèses, des fonctions classiques (MIN, MAX, LOG, SQRT, moyenne, variance), ainsi que de fonctions particulières comme le codage disjonctif automatique.

La disponibilité de fonctions de tirage au hasard permet l'introduction de perturbations aléatoires dans les données, en particulier pour l'étude de la stabilité des résultats (valeurs propres, directions factorielles, coordonnées, graphiques, etc). On peut également créer des variables purement aléatoires (continues ou nominales) s'ajoutant aux données et servant de référence et d'élément de comparaison avec les variables observées.

L'utilisateur a la possibilité de définir des tables pour gérer le découpage des variables continues ou pour créer de nouvelles variables par croisement. La souplesse et la richesse des opérations de recodage tient au fait que l'utilisateur écrit lui-même les instructions à l'aide d'un *langage* simple. Le même langage est utilisé pour sélectionner les individus à l'aide de *filtres* logiques, simples ou complexes. La procédure lit les commandes écrites par l'utilisateur, puis exécute les instructions après contrôle

Les opérations d'archivage réalisées par ESCAL sont assimilables aux opérations de recodage. En transférant des classifications ou des axes de coordonnées factorielles on enrichit le tableau de données avec des informations réutilisables dans les traitements statistiques ultérieurs, qu'ils soient internes ou externes à SPAD.N. Les transferts sont sélectifs: l'utilisateur peut préciser les éléments à copier.

FREGA (découpage en classes de fréquences égales)

La procédure FREGA effectue le découpage d'une variable en un nombre de classes choisi par l'utilisateur, en assurant un effectif égal d'individus dans chaque classe quand cela est possible. La procédure traite simultanément plusieurs variables et engendre automatiquement l'ensemble des libellés des variables et des modalités créées. La facilité d'utilisation permet de tester aisément des découpages alternatifs.

REDRE (calcul de coefficients de redressement)

Cette procédure crée une variable de pondération des individus destinée à "redresser" un échantillon. Le poids, ou coefficient de redressement, est calculé pour assurer une distribution imposée dans une ou plusieurs variables nominales du fichier. On assurera par exemple une répartition imposée sur le sexe, les catégories d'âges et les zones géographiques.

Il n'y a pas de contrainte sur les facteurs entrant dans le calcul des poids. Le calcul est itératif. Les distributions obtenues sont en général très proches des distributions demandées, et la somme des poids est ajustée *in fine* à la taille de l'échantillon.

SELEC (sélection)

Cette procédure est préalable à toute étape d'analyse (les procédures d'archivage ou de recodage comme ESCAL, TABLE, REDRE ou FREGA exceptées). Il s'agit de retenir, parmi les données qui ont été archivées, celles qui seront *utiles* pour la suite du traitement statistique.

SELEC permet de désigner les variables et, éventuellement, les individus à retenir. Le choix des variables s'effectue de façon simple en annonçant les numéros des variables

à retenir, et les catégories dans lesquelles on mettra ces variables ("actives" ou "illustratives" par exemple). On choisit de plus la variable qui sera éventuellement utilisée pour pondérer les calculs à venir.

IDENT	MODALITES LIBELLE	AVANT APUREMENT EFF.	AVANT APUREMENT POIDS	APRES APUREMENT EFF.	APRES APUREMENT POIDS	HISTOGRAMME DES POIDS RELATIFS
	1 . la famille est le seul endroit où l'on se sente bien					
fbi1	- oui	561	561.00	567	567.00	**********
fbi2	- non	431	431.00	433	433.00	********
fbi*	- non-réponse	8	8.00	== VENTILEE ==		
	2 . opinion à propos du mariage					
opm1	- union indissoluble	231	231.00	231	231.00	*******
opm2	- dissout si pb. grave	342	342.00	342	342.00	**********
opm3	- dissout si accord	387	387.00	387	387.00	**********
opm4	- ne sait pas	39	39.00	40	40.00	***
opm*	- non-réponse	1	1.00	== VENTILEE ==		
	3 . à qui incombent les travaux ménagers et les soins enfants ?					
esc1	- incombent à la femme	42	42.00	50	50.00	****
esc2	- plutôt à la femme	336	336.00	347	347.00	**********
esc3	- homme et femme	599	599.00	603	603.00	**********
esc4	- ne sait pas	19	19.00	== VENTILEE ==		
esc*	- non-réponse	4	4.00	== VENTILEE ==		
	6 . les découvertes scientifiques améliorent-elles la vie ?					
sci1	- oui, un peu	509	509.00	509	509.00	**********
sci2	- oui, beaucoup	383	383.00	384	384.00	**********
sci3	- pas du tout	105	105.00	107	107.00	*******
sci*	- non-réponse	3	3.00	== VENTILEE ==		
	7 . comparée aux personnes de votre âge, votre santé est					
san1	- très satisfaisante	267	267.00	274	274.00	**********
san2	- satisfaisante	600	600.00	602	602.00	**********
san3	- peu satisfaisante	115	115.00	124	124.00	********
san4	- pas du tout satisf.	18	18.00	== VENTILEE ==		
san*	- non-réponse	0	.00			
	8 . évolution du niveau de vie de l'enquêté depuis 10 ans					
niv1	- beaucoup mieux	102	102.00	102	102.00	*******
niv2	- un peu mieux	316	316.00	317	317.00	**********
niv3	- c'est pareil	250	250.00	250	250.00	**********
niv4	- un peu moins bien	190	190.00	190	190.00	**********
niv5	- beaucoup moins bien	114	114.00	115	115.00	*******
niv6	- ne sait pas	26	26.00	26	26.00	**
niv*	- non-réponse	2	2.00	== VENTILEE ==		
	9 . opinion sur le fonctionnement de la justice en 1979					
jus1	- très bon	13	13.00	== VENTILEE ==		
jus2	- assez bon	243	243.00	245	245.00	**********
jus3	- assez mauvais	398	398.00	399	399.00	**********
jus4	- très mauvais	256	256.00	257	257.00	**********
jus5	- ne sait pas	65	65.00	70	70.00	*****
jus6	- ne veut pas répondre	25	25.00	29	29.00	**
jus*	- non-réponse	0	.00			

Figure 1 : dictionnaire de variables sélectionnées pour des Correspondances Multiples (CORMU)

Il est possible de *filtrer* les individus, c'est-à-dire de les choisir en fonction des valeurs prises par certaines variables. Les opérations logiques permettant de sélectionner les individus sont très souples et sans limitation.

3. Les procédures d'analyse factorielle

Les procédures d'analyses factorielles sont des outils de statistique descriptive permettant d'étudier simultanément les relations entre variables et les ressemblances entre les unités statistiques appelées individus. Les procédures décrites ici effectuent le cœur des calculs de ces analyses, d'autres procédures du logiciel étant consacrées

à l'exploitation de leurs résultats représentations graphiques, aides à l'interprétation des facteurs, validations, analyses de type "confirmatoire".

L'analyse en Composantes Principales (COPRI) décrit les liaisons entre variables ***continues*** *(dites aussi quantitatives) alors que l'analyse des Correspondances Multiples (CORMU, COREM, CORCO) décrit les liaisons entre variables* ***nominales*** *(appelées encore variables qualitatives). Dans les deux cas toute autre variable connue sur les mêmes individus peut être introduite dans les analyses au titre d'élément* ***illustratif*** *(on dit aussi élément supplémentaire). Enfin l'analyse des correspondances simple ou "Bnaire" (CORBI) étudie les distributions de* ***fréquences*** *en ligne et colonne dans un tableau de contingence, ou plus généralement pour tout tableau de nombres non négatifs*

La détermination des élément dits ***actifs*** *pour réaliser une analyse est un choix important que doit faire l'utilisateur. Ce choix doit satisfaire à certaines conditions dont les principales sont l'homogénéité des variables (elles doivent appartenir à un même point-de-vue ou thème) et l'exhaustivité (elles doivent décrire complètement ce thème). Pour ces raisons les méthodes d'analyse factorielles, associées aux méthodes de classification, constituent un instrument d'observation statistique qu'on appelle* ***"thémascope"***.

Les coordonnées factorielles de tous les éléments entrant dans l'analyse sont consignées dans un fichier repéré par le mot clé NGUS.

COPRI (composantes principales)

Cette procédure effectue l'analyse en composantes principales d'un ensemble de variables continues. On peut effectuer soit une analyse *normée* (analyse de la matrice des corrélations entre variables), soit une analyse *non normée* (analyse de la matrice des covariances).

L'analyse, réalisée sur un groupe de variables continues *actives*, permet d'introduire en éléments *illustratifs* des variables continues aussi bien que des variables nominales. Les calculs seront réalisés en tenant compte éventuellement d'un poids de redressement et pourront être restreints *par filtre* à un sous ensemble de l'échantillon. Le programme gère automatiquement les données manquantes éventuelles.

Le listage des résultats fournit les statistiques usuelles sur les variables analysées (moyennes, écarts types, extrema) ainsi que la matrice de corrélations ou la matrice de covariances. On édite ensuite toutes les valeurs propres de l'analyse et le tracé de l'histogramme de décroissance des valeurs afin d'en étudier la forme. On peut évaluer l'importance des écarts entre valeurs propres à l'aide du graphique des intervalles de confiance asymptotiques calculés pour chaque valeur propre.

L'utilisateur peut demander une nouvelle édition de la matrice de corrélations ou de covariances où les variables sont rangées dans l'ordre de leurs coordonnées sur le premier axe factoriel. Dans le cas fréquent d'un *facteur taille*, cette édition range les variables dans la matrice de telle sorte que la proximité entre variables est à l'image de la corrélation.

Pour un individu (ou ligne du tableau des données), le programme édite son poids relatif, sa distance à l'origine, ses coordonnées sur un nombre quelconque d'axes factoriels, les contributions (dites absolues) et les cosinus carrés (ou contributions relatives). Ces éditions sont faites à la demande pour les individus actifs et supplémentaires.

Pour les variables actives et illustratives, on édite les coordonnées sur un nombre quelconque d'axes, les corrélations variable-facteur (dont le carré est l'analogue des contributions relatives), et les projections des anciens axes unités des variables actives (dont le carré est l'analogue des contributions absolues).

HISTOGRAMME DES 21 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	.2346	8.94	8.94	**********
2	.1789	6.81	15.75	*******
3	.1551	5.91	21.66	******
4	.1507	5.74	27.40	******
5	.1414	5.39	32.79	*****
6	.1392	5.30	38.09	*****
7	.1334	5.08	43.17	*****
8	.1322	5.04	48.21	*****
9	.1256	4.78	52.99	*****
10	.1246	4.75	57.74	*****
11	.1186	4.52	62.26	*****
12	.1165	4.44	66.69	*****
13	.1111	4.23	70.93	*****
14	.1099	4.19	75.11	*****
15	.1065	4.06	79.17	*****
16	.1043	3.97	83.14	*****
17	.1026	3.91	87.05	*****
18	.0979	3.73	90.78	*****
19	.0927	3.53	94.31	*****
20	.0830	3.16	97.47	*****
21	.0664	2.53	100.00	*****

Figure 2 : Décroissance des valeurs propres

Les modalités des variables nominales sont traitées comme centres de gravité des individus qui les composent et constituent des individus illustratifs fictifs. On édite pour chaque modalité ses coordonnées sur les axes ainsi qu'un critère appelé *valeur-test* servant à évaluer sur l'axe la distance à la moyenne générale (en nombre d'écarts types d'une loi normale). La valeur-test mesure donc l'intérêt du point illustratif représentant la modalité sur l'axe.

La procédure effectue les calculs principaux de l'analyse en composantes principales mais ne fournit pas les graphiques ni certaines aides spécifiques pour les interprétations. Le fichier des coordonnées factorielles (noté NGUS) créé par COPRI permet de transporter l'ensemble des résultats vers les procédures d'aide: GRAPH pour réaliser une grande variété de graphiques factoriels, DEFAC pour séléctionner les éléments caractéristiques sur chaque axe.

La procédure ESCAL permet d'archiver les coordonnées factorielles au sein des données initiales, autorisant ainsi tout autre traitement statistique ultérieur par SPAD.N. Les procédures BIFOR et ASCII exportent ces résultats en fichiers "textes" Dans la version PC du logiciel, la procédure ASCII communique les résultats au module graphique interactif.

CORBI (correspondances binaires)

La procédure CORBI réalise l'analyse des correspondances d'un tableau de contingence (ou tableau de fréquences), ou de tout tableau de valeurs non négatives (comme les tableaux disjonctifs, complets ou non). S'il s'agit de fréquences, le tableau d'entrée pourra être un tableau de fréquences déjà connues, ou un tableau créé par le

logiciel lui-même à partir de croisements de plusieurs variables nominales observées sur des individus (procédure TABLE).

Dans tous les cas, on gère de façon simple les lignes et colonnes actives et illustratives du tableau avec les outils usuels du logiciel (procédure SELEC). On peut également positionner en éléments illustratifs les modalités de variables nominales. L'analyse imprime les valeurs propres et l'histogramme de leur décroissance. Pour les très grands tableaux, on peut opter pour un algorithme efficace de calcul des seules premières valeurs propres (par "approximation stochastique").

Le logiciel édite sur un nombre quelconque d'axes les coordonnées des lignes et des colonnes du tableau (actives et illustratives), les cosinus carrés (ou contributions relatives) et les contributions (dites aussi absolues). On connait pour chaque élément son poids relatif et le carré de sa distance au centre de gravité. Les modalités des variables nominales peuvent être introduites en éléments illustratifs, pour lesquels on calcule les coordonnées et les *valeurs-tests* sur les axes.

La procédure crée un fichier de résultats (repéré par le mot clé NGUS) contenant en particulier les coordonnées factorielles. Ce fichier est utilisé pour interpréter les axes factoriels (procédures DEFAC), tracer une grande variété de graphiques (procédure GRAPH) ou archiver ces résultats avec les données initiales pour d'autres traitements ultérieurs (procédure ESCAL).

Les procédures BIFOR ou ASCII exportent ces résultats en fichiers "textes" pour une exploitation éventuelle par d'autres logiciels. Dans la version PC du logiciel, la procédure ASCII communique les résultats au module graphique interactif.

CORMU (correspondances multiples)

Alors que COPRI réalise l'analyse d'un tableau de variables continues, la procédure CORMU analyse un tableau analogue de variables nominales. Dans le cas d'une enquête par exemple, chaque individu (ligne du tableau) est caractérisé par ses réponses à une série de *questions*. Les questions où l'individu répond en choisissant une modalité de réponse sont des variables nominales. La procédure CORMU, étape essentielle d'analyse descriptive des relations entre plusieurs variables nominales, réalise ce qu'on appelle *analyse des correspondances multiples*.

La sélection des variables nominales actives, des variables illustratives (nominales et continues) ainsi que des individus actifs et illustratifs est réalisée avec les outils usuels (procédure SELEC). Le cas échéant on tiendra compte d'une pondération attribuée à chaque individu. Le programme édite sous une forme très compacte l'ensemble des croisements deux à deux des variables nominales actives: c'est le *tableau de correspondances multiples*, appelé aussi *tableau de Burt*. Dans un même tableau, on peut obtenir sous forme compacte les effectifs et les pourcentages lignes et colonnes.

Pour se prémunir contre les effets des modalités à faible effectif sur le calcul des axes factoriels, le programme met en œuvre une procédure originale de protection ("ventilation" au hasard des réponses concernées). Il suffit d'annoncer le seuil à partir duquel on considère qu'un effectif est trop faible.

Sur les axes demandés, on édite les coordonnées de tous les éléments introduits dans l'analyse (individus actifs et illustratifs, variables nominales actives et illustratives, variables continues illustratives), leurs contributions et leurs cosinus carrés. La localisation des variables continues se fait par calcul des corrélations avec les axes,

avec la même interprétation qu'en analyse en composantes principales. On obtient également le poids relatif et le carré de la distance au centre pour chaque élément.

Les modalités étant centre de gravité des individus qui les composent, on peut associer aux modalités illustratives le critère appelé *valeur-test* qui évalue sur chaque axe la "distance" au centre de gravité en nombre d'écarts types d'une loi normale (test d'une moyenne confondue avec la moyenne générale). Ceci permet d'évaluer si un sous groupe d'individus a une localisation significative dans une direction factorielle (par extension, le critère est évalué aussi pour les modalités actives).

EDITION DU TABLEAU DE CORRESPONDANCES MULTIPLES

	fbi1	fbi2	opm1	opm2	opm3	opm4	esc1	esc2	esc3	sci1	sci2	sci3	san1	san2	san3	niv1	niv2	niv3	niv4	niv5	niv6
fbi1	567	0																			
fbi2	0	433																			
opm1	197	34	231	0	0	0															
opm2	227	115	0	342	0	0															
opm3	128	259	0	0	387	0															
opm4	15	25	0	0	0	40															
esc1	39	11	25	11	12	2	50	0	0												
esc2	215	132	92	155	84	16	0	347	0												
esc3	313	290	114	176	291	22	0	0	603												
sci1	296	213	120	161	209	19	27	167	315	509	0	0									
sci2	219	165	87	157	127	13	17	153	214	0	384	0									
sci3	52	55	24	24	51	8	6	27	74	0	0	107									
san1	136	138	67	80	115	12	8	100	166	132	111	31	274	0	0						
san2	356	246	135	216	232	19	35	201	366	315	231	56	0	602	0						
san3	75	49	29	46	40	9	7	46	71	62	42	20	0	0	124						
niv1	60	42	25	36	39	2	4	42	56	47	50	5	40	57	5	102	0	0	0	0	0
niv2	184	133	75	106	125	11	14	101	202	151	136	30	83	195	39	0	317	0	0	0	0
niv3	140	110	68	84	86	12	16	100	134	129	98	23	73	146	31	0	0	250	0	0	0
niv4	111	79	41	73	66	10	10	63	117	109	62	19	38	125	27	0	0	0	190	0	0
niv5	60	55	21	33	57	4	5	34	76	57	32	26	33	63	19	0	0	0	0	115	0
niv6	12	14	1	10	14	1	1	7	18	16	6	4	7	16	3	0	0	0	0	0	26

Figure 3 : Tableau des correspondances multiples (Tableau de Burt)

Tous les résultats utiles sont copiés dans un fichier (repéré par le mot clé NGUS) et transmis aux procédures d'exploitation graphique (GRAPH), d'aide à l'interprétation des axes (DEFAC), ou d'archivage (ESCAL) permettant tous traitements statistiques faisant intervenir simultanément ces résultats et les données initiales. Les procédures BIFOR ou ASCII assurent la transmission des résultats à tout autre logiciel externe. Dans la version PC du logiciel, la procédure ASCII communique les résultats au module graphique interactif.

COREM (corresp. multiples avec choix des modalités)

Cette procédure réalise comme CORMU l'analyse d'un tableau de variables nominales. Avec CORMU l'utilisateur choisit les *variables* actives de l'analyse, ou plus précisément, en choisissant les variables, il sélectionne l'ensemble des modalités actives de l'analyse. Avec la procédure COREM, l'utilisateur a la possibilité de choisir *une à une* les modalités actives de l'analyse, sans devoir prendre toutes les modalités d'une variable active.

On peut ainsi éliminer non seulement les modalités dont l'effectif est inférieur à un seuil (comme dans CORMU), mais aussi certaines modalités choisies par l'utilisateur, par exemple les modalités "donnée manquante". Cette procédure fournit les mêmes possibilités que CORMU et édite les mêmes résultats. Bien que ne possèdant pas

exactement les propriétés de l'analyse réalisée par CORMU, les résultats s'interprêteront avec des règles de lecture analogues (références *Escofier* et *Benali)*.

CORCO (correspondances multiples conditionnelles)

Cette procédure réalise, comme CORMU ou COREM, l'analyse d'un tableau de variables nominales. Elle sert donc à étudier les inter-relations entre plusieurs variables qualitatives. L'analyse des correspondances multiples appelée *conditionnelle* permet d'étudier ces liaisons en éliminant l'influence d'une autre variable, sous réserve qu'elle soit aussi qualitative.

Si par exemple les observations sont faites quatre années successives, et si les variables nominales observées ont une évolution temporelle importante, l'analyse classique décrira essentiellement leur évolution conjointe dans le temps. Par contre l'analyse avec CORCO, en conditionnant l'étude par la variable "temps" (à 4 modalités ici), décrira les liaisons intrinsèques entre variables, c'est-à-dire les liaisons qui existent indépendamment du temps.

Pour l'essentiel les résultats édités par le programme sont analogues à ceux fournis par CORMU. De plus la procédure permet de réaliser, comme COREM, la sélection directe des *modalités* actives de l'analyse (références *Escofier* et *Benali)*.

4. Les procédures de classification

Les procédures de classification constituent une autre famille d'outils pour l'instrument d'observation statistique dénommé "thémascope". Les analyses factorielles positionnent les objets à décrire les uns par rapport aux autres sur des graphiques, et fournissent donc des représentations spaciales continues. Cependant ces méthodes prennent difficilement en compte des interactions d'ordre élevé. De plus la complexité de certaines structures est telle que des projections dans des sous espaces peuvent être insuffisantes. Les procédures de classification cherchent à regrouper les objets pour définir des groupes homogènes. Une typologie est obtenue quand on a fait de chaque groupe d'objets un "type", une entité dont on connait les caractéristiques. Une typologie est souvent un moyen commode d'observation au delà des premières dimensions d'une analyse factorielle

Les algorithmes de construction des classes travailleront sur les coordonnées factorielles d'une analyse préalable, ce qui présente plusieurs avantages. On assure de cette façon la compatibilité des calculs et donc des résultats car on travaille dans les deux cas sur les mêmes configurations initiales des objets. On peut "lisser" ces configurations et en général obtenir des catégories mieux typées en abandonnant les derniers axes factoriels, souvent porteurs des composantes aléatoires (non systématiques) des données. Enfin on réduit le volume des opérations en travaillant sur les seuls premiers axes et en profitant de leur orthogonalité.

*La classification d'objets en groupes homogènes est une procédure complexe et il est difficilement imaginable qu'elle soit le résultat brut d'un algorithme de calcul. Le logiciel propose non pas un ou plusieurs algorithmes de classifications, mais plusieurs **stratégies**. Une stratégie met en jeu plusieurs algorithmes et combinaisons d'opérations à décider au fur et à mesure par l'utilisateur, en fonction des résultats déjà obtenus. Une des stratégies, SEMIS, est bien adaptée à la classification des très grands ensembles de données; l'autre, RECIP, travaillera sur des ensembles plus petits.*

On notera la possibilité de réaliser une classification sous contrainte de contiguité, c'est-à-dire en respectant le lien de voisinage des individus dans une classe. L'utilisation principale concerne le voisinage géographique (avoir une frontière commune), mais s'étend aisément en déclarant contigus des individus à distance inférieure à un seuil fixé.

Comme pour les autres analyses, les individus à classer peuvent être sélectionnés, soit par liste soit en utilisant des filtres sur les variables. Ils peuvent être munis d'un poids de redressement. Les résultats des classifications sont archivables pour une utilisation ultérieure dans d'autres procédures ou à l'extérieur de SPAD N.

RECIP (classification hiérarchique directe)

Cette procédure construit un arbre d'agrégation hiérarchique (ou dendrogramme) des individus caractérisés par leurs coordonnées factorielles, en utilisant le critère d'agrégation de *Ward*. Ce critère, basé sur la réduction minimale de variance par agrégation, est homogène au critère d'inertie utilisé pour la détermination des axes factoriels. Il possède des propriétés générales qui intéressent la majorité des applications pratiques. L'arbre lui-même est construit avec l'algorithme rapide de recherche en chaîne des voisins réciproques, ou algorithme de *Benzécri*.

CLASSIFICATION ASCENDANTE HIERARCHIQUE

NUM.	AINE	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
31	18	26	2	28.00	.00188	**
32	9	30	2	47.00	.00224	**
33	29	25	2	22.00	.00228	**
34	5	21	2	65.00	.00285	***
35	4	23	2	82.00	.00412	****
36	31	22	3	42.00	.00431	****
37	19	27	2	27.00	.00449	****
38	10	24	2	52.00	.00452	****
39	34	12	3	98.00	.00477	****
40	14	7	2	77.00	.00538	*****
41	40	15	3	101.00	.00622	*****
42	8	6	2	89.00	.00764	******
43	38	32	4	99.00	.00800	******
44	35	13	3	114.00	.00809	*******
45	33	20	3	38.00	.00892	*******
46	16	28	2	40.00	.01013	********
47	39	2	4	178.00	.01085	********
48	46	45	5	78.00	.01599	**********
49	3	36	4	118.00	.01722	**********
50	41	43	7	200.00	.02466	********************
51	44	42	5	203.00	.02587	********************
52	49	17	5	140.00	.02810	**********************
53	51	1	6	340.00	.04208	**********************************
54	48	37	7	105.00	.05092	**
55	11	54	8	142.00	.05823	**
56	52	53	11	480.00	.06308	**
57	55	56	19	622.00	.08793	**
58	50	57	26	822.00	.09571	**
59	47	58	30	1000.00	.10732	**

Figure 4: Histogramme des indices pour le choix de la partition

L'arbre hiérarchique est ici un intermédiaire de calcul vers le choix d'une partition de l'ensemble des objets. En effet les algorithmes de partitionnement direct nécessitent de connaître le nombre des classes finales. En construisant un arbre préalable et en étudiant sa forme, l'utilisateur acquiert une certaine connaissance sur le nombre probable de classes dans la population: les zones denses de points homogènes s'agrègent aux niveaux bas de l'arbre et se séparent en classes naturelles quand les branches de raccordement s'allongent. La procédure RECIP édite donc l'arbre (ainsi

que ses caractéristiques statistiques), pour faciliter le choix des partitions les plus pertinentes.

L'arbre obtenu doit ensuite être "coupé" par l'utilisateur pour créer une ou plusieurs partitions. La procédure PARTI réalise ces coupures et les décrit.

SEMIS (classification mixte pour les grands tableaux)

La procédure SEMIS remplacera RECIP dans le cas des grands tableaux. Elle est en effet plus rapide et nécessite moins de mémoire centrale. Comme RECIP, elle opère sur des coordonnées factorielles, mais ne nécessite pas que le tableau des coordonnées factorielles soit recopié dans la mémoire centrale de l'ordinateur.

Une première étape d'agrégation autour de centres mobiles (type "*k-means*" ou "*nuées dynamiques*") conduit à la construction rapide d'une partition contenant un grand nombre de petits groupes (une centaine par exemple). Ces groupes sont sensés être des morceaux de classes "réelles" que l'algorithme de partitionnement a éclatées.

Pour obtenir d'emblée une partition préalable de bonne qualité, on y intègre une procédure d'*auto-validation*. Celle ci consiste à réaliser plusieurs partitions successives (les "partitions de base") puis à les croiser. On retient comme classes finales les *groupes stables* (appelés aussi "formes fortes") constitués par les groupes d'individus classés ensemble dans les partitions de base.

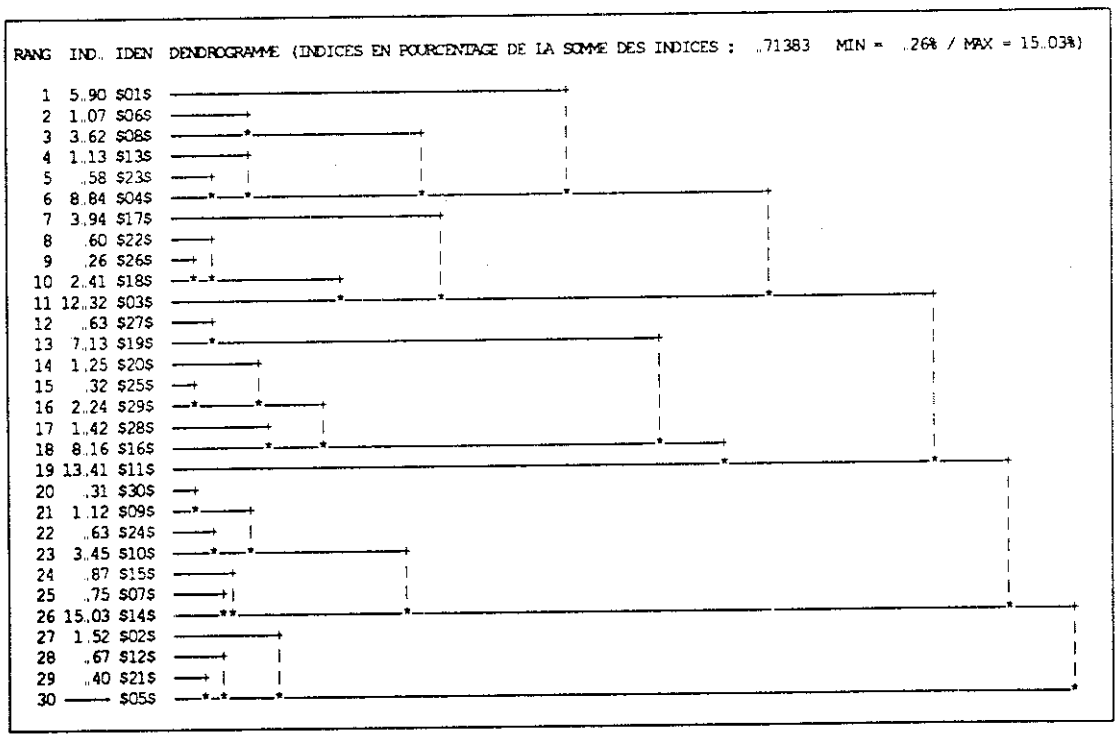


Figure 5 : Arbre de la hiérarchie

Dans une seconde étape, on construit un arbre hiérarchique à partir des centres de ces groupes stables. Cette construction est très rapide car les éléments à agréger sont peu nombreux.

La procédure SEMIS dispose de plusieurs paramètres de commande permettant de piloter de façon souple la construction du dendrogramme: choix du nombre d'axes

factoriels utilisés, choix du nombre de partitions de base à croiser pour déterminer les groupements stables, choix du nombre d'itérations pour construire chaque partition de base, démarrage des itérations par tirage au hasard initial ou par choix a priori des éléments de départ, création ou non d'une classe résiduelle regroupant les individus "non stables" dans les partitions successives. Après quelques essais, l'utilisateur trouve généralement la combinaison des commandes qui fournit l'arbre hiérarchique à retenir.

Comme pour la procédure RECIP, ce dendrogramme est un intermédiaire pour que l'utilisateur choisisse la ou les partitions qui lui paraissent s'imposer. On appellera ensuite la procédure PARTI pour créer la partition finale.

PARTI (partition par coupure de l'arbre d'agrégation)

Cette procédure permet d'effectuer la coupure d'un arbre (créé par exemple par RECIP ou SEMIS) pour obtenir une partition en un certain nombre de classes. En fait on peut réaliser et archiver plusieurs partitions simultanément car il est rare qu'une seule partition de la population s'impose. Les caratérisations statistiques ultérieures (procédure DECLA) ou d'autres considérations permettront d'affiner le choix ensuite.

Une partition définie par coupure d'un arbre est obtenue sous contrainte d'emboitement des partitions. Elle peut donc être améliorée en relachant cette contrainte. Une procédure itérative de calcul, appelée *consolidation*, en option dans PARTI, conduit ainsi à une partition de qualité optimale pour le critère d'homogénéité des classes.

COUPURE 'a DE L'ARBRE EN 4 CLASSES

CONSOLIDATION DE LA PARTITION AUTOUR DES 4 CENTRES DE CLASSES,
REALISEE PAR 10 ITERATIONS A CENTRES MOBILES

	INERTIES		EFFECTIFS		POIDS		DISTANCES	
	AVANT	APRES	AVANT	APRES	AVANT	APRES	AVANT	APRES
INERTIE INTER-CLASSES	.2910	.3522						
INERTIES INTRA-CLASSE								
CLASSE 1 / 4	.4531	.2591	480	339	480.00	339.00	.0752	.1699
CLASSE 2 / 4	.2657	.2594	142	146	142.00	146.00	.5788	.6416
CLASSE 3 / 4	.1530	.1492	200	200	200.00	200.00	.4222	.4894
CLASSE 4 / 4	.1028	.2456	178	315	178.00	315.00	.4956	.3272
INERTIE TOTALE	1.2655	1.2655						

Figure 6. Optimisation de la partition en 4 classes

La procédure permet de lister le contenu des classes, et fournit diverses éditions permettant d'apprécier la nature de la partition finale (inerties, distances à l'origine, coordonnées et valeurs-tests sur les axes).

Un paramètre particulier permet de demander la liste des *parangons* de chaque classe. Les parangons sont les individus les plus proches du centre de gravité de la classe, et à ce titre assimilables à des représentants typiques de la classe. D'un point de vue technique, on pourra par exemple les utiliser pour amorcer les itérations de procédures ultérieures de partitionnement directe.

On peut également demander d'éditer les individus qui ont les plus fortes contributions à l'inertie interne de la classe. Si les poids sont égaux, on obtient ainsi les individus les plus éloignés du centre de la classe.

TYTRA (observation et typologie de trajectoires)

On considère des variables nominales observées par exemple à différentes dates. Leurs modalités (par exemple une catégorie "A" aux dates 1,2,3,4) dessinent avec le temps des trajectoires que l'on retrouve dans les plans d'une analyse factorielle. La procédure TYTRA permet d'observer les formes et les positions de ces trajectoires dans les plans d'une analyse préalable. Le problème est souvent de repérer les catégories ayant des trajectoires analogues.

Une trajectoire est en fait le croisement d'une variable avec les catégories de la variable privilégiée (les dates par exemple). Toutes les trajectoires sont définies de façon automatique, et positionnées immédiatement dans les graphiques.

De plus TYTRA prépare des fichiers internes directement ré-utilisables par les procédures du logiciel. On fera essentiellement appel aux procédures décrites ci-dessus pour réaliser une typologie des formes de trajectoires (soit RECIP soit SEMIS pour hiérarchiser les trajectoires, puis PARTI pour en tirer une ou plusieurs typologies alternatives,et DECLA pour caractériser statistiquement le contenu des classes)

ARMAC (archivage d'une matrice de contiguïté)

Cette procédure est préalable à la procédure de classification sous contrainte de contiguïté CAMAC décrite plus loin. Le plus souvent la contrainte de contiguïté est une contrainte de voisinage géographique (avoir une frontière commune)

Si les individus proches se ressemblent, il est intéressant de chercher des classes homogènes qui ne contiennent que des individus voisins géographiquement.

L'objet de la procédure ARMAC est la lecture d'une matrice dite de contiguïté, matrice qui décrit pour chaque individu la liste de ses voisins contigus. Il s'agira donc souvent d'une contiguïté géographique. Mais la contiguïté peut être définie à partir d'autres considérations. En partant par exemple du calcul d'une distance entre individus, on peut déclarer contigus les individus dont la distance est inférieure à un seuil donné.

Cette matrice de contiguïté est vérifiée et éventuellement corrigée en cas d'erreur logique de définition des voisinages. Un paramètre permet de mettre en œuvre plusieurs définitions de la contiguïté: au premier niveau si on a une frontière commune, au deuxième niveau si on touche un individu qui a une frontière commune, etc.

CAMAC (classification sous contrainte de contiguïté)

Cette procédure réalise une classification hiérarchique sur les lignes d'un tableau, lorsque les individus sont liés par une contrainte de contiguïté. Il s'agit en fait de réaliser le travail décrit pour la procédure RECIP, mais en tenant compte de la contrainte de contiguïté au moment de l'agrégation des individus: on ne peut agréger que des individus contigus.

On notera que l'arbre hiérarchique construit par CAMAC peut présenter des *inversions* c'est-à-dire des incohérences dans le rangement des distances entre points et classes. Un arbre présentant beaucoup d'inversions traduit le fait que le phénomène étudié présente une assez forte hétérogénéité spaciale. Peu ou pas d'inversion caractérise une autocorrélation spaciale du phénomène observé.

L'algorithme permet de borner les effectifs des classes pour éviter l'absorption de tous les éléments par une classe dominante par suite d'un effet de chaîne. Ainsi on peut refuser d'agréger un individu à une classe déjà trop volumineuse. Une classe dont l'effectif dépasse un seuil donné devient une classe finale (ou nœud terminal). Cette procédure qui crée des classes finales peut donc rendre l'arbre *non connexe* (certaines branches ne sont pas reliées au reste de l'arbre).

Lorsque l'on coupe l'arbre pour définir une partition (procédure PARTI) il faut s'assurer, s'il y a des nœuds terminaux, de demander plus de classes que de nœuds terminaux. La partition obtenue sera traitée ensuite par le logiciel de façon classique: caractérisation statistique par DECLA, archivage par ESCAL, positionnement graphique par GRAPH, etc.

5. Les procédures de description

Ces procédures sont des outils de description des individus et des variables disponibles dans le tableau de données. L'objectif commun à tous les traitements proposés est le suivant: caractériser de façon ***rapide*** *et* ***complète*** *des données* ***volumineuses****. Par exemple une seule commande permettra d'éditer tous les tris à plat et tous les histogrammes concernant un sous échantillon. On obtiendra dans le même esprit la caractérisation statistique automatique des classes d'une partition, cette caractérisation faisant intervenir systématiquement toutes les variables disponibles dans les données.*

Dans toutes ces procédures, la notion clé pour le tri des éléments caractéristiques est la notion de ***valeur-test*** *dont le principe est le suivant. On évalue l'intérêt d'un élément pour caractériser une catégorie d'individus à partir d'une statistique calculée sur l'échantillon. Considérons par exemple l'utilisation de la moyenne pour déterminer les variables continues qui caractérisent un groupe d'individus. Si l'écart entre la moyenne calculée dans le groupe et sa valeur calculée sur la population est attribuable au hasard seul, la variable ne caractérise pas le groupe. Si la moyenne s'écarte de façon significative de la moyenne générale, on dira que les individus du groupe sont caractérisés par cette variable.*

La valeur-test proprement dite est le critère qui évalue statistiquement l'écart entre la mesure sur le groupe et la mesure sur la population. Il est exprimé dans une unité commode pour l'interprétation: c'est un nombre d'écarts-types d'une loi normale.

Par exemple une valeur-test égale à 4 s'interprète ainsi: si la variable n'était pas caractéristique, la probabilité d'un écart entre moyennes au moins aussi élevé que l'écart observé est égale à la probabilité de tirer une observation à 4 écarts-types dans une loi normale. Plus la valeur-test est grande (et supérieure au seuil usuel de 2 écarts-types), mieux l'élément caractérise la catégorie d'individus. Les éléments caractéristiques sont triés par valeurs-tests décroissantes, donc par ordre d'intérêt décroissant.

DEMOD (caractérisation des variables nominales)

Cette procédure fournit la caractérisation statistique détaillée d'une ou plusieurs variables nominales, en utilisant toutes les informations du fichier et en les rangeant par ordre d'intérêt.

Caractérisation d'une catégorie par les modalités
1. caractérisation par les différences (valeurs-tests)

La variable à caractériser est par exemple l'âge codé en catégories. Considérons la catégorie des jeunes. Une modalité d'une autre variable nominale est un certain groupe d'individus, par exemple le groupe des garçons (variable sexe). Pour évaluer si la modalité *garçons* caractérise la catégorie *jeunes* (pour savoir s'il y a sur-représentation des garçons dans la catégorie des jeunes), on compare la proportion de garçons chez les jeunes à la proportion générale de garçons. La valeur-test (évaluée ici à partir de la loi hypergéométrique) mesure l'importance de l'écart en nombre d'écarts-types de la loi normale.

Les valeurs-tests sont calculées pour toutes les modalités de toutes les variables nominales. Donc toutes les modalités sont rangées en fonction des valeurs-tests décroissantes pour caractériser, dans l'exemple, la catégorie des garçons. L'utilisateur est donc assuré qu'aucune caractéristique significative ne peut lui échapper.

Caractérisation d'une classe par les modalités
2. caractérisation par le contenu ("MOD/CLA")

La procédure permet aussi un autre type de caractérisation. Le classement fourni par les valeurs-tests range les modalités à partir d'un critère statistique qui évalue l'importance de l'*écart* entre deux proportions, quelque soit l'importance de ces proportions. On peut utiliser deux autres critères *en volume*, mettant en jeu les proportions elles-mêmes.

On peut tout d'abord demander le classement des modalités en fonction de leur *abondance* dans la catégorie à décrire. Cette catactérisation est fonction de la part de la modalité dans la classe à décrire (il est noté "MOD/CLA"). La première caractéristique est la modalité la mieux représentée dans la catégorie. On aura une liste du type: cette catégorie contient, par ordre décroissant, 90% de ceci, 75% de cela, 68% de cet autre attribut, etc.

Caractérisation d'une classe par les modalités
3. caractérisation par le contenant ("CLA/MOD")

Enfin on peut caractériser une classe d'individus par la façon dont elle "absorbe" les attributs décrivant les individus. Par exemple on dira que la classe *jeunes* contient tous les *célibataires*, les 2/3 des *parisiens*, 55% des *gauchers*, etc. Ce critère range toutes les modalités caractérisantes dans l'ordre où les attributs sont inclus dans la classe d'individus à décrire (il est noté "CLA/MOD").

Caractérisation d'une classe par les variables continues

Une variable continue caractérise une catégorie d'individus si la moyenne dans cette catégorie diffère significativement de la moyenne générale. La catégorie des *jeunes* par exemple sera caractérisée d'abord par un revenu faible (comparé à la moyenne), ensuite par un poids inférieur à la moyenne, un nombre d'heures d'activités sportives supérieur à la moyenne, etc.

Pour ranger les variables continues par ordre d'importance, le logiciel évalue les écarts entre moyennes en terme de valeurs-tests. On teste l'écart entre la moyenne dans le groupe et la moyenne générale. Plus l'écart est significatif, plus la valeur-test associée est grande. Le test mis en œuvre est un test non paramètrique de comparaison de moyennes. On classe l'ensemble des variables continues dans l'ordre des valeurs-tests décroissantes. La procédure prend en compte toutes les variables continues

disponibles, de sorte qu'on est assuré d'avoir classé toutes les caractéristiques du groupe d'individus.

Caractérisation d'une variable nominale par les continues

Considérons la variable "mode de locomotion: à pied / vélo / voiture / métro" et les trois paramètres: âge de l'individu, revenu annuel, kilométrage hebdomadaire. Pour ranger par ordre d'intérêt les 3 paramètres continus, on réalise 3 analyses de la variance successives, avec le même facteur "locomotion" à 4 modalités.

La "meilleure" analyse de la variance est celle qui correspond à la statistique de Fisher la plus significative. Elle correspond au paramètre continu le mieux prévisible à l'aide du facteur. Ce paramètre arrivera en tête des 3 paramètres caractérisant la variable nominale "locomotion".

Pour classer les variables continues par ordre d'intérêt dans la caractérisation d'une variable nominale, on effectue toutes les analyses de variance. Pour chaque statistique de Fisher, on calcule la probabilité d'être dépassée. La *valeur-test* est la valeur d'une variable normale qui a la même probabilité d'être dépassée. On range les statistiques de Fisher, donc les variables continues, dans l'ordre des valeurs-tests décroissantes.

Caractérisation d'une variable nominale par les autres nominales

On calcule la statistique du χ^2 associée au croisement de deux variables nominales, ainsi que la probabilité de dépasser la valeur calculée. La valeur de la loi normale qui a la même probabilité d'être dépassée est appelée *valeur-test*. Plus la valeur-test est forte, plus le tableau de croisement est intéressant.

On calcule tous les tableaux croisant la variable à caractériser avec les autres variables nominales et on range ces variables nominales dans l'ordre des valeurs-tests décroissantes. On obtient ainsi la caractérisation complète de la variable nominale.

Autres caractérisations

Une classe d'individus peut être "caractérisée" par les variables nominales de la façon suivante. On compare à l'aide de la statistique du χ^2 le profil de répartition du groupe d'individus dans la variable au profil global de la variable. On range ainsi l'ensemble des variables nominales en fonction des valeurs-tests décroissantes.

Une variable nominale peut être "caractérisée" par les modalités des autres variables de la façon suivante. Considérons les individus appartenant à une modalité d'une autre variable. Ces individus sont répartis dans les modalités de la variable à caractériser. On compare à l'aide la statistique du χ^2 le profil des individus répartis au profil global de la variable à caractériser. Les modalités qui caractérisent le mieux la variable nominales sont rangées par ordre décroissant des valeurs-tests.

Une variable nominale ou une modalité d'une variable nominale peuvent aussi être caractérisées par des variables de *fréquences* associées aux individus.

Tableaux d'éditions statistiques

Outre les caractérisations statistiques évoquées ci-dessus, la procédure DEMOD sera utilisée pour réaliser l'édition systématique de certains tableaux de résultats statistiques.

Pour une variable nominale donnée, on éditera tout ou partie des tableaux de croisement avec les autres variables. Les tableaux sont sélectionnés et édités automatiquement du plus "caractéristique" au moins caractéristique, dans l'ordre décroissant des valeurs-tests associées aux χ^2.

Concernant la même variable nominale, on peut demander l'édition d'un tableau donnant les moyennes, écarts-types, minima et maxima d'une variable continue pour chaque modalité. A chaque variable continue correspondra un tableau de statistiques. L'ordre des éditions de ces tableaux sera celui des valeurs-tests qui classent les variables continues par ordre d'importance vis-à-vis de la variable nominale à décrire.

DECLA (description des classes d'une partition)

Cette procédure fournit une caractérisation statistique détaillée de chaque classe d'une partition, analogue à la caractérisation fournie par DEMOD. Une partition est en effet assimilable à une variable nominale, et une classe d'une partition assimilable à une modalité de variable nominale.

MODALITES CARACTERISTIQUES		IDEN	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	POIDS	V.TEST
CLASSE 1 / 4		aala			33.90	339	
la famille est le seul endroit où l'on se sente bien	oui	fbi1	50.27	83.19	56.10	561	12.73
la société francaise a-t-elle besoin de se transformer ?	oui	tsol	42.56	95.28	75.90	759	11.19
opinion à propos du mariage	dissout si pb. grave	opm2	52.92	53.39	34.20	342	9.01
opinion sur le fonctionnement de la justice en 1979	assez mauvais	jus3	47.74	56.05	39.80	398	7.42
opinion à propos du mariage	union indissoluble	opm1	53.68	36.58	23.10	231	7.02
diplôme d'enseignement général le plus élevé obtenu	CEP ou fin études	dip2	45.79	43.36	32.10	321	5.34
à qui incombent les travaux ménagers et les soins enfants ?	plutôt à la femme	esc2	43.45	43.07	33.60	336	4.44
comparée aux personnes de votre âge, votre santé est	satisfaisante	san2	39.33	69.62	60.00	600	4.42
âge de l'enquêté(e) en classes	65 ans et plus	âge5	49.11	24.48	16.90	169	4.41
évolution du niveau de vie de l'enquêté depuis 10 ans	un peu moins bien	niv4	47.37	26.55	19.00	190	4.20
à qui incombent les travaux ménagers et les soins enfants ?	incombent à la femme	esc1	61.90	7.67	4.20	42	3.63
évolution du niveau de vie de l'enquêté depuis 10 ans	c'est pareil	niv3	42,40	31.27	25.00	250	3.17
comparée aux personnes de votre âge, votre santé est	peu satisfaisante	san3	46.96	15.93	11.50	115	2.99
statut d'occupation du logement	propriétaire	log2	40.34	34.51	29.00	290	2.66
CLASSE 4 / 4		aa4a			31.50	315	
opinion à propos du mariage	dissout si accord	opm3	69.25	85.08	38.70	387	20.82
la famille est le seul endroit où l'on se sente bien	non	fbi2	60.09	82.22	43.10	431	17.25
la société francaise a-t-elle besoin de se transformer ?	oui	tsol	40.45	97.46	75.90	759	12.18
opinion sur le fonctionnement de la justice en 1979	très mauvais	jus4	61.33	49.84	25.60	256	11.54
à qui incombent les travaux ménagers et les soins enfants ?	homme et femme	esc3	44.57	84.76	59.90	599	11.28
âge de l'enquêté(e) en classes	20 - 29 ans	âge2	50.61	39.68	24.70	247	7.21
taille d'agglomération (en nombre d'habitants)	Paris	agg5	46.63	48.25	32.60	326	7.00
statut d'occupation du logement	locataire	log3	40.54	67.30	52.30	523	6.42
les découvertes scientifiques améliorent-elles la vie ?	pas du tout	sci3	55.24	18.41	10.50	105	5.23
évolution du niveau de vie de l'enquêté depuis 10 ans	beaucoup moins bien	niv5	52.63	19.05	11.40	114	4.89
diplôme d'enseignement général le plus élevé obtenu	université,gde école	dip6	49.30	22.22	14.20	142	4.70
diplôme d'enseignement général le plus élevé obtenu	baccalauréat (1/2)	dip4	42.59	21.90	16.20	162	3.17
évolution du niveau de vie de l'enquêté depuis 10 ans	ne sait pas	niv6	57.69	4.76	2.60	26	2.60
diplôme d'enseignement général le plus élevé obtenu	BEPC-BE-BEPS	dip3	39.87	20.00	15.80	158	2.35

NUM.LIBELLE VARIABLES CARACTERISTIQUES	IDEN	MOYENNES CLASSE	MOYENNES GENERALE	ECARTS TYPES CLASSE	ECARTS TYPES GENERAL	V.TEST
CLASSE 1 / 4	aala		(POIDS =	339.00	EFFECTIF = 339)	
37.âge de l'enquêté(e)	âge	48.000	42.680	17.315	17.496	6.88
47.revenu personnel souhaité	rsou	6658.096	7244.479	3555.871	4756.783	-2.70
48.estimation du revenu minimum d'une famille de 2 enfants	rmin	5236.242	5561.887	1822.149	2423.403	-2.84
CLASSE 4 / 4	aa4a		(POIDS =	315.00	EFFECTIF = 315)	
48.estimation du revenu minimum d'une famille de 2 enfants	rmin	6237.507	5561.887	3139.295	2423.403	5.90
47.revenu personnel souhaité	rsou	7837.292	7244.479	6107.843	4756.783	2.60
37.âge de l'enquêté(e)	âge	34.349	42.680	14.006	17.496	-10.21

Figure 7: Caractérisation statistique des classes 1 et 4

Il s'y ajoute cependant la caractérisation par les axes factoriels de l'analyse qui a précédé la classification. D'une part on classe les axes par ordre d'importance pour caractériser globalement la partition (par les valeurs-tests associées aux analyses de variance, analyses dans lesquelles les coordonnées factorielles constituent la variable"y" et la partition constitue le "facteur" à tester).

D'autre part pour chaque classe d'individus, on évalue l'intérêt de sa localisation sur un axe en terme de valeur-test. La coordonnée d'une classe sur un axe étant la moyenne des coordonnées des individus, le rangement par les valeurs-tests correspond à la caractérisation d'un groupe d'individus par des variables continues. Le principe est identique à celui qui est utilisé dans la procédure DEMOD.

DESCO (description d'une variable continue)

La procédure DESCO fournit la caractérisation statistique automatique d'une variable continue. Les éléments caractéristiques peuvent être les autres variables continues On compare alors les corrélations entre elles en les rangeant en fonction de valeurs-tests. La valeur-test associée à une corrélation découle du test de nullité de la corrélation: plus la valeur-test est grande, plus l'hypothèse d'une corrélation nulle est facile à rejeter.

Une variable continue est caractérisable aussi par les variables nominales.On effectue toutes les analyses de variance où "y" est la variable continue à caractériser, et le facteur est chaque variable nominale successivement On range les résultats en fonction des statistiques de Fisher, transformées en valeurs-test pour les rendre comparables.

Enfin on peut caractériser la variable continue par les modalités des variables nominales. On détecte les modalités où la moyenne est caractéristique à l'aide d'un rangement par valeurs-tests (comparaison de moyennes).

Pour tous les éléments de caractérisation, on peut en fait demander une édition soit en fonction du caractère significatif de l'élément (valeurs-tests), soit en *valeur*: rangement des modalités dans l'ordre des moyennes décroissantes, rangement des continues dans l'ordre des corrélations décroissantes. Une option d'édition permet de lister les individus appartenant à chaque classe.

DEFAC (description des axes factoriels)

La procédure DEFAC permet de caractériser statistiquement les axes issus d'une analyse factorielle en utilisant l'ensemble des informations disponibles: les individus, les variables continues, les modalités des variables nominales et éventuellement les fréquences Les éléments les plus caractéristiques sont sélectionnés automatiquement et rangés pour faciliter la lecture.

Dans le cas des modalités, une option d'édition permet de ranger ces éléments caractéristiques soit selon leurs coordonnées sur l'axe, soit selon le critère statistique des valeurs-tests.

GRAPH (graphiques)

Cette procédure permet de tracer une grande variété de graphiques plans: soit des représentations planes de deux variables, soit les plans factoriels d'une analyse préalable, soit les résultats d'une classification.

La procédure possède de nombreuses fonctionalités, pilotées par 21 commandes ayant toutes des valeurs par défaut. Seules les principales sont évoquées ici.

Tous les graphiques sont tracés en mode texte et non en mode graphique; le programme est donc indépendant du matériel d'impression. On peut cependant annoncer la hauteur des caractères utilisés si on souhaite assurer l'égalité des échelles sur les axes.

Les points marqués sur le graphique peuvent être représentés par un caractère au choix, ou par des libellés qui peuvent avoir jusqu'à 60 caractères, ou encore par des symboles choisis pour représenter des catégories (H pour homme, F pour femme par exemple).

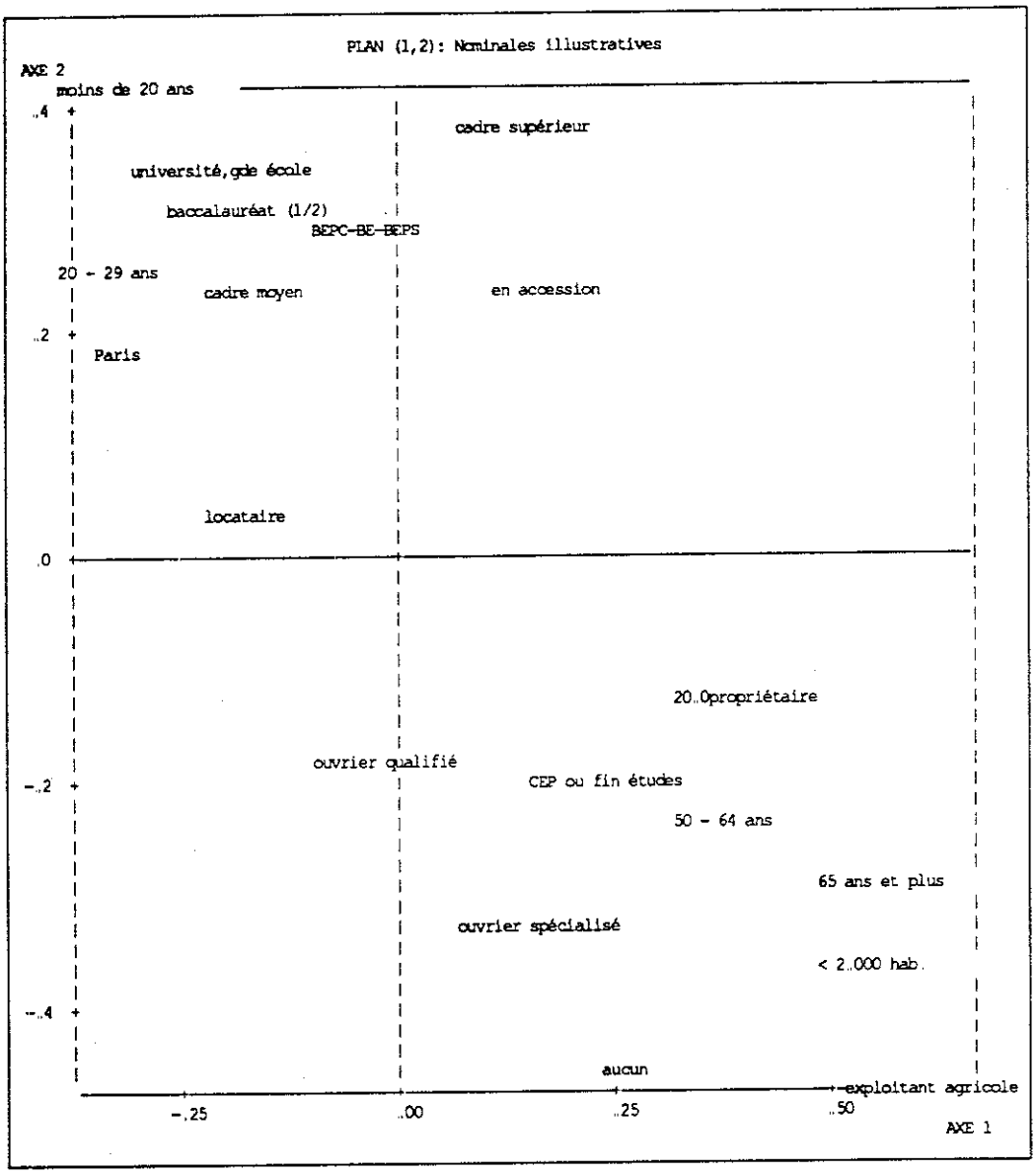


Figure 8: Graphique factoriel de variables nominales

On peut demander en option de déplacer légèrement des points qui seraient superposés. Sinon les points doubles seront clairement identifiés. La hauteur et la largeur du graphique sont choisies sans limitation, en nombre de pages ou en nombre de caractères.

On peut donner un titre au graphique et un libellé pour chaque axe. Les graduations marquées sur les axes sont exprimées en valeurs arrondies, faciles à lire. On peut ajouter ou supprimer le cadre du graphique, choisir les valeurs extrêmes sur les axes et matérialiser les droites passant par le centre du nuage de points. On peut effectuer des zooms sur le centre du graphique, éliminer les points extérieurs ou les rapporter sur les bords.

Un des deux axes de coordonnées peut être une variable nominale, ce qui permet de comparer les distributions d'une variable dans différents groupes d'individus.

Pour les représentations liées aux plans factoriels et aux classifications, la procédure possède de nombreuses fonctionalités spécifiques. La plus grande souplesse est permise pour le choix des axes factoriels, le choix des échelles sur chaque axe et le choix des points à placer dans le plan.

Toutes les représentations simultanées permises par les méthodes factorielles sont possibles à l'aide de commandes simples concernant les individus (actifs ou illustratifs), les modalités (actives ou illustratives), les variables continues et fréquences (actives et illustratives) Mais on peut également réaliser des graphiques ne contenant qu'un groupe de points spécifiés par une liste (par exemple, les points représentant les modalités des variables 5 à 10). Pour que les graphiques soient superposables, on peut demander qu'ils aient la même échelle.

Dans le cas où les éléments à positionner sur un graphique sont très nombreux, on rendra le graphique plus lisible en sélectionnant les points. La sélection automatique peut combiner deux critères: d'une part l'inertie du point (ou contribution absolue), d'autre part le cosinus-carré (pour éliminer les points mal représentés).

Notons enfin que, si une partition a été réalisée, on peut représenter simultanément sur les graphiques factoriels les centres de gravité des classes. On peut visualiser la dispersion relative des individus de chaque classe sur un graphique de densité où chaque individu est représenté par le numéro de sa classe.

POSIT (positionnement graphique a posteriori)

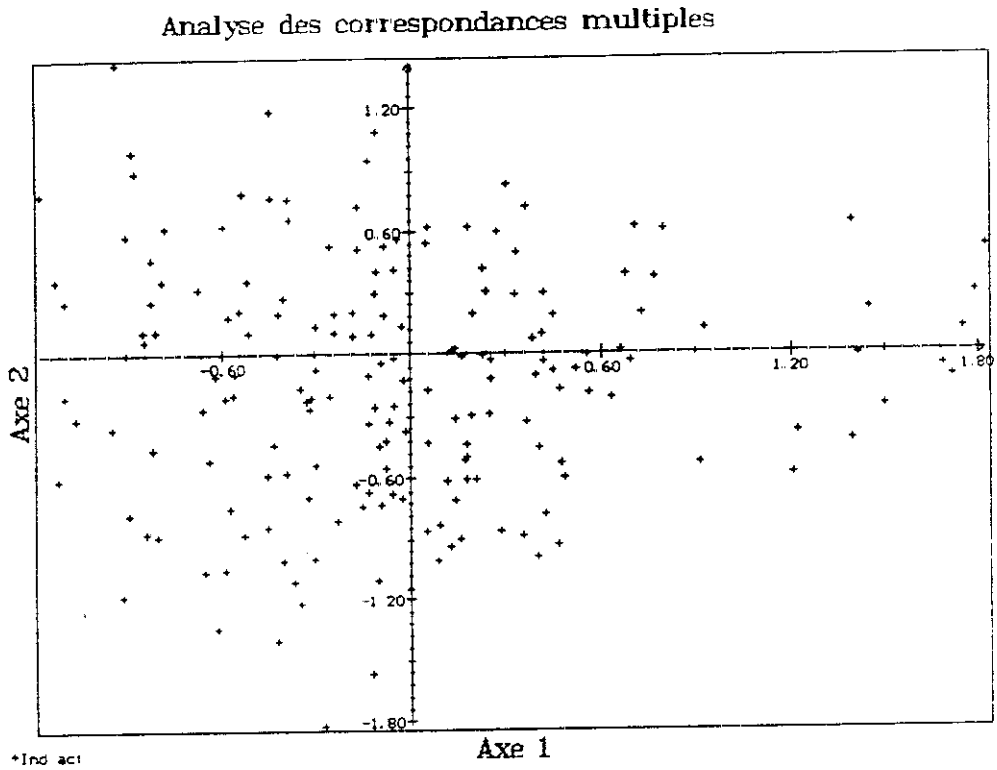
Cette procédure permet de positionner *a posteriori* des modalités sur des plans factoriels existant, même si les variables correspondantes n'ont pas été sélectionnées pour l'analyse On édite les coordonnées des modalités, les valeurs-tests indiquant l'intérêt de chaque point sur chaque axe, et les graphiques factoriels.

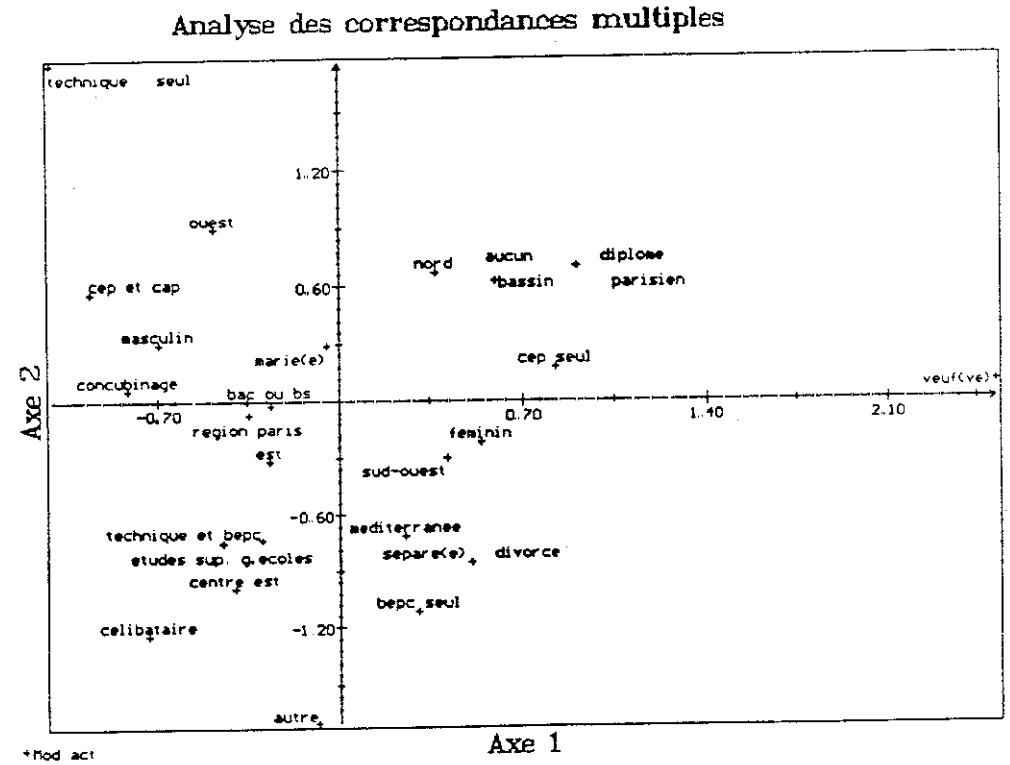
STATS (statistiques sommaires des variables)

Cette procédure fournit une description rapide et automatique des variables retenues par l'utilisateur: tris-à-plat pour les variables nominales, statistiques classiques et histogrammes pour les variables continues. Ses fonctionalités sont prévues pour fournir rapidement l'essentiel des informations sur chaque variable séparément.

Les tris-à-plat sont accompagnés d'un histogramme des fréquences dans chaque modalité. Les histogrammes des variables continues sont tracés horizontalement et fournissent les bornes et moyennes de chaque classe. On peut éliminer automatiquement les points aberrants pour ne pas "écraser" les histogrammes. La forme et l'étendue des histogrammes sont contrôlables par l'utilisateur.

Pour les variables continues, on calcule les moyennes, écarts-types et extrema. Si la variable continue prend peu de valeurs distinctes, on peut tracer un histogramme de





Deux exemples de sorties du Module Graphique (Version 2.0 pour PC)

"discrétisation": les classes du graphique correspondent aux valeurs distinctes prises par la variable.

TABLE (tableaux croisés)

La procédure TABLE produit des tableaux croisant entre elles des variables nominales. Chaque case d'un tableau peut contenir un ou plusieurs des éléments suivants: l'effectif des individus, le pourcentage en ligne et en colonne, la moyenne et l'écart-type d'une variable continue.

Les effectifs dans les cases peuvent être pondérés par une variable de redressement et les données manquantes sont traitées automatiquement.

Tout tableau peut concerner la totalité des individus, ou un sous échantillon obtenu par filtre logique (être un homme, ayant plus de 20 ans, etc). La puissance de la procédure provient de la facilité de commande des croisements d'un groupe de variables par un autre groupe de variables.

Les tableaux obtenus sont empilables et juxtaposables. Le résultat est "archivable" et récupérable par le logiciel pour réaliser des analyses ultérieures. Par exemple les tableaux d'effectifs peuvent être soumis à des analyses de correspondances, et les tableaux de moyennes, à des analyses en composantes principales.

6. Les procédures d'ajustements linéaires

*Ce chapitre concerne des procédures dites de **"décision"**. L'utilisateur choisit un modèle de dépendance entre les variables et utilise les observations pour estimer les paramètres inconnus du modèle. La statistique classique permet d'évaluer la qualité des estimations. Le modèle peut ensuite être utilisé pour extrapoler ou pour prévoir des observations non réalisées. Prévoir une valeur est le support de la décision (par exemple prévoir à quelle classe appartient un individu). Dans le cadre de certaines hypothèses de travail, le modèle peut être aussi utilisé pour tester la réalité des dépendances supposées.*

*Les modèles utilisés ici sont linéaires, non pas au titre des dépendances, mais du point de vue des coefficients inconnus à estimer. Si la variable à prévoir est du type continue, on parle de **régression multiple**. Dans cette famille on trouvera les méthode d'analyse de la variance et de la covariance utilisées en particulier pour le traitement des plans d'expérience.*

*Si la variable à prévoir est de type nominal, on parlera **d'analyse discriminante**. On se restreint ici au cas de la discrimination entre deux groupes d'individus. Outre la plus grande facilité d'interprétation des résultats, ce cas présente l'avantage d'une analogie parfaite avec la régression multiple. Une application particulièrement intéressante de la discriminante est le "scoring" auquel une procédure particulière est consacrée.*

Le choix d'un modèle d'ajustement linéaire est l'opération la plus délicate pour le statisticien. Il peut être guidé dans cette voie par des analyses exploratoires de type factorielles et de classifications permettant de décrire les dépendances entre paramètres. Ensuite il restera à choisir, dans un champ plus restreint de modèles, celui ou ceux qu'il faudra retenir. La procédure FUWIL peut être utilisée dans cette phase de choix.

VAREG (régression et analyse de variance)

Il s'agit d'une procédure générale d'ajustement de modèles linéaires permettant de réaliser une très grande variété d'analyses statistiques. Citons:

- les régressions simples et multiples
- les analyses de la variance pour traiter tout plan d'expérience
  - nombre quelconque de facteurs et de niveaux par facteur
  - nombre quelconque d'interactions d'ordre 2 ou 3
  - nombre de répétions quelconques, non égales
  - blocs équilibrés, carrés latins, etc
- les analyses de covariance sans limitation sur le nombre de covariables
  - sans limitation sur les interactions de facteurs
  - sans limitation sur le nombre d'observation par case, etc.

Les modèles à ajuster sont écrits par l'utilisateur avec une notation "algébrique" très simple, telle que: V4 = V1 + V2 + V7 + V1*V2. La notation "V1*V2" indique l'introduction de l'interaction des facteurs V1 et V2.

Quelle que soit l'analyse, on peut sélectionner un sous groupe d'individus, et on peut faire intervenir une pondération dans tous les calculs. L'estimation d'un coefficient est toujours accompagnée de l'estimation de son écart-type, de la valeur du *t* de Student, de la probabilité critique correspondante et de son expression en terme de valeur-test. La procédure contient en option un traitement automatique des données manquantes.

Chaque test *F* de Fisher associé à la décomposition de la variance est édité avec sa probabilité critique et la valeur-test correspondante pour faciliter l'évaluation des effets des facteurs et des interactions présents dans l'analyse. On édite les statistiques classiques des ajustements: somme des carrés d'écarts, estimation de la variance résiduelle, coefficient de corrélation multiple, test global de nullité de tous les coefficients et valeur-test associée.

On peut éditer les statistiques usuelles caractérisant les variables entrant dans le modèle, ainsi que la matrice des corrélations et la matrice des covariances. En sortie de la procédure, on peut créer un fichier "texte" contenant les principaux résultats, en particulier les coefficients de l'ajustement. Ce fichier est utilisé par le *module graphique* disponible sur la version PC.

DIS2G (analyse discriminante à 2 groupes)

La procédure DIS2G réalise une analyse discriminante linéaire à 2 groupes selon la méthode classique de Fisher. Dans sa forme usuelle la variable "y" à prédire est une variable nominale à 2 modalités, et les variables "x" sont des variables continues. Dans le cas de 2 modalités, l'analyse discriminante est formellement équivalente à une régression, ce qui justifie certains calculs réalisés par la procédure.

L'analyse peut être réalisée en utilisant comme variables "x" les axes factoriels d'une analyse préalable. Dans le cas d'une analyse en *composantes principales*, cette procédure permet de choisir les axes les plus intéressants pour établir la fonction discriminante. On pourra par exemple éliminer des axes éloignés, porteurs de fluctuations aléatoires ou sans intérêt pour la discrimination. Après calcul des coefficients relatifs aux axes factoriels, le programme les transforme pour établir la fonction discriminante finale sur les variables d'origine.

Dans le cas d'une analyse des *correspondances multiples*, le programme établit la fonction discriminante sur les axes choisis par l'utilisateur, puis calcule les coefficients attribués aux modalités des variables nominales de l'analyse. Cette méthode permet donc de réaliser de façon naturelle les analyses discriminantes sur variables nominales.

La procédure DIS2G possède plusieurs méthodes de validation des résultats Une méthode consiste à scinder l'échantillon en deux parties: l'une pour calculer la fonction discriminante (échantillon*"d'apprentissage"*), l'autre pour évaluer la qualité de la discrimination (échantillon*"test"*). L'échantillon test peut être construit automatiquement par tirage au hasard, ou défini par l'utilisateur soit par liste soit par filtre logique sur les variables.

Le programme intègre de plus une procédure de validation des résultats par "*bootstrap*". On obtient ainsi des estimations sans biais pour les coefficients, les écarts-types des coefficients, les corrélations de la fonction discriminante avec les variables d'origine, les pourcentages de bien et mal classés et les écarts-types associés à ces pourcentages.

Pour tout individu *anonyme* (dont on ne connait pas le groupe), la procédure calcule la probabilité d'appartenance à chaque groupe.

Dans tous les cas on peut, avant de calculer la fonction discriminante, introduire des probabilités a priori d'appartenance aux groupes, ainsi qu'une matrice de coûts a priori. On peut sélectionner aisément les individus qui participeront aux calculs et les munir d'un poids de redressement (voir la procédure SELEC).

La procédure fournit de nombreuses éditions de résultats, en particulier les affectations des individus dans les groupes, accompagnées de la probabilité d'appartenance, les statistiques des variables pour chaque groupe, ainsi que les matrices de corrélations, et les histogrammes superposés des distributions des individus dans les groupes

La procédure DIS2G crée deux fichiers de résultats permettant de communiquer avec l'extérieur. Un premier fichier contient le reclassement des individus par la fonction discriminante en 4 catégories: bien ou mal classé dans chaque groupe Ce fichier est un fichier interne de SPAD.N, donc archivable et récupérable pour tout traitement statistique interne au logiciel.

Un second fichier, de type texte, contient les coefficients de la fonction discriminante. Ce fichier est utilisé comme véhicule des résultats vers la procédure SCORE lorsque l'on veut étudier une fonction de score.

SCORE (création et étude de scores)

La procédure SCORE est exécutable après une analyse discriminante à 2 groupes réalisée sur des variables nominales (procédure CORMU suivie de DIS2G). Le score attribué à un individu s'obtient en additionnant les coefficients associés aux modalités de l'individu.

Les coefficients sont automatiquement récupérés après l'analyse discriminante. Il est possible cependant d'étudier l'effet d'une modification des coefficients ou l'effet des arrondis sur les valeurs en introduisant soi-même les coefficients corrigés de la fonction. Tous les calculs seront réalisés avec ces coefficients.

L'introduction d'une tolérence d'erreur de classement permet de définir trois zones de décision sur la fonction de score: la *zone verte* du côté des scores forts, la *zone rouge* du côté des scores faibles, et la zone intermédiaire ou *zone d'indécision*. Un graphique permet d'apprécier comment la zone d'indécision diminue quand la tolérance d'erreur augmente.

Le tableau des coefficients est édité en rangeant les variables dans l'ordre décroissant de leur participation maximale au score. Dans chaque variable, les modalités sont rangées dans l'ordre décroissant de leur contribution au score.

Indépendamment de la pondération générale pour tous les calculs (procédure SELEC) on peut utiliser ici une pondération sur les deux groupes de la variable à discriminer pour rendre les effectifs représentatifs de ceux de la population.

Une abaque permet de lire, pour chaque valeur du score, l'estimation de la *probabilité conditionnelle* d'appartenir à chaque groupe. Des graphiques fournissent également la répartition des groupes et l'estimation de la répartition de la population en fonction de la valeur du score.

La procédure crée un fichier de résultats contenant les scores calculés pour chaque individu. Ces données sont archivables dans SPAD.N et donc réutilisables pour tout traitement ultérieur (par exemple pour des graphiques).

FUWIL (sélection des ajustements optimaux)

La procédure FUWIL est utilisée pour aider au choix des "meilleures" variables à introduire dans un modèle d'ajustement linéaire. Elle servira ausi bien dans le cas de la *régression multiple* que dans le cas de la *discriminante à deux groupes*, formellement équivalente à une régression.

L'utilisateur dispose de trois critères de comparaison globale des ajustements:

- le coefficient "R^2"de corrélation multiple
- le coefficient de corrélation multiple "corrigé"
- le coefficient "Cp" de Mallows

La notion de "meilleur" ajustement est relative ici au critère choisi. En fait l'utilisateur fera intervenir en général beaucoup d'autres considérations au moment de la sélection du modèle final (en particulier les valeurs-tests des coefficients).

Le programme édite les ajustements avec une seule variable, de la meilleure à la moins bonne. Puis il édite les ajustements à deux variables, du meilleurs couple de variables au moins bon. Ensuite il édite les ajustements à trois variables, etc.

Ces éditions sont accompagnées de la valeur du critère global, ainsi que des principales informations sur les coefficients: valeurs, écarts-types, probabilités critiques et valeurs-tests. L'utilisateur dispose ainsi des éléments essentiels pour le choix de son modèle. Un graphique de synthèse montre comment le critère globale évolue d'un ajustement à l'autre.

Le choix d'un modèle optimal, régression multiple ou discriminante, peut être réalisé non pas sur les variables d'origine mais sur les facteurs issus d'une analyse préalable. Il suffit pour cela de créer un fichier texte des coordonnées factorielles (procédures BIFOR ou ASCII) et d'utiliser ces fichiers comme fichiers de départ.

EXEMPLES D'ENCHAINEMENTS

Dans cette partie, on présente quelques enchaînements produisant des analyses usuelles. Pour chaque analyse, on donne quelques indications sur les procédures utilisées, les options choisies, et les résultats qu'on obtient. Dans ce qui suit, on suppose que les procédures de lecture et d'archivage ARDIC et ARDON ont été exécutées préalablement

1. Correspondances multiples et classification

L'analyse des correspondances multiples fournit une description d'un tableau de variables nominales observées sur des individus (réponses à un questionnaire d'enquête, par exemple). Les éventuelles variables continues sont utilisables pour enrichir les aides à l'interprétation. Afin d'étendre la description au delà des tous premiers axes factoriels de l'analyse, on complète souvent l'étude par une classification des individus.

SELEC On sélectionne les variables nominales actives et illustratives utiles pour l'analyse. On peut éventuellement déclarer des variables continues illustratives, et sélectionner une partie seulement des individus.

CORMU On exécute les principaux calculs de l'analyse des correspondances multiples. Les valeurs propres et les coordonnées des modalités (et éventuellement des individus) sont éditées.

GRAPH On édite les graphiques factoriels, pour les modalités actives et illustratives, ainsi que des graphiques de densité des individus

DEFAC Caractérisation statistique automatique des facteurs. Aides à l'interprétation.

RECIP Cette procédure réalise une agrégation hiérarchique des individus caractérisés par leurs premières coordonnées factorielles. Pour des grands tableaux, on utiliserait la procédure SEMIS.

PARTI Au vu des résultats de la procédure précédente, on demande la coupure de l'arbre en un certain nombre de classes, pour autant de partitions que nécessaire.

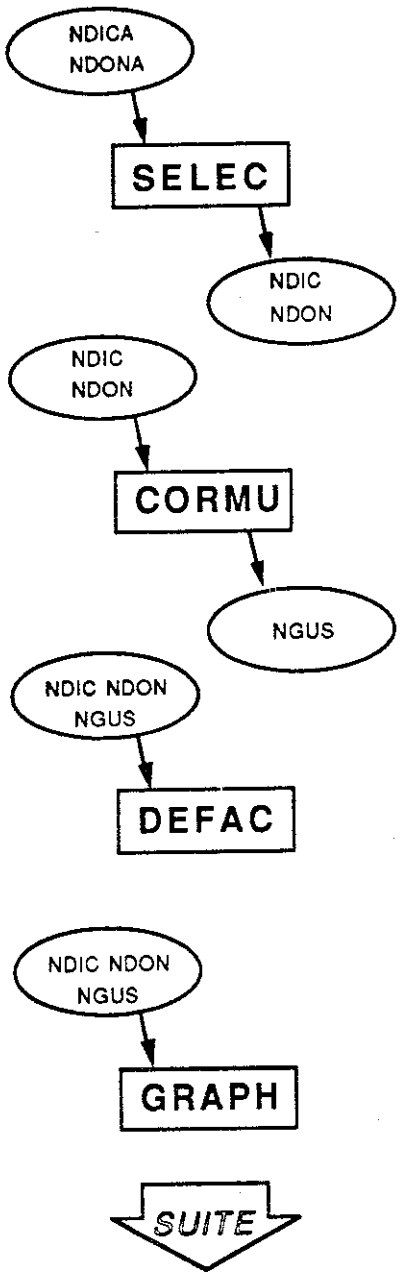
DECLA On demande la caractérisation statistique des classes, par les modalités des variables nominales et par les axes de l'analyse (par exemple).

GRAPH Déjà utilisée pour éditer les graphiques factoriels, cette procédure permet ici de visualiser la position des classes sur les graphiques, ainsi que le nuage de densité des individus repérés par le numéro de leur classe.

2. Composantes principales et classification

Cet enchaînement est analogue au précédent, mais concerne l'exploration d'un tableau de données individuelles où les variables observées sont continues. Les variables nominales, qui définissent des groupements d'individus, peuvent être utilisées pour enrichir les interprétations. Ici encore, l'analyse peut être suivie d'une classification descriptive. L'enchaînement sera par exemple:

SELEC -- COPRI -- DEFAC - GRAPH -- RECIP -- PARTI -- DECLA -- GRAPH



Enchaînement (début)

CORRESPONDANCES MULTIPLES ET CLASSIFICATION

➡ SELEC: choix des éléments

SELEC sélectionne les éléments entrant dans l'analyse:

- nominales actives et illustratives
- continues illustratives
- individus actifs et illustratifs
- poids des individus

➡ CORMU: calculs principaux

CORMU effectue les principaux calculs de l'analyse des correspondances multiples. La procédure édite:

- les valeurs propres avec leur histogramme
- les coordonnées, contributions, cosinus carrés et valeurs-tests de tous les éléments.

➡ DEFAC: description des facteurs

DEFAC fournit des aides à l'interprétation des facteurs. Chaque axe est caractérisé statistiquement par ses éléments les plus significatifs.

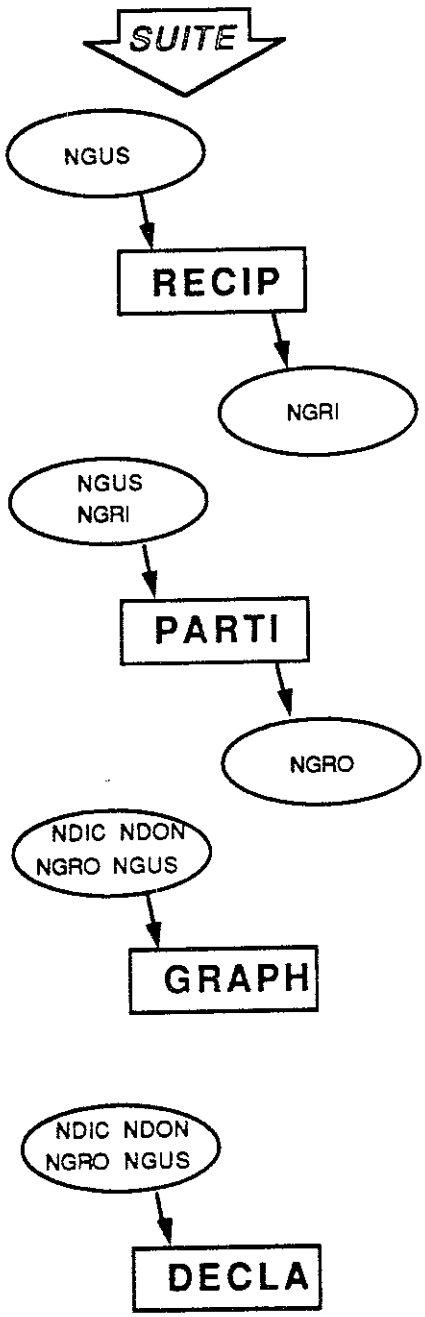
L'utilisateur choisit les axes à caractériser, les éléments utilisés pour la caractérisation et le critère de sélection des éléments les plus caractéristiques.

➡ GRAPH: graphiques factoriels

GRAPH permet la représentation sur les plans factoriels de tous les éléments ayant participé à l'analyse:

- modalités, actives et illustratives
- variables continues illustratives
- individus identifiés par leur nom ou sous forme de graphiques de densité.

Ces graphiques sont pilotés par plusieurs paramètres fournissant une grande variété de tracés.



Enchaînement (fin)

CORRESPONDANCES MULTIPLES ET CLASSIFICATION

➡ RECIP: classification des individus

RECIP effectue une agrégation hiérarchique des individus. Les calculs sont effectués sur les coordonnées factorielles. On utilise le critère de Ward. La procédure édite l'arbre hiérarchique pour faciliter le choix des partitions.

Si le nombre d'individus est très important, on utilise la procédure de classification mixte SEMIS plus économique (construction de classes stables autour de centres mobiles).

➡ PARTI: partition par coupure

PARTI détermine des partitions par coupure de l'arbre hiérarchique. Chaque partition est consolidée, c'est-à-dire optimisée par itérations autour de centres mobiles. On peut créer et archiver plusieurs partitions (fichier NGRO).

On peut éditer le contenu des classes ainsi que les parangons (indivius types) de chacune. Les classes sont positionnées et caractérisées sur les axes factoriels.

➡ GRAPH: graphiques des classes

GRAPH permet de positionner sur les graphiques factoriels les centres des classes et les individus repérés par le numéro de classe.

L'utilisateur choisit les éléments à représenter ainsi que les options de tracé des graphiques.

➡ DECLA: caractérisation des classes

DECLA fournit la caractérisation statistique automatique des classes d'une partition. Les éléments les plus significatifs sont recherchés dans la totalités des données disponibles et rangés selon le critère des valeurs-tests.

3. Analyse des correspondances d'un croisement de variables

On dispose encore d'un tableau de données individuelles, contenant des variables nominales et continues. On se propose d'analyser le tableau de contingence croisant deux des variables nominales. D'une façon plus générale, l'enchaînement proposé permet d'étudier un *empilement* de tableaux de contingence, et même une juxtaposition de tableaux empilés ("tranches" d'un tableau de correspondances multiples).

TABLE La procédure calcule et édite les tableaux de contingence à analyser.

SELEC On sélectionne les colonnes et les lignes, actives et illustratives.

CORBI La procédure effectue les principaux calculs de l'analyse des correspondances. Les valeurs propres, les coordonnées et les contributions sont éditées.

GRAPH On édite les graphiques factoriels pour les lignes et les colonnes, actives et illustratives.

Notons également la possibilité de remplacer, dans le tableau de contingence, les effectifs par les moyennes d'une variable continue. On crée ainsi un tableau "synthétique" de mesures qu'on peut soumettre à une analyse en composantes principales. Dans l'enchaînement précédent, COPRI remplacerait alors CORBI.

4. Description élémentaire des données

On demande les histogrammes des variables continues (accompagnées des moyennes et écarts-types), et les tris à plat des variables nominales. On décrit l'association deux à deux des variables continues à l'aide de graphiques plans.

SELEC On sélectionne les variables à décrire.

STATS La procédure dessine les histogrammes et édite les statistiques et les tris à plat

GRAPH On édite les graphiques représentant les individus en fonction des valeurs de deux variables.

On peut décrire l'association entre variables nominales par des tableaux de contingence, à l'aide de la procédure TABLE (pouvant être utilisée seule).

5. Description directe d'une variable nominale

On désire caractériser statistiquement chaque groupe d'individus défini par une modalité de variable nominale, à l'aide des variables observées sur ces individus. On caractérisera par exemple le groupe des hommes et celui des femmes, définis par les deux modalités de la variable "sexe de l'enquêté".

SELEC Sélection des variables utiles à la caractérisation.

DEMOD Description des modalités de certaines variables.

Bibliographie d'analyse statistique des données

Benzécri J-P. (1973) *L'Analyse des Données, Tome 1: La Taxinomie, Tome 2: L'Analyse des Correspondances* Dunod, Paris (2de éd. 1976)

Bouroche J-M. , Saporta G (1983). *L'analyse des Données* P.U.F. , Collection "Que sais-je", Paris.

Caillez F., Pagès J-P. (1976). *Introduction à l'Analyse des Données*. Smash, Paris.

Cibois P. (1984). *L'analyse des Données en Sociologie*. P.U.F., Paris.

Celeux G , Diday E., Govaert G , Lechevallier Y , Ralambondrainy H. (1989) *Classification Automatique des Données*. Dunod, Paris

Diday E. (1983). *Eléments d'Analyse des Données*. Dunod, Paris.

Escofier B., Pagès J. (1988). *Analyses Factorielles Simples et Mutiples*. Dunod, Paris

Fénelon J-P. (1981). *Qu'est-ce que l'Analyse des Données* Lefonen, Paris.

Gifi A. (1981). *Non linear Multivariate Analysis*. Department of Data Theory, University of Leiden, Leiden.

Greenacre M. (1984) Theory and Application of Correspondence Analysis. Academic Press, London.

Jambu M. , Lebeaux M-O (1978) *Classification Automatique pour l'Analyse des Données Tome 1: Méthodes et Algorithmes, Tome 2: Logiciels* Dunod, Paris

Lagarde J. (1983) *Initiation à l'Analyse des Données* Dunod, Paris.

Lambert T. (1986). *Réalisation d'un Logiciel d'Analyse de Données*. Université de Paris-Sud, Dép. Statistique, Orsay.

Lebart L , Morineau A. (1982). *SPAD Système Portable pour l'Analyse des Données* CESIA, 82 rue de Sèvres, 75007 Paris

Lebart L., Morineau A., Fénelon J-P. (1979). *Traitement des Données Statistiques, Méthodes et Programmes* Dunod, Paris (2de éd. 1982)

Lebart L., Morineau A , Tabard N. (1977). *Techniques de la Description Statistique, Méthodes et Logiciels pour l'Analyse des Grands Tableaux.* Dunod, Paris

Lebart L., Morineau A., Warwick K.W. (1984) *Multivariate Descriptive Statistical Analysis, Correspondence Analysis and Related Techniques for Large Matrices*. Wiley, New York.

Lerman I. C. (1981). *Classification et Analyse Ordinale des Données*. Dunod. Paris.

Nishisato S. (1980). *Analysis of Categorical Data, Dual Scaling and its Applications* Universty of Toronto Press, Toronto.

Roux M. (1985). *Algorithmes de Classification* Masson, Paris.

Tomassone R., Danzart M., Daudin J J, Masson J P. (1988) *Discrimination et Classement* Masson, Paris.

Tomassone R , Lesquoy E., Millier C. (1983) *La regression: Nouveaux regards sur une Ancienne Méthode Statistique* Masson, Paris.

Volle M. (1981). *Analyse des Données*. Economica, Paris