UNIVERSITÉ DE NEUCHÂTEL
FACULTÉ DE DROIT ET DES SCIENCES ÉCONOMIQUES

Méthodes pratiques de dépouillement de questionnaires

THÈSE

PRÉSENTÉE À LA FACULTÉ DE DROIT ET DES SCIENCES ÉCONOMIQUES
POUR OBTENIR LE GRADE DE DOCTEUR ÈS SCIENCES ÉCONOMIQUES

PAR

PIERRE-ANDRÉ CHARDON

IMPRIMERIE DE L'OUEST SA, 2034 PESEUX
1981

Monsieur Pierre-André CHARDON est autorisé
à imprimer sa thèse de doctorat ès sciences économiques intitulée

« Méthodes pratiques de dépouillement de questionnaires ».

Il assume seul la responsabilité des opinions énoncées.

Neuchâtel, 16 juin 1980.

Le doyen
de la Faculté de droit
et des sciences économiques

Roland Ruedin

Méthodes pratiques de dépouillement de questionnaires

I. Introduction 11
1. Le questionnaire 11
2. Les phases du dépouillement 12
A) Confection du questionnaire 12
* Présentation du document 13
* Questions ouvertes et semi-ouvertes 13
* Structure et contenu 15
B) Préparation du dépouillement 16
* Création des fichiers 16
* Etablissement de documents utiles 17
C) Le dépouillement proprement dit 17
3. Représentation d'un questionnaire 18
4. Présentation du questionnaire d'où sont tirés la plupart des exemples 19

II. Codification des données 20
1. Questions dont les réponses sont qualitatives 20
* Variables nominales 21
* Variables ordinales 23
2. Questions dont les réponses sont quantitatives 23
3. Dedoublement des variables 25
* Variables ordinales 25
* Variables binaires 26
4. Exemples d'erreurs à éviter 27

III. Les tris 29
1. Tris à plat
2. Tris croisés 31
* La notion de filtre 34
3. Tests du khi-carré 36
A) Comparaison d'une distribution à une autre 36
B) Tests d'indépendance et tris croisés 38

IV. L'analyse factorielle 41

1. Principes de la méthode 41

A) Analyse factorielle générale 44

* Analyse en composantes principales sur données centrées 47

* Idem sur données centrées et réduites 47

B) Analyse factorielle des correspondances 48

C) Représentation simultanée des observations et des variables 50

D) Quelques raisons de préférer l'analyse des correspondances 53

2. Interprétation des résultats 54

A) Aides à l'interprétation 54

B) Quelques considérations quant à l'interprétation 58

3. Observations et variables supplémentaires 59

4. Conditions d'application de l'analyse des correspondances 60

5. Exemples d'applications 62

A) Premier exemple 63

B) Deuxième exemple 73

C) Troisième exemple 76

V. Classification automatique 83

1. La Classification hiérarchique 83

A) Le principe de la classification hiérarchique ascendante 83

B) Distances entre points 87

C) Critères d'agrégation 89

D) Choix de la distance et du critère d'agrégation 91

2. Les méthodes de partitionnement 94

3. Exemples d'utilisation de la classification hiérarchique 96

A) Classification des loisirs après analyse factorielle (partitionnement) 96

B) Classification d'exploitations agricoles (classification hiérarchique ascendante) 100

VI. Exploitation des résultats fournis par l'analyse factorielle et la classification automatique 103

1. Les tris 103

A) Tris portant sur des variables existantes 104

B) Tris portant sur de nouvelles variables 105

C) Exemple: suite de l'étude des exploitations agricoles 106

2. Association d'observations à des classes de variables 112

A) Critères d'association 113

B) Exemple: observations associées aux classes de loisirs 116

VII. Programmes 119

1. Liste des programmes utilisés 119

2. Fichiers utilisés par les programmes 121

3. Remarques sur l'utilisation et la création de programmes 123

Annexe A. Exemple de dialogue pour éditer des tableaux 131

Annexe B. Listes de programmes 137

Annexe C. Bibliographie 142

PREFACE

Cette étude a pour but de montrer l'utilité de la statistique multivariée associée aux méthodes habituelles de tris pour celui qui est appelé à extraire et analyser l'information contenue dans un questionnaire. Elle s'adresse donc en particulier au sociologue et au psychologue.

Le but n'est donc pas de présenter en détail les aspects mathématiques des méthodes décrites, ni de proposer un système informatique de dépouillement de questionnaires. La présentation des outils statistiques est faite dans l'unique espoir de permettre au non-spécialiste de les comprendre de manière intuitive, mais néanmoins correcte, en sorte qu'il puisse les utiliser avec profit.

Les deux premiers chapitres sont relatifs aux données, à leur saisie et à leur représentation. Après quelques rappels concernant les tris, l'analyse factorielle et la classification automatique sont décrites, ainsi que l'usage de leurs résultats. Enfin, un chapitre est consacré aux programmes utilisés, et quelques indications sont données sur ce que devrait offrir à l'utilisateur un système de dépouillement.

Les exemples d'application ne sont que des ébauches d'études. Il ne faut donc pas en attendre des résultats complets au sens d'une étude sociologique fouillée.

I. INTRODUCTION

1. Le Questionnaire

Le questionnaire est un moyen de réunir un certain nombre d'informations sur des gens, des entreprises, des régions, etc. Ces renseignements concernent un domaine particulier, tel que la pratique de loisirs, la possession d'un matériel, les opinions sur un sujet, les raisons d'un comportement, ou n'importe. Ce qui distingue l'interrogation par questionnaire de l'interview, c'est que les questions sont standards et, sauf exceptions, les possibilités de réponses sont limitées par une liste exhaustive. L'individu n'a donc pas l'occasion de s'exprimer en répondant à un questionnaire, car celui-ci ne vise pas à collecter des témoignages. En conséquence, aucun enquêté ne va apparaître personnellement dans le dépouillement, mais en tant qu'élément d'un groupe.* Un certain nombre de groupes sont donnés à priori par des attributs tels que l'âge, le sexe, l'état civil ou le revenu. D'autres groupes seront formés en fonction des renseignements du champ de l'étude.

Pour qu'un groupe ait de l'intérêt, il faut qu'un nombre suffisant d'unités le compose, sinon il n'a pas de valeur représentative. Cela suppose, entre autre, un échantillon d'une

* *Le mot groupe est utilisé dans le sens général de "catégorie sociale" ou "ensemble statistique", et non pas nécessairement de "groupement" au sens proprement sociologique.*

certaine taille. Comme le nombre de questions posées est souvent assez grand, on a affaire à une grande quantité de données, d'où le traitement par ordinateur. Les caractéristiques essentielles des questionnaires que nous envisageons ici sont donc l'uniformité, en ce sens qu'il s'agit de questionnaires fermés, et l'importance du volume des informations à traiter. A cela s'ajoute le fait que la plupart des données sont qualitatives, ce qui rend inutilisable la plupart des techniques de la statistique classique.

Les questionnaires contiennent toujours deux parties. La première concerne les renseignements dits socio-administratifs, comme l'âge, le sexe ou l'état civil, qui constituent en quelque sorte la carte d'identité de la personne interrogée. Ces données peuvent provenir d'autres sources que l'enquêté lui-même. Par exemple, le revenu peut être obtenu en consultant le registre des impôts, la classe sociale être estimée par l'enquêteur. La seconde partie contient les questions propres à chaque enquêté. Elles forment le champ principal de la recherche.

Pour la plupart des questionnaires sociologiques, une partie du dépouillement consistera à chercher les relations entre les variables du second groupe et celles du premier.

2. Les phases du dépouillement

A) Confection du questionnaire

C'est une phase importante. Un questionnaire mal préparé au type de dépouillement que nous envisageons ici crée des difficultés.

* Présentation du document

Il faudra transférer les informations collectées sur un support magnétique de l'ordinateur. Il existe une possibilité de lecture directe de marques par la machine. Si on dispose d'un tel lecteur, il suffit de prévoir les endroits ou l'enquêté doit faire une coche pour donner sa réponse à la question. Cette méthode a été utilisée pour le recensement fédéral de 1970. Sinon, il faudra retranscrire les réponses soit sur des cartes perforées, soit directement sur un support magnétique par l'intermédiaire d'une console. Pour que ce travail soit facile, ce qui évite bon nombre d'erreurs et permet une relative rapidité (donc une diminution des coûts), il faut que les valeurs soient facilement lisibles et qu'il ne puisse pas y avoir de confusion ni sur l'ordre dans lequel elles doivent être retranscrites, ni sur l'endroit où elles doivent figurer. Une façon de faire qui respecte ces conditions consiste à laisser une grande marge sur chaque feuille avec des carrés qui représentent les colonnes des cartes à perforer (fig. 1). Ainsi on évite également une première retranscription des valeurs sur des documents de perforation séparés, qui occasionnent une perte de temps et des difficultés lors des contrôles ultérieurs.

* Questions ouvertes et semi-ouvertes

Une question ouverte ne comporte pas de limitation dans les possibilités de réponses. Les méthodes de dépouillement envisagées ici ne permettant de traiter que des données représentables par des nombres (en général des codes), il est nécessaire de faire un premier dépouillement manuel pour exploiter ce genre de questions. Il faut créer des catégories de réponses, auxquelles on fait correspondre des codes. On attribue ensuite à chaque enquêté le code correspondant à la catégorie dans

II. CARACTERISTIQUES PERSONNELLES

3. Etes-vous ___ ?

1 Célibataire
2 Marié (e)
3 Veuf (veuve)
4 Divorcé (e) / Séparé (e)
0 Refus de réponse

26- ☐

4. a) Quelle est votre religion ?

1 Protestant (e)
2 Catholique (romsin ou catholique chrétien)
3 Israélite
4 Autre religion ; spécifier : _____
5 Sans confession
0 Refus de réponse

27- ☐

b) Question pour ceux qui ont une religion

Vous considérez-vous comme pratiquant (e) ?

3 Non
4 Oui, assez
5 Oui, très
1 Refus de réponse
2 Ne sait pas

28- ☐

Figure 1

laquelle on le range. Il est évident que ce travail n'est possible que si la question suppose une réponse relativement succinte, ce qui diminue d'autant son intérêt.
Une question semi-ouverte offre le choix entre des réponses standards et une réponse libre. Le même traitement qu'aux questions ouvertes doit donc y être appliqué. En général le travail est plus simple, car la réponse libre est la plupart du temps donnée en un mot.
Notons que si on craint d'oublier dans la liste des réponses proposées une possibilité importante, on peut toujours prévoir le cas "autre" en demandant de préciser, ce qui permet de compléter éventuellement la liste au vu des réponses.
C'est la nécessité d'un pré-dépouillement manuel qui doit inciter à limiter au maximum les questions ouvertes.

* Structure et contenu

Il arrive souvent qu'une question ne s'adresse qu'à une partie des enquêtés. Un code qui signifie "sans objet" sera attribué pour cette question à ceux qui n'ont pas à y répondre. Mais il arrive aussi qu'il s'agisse de tout un groupe de questions, qui forment un champ d'étude particulier (par exemple les investissements à l'étranger dans un questionnaire sur les investissements). Dans ce cas il faut prévoir une quesion qui indique s'il y a des réponses à ce groupe ou non. Dans l'exemple considéré, il vaut donc mieux poser la question "faites vous des investissements à l'étranger ? " suivie du commentaire "Si non, passer à la question x", que de donner l'instruction "Si vous n'investissez pas à l'étranger, passer à la question x". De cette façon, il est facile d'isoler les enquêtés concernés par un tel groupe de questions.
Dans le même ordre d'idées, il faut absolument éviter que des questions qui ont le même numéro soient différentes selon le

genre d'enquêtés auxquelles elles s'adressent, ce qu'on peut être tenté de faire lorsqu'un questionnaire comporte une partie commune suivie de plusieurs parties spécifiques.
Le contenu est l'affaire du spécialiste. Toutefois une remarque s'impose: il faut soigneusement éviter d'allonger un questionnaire en demandant des renseignements qui ne seront pas exploitables lors du dépouillement.

B) Préparation du dépouillement

Un certain nombre de travaux sont nécessaires pour mettre en oeuvre les méthodes décrites plus loin. Nous avons déjà vu le problème des questions ouvertes, et nous verrons plus loin le problème de la codification des données.

* Création des fichiers

L'enregistrement des données se fait à l'aide d'un petit programme, dont l'exemple le plus simple est donné en annexe.
Dans le but de détecter d'éventuelles erreurs, il est recommandé de compléter la programme de lecture et d'enregistrement par une procédure de recherche du maximum et du minimum de chaque variable, ainsi que, si il y a plusieurs cartes perforées par enquêté, d'une procédure de contrôle de la séquence des cartes.
On crée ainsi un fichier qui contient toutes les informations reportées sur les cartes perforées. C'est à partir de là, et non des cartes, que seront extraits d'autres fichiers, de taille plus réduite, en vue d'appliquer certains des programmes

décrits plus loin. Des fichiers annexes devront être créés, en fonction des programmes utilisés.

* Etablissement de documents utiles

Il est recommandé d'établir une liste des fichiers qui indique leur contenu, ainsi qu'une liste des variables dans laquelle on trouve la signification de chaque code. D'autre part, on peut signaler l'utilité d'établir un plan de dépouillement, même si plus tard il ne sera pas suivi à la lettre en raison du caractère empirique de ce genre de recherche.

C) Le dépouillement proprement dit

Le cycle proposé est le suivant :

- Edition d'un certain nombre de tableaux
- Extraction d'un choix de variables

  Ce point peut être rendu nécessaire si le questionnaire contient beaucoup de variables et qu'on veut disposer de fichiers plus petits et donc plus faciles à manipuler pour faire une recherche sur un domaine particulier du champ de l'enquête. Il est aussi possible que cette phase soit rendue obligatoire par les limitations des programmes utilisés.

- Recodification de quelques-unes des variables

  Il arrive qu'on doive codifier des variables autrement pour un traitement qu'elles ne l'étaient au moment de leur saisie.

Cela arrive en particulier quand on a saisi des variables avec un codage pratique pour les tris et économe en place, et qu'on veut se servir de ces mêmes variables pour faire une analyse factorielle ou une classification automatique.

- Application(s) de l'analyse factorielle et/ou de la classification automatique

- Réédition de tableaux, pour étudier les classes formées ou pour confirmer et quantifier les résultats fournis par l'analyse factorielle et la classification

Il est bien entendu que l'ordre suivi dépend du questionnaire et des résultats attendus.

3. Représentation d'un questionnaire

On peut toujours représenter les données numériques issues d'un questionnaire sous la forme d'un tableau, chaque ligne contenant toutes les informations relatives à un enquêté, et chaque colonne une information sur chacune des personnes. Comme les traitements décrits ici s'appliquent toujours à des tableaux, qu'il s'agisse de celui qui contient toutes les données de l'enquête ou seulement d'un extrait, nous appellerons, par commodité, "observations" les lignes et "variables" les colonnes des tableaux à traiter.

4. Présentation du questionnaire d'où sont tirés la plupart des exemples

Ce questionnaire, "Le comportement culturel des Chaux-de-Fonniers", fait partie d'une recherche menée à La Chaux-de-Fonds de 1972 à 1974 par une équipe de sociologues, sous la responsabilité de Madame Josette Coenen-Huther et la haute direction de l'Institut de sociologie et de science politique de l'Université de Neuchâtel (Prof. Maurice Erard et l'un des assistants, Jacques Amos). Les enquêtés ont été sélectionnés par tirage au sort dans le fichier de la police des habitants. Il a fallu 791 adresses pour obtenir 492 questionnaires utilisables. Les étrangers, les moins de 16 ans et ceux qui avaient l'âge de la retraite ont été exclus de l'enquête. Les distributions des gens selon l'âge, le sexe et l'état civil correspondent à celles observées dans la population chaux-de-fonnière.

Après les questions relatives aux caractéristiques personnelles, le questionnaire contient une liste très étendue d'activités de loisirs. Pour chacune de ces activités, on demande quand, à quelle fréquence, à quelle saison, avec qui et dans le cadre de quelle association elle est pratiquée. Des questions portent sur des loisirs particuliers (cinéma, théâtre, spectacles de variétés, spectacles sportifs, concerts, etc); d'autres cherchent à cerner les désirs et les insatisfactions des gens.

Le volume de ce questionnaire (1338 variables) a posé quelques problèmes pour le traitement, compte tenu de la petite taille de la machine utilisée (IBM 1130 - 16 K-mots; capacité d'un disque : 512000 mots, soit 1 M-byte).

II. CODIFICATION DES DONNEES

Les données sont manipulées plusieurs fois, d'abord au moment de leur saisie, puis au cours des différents traitements. Une codification parfaitement adaptée dans un cas n'est pas forcément bonne dans un autre, même pour un type donné de traitement.

Au moment de la saisie, la préoccupation sera d'assurer la simplicité des tris et l'occupation minimale de place sur le support magnétique. Autrement dit, on essayera de représenter les réponses sans ambiguité et avec un nombre minimal de variables.

1. Questions dont les réponses sont qualitatives

Les questions à réponses qualitatives sont les plus courantes dans les questionnaires sociologiques. Dans ce cas, les codes n'ont aucune signification en soi : il est indifférent d'attribuer le code 0 à "homme" et 1 à "femme" plutôt que le contraire.

* Variables nominales

On distingue deux cas, selon que les modalités de la réponse s'excluent mutuellement ou non.

Si les n modalités s'excluent, on peut attribuer à chacune d'entre elles un nombre entier, de 0 à n-1.

Exemple : état civil. On a prévu les modalités "célibataire", "marié", "autre", et la possibilité qu'il n'y ait pas de réponse (4 modalités). On pourrait coder

sans réponse	0
célibataire	1
marié	2
autre	3

Ainsi l'état civil est représenté par une seule variable. Si les n modalités ne s'excluent pas, il faut alors représenter les réponses à la question par n ou n-1 variables binaires, qui auront la valeur 1 si la qualité est présente, et 0 autrement.

Cas typique : on présente une liste de films et on demande lesquels ont été vus.

L'exemple précédent (état civil) pourrait être codé de cette façon, ce qui donnerait :

sans réponse	0 0 0
célibataire	1 0 0
marié	0 1 0
autre	0 0 1

On utilise alors trois variables pour représenter l'état civil au lieu d'une. Dans un cas semblable, cette codification est déconseillée au moment de la saisie des données, mais pour certains traitements, elle peut être la seule qui convienne.

* Variables ordinales

Certaines réponses, sans être quantitatives, sont logiquement ordonnées. C'est principalement le cas des échelles, où on demande, par exemple, de choisir entre "refus", "indétermination" et "acceptation", ou d'exprimer une appréciation, avec plus ou moins de nuances, entre des extrêmes tels que "très mauvais" et "très bon", ou de donner une note de 0 à 9.

Ces réponses peuvent être représentées par des variables binaires, mais le plus simple est de les coder en 0 - n, en respectant l'ordre logique des modalités.

2. Questions dont les réponses sont quantitatives

Si on pense utiliser les quantités telles quelles, on peut saisir ces réponses sans codification. Sinon, on va transformer ces variables en variables qualitatives, en formant des classes, selon une des trois méthodes suivantes :

a) Choisir des seuils tels que les longueurs des classes créées soient toujours les mêmes.
   Exemple : une variable a des valeurs comprises entre 0 et 99. On crée cinq classes de longueur 20, avec les seuils 20, 40, 60 et 80.

b) Choisir des seuils tels que les effectifs des différentes classes soient les mêmes, ou à peu près.

c) Choisir des seuils jugés significatifs, sans se préoccuper des longueurs ni des effectifs des classes ainsi définies.

Pour choisir le nombre de classes, il faut faire un compromis entre la perte d'information, qui augmente avec la diminution du nombre de classes, et l'importance des effectifs dans chaque classe. Des effectifs trop faibles conduisent à des résultats aléatoires, ce qui devrait déjà inciter à choïsir un nombre de classes assez faible. D'autre part, des classes de largeur trop petite risquent de poser des problèmes d'interprétation là où il ne devrait pas y en avoir, simplement parce que l'appartenance à une de ces classes ou à une des classes contiguës n'implique pas de différence significative de la grandeur mesurée. Généralement, on choisit un nombre de classes compris entre trois et dix.

Notons qu'il est toujours possible de stocker à la fois les valeurs telles qu'elles sont données dans la réponse et les codes correspondant aux classes.

On peut coder les n classes en une variable qui prend les valeurs 0 à n-1, ou en n (ou n-1) variables binaires, comme les variables nominales. Une autre représentation par des variables binaires, qui peut aussi convenir à des variables ordinales, est la suivante :

Si une observation appartient à la k-ème classe, on donne la valeur 1 aux k (ou k-1) premières variables binaires qui représentent la réponse.

Exemple : on a retenu pour une réponse trois classes. On peut coder cette réponse ainsi :

classe	codage en 0-n	codage par des variables binaires	
	1 variable	3 variables	2 variables
petit	0	1 0 0	0 0
moyen	1	1 1 0	1 0
grand	2	1 1 1	1 1

De cette façon, la distance entre deux observations qui appartiennent à deux classes consécutives est plus petite que celle qui sépare deux observations appartenant à deux classes non contiguës.

3. Dédoublement des variables

* Variables ordinales

Lorsqu'on fait l'analyse des correspondances sur des variables ordinales et qu'on veut attribuer la même importance aux deux valeurs extrêmes, on peut dédoubler ces variables. Pour chaque variable concernée, on en crée une nouvelle de la façon suivante : si MAX est le maximum de la variable et que l'observation a la valeur k pour cette variable, on attribuera à cette observation la valeur MAX - k pour la nouvelle variable.

Exemple à 5 modalités :

	variable initiale	nouvelle variable	
très mauvais	0	4	(4 = 4 - 0)
mauvais	1	3	(3 = 4 - 1)
moyen	2	2	
bon	3	1	
très bon	4	0	

Si la possibilité "sans réponse" existe, et que ce cas a le code 0 dans la variable d'origine, on peut dédoubler cette variable en suivant la règle : si k est la valeur de la première variable et que k = 0, alors la nouvelle variable prend aussi la valeur 0. Sinon, la nouvelle variable vaudra MAX - k + 1. Ce qui donne, par exemple :

sans réponse	0	0
mauvais	1	4
assez mauvais	2	3
plutôt bon	3	2
bon	4	1

* Variables binaires

Si on a des attributs représentés par des variables binaires et qu'on veut donner autant d'importance à l'absence de la qualité qu'à sa présence, on peut dédoubler ces variables.

Exemple : Avez-vous la télévision ?

sans réponse	0	0
oui	1	0
non	0	1

Remarque : la première de ces deux variables représente la possession de la télévision. Si on devait y donner un nom court et explicite (pour une représentation graphique par exemple), on la désignerait probablement par TV. Comme la deuxième représente la non-possession de la télévision, il serait commode de la nommer NTV (ou TVN si on préfère).

4. Exemples d'erreurs à éviter

Donnons enfin deux exemples de mauvaises codifications.

1)

01	revenu simple de moins de 32'000.-
02	revenu simple de 32'000 à 64'000.-
03	revenu simple de plus de 64'000.-
10	revenu double de moins de 32'000.-
20	revenu double de 32'000 à 64'000.-
30	revenu double de plus de 64'000.-

On distingue d'un seul coup d'oeil les revenus simples des revenus doubles, ce qui est très bien. Mais l'ordinateur ne visualise pas ce qui est écrit. Cet exemple devrait être transcrit à l'aide de deux variables, l'une indiquant par 0 ou 1 si le revenu est simple ou double , et l'autre en 0-2 donnant le montant.
Signalons à cette occasion qu'il est parfaitement inutile d'écrire (et de retranscrire sur cartes perforées) 00 ou 01 plutôt que 0 ou 1. La machine lit de toute façon 0 et 1. Si on faisait cela pour 20 variables et 500 enquêtés, on écrirait

10'000 caractères inutiles, soit l'équivalent de 125 cartes pleines, ce qui n'est pas négligeable.

2) Question : Dans quelle faculté êtes-vous inscrits ?

faculté	codification	
	mauvaise	correcte
A	0	1 0 0
B	1	0 1 0
C	2	0 0 1
A et B	3	1 1 0
A et C	4	1 0 1
B et C	5	0 1 1

La codification par une variable en 0-5 est mauvaise, car si on s'intéresse à ceux qui sont inscrits, par exemple, dans la faculté A, il faudra isoler les enquêtés qui ont le code 0 ou 3 ou 4, ce qui est peu pratique. L'exemple ci-dessus devrait être représenté par trois variables binaires, comme indiqué. Par la suite, si on s'intéresse à ceux qui sont inscrits dans deux facultés, on crée une nouvelle variable qui prend la valeur 1 pour les enquêtés qui sont dans ce cas, et 0 autrement.

On peut être tenté de commettre ce genre d'erreur lorsqu'on n'a pas prévu au départ que les réponses à une question pouvaient ne pas s'exclure mutuellement. Si cela arrive, il faut prendre des mesures aussi précoces que possible (avant de remplir les documents de perforation, si ce n'est pas trop tard quand on s'aperçoit de l'erreur, ou au moins au moment de la saisie sur l'ordinateur). En effet, de telles corrections nécessitent un programme ad-hoc.

III. LES TRIS

C'est la méthode la plus simple et la plus souvent utilisée pour dépouiller des questionnaires. Elle consiste à compter le nombre d'observations qui ont telle(s) ou telle(s) caractéristique(s). Les résultats se présentent sous forme de tableaux qui contiennent les effectifs des classes définies par les caractéristiques utilisées.

Nous appelons fréquences les résultats des divisions des effectifs de chaque classe par la somme des effectifs, et profil l'ensemble de ces fréquences.

Exemple : célibataires	mariés	autres	total	
119	341	32	492	(effectifs)
24.2	69.3	6.5	100.	(fréquences, en %)

profil de l'état civil

1. Tris à plat

L'exemple ci-dessus constitue un tri à plat, en ce sens qu'on compte simplement l'apparition des différentes modalités d'une seule variable. Ce genre de tableaux permet de répondre à la question-type "combien", sans donner d'indication sur les relations entre les variables, donc sans apporter d'éléments explicatifs.

Ces décomptes devraient être faits pour toutes les variables pour lesquelles c'est possible au début du dépouillement, en

vue, notamment, de comparer les profils des variables d'identification des enquêtés aux profils de ces mêmes variables dans la population étudiée, afin de s'assurer de la représentativité de l'échantillon. D'autre part, ce travail permet d'éliminer du dépouillement des questions inintéressantes parce qu'elles sont restées la plupart du temps sans réponse, ou que les enquêtés y ont presque tous répondu de la même façon.

Une autre utilisation du tri à plat est intéressante : quand, d'une manière ou d'une autre, on a affecté une partie des individus de l'échantillon à un groupe, on peut chercher efficacement les caractéristiques de ce groupe en faisant les mêmes tris à plat sur le groupe et sur l'ensemble de l'échantillon pris comme population de référence.

Par exemple, si on obtient les chiffres suivants :

célibataires	mariés	autres	
24.2	69.3	6.5	fréquences dans l'échantillon
35.8	62.2	2.1	fréquences dans le groupe

on voit que le groupe se caractérise par la forte proportion de célibataires.

En répétant ce travail avec d'autres variables, on peut se faire une bonne image du groupe étudié.

2. Tris croisés

Par cette démarche, on met un ensemble de caractéristiques en relation avec un autre ensemble de caractéristiques.

Exemple : Formation (plus haut titre obtenu) - Sexe

	Homme	Femme	Total
Sans objet	20	22	42
Aucun	53	103	156
App. arts et métiers	69	20	89
App. de commerce	18	40	58
Maîtrise + école prof.	48	44	92
Maturité	4	16	20
E.T.S., Uni., Poly.	18	1	19
Autre	8	8	16
Total	238	254	492

La case (i,j) de ce tableau indique le nombre de personnes qui ont à la fois la i-ème modalité de la variable "formation", et la j-ème de la variable "sexe".

On complète ce genre de tableaux par ceux des pourcentages sur les lignes et les colonnes, soit

	Homme	Femme	Total
Sans objet	47.6	52.4	100.
Aucun	34.0	66.0	100.
App. arts et métiers	77.5	22.5	100.
App. de commerce	31.0	69.0	100.
Maîtrise + école prof.	52.2	47.8	100.
Maturité	20.0	80.0	100.
E.T.S., Uni., Poly.	94.7	5.3	100.
Autre	50.0	50.0	100.
Total	48.4	51.6	100.

et

	Homme	Femme	Total
Sans objet	8.4	8.7	8.5
Aucun	22.3	40.6	31.7
App. arts et métiers	29.0	7.9	18.1
App. de commerce	7.6	15.7	11.8
Maîtrise + école prof.	20.2	17.3	18.7
Maturité	1.7	6.3	4.1
E.T.S., Uni., Poly.	7.6	0.4	3.9
Autre	3.4	3.1	3.3
Total	100.0	100.0	100.0

Nous avons ainsi le profil de la variable "sexe" dans les différents groupes définis par le plus haut titre, ainsi que dans l'ensemble des enquêtés. Réciproquement, le deuxième tableau de pourcentages nous donne les profils de la variable "formation" dans les groupes "hommes" et "femmes".

La lecture de ces tableaux nous apprend, par exemple, que 77.5 % de ceux qui ont fait un apprentissage aux arts et

métiers sont des hommes, que 69.0 % de ceux qui ont fait un apprentissage de commerce sont des femmes. Nous voyons par ailleurs que 29.0 % des hommes interrogés ont fait un apprentissage aux arts et métiers, et que 15.7 % des femmes ont fait un apprentissage de commerce.

Dans cet exemple, les qualités sont données par toutes les modalités de deux variables. Le procédé peut être appliqué au croisement de deux séries quelconques de qualités données par un nombre quelconque de variables. On pourrait par exemple croiser une liste de films avec une liste de pièces de théâtre. Dans ce cas, les lignes et les colonnes du tableau seraient formées en prenant la modalité "vu" de chacune des variables qui représentent les films(pour les lignes) et les pièces de théâtre (pour les colonnes).

Les tris croisés permettent principalement de montrer ou de vérifier les relations qui existent entre deux variables. Ils ont l'avantage de fournir des résultats simples, très proches des données de base. Ils se prêtent beaucoup moins bien à la recherche de relations, car cette recherche suppose l'édition d'une quantité importante de tableaux. En effet, si on croise n variables deux à deux, le nombre de tableaux est de n*(n-1)/2, soit 190 pour n = 20, et il est rapidement difficile de faire la synthèse de l'information obtenue à l'aide de cette seule technique*. Une autre faiblesse de la méthode, c'est qu'elle ne permet pas d'appréhender les interrelations multiples, c'est-à-dire celles qui mettent en jeu plusieurs variables.

* *Les notations sont celles du Fortran: * est le signe de multiplication, / le signe de division.*

* La notion de filtre

En plus des variables qui forment les lignes et les colonnes, on peut prendre en compte un troisième jeu de variables qui servent à limiter l'entrée dans le tableau, en sorte que celui-ci ne porte plus que sur une partie déterminée de l'échantillon. Ces variables forment un filtre. On l'utilise :

- Pour éliminer des tableaux les observations qui ne sont pas concernées. Ainsi, par exemple, on éliminera les entreprises qui ne font pas d'investissements à l'étranger dans l'étude de ce type d'investissement.

- Pour éliminer d'un tableau les observations inintéressantes du point de vue de l'une ou l'autre variable utilisée pour construire ce tableau. Par exemple, on croise les variables "classe sociale" et "revenu". On élimine la modalité "indéterminé" de la variable "classe sociale".

- Pour étudier indépendamment divers groupes de l'échantillon, soit dans le but d'étudier cette sous-population en soi, soit pour essayer de mettre en lumière des relations qui peuvent exister à l'intérieur de certains groupes et disparaître au niveau de l'échantillon entier. Prenons un exemple extréme : deux variables ont été croisées, et on a obtenu le tableau a) ci-dessous. Les tableaux b) et c) mettent en relation les mémes variables, mais dans les sous-populations "hommes" et "femmes".

a) échantillon entier

var. 1 / var. 2	0	1	total
0	20	20	40
1	80	80	160
total	100	100	200

b) hommes

var. 1 / var. 2	0	1	total
0	20	0	20
1	0	80	80
total	20	80	100

c) femmes

var. 1 / var. 2	0	1	total
0	0	20	20
1	80	0	80
total	80	20	100

- Pour ajouter une troisième dimension au tableau, en donnant successivement différentes valeurs au filtre pour les mêmes variables en lignes et en colonnes.

3. Tests du khi-carré

A) Comparaison d'une distribution à une autre

Reprenons un exemple considéré plus haut, avec les effectifs au lieu des fréquences :

célibataires	mariés	autres	total	
119	341	32	492	effectifs dans l'échantillon
86	149	5	240	effectifs dans le groupe

La question qui se pose est de savoir si la distribution dans le sous-échantillon est la même que dans l'échantillon global, pris comme population de référence. Pour y répondre, nous utilisons le test d'ajustement du khi-carré.

Principe : Les fréquences 119/492, 341/492 et 32/492 peuvent être assimilées aux probabilités d'extraire respectivement un célibataire, une personne mariée ou un "autre" lors du tirage au hasard d'une personne dans l'échantillon. Si on prenait au hasard 240 personnes, et qu'on répète l'expérience un grand nombre de fois, on obtiendrait en moyenne 240*(119/492), soit 58.0 célibataires, 166.3 mariés et 15.6 "autres". Toutefois, à chaque expérience apparaîtrait une différence due au hadard entre ces nombres (dits effectifs théoriques) et les effectifs obtenus, différence qu'on peut mesurer. Le test consiste à faire comme si le sous-échantillon avait été tiré au hasard, et à mesurer la différence entre la distribution effective et la distribution théorique. Si cette mesure donne un résultat suffisamment petit, nous dirons que les deux distributions sont les mêmes, sinon nous rejetterons cette hypothèse.

Calculs : Soient

E = 492	l'effectif dans l'échantillon
E(i)	l'effectif de la modalité i dans l'échantillon
e = 240	l'effectif du groupe
e(i)	l'effectif de la modalité i dans le groupe
n = 3	le nombre de modalités

alors $p(1) = E(i)/E$ est la probabilité de la modalité i

et $t(i) = p(i)*e$ l'effectif théorique de la modalité i dans le groupe.

La mesure de la différence entre la distribution théorique et la distribution observée sera calculée par la formule

$$X^2 = \sum_{i=1}^{n} \frac{(e(i) - t(i))^2}{t(i)} = 22.5$$

Il faut maintenant comparer cette valeur à celle donnée par une table du khi-carré. Pour cela, on détermine le nombre de degrés de liberté (noté ddl) et on choisit un seuil raisonnable (noté a).

Dans l'exemple, ddl = 2 = (n - 1) = (nombre de modalités - 1). En effet, l'effectif du groupe étant donné, si deux des e(i) sont connus, le troisième est déterminé.

On choisit généralement a = 0.05, ce qui signifie que si l'hypothèse d'égalité des distributions est vraie, la probabilité que X^2 dépasse la valeur lue dans la table est de 5 %.

Pour ddl = 2 et a = 0.05, la table donne la valeur 5.99. Nous

rejetons donc l'hypothèse d'égalité des deux distributions, ou, en d'autres termes, nous admettons qu'elles sont significativement différentes.

B) Test d'indépendance et tris croisés

Lorsqu'on a construit un tableau de contingence, on peut se demander si les qualités mises en relation 2 à 2 peuvent être considérées comme statistiguement indépendantes ou non. Pour le savoir, nous pouvons utiliser le test du khi-carré.

Principe : Appelons I l'ensemble des qualités qui forment les lignes du tableau et J celui des qualités en colonnes. Nous faisons l'hypothèse que I et J sont indépendants. Nous pouvons alors calculer les effectifs théoriques des cases du tableau et, comme précédemment, mesurer la différence entre ces effectifs et les effectifs observés. Nous dirons que l'hypothèse d'indépendance est fausse si la mesure de cette différence dépasse une certaine valeur, lue dans une table du khi-carré.

Calculs : Soient $e(i,j)$ l'effectif de la i-ème ligne de la j-ème colonne

m le nombre de lignes du tableau

n le nombre de colonnes du tableau

$e(i,.) = \sum_j e(i,j)$ l'effectif de la i-ème ligne

$e(.,j) = \sum_I e(i,j)$ l'effectif de la j-ème colonne

$E = \sum_I e(i,.) = \sum_j e(.,j) = \sum_I \sum_j e(i,j)$ l'effectif total

Comme avant, nous dirons que

$p(i,.) = e(i,.)/E$ est la probabilité d'avoir la qualité i de I

$p(.,j) = e(.,j)/E$ est la probabilité d'avoir la qualité j de J

Si l'hypothèse d'indépendance est vraie, alors

$p(i,j) = p(i,.)*p(.,j)$ est la probabilité d'avoir à la fois les qualités i de I et j de J

et

$t(i,j) = p(i,j)*E$ est l'effectif théorique de la case (i,j) du tableau.

Nous pouvons alors calculer

$$X^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(e(i,j) - t(i,j))^2}{t(i,j)}$$

Si X^2 est plus petit que la valeur fournie par la table du khi-carré à $(m - 1)*(n - 1)$ degrés de liberté pour un seuil donné, nous admettons l'hypothèse d'indépendance, que nous rejetons autrement. On a $(m - 1)*(n - 1)$ degrés de liberté, car dans un tableau à m lignes et n colonnes dont les totaux sont donnés, une ligne et une colonne sont fixées dès que $(m - 1)*(n - 1)$ éléments sont donnés.

Pour le tableau "Formation - Sexe", $X^2 = 73.6$. Nous rejetons l'hypothèse d'indépendance, car la table du khi-carré à 7 degrés de liberté fournit la valeur 14.1 pour un seuil de 5 %.

Remarque : La mesure de la différence entre les effectifs observés et les effectifs théoriques, si l'hypothèse d'indépendance est vraie, est une variable aléatoire qui suit asymptotiquement une loi du khi-carré à un nombre donné de degrés de liberté. Autrement dit, cette variable aléatoire suit d'autant mieux une loi du khi-carré que les effectifs théoriques sont grands. Il faut donc s'assurer que les effectifs théoriques de chaque classe sont suffisamment grands (en pratique plus grand que 5) avant d'appliquer le test, pour s'assurer de sa validité. Si des classes avaient des effectifs théoriques trop petits, il faudrait les grouper entre elles ou avec d'autres classes.

IV. L'ANALYSE FACTORIELLE

1. Principes de la méthode

Soit x le tableau à 14 lignes et 4 colonnes que nous voulons analyser. Nous appelons "variables" les colonnes de ce tableau, et "observations" les lignes.

	V1	V2	V3	V4
1	1	2	5	4
2	6	2	6	2
3	6	7	3	9
4	7	7	1	8
5	9	2	5	9
6	1	3	6	6
7	8	1	8	0
8	0	7	4	5
9	8	2	6	0
10	1	3	5	5
11	6	0	7	1
12	8	8	4	8
13	5	8	0	8
14	7	0	9	1

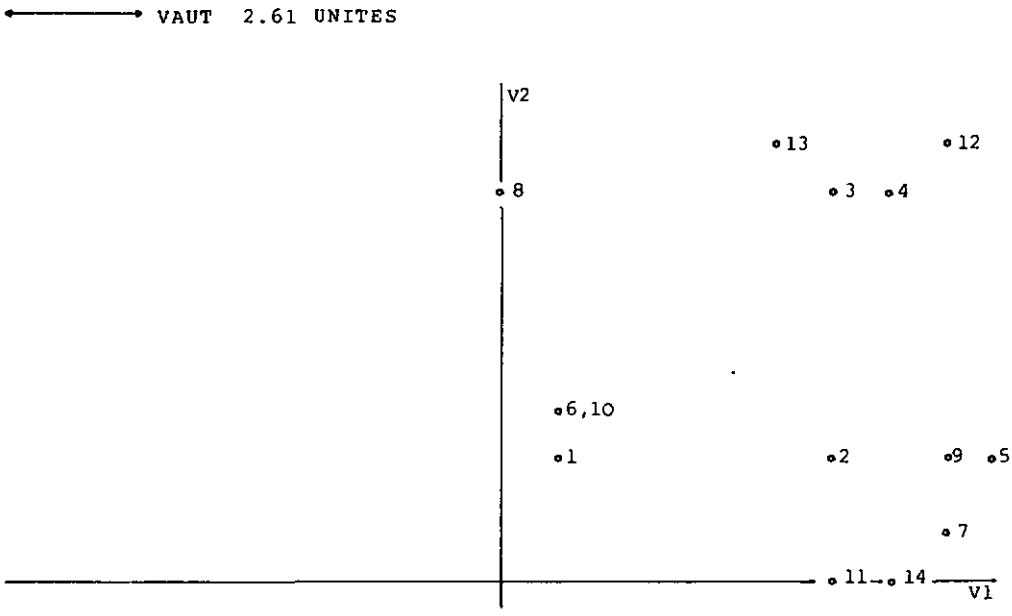
Supposons que ces variables sont ordinales (par exemple des notes de quatre branches attribuées à des élèves, des appréciations de qualités obtenues par différents produits, etc.). Analyser ce tableau, c'est essayer de répondre à des questions telles que :

- Quelles sont les observations (élèves, produits ...) qui se ressemblent ?
- Quelles sont celles qui sont les plus différentes ?
- En quoi se ressemblent-elles ou se différencient-elles ?
- Peut-on former des groupes ?

Avec deux variables, ce travail serait aisé. Il suffirait de considérer chaque observation comme un point dont les coordonnées seraient les valeurs des variables, et de dessiner ces points. Ainsi, en prenant les deux premières colonnes du tableau, on obtient le dessin de la figure 2. Nous pouvons ainsi former quatre groupes de points, dont les caractéristiques sont respectivement d'avoir des grandes valeurs sur les deux axes, sur un seul, et sur aucun axe.

Remarques :

1) Le dessin contient toute l'information donnée dans le tableau pour les 14 observations et les deux premières variables, puisque les observations sont dans un espace à deux dimensions. L'interprétation serait facile, car chaque axe représente simplement une variable. Considérons, par exemple, que les observations sont des élèves, et que les variables sont les notes obtenues dans les branches littéraires pour la première et dans les branches scientifiques pour la seconde. Alors, plus la projection du point qui représente un élève est grande sur le premier axe, plus cet élève est fort dans les matières littéraires. L'élève numéro 5 est le plus fort dans ces matières, le 8 est le plus faible. De même, en ce qui concerne les branches scientifiques, les élèves 12 et 13 sont les plus forts, et 11 et 14 les plus faibles. Les quatre groupes de points auxquels il est fait allusion plus haut contiennent respectivement les élèves forts, moyens et faibles. On constate encore sur le dessin que tous les élèves moyens sont forts dans une matière



. Figure 2

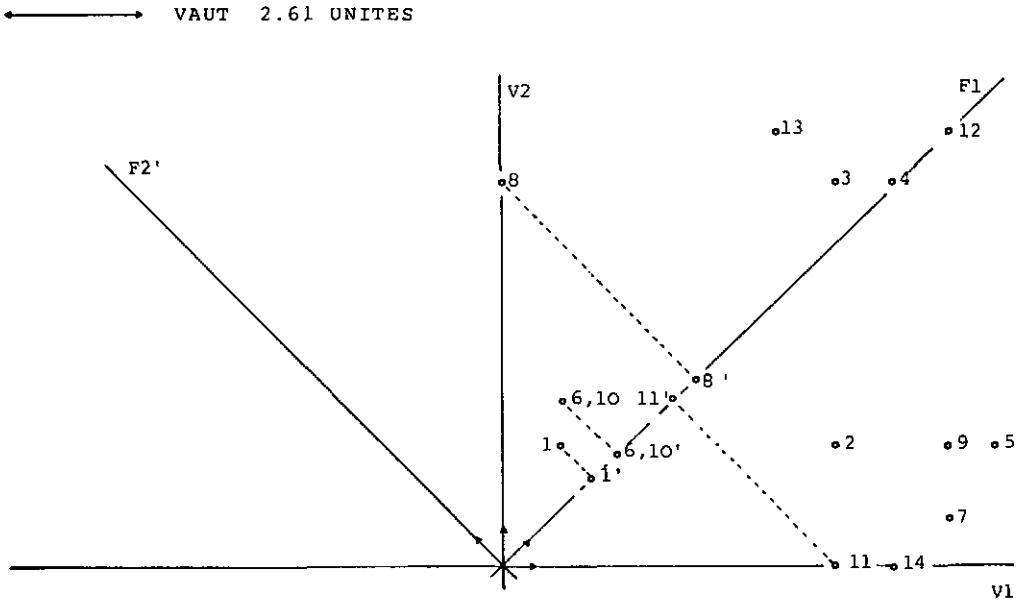


Figure 3

et faibles dans l'autre.

2) Les vecteurs de base (qui engendrent les axes V1 et V2) ont pour composantes (1 0) et (0 1). Si on change la base, par exemple en faisant une rotation autour de l'origine, on ne transforme pas le nuage de points. D'autre part, les vecteurs de la nouvelle base s'expriment (forcément) par rapport à ceux de l'ancienne. Les axes qu'ils engendrent représentent ainsi une combinaison, un "compromis" entre les variables. Dans l'exemple de la figure 3, le vecteur unitaire qui engendre l'axe F1 a pour composantes ($1/\sqrt{2}$, $1/\sqrt{2}$). On peut remarquer en passant que le même axe serait engendré par le vecteur unitaire de composantes ($-1/\sqrt{2}$, $-1/\sqrt{2}$). Si on n'examine que les projections des points sur cet axe, 1', 6' ... (on regarde alors un dessin à une dimension), on ne retient plus qu'une partie de l'information, la même qu'on aurait en calculant la moyenne de chaque observation pour les deux variables utilisées.

A) Analyse factorielle générale

Le principe de l'analyse factorielle est le suivant : puisque les points sont dans un espace à plus de deux dimensions et qu'on ne peut pas les dessiner, on va choisir un petit nombre d'axes qu'on prendra deux à deux pour former des plans, et on dessinera les projections des points sur ces plans. Le choix de ces axes sera fait de façon à perdre le moins d'information possible. On cherche donc un premier axe (facteur) qui passe par l'origine et qui soit tel que l'inertie des projections des points sur cet axe soit maximal. Puis on cherche un deuxième axe selon les mêmes critères et qui soit perpendiculaire au premier, puis un troisième perpendiculaire aux deux pre-

miers, etc. (Les vecteurs unitaires qui engendrent ces axes forment ainsi une nouvelle base orthonormée).

Dans l'analyse factorielle générale, qui est la méthode décrite ici, on définit l'inertie comme la somme des carrés des distances des points à l'origine, la distance utilisée étant la distance euclidienne classique.

En pratique, on ne calcule pas plus de cinq facteurs, et on ne se sert souvent que des deux ou trois premiers.

Les projections sur le plan des deux premiers facteurs obtenus par l'analyse du tableau précédent (avec les quatre variables) donnent le dessin de la figure 4.

Remarques :

1) L'origine correspond aux points dont les coordonnées seraient toutes nulles.
2) Cette représentation n'est pas très satisfaisante pour plusieurs raisons :
   Toutes les projections sur le premier facteur sont positives; cela traduit simplement le fait que les données sont toutes positives, ce que nous savons avant toute analyse.
   Nous voyons la position du nuage des points dans l'espace, mais cela ne nous intéresse pas ; en effet, si on multipliait toutes les valeurs du tableau par deux, par exemple en notant les objets, élèves ou produits, sur 20 au lieu de 10, les points seraient plus éloignés de l'origine alors que le phénomène observé est le même.

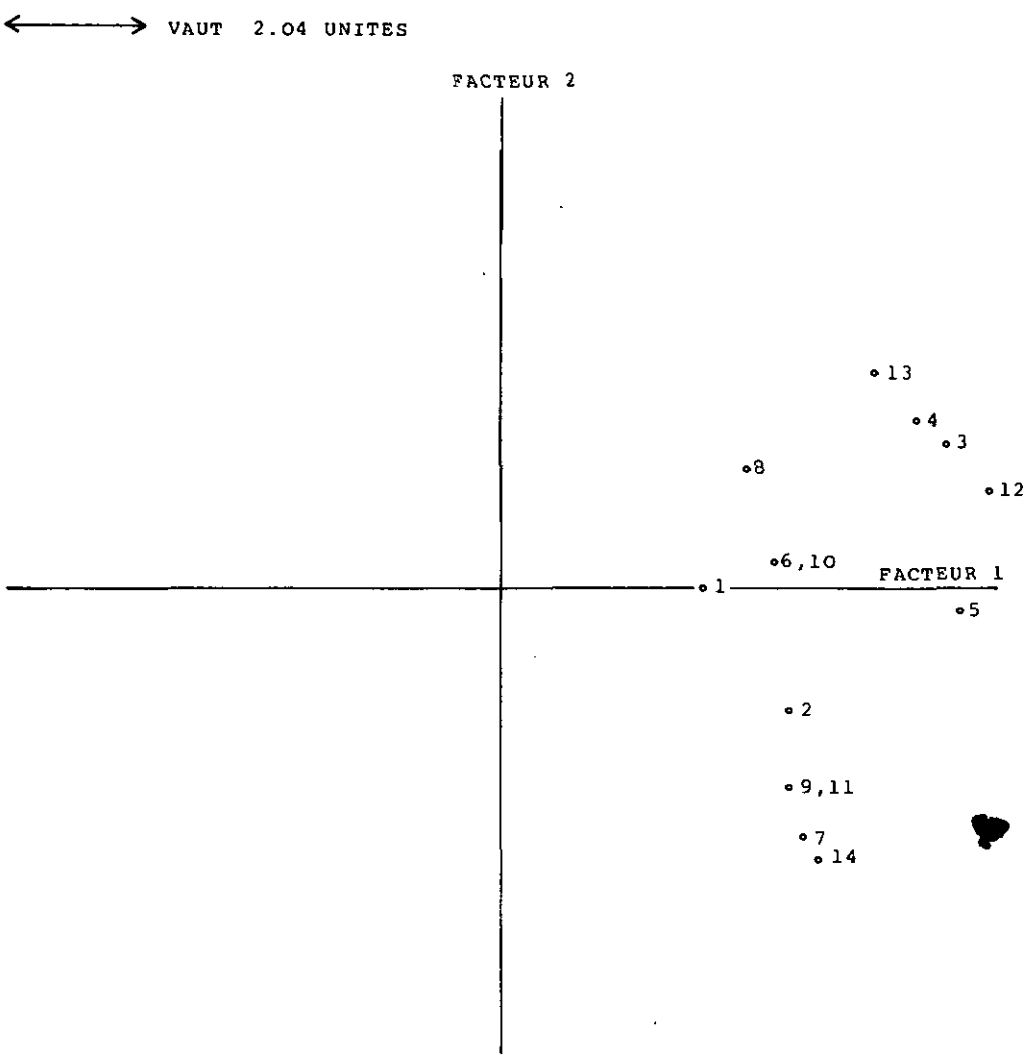


Figure 4

* Analyse en composantes principales sur données centrées

Notons une transformation préalable intéressante des données, qui consiste à soustraire aux éléments du tableau la moyenne de la colonne dans laquelle ils sont (donc à soustraire à chaque variable sa moyenne, donc à centrer les variables). Cette opération a pour effet de déplacer l'origine sur le point dont les coordonnées sont les moyennes des variables. Elle évite un défaut important de l'analyse générale : si une des variables a une moyenne beaucoup plus grande que les autres, la part d'inertie totale qui lui est due est aussi très grande, si bien que le premier facteur est quasiment confondu avec l'axe défini par cette variable.

* Analyse en composantes principales sur données centrées et réduites

Le même défaut peut se retrouver après avoir centré les variables : les données étant les écarts à la moyenne des variables, si une de celles-ci a une dispersion beaucoup plus grande que les autres, son influence sera prédominante dans la détermination du premier facteur. Ce défaut se manifesterait, par exemple, si on mesurait des phénomènes comparables dans des unités très différentes telles que centimètres et kilomètres. D'où cette transformation préalable des données qui consiste à diviser les variables centrées par leur écart-type. La réduction est recommandée pour des variables non homogènes. Toutefois, donner à chaque variable la même dispersion peut faire perdre une information utile, dans la mesure où les variables en jeu peuvent avoir par nature des dispersions différentes.

B) Analyse factorielle des correspondances

C'est le type d'analyse factorielle qui a été utilisée dans les exemples d'application.

Les différences avec l'analyse factorielle générale sont les suivantes :

1) Les coordonnées des points sont les éléments de leur profil. Autrement dit, les points représentés sont les profils (des lignes ou des colonnes). Deux points sont donc proches si leurs profils sont à peu près les mêmes, indépendamment de leurs niveaux. On élimine par là l'effet de taille. L'intérêt est particulièrement évident si on pense à un tableau dont les lignes seraient formées par des groupes contenant plus ou moins d'individus. Prenons par exemple deux observations qui donnent la ventilation par classe de revenu des habitants d'une petite commune (i) et d'une plus grande (j).

	revenu			nombre total
	faible	moyen	élevé	d'habitants
obs. i	44	128	18	190
obs. j	110	320	45	475

On voit bien que dans les communes i et j, la distribution des individus selon les classes de revenus sont strictement les mêmes. Dans l'analyse factorielle des correspondances, les deux points qui représentent ces communes seraient confondus, puisqu'elles ont le même profil (0.23 = 44/190 = 110/475, 0.67, 0.10).

2) Ces points sont munis d'un poids.
Soit $d^2(i,0)$ le carré de la distance du point i à l'origine.
L'inertie du nuage est définie par

$$I = \sum_i d^2(i,0)*p(i)$$

où p(i) est le poids de i.
Le poids de la ligne i est égal à sa somme (son effectif) divisé par la somme des éléments du tableau. La somme des poids est donc égale à 1.

Dans l'analyse générale, l'inertie est définie par

$$I = \sum_i d^2(i,0)$$

ce qui revient à considérer tous les poids égaux à 1. Dans ces conditions, la part d'inertie due à l'observation i est d'autant plus grande que son effectif (la somme de ses composantes; son éloignement de l'origine) est plus grand. Or, les différences d'effectifs entre les observations disparaissent après la transformation en profils (la somme des éléments d'un profil est toujours égale à 1). Les pondérations sont donc introduites pour tenir compte dans le calcul des facteurs des différences d'effectifs. Notons en passant que le fait de dédoubler les variables donne à chaque ligne le même poids.

3) La distance utilisée est une distance pondérée.
Soit x(i,j) le terme général du tableau X

$x(i,.) = \sum_j x(i,j)$ la somme de la ligne i

$x(.,j) = \sum_i x(i,j)$ la somme de la colonne j

$x = \sum_i \sum_j x(i,j)$ la somme des éléments du tableau

alors le carré de la distance euclidienne classique entre les profils de deux observations m et n vaut

$$d^2(m,n) = \sum_j \left(\frac{x(m,j)}{x(m,.)} - \frac{x(n,j)}{x(n,.)} \right)^2$$

Si x(.,t) est très grand par rapport aux autres sommes de colonnes, alors le terme

$$\left(\frac{x(m,t)}{x(m,.)} - \frac{x(n,t)}{x(n,.)} \right)^2$$

va être prépondérant dans la somme ci-dessus. D'où l'utilisation de la distance pondérée, appelée distance du khi-carré

$$d^2(m,n) = \sum_j \left(\frac{x(m,j)}{x(m,.)} - \frac{x(n,j)}{x(n,.)} \right)^2 / \frac{x(.,j)}{x}$$

4) L'origine (le centre d'inertie) correspond au point de profil moyen (profil de la population). Avec les mêmes notations que ci-dessus, le profil moyen est l'ensemble des

$$\frac{x(.,j)}{x} \quad \text{pour } j=1,2, \ldots$$

Il y a différentes raisons d'utiliser l'analyse factorielle des correspondances plutôt que l'analyse en composantes principales dans les dépouillements de questionnaires sociologiques. Nous y reviendrons.

C) Représentation simultanée des observations et des variables

Si on considère le tableau X', transposé du tableau X précédent, nous pouvons y appliquer les mêmes méthodes, en ayant cette fois 4 points dans un espace à 14 dimensions. Dans ce cas, les observations sont les anciennes variables, soit les notes, dans

l'espace engendré par les objets (élèves ou produits). Bien que les observations des tableaux X et X' soient des points qui appartiennent à des espaces différents, il existe des relations telles entre les facteurs de même rang des deux analyses que nous pouvons superposer ces facteurs et avoir ainsi sur le même dessin les projections des lignes et des colonnes de X. Cette superposition se justifie de façon un peu différente selon le type d'analyse :

Dans les deux types d'analyses, l'inertie portée par le k-ème facteur du nuage des observations est la même que celle portée par le k-ème facteur du nuage des variables. Dans l'analyse factorielle des correspondances, la projection de la variable j sur le facteur k est obtenue en calculant la moyenne pondérée des projections des observations sur ce facteur, les coefficients étant les éléments du profil de la variable j, puis en divisant cette moyenne par la racine carrée de l'inertie sur le k-ème facteur. Ce qui se traduit par les formules suivantes :

Soit $g(j,k)$ la projection de la variable j sur le facteur k
$f(i,k)$ la projection de l'observation i sur k
$I(k)$ l'inertie sur le facteur k
$x(.,j)$ l'effectif de la j-ème variable
$x(i,.)$ l'effectif de la i-ème observation

alors

$$g(j,k) = \left(\sum_i \frac{x(i,j)}{x(.,j)} * f(i,k) \right) / \sqrt{I(k)}$$

et réciproquement

$$f(i,k) = \left(\sum_j \frac{x(i,j)}{x(i,.)} * g(j,k) \right) / \sqrt{I(k)}$$

Nous renonçons à commenter le dessin des projections des points (variables et observations) sur le plan formé par les deux premiers facteurs de l'analyse des correspondances du tableau X (fig. 5).

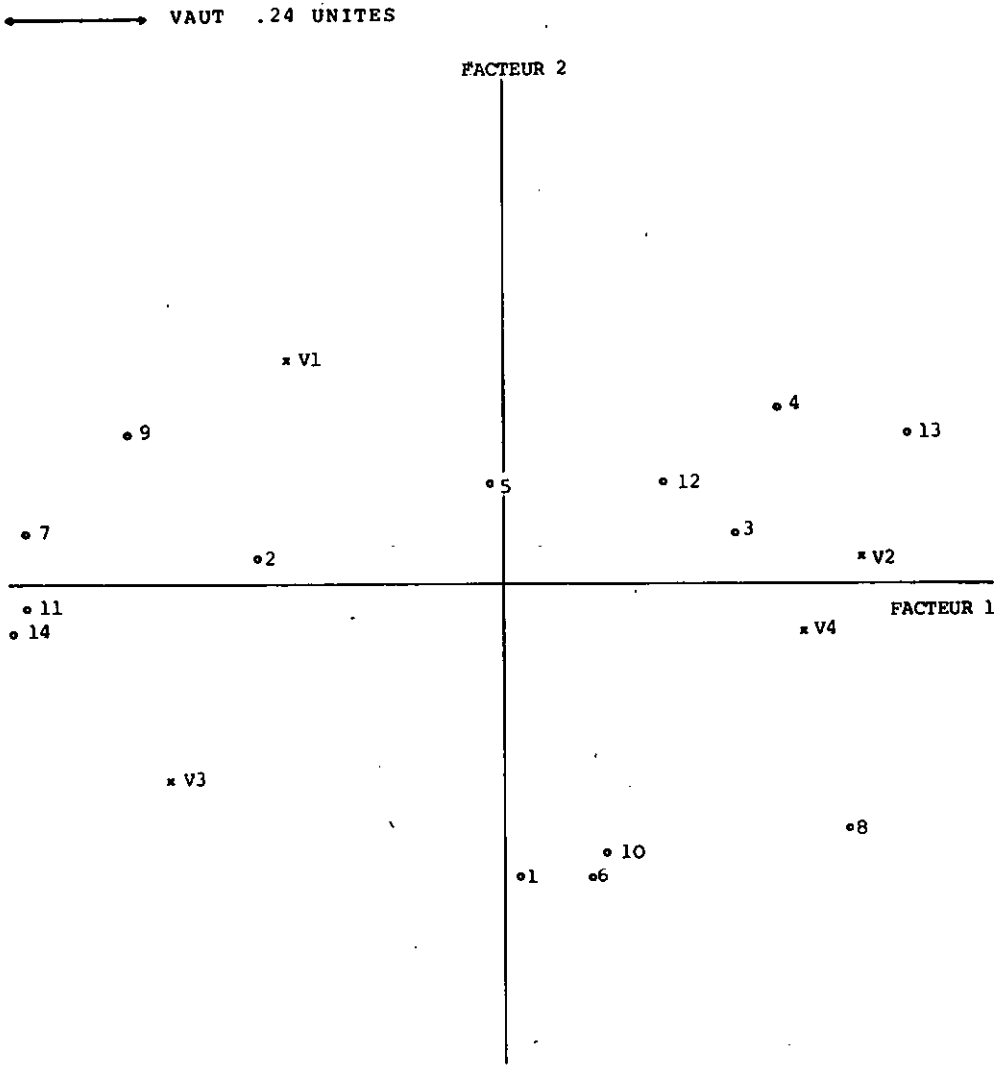

VAUT .24 UNITES
FACTEUR 2
FACTEUR 1
V1
V2
V3
V4
1
2
3
4
5
6
7
8
9
10
11
12
13
14

Figure 5

D) Quelques raisons de préférer l'analyse des correspondances

Outre l'avantage qu'il y a la plupart du temps à éliminer les effectifs pour ne plus comparer que des profils, il convient de noter encore les particularités suivantes dans l'analyse des correspondances :

- Elle donne de meilleurs résultats dans l'analyse des tableaux en 0-1.

- Elle est parfaitement symétrique, en ce sens que les analyses des tableaux X et X' sont parfaitement équivalentes, ce qui n'est pas le cas de l'analyse en composantes principales. Le rôle joué par les lignes et les colonnes est donc le même, ce qui est satisfaisant, surtout si on analyse un tableau de contingence. En effet, dans ce cas, la distinction entre variables et observations est purement formelle. Si deux observations ont le même profil, on peut les remplacer par une seule observation égale à leur somme. Cela ne change rien à l'analyse. Il en est de même pour les variables. Cette propriété explique la stabilité des résultats, lorsqu'on agrège des points de profils semblables. Ceci assure une relative indépendance vis-à-vis du choix nécessairement arbitraire d'une nomenclature. C'est particulièrement appréciable dans les questionnaires lorsqu'il faut former des classes (de revenus, d'âge, ...). En effet, si on déplace légèrement les seuils des classes, on forme des nouvelles classes pas très différentes des précédentes, et ce changement sera peu perceptible dans l'analyse. De même, si on supprime un seuil, on regroupe alors deux classes contiguës, qui sont la plupart du temps assez semblables pour que ce regroupement influence peu l'analyse.

2. Interprétation des résultats

Les dessins des projections des points sur les différents axes factoriels ne donnent pas de résultats à proprement parler. Ils représentent simplement sous une forme synthétique la structure des données. Ils demandent donc à être interprétés. Cette interprétation conduit souvent à formuler des hypothèses plutôt qu'à énoncer des conclusions. En plus des grandeurs décrites dans le paragraphe suivant, les programmes d'analyse factorielle donnent les valeurs propres, qui représentent l'inertie sur les différents facteurs. L'utilisation de ces valeurs pose des problèmes. Il y sera fait allusion dans les remarques qui suivent le premier exemple d'application de l'analyse des correspondances.

A) Aides à l'interprétation

Les programmes d'analyse factorielle fournissent les renseignements décrits ci-dessous, qui facilitent le travail d'interprétation.

1) Pourcentage d'inertie sur un facteur

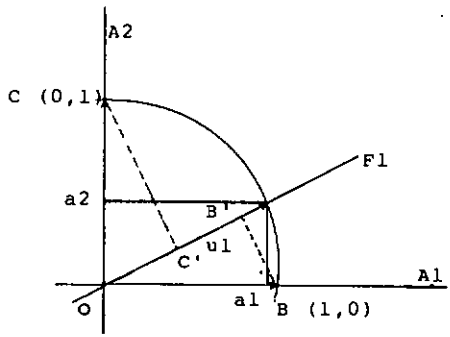
C'est la part d'inertie représentée par un facteur, autrement dit l'inertie sur un facteur divisée par l'inertie totale. C'est donc une mesure de la capacité d'un facteur à illustrer (à expliquer) la structure des données. Il n'y a pas de régles strictes qui permettent d'affirmer que le pourcentage expliqué par un facteur est grand ou petit. En particulier, on ne peut pas décider d'une maniére automatique à partir de quel facteur il faut arréter l'interprétation. Il s'agit donc de juger de cas en cas, compte tenu du fait que les deux premiers facteurs

sont toujours interprétés, en s'aidant éventuellement de règles heuristiques telles que :

- Si on interprête le k-ème facteur, et que le (k+1)-ème a un pourcentage de peu inférieur, il faut aussi essayer de l'interpréter. En effet, si les pourcentages de deux facteurs sont très proches, une légère modification dans la structure des données pourrait inverser l'ordre de ces deux facteurs.
- Symétriquement, si le pourcentage du (k+1)-ème facteur est beaucoup plus petit que celui du k-ème, il ne vaut probablement pas la peine de retenir le (k+1)-ème facteur.
- On arrête l'interprétation au k-ème facteur si on ne trouve pas la signification du (k+1)-ème (on n'interprète donc pas non plus le (k+2)-ème).

2) Contributions des variables (ou observations) aux facteurs (dites contributions absolues)

Cette indication est une mesure de l'influence de chacune des variables dans la détermination d'un axe. Rappelons que les axes factoriels du nuage des observations sont engendrés par des vecteurs unitaires dont les composantes s'expriment par rapport à la base canonique, dans laquelle chaque axe est assimilé à une variable, et voyons un exemple à deux dimensions



Sur le dessin, on se rend compte que l'axe F1 a une direction plus proche de celle de l'axe A1 que de celle de A2. Le carré de la longueur u1, qui vaut 1 par construction, est égal à la somme des carrés de ses composantes sur A1 et A2 (théorème de Pythagore).

Donc $\|u1\|^2 = a1 + a2 = 1$

Par définition, la contribution de A1 à F1 est

$$a1^2 / \|u1\|^2 = a1^2$$

On a ici a1 = 0.9 et a2 = 0.436, donc les contributions de A1 et A2 à F1 sont respectivement 0.81 et 0.19 (81 % et 19 %). On peut voir cette définition de manière différente en considérant les points B et C et leurs projections sur F1, B' et C'. Les distances de B' et C' à l'origine (l'inertie des projections des points B et C sur F1) vaut 1.

Par définition, la contribution de B à F1 est

$$d^2(B',0) / (d^2(B',0) + d^2(C',0)) = 0.81$$

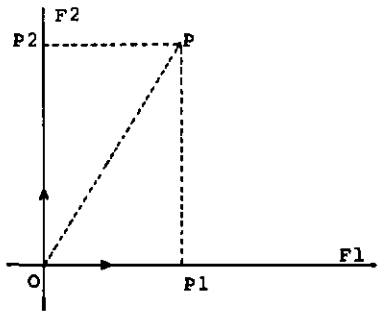
Nous pouvons exprimer cette définition de la façon suivante : La contribution d'un point (variable ou observation) à un facteur est la part de l'inertie sur ce facteur due à la projection de ce point.

Remarque : les poids des points B et C de cet exemple valent 1, mais cette définition reste bien entendu valable pour des points de poids différent (cas de l'analyse des correspondances).

3) Contributions des facteurs aux variables (ou observations)
(dites contributions relatives)

La contribution d'un facteur à un point (observation ou variable) mesure la capacité d'un facteur à rendre compte de la position "réelle" du point.

Exemple à deux dimensions :



Le carré de la distance du point P à l'origine est égal à la somme des carrés de ses projections sur F1 et F2.

Donc $$\|OP\|^2 = P1^2 + P2^2$$

La contribution de F1 à P vaut, par définition

$$P1^2 \ / \ \|OP\|^2$$

Ici on a P1 = 2 et P2 = 4. $\|OP\|^2 = 20$, et les contributions de F1 et F2 à P sont respectivement 0.2 et 0.8 (20 % et 80 %).

B) Quelques considérations quant à l'interprétation

Il n'y a pas de procédure rigide d'interprétation, toutefois, avant de voir des exemples d'interprétation sur des applications concrètes, on peut signaler quelques faits :

- Il est généralement plus facile d'interpréter facteur par facteur que de chercher à identifier directement différentes zones dans le plan.
- Pour interpréter un facteur, il est souvent préférable de sélectionner un petit nombre de variables de part et d'autre de l'origine, celles qui ont les contributions les plus fortes. On cherche alors à qualifier la ressemblance des variables à l'intérieur de chacun des deux groupes, et les différences entre les deux groupes, pour voir ce qui s'oppose sur le facteur.
- On peut interpréter directement la proximité de deux variables ou de deux observations, en prenant garde toutefois que les deux points proches sur un plan se ressemblent d'un certain point de vue, mais qu'ils peuvent être éloignés sur d'autres facteurs.
- En principe, l'interprétation de la position d'une observation se fait par rapport à la position de toutes les variables. Mais si une observation est très proche d'une variable, on peut affirmer que la valeur de l'observation pour cette variable est élevée comparativement aux valeurs des autres variables pour cette observation et, réciproquement, que la valeur de cette variable pour cette observation est grande par rapport aux valeurs qu'elle prend dans les autres observations.
- Toute tentative d'interprétation doit se faire avec la plus grande ouverture d'esprit possible, en faisant intervenir au maximum les connaissances acquises par ailleurs sur le phénomène étudié.

- Si le premier facteur est déterminé très fortement par une seule variable qui s'oppose à toutes les autres, il faut noter le fait, éliminer la variable et recommencer l'analyse. En général cela se produit lorsqu'il y a dans les données une structure évidente. Notons à ce propos que les évidences ne se révèlent souvent comme telles qu'après un traitement statistique adéquat ...
- On dira que deux analyses fournissent les mêmes résultats si les interprétations auxquelles elles conduisent sont les mêmes.
- Il ne faut pas oublier, quand on compare deux analyses, qu'un facteur est toujours défini au signe près.

3. Observations et variables supplémentaires

Une fois que les axes factoriels sont calculés, on peut y projeter des observations et des variables supplémentaires, c'est-à-dire des points qui sont soit des nouveaux points (qui n'ont pas participé à l'élaboration des facteurs), soit des agrégats d'observations ou de variables utilisées dans l'analyse.

La projection de nouvelles observations n'est pas très fréquente dans les dépouillements de questionnaires sociologiques. Un cas typique de l'utilisation de ce procédé se trouve dans l'élaboration de diagnostics : une analyse factorielle est faite, dont les observations sont des malades et les variables leurs résultats à différents tests. Cette population, grâce à laquelle les axes factoriels sont définis, et qui est connue par ailleurs, est utilisée comme population de référence. Un nouveau malade est alors traité comme une observation supplémentaire dont les composantes sont ses résultats aux mêmes tests.

Les projections sur les axes sont alors utilisées pour guider l'établissement de son diagnostic.

La projection d'agrégats est d'une utilisation constante. En effet, chaque fois que le nombre d'observations est grand, on renonce à dessiner leurs projections pour ne plus dessiner que des points représentatifs de groupes, tels que "hommes", "jeunes", etc. Ce genre de projections sera d'ailleurs fait même si le nombre d'enquêtés est assez faible pour que tous les individus puissent être dessinés, dans la mesure où les enquêtes par questionnaires ont pour objet l'étude de groupes de personnes, et pas d'individus particuliers. Dans l'analyse factorielle des correspondances, une observation supplémentaire se projette au centre de gravité des observations qui la constituent. Il en est de même pour les variables supplémentaires. Exemple : voir la figure 6.

4. Conditions d'application de l'analyse des correspondances

Il est possible, algébriquement, de traiter tout tableau dont les éléments sont positifs. Toutefois, pour que l'analyse ait un sens, il ne faut pas mélanger n'importe quoi. Un critère simple et efficace est de former (mentalement) les sommes des lignes et des colonnes du tableau. Si ces sommes ont un sens sémantique, on peut analyser le tableau. En effet, comme on considère les profils des points, les tableaux de pourcentages sur les lignes et les colonnes doivent avoir une signification. On peut toujours interpréter les sommes des lignes et des colonnes d'un tableau composé de 0 et de 1 : supposons que les observations sont des individus et les variables des qualités,

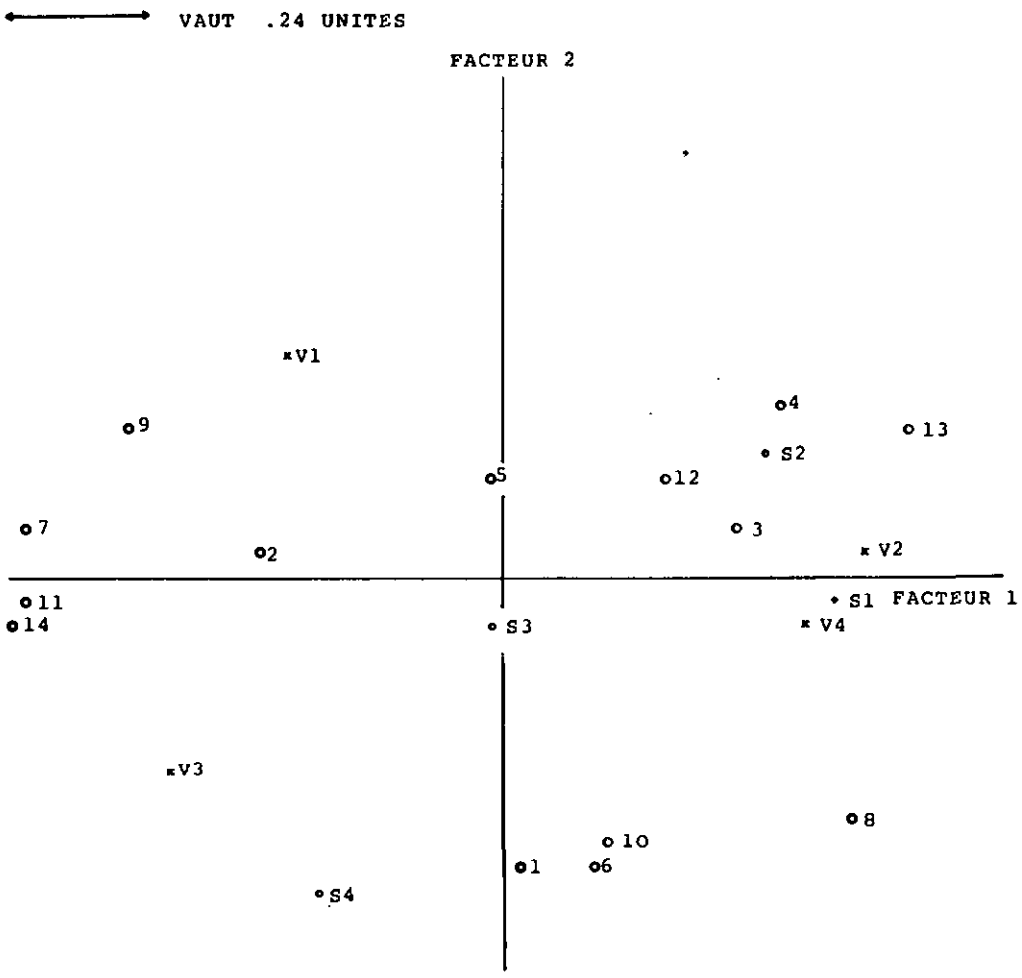


Figure 6

S1 eat une variable supplémentaire composée des variables V2 et V4.

S2 est une observation supplémentaire constituée par les observations 3, 4, 12 et 13.

S3 est constituée dea observations 8 et 9.

S4 est une nouvelle observation, dont les valeurs pour les variables sont 2 2 8 3.

le 1 indiquant que la qualité est présente. Les sommes des lignes donnent le nombre de qualités des individus et les sommes des colonnes le nombre de fois que les qualités sont présentes. Dans ce cas, il faut encore que les variables forment un champ d'investigation homogène pour que l'analyse ait un sens. Notons que les variables exclues de l'analyse n'en sont pas moins utiles : on pourra les projeter en tant que variables supplémentaires ou s'en servir pour définir des observations supplémentaires. Un cas classique : on a des variables qui représentent des activités de loisirs (elles forment le champ d'investigation), et des variables de caractérisation de l'enquêté (sexe, âge, etc). Il ne faudrait pas calculer les facteurs en prenant toutes les variables, mais seulement celles représentant les activités. Les caractéristiques personnelles serviront ensuite à la projection de points représentatifs des différents groupes définis par ces variables ("hommes", "femmes", "jeunes", etc.). Voir le troisième exemple d'application.

5. Exemples d'applications de l'analyse factorielle des correspondances

Pour clore ce chapitre, nous allons voir trois exemples d'applications de la méthode décrite ci-dessus.

Le premier met en jeu un petit nombre de données, et il sera traité en détail. Les autres sont des exemples tirés de l'enquête citée.

A) Premier exemple

Il s'agit d'un tableau à 7 lignes et 6 colonnes qui contient le nombre d'étudiants inscrits à l'Université de Neuchâtel pendant le semestre d'hiver 1977-1978, selon la section et la situation professionnelle des parents (tableau 1).*

Nous avons donc les variables

LETT	faculté des lettres
SCIE	faculté des sciences, sauf médecine
MEDE	médecine
DROI	droit
SEPS	sciences économiques, politiques et sociales
THEO	théologie

et les observations

LIB	professions libérales, directeurs
ENS	enseignants
ART	artisans et commerçants
EMP	employés et fonctionnaires
AGR	agriculteurs et viticulteurs
OUV	ouvriers
AUT	autres (indéterminés, étrangers, parents décédés)

Les principaux résultats fournis par l'analyse factorielle sont reproduits dans le tableau 2. Les composantes des variables et des observations ne nous intéressent pas en tant que telles.

* *Tableau tiré de "La nouvelle donne", étude réalisée par MM. F. Hainard et A. Jeannin, Cahiers de l'ISSP, 1979.*

ETUDIANTS, SELON LA SECTION ET LA SITUATION PROFESSIONNELLE DES PARENTS

	LETT	SCIE	MEDE	DROI	SEPS	TREO	TOTAL
LIB	151	82	33	83	53	7	409
ENS	44	29	3	16	13	2	107
ART	79	36	9	23	31	4	182
EMP	158	88	18	66	42	8	380
AGR	13	11	1	4	4	1	34
OUV	61	54	9	17	25	5	171
AUT	81	40	8	32	26	4	191
TOTAL	587	340	81	241	194	31	1474

	LETT	SCIE	MEDE	DROI	SEPS	TREO	TOTAL
LIB	36.9	20.0	8.1	20.3	13.0	1.7	100.0
ENS	41.1	27.1	2.8	15.0	12.1	1.9	100.0
ART	43.4	19.8	4.9	12.6	17.0	2.2	100.0
EMP	41.6	23.2	4.7	17.4	11.1	2.1	100.0
AGR	38.2	32.4	2.9	11.8	11.8	2.9	100.0
OUV	35.7	31.6	5.3	9.9	14.6	2.9	100.0
AUT	42.4	20.9	4.2	16.8	13.6	2.1	100.0
TOTAL	39.8	23.1	5.5	16.4	13.2	2.1	100.0

	LETT	SCIE	MEDE	DROI	SEPS	TREO	TOTAL
LIB	25.7	24.1	40.7	34.4	27.3	22.6	27.7
ENS	7.5	8.5	3.7	6.6	6.7	6.5	7.3
ART	13.5	10.6	11.1	9.5	16.0	12.9	12.3
EMP	26.9	25.9	22.2	27.4	21.6	25.8	25.8
AGR	2.2	3.2	1.2	1.7	2.1	3.2	2.3
OUV	10.4	15.9	11.1	7.1	12.9	16.1	11.6
AUT	13.8	11.8	9.9	13.3	13.4	12.9	13.0
TOTAL	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Tableau 1

Facteurs	1	2	3	4	5
Valeurs propres	.015	.005	.004	.000	.000
Pourcentage	.610	.209	.170	.008	.002
Pourcentage cumulé	.610	.819	.989	.998	1.000

COMPOSANTES DES VARIABLES

	1	2	3	4	5
LETT	-.01	-.07	-.01	.01	.00
SCIE	-.15	.08	-.04	-.01	.00
MEDE	.23	.18	.11	.03	-.01
DROI	.20	.01	-.07	-.01	.01
SEPS	-.03	-.02	.13	-.02	.00
THEO	-.16	.04	.03	.06	.04

COMPOSANTES DES OBSERVATIONS

	1	2	3	4	5
LIB	.16	.06	.02	.00	.00
ENS	-.12	-.03	-.08	-.04	-.01
ART	-.04	-.10	.13	.00	-.01
EMP	.01	-.03	-.07	.02	.00
AGR	-.24	.06	-.08	.00	.02
OUV	-.22	.12	.04	.00	.00
AUT	.00	-.08	-.01	-.01	.02

CONTRIBUTIONS DES VARIABLES AUX FACTEURS

	1	2	3	4	5
LETT	.001	.364	.005	.129	.102
SCIE	.337	.281	.103	.034	.015
MEDE	.194	.335	.155	.168	.094
DROI	.422	.003	.182	.112	.118
SEPS	.010	.012	.552	.247	.047
THEO	.036	.005	.003	.310	.625

CONTRIBUTIONS DES FACTEURS AUX VARIABLES

	1	2	3	4	5
	.011	.962	.010	.014	.000
	.728	.208	.062	.001	.000
	.547	.324	.122	.007	.001
	.887	.002	.107	.003	.001
	.057	.024	.898	.020	.001
	.798	.038	.020	.095	.049

CONTRIBUTIONS DES OBSERVATIONS AUX FACTEURS

	1	2	3	4	5
LIB	.447	.204	.024	.032	.001
ENS	.066	.016	.112	.525	.208
ART	.016	.240	.479	.013	.078
EMP	.001	.036	.294	.359	.045
AGR	.088	.015	.035	.000	.114
OUV	.382	.322	.055	.010	.005
AUT	.000	.166	.001	.061	.551

CONTRIBUTIONS DES FACTEURS AUX OBSERVATIONS

	1	2	3	4	5
	.853	.134	.013	.001	.000
	.595	.051	.282	.065	.007
	.069	.353	.576	.001	.001
	.006	.123	.820	.049	.002
	.853	.049	.095	.000	.004
	.752	.217	.030	.000	.000
	.004	.946	.005	.014	.032

Tableau 2

Nous ne nous en servons que pour faire les dessins de la figure 7. Nous constatons, en examinant les pourcentages d'inertie portés par les différents facteurs, que les deuxième et troixième axes ont des valeurs assez proches (20.9 et 17.0 %), et que ces valeurs sont grandes par rapport à celle relative au quatrième facteur. Ceci doit nous inciter à priori à chercher une interprétation du troisième facteur aussi. Pour interpréter le premier axe, nous retiendrons les trois variables qui contribuent le plus fortement au facteur, soit DROI (42.2 %), SCIE (33.7 %) et MEDE (19.4 %), ainsi que les observations LIB (44.7 %), et OUV (38.2 %). Nous constatons que les points SCIE et OUV s'opposent aux points DROI, MEDE et LIB. Nous pouvons tirer à partir de là un certain nombre de conclusions "mécaniques" :

- Les profils de DROI et MEDE sont assez semblables; ils diffèrent beaucoup du profil de SCIE.

- La proximité des points OUV et SCIE suggère que les enfants d'ouvriers s'inscrivent dans une proportion relativement plus grande en sciences qu'ailleurs, ou réciproquement, que les sciences "recrutent" plus d'enfants d'ouvriers que d'autres catégories. On obtient la confirmation de cela en regardant les tableaux en pourcentages sur les lignes et les colonnes associés au tableau analysé : 23.1 % de l'ensemble des étudiants sont inscrits en sciences, alors que cette proportion est de 31.6 % chez les fils d'ouvriers. Ces derniers représentent 11.6 % du total des étudiants, et 15.9 % des étudiants en sciences.

- On peut faire les mêmes remarques pour l'association DROI, MEDE et LIB.

Constatons encore que OUV correspond à des étudiants dont les parents ne sont pas licenciés, au contraire de LIB. Nous

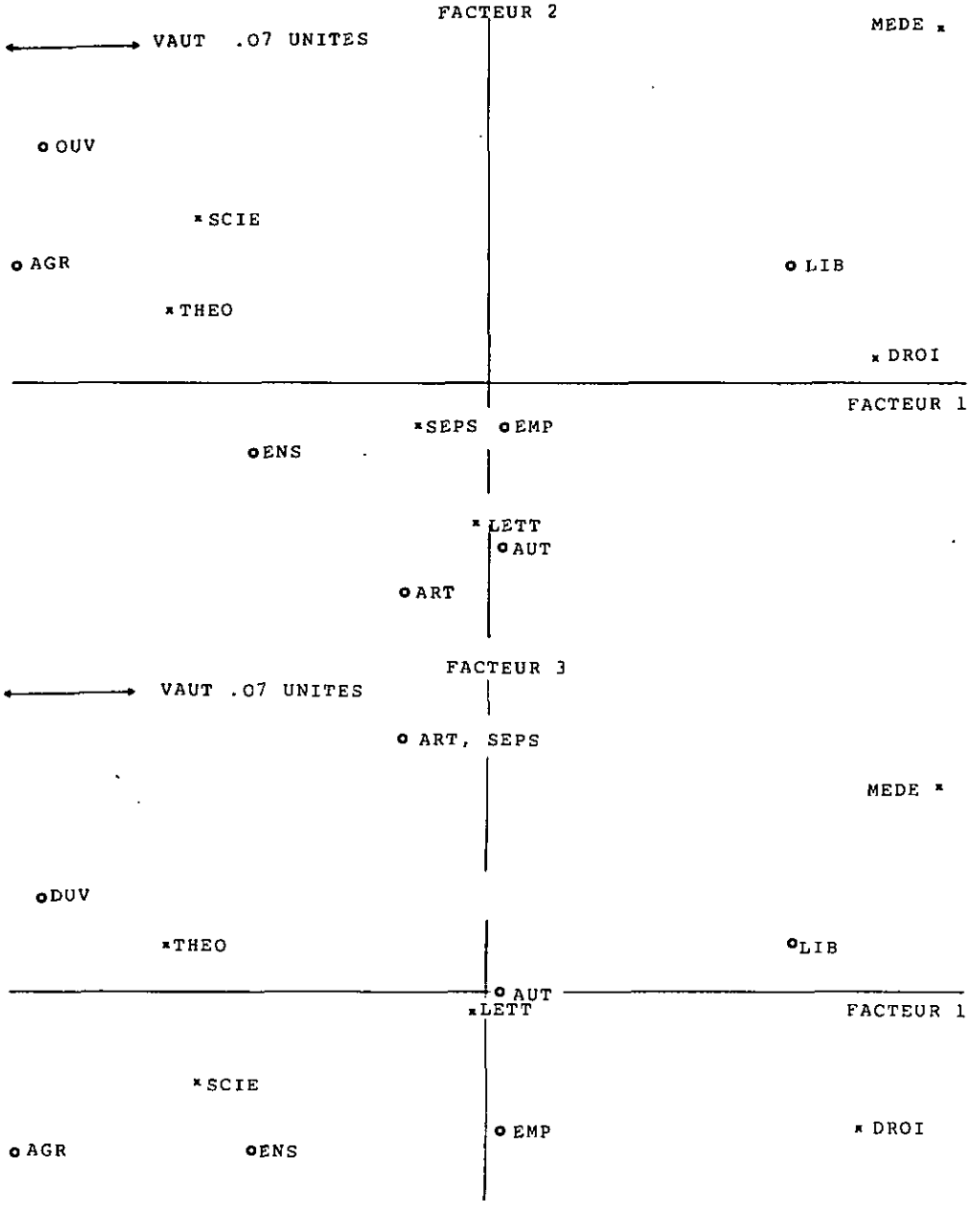

FACTEUR 2
VAUT .07 UNITES
MEDE
OUV
SCIE
AGR
LIB
THEO
DROI
FACTEUR 1
SEPS
EMP
ENS
LETT
AUT
ART
FACTEUR 3
VAUT .07 UNITES
ART, SEPS
MEDE
DUV
THEO
LIB
AUT
LETT
FACTEUR 1
SCIE
EMP
DROI
AGR
ENS

Figure 7

pouvons donc interpréter le premier facteur comme un axe de mobilité sociale, la voie préférentielle du changement étant les sciences, et les voies préférentielles de la reproduction (changement nul) étant le droit et la médecine. Cette interprétation est confirmée par la situation du point AGR, qui est très bien représenté par sa projection sur le premier facteur (la contribution de ce facteur à cette observation est de 85.3 %). Nous pouvons encore constater la ressemblance des points SCIE et THEO, ce dernier étant lui aussi bien situé par sa projection sur le premier axe. Une démarche semblable nous permet d'interpréter le deuxième facteur comme un axe de plus ou moins grand déterminisme (déterminisme de recrutement pour les sections, et de choix pour les catégories). Nous remarquerons simplement à ce propos que le point AUT, qui représente une catégorie composite par excellence, est représenté à 94.6 % par sa projection sur cet axe, et qu'il est situé, bien entendu, du côté de l'axe qui correspond au déterminisme le plus faible. Le troisième facteur opère la distinction entre les anciennes classes moyennes (artisans et commerçants, associées à SEPS), et les nouvelles (employés et fonctionnaires, associées à DROI). On notera que les projections des points EMP et AGR sont trés proches sur ce facteur, mais que AGR ne doit pas intervenir dans l'interprétation, au contraire de EMP, qui est pourtant plus prés de l'origine. Ceci découle des contributions respectives de ces points à ce facteur.

Remarques importantes :

1) L'analyse des correspondances met en lumière des structures sociologiquement significatives que la lecture du tableau, pourtant petit, ne permet pas de déceler facilement. Conclusion (heureuse) : la méthode est utile.

2) La lecture des tableaux montre de façon immédiate que la faculté des lettres est celle qui compte le plus grand nombre d'étudiants (4/10), et que cette majorité relative est vraie dans toutes les catégories d'étudiants. Cette information disparaît dans l'analyse.

3) La position du point AUT montre à posteriori que cette catégorie est bien aussi composite qu'on pouvait le penser à priori. Si ce n'était pas le cas, ce point se rapprocherait d'un des points représentant une catégorie bien définie.

4) On peut constater à la lecture des tableaux en pourcentage que les structures dégagées par l'analyse ne sont pas très marquées. Ceci nous donne l'occasion de faire les observations suivantes :
Le calcul du khi-carré (après avoir supprimé la ligne AGR et la colonne THEO, dans lesquelles certains effectifs théoriques sont inférieurs à 5) donne 33.0 pour 20 degrés de liberté, et la valeur de la table est de 31.4 au seuil de 5 %. L'hypothèse d'indépendance des lignes et des colonnes est donc tout juste rejetée, du point de vue de ce test. Comment faire pour augmenter la valeur du khi-carré en changeant artificiellement les données, mais sans les dénaturer ? La première possibilité est de multiplier tous les éléments du tableau par une constante plus grande que 1. Dans ce cas, l'augmentation de la valeur du khi-carré se comprend aisément : en effet, si les structures qu'on observe ici sur 1474 individus étaient exactement les mêmes dans une population par exemple cinq fois plus grande (7370 personnes), la probabilité que ces structures ne soient dues qu'au hasard diminue considérablement. Or, si on fait l'analyse des correspondances du tableau ainsi transformé, on obtient exactement les mêmes résultats qu'avant. On peut donc déjà déduire de cela que les structures sont dégagées par l'analyse des correspondances indépendamment de leur validité statistique, au sens du khi-carré. Un autre moyen d'augmenter

la valeur du khi-carré est d'augmenter le contraste du tableau, sans changer les sommes des lignes ni des colonnes, en augmentant ou en diminuant l'effectif de chaque case du tableau de la façon suivante : on calcule pour chaque case la différence entre l'effectif observé et l'effectif théorique (toujours au sens du khi-carré). On ajoute ensuite à l'effectif théorique cette différence multipliée par une constante plus grande que 1 pour obtenir l'effectif transformé de la case. Ainsi, les cases du tableau dont l'effectif est plus grand que l'effectif théorique vont être augmentées (la différence calculée est positive), tandis que les autres cases auront leurs effectifs diminués, et cela d'autant plus que la différence calculée était grande. On accentue ainsi les tendances de déviation entre l'observation réelle et la situation théorique en cas d'indépendance parfaite des lignes et des colonnes (tableau 3). Il est de nouveau aisé de comprendre que la valeur du khi-carré augmente à la suite de cette transformation. Si nous comparons les résultats de l'analyse des correspondances sur ce tableau à ceux que nous avions avant (tableau 4 et 2), nous constatons que les valeurs propres ont toutes été multipliées par un facteur $\sqrt{k}$ (environ 4), et que toutes les projections ont été multipliées par k. Les autres résultats n'ont pas changé. Les dessins auraient exactement la même apparence qu'avant (seules les échelles ont changé). Ceci entraîne que l'interprétation des résultats de cette analyse va être parfaitement la même que celle trouvée à propos du tableau non transformé. On peut tirer de cela une deuxième déduction, à savoir que les résultats utilisables fournis par l'analyse des correspondances (dessins, pourcentages d'inertie et contributions) sont indépendants du contraste qu'il y a dans le tableau étudié, pour une structure donnée. Pourquoi alors ne pas se servir des valeurs propres, qui, rappelons-le, donnent l'inertie sur les facteurs, comme mesure de contraste des structures dégagées ? En fait, il n'est pas conseillé de le faire, car l'inertie ne dépend pas seulement des contrastes réels que présente le phénomène étudié,

mais aussi, dit d'une manière très grossière, de la codification des données.

Conclusions pratiques de cette discussion :

- L'intensité des tendances dégagées par l'analyse factorielle doit être recherchée en dehors de cette méthode (dans cet exemple, en relisant à la lumière des résultats de l'analyse les tableaux en pourcentages).
- Les résultats d'une analyse sont réputés valides si le spécialiste qui fait l'interprétation les juge satisfaisants, que ce soit pour tirer des conclusions ou formuler des hypothèses.

B) Deuxième exemple

Il s'agit d'une étude de la composition de la population selon l'enquête de La Chaux-de-Fonds. Il porte sur 7 caractéristiques personnelles, soit le sexe, l'âge, le revenu, l'état civil, l'activité, la classe sociale et la formation, indiquée par le plus haut titre obtenu. Ces caractéristiques ont fourni 42 variables binaires, décrites dans le tableau 5. Les observations sont les 492 enquêtés. Comme on peut le voir sur le dessin de la figure 8, le premier facteur oppose les jeunes (16-19 ans, apprentis, étudiants, célibataires) au reste de la population. Le deuxième facteur est un axe d'échelle sociale : on a d'un côté les revenus élevés, les études supérieures, les bourgeois, et de l'autre les revenus inférieurs, le niveau d'instruction primaire et les ouvriers. On peut noter que les apprentis et les étudiants, très proches sur le premier facteur, sont éloignés sur le deuxième. Comme on pouvait s'y attendre, les étudiants sont du côté "catégories sociales élevées" du deuxième facteur. On remarque encore que les points "veufs" et "divorcés-séparés" sont très proches sur les trois premiers axes, du côté "ouvrier" du deuxième facteur.

TABLEAUX APRES AUGMENTATION DU CONTRASTE

	LETT	SCIE	MEDE	OROI	SEPS	THEO	TOTAL
LIB	138.6	69.1	44.0	99.8	52.1	5.3	409.0
ENS	45.4	33.5	.0	14.4	11.9	1.7	107.0
ART	85.8	29.8	8.0	16.0	38.3	4.2	182.0
EMP	164.9	88.4	15.0	70.0	33.7	8.0	380.0
AGR	12.4	14.3	.1	2.4	3.5	1.3	34.0
OUV	53.6	69.2	8.6	5.6	27.6	6.5	171.0
AUT	86.1	35.8	5.4	32.8	26.9	4.0	191.0
TOTAL	587.0	340.0	81.0	241.0	194.0	31.0	1474.0

	LETT	SCIE	MEDE	OROI	SEPS	THEO	TOTAL
LIB	33.9	16.9	10.7	24.4	12.7	1.3	100.0
ENS	42.5	31.3	.0	13.5	11.1	1.6	100.0
ART	47.1	16.4	4.4	8.8	21.1	2.3	100.0
EMP	43.4	23.3	3.9	18.4	8.9	2.1	100.0
AGR	36.6	42.0	.3	7.0	10.3	3.8	100.0
OUV	31.3	40.4	5.0	3.3	16.1	3.8	100.0
AUT	45.1	18.7	2.8	17.2	14.1	2.1	100.0
TOTAL	39.8	23.1	5.5	16.4	13.2	2.1	100.0

	LETT	SCIE	MEDE	OROI	SEPS	THEO	TOTAL
LIB	23.6	20.3	54.3	41.4	26.9	17.2	27.7
ENS	7.7	9.9	.0	6.0	6.1	5.6	7.3
ART	14.6	8.8	9.8	6.6	19.8	13.5	12.3
EMP	28.1	26.0	18.5	29.1	17.3	25.8	25.8
AGR	2.1	4.2	.1	1.0	1.8	4.2	2.3
OUV	9.1	20.3	10.6	2.3	14.2	20.8	11.6
AUT	14.7	10.5	6.7	13.6	13.9	12.8	13.0
TOTAL	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Tableau 3

Facteurs	1	2	3	4	5
Valeurs propres	.063	.021	.018	.001	.000
Pourcentage	.610	.209	.170	.008	.002
Pourcentage cumulé	.610	.819	.989	.998	1.000

COMPOSANTES DES VARIABLES

	1	2	3	4	5
LETT	-.01	-.14	-.01	.02	-.01
SCIE	-.30	.16	-.09	.01	.00
MEDE	.47	.36	.22	.04	-.02
DROI	.40	.02	-.14	.02	.01
SEPS	-.07	-.04	.27	.04	.01
THEO	-.33	.07	.05	.11	.08

COMPOSANTES DES OBSERVATIONS

	1	2	3	4	5
LIB	.32	.13	.04	-.01	.00
ENS	-.24	-.07	-.16	-.08	-.03
ART	.09	.20	.26	.01	.01
ENP	.01	-.05	-.14	.03	.01
AGR	-.49	.12	-.16	.00	.03
OUV	-.45	.24	.09	.01	.00
AUT	.01	-.17	.01	.02	.03

CONTRIBUTIONS DES VARIABLES AUX FACTEURS

	1	2	3	4	5
LETT	.001	.364	.005	.129	.102
SCIE	.337	.281	.103	.034	.015
MEDE	.194	.335	.155	.168	.094
DROI	.422	.003	.182	.112	.118
SEPS	.010	.012	.552	.247	.047
TREO	.036	.005	.003	.310	.625

CONTRIBUTIONS DES FACTEURS AUX VARIABLES

	1	2	3	4	5
	.011	.962	.010	.014	.003
	.728	.208	.062	.001	.000
	.547	.324	.122	.007	.001
	.887	.002	.107	.003	.001
	.057	.024	.898	.020	.001
	.798	.038	.020	.095	.049

CONTRIBUTIONS DES OBSERVATIONS AUX FACTEURS

	1	2	3	4	5
LIB	.447	.204	.024	.032	.001
ENS	.066	.016	.112	.525	.208
ART	.016	.240	.479	.013	.078
EMP	.001	.036	.294	.359	.045
AGR	.088	.015	.035	.000	.114
OUV	.382	.322	.055	.010	.005
AUT	.000	.166	.001	.061	.551

CONTRIBUTIONS DES FACTEURS AUX OBSERVATIONS

	1	2	3	4	5
	.853	.134	.013	.001	.000
	.595	.051	.282	.065	.007
	.069	.353	.576	.001	.001
	.006	.123	.820	.049	.002
	.853	.049	.095	.000	.004
	.752	.217	.030	.000	.000
	.004	.946	.005	.014	.032

Résultats de l'analyse des correspondances du tableau contrasté

Tableau 4

No	Caractéristiques	Modalités	Abréviations
1	Sexe	Hommes	MASC
2		Femmes	FEMI
3	Age	60 - 64 ans	AGE0
4		55 - 59 ans	AGE1
5		50 - 54 ans	AGE2
6		45 - 49 ans	AGE3
7		40 - 44 ans	AGE4
8		35 - 39 ans	AGE5
9		30 - 35 ans	AGE6
10		25 - 29 ans	AGE7
11		20 - 25 ans	AGE8
12		16 - 19 ans	AGE9
13	Revenu	- 10'000	REV0
14		10 - 15'000	REV1
15		15 - 20'000	REV2
16		20 - 25'000	REV3
17		25 - 30'000	REV4
18		30 - 35'000	REV5
19		35 - 40'000	REV6
20		40 - 50'000	REV7
21		50 - 100'000	REV8
22		100'000 et plus	REV9
23	Etat-civil	Célibataires	EC1
24		Mariés	EC2
25		Veufs	EC3
26		Divorcés - Séparés	EC4
27	Formation	Ménagères	A1
28		Etudiants	A2
29		Apprentis	A3
30		Ex. un métier	A4
31		Divers	A5
32	Classe	Classe indéterminée	C0
33		Bourgeois	C1
34		Classe moyenne	C2
35		Ouvriers	C3
36	Titre	Aucun titre	T1
37		Apprentissage arts et métiers	T2
38		Apprentissage de commerce	T3
39		Maîtrise	T4
40		Ecole professionnelle	T5
41		Maturité	T6
42		E.T.S.	T7
43		Uni. - Poly.	T8
44		Autres	T9

Tableau 5

CARACTERISTIQUES PERSONNELLES

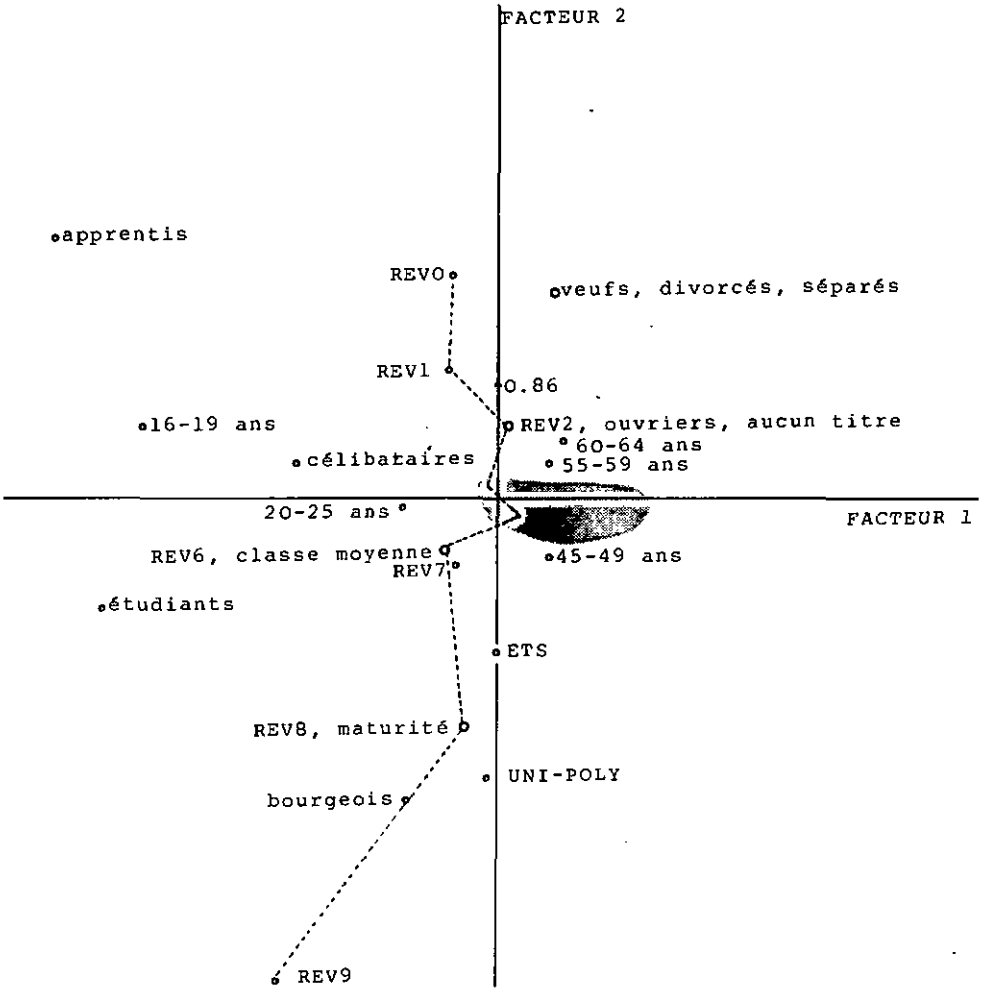


Figure 8

Le troisième facteur, qui n'est pas reproduit ici, distingue les hommes des femmes.

Remarques :

1) Une caractéristique comme la jeunesse est en fait saisie à travers plusieurs questions : l'âge, bien sûr, mais aussi l'activité (modalités "étudiants" ou "apprentis"), le plus haut titre ("sans objet") et l'état civil ("célibataires").

2) La variable "revenu", qui est par essence une variable ordinale, a été codée ici en dix variables binaires comme si c'était une variable nominale. Néanmoins, ces dix variables se projettent bien en ordre de revenu décroissant le long du deuxième facteur (ces variables sont reliées par un trait sur le dessin de la figure 8).

C) Troisième exemple

Une liste de 92 activités de loisirs a été soumise aux enquêtés de La Chaux-de-Fonds, qui étaient invités à donner un certain nombre d'indications sur chacune de ces activités, dont le fait de pratiquer l'activité ou non. Les activités pratiquées par moins de 10 personnes (environ 2 % de la population) ont été éliminées de l'étude. Les observations de cet exemple sont donc les 492 enquêtés, et les variables les 68 activités de loisirs pratiquées par 10 personnes au moins. Le code des variables est 1 si l'activité est pratiquée, 0 autrement.
Les personnes pratiquent en moyenne 22.5 activités, et chaque activité est pratiquée en moyenne par 120 personnes. Les activités les plus pratiquées sont la promenade (438 personnes), la lecture des journaux et revues (426), les vacances hors du domicile (423), l'écoute de la radio et de la télévision (403 les deux). On trouve sur le premier facteur (fig. 9) la distinction entre les loisirs populaires et les loisirs qu'on pourrait appeler "élitistes". Les variables qui déterminent le plus

PLAN DES LOISIRS

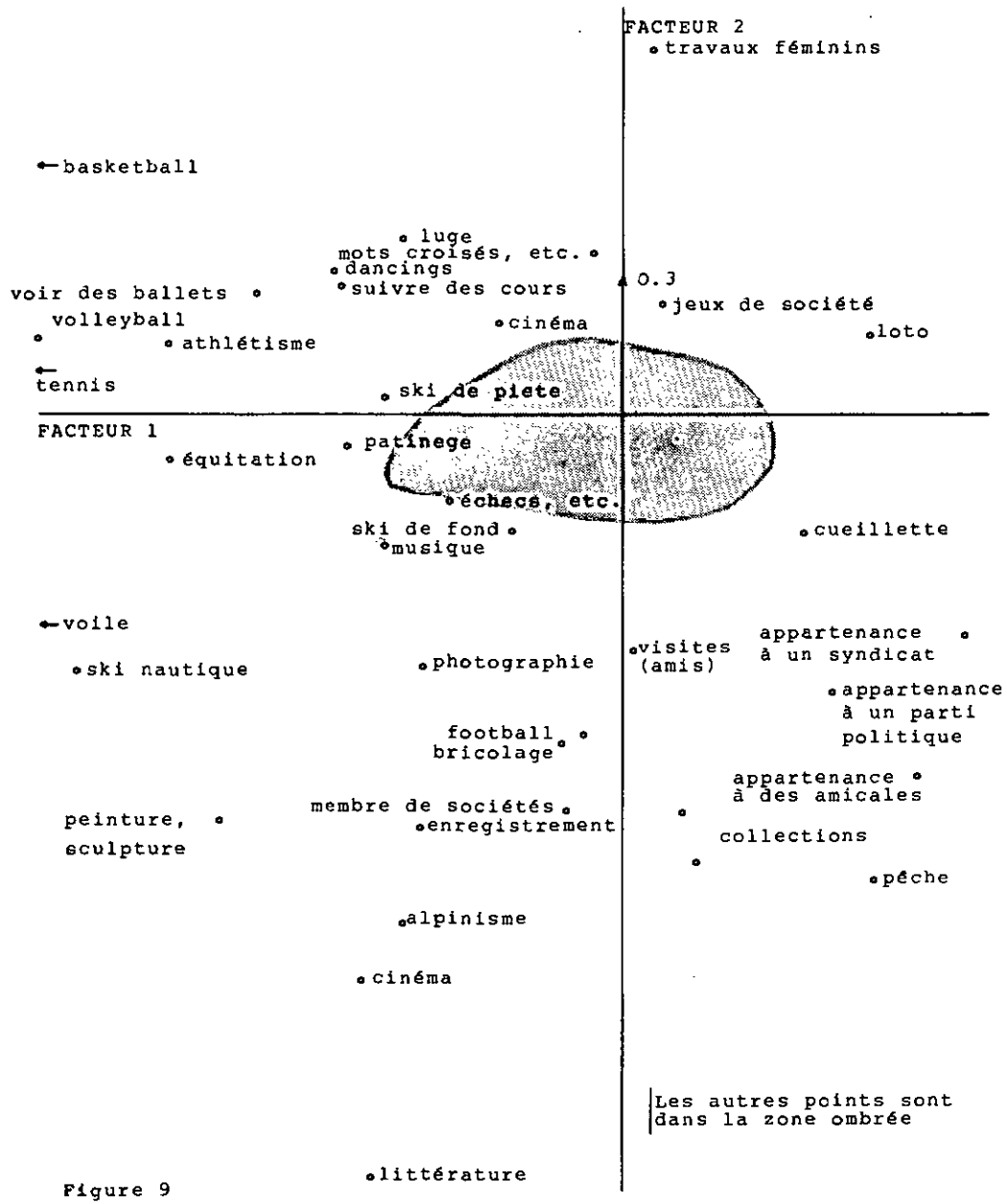


Figure 9

fortement ce facteur sont d'un côté la fréquentation des syndicats, des amicales, des lotos et des fêtes, la cueillette, la télévision et les déplacements de moins d'un jour. De l'autre côté, on trouve le tennis, le ski de piste, les dancings, les spectacles de ballets, les arts plastiques, la natation, etc. On remarque en outre que tous les sports sont du même côté de ce facteur, sauf la marche. Cette exception s'explique probablement par le fait que les enquêtés ont assimilé marche et promenade. Le deuxième axe oppose les activités plutôt féminines des autres. Exemple : travaux féminins, mots croisés, scrabble, etc. contre amicales, bricolage, enregistrement, etc. Les différentes modalités des caractéristiques "sexe", "âge" et "état civil" projetées sur le premier plan factoriel (figure 10), montrent sur le premier facteur une opposition entre les jeunes, les célibataires et les revenus élevés d'un côté, et de l'autre les gens âgés, les divorcés, séparés et veufs, ainsi que les revenus faibles.

Au lieu de parler de loisirs "élitistes" contre loisirs "populaires", on pourrait parler de loisirs "actifs" et "non-actifs" (ou "moins actifs"). Rappelons que ces points sont des observations supplémentaires qui se projettent au centre de gravité des observations qui ont la caractéristique représentée. Ces projections donnent un début de réponse à la question "qui fait quoi ?". On peut faire un lien entre les deux premiers facteurs de cette analyse et les trois premiers de l'analyse sur les caractéristiques personnelles. En effet, le premier facteur de cette analyse semble être un mélange des deux premiers de l'analyse sur les caractéristiques personnelles (gens jeunes et aisés, opposés aux gens âgés et modestes), tandis que le deuxième ressemble au troisième de l'analyse précédente (opposition hommes - femmes).

Sur le troisième axe, (figure 11) on trouve des loisirs qu'on pourrait qualifier de "culturels-utiles" d'un côté, et des

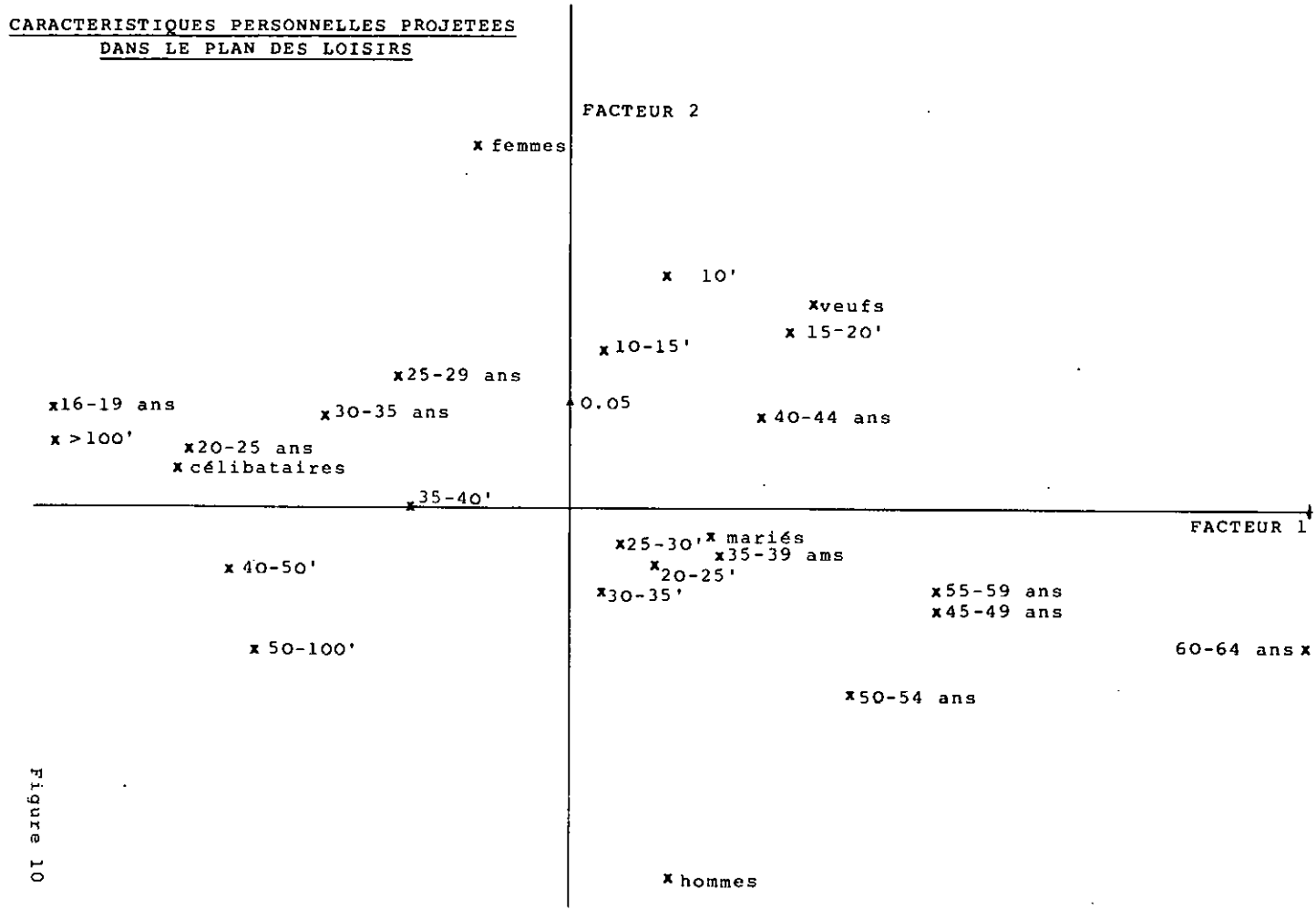

CARACTERISTIQUES PERSONNELLES PROJETEES
DANS LE PLAN DES LOISIRS
FACTEUR 2
FACTEUR 1
femmes
10'
veufs
15-20'
10-15'
25-29 ans
16-19 ans
30-35 ans
0.05
40-44 ans
>100'
20-25 ans
célibataires
35-40'
25-30'
mariés
35-39 ams
20-25'
30-35'
40-50'
55-59 ans
45-49 ans
50-100'
60-64 ans
50-54 ans
hommes

Figure 10

loisirs de pure distraction de l'autre : participation à des sociétés savantes, à des partis politiques et des associations professionnelles, à des conférences-débats et des représentations théâtrales, contre les salles de jeux, les lotos, les jeux de société, le football, le basketball, les collections d'objets divers et la pêche. Le quatrième facteur a aussi une interprétation : il oppose les activités d'équipe à des activités individuelles (même si elles sont pratiquées en groupe, comme le loto). Les variables les plus significatives sur ce facteur sont peinture-sculpture, travaux féminins, littérature, collections, enregistrement, bricolages et loto, opposées aux variables associations professionnelles, football, amicales, syndicat.

PLAN DES LOISIRS

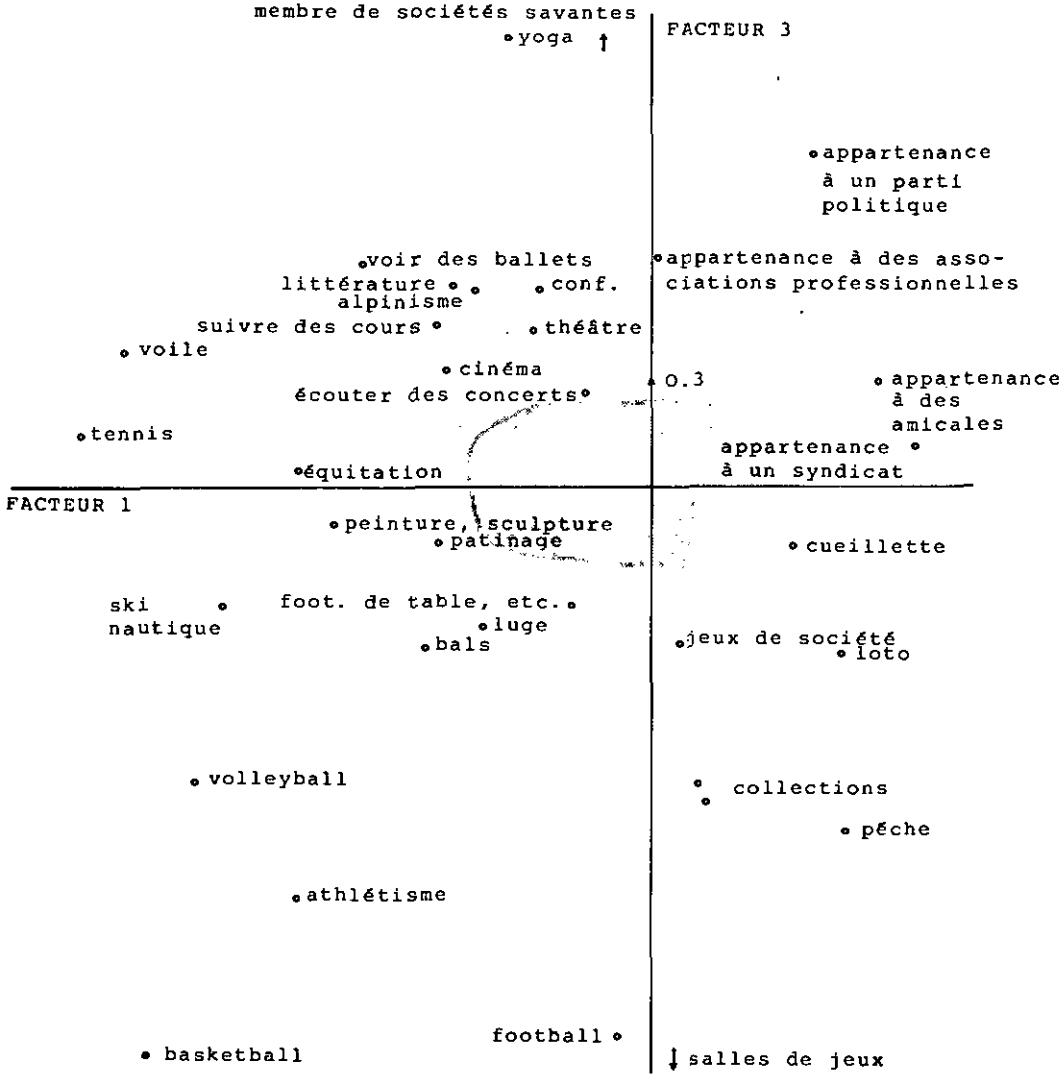


Figure 11

V. CLASSIFICATION AUTOMATIQUE

Le but des méthodes de classification automatique est de créer des classes (des groupes) d'objets (personnes, activités, entreprises ...) qui se ressemblent. Comme pour l'analyse factorielle, on considère que les objets sont des points dans un espace à n dimensions. Deux objets se ressemblent d'autant plus que les points qui les représentent sont proches l'un de l'autre. Les objets peuvent être les lignes ou les colonnes d'un tableau, ou, dans les termes utilisés jusqu'ici, des observations ou des variables. Les composantes dans les différentes dimensions des objets à classer sont appelées caractéristiques. Il y a deux types principaux de méthodes : les méthodes de classification hiérarchique et les méthodes de partitionnement. Nous ne décrirons ici qu'un algorithme de chaque type.

1. La classification hiérarchique

A) Le principe de la classification hiérarchique ascendante

Au départ de l'algorithme, on a autant de classes que d'objets, chaque classe contenant exactement un objet.

1) On calcule toutes les distances entre les objets.
2) On agrège en une seule classe les deux classes séparées par la plus petite distance, c'est-à-dire les deux classes qui

	A1	A2
1	4	7
2	9	4
3	2	1
4	9	1
5	5	8
6	4	10
7	8	4
8	1	2
9	1	3
10	2	3
11	10	2
12	8	1
13	9	2
14	6	10
15	5	9

Tableau 6

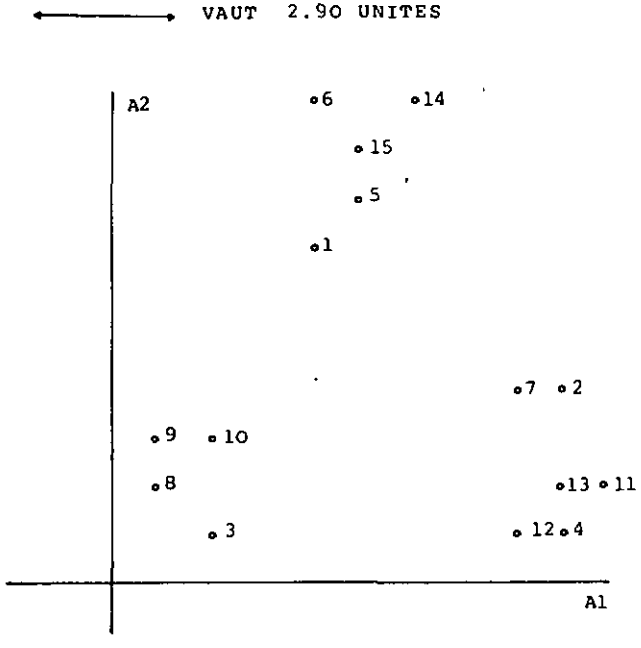


Figure 12

ITERATION	DISTANCE	CLASSES AGREGEES	
1	1.0	2	7
2	1.0	8	9
3	1.0	4	12
4	1.0	11	13
5	1.0	5	15
6	1.5	8	10
7	2.5	4	11
8	3.5	1	5
9	3.7	3	8
10	4.0	6	14
11	6.0	1	6
12	7.5	2	4
13	56.7	1	3
14	59.7	1	2

Tableau 7

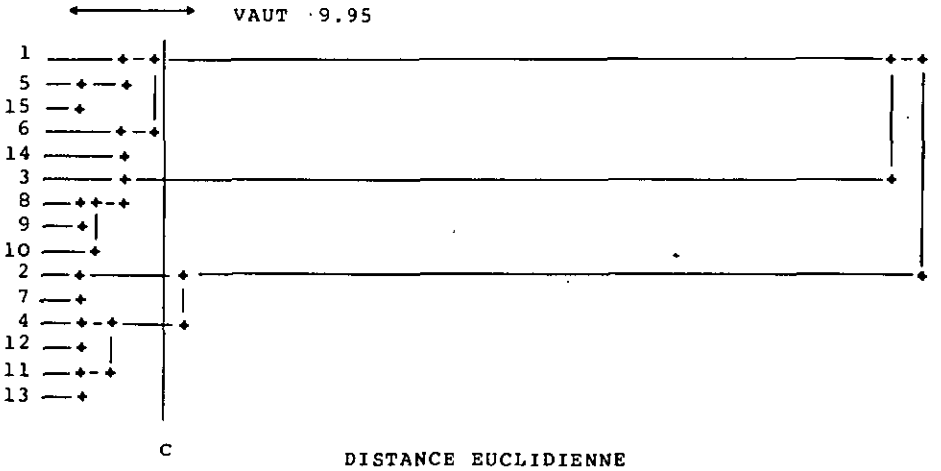


Figure 13

se ressemblent le plus. On diminue ainsi le nombre de classes de un.

3) On calcule les distances entre cette nouvelle classe et les autres. (Les distances des autres classes entre elles n'ont pas changé).
4) On recommence au point 2), tant qu'il reste plus d'une classe.

A la fin de l'algorithme, nous avons une liste qui contient pour chaque étape les noms des deux classes agrégées, ainsi que la distance qui les séparait avant l'agrégation. Ceci permet de dessiner un arbre hiérarchique qui nous sert à visualiser les classes formées par le calcul.

Illustrons ceci par un petit exemple. Soient 15 points dans un espace à deux dimensions (tableau 6). Le dessin de ces points nous révèle trois groupes (figure 12). Les calculs donnent les résultats reproduits dans le tableau 7, le nom d'une classe étant le numéro le plus petit parmi les numéros des points qui appartiennent à la classe. On peut constater sur le dessin de l'arbre hiérarchique (figure 13) l'existence de trois classes bien distinctes, par le fait qu'il y a trois branches qui s'agrègent alors qu'elles sont séparées par des distances élevées, tandis que les agrégations sur chacune de ces branches se font à des niveaux de beaucoup inférieur. On peut couper l'arbre de façon à avoir un nombre quelconque de classes. Par exemple, si on désirait partager les points en 4 classes, il faudrait couper l'arbre comme l'indique le trait C. Les classes ainsi délimitées sont formées respectivement des points (1, 5, 15, 6 et 14), (3, 8, 9 et 10), (2 et 7), et (4, 11, 12 et 13).

Deux notions doivent être explicitées à propos de cette méthode : celle de distance entre objets et celle de critère d'agrégation.

B)Distance entre points

Il existe un grand nombre de mesures possibles de la distance entre deux points. Nous n'en citerons que 4, souvent utilisées et faciles à interpréter.

1. Le carré de la distance euclidienne classique

C'est la distance utilisée dans l'exemple ci-dessus. On retrouve par le calcul les distances, et donc aussi les classes, perceptibles à l'oeil sur le dessin. Pour que deux objets soient à une distance nulle, pour qu'ils soient identiques, il faut qu'ils aient les mêmes composantes.

2. La distance du khi-carré

Rappelons simplement que c'est une distance pondérée entre profils. Elle est donc nulle entre deux objets qui ont le même profil (voir ch. IV, paragraphe 1, analyse des correspondances). Appliquée à l'exemple précédent, elle fait apparaître deux classes très distinctes (figure 14). Ces deux classes contiennent respectivement les points au-dessous et au-dessus de la bissectrice des axes. On peut faire apparaître sur le dessin (figure 15) ces deux groupes en dessinant les projections des profils sur une droite parallèle à la droite d'équation A1 + A2 = 1, où se trouveraient tous les objets si on prenait comme coordonnées les éléments de leurs profils (les pourcentages en ligne).

3. Indices de ressemblance

Si les variables sont binaires, c'est-à-dire si les coordonnées des objets à classer sont des 0 et des 1 uniquement, on peut calculer des indices de ressemblance, à partir desquels on obtient des indices de distances (ou même des distances proprement dites). Désignons par O le nombre de fois que deux objets ont simultanément des composantes égales à 1 (oui), par N le

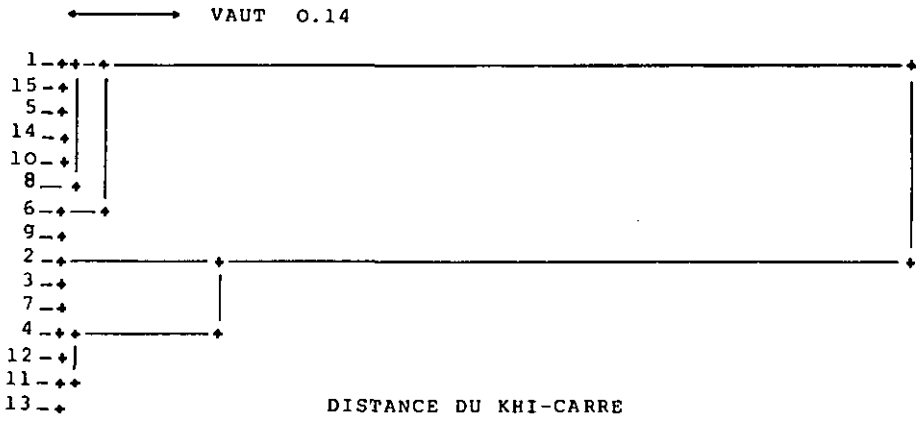


Figure 14

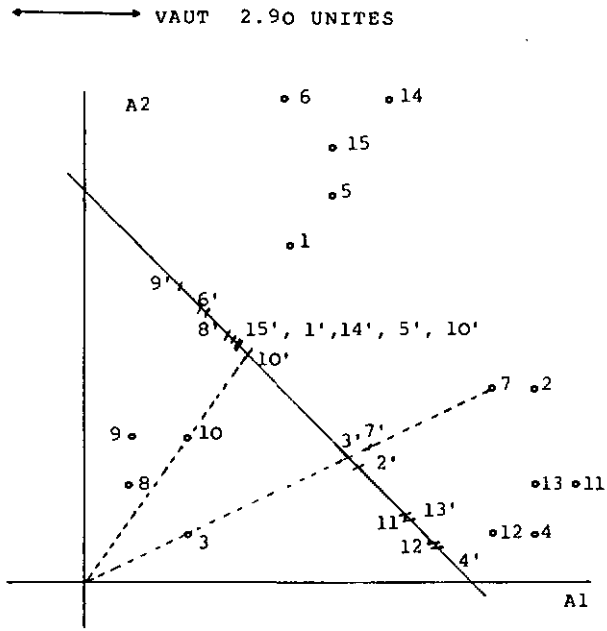


Figure 15

nombre de fois que les deux objets ont en même temps des composantes nulles (non) et par p la dimension de l'espace dans lequel sont ces points. Les indices de ressemblance sont calculés en fonction de O, N et p, et éventuellement en fonction de D = p - (N + O), le nombre de fois que les composantes des deux objets sont différentes.

3.1 Russel et Rao

L'indice de ressemblance est défini par O/p, grandeur qui varie de 0 à 1, comme l'indice de distance, défini par 1 - (O/p). On voit que seuls les 1 communs comptent. Deux objets qui ont les mêmes composantes, mais avec un nombre de composantes non nul faible, vont donc être considérés comme assez éloignés. Au contraire, deux objets dont le nombre de 1 communs est assez grand, mais dont les autres composantes sont systématiquement opposées seront proches.

3.2 Sokal et Michener

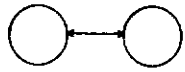
L'indice de ressemblance est défini par (O + N)/p. Il varie de 0 à 1, comme la distance définie par 1 - (O + N)/p. Les 0 et les 1 jouent des rôles symétriques. Cette distance est équivalente, pour des variables binaires, à la distance euclidienne.

C) Critères d'agrégation

Il s'agit de savoir de quelle façon mesurer (quelle que soit la distance entre objets choisie) la distance entre deux classes, dont une au moins contient plus d'un point. Il y a plusieurs possibilités. Les trois décrites ci-dessous peuvent être appliquées dans tous les cas.

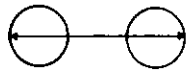
1. Chaînage simple

La distance entre deux classes est définie par la distance la plus petite qu'on puisse trouver entre un objet appartenant à une des classes et un objet de la seconde.

Schématiquement : 

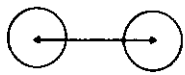
2. Chaînage complet

C'est la distance la plus grande entre deux points appartenant à chacune des deux classes qui définit la distance entre ces classes.

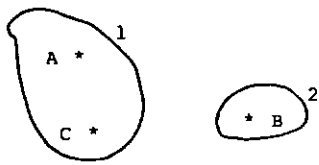
Schématiquement : 

3. Distance moyenne

La distance entre deux classes est définie par la distance moyenne des objets de la première classe aux objets de la seconde.

Schématiquement : 

Illustrons ceci par un petit exemple :



On a la classe 1, formée des objets A et C, et la classe 2, qui ne contient que le point B.

D(1,2), la distance entre les classes 1 et 2, vaut respectivement

D(C,B) si on choisit le chaînage simple,
D(A,B) dans le cas du chaînage complet, ou
0.5 * (D(A,B) + D(A,C)).

D) Choix de la distance et du critère d'agrégation

Une des difficultés de l'utilisation de la classification hiérarchique ascendante provient de la possibilité, donc de l'obligation, de choisir une distance et un critère d'agrégation. Or, de ce choix vont dépendre les résultats.

Il est difficile de formuler des critères généraux de choix d'une distance. On peut toutefois faire quelques remarques à ce propos.

- Le choix entre la distance euclidienne et la distance du khi-carré peut se faire selon les mêmes critères qui conduisent à faire une analyse factorielle en composantes principales plutôt que de faire une analyse des correspondances. Il s'agit essentiellement de savoir si on s'attache aux niveaux des caractéristiques des objets à classer, ou seulement aux profils de ces objets.

- La distance du khi-carré ne s'applique qu'à des données positives. On ne peut donc pas, en particulier, utiliser cette distance pour des données centrées, ni pour classer des objets dont les caractéristiques sont leurs projections sur

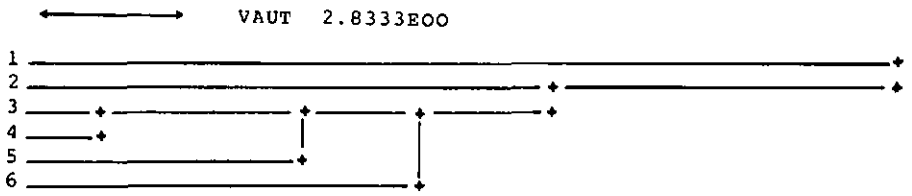
des axes factoriels. Ce cas se prête particulièrement bien à la distance euclidienne, puisque, si on fait une classification sur de tels objets, on espère trouver par le calcul des classes conformes à l'image des points perçue sur le dessin des différents plans factoriels.

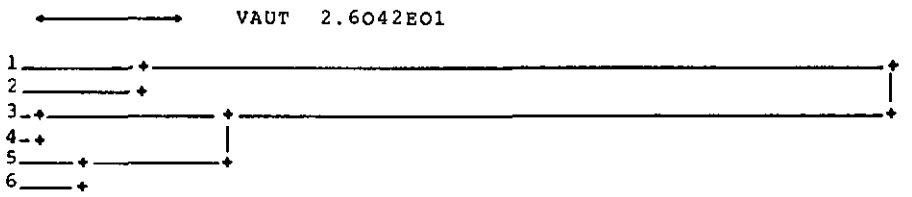
- Dans le cas de variables binaires, il faut se demander si on veut que la proximité entre deux objets n'augmente qu'avec le nombre de 1 communs, ou aussi en fonction du nombre de 0 communs. Dans le premier cas, il faudrait utiliser l'indice de distance de Russel et Rao, dans le second, la distance de Sokal et Michener.

En ce qui concerne le critère d'aqrégation, les trois critères présentés ci-dessus sont utilisables, comme nous l'avons dit, quelle que soit la distance choisie. La différence essentielle entre ces critères provient du fait suivant :

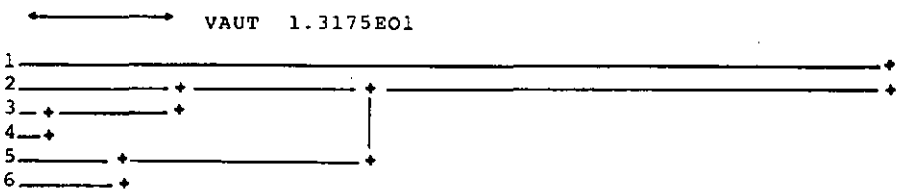
Si le critère de "distance moyenne" est neutre, le chaînage simple a tendance à contracter l'espace, tandis que le chaînage complet a tendance à le dilater. En effet, le chaînage simple va souvent ajouter à des classes existantes des points ou des classes proches, plutôt que de former de nouveaux groupements, comme le chaînage complet le ferait. On peut donc dire que des classes bien séparées exhibées par le chaînage simple ont une existence sûre, puisque la méthode tend à fournir des résultats peu nets. En revanche, le chaînage complet tend à créer des classes bien séparées, mais d'une manière quelque peu artificielle. Pourquoi alors ne pas toujours choisir le critére de "distance moyenne" ? Parce qu'il peut étre utile, pour chercher à interpréter des résultats peu clairs obtenus en utilisant ce critère d'augmenter le contraste de l'arbre hiérarchique en utilisant le chaînage complet (en étant conscient de l'effet séparateur de cette méthode). D'un

	V1	V2
1	.0	.0
2	1.0	4.0
3	2.0	7.0
4	2.0	8.0
5	1.0	10.0
6	.0	12.5





CHAINAGE COMPLET



DISTANCE MOYENNE

Figure 16

autre côté, il peut être intéressant de chercher la confirmation de l'existence de certaines classes en utilisant le chaînage simple.

L'effet de ces trois critères d'agrégation est illustré par le petit exemple de la figure 16 :

Quelle que soit la stratégie, la première itération réunit les points 3 et 4. Dans le cas du chaînage simple, les points 5, 6, 2 et 1 viennent successivement s'ajouter à cette première classe, tandis qu'avec le chaînage complet, les trois premières itérations forment trois classes contenant chacune deux des six points. On ne manquera pas d'observer les changements d'échelles entre les trois dessins. Dans le cas du chaînage simple, l'échelle est 10 fois plus grande que dans celui du chaînage complet, donc la dernière agrégation se fait à une distance 10 fois plus petite. L'espace est donc plus petit ! Il faut encore noter que dans le cas du premier exemple, le critère utilisé est la "distance moyenne", mais que les autres critères appliqués à ces données ne changent rien au résultat, au moins en ce qu'il a de plus important, soit de montrer l'existence de trois classes bien distinctes.

2. Les méthodes de partitionnement

L'idée du partitionnement est la suivante : étant donnés des objets à classer en n classes, trouver la n-partition (c'est-à-dire la répartition des objets dans les n classes) la meilleure possible, soit celle qui fournit les classes les plus homogènes. Cette idée est irréalisable, car il faudrait essayer toutes les partitions possibles, ce qui entraînerait

un nombre considérable de calculs. (Si n n'était pas fixe, la meilleure partition serait celle où on a autant de classes que d'objets, et un objet par classe. L'homogénéité des classes serait alors parfaite). On peut toutefois approcher ce résultat (la meilleure partition) par des méthodes heuristiques comme KMEAN, décrite ci-dessous :

Avant tout calcul, l'utilisateur choisit le nombre de classes qu'il désire former. Au départ de l'algorithme, les objets sont affectés,généralement au hadard, aux différentes classes. A chaque itération, on essaye de déplacer un objet, c'est-à-dire de l'affecter à une autre classe, de façon à améliorer la partition. Si une amélioration est possible, le changement est fait, et une nouvelle itération débute. Sinon, le calcul s'arrête. Le critère d'amélioration est généralement la somme des variances à l'intérieur des classes : la partition est d'autant meilleure que cette somme, dite variance intra-classe, est plus petite. A la fin de l'algorithme, chaque classe contient au moins un objet, si bien que le nombre de classes ne dépend que du choix initial de l'utilisateur. Contrairement à la classification hiérarchique, où le choix du nombre de classes est fait à la fin du calcul (en coupant l'arbre), ce choix est donc fait ici à priori. Mais il est possible, et même recommandé, de faire plusieurs classifications, en changeant le nombre de classes.

Remarque : on ne sait pas comparer mathématiquement une n-partition et une (n+k)-partition. Le nombre de classes à retenir ne dépend donc que du jugement de l'utilisateur. Les résultats fournis par les programmes de partitionnement se présentent sous la forme de listes des points affectés à la fin de l'algorithme à chacune des classes, plus éventuellement les coordonnées des centres de gravité des classes. Appliquée à l'exemple des 15 points dans un plan (tableau 6, figure 12), en choisissant 3 classes, cette méthode fait apparaître les

mêmes groupements que la classification hiérarchique ascendante. Il en est de même si on choisit 4 classes.

Notons encore que les méthodes de partitionnement demandent beaucoup moins de mémoire centrale que la classificaction hiérarchique ascendante.

3. Exemples d'utilisation de la classificaction automatique

A) Classification des loisirs après analyse factorielle (Partitionnement)

Comme on peut le voir sur les dessins des figures 9 et 11 du chapitre précédent, l'analyse factorielle dans laquelle les variables sont les pratiques d'activités de loisirs ne donne pas des résultats très nets. D'autre part,la typologie des activités suivant les différents axes n'est pas très satisfaisante en ce sens qu'elle ne permet pas facilement de décider de quel type font partie les activités. Si on tient encore compte du fait que les pourcentages d'inertie représentés par les cinq premiers facteurs ne sont pas très différents, il semble opportun de chercher à grouper les activités en classes. Pour cela, on a considéré que les objets à classer sont les variables de l'analyse factorielle, soit les activités de loisirs, dont les caractéristiques sont leurs projections sur les cinq premiers facteurs. Un partitionnement en huit classes a fourni les résultats reproduits dans le tableau 8. Ces classes semblent, intuitivement, assez homogènes. Prenons les classes dans l'ordre décroissant de leurs effectifs :

- La classe 4, qui regroupe 25 activités, peut être qualifiée de classe des loisirs "populaires", ou "standards". On y trouve toutes les activités pratiquées par presque tout le monde.

- La classe 2 regroupe des loisirs eux aussi communs, mais la lecture de la liste suggère une idée de jeunesse, ce qui n'est guère le cas pour la classe 4.

- La classe 8 semble assez nettement regrouper des loisirs qu'on pourrait qualifier d'"intellectuels".

- La classe 5 regroupe des activités créatives (sauf l'alpinisme ; probablement que dans l'échantillon, les personnes qui pratiquent ce sport ont par ailleurs ce type d'activités).

- La classe 6 est une classe de "sports d'élite" (ou "de l'élite" ?).

- Les activités de la classe 3 suggèrent l'idée de loisirs pratiqués en solitaire.

- Les sports groupés dans la classe 1 demandent un engagement physique important. Ils constituent des spectacles de masse.

- La classe 7 suggère pour ceux qui pratiquent ces activités une certaine érudition, et peut-être aussi un certain âge.

Il restera à déterminer qui sont les personnes qui pratiquent les activités de ces différentes classes. Nous verrons plus loin comment il est possible d'y parvenir.

CLASSIFICATION DES LOISIRS APRES ANALYSE FACTORIELLE

8 CLASSES

CLASSES	PRATIQUANTS	DENOMINATION DES OBJETS
1	22	Cyclisme
	12	Athlétisme
	27	Football
2	308	Plage - Piscine
	243	Cinéma (voir)
	88	Dancings
	148	Echecs, Dames, Chars
	184	Foot. de table, Billard, Minigolf, Quilles, ...
	115	Gymnastique
	214	Natation
	173	Ski de piste
	135	Ski de fond
	13	Patinage
	86	Luge
3	58	Collections d'objets divers
	25	Collections d'objets de la nature
	36	Pêche
	24	Salles de jeux
4	438	Promenade
	389	Pique-nique
	403	Radio
	403	Television
	426	Journaux et revues
	317	Livres
	201	Cueillette
	232	Variétés
	213	Spectacles sportifs
	117	Bals
	247	Cartes
	130	Jeux de société
	121	Lotos
	215	Mots croisés, Scrabble, Puzzles
	323	Déplacements de moins de un jour
	309	Déplacements de plus de un jour
	423	Vacances hors du domicile
	375	Visites (amis, voisins, famille ...)

Tableau 8

	278	Bistrots
	101	Amicales
	160	Syndicats
	339	Visites de musées et expositions
	173	Travaux féminins
	252	Marche
	380	Fêtes publiques
5	42	Peinture - Sculpture
	27	Cinéma (faire)
	104	Photo
	62	Enregistrement
	12	Littérature (faire)
	105	Bricolage
	32	Alpinisme
6	13	Ski nautique
	12	Voile
	22	Tennis
	13	Basketball
	17	Volleyball
7	34	Partis politiques
	10	Sociétés savantes
8	12	Yoga
	181	Théâtre
	72	Associations professionnelles
	51	Suivre des cours
	157	Conférences et débats
	52	Spectacles de ballets
	204	Concerts
	88	Musique (faire)
	27	Equitation

Tableau 8 (suite)

B) Classification d'exploitations agricoles *
(Classification hiérarchique ascendante)

Les objets à classer sont 291 exploitations agricoles situées dans le Val-de-Travers. Les renseignements proviennent du recensement fédéral des entreprises agricoles du 30 juin 1975.

De nombreuses études ont été faites, qui tentaient d'établir une discrimination des entreprises agricoles sur la base de critères envisagés séparément. Les caractéristiques retenues ici sont celles qui paraissent les plus pertinentes pour la région étudiée et le type d'agriculture qui lui est spécifique. La méthode utilisée permet de tenir compte de toutes ces caractéristiques en même temps. Les treize caractéristiques retenues sont les suivantes :

1) La zone d'implantation
   Le Val-de-Travers est divisé en deux zones dont la délimitation dépend principalement de l'altitude. La zone 1 est plus favorable que la zone 2.
2) L'activité principale du chef d'exploitation
   On distingue si l'agriculture est son activité principale ou non.
3) La main-d'oeuvre
   Il s'agit du nombre de personnes employées à titre principal ou accessoire dans l'exploitation.
4) Utilisation en commun de machines agricoles (oui - non).
5) Formation de l'exploitant
   Aucune titre, apprentissage, école spécialisée, examen de capacité, maîtrise ou technicum, autre.

* *Données utilisées dans une recherche en cours de M. F. Hainard, assistant en sociologie, sur la paysannerie du Val-de-Travers.*

6) Nombre de parcelles de terres cultivées
C'est une indication du morcellement des terres cultivées de l'exploitation.

7) Surface de terre exploitée à ferme ou en usufruit
Les forêts et alpages ne sont pas compris.

8) Surface de terres en propriété
Forêts et alpages non compris.

9) Surface totale de l'exploitation, sans les forêts
C'est la somme des deux grandeurs ci-dessus.

10) Surface de forêts mises en valeur

11) Surface de terres ouvertes
C'est la surface des terres assolées moins les prairies artificielles.

12) Bétail bovin
Nombre de vaches par exploitation.

13) Nombre de machines utilisées
Il s'agit de grosses machines agricoles.

Les quantités ont été saisies après découpage en classes. Toutes les caractéristiques ont été représentées par des variables binaires, les caractéristiques quantitatives de la façon indiquée à la page 23 La surface totale (no 9) n'a pas été utilisée pour construire l'arbre hiérarchique, puisqu'elle est obtenue en additionnant les surfaces 7 et 8. La classification a été faite en utilisant la distance euclidienne et le critère d'agrégation "distance moyenne". L'arbre, que nous ne pouvons pas reproduire ici, a été coupé de façon à retenir 9 classes qui contiennent respectivement 79, 75, 67, 56, 5, 4, 2, 2 et 1 exploitations. Les quatre classes dont l'effectif est le plus faible n'ont évidemment pas été étudiées. On a simplement remarqué, à titre anecdotique, que l'exploitation qui forme une classe à elle seule se démarque véritablement par rapport aux autres exploitations du Val-de-Travers. La classe qui contient cinq exploitations a été retenue pour

la suite de l'étude, car on s'est aperçu qu'elle est très homogène et très particulière. Nous verrons plus loin quelles sont les caractéristiques des classes prises en considération (qui groupent 97 % des exploitations), et comment ces caractéristiques ont été décelées.

VI. EXPLOITATION DES RESULTATS FOURNIS PAR L'ANALYSE FACTORIELLE ET LA CLASSIFICATION AUTOMATIQUE

Comme il a été dit plus haut, l'analyse factorielle conduit souvent à formuler des hypothèses plutôt qu'à énoncer des conclusions. Par ailleurs, la classification automatique en soi apporte rarement des résultats directement exploitables. En effet, si on a fait une classification des questionnaires, on connaît des groupes, mais on ne sait pas ce qui différencie ces groupes de la population de référence. Si on a constitué des classes de variables, il est probable qu'on souhaite établir des liaisons entre les variables des différentes classes et d'autres variables (en particulier les variables socio-administratives, ou caractéristiques personnelles), ou avec les sous-populations d'enquêtés qu'on peut rapprocher de ces classes. En tous les cas, le problème consiste, à partir des résultats obtenus grâce aux méthodes multidimensionnelles, à revenir aux données brutes.

1. Les tris

Le retour aux données sera fait la plupart du temps en réexaminant des tableaux de tris à la lumière des résultats obtenus jusque là, ou en éditant de nouveaux tableaux de tris.

A) Tris portant sur des variables existantes

L'édition de nouveaux tableaux de tris portant sur des variables qui existaient déjà au moment de la saisie des données peut être utile pour différentes raisons. Par exemple :

- Deux variables sont proches sur les différents plans factoriels. On sait donc qu'elles ont des profils "semblables", qu'elles sont "souvent" pratiquées par les mêmes personnes. On peut alors chercher à quantifier ces relations en croisant les deux variables en question.

  Par exemple : activités I et J

act. J / act. I	pratiquée	non pratiquée	total
pratiquée	75	25	100
non pratiquée	75	325	400
total	150	350	500

  On voit que 75 % de ceux qui pratiquent I pratiquent aussi J; que 50 % de ceux qui pratiquent J pratiquent aussi I; que ces activités sont pratiquées simultanément par 21.4 % de la population ...
- Quelques variables ont un comportement particulier (leurs projections sont très excentrées; elles forment une classe intéressante). On peut chercher de nouvelles informations en croisant ces variables entre elles, ou avec d'autres variables (en particulier les caractéristiques personnelles).
- En bref, chaque fois qu'une méthode multidimensionnelle suggère des relations entre variables, il est conseillé de

chercher à quantifier ces relations par de nouveaux tris.

B) Tris portant sur de nouvelles variables

Très souvent, compte tenu des résultats obtenus par les méthodes d'analyse multidimensionnelles et les premiers tris, on a suffisamment d'informations pour souhaiter travailler avec de nouvelles variables qui représentent chacune une combinaison de plusieurs caractères. Ces nouvelles variables peuvent être construites "manuellement" ou être le produit de l'application des méthodes multidimensionnelles. La construction manuelle de nouvelles variables peut se faire de n'importe quelle façon jugée pertinente. Par exemple, dans la recherche chaux-de-fonnière, on a soumis aux enquêtés une liste d'appareils électro-ménagers pour savoir lesquels ils possèdent. On peut créer une variable "nombre d'appareils détenus" par addition. Autre exemple : on veut disposer d'une variable dont les modalités soient "plus de 30 ans", "moins de 30 ans et mariés, divorcés, veufs ou séparés", "moins de 30 ans et célibataires", "indéterminé". On met alors en jeu les caractéristiques "âge" et "état civil", avec la règle de formation suivante :

Age	Etat civil	Nouvelle variable
indéterminé	-	0
-	indéterminé	0
< 30	célibataire	1
< 30	marié, divorcé ...	2
> 30	-	3

Il faut être attentif au fait que ce genre de manipulations réclame une programmation ad hoc, ce qui entraîne des coûts et des délais. Les nouvelles variables les plus intéressantes sont celles fournies par les méthodes multidimensionnelles. Rappelons que les axes factoriels sont des combinaisons des variables qui participent à l'analyse. Les composantes des observations sur les facteurs peuvent donc être considérées comme des scores obtenus par les observations sur les différents axes, qui représentent tous une synthèse, selon différents points de vue, de l'information contenue dans les variables utilisées. Les composantes peuvent donc être considérées comme de nouvelles variables, que ce soit pour faire des tris ou de la classification automatique. Selon l'utilisation, on se servira des valeurs sans les modifier, ou on formera des classes, comme pour n'importe quelle variable quantitative. La classification automatique dont les objets sont les observations associe à chacune de celles-ci le numéro de la classe à laquelle elle est affectée. Ce numéro de classe (l'affectation à une classe) dépend des valeurs de toutes les caractéristiques pour l'observation. Il constitue donc une excellente synthèse des variables utilisées pour la classification.

C) Exemple : suite de l'étude des exploitations agricoles

Comme nous l'avons vu plus haut, la classification hiérarchique ascendante nous a permis de mettre en évidence 5 classes d'exploitations agricoles. Il fallait alors chercher à qualifier ces classes. Ce travail a été fait de la manière suivante : à chaque exploitation a été associé son numéro de classe, considéré comme une nouvelle variable. Ce numéro a été ensuite utilisé comme filtre, ce qui a permis de faire des tris à plat de toutes les variables dans les 5 groupes d'exploitations, chaque tri étant comparé au tri correspondant

pour l'ensemble des exploitations. Nous avons ainsi obtenu une série de tableaux (13 * 5), dont le premier examen est facilité par le calcul du khi-carré : les tableaux pour lesquels cette valeur est petite sont ignorés. Un extrait de ces tris pour une classe est donné dans les tableaux 9, 10 et 11. Les 67 exploitations de cette classe représentent le 23 % du total des exploitations.

Commentaire des tris reproduits :

1) Les tris des variables indiquant les surfaces de terres en propriété et à ferme ont été reproduits parce qu'ils sont les plus frappants pour cette classe. Ils montrent que les exploitants de cette classe ne possèdent pratiquement pas de terres et que la surface des terres en location est grande (plus de 20 ha dans 94 % des cas ; la surface totale ne dépasse 20 ha que pour 53 % des exploitations étudiées).

2) Les tableaux relatifs à la zone habitée et à l'utilisation en commun de machines agricoles ont été reproduits pour montrer la nécessité de comparer les caractéristiques des classes à celles de l'ensemble des exploitations. En effet, sur le premier de ces tableaux, nous constatons que plus de 76 % des exploitations de la classe étudiée sont situées en zone 2. On pourrait donc en conclure que ces exploitations se caractérisent par le fait qu'une forte majorité d'entre elles sont situées dans cette zone. Cette conclusion est fausse, car cette proportion est de 69 % dans l'ensemble. La différence n'est pas significative au sens du khi-carré, qui vaut 1.56, alors que la table donne 3.86 au seuil de 5 %. Le deuxième de ces tableaux indique que dans la classe étudiée, les exploitants se divisent par moitié entre ceux qui utilisent des machines en commun et les autres. On pourrait donc en conclure ... qu'il n'y a rien à dire à ce propos. En réalité, comme seuls 34 % du total des exploitants utilisent des machines en commun, nous devons constater que cette pratique est significativement

TERRES EN PROPRIETE

	-	<1ha	1-5ha	5-10ha	10-20ha	20-30ha	30-50ha
E	117	16	20	23	59	46	10
	40.2	5.5	6.9	7.9	20.3	15.8	3.4
G	63	1	1	1	1	0	0
	94.0	1.5	1.5	1.5	1.5	0.0	0.0
G/E	53.8	6.2	5.0	4.3	1.7	0.0	0.0

TERRES A FERME

	-	<1ha	1-5ha	5-10ha	10-20ha	20-30ha	30-50ha	50ha
E	78	18	37	31	43	44	30	10
	26.8	6.2	12.7	10.7	14.8	15.1	10.3	3.4
G	0	0	0	0	4	28	27	8
	0.0	0.0	0.0	0.0	6.0	41.8	40.3	11.9
G/E	0.0	0.0	0.0	0.0	9.3	63.6	90.0	80.0

Tableau 9

ZONE HABITEE

	Zone 1	Zone 2
Ensemble	90	201
	30.9	69.1
Groupe	16	51
	23.9	76.1
G / E	17.8	25.4

UTILISATION EN COMMUN DE MACHINES AGRICOLES

	non	oui
E	192	99
	66.0	34.0
G	34	33
	50.7	49.3
G/E	17.7	33.3

Tableau 10

TERRES OUVERTES

	-	<1ha	1-5ha	5-10ha	10-20ha
E	121	69	81	16	4
	41.6	23.7	27.8	5.5	1.4
G	14	15	30	7	1
	20.9	22.4	44.8	10.4	1.5
G/E	11.6	21.7	37.0	43.7	25.0

BETAIL BOVIN

	-	1-10	11-20	21-30	31-50	51 et plus
E	23	104	127	26	9	2
	7.9	35.7	43.6	8.9	3.1	0.7
G	0	7	43	13	4	0
	0.0	10.4	64.2	19.4	6.0	0.0
G/E	0.0	6.7	33.9	50.0	44.4	0.0

Tableau 11

TERRES OUVERTES, SANS LES EXPLOITATIONS QUI N'EN N'ONT PAS.

	moins de 1ha	1-5ha	5-10ha	10-20ha	
Ensemble	69	81	16	4	170
	40.6	47.6	9.4	2.4	100%
Groupe	15	30	7	1	53
	28.3	56.6	13.2	1.9	100%

Le Khi-carre a été calculé en groupant les deux dernières catégories. Il vaut 3.36, alors que la table donne 5.99 pour 2 degrés de liberté au seuil de 5%.

Tableau 12

plus répandue dans ce groupe que dans l'ensemble de la population (le khi-carré vaut 6.93), même si la différence n'est pas spectaculaire.

3) Les renseignements quantitatifs fournis par les deux derniers tableaux, relatifs aux terres ouvertes et au bétail bovin, doivent être nuancés. On peut affirmer que les troupeux sont grands par rapport à l'ensemble : 11 têtes ou plus dans 89 % des exploitations du groupe, alors que cette grandeur n'est atteinte que dans 61.2 % de l'ensemble des exploitations qui ont du bétail bovin. En revanche, il ne serait pas tout à fait correct d'affirmer que les surfaces de terres cultivées sont plus grandes dans cette classe que dans l'ensemble. Il est vrai que la proportion d'exploitations qui ont des terres ouvertes y est plus élevée, mais si on compare les surfaces cultivées après avoir éliminé les exploitations qui ne font pas de culture, on voit que la différence n'est pas significative au sens du khi-carré (tableau 12).

Une première lecture des tableaux relatifs aux différentes classes permet une description sommaire de celles-ci. Cette description est donnée ci-dessous par un qualificatif, insuffisant par sa brièveté (une étiquette), et par les caractéristiques les plus marquantes.

Classe 1. 67 exploitations (23 %)

"Fermiers prospères"
- Main-d'oeuvre "importante"
- Utilisation de machines en commun assez fréquente
- Proportion d'exploitants diplômés un peu supérieure à la moyenne
- Pas de terres en propriété
- Domaines assez grands
- Pas d'exploitation de forêts

- Terres ouvertes dans une grande proportion des exploitations
- Bétail bovin assez important
- Bien équipés

Les tableaux donnés en exemple sont relatifs à cette classe.

Classe 2. 56 exploitations (19 %)

"Petite paysannerie à plein temps"
- Morcellement important
- Domaines de grandeur moyenne (10 à 30 ha)
- Les exploitants sont soit propriétaires, soit fermiers
- Peu de terres ouvertes
- Tous possèdent du bétail bovin, mais en petites quantités

Classe 3. 75 exploitations (26 %)

"Mini-exploitations"
- Plutôt en zone 1
- Regroupe quasi toutes les exploitations dont le chef est agriculteur à titre accessoire
- Main-d'oeuvre réduite
- Pas d'utilisation de machines en commun
- Mal équipées
- Domaines les plus petits (97 % des domaines de moins de 10 ha)
- Presque pas de vaches (91 % de ceux qui n'en ont pas sont dans cette classe; ceux qui en possèdent en ont moins de 10)

Classe 4. 79 exploitations (27 %)

"Paysannerie conforme à l'image traditionnelle"
- Plutôt en zone 2
- Main-d'oeuvre "importante"
- Utilisation plus fréquente qu'en moyenne de machines en commun
- Terrains en propriété, assez grands
- Pas ou peu de terres à ferme

- Exploitation de forêts fréquente (dans 80 % des cas)
- Terres ouvertes de petites dimensions
- Bétail bovin d'importance moyenne (68 % entre 11 et 20 vaches)
- Assez bien équipées

Classe 5. 5 exploitations (1.7 %)

"Exploitations idéales, économiquement parlant"

- Situées en zone 1
- Les domaines les plus grands
- Grandes surfaces en propriété, auxquelles s'ajoutent des terres à ferme
- Surface importante de terres ouvertes
- Bétail bovin important (pas moins de 20 bêtes)
- Très bien équipées

2. Association d'observations à des classes de variables

Une partition d'un ensemble de variables étant donnée, un bon moyen d'étudier ces classes de variables est d'étudier les sous-populations qui s'y rattachent. Pour terminer ce chapitre, nous allons donc voir un moyen de former des classes d'observations non disjointes par association à des classes de variables. Nous avons vu comment former des classes à l'aide des méthodes multidimensionnelles en fonction des valeurs des variables dans les observations. Les classes de variables que nous traitons ici peuvent être formées de n'importe quelle manière jugée pertinente. Rappelons qu'il n'y a pas de classifications vraies, il n'y a que des classifications correctes compte tenu

du critère choisi. Il est tout à fait possible qu'on souhaite traiter des variables en les groupant selon des critères à priori, parfaitement valables, quelles que soient les distributions de ces variables dans l'échantillon. Considérons par exemple des notes obtenues dans différentes branches : on peut très bien vouloir faire des catégories telles que "branches relatives à la langue française", "langues étrangères", "sciences exactes", "sciences expérimentales", et cela quels que soient les résultats obtenus par les élèves dans les branches qui forment ces catégories. Ou, pour reprendre l'exemple des activités de loisirs que nous avons vu plus haut, nous pourrions souhaiter diviser les loisirs en "activités de plein-air", "activités d'intérieur", "autres", ou encore en "activités subventionnées", "autres", ou de n'importe quelle façon jugée intéressante, telle que la typologie élaborée par M. Erard pour cette enquête.

A) Critères d'association

On peut associer les observations aux classes de variables de n'importe quelle manière jugée significative. Par exemple, dans le cas des notes, on pourrait associer un élève à une catégorie de branches si la somme des notes obtenues par l'élève dans les branches de cette catégorie dépasse une valeur fixée, ou bien si une de ses deux meilleures notes a été obtenue dans une des branches de cette catégorie. Dans le cas des activités, un individu pourrait être associé à un groupe de loisirs dès qu'il pratique un nombre donné de loisirs de ce type. Quel que soit le critère choisi, le principe reste toujours le même : on associe une observation à une classe de variables si cette observation obtient un certain score dans les variables qui forment la classe. On peut remarquer que le même score ne peut généralement pas s'appliquer à toutes les classes, d'abord par le fait que les différentes classes

ne contiennent pas forcément le même nombre de variables, et ensuite parce que les sommes des variables peuvent être très différentes. Ainsi, si on reprend l'exemple des classes de loisirs (tableau 8), on voit bien que la signification d'un critère tel que "deux activités au moins de la classe doivent être pratiquées" n'est pas la même s'il s'applique à la classe 4 ou à la classe 7 : dans le premier cas, on peut s'attendre à ce que toutes les observations soient associées à cette classe (le critère n'est pas sélectif), tandis que dans le second cas, il serait trop restrictif. Le score doit donc varier selon les classes de variables. Remarquons encore que le même score peut avoir des significations diverses selon les observations : une personne qui pratique, par exemple, deux activités d'une classe, alors qu'elle déclare pratiquer 50 activités n'a probablement pas la même assiduité dans la pratique de ces deux activités qu'une autre qui en pratique dix en tout. Le score devrait donc aussi varier selon les observations. La technique que nous proposons pour décider de l'affectation d'une observation à chacune des classes de variables est la suivante :

Soit	ST	la somme de toutes les variables
	S(i)	la somme des variables de la classe i
alors	F(i)=S(i)/ST	est la fréquence relative des variables de la classe i. Cette valeur dépend à la fois du nombre de variables dans la classe, et de leurs effectifs. Elle s'interprète comme l'importance relative des variables de la classe i.

D'autre part, soit		la somme de toutes les variables pour
	st(j)	l'observation j
		la somme des variables de la classe i
	s(i,j)	pour l'observation j

alors f(i,j)=s(i,j)/st(j) est la fréquence relative des variables de la classe i pour l'observation j. Cette valeur a d'autant plus de chances d'être grande que S(i) est grande, puisque s(i,j) vaut en moyenne S(i)/n, où n est le nombre d'observations.

On peut alors donner la règle d'affectation :

L'observation j est affectée à la classe i
si f(i,j) > c * F(i)

On peut donner au coefficient c une valeur positive quelconque. Si c vaut 1, cette règle s'exprime ainsi : l'affectation est réalisée si l'importance relative des variables de la classe est plus grande pour l'observation que pour l'ensemble de l'échantillon. Si c est plus grand que 1, la condition est plus restrictive. Par exemple, si c = 1.5, l'importance relative des variables de la classe doit être au moins 50 % plus grande pour les observations qui y seront associées que pour l'ensemble des observations. Si c vaut 0, seules les observations qui n'ont que des zéros pour l'ensemble des variables d'une classe n'y sont pas affectées. Les classes d'observations qu'on obtient ne sont pas disjointes, car l'affectation d'une observation à une classe selon le critère ci-dessus n'empêche pas son affectation à d'autres classes. Autrement dit, nous n'obtenons pas une partition des observations. Ceci n'est pas génant, et cela peut même être intéressant dans un cas comme celui qui est donné en exemple ci-dessous : en effet, il semble beaucoup plus normal qu'un individu puisse être associé à plusieurs types de loisirs qu'à un seul.

B) Exemple : observations associées aux classes de loisirs

Nous avons repris les 8 classes de loisirs décrites dans le chapitre précédent, et nous avons créé des classes d'observations selon la technique qui vient d'être décrite. Ces classes d'observations ont été ensuite étudiées à l'aide de tris à plat, de la même manière que dans l'exemple sur les classes d'exploitations agricoles. Si on s'en tient aux caractéristiques personnelles, on obtient les résultats sommairement résumés suivants :

Classe 1. 54 questionnaires (11 %)

Loisirs : cyclisme, athlétisme, football
Majorité d'hommes (76 %), jeunes, célibataires, dont les revenus sont quelconques. Etudiants et surtout apprentis sont sur-représentés.

Classe 2. 232 questionnaires (47 %)

Loisirs : plage-piscine, cinéma, dancings, ...
Hommes et femmes en nombres égaux. Plutôt jeunes et célibataires. Sous-représentation des ménagères et sur-représentation des étudiants et apprentis. En ce qui concerne la formation, sous-représentation de ceux qui ont un niveau d'instruction primaire.

Classe 3. 111 questionnaires (23 %)

Loisirs : collections d'objets divers et d'objets de la nature, pêche, salles de jeux
Majorité d'hommes (72 %) de tous âges, plutôt célibataires, plus d'apprentis que d'étudiants. Formation quelconque.

Classe 4. 268 questionnaires (54 %)

Loisirs : promenade, pique-nique, radio, TV, ...
Légère sur-représentation des femmes. Personnes plutôt âgées. Prédominance des bas revenus. Peu de célibataires, tous les veufs, 72 % des divorcés-séparés, 70 % des ménagères. Peu d'étudiants et apprentis. Sur-représentation de la classe ouvrière. Forte sur-représentation du niveau d'instruction primaire et forte sous-représentation du niveau supérieur.

Classe 5. 194 questionnaires (39 %)

Loisirs : peinture-sculpture, faire du cinéma, de la photo, des enregistrements, littérature, bricolage, alpinisme.
Majorité d'hommes. Les ménagères et les apprentis sont sous-représentés au profit des étudiants et des actifs. Légère sur-représentation des classes sociales supérieures. Formations secondaires et supérieures sur-représentées.

Classe 6. 57 questionnaires (12 %)

Loisirs : ski-nautique, voile, tennis, basket, volley.
Hommes et femmes en proportions égales. Plutôt jeunes et célibataires, apprentis et étudiants. Revenus et classe sociale élevés. Niveaux d'instruction supérieurs sur-représentés.

Classe 7. 42 questionnaires (8.5 %)

Loisirs : partis politiques, sociétés savantes.
Majorité d'hommes. Sous-représentation des moins de 30 ans, et surtout des moins de 25 ans. Sur-représentation des revenus élevés et des niveaux d'instruction supérieurs.

Classe 8. 197 questionnaires (40 %)

Loisirs : yoga, théâtre, associations professionnelles, cours, conférences-débats, ballets, concerts, musique, équitation.

Sexe et âge indifférents. Revenus élevés. Sur-représentation des étudiants et des couches sociales supérieures. Forte sur-représentation de la classe moyenne. Niveau d'instruction élevé.

VII. LES PROGRAMMES

Tous les programmes utilisés ont été écrits en Fortran II très rudimentaire. Ils ont été créés pour fonctionner sur une petite machine de conception ancienne (traitement par lots uniquement, un seul utilisateur à la fois, pas de mémoire virtuelle). Il n'est pas très intéressant de donner ici les listes complètes des sources de ces programmes, car l'acquisition d'un nouvel ordinateur par l'Université de Neuchâtel est trop récente pour qu'ils aient eu le temps d'être remaniés de façon à profiter des possibilités actuelles. En fait, ils ont simplement été transportés d'une machine à l'autre, en étant simplement adaptés pour qu'ils fonctionnent. Nous allons donc seulement donner une liste des programmes utilisés pour cette étude, puis nous verrons de quoi ils ont besoin pour fonctionner, enfin nous ferons quelques remarques sur ce que devrait offrir un système de dépouillement en vue notamment de simplifier le travail des utilisateurs.

1. Liste des programmes utilisés

La plupart des programmes ont été conçus et réalisés grâce à la participation de plusieurs personnes dans des proportions variables. Nous ne citerons pas de noms, mais il s'agit toujours du Professeur A. Strohmeier et/ou d'un(e) de ses collaborateurs(trices).

1) Programmes d'analyse factorielle

- AN1 Analyse factorielle des correspondances
- AO1 Projections d'observations supplémentaires après analyse des correspondances
- AV1 Projections de variables supplémentaires
- CP1 Analyse factorielle en composantes principales
- CO1 Projections d'observations supplémentaires après analyse en composantes principales
- CV1 Projections de variables supplémentaires
- DI1 Dessins sur imprimante de points dans les plans factoriels
- DE1 Dessins de ces points sur traceur

2) Programmes de classification automatique

- CLHB Classification hiérarchique ascendante. Les caractéristiques des objets à classer sont les réponses aux questions.
- CLHP Classification hiérarchique ascendante. Les caractéristiques des objets à classer sont leurs projections sur des axes factoriels.
  Ces deux programmes permettent de dessiner l'arbre hiérarchique sur une imprimante.
- CLAFI Partitionnement, méthode KMEAN.

3) Programmes de tris

- FIO Programme de tris à plat, prévu pour des tris systématiques. Plusieurs possibilités de filtre(s).
- SOCIO Permet de faire 8 types standards de tableaux. Les lignes et les colonnes ont des entêtes.
- TOCE Permet de faire des tris croisés quelconques : chaque ligne et chaque colonne du tableau à construire doit être décrite.

- ITDCE Permet l'impression des tableaux construits avec TDCE.

4) Programmes divers

- AFOBS Forme des classes d'observations par association à des classes données de variables
- COBIN Permet de passer d'une codification en 0-n à une codification en n ou n-1 variables binaires.

2. Fichiers utilisés par les programmes

Quel que soit le système de dépouillement utilisé, qu'il soit composé d'un seul gros programme ou de plusieurs petits, il est nécessaire de faire un travail préliminaire de préparation qui consiste à créer le ou les fichiers contenant les informations relatives à l'enquête. Pour donner une idée du genre et du contenu des fichiers qui peuvent être utilisés pour un dépouillement, nous allons décrire les fichiers utilisés par le programme SOCIO.

Rappelons tout d'abord ce que nous entendons ici par fichier : il s'agit d'un ensemble d'enregistrements stockés sur disque magnétique et contenant de l'information. Si les enregistrements ont tous la même structure, on peut se représenter un fichier sous la forme d'un tableau : le tableau est le fichier, une ligne du tableau est un enregistrement et les colonnes correspondent aux zones des enregistrements. Il découle de cela que la description d'un tel fichier est complète si on a décrit un enregistrement et qu'on indique leur nombre. Les enregistrements sont numérotés, ils sont repérés par leur numéro.

1) Fichier des données

C'est le fichier dans lequel sont reportées les informations contenues dans chacun des questionnaires. Toutes les zones ont la même longueur, la première contient l'identification du questionnaire, les suivantes les variables, qui sont repérées par leur numéro d'ordre dans l'enregistrement, et les deux dernières peuvent ne rien contenir. Elles ne sont utilisées que par les programmes d'analyse factorielle.

2) Fichier des renseignements sur l'enquête

Ce fichier n'a qu'un enregistrement, qui indique le nombre de questionnaires, le nombre de variables par questionnaire et le titre général de l'enquête.

3) Fichiers des en-têtes

Les lignes et les colonnes des tableaux construits par le programme SOCIO ont une en-tête. Les en-têtes des lignes correspondent aux différentes modalités de la variable en ligne. Il en est de même pour les en-têtes des colonnes. A chaque variable est donc associé un groupe d'en-têtes utilisées lorsque les lignes ou les colonnes d'un tableau sont formées par ses différentes modalités. Ces groupes d'en-têtes sont stockés dans un fichier. Chaque enregistrement de ce fichier contient un de ces groupes. Il n'y a généralement pas autant de groupes d'en-têtes que de variables, car certains groupes peuvent s'appliquer aux modalités de plusieurs variables. C'est le cas, notamment, d'un groupe d'en-têtes tel que "SANS REPONSE", "NON", "OUI", qui convient aussi bien aux modalités de la réponse à la question "AVEZ-VOUS LA RADIO ?", qu'aux modalités découlant d'une question telle que "VOS ENFANTS SONT-ILS TOUS EN AGE DE SCOLARITE ?". La liaison entre une variable et le groupe des en-têtes qui correspondent à

ses modalités est faite en indiquant le numéro de l'enregistrement où se trouve ce groupe d'en-têtes. Cette indication est donnée dans le fichier de description des variables, mentionné ci-dessous.

4) Fichier de description des variables

Chaque enregistrement est relatif à une variable. Il contient le numéro de la variable, son maximum, le numéro de l'enregistrement qui contient les en-têtes relatives à cette variable, et son libellé (par exemple : ETAT CIVIL).

En plus de ces différents fichiers, le programme doit encore recevoir les indications qui lui feront faire ce qu'on attend de lui. Ces indications sont les paramètres qui indiquent les numéros des variables à traiter, les types de tableaux en pourcentages désirés, le titre du tableau, et d'autres indications éventuelles. Ces informations n'ont pas besoin d'être mémorisées de manière permanente puisque chaque ensemble de paramètres est relatif à un tableau particulier.

3. Remarques sur l'utilisation et la création de programmes

Dans le passé, le seul moyen pour un utilisateur d'accéder à un ordinateur était de perforer des cartes contenant les indications nécessaires pour faire exécuter le traitement souhaité. Ces indications étaient composées d'instructions données à la machine, et des paramètres fournis au programme utilisé. Il fallait ensuite donner ces cartes à un opérateur et attendre la sortie des résultats, ce qui impliquait des délais très longs par rapport aux temps de calculs. Pour donner

les bons paramètres dans le bon ordre, il fallait se référer constamment à un mode d'emploi du programme utilisé. Dans ces conditions, il était difficile pour un sociologue de se passer du concours permanent d'un informaticien pour dépouiller une enquête à l'aide d'un ordinateur.

A l'heure actuelle, l'accès normal à une machine se fait par l'intermédiaire de consoles (clavier et écran) accessibles au public. Si les programmes sont bien faits, il suffit de connaître quelques opérations élémentaires (connecter la console, lancer l'exécution d'un programme, répondre à une question affichée sur l'écran, afficher et/ou imprimer des résultats, quitter la console) pour utiliser la machine. Cet apprentissage demande un effort minime (moins de deux heures), et permet au sociologue de ne plus faire appel à un conseiller au cours du dépouillement, sauf exceptions. L'avantage est évident : l'utilisateur peut répondre lui-même et tout de suite à ses demandes, d'où réduction des coûts et surtout des délais. Nous l'avons dit, pour que le chercheur puisse utiliser l'ordinateur aussi facilement qu'une machine à calculer, il faut des programmes adaptés à leur usage par un non-spécialiste. Ceci implique que ces programmes :

- Prennent en charge tous les aspects techniques (noms des fichiers, leurs structures, où mettre les résultats, etc.).
- Guident l'utilisateur en posant des questions qui demandent tous les paramètres nécessaires, mais aucun paramètre inutile, compte tenu des paramètres déjà donnés.
- Renseignent l'utilisateur, en affichant sur demande la signification de la question posée et les possibilités de réponses.
- Contrôlent, dans la mesure du possible, la validité des réponses.

On peut construire un tel système en adoptant la structure logique décrite dans la figure 17. Le système est composé d'un programme (A) qui fait l'aiguillage vers le traitement désiré et qui prend en charge les problèmes techniques. Chaque traitement se décompose en deux parties : dans la première (P), l'utilisateur est invité à donner les paramètres nécessaires, tandis que les calculs sont faits dans la seconde partie (C). Dans notre cas, les possibilités de calculs des différents traitements seraient celles correspondant aux programmes de la liste donnée ci-dessus. Cette structure présente un certain nombre d'avantages :

1) La modularité

Il est facile d'implanter une nouvelle application, car la partie C (calculs) puis la partie P d'introduction des paramètres peuvent être développées indépendamment du système déjà existant, et y être ajouté au prix d'une modification simple de la partie A.

2) Le choix du mode de calcul

Une fois les paramètres introduits à l'aide de P, l'exécution du programme de calcul C peut se faire en interractif (la console est bloquée jusqu'à la fin de l'exécution) ou en batch (l'exécution est prise en charge par le système de l'ordinateur, qui libère la console pour d'autre tâches). L'utilisateur peut choisir en fonction du temps présumé des calculs.

3) Séparation des programmes selon leurs fonctions

Chaque traitement met en jeu 3 programmes, soit le programme d'aiguillage, commun, le programme de remplissage des paramètres et le programme de calcul, spécifiques.

STRUCTURE D'UN SYSTEME DE DEPOUILLEMENT

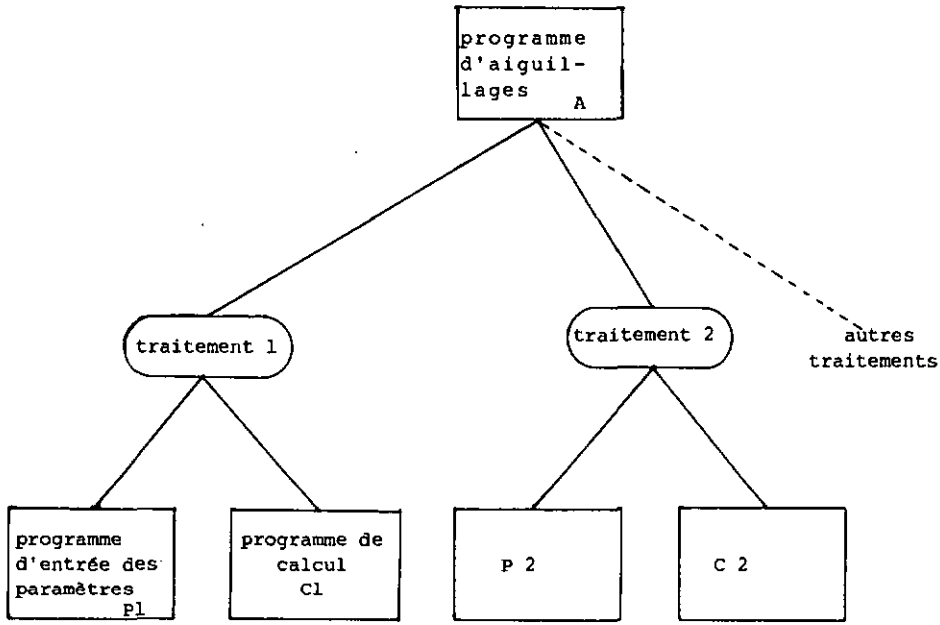


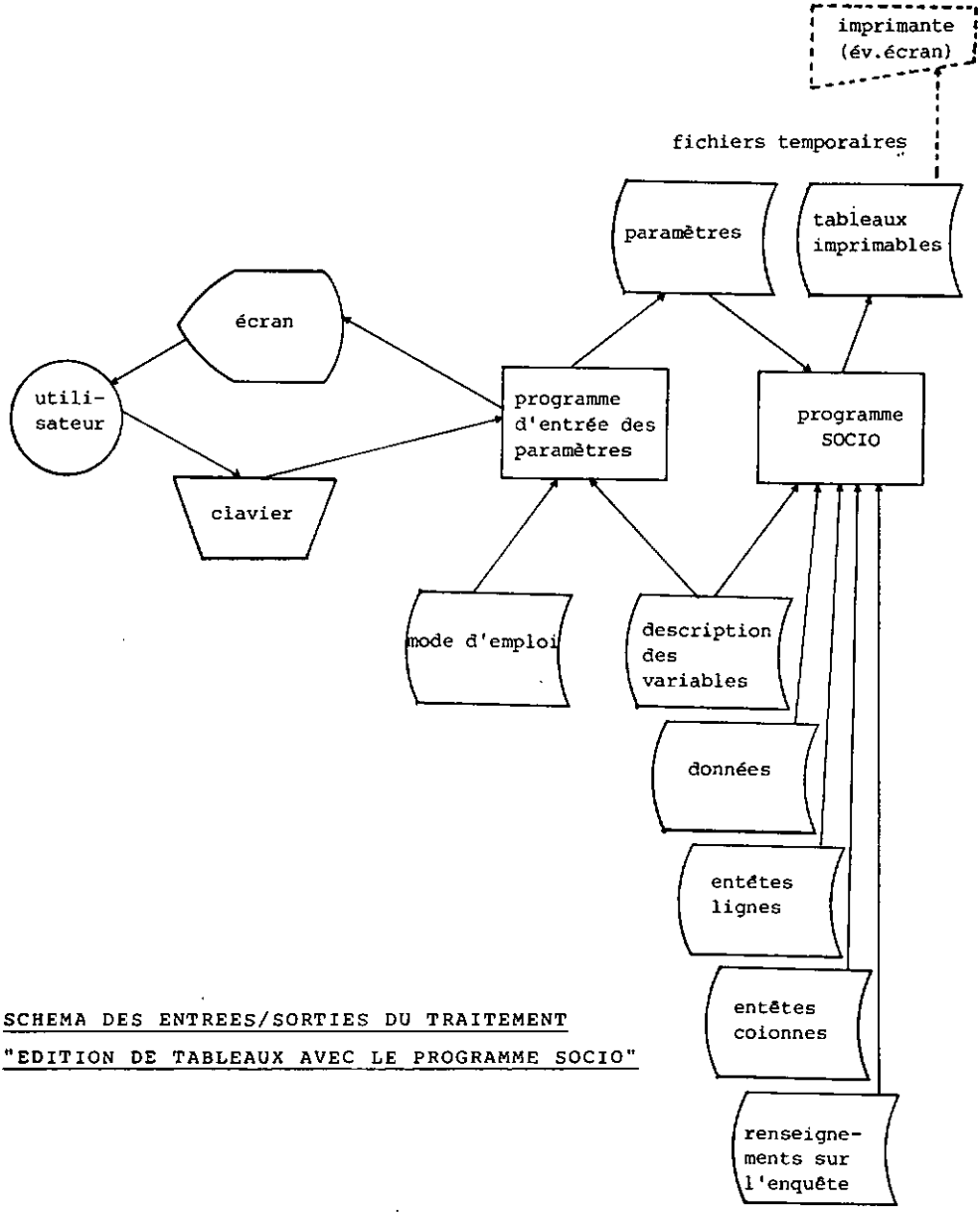
Figure 17

Le programme d'aiguillage, qui prend en charge les aspects techniques du traitement, contient forcément des instructions propres à la machine utilisée. Il est donc particulièrement intéressant de l'écrire dans le langage de commande de celle-ci. Le programme de calcul est écrit dans un langage évolué, comme le Fortran. On peut considérer qu'il est au point dès qu'il fournit des résultats corrects. Une fois testé, il n'y a pratiquement pas lieu d'y apporter des changements. L'utilisateur n'en attend que des résultats; il ne s'intéresse pas à la manière dont sont faits les calculs. Le programme de remplissage des paramètres est celui avec lequel l'utilisateur est en contact, puisque ce programme pose des questions auxquelles l'utilisateur répond. Pour être au point, ce programme ne doit pas seulement fournir au programme de calcul les bons paramètres, mais il doit permettre à l'utilisateur de les donner. Pour cela, il faut que les questions qu'il pose soient claires, ainsi que les explications qu'il fournit. Or, c'est un aspect que maîtrise assez mal celui qui a fait le programme de calcul, et qui connaît donc très bien la signification des paramètres. Ainsi ce programme ne pourra avoir sa forme définitive qu'après avoir été utilisé un certain nombre de fois, et avoir été critiqué par ceux qui l'utilisent. Il est donc souhaitable que ce programme soit aussi simple que possible, car il devra probablement être modifié plusieurs fois au début de sa mise en service.

Notons encore que ce programme doit afficher deux sortes de textes : les questions, qui sont courtes, et les explications, qui prennent généralement plusieurs lignes. Pour faciliter le travail de mise au point et d'amélioration, il est préférable que les textes explicatifs(dont l'ensemble forme en quelque sorte le mode d'emploi du programme de calcul) soient complètement extérieurs au programme, dans un fichier directement accessible et modifiable à l'aide de l'éditeur de la machine.

Ainsi, il n'est même pas nécessaire de toucher au programme d'introduction des paramètres pour compléter les explications fournies à l'utilisateur.

La figure 18 schématise les entrées/sorties d'un traitement particulier, l'édition de tableaux à l'aide du programme SOCIO. Un exemple de dialogue relatif à ce traitement, ainsi que les résultats obtenus, sont donnés en annexe.



SCHEMA DES ENTREES/SORTIES DU TRAITEMENT
"EDITION DE TABLEAUX AVEC LE PROGRAMME SOCIO"

Figure 18

ANNEXE A : EXEMPLE DE DIALOGUE POUR UTILISER LE PROGRAMME "SOCIO"

La première page montre un exemple de dialogue entre le programme qui permet d'introduire les paramètres du programme SOCIO et un utilisateur. Les réponses de l'utilisateur sont soulignées.

Les tableaux qui suivent sont la reproduction exacte des tableaux dont l'édition a été demandée dans l'exemple de dialogue.

```
$ SOCIO
NOM DES DONNEES: HAINARD

ENTREE DES PARAMETRES POUR LE PROGRAMME
D'EDITION DE TABLEAUX.

LIGNES : ?

 IL Y A PLUSIEURS POSSIBILITES:

 0    STOP L'ENTREE DE PARAMETRES

 N    LES LIGNES DU TABLEAU SERONT FORMEES PAR LES DIFFERENTES
      MODALITES DE LA VARIABLE DONT LE NUMERO D'ORDRE EST N.

 -N   LES LIGNES SERONT FORMEES PAR N VARIABLES, DONT LES NUMEROS
      D'ORDRE SERONT DEMANDES PLUS LOIN. LA CONDITION D'ENTREE
      DANS UNE LIGNE EST LA SUIVANTE: L'OBSERVATION ENTRE DANS LA
      LIGNE SI LA VALEUR DE LA VARIABLE QUI FORME CETTE LIGNE EST
      COMPRISE ENTRE UN MINIMUM ET UN MAXIMUM, DEMANDES PLUS LOIN.
      CETTE CONDITION EST LA MEME POUR TOUTES LES LIGNES.

 ON PEUT EDITER DES TABLEAUX STANDARDS (CROISEMENT DE DEUX VARIABLES,
 TITRE FORME PAR LE NOM DES VARIABLES, POURCENTAGES SUR LES LIGNES
 ET LES COLONNES) EN DONNANT LE LIBELLE (LE NOM) DES VARIABLES EN
 LIGNES ET (A LA QUESTION SUIVANTE) EN COLONNES, AU LIEU DU NUMERO.
 DANS CE CAS, SI UN FILTRE EST DEMANDE, IL SERA FORME SUCCESSIVEMENT
 PAR TOUTES LES MODALITES DE LA VARIABLE UTILISEE (D'OU L'EDITION
 D'AUTANT DE TABLEAUX QUE LA VARIABLE EN FILTRE A DE MODALITES).

LIGNES : 6
COLONNES : 11
FILTRE : 0
NBRE DE LIGNES POUR LE TITRE (1 OU 2) : 1
NBRE DE TABLEAUX DE POURCENTAGE (0 A 5) : ?

 ON PEUT EDITER 0 1 2 3 4 OU 5 TABLEAUX DE POURCENTAGES POUR CHAQUE
 TABLEAU EN NOMBRES ABSOLUS. IL FAUT INDIQUER ICI LE NOMBRE SOUHAITE
 DE CES TABLEAUX. LES CODES DES TYPES DE POURCENTAGES SERONT
 DEMANDES PLUS LOIN.


 LES DIFFERENTS TYPES DE TABLEAUX EN POURCENTAGES ET LEURS CODES
 SONT LES SUIVANTS:

 CODE     TYPE
  1       POURCENTAGES CALCULES SUR LA SOMME DU TABLEAU.
  2       POURCENTAGES EN LIGNES.
  3       POURCENTAGES EN COLONNES.
  4       POURCENTAGES CALCULES SUR LE NOMBRE DE QUESTIONNAIRES
  5       POURCENTAGES CALCULES SUR LE NOMBRE D'OBSERVATIONS ENTREES
          DANS LE TABLEAU (UNE OBSERVATION PEUT ENTRER PLUSIEURS FOIS
          SI LES LIGNES ET/OU LES COLONNES SONT FORMEES DE PLUSIEURS
          VARIABLES.

NBRE DE TABLEAUX DE POURCENTAGE (0 A 5) : 0
TITRE DU TABLEAU, 1 LIGNE(S) :
NOMBRE DE PARCELLES EN FONCTIO DE LA SURFACE DE TERRES OUVERTES
LIGNES : MAIN-D'OEUVRE
COLONNES : MCHINES
 PAS TROUVE DE VARIABLE  MCHINES
COLONNES : MACHINES
FILTRE : 0
LIGNES : 0
CALCULS EN BATCH OU EN INTERACTIF ? (B/I): I
LES RESULTATS SONT DANS LE FICHIER HAINARD.RES;13
$
```

TABLEAU ENQUETE VAL-DE-TRAVERS

NOMBRE DE PARCELLES EN FONCTIO DE LA SURFACE DE TERRES OUVERTES

	-	< 100	1 - 5 HA	5 -10 HA	10-20 HA	TOTAL
S.R.	13	5	2	1	0	21
1 PARCE.	32	23	26	1	1	83
2 - 4	53	26	28	5	0	112
5 - 9	18	11	16	4	2	51
10 ET +	5	4	9	5	1	24
TOTAL	121	69	81	16	4	291

TABLEAU ENQUETE VAL-DE-TRAVERS

MAIN-D'OEUVRE / MACHINES

	S.R.	1 MACNI.	2	3 - 4	5 - 6	TOTAL
S.R.	0	0	0	0	0	0
1 PERS.	16	12	6	5	0	39
2	25	17	35	59	28	164
3 - 4	3	7	16	32	19	77
5 +	1	0	1	6	3	11
TOTAL	45	36	58	102	50	291

TABLEAU ENQUETE VAL-DE-TRAVERS

MAIN-D'OEUVRE / MACNINES

	S.R.	1 MACNI.	2	3 - 4	5 - 6	TOTAL
S.R.	0.0	0.0	0.0	0.0	0.0	0.0
1 PERS.	41.0	30.8	15.4	12.8	0.0	100.0
2	15.2	10.4	21.3	36.0	17.1	100.0
3 - 4	3.9	9.1	20.8	41.6	24.7	100.0
5 +	9.1	0.0	9.1	54.5	27.3	100.0
TOTAL	15.5	12.4	19.9	35.1	17.2	100.0

TABLEAU ENQUETE VAL-DE-TRAVERS

MAIN-D'OEUVRE / MACHINES

	S.R.	1 MACHI.	2	3 - 4	5 - 6	TOTAL
S.R.	0.0	0.0	0.0	0.0	0.0	0.0
1 PERS.	35.6	33.3	10.3	4.9	0.0	13.4
2	55.6	47.2	60.3	57.8	56.0	56.4
3 - 4	6.7	19.4	27.6	31.4	38.0	26.5
5 +	2.2	0.0	1.7	5.9	6.0	3.8
TOTAL	100.0	100.0	100.0	100.0	100.0	100.0

ANNEXE B : LISTES DE PROGRAMMES

La liste du programme SOCIO et de l'un de ses sous-programmes est donnée à titre informatif. On notera seulement les lignes 12 à 17, qui permettent de remplir le tableau. Ce remplissage se fait sans aucun test, ce qui n'est possible que si les variables sont codées à partir de 0, et qu'on a contrôlé que leurs maximums effectifs sont égaux aux maximums indiqués dans le fichier de description des variables. Cette économie de test permet bien sûr une économie en temps de calcul.

```
      SUBROUTINE SOC3(X,IA,INDL,INDC,NL,NC,QP)
      DIMENSION IA(41,11),X(1)
      COMMON NU1,NQS,NVAR,TITRE(4),NASSL(2),NASSC(2)
      READ(3'INDL) NASSL
      READ(3'INDC) NASSC
      NL=NASSL(1)+1
      NC=NASSC(1)+1
      DO 1 I=1,41
      DO 1 J=1,11
1     IA(I,J)=0
      NU1=1
      DO 2  I=1,NQS
            READ(1'NU1) ANOM,(X(N),N=1,NVAR),T,IK
            IC=X(INDC)+1
            IL=X(INDL)+1
            IA(IL,IC)=IA(IL,IC)+1
2           CONTINUE
      CALL SOCB(NL,NC,IA)
      QP=FLOAT(IA(NL+1,NC+1))
      RETURN
      END
```

```
      PROCRAM SOCIO
      INTEGER*2 EL(8,41),EC(4,11,2)
      DIMENSION X(1500),IA(41,11) ,PO(41,11),ISC(40),ISL(40),ISF(10)
      DIMENSION T(20,2),NPO(5),MINF(10).MAXF(10),FIL(4,2)
      COMMON NU1,NQS,NVAR,TITRE(4),NASSL(2),NASSC(2)
C
C  INITIALISATIONS
C
      DATA BLANC/'    '/
      DEFINE FILE 3(1500,24,U,NU3),2(1,12,U,NU2)
      DEFINE FILE10(200,80,U,NU10),20(200,320,U,NU20)
      LEC=5
      IMP=6
      READ(2'1) NQS,NVAR,TITRE
      NVAR2=2*(NVAR+3)
      DEFINE FILE 1(NQS,NVAR2,U,NU1)
      QS=FLOAT(NQS)
10    NC=0
      NL=0
      DO 11 I=1,4
      DO 11 J=1,2
11    FIL(I,J)=BLANC
C
C  LECTURE DES PARAMETRES
C
      READ(LEC,1000) INDL,INDC,INDF,NLTI,(NPO(I),I=1,5),MINL,MAXL,MINC,M
     VAXC,MINF(1),MAXF(1),TAB
      IF(INDL) 30,20,30
20    CALL EXIT
30    READ(LEC,1001)((T(I,J),I=1,20),J=1,NLTI)
      NF=0
      ISL(1)=INDL
      IF(INDL.GE.0) GOTO 40
      NL=IABS(INDL)
      READ(LEC,1002) (ISL(I),I=1,NL)
40    ISC(1)=INDC
      IF(INDC.GE.0) GOTO 50
      NC=IABS(INDC)
      READ(LEC,1002) (ISC(I),I=1,NC)
50    IF(INDF.EQ.0) GOTO 70
      NF=1
      ISF(1)=INDF
      IF(INDF.GT.0) GOTO 60
      NF=IABS(INDF)
      READ(LEC,1003)(ISF(I),MINF(I),MAXF(I),I=1,NF)
60    READ(LEC,1001)((FIL(I,J),I=1,4),J=1,2)
70    CONTINUE
C
C  AIGUILLAGE
C
      IF(INDC.LT.0 .AND.INDL.GT.0) GOTO 300
      IF(INDC.NE.0) GOTO 80
      IF(INDL.GT.0) NTYP=1
      IF(INDL.LT.0) NTYP=2
      GOTO 90
80    K=0
      IF(INDL.LT.0) K=K+2
      IF(INDC.LT.0) K=K+2
      IF(INDF.NE.0) K=K+1
      NTYP=3+K
90    GOTO (100,110,120,130,140,150,160,170),NTYP
```

```
C
C  REMPLISSAGE ET IMPRESSION DES DIVERS TYPES DE TABLEAUX
C
C  TRI A PLAT, UNE VARIABLE
100    CALL SOC1(INDL,NLTI,X,IA,PO,T,       TAB,EL,EC)
       GOTO 10
C  TRI A PLAT. PLUSIEURS VARIABLES
110    CALL SOC2(X,IA,PO,T,ISL.NL,NLTI,       TAB,EL,EC)
       GOTO 10
C  TRI CROISE, DEUX VARIABLES, SANS FILTRE
120    CALL SOC3(X,IA,INDL,INDC,NL,NC,QP)
       PC=QP
       CALL SIMP1(T,IA,NLTI,TAB,NL,NC,20,10,EL,EC,FIL,ISL,ISC)
       GOTO 200
C  TRI CROISE, DEUX VARIABLES, AVEC FILTRE
130    CALL SOC4(X,IA,INDL,INDC.ISF,MINF,MAXF,NF,NL,NC,QP)
       PC=QP
       CALL SIMP1(T,IA,NLTI,TAB,NL,NC,20,10,EL,EC,FIL,ISL,ISC)
       GOTO 200
C  TRI CROISE, UNE VARIABLE EN COLONNE, PLUSIEURS EN LIGNE, SANS FILTRE
140    CALL SOC5(X,IA,ISL,MINL,MAXL,NL,INDC,NC,PC,QP)
       CALL SIMP1(T,IA,NLTI,TAB,NL,NC,30,10,EL,EC,FIL,ISL,ISC)
       GOTO 200
C  TRI CROISE, UNE VARIABLE EN COLONNE, PLUSIEURS EN LIGNE, AVEC FILTRE
150    CALL SOC6(X,IA,ISL,MINL,MAXL,NL,INDC,NC,ISF,MINF,MAXF,NF,PC,QP)
       CALL SIMP1(T,IA,NLTI,TAB,NL,NC,30,10,EL,EC,FIL,ISL,ISC)
       GOTO 200
C  TRI CROISE, PLUSIEURS VARIABLES EN COLONNE ET EN LIGNE, SANS FILTRE
160    CALL SOC7(X,IA,ISL,MINL,MAXL,NL,ISC,MINC,MAXC,NC,PC,QP)
       CALL SIMP1(T,IA,NLTI,TAB,NL,NC,30,40,EL,EC,FIL,ISL,ISC)
       GOTO 200
C  TRI CROISE, PLUSIEURES VARIABLES EN LIGNE ET EN COLONNE, AVEC FILTRE
170    CALL SOC8(X,IA,ISL,MINL,MAXL,NL,ISC,MINC,MAXC,NC,ISF,MINF,MAXF,NF
     *,PC,QP)
       CALL SIMP1(T,IA,NLTI,TAB,NL,NC,30,40,EL,EC,FIL,ISL,ISC)
C
C  TABLEAUX DE POURCENTAGES ET LEURS IMPRESSIONS
C
200    I=0
       MISP=0
210    I=I+1
       IF(NPO(I).LE.0.OR.I.GT.5) GOTO 10
       NPOI=NPO(I)
       GOTO(220,250,250,230,240),NPOI
220    R=QP
       GOTO 250
230    R=QS
       GOTO 250
240    R=PC
250    CALL SPOUR(IA,PO,NC,NL,NPO(I),R)
       CALL SIMP2(T,PO,NLTI,TAB,NL,NC,EL,EC,FIL,       MISP)
       GOTO 210
C
300    WRITE(IMP,1004)
       GOTO 10
C
1000   FORMAT(4I5,5I1,6I5,A4)
1001   FORMAT(20A4)
1002   FORMAT(16I5)
1003   FORMAT(5(I5,2I4,2X))
1004   FORMAT(/1X,'PARAMETRES INCOMPATIBLES')
       END
```

```
      PROGRAM LECTURE
C
C  CE PROGRAMME PERMET DE LIRE DES DONNEES ET DE LES STOCKER
C  EN CODE BINAIRE, DIRECTEMENT UTILISABLE PAR LA MACHINE,
C  SANS AUCUN CONTROLE NI RECODAGE.
C
      DIMENSION X(nombre de variables + 2)
      LEC = unite d'entree
      IMP = unite de sortie
      NVAR2 = 2 * (nombre de variables + 3)
      DEFINE FILE IMP(32000.NVAR2,U,NO)
      NO = 1
1     READ(LEC,100,END=2) ANOM,(X(I),I=1,nombre de variables)
      WRITE(IMP'NO) ANOM,(X(I),I=1,nombre de variables + 2)
      GOTO 1
2     STOP
      END
```

ANNEXE C : BIBLIOGRAPHIE

Statistique multivariée

Benzécri J.-P.
L'analyse des données (2 volumes)
Dunod, Paris, 1973
Ouvrage de base sur l'analyse des correspondances (2è volume).
Abord difficile. Contient des exemples d'applications.

Lebart L. et Fénelon J.-P.
Statistique et informatique appliquées
Dunod, Paris, 2è édition, 1973
Développement mathématique complet sur l'analyse factorielle.

Lebart L., Morineau A. et Tabard N.
Techniques de la description statistique
Dunod, Paris, 1977
On y trouve l'application de l'analyse des correspondances au dépouillement d'enquêtes. Contient, par ailleurs, un chapitre intéressant sur la validité des résultats. Nombreux listings de programmes en Fortran.

Gendre F.
L'analyse statistique multivariée
Droz, Genève, 1976

Strohmeier A.
L'analyse factorielle des correspondances : les champs d'application de la méthode et la codification des données.
Cahiers de Méthode quantitative, Neuchâtel, 1975

Strohmeier A.
L'analyse factorielle : aspect mathématique
Cahiers de Méthode quantitative, Neuchâtel, 1977

Volle M.
Analyse des données
Economica, Paris, 1978

Anderberg M.-R.
Cluster analysis for applications
Academic Press, New-York, 1973

Hartigan J.-A.
Clustering, algorithms
J.Wiley, New-York, 1975

Späth N.
Cluster - Analyse - Algorithmen zur Objektklassifizierung
R. Oldenbourg, München, 1975
Contient des listings de programmes Fortran de classification automatique.

Graf-Jacottet M.
Classification automatique : aspects mathématiques
Cahiers de Méthode quantitative, Neuchâtel, 1979

Strohmeier A.
La classification automatique : les bases
Cahiers de Méthode quantitative, Neuchâtel, 1977
Très accessible. Excellente introduction aux notions utilisées en classification automatique.

Divers

Burgat P.
Compléments de statistique mathématique
Cahiers de Méthode quantitative, Neuchâtel, 1975

Pfaffenberger R.-C.et Patterson J.-H.
Statistical methods for business and economics
Irwin, Homewood, 1977
Présentation claire et complète de la statistique classique.

Cherry
Pratique des enquêtes statistiques
Puff, Paris, 1961
Vieux (on n'y parle pas d'ordinateur).

Flament C.
L'analyse booléenne de questionnaire
Mouton, Paris, 1976
Purement théorique.

Javeau C.
L'enquête par questionnaire
Editions d'organisation, 2è édition, 1978
L'enquête de A jusqu'à Z, mais le dépouillement n'est fait que de tris (sur diverses machines : l'ordinateur ou ... la trieuse).

Il existe un certain nombre de packages commercialisés de traitements statistiques, comme P-STAT, GENSTAT, SPSS, GLIM. Tous demandent un effort assez important d'apprentissage.