A.D.I.S.H.
Association pour le Développement de
l'Informatique dans les Sciences de l'Homme

12, rue de l'Ecole de Médecine
75270 PARIS - Cédex 06

Jean-Paul GRÉMY

avec la collaboration de Marie-Joëlle LE MOAN

Analyse de la démarche de construction de typologies dans les sciences sociales

Compte-rendu de fin d'étude
d'une recherche financée par la Délégation
Générale à la Recherche
Scientifique et Technique

Action complémentaire coordonnée
"Informatique et Sciences Humaines"

Septembre 1976

Décision d'aide n° 74 7 0345

Republié dans *Informatique et sciences humaines, n° 35* (1977)

RÉSUMÉ

Ce rapport analyse la manière dont procèdent les chercheurs en sciences sociales pour élaborer une typologie "à la main" (sans aide informatique). Les procédures décrites diffèrent sensiblement des algorithmes mis en oeuvre dans les programmes de classification automatique ; en particulier, elles prennent en compte non seulement les informations statistiques (au niveau des données d'observation), mais également et simultanément les informations sémantiques (au niveau des concepts et des théories).

Automatiser les procédures manuelles présenterait à la fois de grandes difficultés pour l'informaticien, et peu d'intérêt pratique pour l'utilisateur. Par contre, il est possible soit d'amender les programmes existants en les rendant conversationnels, soit de mettre au point un système interactif d'aide à la construction de typologies : la description d'un tel système constitue la conclusion de ce rapport.

SUMMARY

This report analyses the way researchers in social sciences go about classifying data without computer assistance. The processes described here differ notably from the algorithms used in automatic classification programs. In particular, they take into account not only numerical data (at the level of observed facts), but also theoretical information (at the implicit level of concepts and theories).

The simple automation of such processes would raise serious problems for the computer programmer, and prove of little practical value to the social scientist. On the other hand, it seems possible to improve existing programs by making them interactive. But it would be equally possible to apply some of the conclusions of this analysis by developping a new interactive system, which could perform a variety of elementary data transformations, as an assistance to the construction of typologies. Such a system is described in the last chapter of this report.

TABLE DES MATIÈRES

Introduction 3

0.1. Motivations de la recherche 3

0.2. Objectifs de la recherche 4

0.3. Principales conclusions 6

1 - La méthode utilisée 8

1.1. Position du problème 8

1.2. L'analyse bibliographique 9

1.3. L'enquête auprès des chercheurs 12

1.4. La comparaison avec les procédures automatisées 14

2 - Description de la démarche de construction de typologies dans les sciences sociales 15

2.1. Dans quels cas élabore-t-on une typologie ? 15

2.2. Comment procède-t-on pour obtenir des types ? 17

2.2.1. La constitution de types-idéaux 18

2.2.2. La réduction d'un espace d'attributs 23

2.2.3. L'agrégation autour d'unités-noyaux 33

2.3. Comparaison des trois procédures : "systématisme", pragmatisme, empirisme 46

3 - Un système informatique d'aide à la construction de typologies 49

3.1. Comparaison des procédures manuelles et des procédures automatisées 49

3.2. Description d'un système de manipulation de données 53

3.2.1. Caractéristiques générales du système 54

3.2.2. La gestion des fichiers 56

3.2.3. Les opérations sur les données 60

3.2.4. Mise en oeuvre du système 64

3.3. Perspectives d'application 67

Ouvrages consultés 68

I N T R O D U C T I O N

Cette recherche a été entreprise en septembre 1974, dans le cadre de l'Action Complémentaire Coordonnée "Informatique et Sciences Humaines". Son thème correspond au titre II B de l'appel d'offres : "Traitement des données : les méthodes, leur fondement, leur justification". Elle a été terminée en juillet 1976.

Nous retraçons brièvement ci-après :

- les motivations de la recherche,
- les objectifs qui lui étaient assignés,
- les principales conclusions auxquelles nous avons abouti.

0.1. MOTIVATIONS DE LA RECHERCHE.

A l'origine de cette recherche se trouve une interrogation sur les raisons du peu d'impact de l'informatique sur les sciences sociales. En dehors du dépouillement d'enquêtes, où l'ordinateur n'a fait que prendre la succession de la mécanographie, il n'y a guère de domaines des sciences sociales où l'informatique se soit imposée. Les seules exceptions notables sont l'analyse des données et, à un degré moindre, l'élaboration de modèles (techniques de simulation).

L'évolution des méthodes d'analyse des données est intéressante à étudier. Dans une première phase, ont été automatisées des procédures statistiques déjà utilisées, mais très lourdes à manier avec les moyens traditionnels (analyse factorielle en composantes principales, p. ex.). Le progrès très sensible qui en est résulté, et la satisfaction des utilisateurs, ont suscité, dans une deuxième phase, l'apparition de besoins nouveaux. Besoins quantitatifs d'abord, le recours à l'analyse des données et la comparaison de méthodes étant mis à la portée d'un plus grand nombre de chercheurs ; besoins qualitatifs également, la demande d'outils nouveaux étant stimulée par la comparaison des méthodes d'analyse, et la

facilité d'emploi des programmes correspondants. Une troisième phase a été marquée par une floraison de méthodes originales, et de procédures mieux adaptées à des problèmes scientifiques. La quatrième phase est celle dans laquelle nous nous trouvons actuellement. Elle se caractérise par un tassement de la demande des utilisateurs, qui ne s'est pas développée proportionnellement au nombre de techniques offertes, et surtout par l'apparition d'une attitude critique contrastant avec l'engouement observé pendant la deuxième phase.

Il semble par conséquent que l'on traverse une période de relative désaffection des praticiens et chercheurs en sciences sociales vis à vis de certains apports de l'informatique. Il nous a paru utile d'en rechercher les causes, afin, si possible, d'y proposer des remèdes. Nous avons choisi de limiter cette recherche à un domaine restreint de l'analyse des données, celui de la construction de typologies dans les sciences sociales.

0.2. OBJECTIFS DE LA RECHERCHE

Le point de départ de cette recherche ne se réduit pas à la constatation d'une certaine désaffection à l'égard de l'informatique. Nous avons en outre porté le diagnostic (provisoire) suivant. La méthodologie des sciences sociales est encore en train de se constituer ; elle n'a pas atteint un degré de développement tel qu'elle puisse se traduire par un ensemble de règles et de préceptes suffisant pour guider le praticien et le chercheur. C'est pourquoi ceux-ci éprouvent souvent de grandes difficultés à exprimer clairement la manière dont ils conçoivent la validation de leurs hypothèses, et à décrire avec précision les opérations qu'il convient de faire subir à leurs données. Tant que les techniques statistiques d'analyse ont demandé de gros efforts pour leur mise en oeuvre, elles sont restées l'apanage de chercheurs ayant poussé fort loin la théorisation de leur problème ; le recours de ces techniques avait alors de fortes chances d'être fondé, et de donner satisfaction aux chercheurs. Ultérieurement, le progrès qu'a constitué la mise de ces techniques à la disposition d'un public plus vaste, grâce à l'informatique n'a pas été accompagné d'une élévation du niveau méthodologique de ce même public, et d'une mise en garde relative aux mauvais usages des outils mathématiques et informatiques. Cet état de choses a entraîné deux séries de conséquences :

1) Un nombre important d'utilisateurs insuffisamment avertis ou préparés a eu accès à des techniques trop puissantes ou trop spécifiques par rapport à leurs besoins ou leurs attentes. Il est heureusement de plus en plus rare que l'on demande à la machine de faire apparaître *la* structure immanente que recèlent les données. Il est, par contre, encore fréquent que l'utilisateur ignore les conditions précises d'utilisation d'un programme, l'existence de programmes concurrents entre lesquels il aurait dû avoir à choisir (et selon quels critères ?), l'incidence sur la typologie obtenue du codage des données ou du type de distance défini entre les individus, etc. D'où, après une phase d'engouement liée au prestige de l'ordinateur, une vague d'insatisfaction pouvant se traduire par un rejet d'autant plus brutal que la décision d'utiliser les moyens informatiques aura été plus irrationnelle.

2) Les mathématiciens et les informaticiens, encouragés par leurs premiers succès, ont cherché à améliorer leurs programmes et à créer des outils nouveaux. Si la méthodologie des sciences sociales avait atteint un développement suffisant, il aurait été facile de fixer les normes auxquelles ces outils auraient dû satisfaire (quitte à démontrer éventuellement ensuite que certaines d'entre elles étaient trop strictes ou trop contraignantes). Mais en l'état actuel de ce développement, il était dans la plupart des cas difficile, voire impossible, d'exprimer une demande précise pour un outil nouveau. Cette insuffisance de la part des spécialistes en sciences sociales, loin de tempérer le zèle des mathématiciens et informaticiens, leur a permis de donner libre cours à leur esprit d'invention. Ce faisant, ils ont en fait concédé aux considérations d'ordre mathématique ou informatique la priorité sur les exigences (d'ailleurs informulées) des sciences sociales. Dans ces conditions, le risque était grand que les outils nouveaux, quelles que soient leur qualités techniques, ne puissent satisfaire les utilisateurs potentiels.

Ces hypothèses, relatives à la distance méthodologique qui sépare les programmes de classification ou de segmentation de leurs utilisateurs, nous ont été suggérées par une série d'entretiens que nous avons eu avec divers praticiens ou chercheurs ayant fait usage de tels programmes. Elles ont été renforcées par l'analyse de comptes-rendus de recherches comportant le recours à la classification automatique ou la segmentation : dans

la majorité des cas, les résultats apportés par le programme n'ont pu être intégrés à l'exposé des résultats de la recherche, et ont dû être publiés tels quels, parfois rejetés en annexe ou même simplement omis.

C'est pourquoi nous nous sommes assigné comme objectif de décrire ce que cherche le spécialiste des sciences sociales lorsqu'il élabore une typologie, et les procédés qu'il emploie pour y parvenir (exception faite des procédures automatiques). Une fois cette démarche formalisée, il devient possible de la comparer aux algorithmes mis en oeuvre dans les procédures automatisées, d'évaluer l'adéquation de l'outil informatique aux besoins réels des utilisateurs, et de définir, s'il y a lieu, le cahier des charges d'outils nouveaux, mieux adaptés aux attentes des praticiens et chercheurs en sciences sociales.

0.3. PRINCIPALES CONCLUSIONS.

Dans les chapitres qui suivent, nous exposons la méthode utilisée (ch. 1), les résultats obtenus (ch. 2) et les conclusions pratiques auxquelles cette recherche a conduit (ch. 3). Il nous paraît utile de signaler dès maintenant que cette étude a mis en évidence les points suivants :

1) S'il est difficile de formaliser complètement la procédure suivie pour construire une typologie manuellement (ce qui explique qu'aucune description complète n'en soit donnée dans la littérature spécialisée), il est relativement aisé par contre d'isoler et de décrire la plupart des opérations sur les données qui interviennent dans ce processus.

2) L'enchaînement de ces opérations sur les données n'est pas rigide. Il dépend très largement de la nature des données et des hypothèses, de la personnalité et de la culture du chercheur, et des intuitions de celui-ci au cours du déroulement de la recherche. Au cours de l'élaboration d'une typologie intervient tout un ensemble de connaissances et de savoir-faire propre au chercheur, que nous avons appelé le *savoir implicite ;* en raison de son volume et de sa complexité, ce savoir implicite ne paraît pas pouvoir être formalisé en vue de son introduction en machine comme préalable à un traitement automatique des données.

3) Le chercheur procède par essais et erreurs : il décide de la suite des opérations à effectuer selon les résultats apportés par les opérations précédentes. Si l'on renonce à introduire dans la mémoire de l'or-

dinateur le savoir implicite du chercheur, un programme de construction de typologie devrait en principe être conversationnel.

4) La majorité des programmes de classification automatique ou de segmentation ne permettent pas une utilisation pas à pas, au vu des résultats intermédiaires du traitement. Ils ne correspondent pas aux exigences du chercheur pour la construction complète d'une typologie satisfaisante pour l'utilisateur. Ils ont cependant leur place en tant que chaînes d'opérations sur les données, dans l'ensemble du processus d'élaboration des types.

5) A côté de ces programmes relativement lourds, il est possible de concevoir un système conversationnel de *manipulation de données* qui constitue une aide informatique efficace dans la construction de typologies. Outre les services qu'il pourrait rendre dans l'immédiat, un tel système peut également constituer un moyen de perfectionnement méthodologique des spécialistes en sciences sociales, et un outil de recherche pour la conception d'outils informatiques nouveaux.

1. LA MÉTHODE UTILISÉE

Le problème central de cette recherche a été la description de la procédure effectivement suivie par un spécialiste des sciences sociales construisant une typologie sans utiliser de moyens informatiques. Nous exposons successivement :

- les termes dans lesquels se posait le problème ;
- les principes qui ont présidé à la sélection et à l'analyse de la bibliographie sur le sujet ;
- les règles que nous avons suivies pour l'observation et la formalisation de l'activité du "typologue".

1.1. POSITION DU PROBLÈME.

La classification des individus, des groupes sociaux, des institutions, des comportements, des relations entre individus ou sociétés, est une démarche extrêmement courante dans les sciences sociales. En règle générale, elle constitue même un préalable tant à la conduite d'une recherche sur le terrain qu'à l'élaboration d'une théorie. On pourrait s'attendre à ce que les manuels contiennent des indications descriptives, relatant comment on dégage une typologie d'un ensemble d'observations, ou même normatives, indiquant la meilleure procédure à suivre pour définir des types. Il n'en est rien. De même, la littérature spécialisée ne contient guère que des réflexions théoriques sur la fonction des typologies et sur la notion de type (ou ses variantes : strate, classe, etc.), ou des présentations de classifications sans que la procédure suivie pour les réaliser soit complètement décrite. On peut affirmer par conséquent, qu'il n'existe pratiquement pas de travaux publiés sur la description de la démarche réelle ayant abouti à l'élaboration d'une typologie.

Dans ces conditions, la meilleure approche est l'observation de la manière dont procèdent les chercheurs dans une situation concrète. Cette méthode présente de nombreuses difficultés sur lesquelles nous reviendrons. La principale nous paraît être qu'une part très importante du

travail d'analyse des données échappe à l'observation : il s'agit des opérations mentales, qui peuvent se traduire par des intuitions subites, des restructurations quasi-instantanées de la manière dont le chercheur perçoit et organise ses données. C'est pourquoi nous avons prévu, là où l'observation se révélait difficile, de recourir à des entretiens approfondis avec les chercheurs.

Cette approche fait évidemment la part trop belle à l'observation, et à l'introspection du chercheur. Faute d'avoir pu en imaginer une meilleure (l'expérimentation paraissant difficile à conduire à partir de données réelles), nous nous sommes résignés à partir d'informations assez pauvres et surtout fragmentaires, pour élaborer un modèle explicatif. Aussi, nous a-t-il semblé opportun de ne pas négliger les apports possibles de l'analyse bibliographique, les descriptions, même complaisantes et incomplètes, de la démarche suivie pouvant nous apporter des indications intéressantes à condition de les aborder avec un certain esprit critique.

En conséquence, nous avons choisi de procéder à l'analyse d'une trentaine d'exemples de construction de typologies "à la main" publiés dans la littérature spécialisée ou sous forme de rapport, et de faire suivre cette analyse bibliographique d'entretiens approdondis, accompagnés si possible d'observations, avec une dizaine de chercheurs réalisant ou ayant réalisé récemment une classification.

1.2. L'ANALYSE BIBLIOGRAPHIQUE.

Le choix des textes à analyser a été fait en recherchant au maximum la diversité, à la fois sur le plan de la théorie sous-jacente et sur celui du domaine d'application. Les principes qui nous ont guidés sont que les exemples de construction de typologie devaient :

- être empruntés pour un tiers, aux auteurs classiques (de Marx à Gurvitch ou Parsons), et pour deux tiers à la littérature courante, en y incluant si possible des applications industrielles ou commerciales ;

- concerner des problèmes variés, tels que la classification des types de société ou de culture, des courants religieux, des groupes de pression, des formes de criminalité, des phénomènes migratoires, des changements sociaux, etc. ;

- avoir comme unités de base des individus, des groupes constitués (familles, villes, organisations), des phénomènes sociaux (relations dans les groupes, déplacements), ou des objets au sens large (oeuvres d'art, outils, techniques) ;

- illustrer des procédures, des intentions et des arrière-plans théoriques aussi variés que possible.

L'analyse des textes sélectionnés a pour unique objectif la reconstitution de la procédure de classification utilisée. Les règles que nous avons suivies sont :

1) Si l'auteur donne une description de la méthode utilisée, et qu'il présente ses données (ou si celles-ci sont accessibles par un autre moyen), commencer par appliquer la méthode décrite aux données.

2) Si non, ou si les résultats de l'application de la méthode décrite ne correspondent pas à la typologie présentée, rechercher dans le texte des indices permettant de reconstituer la méthode réellement utilisée.

3) Elaborer un modèle de procédure susceptible de fonctionner réellement, incorporant toutes les informations recueillies sur la méthode réelle, et comblant les lacunes de ces informations par des hypothèses plausibles.

4) Appliquer cette procédure aux données et comparer à la typologie présentée.

Pour mener à bien la reconstitution de la méthode suivie, une première étape consiste à recourir à ce que A. BARTON appelle *substruction*. Très schématiquement, la procédure de substruction consiste à :

- déterminer les caractères effectivement présents dans la typologie présentée ;

- rattacher ces caractères à des dimensions (ou, au minimum, à la variable présence/absence du caractère) ;

- faire le produit cartésien des partitions définies sur l'ensemble des unités par chacune des dimensions identifiées ;

- comparer la partition définie par la typologie à celle obtenue par le produit cartésien, et expliciter les règles de passage de

la seconde à la première (*réduction de l'espace d'attributs* dans la terminologie de BARTON).

Pour qui a déjà réalisé des typologies "à la main", il est évident qu'une telle reconstitution ne reproduit pas la démarche réelle du chercheur ; c'est tout au plus un raccourci commode entre les données de départ de la recherche et l'aboutissement de celle-ci, et un moyen agréable d'exposition de la typologie. Dans les textes où l'auteur tente de donner un aperçu de sa démarche, c'est en réalité à ce mode d'exposition qu'il a recours ; il est extrêmement rare qu'un chercheur fasse état des hypothèses qu'il a abandonnées en cours de route, des impasses dans lesquelles il s'est engagé, des tâtonnements qui ont jalonné sa progression. C'est pourquoi nous n'avons jamais admis sans esprit critique les déclarations des auteurs sur leur méthode. Non que nous accusions les chercheurs d'avoir cherché à tromper le lecteur : d'une part, comme les entretiens que nous avons eus ultérieurement l'on confirmé, il est très difficile de décrire les opérations mentales mises en jeu dans une recherche ; d'autre part, il existe une tradition qui veut que dans la rédaction des comptes-rendus de recherche, on présente la procédure idéale, que l'on aurait dû appliquer (estime-t-on), plutôt que la démarche réellement suivie.

C'est pourquoi, nous nous sommes attachés, dans l'analyse de certains textes, à recenser ce qui pouvait constituer des indices du cheminement mental de l'auteur, sans nous satisfaire des seules explications "évidentes". A titre d'exemple, il est clair que Maurice HALBWACHS, dans son *Esquisse d'une psychologie des classes sociales* (1938), s'inspire très largement de la classification des groupes socio-professionnels de François SIMIAND. Faut-il considérer l'oeuvre de SIMIAND (en particulier son *Cours d'économie politique* (1931)) comme l'unique source de la typologie d'HALBWACHS ? Pour répondre à cette question, il est nécessaire de confronter les deux classifications, et d'analyser leurs différences ; et en particulier celles qui ont trait aux fondements théoriques. Mais une lecture attentive du texte de HALBWACHS, en mettant provisoirement entre parenthèses les considérations abstraites ou très générales, permet déjà de formuler quelques hypothèses à ce sujet. Par exemple, on remarque que parmi la vingtaine de types élémentaires que distingue l'auteur, cinq

sont décrits avec plus de détails et de "vérité" psychologique que les autres. Ce sont le paysan indépendant, possesseur légal de sa terre, ou le fermier, possesseur de fait ; l'employé de commerce ; l'artisan ou le petit commerçant ; le chef d'entreprise ; l'ouvrier de l'industrie. On peut penser que la position sociale de l'auteur lui conférait une certaine familiarité avec les quatre premiers types ; et que ses recherches (*la classe ouvrière et les niveaux de vie*, 1913 ; *L'évolution des besoins dans les classes ouvrières*, 1933) et ses fonctions (au comité mixte de la Société des Nations sur l'alimentation des travailleurs)l'ont amené à bien connaître le cinquième. D'où l'hypothèse qu'à l'origine de la typologie, outre les données théoriques déjà mentionnées, se trouve une connaissance concrète de quelques groupes sociaux, autour desquels s'organiseront les autres groupes. On remarque également dans le texte la mise en parallèle de ces principaux types, conduisant à dégager certaines différences dans leur mode de vie, et certaines oppositions liées à leur fonction sociale. Ces différences et oppositions deviennent des axes qui permettent d'opposer d'une part les milieux ruraux aux milieux urbains, d'autre part la bourgeoisie aux ouvriers d'industrie, etc. Adopter une telle approche dans l'étude d'un auteur classique est une pratique peu usitée dans les sciences sociales, car elle prend le texte à contre-courant ; il nous a paru qu'elle constituait un moyen privilégié pour comprendre la genèse d'une théorie (parallèlement à l'étude historique de la pensée de l'auteur).

1.3. L'ENQUÊTE AUPRÈS DES CHERCHEURS.

Cette phase constitue à la fois la partie la plus importante et la plus délicate de la recherche. Il s'agit en effet d'établir, sur chacun des exemples concrets sélectionnés, le "journal de bord" du chercheur ; journal relatant ses ojbectifs, ses hypothèses de départ, les procédures utilisées pour les matérialiser, les critères d'adoption ou de rejet des hypothèses, la part de l'intuition, le poids des connaissances, le rôle des *a priori* théoriques, et la dynamique d'ajustement des hypothèses aux données.

Nous souhaitions participer à titre d'observateurs à l'élaboration d'une typologie par une équipe de recherche. Ceci n'a pas été possible,

en raison de la perturbation qu'une telle observation aurait occasionnée dans le fonctionnement de l'équipe. Par contre, nous avons pu avoir des entretiens avec des chercheurs en train d'élaborer une typologie, ou venant d'en élaborer une. Dans quelques cas, nous avons même réalisé une reconstitution de la démarche suivie, le chercheur nous ayant donné libre accès à ses documents, et ayant accepté non seulement de nous décrire d'une manière très détaillée la façon dont il avait procédé, mais aussi de discuter avec nous, au cours de plusieurs entretiens ultérieurs, de nos essais de reconstitution.

Cette enquête auprès des chercheurs a posé deux types de problèmes. Le premier est lié à l'objectif même de l'enquête : il est inhabituel que des chercheurs en sciences sociales demandent à d'autres chercheurs en sciences sociales comment ils sont parvenus à tel résultat ; ou plus exactement leur demandent de décrire avec précision les hypothèses qu'ils ont dû abandonner, les impasses dans lesquelles ils se sont fourvoyés, éventuellement les erreurs qu'ils ont commises. Il existe en effet de fortes résistances à dévoiler ce que l'on considère un peu honteusement comme une "cuisine", par opposition à la voie royale de la recherche telle que la décrivent certains auteurs (en général non chercheurs eux-mêmes).

Le second type de problèmes tient à la difficulté d'analyser les processus mentaux. Les conditions optimales pour l'observation se trouvent réunies lorsque le chercheur a remplacé les unités qu'il désire classer par des objets qu'il manipule. Par exemple, lorsque les individus dont il élabore une typologie sont remplacés par des fiches sur lesquelles sont résumées leurs principales caractéristiques, et que le chercheur s'efforce de ranger ses fiches en tas représentant les types d'individus. Une telle situation est proche des expériences sur la formation de concept de HANFMANN et KASANIN, dérivées du test de ACH. Le rôle de l'observateur est alors de relever tout déplacement de fiche, même seulement esquissé, et d'en faire expliciter la raison par le chercheur ; la meilleure procédure, lorsqu'elle est acceptée par le chercheur, consiste à lui demander de verbaliser à haute voix ses manipulations, et à les enregistrer. Or, même dans ces conditions, si l'on peut comprendre les attentes du chercheur et ses jugements sur les résultats de ses opérations

mentales, on obtient difficilement la description de ces opérations mentales elles-mêmes. Cela tient aux limites bien connues du processus introspectif.

C'est pourquoi les données que nous avons recueillies dans nos essais de reconstitution, et à plus forte raison celles que nous ont apportées nos entretiens avec les chercheurs, ne nous ont pas permis d'élaborer une explication complète du processus aboutissant à une typologie. Nous ne nous y attendions d'ailleurs pas. Par contre, nous avons pu, à partir de ces données, dresser la liste des opérations les plus facilement automatisables, et indiquer leurs places respectives dans le processus.

1.4. LA COMPARAISON AVEC LES PROCÉDURES AUTOMATISÉES.

Parallèlement aux analyses bibliographiques et aux entretiens avec les chercheurs, nous avons passé en revue les principaux programmes de classification automatique et de segmentation. Cet aspect de la recherche n'a pas soulevé de difficulté particulière, les programmes recensés faisant tous l'objet de publications (cf. bibliographie en annexe). Il a par conséquent été facile de confronter notre formalisation des procédures de construction de typologies "à la main", avec les procédures automatisées décrites dans la littérature spécialisée.

2. *DESCRIPTION DE LA DÉMARCHE DE CONSTRUCTION DE TYPOLOGIES DANS LES SCIENCES SOCIALES*

Élaborer une typologie consiste à distinguer, au sein d'un ensemble d'unités (individus, groupes d'individus, faits sociaux, etc.), des groupes que l'on puisse considérer comme homogènes d'un certain point de vue. Le contenu de cette notion d'homogénéité varie selon les auteurs et les domaines d'application ; elle se fonde généralement sur une certaine ressemblance définie à partir d'un sous-ensemble des caractéristiques servant à décrire les unités étudiées. En outre, dans la majeure partie des cas que nous avons analysés, une typologie doit satisfaire à deux exigences supplémentaires :

1) elle doit être *exhaustive*, c'est-à-dire que toute unité étudiée doit pouvoir être affectée à au moins un groupe (l'union des groupes est égale à l'ensemble des unités) ;

2) les types doivent être mutuellement *exclusifs*, c'est-à-dire que toute unité étudiée ne peut être affectée qu'à un groupe au plus (l'intersection des groupes est vide).

Il faut noter que ces deux exigences sont indépendantes, ce qui permet de distinguer quatre variétés de typologies, selon qu'elles sont ou non exhaustives, et que les groupes sont ou non exclusifs les uns des autres.

Cette recherche nous a permis de recenser les considérations qui conduisent les spécialistes des sciences sociales à élaborer une typologie et de décrire les formes de démarches suivies pour cette élaboration.

2.1. Dans quels cas élabore-t-on une typologie ?

Au cours de nos analyses bibliographiques et de nos entretiens, nous avons relevé cinq motifs pour lesquels on recourt habituellement à

la construction d'une typologie. Ce sont :

1) *les exigences de l'application.* La plupart des recherches orientées vers l'application ont pour but de définir ou de spécifier les actions à entreprendre auprès d'individus ou de groupes sociaux. Il peut s'agir par exemple de campagne publicitaire, de propagande électorale, d'action éducative, de restructuration d'entreprises ou d'administrations, de modifications juridiques ou institutionnelles. Dans tous les cas, la recherche doit apporter des éléments de réponse aux questions : "que faire ?", et "auprès de qui ?" ou "en faveur de qui ?". Par conséquent, la recherche aboutit naturellement à classer les individus ou les groupes sociaux étudiés en fonction de leur sensibilité ou leur perméabilité aux modes d'action envisageables. Aussi peut-on dire que toute recherche appliquée débouche nécessairement sur une typologie.

2) *l'importance du volume des données à traiter.* Il arrive fréquemment que le volume des données à traiter et leur complexité rende leur analyse malaisée sans regroupement préalable. Il est commode de matérialiser l'ensemble des informations recueillies sous la forme d'une "matrice des données" (*data matrix*), c'est-à-dire d'un tableau rectangulaire comportant en lignes la liste des unités étudiées et en colonnes la liste des caractéristiques servant à les décrire ; à l'intersection d'une ligne i et d'une colonne j figure la $j^{ième}$ caractéristique de $i^{ième}$ unité (par exemple : la réponse à la question j de l'individu i). Dans la plupart des recherches, une telle matrice reste virtuelle : ses dimensions interdisent en effet toute manipulation directe par le chercheur. C'est pourquoi celui-ci s'efforce de réduire cette matrice en réduisant le nombre de caractéristiques à prendre en compte, et/ou en regroupant les unités étudiées en fonction de leurs ressemblances ; on peut parler dans ce deuxième cas de construction de typologie.

3) *l'impossibilité d'aboutir à un modèle unique.* Certaines recherches ont pour but de mettre au point un modèle explicatif permettant de rendre compte d'un phénomène social. C'est ainsi par exemple, que l'on s'interroge sur les causes de la délinquance juvénile, ou plus spécifiquement sur l'influence des croyances religieuses sur le comportement électoral. L'idéal du chercheur est d'élaborer une théorie générale du phéno-

mène étudié ; il n'y parvient pas toujours. En cas d'échec, il lui reste la ressource de segmenter l'ensemble des unités sur lesquelles portent ses observations, et de proposer un modèle explicatif spécifique pour chaque segment.

4) l'inefficacité du modèle explicatif général. Ce cas est en quelque sorte une variante du précédent. Lorsqu'il a été possible de construire un modèle explicatif valable pour l'ensemble des unités étudiées, il arrive que ce modèle soit d'une telle généralité que les lois dégagées ne présentent aucun intérêt pratique. Le chercheur peut alors spécifier le modèle, le particulariser, en adjoignant des lois qui ne valent que pour une partie des unités étudiées. On aboutit ainsi à une famille de modèles auxquels correspond un ensemble de domaines d'application. Chacun de ces domaines est constitué par un sous-ensemble d'unités, et l'ensemble de ces sous-ensembles forme une typologie.

5) la dynamique interne du système étudié, qui impose de penser en termes de typologie. Dans l'étude d'un phénomène évolutif, on est amené le plus souvent à identifier des forces antagonistes, dont les interactions permettent d'expliquer la dynamique du système ; ce peuvent être des intérêts opposés, des aspirations contradictoires, des objectifs incompatibles, etc. L'analyse des coalitions et des conflits conduit naturellement le chercheur à classer les unités qu'il étudie (individus ou groupes sociaux) en groupes distincts, s'opposant actuellement ou susceptibles de s'opposer. Comme dans le premier cas (exigences de l'application), le recours à l'outil typologique est ici une nécessité.

2.2. *Comment procède-t-on pour obtenir des types ?*

Au cours de l'analyse bibliographique, nous avons pu distinguer trois démarches assez distinctes pour la construction de typologies. Ce sont :

- la situation des unités étudiées par rapport à un ensemble de types abstraits ("types-idéaux") ;

- la structuration de l'univers étudié à partir des dimensions servant à décrire les unités (réduction de l'"espace d'attributs") ;

- le regroupement des unités autour d'un petit nombre d'entre elles choisies comme noyaux de la typologie (agrégation des unités).

Les interviews et les séances d'observation auxquels nous avons procédé ensuite n'ont pas fait apparaître d'autre forme de démarche. Par contre, ils ont montré que ces trois formes ne s'excluaient pas nécessairement, mais qu'au contraire elles pouvaient intervenir successivement au cours de l'élaboration d'une typologie, et ce parfois à plusieurs reprises. Aussi pensons-nous que les descriptions fournies par les auteurs dans leurs publication ne reflètent que la forme dominante de leur démarche réelle, ou bien celle qu'ils ont jugé la plus exemplaire parce que la plus consciente de leur part. Pour des raisons de clarté d'exposition, nous avons cependant cru bon de conserver la distinction entre ces trois formes de démarche.

2.2.1. La constitution de types idéaux.

Le terme de "type-idéal" (*Idealtypus*) est emprunté à Max WEBER. Celui-ci a donné plusieurs exposés théoriques de cette notion, et l'a utilisée dans ces travaux. C'est la plus abstraite des procédures de construction de typologie, et par conséquent la plus difficile à formaliser. Nous rappelons tout d'abord la démarche suivie pour constituer des types-idéaux ; nous analysons ensuite les hypothèses implicites qui nous paraissent la sous-tendre.

a) Description de la procédure. L'objectif de cette procédure est de construire des cas "typiques", c'est-à-dire des notions abstraites permettant de rendre compte des phénomènes réels. Ceux-ci sont multiformes : "le même évènement historique peut par exemple avoir par un de ses aspects une structure 'féodale', par un autre 'patrimoniale', par d'autres 'bureaucratique' et par d'autres encore 'charismatique'". (Max WEBER, *Economie et Société*, Paris, Plon, 1971, p_p 17-18). Pour que chacun de ces concepts ait un sens univoque, on est obligé d'élaborer "des types ("*idéaux*") "purs" de chacune de ces sortes de structures qui révèlent alors chacune pour soi l'unité cohérente d'une adéquation *significative* aussi complète que possible, mais qui, pour cette raison, ne se présentent peut-être pas davantage dans la réalité sous cette forme *pure*

absolument idéale, qu'une réaction physique que l'on considère sous l'hypothèse d'un espace absolument vide" (ibid., p. 18)."Par son contenu, cette construction a le caractère d'une utopie que l'on obtient en accentuant *par la pensée* des éléments déterminés de la réalité" (Max WEBER, *Essais sur la théorie de la science*, Paris, Plon, 1965, p. 180). L'ensemble des types-idéaux constitue un modèle abstrait de la réalité étudiée. C'est un schéma d'interprétation (un "tableau de pensée") qui fournit des objets de comparaison *extérieurs* à la réalité, et par rapport auxquels on peut situer les objets réels dans l'univers des possibles défini par les dimensions.

"On obtient un idéaltype *en accentuant* unilatéralement *un ou plusieurs* points de vue et en enchaînant une multitude de phénomènes, donnés *isolément*,diffus et discrets, que l'on trouve tantôt en grand nombre, tantôt en petit nombre et par endroits pas du tout, qu'on ordonne selon les précédents points de vue choisis unilatéralement, pour former un *tableau de pensée* homogène (einheitlich)" (Max WEBER, ibid, p. 181). En d'autres termes, on dispose d'un ensemble d'unités à classer, chaque unité étant décrite par sa position sur plusieurs dimensions. Ces dimensions sont en nombre élevé. Dans l'optique du chercheur, toutes les dimensions n'ont pas la même importance théorique. Certaines d'entre elles occupent dans la théorie une place privilégiée, en ce sens qu'elles se trouvent chacune au centre d'un réseau de relations de type causal entre dimensions ; elles permettent de rendre compte, dans la théorie du chercheur, de la quasi-totalité des dimensions utilisées pour la description des unités. Pour ces dimensions importantes, ou pour certaines d'entre elles seulement, on imagine une unité se situant à une position extrême : maximum théorique ($\geqslant$ maximum observé), correspondant à la possession d'une caractéristique au plus haut degré concevable, ou minimum théorique ($\leqslant$ minimum observé), correspondant à l'absence totale d'un caractère. En pratique, aucune unité réelle (observable) ne se situe à une position extrême pour toutes les dimensions importantes.

Pour que cette unité fictive constitue un type-idéal, il faut en outre qu'elle possède à la fois une cohérence interne et une cohérence externe. La cohérence interne se traduit par l'existence d'un modèle expliquant la conjonction, dans une même unité, de l'ensemble des caracté-

ristiques du type, et la présentant à la fois comme théoriquement possible et logiquement nécessaire. La cohérence externe se manifeste par l'intégration de l'ensemble des types idéaux dans une théorie globale exprimant de manière synthétique les relations postulées entre ces unités abstraites.

Cette procédure de construction est donc très liée à la théorie du chercheur ; et l'on admet que deux chercheurs différents, visant à rendre compte des mêmes phénomènes, puissent légitimement aboutir à des types-idéaux différents.

Un exemple simple (fictif) illustre la démarche suivie pour dégager des types-idéaux. Supposons une population d'élèves auxquels on a affecté une note de 0 à 20 pour un ensemble de matières d'enseignement général. Selon ce qu'on recherche en analysant ces données, on peut concevoir le type-idéal de l'excellent élève (qui obtient 20/20 dans toutes les matières) et le type-idéal du cancre (qui obtient 0/20 dans toutes les matières) ; ou bien le type-idéal du littéraire pur (qui obtient 20/20 en français, langues vivantes, langues mortes, et 0/20 en mathématiques, physique, chimie), et celui du scientifique pur (qui obtient 20/20 là où le littéraire pur obtient 0/20, et réciproquement). On peut imaginer bien d'autres types-idéaux sur ces données. Pour que ces ensembles de types-idéaux constituent une typologie, il faut en outre que leurs définitions soient fondées en théorie : en recourant par exemple ici aux notions d'intelligence générale, de bonne ou de mauvaise adaptation aux exigences scolaires, etc., pour la première typologie ; à la notion d'aptitudes différentielles et complémentaires pour la seconde.

b) Analyse des hypothèses implicites. Quel est le rapport entre les données et la typologie qui doit en rendre compte ? Représentons chaque unité par un point dans un hyperespace à n dimensions (s'il y a n dimensions empiriques servant à décrire les unités). Sur chaque dimension, considérons deux points : celui correspondant au maximum théorique, et celui correspondant au minimum théorique. Ces valeurs théoriques sont déterminées par le chercheur de manière à être compatibles avec les maxima et les minima observés. Ces $2n$ points situés sur les

dimensions constituent les coordonnées de 2^n points de l'hyperespace ; ces derniers sont les angles de l'enveloppe de l'univers des possibles dans l'hyperespace (il s'agit des possibles selon la procédure de description choisie). C'est au sein de l'ensemble de ces 2^n points "théoriques" que seront choisis les types-idéaux.

Si l'on s'en tient à la description de WEBER, la sélection des types-idéaux dans cet ensemble de points obéit à de pures considérations théoriques. En principe, ni la densité des points réels figurant les unités, ni même la présence ou l'absence d'unités dans une zone de l'hyperespace, ne jouent un rôle dans cette sélection, puisque l'on tient compte de phénomènes "que l'on trouve tantôt en grand nombre, tantôt en petit nombre et par endroits pas du tout". En fait, le modèle logique qui décrit un type-idéal s'appuie sur les relations postulées par la théorie entre les dimensions servant à décrire les unités. Si ces relations ne sont pas en contradiction avec les observations, elles doivent se traduire par des corrélations entre les dimensions, c'est-à-dire par des densités de points réels plus grandes en certaines zones de l'hyperespace. C'est pourquoi, si l'on veut que la théorie justifiant la sélection soit cohérente et compatible avec les observations, c'est-à-dire avec la répartition des points réels dans l'hyperespace, on est dans l'obligation de sélectionner des points théoriques qui ne soient pas trop éloignés d'agglomérats de points réels, puisque ces agglomérats sont l'indice des relations entre les dimensions. En outre, le repérage des points réels à partir des points théoriques ainsi sélectionnés en sera facilité.

Il semble donc que, dans l'ensemble des points théoriques, seul un sous-ensemble assez restreint contienne tous les points susceptibles d'être sélectionnés comme types-idéaux. En outre, ce sous-ensemble est déjà structuré par les corrélations observables entre les dimensions : parmi tous les couples, triplets, ..., m-uplets possibles à partir de ce sous-ensemble, seul un très petit nombre satisfait à la condition de résumer de manière économique (avec le minimum de redondance) le maximum de relations entre les dimensions (corrélations). Cette condition est l'équivalent, au niveau des données, de l'exigence de cohérence au ni-

veau théorique. Le nombre de couples, triplets, ..., *m*-uplets satisfaisant à la condition ci-dessus est le nombre de "typologies-idéales" compatibles avec les données. Si le chercheur recherche la compatibilité avec les données, la typologie à laquelle il doit logiquement aboutir est l'une des typologies ainsi déterminées. La part de la théorie dans ce processus se révèle être beaucoup plus réduite qu'il ne paraît de prime abord. Il semble que ce soit au niveau de la sélection des dimensions importantes, qui a pour effet de contracter l'espace de description des données, que le rôle des considérations théoriques soit déterminant : en effet, cette sélection des dimensions réduit considérablement l'ensemble des typologies possibles au sein duquel se fera le choix de la typologie. Si les dimensions retenues ne permettent d'aboutir à aucune des typologies pratiquement possibles, le chercheur ne peut que renoncer à la construction de types-idéaux, ou bien renoncer à une typologie compatible avec ses données, ou encore renoncer à certaines thèses de sa théorie. Dans ce dernier cas, il lui faut procéder à un ajustement, par essais et erreurs, de sa théorie à ses observations.

Si l'on accepte cette représentation de la "typologie-idéale", représentation considérablement éloignée de la description qu'en donne Max WEBER, on constate que cette procédure se fonde sur les hypothèses implicites suivantes :

1) les objets réels (unités) sont décrits à partir de dimensions orientées : variables nominales dichotomiques susceptibles d'une bipolarisation, variables ordinales, variables mesurables. Ceci exclut les variables nominales à plus de deux états, qui ne sont pas à proprement parler des dimensions.

2) ces dimensions sont susceptibles d'être bornées, soit qu'il existe des bornes "naturelles" (zéro absolu), soit que la théorie permette de leur fixer un minimum et un maximum, soit que l'on puisse légitimement prendre en considération le minimum et le maximum observables.

3) ces dimensions sont au moins partiellement liées entre elles dans la réalité : les données mettent en évidence l'existence de corrélations constituant des constellations de dimensions (ces corrélations sont

évidemment dues à une répartition non homogène des points représentant les objets dans l'espace des descripteurs).

4) les dimensions sont au moins partiellement liées entre elles dans la théorie : l'explication proposée par le chercheur établit des relations de type causal entre les présences/absences et/ou les degrés d'intensité conjoints de plusieurs dimensions.

5) il y a un certain parallélisme (homomorphisme) entre les corrélations empiriques et les relations théoriques. Cela revient à dire que la théorie est compatible avec les données, qu'elle rend compte des observations de façon satisfaisante.

6) on observe un effet de complémentarité (de compensation) entre certaines dimensions : un objet qui obtient un rang extrême (le plus élevé ou le plus bas) sur une dimension n'obtient pas également un rang extrême sur toutes les autres dimensions (les angles de l'enveloppe des possibles dans l'espace des descripteurs sont vides). En conséquence, les corrélations entre les dimensions ne se rapprochent en général que très rarement de la valeur absolue maxima : les dimensions ne peuvent pas être réduites à une seule d'entre elles sans que l'on perde beaucoup d'information.

7) on appelle "type-idéal" un objet abstrait construit par le passage à la limite sur l'ensemble des dimensions décrivant les objets réels, et susceptible de résumer certaines des corrélations observées entre les dimensions.

8) l'ensemble des "types-idéaux" permet de résumer l'ensemble des corrélations entre dimensions observées sur les données.

9) les objets réels peuvent être décrits de façon condensée par leurs distances aux types-idéaux. En pratique, cela signifie que les unités observées sont présentées par la théorie comme des composés, dans des proportions variables, des différents "corps purs" que sont les types-idéaux.

2.2.2. La réduction d'un espace d'attributs.

Cette seconde procédure de construction de typologie s'applique principalement lorsque le chercheur s'intéresse à un domaine encore peu

exploré systématiquement, ou lorsqu'il aborde un domaine déjà connu avec une problématique nouvelle.

a) Description de la procédure. La démarche de construction de typologie se décompose en deux phases nettement distinctes. La première est une phase d'analyse des concepts de base en leurs dimensions, afin d'élaborer un cadre de description des unités étudiées ("espace d'attributs"). La seconde est une phase de réduction de l'espace ainsi défini à un petit nombre de dimensions, et de modalités sur ces dimensions, aboutissant à la typologie désirée.

L'objectif de la première phase est d'apporter au chercheur un ensemble de repères aussi nombreux et aussi précis que possible pour décrire les unités qu'il se propose d'étudier. Certes, toute recherche comporte l'élaboration d'un cadre descriptif ; en particulier, dans les recherches sociologiques dites "empiriques", ce cadre descriptif sous-tend au moins en partie la formulation des questions qui seront posées au cours de l'enquête sur le terrain. Mais, dans la plupart des recherches, ce cadre doit avoir un caractère opérationnel : il doit être maniable, c'est-à-dire facile à exposer dans un compte-rendu, et aisément reliable aux questions du questionnaire (problème des relations entre indicateurs et dimensions). Dans la procédure que nous décrivons ici, loin de viser au départ à la parcimonie, à l'économie des moyens qu'impose un cadre opérationnel, le chercheur prétend à l'exhaustivité. Son but premier est d'obtenir le maximum de nuances et de distinctions, fût-ce au prix d'une certaine redondance, ou d'un manque apparent de cohérence. Le produit de cette première phase doit être le catalogue de l'ensemble des caractéristiques possibles des unités.

Evidemment, prétendre réellement à l'exhaustivité serait illusoire. En pratique, cette prétention se traduit par l'absence totale d'exclusive : toutes les dimensions, toutes les oppositions relatives aux unités étudiées sont admises pourvu qu'elles aient un sens. Peu importe la théorie à laquelle elles se rattachent ; tout est admis. La sélection se fera dans la deuxième phase.

La recherche des dimensions descriptives emprunte plusieurs voies : relevé des distinctions et des classifications proposées par les travaux antérieurs, extraction des distinctions implicites recélées par certaines théories, décomposition des notions de base en notions plus élémentaires,

recours à l'expérience du terrain et à la sensibilité propre du chercheur. La dernière voie est difficile à formaliser ; elle s'apparente aux effets de l'intuition. Les trois autres ressortissent plus ou moins à l'analyse sémantique ou à l'analyse de contenu.

Au niveau le moins élaboré, le survol de la littérature sur le problème étudié par le chercheur fournit à celui-ci des descriptions systématiques, dans lesquelles il lui suffit de relever les dimensions et les catégories que l'auteur distingue sur ces dimensions. Une lecture attentive des textes théoriques permet de déceler également des amorces de dimensions descriptives : par exemple, un qualificatif accolé à une catégorie d'unités est l'indice d'une distinction implicite, que l'auteur n'a pas jugé utile de développer ici. Le chercheur peut s'efforcer d'expliciter cette distinction, en recherchant systématiquement les contraires ou les divers degrés d'intensité des qualificatifs retenus. Au niveau le plus abstrait se situe l'analyse des concepts-clés pour en dégager les dimensions latentes. Elle consiste en une analyse sémantique poussée des principales notions utilisées par les théories relatives au domaine étudié (par exemple, notions d'anomie, d'intégration, etc.), et de leurs connotations ; cette analyse conduit à dégager un ensemble de significations différentes quoique voisines qui constitueront autant de dimensions descriptives. A cet exercice s'apparentent le décalque de distinctions propres à une autre discipline, et l'application au problème étudié de catégories à prétention universelle, empruntées à l'arsenal des philosophies ou des autres sciences ; comme par exemple les antinomies : agir/subir, essence/existence, manifeste/latent, permanent/transitoire, etc. Une telle démarche, d'inspiration scolastique, peut avoir une fonction heuristique dans l'analyse des concepts, et conduire à découvrir des dimensions nouvelles.

Cette première phase se termine généralement par une mise en ordre des dimensions ainsi identifiées. La procédure suivie est assez proche de celle de l'analyse de contenu : de même que l'on regroupe sous une même rubrique des énoncés différents se rapportant à un même thème, le chercheur rassemble dans une même catégorie des dimensions dont les significations lui semblent apparentées. Il s'efforce de donner une définition de chaque catégorie ainsi obtenue. S'il réitère cette opération sur les catégories, il obtient une classification à plusieurs niveaux, les niveaux les plus élevés correspondant aux concepts les plus généraux (et par conséquent les plus abstraits).

La seconde phase n'est pas présente dans toutes les recherches. Il arrive en effet que le chercheur borne son effort à cette structuration des dimensions descriptives, et publie le catalogue raisonné des dimensions et de leurs modalités. Mais dans ce cas, le chercheur lui-même présente ses travaux comme une recherche préliminaire, inachevée, comme un débrouissaillage du domaine étudié, préalable à la recherche proprement dite. Cette seconde phase (réduction) est donc une suite naturelle, nécessaire, de la phase d'analyse des dimensions.

En effet, l'ensemble des dimensions relevées est pratiquement inutilisable tel quel. Cet ensemble de dimensions constitue les coordonnées de ce que BARTON appelle un "espace d'attribut" (*property space*), univers que l'on peut matérialiser en énumérant toutes les combinaisons de propriétés possibles pour une unité donnée , dans l'hypothèse où les dimensions sont indépendantes. Un exemple très simple en est fourni par Louis GUTTMAN (*"A structural theory for intergroup beliefs and actions"*, American Sociological Review, 24 (1959), 318-328), avec il est vrai une terminologie différente de celle de BARTON. Reprenant une recherche sur les systèmes de valeurs et les comportements relatifs aux relations entre groupes ethniques, GUTTMAN considère trois dimensions dichotomiques : croyances/action, individu/groupe, comparaison/interaction. Ces trois dimensions déterminent $2^3 = 8$ combinaisons de propriétés, qui constituent l'espace d'attributs des conduites interraciales :

croyance - groupe - comparaison :	stéréotype ("je crois mon groupe supérieur aux autres groupes ethniques")
croyance - groupe - interaction :	norme ("je crois que mon groupe devrait fréquenter les autres groupes")
croyance - individu - comparaison :	sentiment de supériorité ("je crois être supérieur aux membres des autres groupes")
croyance - individu - interaction :	interaction hypothétique ("je crois que, dans telle situation, je fréquenterais les membres des autres groupes")
action - groupe - comparaison :	éducation ("je stimule mon groupe ethnique en le comparant aux autres groupes")
action - groupe - interaction :	prédication ("j'incite mon groupe à fréquenter les autres groupes")
action - individu - comparaison :	comportement de supériorité ("je me stimule en me comparant aux membres des autres groupes")

action - individu - interaction : interaction personnelle ("dans telle situation, j'ai des relations avec les membres d'autres groupes ethniques").

Evidemment, l'exemple d'espace d'attributs emprunté à GUTTMAN est suffisamment réduit pour pouvoir être aisément manipulé. En pratique, lorsque la première phase conduit à retenir de dix à vingt dimensions ayant de deux à cinq états chacune, on obtient des espaces d'attributs comptant de 2^{10} = 1024 combinaisons de propriétés possibles à 5^{20} = environ 95 370 milliards de combinaisons. Il est hors de question de matérialiser un tel espace en énumérant ses combinaisons de propriétés ; il faut donc le réduire.

La réduction de l'espace d'attributs peut se dérouler soit sur le plan sémantique, soit sur le plan statistique. La première forme de réduction se fonde sur l'analyse des significations attachées aux dimensions et/ou aux caractéristiques définies sur ces dimensions ; elle implique par conséquent une référence constante aux théories sous-jacentes. L'analyse sémantique peut par exemple amener à constater que deux dimensions ou deux caractéristiques différentes traduisent en fait la même distinction ; ou bien qu'une distinction se trouve logiquement impliquée par une autre distinction, plus fondamentale ; ou encore qu'il y a incompatibilité entre une caractéristique définie sur une dimension, et une autre caractéristique définie sur une autre dimension. L'opération de réduction consiste alors à éliminer les dimensions ou les caractéristiques redondantes (dans les deux premiers cas), ou à combiner en une seule les dimensions présentant certaines incompatibilités, afin d'éliminer les combinaisons d'attributs dénuées de signification (dans le dernier cas).

La réduction sur le plan statistique présuppose la disponibilité pour le chercheur d'un ensemble d'unités décrites à l'aide de certaines des dimensions relevées (sinon toutes, ce qui est rare). Pour que l'analyse statistique ait un sens, il faut également que le nombre d'unités soit suffisamment élevé, et que l'analyse ne prenne guère en compte simultanément plus de quatre variables (ou dimensions). A titre d'ordre de grandeur, il est souhaitable que le nombre d'unités disponibles soit égal à dix fois le nombre de combinaisons d'attributs possibles pour les dimensions retenues dans l'analyse ; en effet, un ensemble d'unités trop petit produit une redondance qui n'est en réalité qu'un artefact. Lorsque l'analyse statistique révèle par exemple une forte corrélation entre deux variables, c'est-

à-dire lorsque les partitions définies par celles-ci sur l'ensemble des unités sont pratiquement confondues, la réduction se ramène à l'élimination de l'une des deux variables. Lorsque l'on constate qu'une case de l'espace d'attributs est toujours vide, c'est-à-dire qu'aucune des unités observées ne possède simultanément deux caractéristiques correspondant à deux dimensions différentes, la réduction conduit à recodifier les deux dimensions en une seule variable.

La réduction sémantique et la réduction statistique présentent de nombreux points communs. En pratique, il est rare que le chercheur se borne à l'une d'entre elles ; plus généralement, la réduction de l'espace d'attributs s'effectue simultanément sur les deux plans, à condition que le chercheur dispose de données d'observation. En effet, ces deux modes de réduction s'étayent l'un l'autre. Il est naturel par exemple, lorsque l'on constate qu'aucune unité observée ne correspond à une position possible de l'espace d'attributs, de vérifier si cette absence peut correspondre à une impossibilité déductible d'une théorie. De même, il est prudent de ne procéder à l'élimination d'une dimension apparemment redondante qu'après avoir constaté qu'elle était effectivement en forte corrélation avec la dimension conservée.

Ces opérations de réduction ont pour effet de diminuer considérablement le nombre de configurations de caractéristiques possibles ; elles ne suffisent cependant pas à rendre l'espace d'attributs aisément manipulable. En effet, dans toutes les recherches que nous avons analysées, le nombre de types souhaité se situait entre trois et dix (avec éventuellement quelques sous-types additionnels). Or, les opérations de réduction que nous avons décrites permettent rarement d'aboutir à un nombre de types aussi restreint. Force est donc au chercheur d'éliminer les distinctions qui se révèlent pourtant assurées tant sur le plan statistique que sur le plan sémantique. Les règles de sélection des dimensions et/ou des caractéristiques découlent alors de la problématique du chercheur, c'est-à-dire des théories auxquelles il se réfère et des buts qu'il poursuit. En fait, ces règles se traduisent par une hiérarchie des distinctions précédemment établies, selon leur pertinence par rapport au problème posé. Seules seront retenues les plus pertinentes d'entre elles, et ce à condition que le nombre de types engendrés par ces distinctions n'excède pas la dizaine. La typologie proposée apparaît alors clairement comme l'une des typologies

possibles, et le lecteur a théoriquement toujours la possibilité, s'il n'en est pas satisfait, de procéder à sa propre sélection parmi les dimensions recensées auparavant.

b) Analyse des hypothèses implicites. Cette procédure fait succéder à une phase d'expansion maxima de l'espace d'attributs une phase de contraction maxima. Formellement, la première phase n'appelle pas de remarque particulière. Son résultat est un hyperespace comptant un nombre très élevé de dimensions. Quelle que soit leur nature, ces dimensions sont pratiquement considérées comme des variables nominales. Les variables continues se trouvent alors découpées en zones regroupant des valeurs voisines mais distinctes à l'origine, chaque zone recevant une dénomination propre. Ainsi, une variable métrique comme l'éventail des revenus peut être découpée en catégories telles que : revenus très élevés, élevés, ..., très faibles ; une variable ordinale comme une opinion, en : très favorable, favorable, ..., très opposé . Certes, ces zones sont ordonnées, non permutables ; mais cette propriété n'est pas utilisée dans la seconde phase, et la dimension est traitée comme une variable nominale. Cette réduction systématique du quantitatif ou de l'ordinal au qualitatif entraîne une perte de précision dans la description des objets étudiés ; elle est rendue nécessaire par la procédure utilisée dans la seconde phase, pour aboutir à une classification, c'est-à-dire à une variable nominale. Une telle réduction des dimensions continues impose évidemment au chercheur des décisions relatives au nombre de zones, et aux positions des limites de zones ; ces décisions ne peuvent pas ne pas avoir d'implication sur le plan de la théorie, même si les coupures retenues apparaissent comme "naturelles". Chaque dimension étant divisée en zones, l'hyperespace est ainsi partitionné en cases, sorte d'hypercubes servant à ranger selon leurs caractéristiques les objets à décrire. Chaque case est localisée par autant de propriétés que l'hyperespace compte de dimensions ; ces propriétés sont les attributs décrivant les objets rangés dans la case.

La phase de réduction de cet hyperespace commence par la contraction de cet hyperespace, c'est-à-dire par la suppression des cases vides. Une case est réputée vide lorsque l'on considère qu'aucun objet réel ne peut s'y trouver. Pour cela, il faut bien entendu qu'aucun objet ne s'y trouve effectivement lorsque l'on dispose de données d'observation ; mais il faut également que l'analyse des significations attachées aux propriétés

et/ou aux dimensions montre qu'il y a contradiction entre deux au moins d'entre elles, et qu'il est par conséquent impossible qu'un objet réel possède simultanément ces attributs contradictoires. Cela se produit en particulier lorsqu'à un état d'une variable ne peut correspondre qu'une partie seulement des états d'une autre variable (voire un seul), ceci étant vrai simultanément pour les deux variables considérées ; il y a alors redondance (ou corrélation) entre les variables.

A ce point de l'analyse de la procédure, dire qu'une case de l'hyperespace est nécessairement vide, c'est dire que, parmi l'ensemble des caractéristiques servant à décrire les objets, deux au moins sont incompatibles. Cela implique que toutes les cases définies par la présence simultanée de ces caractéristiques sont également vides. Si l'une des caractéristiques est un état d'une variable à m états, et l'autre un état d'une variable à n états, poser leur incompatibilité revient à réduire le nombre de cases de l'hyperespace de $\frac{1}{m.n}$.

Pour décrire formellement la contraction progressive de l'hyperespace, il est commode de négliger provisoirement les dimensions, et de prendre comme point de départ les seuls attributs. Cette présentation n'est que la généralisation de la procédure effectivement suivie ; elle présente l'avantage de mettre en évidence les hypothèses implicites que suppose l'existence de dimensions.

Considérons l'ensemble A des n attributs retenus à l'issue de la première phase pour décrire les objets étudiés, indépendamment des dimensions auxquelles ces attributs se rattachent. Pour chaque objet décrit et pour chaque attribut, l'objet le possède ou ne le possède pas ; chaque attribut prend nécessairement l'une des deux modalités {absence, présence}. On peut considérer l'hyperespace de description comme l'ensemble des parties de A, ou si l'on préfère comme le produit cartésien de n ensembles {absence, présence}, ou plus brièvement $E = \{0, 1\}^n$. La taille de cet hyperespace, $|E| = 2^n$ cases, en interdit pratiquement la manipulation.

Lorsque le chercheur pose qu'un sous-ensemble $A' \subset A$ de m attributs constitue l'ensemble des m valeurs possibles d'une même dimension, il pose les deux affirmations suivantes :

- pour un objet donné, deux attributs de A' quelconques mais dis-

tincts, ne peuvent être vrais simultanément ;

- pour un objet donné, tous les attributs de A' ne peuvent être faux simultanément. Par conséquent, à tout objet appartenant à l'ensemble des objets à décrire, on peut toujours faire correspondre un attribut $a_i \in A'$, et un seulement. Cela revient à affirmer *a priori* que doit nécessairement être vide toute case de l'hyperespace correspondant soit à la présence simultanée de deux attributs ou plus de A', soit à l'absence simultanée de tous les attributs de A'. Affirmer que les m attributs constituent une dimension, c'est réduire le sous-espace $E' \subset E$ ayant $|E'| = 2^m$ positions (cases) à un sous-espace n'ayant que m positions. Les m attributs à deux valeurs {absence, présence} sont remplacés par une dimension à m valeurs : {attribut 1, attribut 2, ..., attribut m}.

La procédure de contraction de l'hyperespace se fonde sur une démarche du même type : élimination des cases nécessairement vides par combinaison de dimensions, ou plus exactement d'échelles nominales. Le raisonnement est le suivant. Soient deux sous-ensembles d'attributs, $Y \subset A$ et $Z \subset A$, distincts ($Y \cap Z = \emptyset$), et leur produit cartésien $C = Y \times Z$. L'ensemble produit C est lui-même une échelle nominale, dont les attributs sont les couples $(y_i \ \& \ z_j)$ correspondant à la présence simultanée chez un même objet d'une caractéristique $y_i \in Y$ et d'une caractéristique $z_j \in Z$. Lorsque l'un de ces couples est logiquement impossible (c'est-à-dire lorsque y_i et z_j sont incompatibles), on l'élimine ; on crée ainsi une nouvelle échelle nominale $C' \subset C$, qui remplace les deux échelles Y et Z. Cette réduction du produit cartésien C entraîne évidemment une réduction proportionnelle de l'hyperespace.

Lorsque cette phase de contraction est achevée, si le nombre de cases de l'espace d'attributs résultant est plus élevé que le nombre de types souhaité, il faut encore réduire cet espace. Cette réduction fait appel à une logique différente de celle de la phase de contraction, puisqu'il ne reste plus aucune case que l'on puisse *a priori* réputer vide. La procédure de réduction va alors consister à regrouper en une même case des cases contigües dans l'hyperespace, en renonçant à la différenciation sémantique entre les attributs qui permettent de les distinguer. Lorsque l'on dispose de données empiriques, deux critères sont envisageables pour décider du regroupement de cases contigües : un critère statistique et un critère théorique. Le *critère statistique* consiste à rechercher la compacité

maxima de l'espace d'attributs, c'est-à-dire la répartition homogène des objets dans l'espace typologique résultant. Lorsque les cases de l'espace typologique sont d'effectifs comparables, la dispersion (variété, entropie) de la distribution est maxima, et le pouvoir explicatif (dans le cas de mise en relation avec d'autres variables) maximum. Le *critère théorique* implique une hiérarchie sur les distinctions sémantiques différenciant les objets appartenant à deux cases voisines. Cette hiérarchie se fonde sur l'importance de ces distinctions au regard des théories admises par le chercheur et/ou sur leur utilité relativement au but de la recherche. Ces deux critères peuvent intervenir dans une même phase de réduction, mais il n'est pas aisé alors de faire le départ entre leurs rôles respectifs.

Cette description plus formelle de la procédure de contraction/réduction d'un espace d'attributs permet d'expliciter les hypothèses suivantes qui la sous-tendent :

1) Les objets étudiés peuvent être décrits à partir de leur appartenance ou leur non-appartenance à des catégories.

2) Ces catégories sont définies en termes de qualités, propriétés, attributs, caractéristiques, que peuvent posséder ou ne pas posséder les objets, ou également d'état d'une variable ou d'une dimension. En pratique, tous ces mots sont synonymes.

3) Un même objet peut posséder (être caractérisé par) plusieurs attributs simultanément.

4) Les significations attachées aux attributs définissent entre eux des relations de compatibilité ou d'incompatibilité. Sont incompatibles deux attributs qui ne peuvent pas être attachés simultanément à un même objet.

5) Deux objets qui possèdent exactement les mêmes attributs et eux seulement sont identiques.

6) Deux objets sont distincts lorsqu'il y a au moins un attribut qu'ils ne possèdent pas en commun.

7) Lorsqu'il existe des relations d'incompatibilité entre attributs,

il est possible de redéfinir les liens sémantiques entre les attributs ou les sous-ensembles d'attributs de manière à éliminer les incompatibilités. Cette opération, appelée *contraction de l'espace d'attributs*, consiste à partir du produit cartésien des bi-partitions définies par les catégories sur l'ensemble des objets étudiés, et à éliminer toutes les parties correspondant à la possession simultanée d'attributs incompatibles. Cette contraction n'entraîne aucune perte d'information.

6) Lorsqu'il n'existe plus d'incompatibilités, il est possible de négliger certaines distinctions sémantiques jugées secondaires, pour réduire l'espace d'attributs contracté à un espace typologique plus maniable parce que plus restreint. Cette réduction entraîne nécessairement une perte d'information.

9) La partition définie sur les objets par l'espace d'attributs d'origine étant une sous-partition de la partition définie par la typologie, si l'espace d'attributs a un sens, la typologie qui en est tirée a également un sens.

10) Les objets réels peuvent être décrits de façon condensée par leur appartenance à l'une des catégories définies par l'espace typologique résultant de la réduction.

2.2.3. L'agrégation autour d'unités-noyaux

Cette troisième procédure de construction de typologie est à la fois la plus fréquemment utilisée dans les recherches empiriques, et la plus rarement décrite dans les textes de méthodologie. Elle est pratiquement la seule procédure manuelle utilisée lorsque l'on dispose d'informations particulièrement riches et structurées, concernant un nombre assez petit d'unités : monographies de communautés, biographies, entretiens individuels en profondeur.

a) Description de la procédure. Il est possible de distinguer deux phases : une phase de condensation des informations relatives à chaque unité étudiée, et une phase d'agrégation des unités au vu de ces informations condensées. Mais, dans le cas le plus couramment rencontré, celui de typologies construites à partir d'entretiens individuels, la phase préalable de

recueil des informations joue également un rôle important dans le processus décrit.

Le recueil des informations est en effet souvent réalisé au moins partiellement par le chercheur lui-même, directement sur le terrain. S'il s'agit d'une recherche recourant à l'entretien approfondi, éventuellement accompagné d'observation, le chercheur effectue ordinairement les premiers entretiens. Cette pratique lui permet d'acquérir une meilleurs connaissance de l'objet de son étude ; le contact direct avec les personnes interrogées lui apporte une impression globale sur les individus et leur environnement, et une compréhension profonde, grâce au dialogue, de la manière dont les répondants vivent leur situation et leurs problèmes. La relation que l'entretien dit "non directif" instaure entre le chercheur et son interlocuteur permet un échange affectif, et peut produire un phénomène d'empathie qui favorise une certaine forme d'identification du chercheur avec son interlocuteur. Cette relative identification exercera ultérieurement une forte influence sur la détermination des premières unités-noyaux.

Ensuite, selon les objectifs de la recherche, les moyens dont il dispose, et les délais qui lui sont impartis, le chercheur peut ou bien effectuer lui-même la totalité des entretiens, ou bien en confier la réalisation à des enquêteurs-psychologues qu'il aura préalablement formés en vue de cette recherche particulière. De la qualité des relations que le chercheur entretient avec les enquêteurs-psychologues de son équipe, et du temps qu'il consacre au compte-rendu par ceux-ci de leurs activités, dépend la possibilité que se produise, avec certains interviewés, une sorte d'empathie par enquêteur interposé.

Les informations recueillies sont évidemment le texte de l'entretien lui-même (en général enregistré sur magnétophone), les renseignements factuels sur la personne interrogée et son environnement, et diverses notations psychologiques plus ou moins subjectives exprimant les impressions de l'interviewé. Ces informations sont par conséquent très riches, et fortement structurées tant par le rythme et l'ordre d'énonciation du discours des locuteurs que par les mots de liaisons placés entre les énoncés ("c'est pourquoi", "à cause de", etc.). D'autre part, en raison de la technique même de l'entretien en profondeur, ces informations présentent, les unes par rapport aux autres, des lacunes : certains thèmes sont abordés par l'un des répondants et non par

les autres. Cette richesse et ce caractère lacunaire constituent des difficultés importantes pour la condensation et la normalisation des informations.

La condensation des informations se fait unité par unité (individu par individu). Il s'agit de résumer toutes les informations relatives à un individu sur une fiche de synthèse. En outre, pour faciliter la comparaison des fiches au cours de la procédure d'agrégation, celles-ci doivent être conçues selon un modèle unique. Cette normalisation, dans la présentation des informations, pose évidemment un problème, dans la mesure où les thèmes abordés et développés par les répondants peuvent être sensiblement différents. La solution communément adoptée consiste à mettre au point une grille de codification pour l'ensemble des informations fournies par la majeure partie (voire la totalité) des répondants, et à réserver sur chaque fiche une place pour les informations spécifiques à chaque individu.

La grille de codification est une sorte de questionnaire que l'on fait subir aux données recueillies. Ce questionnaire est l'expression, sous une forme opératoire, des préoccupations du chercheur et de ses hypothèses (qui ont déterminé la stratégie des entretiens). Il tient compte en outre des thèmes non prévus qui ont pu être abordés spontanément par les répondants (c'est-à-dire de la manière dont les entretiens se sont effectivement déroulés). La grille de codification est construite à partir de l'analyse d'un petit nombre d'entretiens choisis pour leur richesse et leur variété. Elle comporte la liste des thèmes recherchés et, pour chaque thème, la liste des modalités que l'on désire distinguer. S'y ajoute l'indication des moyens d'identifier l'information cherchée, et des règles de codification de celle-ci ; ces préceptes sont accompagnés d'exemples de cas moyens et de cas litigieux.

Les informations correspondant à une rubrique de la grille d'analyse sont codées et reportées sur les fiches individuelles ; les informations spécifiques à un individu sont notées en clair dans une zone de la fiche réservée à cet effet. Elles peuvent comprendre par exemple la description physique de l'individu et de son cadre de vie, son portrait psychologique esquissé par l'enquêteur, un résumé de ses déclarations non codifiées, et quelques phrases caractéristiques tirées de l'entretien.

La confection des fiches a été, dans les cas que nous avons étudiés, réalisée soit par le chercheur lui-même, soit sous son contrôle étroit et avec son active participation ; cette pratique renforce la connaissance en profondeur que le chercheur a de ses données.

La phase d'agrégation des unités apparaît à l'observateur comme l'exécution d'une tâche de classement par essais et erreurs. L'objectif du chercheur est de répartir l'ensemble de ses fiches en un petit nombre de tas homogènes et bien distincts. Il est guidé pour cela par la ressemblance globale qu'il perçoit entre les individus matérialisés par ses fiches ; mais cette impression de ressemblance est sujette, au cours de l'élaboration de la typologie, à des fluctuations et des modifications difficilement explicitables, correspondant à des restructurations des schémas perceptifs du chercheur. D'où le parti que nous avons pris de décrire essentiellement le comportement matériel du chercheur effectuant son classement, sans prétendre expliciter complètement les mécanismes psychologiques sous-jacents.

Au départ de la procédure d'agrégation, il y a ce que nous avons appelé les *unités-noyaux*. Ce sont les cas, peu nombreux, que le chercheur a le sentiment de bien comprendre : non seulement il se représente de manière syncrétique l'ensemble des traits relatifs à ces personnes relevés au cours des entretiens, mais il a l'impression, pour chacune d'elles, d'être capable d'imaginer avec vraisemblance leur comportement dans d'autres situations que celles évoquées au cours des entretiens. En un mot, il les voit vivre. Bien entendu, nous n'avons pas plus que le chercheur les moyens d'évaluer la validité de ces impressions. Au demeurant, cette validité importe peu : d'une part, le sentiment du chercheur n'est ici que l'expression du fait qu'il considère au départ les individus-noyaux comme des cas particulièrement "typiques" ; d'autre part, le fait que ce soient ces individus-là, et non d'autres, qui sont sélectionnés au commencement, ne nous a pas paru avoir une très forte incidence sur la typologie finalement obtenue.

Concrètement, le chercheur dispose les fiches correspondant aux unités-noyaux en les séparant nettement : ce sont les amorces des tas qui constitueront les types. Il prend ensuite une autre fiche, et compare son contenu à celui des fiches déjà classées. Si, globalement, l'individu correspondant lui parait ressembler à l'un des individus déjà classés, il affecte sa fiche au tas correspondant ; si, globalement, il lui parait très diffé-

rent de l'ensemble des individus déjà classés, il crée un nouveau tas et y affecte sa fiche ; si, globalement, le cas lui parait ambigu, ou peu net, il place sa fiche en attente.

Au fur et à mesure que ce processus sa déroule, trois phénomènes affectent la typologie en cours d'élaboration. Ce sont : la création de nouveaux tas, ne comprenant pas (au départ du moins) d'unité-noyau ; la séparation en deux tas ou plus d'un tas existant ; la fusion de tas existants en un seul tas. Lorsque le chercheur crée un tas nouveau, c'est qu'il estime que l'individu dont il examine la fiche présente des traits saillants par lesquels il s'oppose nettement aux individus des autres tas ; la création d'un nouveau type contribue ainsi à structurer l'espace typologique d'une manière explicite, en créant l'amorce de dimensions permettant de distinguer les types. La division d'un tas existant est la conséquence de la non-transitivité de la ressemblance. Bien qu'en principe tous les individus d'un même tas doivent se ressembler, il arrive que la chaîne des ressemblances soit tellement distendue que les individus situés aux extrémités de la chaîne s'opposent sur certains traits importants ; il devient alors naturel de rompre cette chaîne en son point le plus faible, pour constituer deux tas distincts. La fusion, en un seul tas, de tas auparavant séparés traduit une modification dans l'importance relative que le chercheur attribue aux différentes caractéristiques des individus : les traits sur lesquels se fondait la distinction entre deux types lui apparaissent moins pertinents pour la classification que les ressemblances entre ces deux types. Toutes ces opérations (création, division, fusion) sont la manifestation d'une restructuration de l'espace typologique dans la pensée du chercheur.

En principe, la typologie est achevée lorsque toutes les fiches ont été affectées à un tas et un seul. Tant qu'il reste au moins une fiche en attente, le chercheur se trouve contraint soit de remanier sa typologie afin de trouver une place à l'individu non classé, soit de reconsidérer la codification qu'il a faite des informations correspondant à cet individu, et de reprendre l'analyse de l'entretien. Cette dernière éventualité conduit parfois le chercheur à décider de prendre en compte une information dont il n'avait pas jusqu'ici perçu la pertinence, et à recommencer partiellement l'analyse de tous les entretiens. Il est très rare que le chercheur se résigne à laisser un individu en dehors de la typologie. Mais lors même que tous les indi-

vidus ont été classés, la tâche du chercheur n'est pas terminée : il lui faut rendre la typologie opératoire.

Pour que la typologie résultante soit efficace, il est souhaitable : qu'elle permette en principe de classer tout individu nouveau ; que les critères d'affectation d'un individu à un type soient clairs et faciles à appliquer ; qu'elle soit facile à décrire globalement en termes de dimensions. Pour satisfaire à ces critères, ainsi qu'aux exigences communes à toutes les typologies (petit nombre de types, répartition aussi équitable que possible), le chercheur peut être conduit à remanier la première typologie obtenue, et à restructurer ainsi l'espace typologique sous-jacent.

b) Analyse des hypothèses implicites. Comparée aux deux autres procédures de construction de typologie, l'agrégation autour d'unités-noyaux apparait comme très empirique. Il y a peu d'*a priori* explicites : alors que les dimensions permettant de décrire les unités jouaient un rôle central dans les procédures analysées précédemment, elles n'apparaissent guère, au cours du processus d'agrégation, que comme un procédé commode pour résumer une partie de l'information relative aux unités à classer. Il semble que la comparaison des unités entre elles se fasse non au niveau des traits servant à les décrire, mais au niveau de leur structure globale ; cette procédure d'agrégation apparait ainsi comme plus proche des données (et de leur complexité) que les précédentes. Cet empirisme rend particulièrement difficile la description formelle de la procédure suivie. C'est pourquoi nous avons pris le parti de décrire un automate reproduisant de manière simplifiée les comportements observés chez les chercheurs, tout en laissant non définis certains points sur lesquels nous ne disposions que de peu d'hypothèses (en particulier celui de la *fonction de ressemblance*). Telle que nous l'avons élaborée, cette description ne doit être considérée que comme l'approximation discrète d'un processus de pensée continu (et surtout infiniment plus complexe).

Au départ, le chercheur dispose d'un ensemble U d'unités, décrites chacune selon une structure de traits ; un sous-ensemble de ces traits est dit commun à l'ensemble de ces unités, en ce sens, que pour chacun d'eux, il est possible de dire pour toute unité $u \in U$ si elle le possède ou non.

Un sous-ensemble $U' \subset U$ est sélectionné par le chercheur selon une procédure reposant sur sa compétence et son expérience personnelle ; formellement, nous l'assimilerons à un tirage au hasard. En pratique, on a le plus souvent $|U'| < 10$. A ce stade, nous appellerons U' *l'ensemble des unités-noyaux* de la typologie.

On utilisera une fonction de ressemblance ϕ, définie sur U^2 ; cette fonction détermine le degré de ressemblance entre chaque couple d'unités en fonction d'analogies liées à leurs traits descriptifs, et principalement à la structure interne de ces traits pour chaque unité. La fonction ϕ devra satisfaire aux conditions suivantes :

$$\forall u_i, u_j \in U, \quad \phi(u_i,u_j) = \phi(u_j,u_i)$$

$$\forall u_i, u_j \in U, \quad \phi(u_i,u_j) = \max \iff u_i \equiv u_j$$

$$\forall u_i, u_j, u_k, u_l \in U, \quad \phi(u_i, u_j) < \phi(u_k, u_l) \iff u_i \text{ ressemble plus à } u_j \text{ que } u_k \text{ ne ressemble à } u_l.$$

Sur cette fonction de ressemblance, nous déterminerons trois valeurs qui serviront de seuils :

r = valeur minimum de ϕ pour que l'on considère que deux unités se ressemblent ;

δ = seuil de différenciation entre deux mesures de ressemblance ;

d = valeur de ϕ maximum pour que l'on considère que deux unités sont dissemblables.

Nous aurons par conséquent les relations $r > d$ et $r-d > \delta$. En outre, la forme de la fonction ϕ pourra varier au cours du déroulement de la procédure. Au début ($t = t_o$), elle aura la forme ϕ_o, et les seuils vaudront respectivement r_o, δ_o et d_o.

Nous définirons en outre, sur un ensemble quelconque de couples d'unités $\subset U^2$, une fonction d'homogénéité h, liée directement à l'ensemble des valeurs de la fonction ϕ définies sur ces couples. A titre d'exemple, on peut imaginer que la fonction h est la moyenne des valeurs de ϕ. Nous appliquerons cette fonction d'homogénéité à deux cas particuliers de parties de U^2. Le premier cas est l'ensemble des couples d'unités appartenant à une

partie T quelconque de U : la fonction $h(T^2)$ mesure alors l'homogénéité interne du sous-ensemble T. Nous fixerons une valeur θ de h qui sera le seuil d'homogénéité minimum nécessaire pour que T puisse être considéré comme un type. Le second cas est l'ensemble des couples (u_i,u_j) tels que u_i et u_j appartiennent respectivement à deux parties disjointes de U : $u_i \in T_i \subset U$, $u_j \in T_j \subset U$, et $T_i \cap T_j = \emptyset$. La fonction $h(T_i,T_j)$ mesure alors l'homogénéité entre deux sous-ensembles disjoints d'unités. Nous fixerons une valeur η de h qui sera le seuil d'homogénéité maximum pour que T_i et T_j constituent deux types distincts. On aura la relation $\theta \geqslant \eta$. La fonction h dépendant de ϕ, sa forme pourra varier au cours du déroulement de la procédure. Au début, elle aura la forme h_o, et les seuils vaudront respectivement θ_o et η_o.

Enfin, on fixera le nombre minimum m et le nombre maximum n de types recherchés, ainsi que le nombre minimum *min* et le nombre maximum *max* d'unités qu'un type pourra compter à l'issue de la procédure de classement. On aura les contraintes : $m.max \geqslant |U|$, et $n.min \leqslant |U|$.

L'algorithme d'agrégation repose sur deux types de décisions : celles qui portent directement sur les unités, et celles qui concernent les sous-ensembles d'unités. Le premier type de décisions se réduit à l'alternative : ou bien intégrer une unité non encore classée à l'un des types existants, ou bien en faire le point de départ d'un nouveau type. La seconde espèce de décisions regroupe celles qui modifient la partition en types des unités déjà classées : division d'un type jugé trop hétérogène, ou fusion de types voisins en un seul type. Ces décisions ont des effets sur le nombre des types obtenus, leur taille, leur homogénéité interne, et sur l'homogénéité inter-groupe. D'autre part, elles sont influencées par les valeurs de seuil correspondant à ces mêmes paramètres. Pour le bon déroulement des processus de décision, il importe que soit fixé un ordre de priorité dans la prise en considération des différents critères de décision. Cet ordre peut d'ailleurs se trouver modifié en cours de processus. Pour simplifier notre exposé, nous avons supposé que cet ordre restait constant jusqu'à ce que toutes les unités aient été classées, et qu'il se définissait

comme suit (en commençant par le critère intervenant en premier) :

1) seuil d conduisant à la création d'un nouveau type ;
2) seuil r régissant l'intégration à un type d'une nouvelle unité ;
3) seuil η conduisant à la fusion de deux types ;
4) seuil θ conduisant à la division d'un type ;
5) *min* = effectif minimum d'un type ;
6) *max* = effectif maximum d'un type ;
7) m = nombre minimum de types ;
8) n = nombre maximum de types.

Cet ordre de priorité ne joue que lorsque plusieurs critères doivent en principe intervenir simultanément pour une même décision.

La mise en oeuvre de l'algorithme peut être schématisée comme suit. Au moment t_o, on a le sous ensemble U'_o des unités-noyaux qui constitue la typologie de départ. A ce stade, il n'est pas indispensable que $|U'_o| \leq n$.

Au moment t_1, on considère la première unité à classer $u_1 \in \complement_{U_o} U'$ (l'ordre des unités à classer est arbitraire). Pour chaque couple (u_1,u'_i), avec $u'_i \in U'_o$, on évalue $\phi_o(u_1,u'_i)$.

Si $\forall u'_i \in U'_o$, $\phi_o(u_1,u'_i) \leq d_o$, l'unité u_1 est une nouvelle unité-noyau qui s'ajoute à la typologie de départ. Si $\forall u'_i \in U'_o$, $\phi_o(u_1,u'_i) > d_o$ *et* $\phi_o(u_1,u'_i) < r_o$, l'unité u_1 est placée en attente de classement. Si $\exists\, u'_i \in U'_o$, $\phi_o(u_1,u'_i) \geq r_o$, on recherche l'unité-noyau u'_j pour laquelle on a la valeur $\phi_o(u_1,u'_j)$ maximum. S'il n'existe pas d'unité-noyau u'_k telle que $\phi_o(u_1,u'_j)-\phi_o(u_1,u'_k) < \delta_o$, l'unité u_1 est affectée au type T_j auquel appartient u'_j ; sinon, elle est mise en attente de classement entre les types T_j et T_k (ou T_j, T_k et T_l ; etc.).

A un moment en cours de processus t_p, la mise en oeuvre de l'algorithme est un peu plus complexe, en raison du fait que les types ne se réduisent plus à une unité noyau, mais comprennent (en général) plusieurs unités ; le nombre total d'unités classées $|U'_{p-1}|$ est plus élevé. On considère la première unité à classer u_p, et, comme précédemment, on évalue $\phi_{p-1}(u_p,u'_i)$ pour chaque $u'_i \in U'_{p-1}$. On applique les mêmes règles qu'au moment t_1 en ce qui concerne la décision de créer un nouveau type avec u_p comme premier élément, de mettre u_p en attente de classement, ou de l'affecter à un type existant. Simplement, la décision d'affectation est soumise à une condition préalable supplémentaire : c'est qu'il n'existe pas, à l'intérieur du type T_k auquel on envisage d'agréger u_p, une unité u'_j telle que $\phi_{p-1}(u_p,u'_j) < d_{p-1}$. Dans le cas contraire, on peut placer u_p en attente de classement.

En cas d'affectation de u_p à un type T_k, l'addition d'une nouvelle unité modifie la taille $|T_k|$ du type, et (en principe) les valeurs des coefficients d'homogénéité internes $h_{p-1}(T_k^2)$, et externes $h_{p-1}(T_k,T_j)$ pour $T_j \neq T_k$. Quatre cas sont alors à considérer :

1) Si l'on a à la fois $h_p(T_k^2) \geqslant \theta_{p-1}$ et $h_p(T_k,T_j) \leqslant \eta_{p-1}$ pour tous les $T_j \neq T_k$, l'affectation de u_p à T_k est admise (sous réserve de la condition de taille du type T_k, qui est à ce stade une condition assez peu contraignante, et dont nous verrons les effets plus loin).

2) Sinon, si l'on a à la fois $h_{p-1}(T_k^2) \geqslant \theta_{p-1}$ et $\exists T_j,\ h_{p-1}(T_k,T_j) > \eta_{p-1}$, on affecte u_p à T_k et on procède à la fusion de T_k et du T_j pour lequel $h_{p-1}(T_k,T_j)$ = maximum, créant ainsi un type $T_m = T_k \cup T_j$.

3) Si l'on a $h_{p-1}(T_k^2) < \theta_{p-1}$ et $h_{p-1}(T_k,T_j) \leqslant \eta_{p-1}$ pour tous les $T_j \neq T_k$, il faut procéder à la division de T_k en deux types plus homogènes, T_a et T_b. Pour cela, on calcule $\phi_{p-1}(u_p,u'_i)$ pour toutes les unités $u'_i \in T_k$, et l'on ordonne les u'_i par valeurs décroissantes de ϕ_{p-1}. On

peut alors (par exemple) décider d'agréger à u_p toutes les u'_i pour lesquelles la fonction ϕ_{p-1} est supérieure à un seuil donné (compris entre d_{p-1} et r_{p-1}) ; ou bien les $\frac{|T_k|}{2}$ unités classées en tête. On obtient ainsi un type T_a constitué autour de u_p, et un type $T_b = T_k - T_a$. On calcule alors $h_{p-1}(T_a,T_j)$ et $h_{p-1}(T_b,T_j)$ pour tous les T_j tels que $T_j \neq T_a$ *et* $T_j \neq T_b$, et l'on procède s'il y a lieu à des fusions comme ci-dessus.

4) Si l'on a $h_{p-1}(T_k^2) < \theta_{p-1}$ et $\exists\, T_j$, $h_{p-1}(T_k,T_j) > \eta_{p-1}$, on procède comme dans le cas 3) pour diviser T_k en deux types, T_a et T_b. On fusionne ensuite le type T_a, agrégé autour de u_p, et le type T_j pour lequel $h_{p-1}(T_k,T_j)$ = maximum.

A certaines étapes de la mise en oeuvre de l'algorithme (mais non nécessairement à chacun des moments t_i), on procède à un examen général de la typologie obtenue sur les unités déjà classées. Cet examen peut conduire à reconsidérer les valeurs de seuil retenues précédemment ; il peut éventuellement déboucher sur la définition d'une nouvelle forme de la fonction ϕ.

Un contrôle simple consiste à vérifier que la taille $|T_j|$ des types obtenus n'est pas supérieure à la taille maximum acceptable *max*, et que le nombre N de types constitués n'excède pas le nombre maximum fixé au départ n. Si l'on a $|T_j| > max$ pour un au moins des types constitués, on peut agir sur les dimensions des types en élevant un ou plusieurs des seuils θ, η, r ou d. Si l'on a au contraire $N > n$, on peut abaisser un ou plusieurs des mêmes seuils. D'autre part, si l'on considère que le nombre d'unités en attente de classement est trop élevé, on peut diminuer ce nombre en réduisant la différence $r - d$. En cas de modification de la valeur d'un ou de plusieurs seuils, il est nécessaire de remanier la typologie et de ré-examiner le cas des unités en attente de classement en fonction de ces nouvelles valeurs, avant de procéder à l'agrégation d'une nouvelle unité.

Nous avons signalé, en décrivant la procédure, que l'appréciation par le chercheur de la ressemblance entre deux unités évolue au fur et à mesure qu'il progresse dans l'élaboration de sa typologie. Cette évolution peut être simulée par une modification de la fonction ϕ. Le mécanisme réel

conduisant à cette modification est probablement la conséquence à la fois d'une maturation dans les réflexions du chercheur, et de l'examen des types obtenus. Nous supposerons que seul intervient le second facteur, et que la remise en cause de la fonction ϕ a pour but à la fois de maximiser l'homogénéité intragroupe, et de minimiser l'homogénéité intergroupe. Par conséquent, à ce stade, on reconsidère le poids attribué à chaque information sur les unitée et la forme de la relation unissant ces informations à la fonction ϕ, de manière à maximiser une fonction des coefficients $h(T^2)$ et à minimiser une fonction des coefficients $h(T_i,T_j)$. On redéfinit ensuite la fonction h et la valeur des divers seuils en tenant compte de la nouvelle forme de ϕ. On remanie en conséquence la typologie obtenue avant de continuer l'agrégation.

Quand toutes les unités ont été classées, il reste à considérer tout d'abord la taille minimum et le nombre des types obtenus. S'il existe un type T_i tel que $|T_i| < min$ ou si l'on a $N < m$, il est indispensable de reprendre la procédure de fusion ou de division des types obtenus de manière à satisfaire à ces conditions. Il faut ensuite examiner si l'on peut présenter les résultats de cette procédure de classification sous une forme simple et claire, à l'aide d'un petit nombre de dimensions. Pour cela, on recherche, parmi l'ensemble des informations relatives aux unités (y compris celles qui n'ont éventuellement pas été prises en compte pour l'évaluation de la fonction de ressemblance ϕ), celles qui sont communes à toutes les unités d'un type (ou à la majorité d'entre elles). Si l'on ne trouve pas d'informations (de traits) caractéristiques, on peut tenter de définir une nouvelle fonction de ressemblance et recommencer la procédure de classification. Si l'on trouve des traits majoritaires pour chacun des types, encore faut-il que la combinaison (la conjonction) de certains de ces traits permette de caractériser chaque type ; c'est-à-dire qu'il ne doit pas y avoir deux types possédant simultanément les mêmes traits dominants et eux seuls. Si cette dernière condition n'est pas remplie, il faut revoir la procédure de classification.

Si l'on admet que l'algorithme présenté constitue une bonne approximation de la procédure réellement appliquée, on peut dégager les hypothèses suivantes, contenues dans la méthode sous-jacente :

1) Les objets (unités) à classer sont décrits à partir de traits de toute nature (qualitatifs, ordinaux, métriques).

2) Pour un objet donné, des relations de type causal sont postulées entre ces traits ; l'ensemble de ces relations constitue la structure de l'objet considéré.

3) La procédure de classification prend en compte principalement la structure définie sur les objets, plutôt que leurs traits descriptifs considérés isolément.

4) Sur chaque couple d'objets, on peut définir une fonction de ressemblance à partir de la comparaison de leurs structures respectives. Cette fonction prend une valeur d'autant plus élevée que les structures des deux objets sont plus voisines ; elle prend sa valeur maximum lorsque les structures des deux objets considérés sont identiques.

5) On peut fixer une valeur-seuil de cette fonction de ressemblance, telle que, pour tout couple d'objets, si la valeur prise par la fonction est supérieure ou égale à la valeur-seuil, on dira que les objets se ressemblent ; si elle est inférieure à la valeur-seuil, on dira qu'il ne se ressemblent pas. On définit ainsi, sur tout couple d'objets, une relation de ressemblance. Cette relation est réflexive, symétrique, et intransitive (mais non anti-transitive).

6) On postule que, bien que la relation de ressemblance soit intransitive, on peut lui associer une partition de l'ensemble des objets (sans recouvrement).

7) Pour atteindre cet objectif, on peut faire varier la valeur-seuil sur la fonction de ressemblance. Si aucune valeur-seuil ne permet d'obtenir des sous-ensembles disjoints, on peut alors modifier la fonction de ressemblance elle-même.

8) Lorsque l'on est parvenu à définir des sous-ensembles disjoints d'objets (ou types), il est possible de sélectionner un petit nombre de traits descriptifs dont la combinaison permet de caractériser chaque sous-ensemble (type).

9) On peut identifier les types obtenus par les coordonnées du barycentre du nuage de points dans le sous-espace défini par les traits descriptifs retenus.

10) Pour décrire les caractéristiques de chacun des types obtenus, il est possible de définir une structure de traits commune à la majorité des unités classées dans un même type ; dans le sous-espace d'identification des types, cette structure doit se confondre avec le barycentre du type.

2.3. Comparaison des trois procédures : "systématisme", pragmatisme, empirisme

Au cours de notre description des procédures utilisées dans les sciences sociales pour élaborer une typologie, nous sommes passés de la méthode la plus abstraite (la plus systématique) à la méthode la plus proche des données observées (la plus empirique). Le tableau comparatif ci-dessous résume les principales caractéristiques de ces trois procédures. On constatera que la procédure intermédiaire (réduction d'un espace d'attributs) y apparait à la fois comme la plus facile à manier et à exposer, et comme la plus criticable à la fois du point de vue de la théorie, et du point de vue de la fidélité aux observations. S'il fallait résumer d'un qualitatif les propriétés de chacune de ces procédures, nous aurions le triptyque : systématique/pragmatique/empirique.

Caractéristiques des procédures	SYSTEMATIQUE : Constitution de types-idéaux	PRAGMATIQUE : Réduction d'un espace d'attributs	EMPIRIQUE Agrégation autour d'unités-noyaux
Point de départ :	Théorie (lois et concepts).	Cadre de description des observations (dimensions).	Structure interne des unités observées.
Démarche dominante :	Analyse sémantique des concepts et de leurs relations.	Analyses sémantique et empirique des dimensions.	Comparaison empirique des unités observées.
Nature des variables :	Variables orientées, permettant d'ordonner les unités (*ordinal*).	Réduction de toutes les variables au niveau *nominal*.	Mesure de la ressemblance entre les unités (*métrique*).
Principale information prise en compte :	Relations entre les contenus sémantiques des dimensions	Distribution (observée ou hypothétique) des unités dans les zones de l'espace d'attributs.	Relations entre les structures internes des unités observées.

Caractéristiques des procédures	SYSTEMATIQUE : Constitution de types-idéaux	PRAGMATIQUE : Réduction d'un espace d'attributs	EMPIRIQUE : Agrégation autour d'unités-noyaux
Hypothèses sur la relation observation/théorie:	La structure cherchée se trouve dans la représentation (théorique) de la réalité ; pas d'hypothèse sur la réalité elle-même, en dehors de l'inexistance de "types purs".	La structure cherchée est un balisage de la réalité commode pour l'esprit ; il n'est pas indispensable que les lignes de partage existent réellement.	La réalité est complexe, mais non indifférenciée ; il est possible d'en dégager une partition "naturelle".
Relation entre les unités observées et les types :	Distance par rapport à ces extrêmes ("expliquer le normal par le pathologique").	Localisation dans une zone définie abstraitement par la présence ou l'absence d'attributs	Distance par rapport à cas centraux ("individus moyens pour le type considéré").
Appartenance à un type :	Toujours multiple, avec des degrés variables selon les types-idéaux.	Toujours univoque (aucune ambiguïté).	En principe univoque, mais problème des cas frontières entre deux types.
Qualités du résultat :	Grande cohérence théorique, pouvoir d'explication élevé.	Facile à exposer, classement sans problème des unités observées.	Grande fidélité aux observations, respect de la complexité de la réalité.

Nous avons signalé au début du § 2.2., que les démarches réelles que nous avons pu observer ne sauraient être réduites à une seule de ces trois procédures, mais présenteraient plutôt des oscillations de l'une à l'autre. D'ailleurs, nous avons noté incidemment qu'aucune procédure n'était absolument systématique ni absolument empirique. En particulier, nous connaissons bien l'importance du rôle joué par les mesures et les observations dans l'élaboration d'une théorie ; et nous avons signalé, dans le processus de

construction de types idéaux (§ 2.2.1), le nécessaire parallélisme entre les liaisons statistiques observables sur les données, et les relations sémantiques et/ou les liens de causalité dans la théorie. D'autre part, il est admis que les techniques de recueil d'information, et par conséquent la nature des informations recueillies, dépendent étroitement des hypothèses et des théories du chercheur ; et il est clair que des considérations théoriques interviennent dans la procédure empirique d'agrégation des unités autour d'unités-noyaux (§ 2.2.3), tant pour définir la structure interne des unités observées que pour évaluer la fonction de ressemblance entre unités.

Cette absence de frontières entre les trois procédures décrites, que l'on observe dans la pratique classificatoire, s'explique aisément lorsqu'on la replace dans la perspective de toute démarche scientifique. L'élaboration d'une typologie se déroule simultanément sur deux plans parallèles : celui des *observations*, c'est-à-dire celui des dénombrements des mesures, des relations de contiguïté, de simultanéïté, de succession dans le temps ; et celui des *explications*, c'est-à-dire celui des concepts, des lois, des relations de causalité. Ces deux plans dépendent étroitement l'un de l'autre ; nous avons vu, en décrivant les démarches observées, à quel point la statistique et la sémantique (ou, si l'on préfère, l'extension et la compréhension) s'étayent mutuellement pour la mise en ordre des données d'observation. La démarche réelle du chercheur nous parait être un constant va-et-vient entre ces deux plans ; c'est ce mouvement dialectique que les procédures de classification automatique doivent soit reproduire, soit au moins ne pas entraver.

3. UN SYSTEME INFORMATIQUE D'AIDE A LA CONSTRUCTION DE TYPOLOGIES

L'analyse de la démarche du chercheur en sciences sociales lorsqu'il élabore une typologie doit déboucher sur une comparaison des procédures manuelles et des procédures automatisées. Une telle comparaison révèle des différences importantes. Dans ces conditions, il serait tentant de concevoir de nouveaux algorithmes de classification automatique s'inspirant de cette analyse ; nous verrons qu'ils imposeraient à l'utilisateur une tâche d'explication préalable de ses hypothèses et de ses objectifs qui serait en fait pratiquement irréalisable. Reste alors une solution plus modeste : considérer qu'en l'état actuel du développement des sciences sociales comme de l'outil informatique, une part importante des activités du chercheur ne saurait être automatisée ; mais qu'il peut être intéressant, par contre, d'assister et de stimuler la réflexion de celui-ci en recourant à l'informatique pour réaliser, avec rapidité et sécurité, certaines opérations de manipulation des données. L'énumération des opérations élémentaires utiles au chercheur dans l'élaboration d'une classification est l'amorce du cahier des charges d'un système informatique d'aide à la construction de typologies dans les sciences sociales.

3.1. Comparaison des procédures manuelles et des procédures automatisées.

Dans les procédures automatisées de classification que nous avons analysées, nous avons jugé bon d'inclure les méthodes de segmentation. En principe, ces méthodes ont pour but l'exploration des principales liaisons et interactions entre un ensemble de variables dites explicatives, et une variable à expliquer. Mais la procédure suivie pour atteindre cet objectif consiste à rechercher une série de dichotomies sur les unités étudiées, qui soient optimales du point de vue d'un certain critère (maximiser la liaison entre une variable explicative et la variable à expliquer) ; on obtient ainsi une partition de l'ensemble des unités étudiées, selon les valeurs prises par les principales variables explicatives. Ce sous-produit des méthodes de segmentation est souvent considéré comme une typologie des unités étudiées, bien que cette typologie ne soit optimale ni du point de vue de l'homogénéité interne des types, ni du point de vue de l'hétérogénéité entre les types, ni du point de vue de la facilité d'exposition et d'utilisation des résultats.

Selon la manière dont sont obtenus et présentés les résultats de la procédure de classification, on peut distinguer deux grandes familles de méthodes : les classifications monothétiques et les classifications polythétiques. Dans les premières, les types sont décrits à partir de variables intervenant séparément, et déterminant des partitions successives de l'ensemble des objets à classer. Les sciences naturelles en fournissent de nombreux exemples : ainsi, selon Émile LITTRE, les objets étudiés par les naturalistes se subdivisent d'abord en animaux, végétaux, et minéraux ; les animaux, en vertébrés et invertébrés ; les vertébrés, en mammifères, oiseaux, reptiles, et poissons; etc. On obtient ainsi une classification arborescente, qui suppose un ordre *a priori* sur les variables (cf. par exemple la hiérarchie : règne > embranchement > classe > ordre > famille > genre > espèce). Dans les classifications polythétiques, selon Michel ADANSON, "il existe autant d'espèces qu'il y a d'individus différant entre eux d'une ou plusieurs différences quelconques" ; les types se définissent alors par l'ensemble de leurs ressemblances et de leurs dissemblances, c'est-à-dire à partir de l'ensemble des variables servant à décrire les objets à classer (on préfère même parfois les définir en extension, c'est-à-dire en énumérant les objets qui en font partie).

Les méthodes de segmentation sont le type même de la classification monothétique (aux réserves que nous avons formulées sur la valeur du résultat près). Les principales d'entre elles (méthode de BELSON ; programme AID de SONQUIST et MORGAN ; programmes ELISEE, SEGMA, MULTISEG, etc.) ne diffèrent que par les contraintes qu'elles imposent sur la nature des variables, et la mesure de liaison entre variables qu'elles utilisent. Comme toute classification monothétique, ces procédures impliquent l'existence d'un ordre sur les variables servant à classer les objets étudiés. Mais ici, cet ordre n'est pas défini *a priori*, à partir de considérations théoriques ; il résulte de l'importance des liaisons entre ces variables et une variable privilégiée, la variable "à expliquer". Les considérations théoriques interviennent cependant pour le choix de la variable "à expliquer", et la sélection des variables "explicatives" parmi l'ensemble des variables servant à décrire les objets à classer ; elles jouent également un rôle dans la fixation des coupures possibles ou des regroupements autorisés sur les variables. Mais lorsque l'on a fixé ces points, et choisi le coefficient mesurant les corrélations, la procédure de segmentation se déroule automatiquement selon des normes purement statistiques. Si l'on compare les programmes de segmentation aux procédures manuelles de construction de typologies, il n'y a guère que la réduction d'un espace d'attributs qui présente avec elles quelques analogies (sélection de variables descriptives, et de zones sur les variables sélectionnées).

La quasi-totalité des programmes de taxinomie numérique sont des procédures de classification polythétiques. Certaines d'entre elles se fondent sur la notion de distance entre les partitions que déterminent sur l'ensemble des objets à classer chacune des variables servant à les décrire ; on recherche alors une partition centrale (médiane ou moyenne, par exemple) qui définit la typologie (cf. par exemple les programmes PARTITIONS CENTRALES, et ECH.). La procédure suivie consiste à cheminer dans l'ensemble des partitions de l'ensemble des objets étudiés (c'est-à-dire à transférer certains objets d'une partie dans une autre), de manière à minimiser une fonction de la distance entre la partition examinée et les partitions originales. Cette manière d'obtenir un ensemble de types nous a paru sans équivalent parmi les procédures manuelles.

Par contre, les programmes utilisant une distance entre les objets à classer présentent tous certaines ressemblances avec la procédure d'agrégation autour d'unités-noyaux. Ces programmes ont tous pour objectif de rechercher, dans l'hyperespace décrivant les objets à classer, des zones de forte densité (*nuages*) qui constitueront les types. Les principaux algorithmes utilisés peuvent être rangés en trois grandes catégories :

1) Construction d'une ultramétrique (méthode de JOHNSON). On regroupe progressivement les paires d'objets pour lesquelles la distance est la plus faible. Chaque regroupement impose le calcul d'une nouvelle distance entre le type obtenu et chacun des autres types, ou des objets non encore regroupés. Cette nouvelle distance est une fonction des distances originales ; ainsi, une distance entre une paire d'objets {x,y} et un objet z est une fonction des distances d (x,z) et d(y,z). Il y a ainsi déformation de l'hyperespace original (passage d'une métrique à une ultramétrique), se traduisant par une modification qui peut être très importante des distances entre objets. En regroupant progressivement objets et/ou types selon leurs distances, on obtient un ensemble de partitions emboîtées. Celui-ci est usuellement présenté sous la forme d'une arborescence dont les sommets terminaux sont les objets à classer ; la racine, l'ensemble des objets ; les noeuds, les regroupements d'objets (types) ; et les arcs, les relations $\supset$ ("contient") entre regroupements. Cette procédure présente des analogies avec la procédure empirique que nous avons décrite : agrégation par évaluation de la ressemblance sur chaque paire d'objets, et remise en cause des distances originales en cours de processus. Mais il n'y a pas d'unités-noyaux données au départ ; et la modification de la fonction de distance est imposée par l'objectif visé (partitions emboîtées), et est réalisée selon des critères purement mathématiques, sans référence à la signification des mesures utilisées.

2) Alternance de séparation et de regroupements (méthode IPHIGENIE). A partir de la matrice des distances, on recherche la plus petite distance, et on regroupe les objets correspondants ; on recherche également la plus grande distance, et on sépare les objets correspondants. On progresse ainsi jusqu'à ce que l'on ne puisse plus procéder à une séparation sans défaire un regroupement précédent, ni procéder à un regroupement sans annuler une séparation antérieure. Les regroupements obtenus à ce stade constituent les types cherchés. Dans cette procédure, le processus de séparation a pour effet d'éviter un risque présenté par la méthode précédente : celui de regrouper, par une sorte d'effet de transitivité de la ressemblance, des objets qui se ressemblent peu entre eux (effet de chaîne). Dans la procédure empirique que nous avons décrite, cet effet est obtenu par la fixation d'un seuil d'homogénéité minimum pour que l'on ait un type.

3) Regroupement autour d'objets initiaux. Cette méthode comporte deux variantes. Dans l'une (programme TYPOL), on fixe un seuil de distance maximum acceptable entre objets d'un même type. On considère ensuite les individus au fur et à mesure qu'ils se présentent : le premier individu constitue le premier type ; le second lui est agrégé si leur distance est inférieure au seuil ; sinon, il constitue le second type ; etc. Dans l'autre variante (programme NUEES DYNAMIQUES), on choisit au départ un petit nombre d'objets étalons ; on agrège chacun des autres objets à classer à l'objet étalon dont il est le plus proche, obtenant ainsi des types. Pour chaque type, on définit un nouvel objet étalon (en général son barycentre), et on réitère le processus d'agrégation jusqu'à ce que les types obtenus restent stables d'une itération à l'autre (convergence). Dans les deux variantes, le nombre de types désirés est fixé au départ.

Le point commun de l'ensemble de ces méthodes de segmentation ou de classification, lorsqu'on les compare aux procédures manuelles, est leur rigidité. En effet, tous ces programmes *(1)* sont conçus pour être exécutés en une seule fois, sans intervention du chercheur en cours d'exécution (*batch processing*). Cela impose au chercheur d'avoir préalablement acquis une expérience importante dans l'utilisation de ce type de programmes, et en particulier de connaître l'incidence, sur la typologie obtenue, du codage des données, de la forme de la fonction de distance sur les objets, des critères d'optimalité,

---

(1) A quelques très rares exceptions près, dont par exemple la technique de segmentation interactive décrite dans STONE, Philip J. ; DUNPHY, Dexter C.; SMITH, Marshall S. ; OGILVIE, Daniel M. : *The General Inquirer* (The M.I.T. Press, 1966), Pp. 121-133.

des valeurs choisies comme seuils, des algorithmes d'agrégation, etc. Si tel n'est pas le cas, le choix de ces paramètres risque de se faire, sinon tout à fait au hasard, du moins sans que l'utilisateur en ait clairement envisagé les conséquences sur le plan théorique.

En décrivant les procédures manuelles, nous avons souligné que la démarche du chercheur se déroulait simultanément sur les plans statistique et sémantique. Les procédures automatisées ne font appel qu'aux critères statistiques. Pour introduire les considérations sémantiques dans les procédures automatisées, il faudrait d'une part complexifier le déroulement des programmes afin de permettre la modification, au vu de résultats intermédiaires, des valeurs des paramètres, des fonctions choisies, et des critères retenus ; d'autre part, trouver le moyen de formuler, d'une manière opératoire pour l'ordinateur, l'ensemble des considérations qui interviennent au cours de la construction d'une typologie à la main, et que nous avons appelé le *savoir implicite* du chercheur. Ce serait à la fois trop demander aux informaticiens, et trop exiger des chercheurs.

Pour tenir compte des enseignements apportés par l'analyse des procédures manuelles, il nous a semblé qu'il convenait de laisser au chercheur le monopole des considérations d'ordre sémantique, et de rétablir les interactions entre le plan sémantique et le plan statistique. Le moyen d'atteindre cet objectif est de réaliser un système interactif. Une solution simple, aisément réalisable, est la réécriture des programmes usuels de classification dans un langage conversationnel, en incluant la sortie de résultats intermédiaires et l'entrée d'instructions modifiant le déroulement du programme. Une telle solution constituerait nous semble-t-il un moyen de répondre aux attentes actuelles des chercheurs.

Une autre solution serait de renoncer provisoirement à des objectifs relativement ambitieux sur le plan méthodologique, et de concevoir un système d'aide informatique à la construction de typologies qui exécute automatiquement des manipulations sur les données, en laissant à l'utilisateur le maximum d'initiative. Ce sont les grandes lignes d'un tel système que nous décrivons ci-après.

3.2. Description d'un système de manipulation de données.

Les analyses des procédures manuelles que nous avons développées, et les essais que nous avons faits pour reconstituer certaines de ces procédures, nous ont permis d'identifier des séquences d'opérations sur les données dont disposait le chercheur. Les opérations élémentaires sur les données sont tou-

jours très simples dans leur principe ; mais leur exécution à la main peut demander un temps relativement long, et se révèler assez complexe. Leur exécution automatique doit par conséquent rendre des services au chercheur. Nous présentons ci-après la description des opérations élémentaires que nous avons recensées.

Auparavant, nous décrivons les caractéristiques générales du système chargé d'exécuter ces opérations. Nous terminons cette présentation par quelques indications sur la manière d'utiliser le système proposé.

3.2.1. Caractéristiques générales du système.

Pour être adapté à ses objectifs, le système proposé doit satisfaire à un petit nombre de conditions que nous énonçons ci-dessous. Ces conditions déterminent la structure des fichiers que l'utilisateur devra avoir à sa disposition.

a) Les principes de base.

La caractéristique fondamentale du système d'aide à la construction de typologies doit être son étroite adaptation aux problèmes et au mode de pensée du chercheur en sciences sociales, tels que nous les avons décrits dans le chapitre précédent. Compte-tenu des développements actuels tant de la méthodologie des sciences sociales que des techniques automatisables, l'une des principales caractéristiques du système est une rigoureuse division du travail entre le chercheur et la machine. Le premier se réservant le monopole des traitements de type sémantique, c'est-à-dire aussi bien de la codification des données "qualitatives" que de l'interprétation des résultats obtenus,et l'ordinateur prenant en compte à la fois l'ensemble des opérations de type logique ou statistique sur les données numériques, et toutes les tâches relatives au stockage, au classement, et à la sélection des informations manipulées. Pour fonctionner efficacement, ce système doit être interactif.

Le dialogue entre le chercheur et la machine doit s'effectuer dans trois domaines distincts :

- la saisie et la mise à jour des données, comprenant le contrôle des informations enregistrées et la détection des anomalies ;
- l'exécution d'opérations sur les données, et la transmission des résultats de ces opérations, ou la demande d'informations complémentaires nécessaires pour leur exécution, ou l'expression d'un diagnostic sur l'impossibilité de les exécuter ;
- l'apport par la machine d'informations sur les opérations exécutées antérieurement, permettant à l'utilisateur de décider la réitération de séquences

ou d'enchaînements de séquences, éventuellement après modification de celles-ci.

b) L'organisation des fichiers.

L'utilisateur doit disposer de quatre types de fichiers distincts :

- un *fichier des données de base,* comprenant l'ensemble des informations tirées des documents relatifs aux unités étudiées (monographies, entretiens, questionnaires, biographies, etc.). Ce fichier est structuré par unité : à chaque unité correspond un enregistrement, de longueur fixe ou variable selon les cas. La structure de chaque enregistrement dépend de la structure des documents utilisés. Les informations peuvent figurer sous forme codée ou en clair. En principe, ce fichier doit pouvoir être mis à jour par apport d'informations complémentaires sur les unités, ou par addition d'unités complémentaires ; mais on ne doit en aucun cas risquer de détruire ou modifier au cours du traitement les informations qu'il contient.

- un *fichier de travail,* extrait du fichier des données de base, mais essentiellement modifiable. Ce fichier est structuré par unité. Comme le fichier des données de base, il compte en principe autant d'enregistrements que d'unités. Mais à la différence de celui-ci, il peut ne comporter qu'une partie des informations relatives à chaque unité, et les informations retenues peuvent à volonté être recodifiées, combinées, ou effacées. En principe, la structure des enregistrements du fichier de travail est standard (du type questionnaire ou grille d'analyse de contenu), et les informations sont codées. Le fichier de travail est obtenu par sélection, condensation, codification ou combinaison des informations contenues dans le fichier des données de base. Si nécessaire, on doit pouvoir créer plusieurs fichiers de travail pour un même ensemble de données.

- un *fichier des opérations,* servant à enregistrer, dans l'ordre dans lequel elles se sont présentées, toutes les opérations qui ont été exécutées sur les données. Un tel fichier constitue le journal de bord de la procédure de recherche d'une typologie. Il est tenu automatiquement par le système, et l'utilisateur ne peut ni le modifier, ni l'effacer. Il permet au chercheur de dresser le bilan de ses tentatives, et de commander la réitération de séquences antérieures.

- un *fichier des résultats,* dans lequel sont conservés à la demande du chercheur les résultats qu'il juge importants (en vue de comparaison ou d'édition ultérieure), et dans lequel le système range automatiquement les résultats d'opérations telles que leur coût d'exécution excède sensiblement le coût de stockage et de consultation dans le fichier résultat.

3.2.2. La gestion des fichiers.

Pour fixer les idées, nous supposerons que l'utilisateur a en permanence à sa disposition une console avec écran de visualisation, et qu'il a accès à des unités de disques lui permettant la libre manipulation (à distance) de ses fichiers. En outre, nous supposerons qu'il lui est possible, sous certaines conditions, d'utiliser un lecteur de cartes ou un dérouleur de bandes, ainsi qu'une imprimante.

Les opérations de gestion des fichiers concernent la création et la mise à jour du fichier des données de base, la création et la modification du fichier de travail, et l'édition des résultats. Le fichier des opérations est un résumé chronologique géré automatiquement par la machine, et sur lequel l'utilisateur n'a d'autre action possible que sa consultation.

a) Saisie et mise à jour des données de base.

Il peut arriver que les données de base se présentent sous une forme directement lisible par la machine ; c'est en particulier le cas pour les résultats d'enquêtes extensives ou de recensements, pour lesquels les informations recueillies ont été codées et reportées sur cartes perforées ou sur bande magnétique. Dans ces conditions, la création du fichier de base est la simple reproduction de ces données, assortie éventuellement d'un contrôle et d'un nettoyage des données. Quant à la mise à jour du fichier de base, en pratique le problème ne se pose pas pour ce type de données.

Par contre, dans la majeure partie des exemples que nous avons analysés, les données sont soit des monographies ou des études de cas, soit des entretiens enregistrés ou des récits ; les données ne peuvent par conséquent être conservées telles quelles (en l'état actuel des techniques de traitement automatique de données textuelles). Le chercheur procède alors à l'analyse du contenu de ses documents de base ; la saisie des données de base doit par conséquent être adaptée à la procédure d'analyse du contenu.

En schématisant beaucoup, l'analyse de contenu est une technique qui permet d'identifier et de sélectionner des éléments du contenu qui présentent un intérêt pour le chercheur. Elle permet en outre de classer et d'étiqueter les éléments de contenu sélectionnés. A partir de cet étiquetage, il est facile de substituer à ces éléments un code (généralement alphanumérique) correspondant à la catégorie à laquelle ils ont été affectés. L'analyse de contenu a donc pour effet d'extraire et de codifier une partie des informations figurant dans les documents de base.

Il faut ici distinguer deux types de documents : d'un côté, les monographies, les études de cas, les questionnaires standardisés (à questions ouvertes), dont la trame est, dans l'ensemble, prédéterminée par le chercheur ; de l'autre côté, les observations de comportement, les protocoles d'entretiens non structurés, les récits, les biographies, dans lesquels la succession temporelle des éléments de contenu peut varier d'une unité à l'autre. Dans le premier cas, la structure des enregistrements correspondant à chaque unité sera, pour l'essentiel, le reflet d'un questionnaire standardisé, la *grille d'analyse de contenu.* Dans le second cas, les enregistrements comporteront le code correspondant à chaque élément de contenu sélectionné, dans l'ordre d'apparition de celui-ci dans le document de base.

La procédure de saisie n'est évidemment pas la même dans les deux cas. Une grille d'analyse de contenu est constituée par l'énumération systématique de toutes les catégories d'information recherchées ; pour une unité donnée, le chercheur relève la présence (ou l'absence) de l'élément de contenu correspondant à chacune des catégories, voire sa fréquence, son intensité, ou toute autre caractéristique opératoirement définie. Pour la saisie sur console, l'utilisateur doit tout d'abord entrer la grille d'analyse qu'il a mise au point, et qui constitue la structure de la partie fixe de chaque enregistrement. La saisie des éléments de contenu concernant une unité donnée peut ensuite s'effectuer de deux manières, au gré du chercheur. Ou bien chaque rubrique de la grille apparaît sous la forme d'une question, et l'utilisateur entre en réponse le code ou la quantité correspondant au contenu à enregistrer ; c'est alors la grille d'analyse qui guide la saisie des données, ce qui suppose que l'unité ait été au préalable entièrement analysée. Ou bien, au fur et à mesure qu'il déchiffre les documents de base, le chercheur entre l'identificateur de la rubrique et le code de l'élément de contenu qu'il vient de détecter ; la procédure de saisie est alors plus souple, mais impose un contrôle (automatique ou manuel) plus poussé après enregistrement d'une unité *(1)*. Lorsque l'ordre d'apparition des éléments de contenu est variable d'une unité à l'autre, il est préférable de conserver la trace de cet ordre d'apparition et d'entrer l'identificateur de la rubrique et le code correspondant aux éléments de contenu au fur et à mesure que ceux-ci se présen-

---

(1) Il importe de vérifier par exemple que ne coexistent pas deux codes possibles mais incompatibles à l'intérieur d'une même unité ; que l'absence d'un code correspond bien à l'absence dans le document de base, de tout élément rattachable à la rubrique correspondante ; que le décompte des fréquences, ou la mesure d'intensité globale, ont été bien effectués par la machine ; etc. Il est donc souhaitable soit de prévoir des contrôles automatiques pendant la saisie (détection des codes multiples ou des absences de code pour une rubrique, p. ex.), soit, après chaque enregistrement d'une unité, de passer en revue la totalité de l'enregistrement réalisé en faisant apparaître la grille d'analyse remplie.

tent. L'enregistrement est alors constitué par une séquence de symboles, de longueur variable, sans qu'il soit besoin d'une grille d'analyse.

En outre, pour chaque unité, il est possible d'ajouter à la partie codée de l'enregistrement toute information spécifique que le chercheur juge bon d'enregistrer (citations en clair, indication d'impressions personnelles, références particulières, etc.). Le statut de telles informations dans le fichier est analogue au statut des commentaires dans un langage de programmation : elles peuvent apparaître à la demande sur un organe de sortie (écran, imprimante), mais elles ne peuvent en aucun cas être traitées.

En principe, le fichier de base ne doit pas pouvoir être modifié au cours de l'élaboration de la typologie ; c'est pourquoi les manipulations de données ne peuvent avoir lieu que sur le fichier de travail. Nous avons vu cependant qu'en cas d'échec de la procédure de construction de typologie, le chercheur pouvait être conduit à rechercher, dans les documents de base, des éléments de contenu qu'il avait auparavant négligés. Il est donc nécessaire de réserver la possibilité d'introduire de nouvelles informations dans le fichier de base. D'autre part, il serait imprudent d'effacer, pour ce faire, des informations que l'on estime alors sans intérêt ; il est possible qu'ultérieurement elles se révèlent indispensables. C'est pourquoi la procédure de mise à jour du fichier des données de base doit permettre d'insérer dans chaque enregistrement des éléments nouveaux, sans altération des éléments précédemment enregistrés. En outre, le fichier de base doit pouvoir également être enrichi par l'introduction d'unités nouvelles.

b) Création et modification du fichier de travail.

Le fichier de travail est obtenu à partir du fichier des données de base. Il est, comme ce dernier, constitué d'enregistrements correspondant chacun à une unité. Ce fichier peut être une simple copie, modifiable, du fichier original. En pratique, ce devrait être un fichier plus maniable que l'original (au départ du moins), obtenu par sélection des informations jugées les plus utiles. De plus, pour la quasi-totalité des applications envisageables, ce devrait être un fichier dont les enregistrements seraient de format rigoureusement identiques, c'est-à-dire systématiquement comparables. Néanmoins, ces enregistrements doivent être extensibles, de façon qu'il soit possible d'y ajouter d'autres informations.

La création du fichier du travail a pour première étape la fixation du format des enregistrements. En d'autres termes, le chercheur doit décider des

informations qu'il estime nécessaire d'extraire du fichier de base, et de l'ordre dans lequel celles-ci doivent être enregistrées. La procédure à suivre pour cette opération dépend de la structure des enregistrements du fichier de base. Si chaque enregistrement de base a un format rigoureusement standard (réponses à un questionnaire, analyse de contenu systématique sans adjonction d'informations non codées), il doit suffire à l'utilisateur d'indiquer quelles zones de l'enregistrement de base doivent être reproduites, et dans quel ordre ; le système peut alors prendre totalement en charge la création du fichier de travail. Si chaque enregistrement de base est partiellement structuré selon une grille d'analyse de contenu, et comporte une zone "libre" contenant des informations spécifiques à l'unité considérée (appréciation globale de l'enquêteur ou de l'analyste), l'utilisateur doit pouvoir obtenir la consultation (sur écran, au coup par coup, ou sur listing pour l'ensemble des unités) d'une partie au moins du contenu du fichier de base ; il pourra ainsi estimer dans quelle mesure il prendra en compte les données non préalablement codées, et procèdera ensuite à leur codification enregistrement par enregistrement. Si chaque enregistrement est constitué par une séquence,de longueur variable, d'informations codifiées (biographies, récits, entretiens non directifs), le chercheur doit pouvoir faire exécuter par la machine la sélection et la condensation des informations dont il a besoin : dénombrement des fréquences d'apparition des thèmes, ou des fréquences d'enchaînement de deux thèmes donnés.

En règle générale, le chercheur doit avoir la possibilité :

- de sélectionner un ou plusieurs enregistrements de base, afin de les consulter (sur écran ou sur listing) ; les critères de sélection peuvent être soit les références de l'enregistrement (numéro ou nom de l'unité correspondante), soit la longueur d'une partie ou la totalité de l'enregistrement (richesse des informations enregistrées), soit une combinaison de caractéristiques des unités à analyser (pattern de réponses à un questionnaire, par exemple).

- de procéder par essais et erreurs (sur écran) à la constitution d'un enregistrement de travail à partir d'un enregistrement de base (sélection de variables, dénombrement de thèmes, calcul d'indices, recodification et combinaison de variables).

- de faire reproduire automatiquement par la machine, sur l'ensemble des enregistrements de base, la séquence d'opérations retenue pour la constitution du premier enregistrement de travail. Le système devrait alors fournir un diagnostic sur les cas d'impossibilité (informations manquantes, p. ex.).

- éventuellement, de constituer la totalité du fichier de travail manuellement, c'est-à-dire enregistrement par enregistrement. Une telle option risque d'impliquer soit l'abandon de fait du format standard pour les enregistrements de travail, soit l'obligation de prévoir un contrôle au cours, ou à l'issue, de la constitution du fichier de travail, pour détecter et signaler les omissions ou les adjonctions de variables d'un enregistrement à l'autre.

Les opérations de modification du fichier de travail sont peu différentes des opérations de création. Elles doivent permettre :

- la suppression de variables ou d'unités faisant partie du fichier de travail ;
- l'adjonction de variables ou d'unités extraites du fichier de base ;
- la permutation et le reclassement des variables (à l'intérieur des enregistrements), et des enregistrements (dans le fichier) ;
- l'adjonction de variables supplémentaires obtenues par combinaison d'informations existantes (variables composites), par calcul (indices), ou matérialisant l'intuition du chercheur (identification subjective de types, hypothèses relatives à des informations manquantes) ;
- l'adjonction au fichier d'unités fictives, construites à partir des unités existantes (passage à la limite sur certaines variables, valeurs moyennes, traits communs à un sous-ensemble d'unités), ou définies directement par le chercheur ;
- la sélection d'unités à partir de certaines de leurs caractéristiques.

c) Edition des résultats

Le système doit permettre au chercheur d'éditer sur imprimante les résultats de ses opérations, soit tels qu'il les a enregistrés dans le fichier résultat, soit après modification sur écran de leur présentation (mise en page, permutations, etc.), soit après une mise en forme automatique (graphe correspondant à une matrice binaire, p. ex.).

3.2.3. Les opérations sur les données.

Si l'on appelle E l'ensemble des enregistrements du fichier de travail, et I l'ensemble des indicateurs (variables, propriétés, attributs, indices) figurant dans chaque enregistrement, on peut classer les opérations sur les données en trois groupes :

- les opérations sur la "matrice des données", c'est-à-dire sur le produit cartésien E × I ;

- les opérations relatives aux unités, c'est-à-dire essentiellement sur E^2 ;

- les opérations relatives aux variables, c'est-à-dire sur I^2.

En outre, dans chaque groupe, on peut distinguer les opérations locales, portant sur un sous-ensemble de E × I, de E^2, ou de I^2, et les opérations globales concernant l'ensemble dans sa totalité.

a) Opérations sur la "matrice des données"

Matériellement, la "matrice des données" est un état des enregistrements de travail. Elle fournit, pour chaque unité considérée, la valeur prise par chacune des variables retenues.

Les opérations locales sur E × I sont essentiellement des opérations de comparaison, portant sur un petit nombre de lignes ou de colonnes de la "matrice". Les opérations portant sur les lignes (correspondant aux unités) ont pour but de rechercher :

- quelle est l'unité qui ressemble le plus à une unité donnée x ?

- quelles sont les deux unités qui ont le plus de ressemblance entre elles ?

- des deux unités x et y, à laquelle l'unité z ressemble-t-elle le plus ?

- comment maximiser la ressemblance entre x et y, en modifiant l'importance attribuées aux variables ?

- quels sont les traits communs aux unités x et y ?

- quelles sont les unités qui présentent simultanément tels ou tels traits ?

Ces opérations peuvent s'effectuer de deux manières : par examen visuel sur écran, ou par calcul. La première est la plus proche des comportements que nous avons observés. Elle suppose la possibilité de visualiser sur l'écran l'ensemble des caractéristiques d'une unité donnée ; par exemple sous forme de profil de réponses, d'histogramme, de chaîne de symboles graphiques, ou de bande d'intensité variable (cf. les travaux de sémiologie graphique). L'utilisateur doit avoir la possibilité de juxtaposer ou superposer les représentations de plusieurs unités présentées automatiquement selon leur place dans le fichier ou sélectionnées par lui, et de faire varier sur l'écran la nature des variables retenues (dont le nombre est limité par les dimensions de l'écran), l'ordre de présentation des variables, ainsi que l'amplitude correspondant à chaque variable. Les manipulations graphiques portant sur l'amplitude des réponses ont pour

effet de passer des profils en valeur absolue aux profils relatifs, et de modifier le poids attribué à chaque variable dans l'évaluation de la ressemblance entre unités. Evidemment, ces opérations peuvent également être effectuées sans support visuel, par calcul d'un indice de ressemblance. Dans cette option, seuls les résultats devront apparaître sur l'écran.

Les opérations locales portant sur les colonnes de la "matrice des données" concernent les relations entre les variables. Lorsque cela a un sens, il doit être possible de procéder aux comparaisons de profils de variables comme pour les profils d'unités. En outre, il est utile d'effectuer des dénombrements relatifs à deux variables (ou plus) considérées simultanément ("tableaux croisés" des dépouillement d'enquête), et de manipuler les tableaux ainsi obtenus (permutations de lignes ou de colonnes, regroupements, totalisations, division par une constante, etc.).

Les opérations globales sur $E \times I$ ont pour objectif de condenser la "matrice des données", en réduisant le nombre des lignes et/ou des colonnes. La réduction du nombre de lignes peut être réalisée par l'élimination d'unités jugées "déviantes" ou "inclassables", par la condensation d'unités "voisines" (en ne retenant que leurs valeurs centrales, p. ex.), ou par remplacement de celles-ci par un type abstrait. La réduction du nombre de colonnes peut être obtenue par l'élimination de variables jugées "redondantes" ou "non pertinentes", ou par leur remplacement par des "facteurs" abstraits.

Ces opérations peuvent être effectuées directement par le chercheur, lorsque les dimensions de la "matrice des données" lui permettent d'être visualisée. On peut également recourir soit à l'extension, à l'ensemble des unités ou des variables, d'opérations locales effectuées précédemment, soit à l'application de techniques statistiques usuelles, telles que l'analyse hiérarchique, l'analyse factorielle, l'analyse de la variance, et leurs dérivés.

b) Opérations sur E^2

La matrice carrée E^2 permet de résumer le résultat de comparaisons effectuées sur les unités à partir de E × I. Les exemples les plus courants sont les indications de ressemblance ou de distance entre les unités.

La constitution de E^2 peut être prise en compte directement par le chercheur lui-même, lorsque le nombre d'unités est suffisamment réduit. Il lui suffit d'enregistrer, pour chaque paire d'unités, leur degré de ressemblance (ou plus simplement l'existence ou non d'une ressemblance). La matrice de ressemblance (ou de distance) doit également pouvoir être constituée automatiquement

à partir d'un indice défini par l'utilisateur, soit a priori, soit au cours des opérations de comparaison dans E × I. En outre, il doit lui être possible de ne considérer éventuellement qu'un sous-ensemble de E, sélectionné par lui.

On doit pouvoir effectuer sur E^2 des opérations locales, telles que la recherche des paires d'unités les plus proches ou les plus dissemblables. Mais les opérations sur E^2 sont pour l'essentiel des opérations globales. Elles sont de trois types :

- comparaison et combinaison de matrices de ressemblance correspondant à des indices différents.
- transformation des valeurs d'indice de E^2. Les plus intéressantes nous paraîssent être le remplacement de la métrique sur E^2 par une ultramétrique, et le regroupement des valeurs d'indices en classes. Lorsque le nombre de classes se réduit à deux, on obtient une famille de matrices binaires, variant en fonction du seuil choisi comme limite entre les classes.
- recherche de configuration particulières dans le graphe correspondant à une transformation de E^2 : chaînes, composantes connexes, etc.

Ces opérations ont pour but de dégager des groupes homogènes d'unités, c'est-à-dire de réduire E^2. Afin de faciliter la comparaison entre plusieurs réductions possibles, l'utilisateur doit pouvoir construire des indices mesurant l'homogénéité intra et inter-groupe, et le pouvoir discriminant de la classification obtenue.

c) Opérations sur I^2

La matrice carrée I^2 résume les relations entre les variables. Elle peut être constituée à partir de relations définies par le chercheur, telles que l'incompatibilité mutuelle entre caractéristiques, l'implication, ou l'équivalence. Elle peut également être construite à partir d'une mesure de la liaison entre variables (coefficients de corrélation, d'association, de colligation, de contingence, etc.).

Les opérations locales sur I^2 sont comme précédemment des opérations de comparaison et de sélection. Les opérations locales peuvent avoir pour objectif :

- la comparaison et la combinaison de matrices obtenues à partir de relations ou de coefficients différents ;
- la recherche de configurations particulières : identification de *clusters* de variables, *path analysis*.

3.2.4. Mise en oeuvre du système.

Il est facile de dépeindre la manière d'utiliser le système que nous venons de présenter succinctement, en prenant pour points de départ les descriptions que nous avons données des procédures manuelles de construction de typologies (§ 2.2.1. à 2.2.3.). Ce faisant, il convient de se rappeler que les procédures ainsi formalisées constituent plutôt des styles d'approche des problèmes typologiques, que des algorithmes rigides s'excluant mutuellement. Par conséquent, le chercheur pourra passer de l'une à l'autre selon l'évolution de ses hypothèses et de sa connaissance des données. D'autre part, la consultation du fichier des opérations permet à l'utilisateur d'établir le lien entre les résultats obtenus et la démarche suivie ; cette réflexion sur la méthode devrait entraîner une plus grande souplesse d'utilisation du système, et une plus grande adéquation des résultats aux objectifs du chercheur.

a) La constitution de types-idéaux

Le recours au système informatique d'aide à la construction de typologies pour définir les types-idéaux n'a de sens que si le chercheur dispose de données d'observation. Dans cette hypothèse, le processus de base d'utilisation du système comporterait les phases suivantes :

- sélection a priori dans I d'un sous ensemble $J \subseteq I$ de variables importantes.

- examen, dans J^2, des relations liant les variables entre elles : sélection de variables orthogonales (non liées entre elles) ; recherche de *clusters* de variables correlées à l'une des variables orthogonales, et définition sémantique de la dimension correspondante ; éventuellement, sélection ou construction d'une nouvelle variable mieux adaptée à la dimension ainsi définie.

- passage à la limite sur les dimensions retenues ; définition des types correspondant à chaque configuration de valeurs extrêmes sur les dimensions ; sélection d'un petit nombre de types-idéaux parmi ceux-ci.

- adjonction des types-idéaux à la sous-matrice des données E × J.

- comparaison des unités observées aux types idéaux ; regroupement des unités les plus proches des types-idéaux, et des unités hybrides (à égale distance de deux types-idéaux).

- examen des unités non classées : recherche des traits communs à certaines d'entre elles, et comparaison avec les types extrêmes non retenus ; éventuellement, sélection parmi ceux-ci d'un nouveau type idéal.

- éventuellement, abandon d'un type idéal trop éloigné des unités observées.

- examen des groupes relativement purs et des groupes hybrides obtenus, en fonction de l'ensemble de leurs caractéristiques (dans E × I) ; évaluation

de la pertinence des types-idéaux par rapport aux données.

Naturellement, cette séquence peut être réitérée, entièrement ou partiellement, autant de fois qu'il est nécessaire pour obtenir des types-idéaux satisfaisants. Chaque réitération aura pour point de départ un changement dans les paramètres de la procédure : choix des variables prises en compte, définition des dimensions fondamentales, sélection des types extrêmes promus types-idéaux, etc. La condition minimale d'arrêt de la procédure pourra être par exemple l'obtention d'un petit nombre de types-idéaux tels qu'aucune unité observée ne possède moins de la moitié des traits caractérisant un type. Il faudra en outre que ces types soient sémantiquement pertinents, c'est-à-dire compatibles avec les théories du chercheur.

b) La réduction d'un espace d'attribut

Cette procédure, nous l'avons vu, comporte deux phases successives : une phase d'expansion maxima, et une phase de contraction maxima.

La première phase se ramène à une saisie des données aussi complète que possible, et à la constitution d'un fichier de travail tel que chacune des variables soit un attribut pouvant être présent ou absent dans chaque unité. Pour constituer le fichier de travail, le chercheur devra :

- décider des coupures à faire sur les variables continues, à partir de considérations théoriques et/ou de l'observation de la distribution correspondant à chaque variable ;
- transformer chaque variable originale, qu'elle soit mesurable, ordinale, ou nominale, en autant de variables dichotomiques qu'il distingue d'états sur la variable d'origine ; on obtient ainsi des attributs dont la présence ou l'absence permettent de décrire chaque unité. Cette transformation peut être automatisée. Les enregistrements de travail sont donc des vecteurs à valeurs dans {0,1}. Si le fichier de base comportait n variables ayant chacune m_i états, chaque enregistrement du fichier du travail compte $\prod_{i=1}^{n} m_i$ variables binaires.

La phase de contraction de l'hyperespace des observations fait appel aux relations d'exclusion mutuelle et d'implication, définies sur I^2 entre les indicateurs binaires. Soit un sous ensemble $J \subset I$ tel que, pour toute paire d'indicateurs $\{k,l\}$ éléments de J, ou ait k et l mutuellement exclusifs ; c'est-à-dire que, pour toute unité observée, on ait au plus un seul attribut de J qui soit présent. En première approximation, on considèrera J comme une dimension. (On vérifiera ensuite que, pour toute unité, l'un au moins des attributs de J est toujours présent). D'autre part, soit un sous-ensemble $K \subset I$, $K \cap J = \emptyset$;

soit $i \in J$ un état de la dimension J. Si l'on a pour tout $j \in J$ tel que $j \neq i$, et pour tout k élément de K, j et k mutuellement exclusifs ; si l'on a en outre, pour tout $k \in K$, k implique i (c'est-à-dire que toute unité possédant l'un des attributs de K possède aussi l'attribut i) ; alors on considérera que le sous-ensemble d'attributs K peut être substitué à l'attribut i, et que $(J-i) \cup K$ est une dimension. De telles opérations, aisément exprimables en termes d'opérations booléennes, conduisent à la détermination de nouvelles dimensions, définissant un hyperespace plus compact.

Ces opérations sont facilement automatisables. Il est indispensable qu'elles s'effectuent sous le contrôle du chercheur, afin que celui-ci donne, lorsqu'il l'estime nécessaire, la primauté aux considérations théoriques. Quand l'hyperespace a été contracté selon cette procédure, le chercheur a la possibilité soit de réduire cet hyperespace directement par regroupement d'états, et combinaison ou suppression de dimensions, au vu des distributions correspondant à ces dimensions ; soit de définir de nouvelles relations d'exclusion mutuelle et d'implication, en égalant à zéro les fréquences inférieures à un seuil donné. (Ainsi, on dira que deux attributs sont mutuellement exclusifs si moins de m unités les possèdent simultanément ; on dira qu'un attribut en implique un autre si moins de n unités possèdant le premier ne possèdent pas le second). En combinant ces deux stratégies au gré de ses exigences théoriques, le chercheur doit en principe aboutir à une réduction de l'espace d'attributs satisfaisante.

c) L'agrégation autour d'unités-noyaux.

Cette dernière procédure est certainement la plus complexe de toutes ; c'est celle qui impose le plus de manipulations et de tâtonnements, donc celle pour laquelle il devrait y avoir le plus de réitérations du processus de base. Celui-ci est relativement simple :

- au départ, le chercheur a sélectionné un petit nombre d'unités-noyaux : il compare, dans E × I, chaque unité non sélectionnée à chaque unité noyau, et rattache celle-ci s'il y a lieu à l'une des unités-noyaux. Ceci se traduit par l'addition, dans l'enregistrement correspondant, d'un code particulier. Eventuellement, il peut retenir une unité non précédemment sélectionnée comme nouvelle unité noyau.

- il définit ainsi dans E^2 une relation de ressemblance. Il peut rechercher, dans le graphe correspondant, les composantes connexes et les isolées.

- il peut également définir, à partir de comparaisons locales dans E × I entre les unités-noyaux, une mesure de ressemblance dans E^2, et chercher à agréger les unités en fonction de cette mesure de ressemblance.

Par ajustement progressif des paramètres et réitération de ce processus, le chercheur peut aboutir à une partition satisfaisante de E.

3.3. Perspectives d'application.

Nous venons d'esquisser les grandes lignes d'un système de manipulation de données orienté vers la constitution de typologies. Ce système ne devrait en principe réaliser que des opérations élémentaires extrêmement simples, et parfaitement claires pour l'utilisateur : opérations de classement, sélection, dénombrement d'unités ; recodification, combinaison, suppression de variables ; calculs arithmétiques usuels ; calculs logiques élémentaires ; identification de structures particulières dans un graphe. Ces opérations ne sont pas propres à la seule construction de typologies ; on les retrouve également par exemple dans les démarches qui ont pour but non plus de classer des unités, mais d'expliquer des phénomènes et de définir des relations de causalité entre des variables. C'est pourquoi il serait probablement facile de compléter le catalogue d'opérations que nous avons dressé, en vue d'en faire un outil moins étroitement spécialisé.

L'avantage majeur d'un système de *manipulation* de données sur les programmes d'*analyse* des données est qu'il préserve l'initiative du chercheur. Il requiert peu de connaissances logico-mathématiques ou informatiques préalables, et devrait être maîtrisé assez rapidement par l'utilisateur. En contrepartie, un tel système est peu puissant. Ce défaut originel peut être, pour l'utilisateur, l'occasion d'un apprentissage méthodologique adapté à ses possibilités : la réitération d'une séquence d'opérations s'apparente en effet à l'utilisation d'une macro-instruction, et rien n'interdit au chercheur de se forger des séquences d'opérations standard correspondant à ses besoins. Un outil de ce type peut ainsi permettre l'expérimentation et la mise au point d'algorithmes nouveaux, répondant mieux aux besoins des sciences sociales.

Par contre, pour réaliser un système de ce type, il faudrait résoudre quatre problèmes : la définition formelle complète du système, en liaison avec des chercheurs et praticiens en sciences sociales ; la programmation du système ainsi défini ; l'abaissement du coût d'utilisation du mode conversationnel ; la formation des utilisateurs potentiels. Ce dernier point nous paraît crucial. Nous l'avons dit, les connaissances requises sont minimes. Par contre, un tel système risque de bouleverser des habitudes de travail solidement acquises. L'observation de chercheurs manipulant leurs données met en évidence l'importance que revêt, pour la réflexion, le contact physique avec le matériel recueilli ou les résumés codifiés, ainsi que le rôle que joue, dans la maturation des hypothèses, la lenteur des opérations manuelles. Dans ce domaine, le passage à l'informatique implique peut-être le risque de perte du sens du concret, et de précipitation dans l'élaboration théorique. Ces écueils peuvent-ils être évités ? Seule une expérimentation d'un système de ce type auprès d'un petit nombre de chercheurs permettra de répondre à cette interrogation.

OUVRAGES CONSULTÉS

Nous donnons ci-après la liste des comptes-rendus de typologies que nous avons analysés (§ 3). Ne figurent dans cette énumération que les textes qui ont fait l'objet d'une diffusion officielle ; par conséquent, quelques rapports d'études, couverts par le secret commercial, n'ont pu être mentionnés.

D'autre part, nous avons également indiqué les ouvrages et articles relatifs à la méthode typologique dans les sciences sociales (§ 2), ainsi qu'à la procédure de classification en général (§ 1), que nous avons utilisés.

Enfin, nous donnons la liste de nos principales sources concernant les procédures de classification automatique (§ 4).

---

1. Textes généraux sur les problèmes de classification.

CANGUILHEM, Georges — *Le normal et le pathologique* (Paris, PUF, 1966).

DAGONET, François — *Le catalogue de la vie* (Paris, PUF, 1970).

DARWIN, Charles — *L'origine des espèces* (Verviers, Gérard et Cie, 1973), en particulier chapitre 13.

DURAND DE GROS, J.-P. — *Aperçus de taxinomie générale* (Paris, Alcan, 1899).

ENGELS, Friedrich — *Dialectique de la nature* (Paris, Editions sociales, 1968), en particulier l'introduction (Pp 29-46) et les Pp 216-223.

GUYENOT, Émile — *Les sciences de la vie aux XVIIe et XIIIe siècles* (Paris, Albin Michel, 1957), en particulier livre 1.

HEIDBREDER, Edna — "Etude de la pensée humaine" in : ANDREWS, T.G. *Méthodes de la psychologie* (Paris, PUF, 1952), chapitre 4.

LAMARCK, Jean-Baptiste — *Philosophie zoologique* (Paris, Union générale d'éditions, 1968).

LEEPER, Robert — "Cognitive Processes", in : STEVENS, S.S., *Handbook of Experimental Psychology* (New-York, Wiley, 1951), chapitre 19.

O'NEIL, W.M. — *Faits et théories* (Paris, A. Colin, 1972), 3ème partie.

WOODWORTH, Robert S. — *Psychologie expérimentale* (Paris, PUF, 1949), 2ème partie, chapitre 30.

2. Considérations théoriques sur les typologies dans les sciences sociales.

BARTON, Allen — "Le concept d'espace d'attributs en sociologie", in BOUDON, Raymond, et LAZARSFELD, Paul, *Le vocabulaire des sciences sociales* (Paris, Mouton, 1967).

BOUDON, Raymond — *L'analyse mathématique des faits sociaux* (Paris, Plon, 1967), chapitre 6.

CAPECCHI, Vittorio — "On the Definition of Typology and Classification in Sociology", *Quality and Quantity, 2* (1968), 1, 9-30.

FELDMAN, J., EL HOURI, M. — "A propos de classifications sociales" (G.E.M.A.S., sans date).

GRANGER, Gilles-Gaston — *Pensée formelle et sciences de l'homme* (Paris, Aubier-Montaigne, 1967), en particulier chapitres 4 et 5.

JOLLIVET, Marcel — "D'une méthode typologique par l'étude des sociétés rurales", *Revue Française de Sociologie, 6* (1965), 33-54.

MAHO, J., NETCHINE, G., ROLLE, P. — "Types, classes et objet en sociologie et en psychologie", *Epistémologie sociologique, 12* (2e semestre, 1971).

MAÎTRE, Jacques — *Sociologie religieuse et méthodes mathématiques* (Paris, PUF, 1972), chapitre 2.

RÉGNIER, André — *La crise du langage scientifique* (Paris, Anthropos, 1974), chapitres 8 et 12.

REUCHLIN, Maurice — *Les méthodes quantitatives en psychologie* (Paris, PUF, 1962), en particulier chapitre 5.

WEBER, Max — *Essais sur la théorie de la science* (Paris, Plon, 1965).

WEBER, Max — *Economie et Société* (Paris, Plon, 1971), 1ère partie.

Collectif : — "Classes, classification et sociologie", *Epistémologie sociologique, 1-5* (1964-68).

3. Exemples de typologies (non compris les études couvertes par le secret commercial)

ABBOUD, Nicole — "Les grèves et les changements de rapports sociaux", *Sociologie du travail, 15* (1973), 4, 428-439.

ADORNO, T.W., FRENKEL-BRUNSWIK, Else, LEVINSON, Daniel J., SANFORD, R. Newitt, *The authoritarian personnality* (New-York, Wiley, 1964).

BÉNASSY-CHAUFFARD, C., PELNARD, J. — "Loisirs de jeunes travailleurs, reflet ou compensation de caractères psychologiques ou sociaux", *Enfance, 11* (1958), 4-5, 381-405.

BENDIX, Reinhard, LIPSET, Seymour Martin *Class, Status, and Power* (Glencoe, Free Press, 1953), en particulier les textes de BENDIX et LIPSET, WEBER, SCHUMPETER, SOROKIN, et PARSONS.

BENOIT, Odile — "Statut dans l'entreprise et attitudes syndicales des ouvriers", *Sociologie du travail*, 1962, 230-242.

BEROUTI, Monique — *Animation urbaine et vie quotidienne* (Thèse de 3e cycle en Sociologie, Université René Descartes, Paris, 1976).

BLUM, Gerald S. — *Les théories psychanalytiques de la personnalité* (Paris, PUF, 1955).

BROWN, Roger W. — "Mass Phenomena", in : LINDZEY, G. (Ed.) , *Handbook of Social Psychology* (Reading, Addison - Wesley, 1954), chapitre 23.

CAZENEUVE, Jean — "A propos de la typologie des sociétés globales", *Revue de Psychologie des peuples, 17* (1967), 1, 39-46.

Centre National de la Cinématographie : *Cinéma français - Perspectives 1970* (numéro spécial du *Bulletin d'information*, février 1965), § 2.3.

CHAMPAUD, J. — "Genèse et typologie des villes du Cameroun de l'Ouest", *Cahiers de l'ORSTOM, Série Sciences Humaines, 9* (1972), 3, 325-336.

Commissariat Général du Plan d'Equipement et de la Productivité : *Modalité de répartition des fonctions de service entre les villes d'une région* (2 tomes ronéotypés, Paris, SEMA, 1966).

CORNU, Roger, LAGNEAU, Janina — *Hiérarchies et classes sociales* (A. Colin, 1969).

CORNU, Roger, MAURICE, Marc — "Revendication, orientation syndicale et participation des cadres", *Sociologie du travail*, 1970, 328-337.

CORNU, Roger — *Rapport d'activité* (CNRS, LEST, 1974), en particulier § 2 et 3.

DAHRENDORF, Ralf — *Class and Class Conflicts in an Industrial Society* (Londres, Rontledge and Kegan Paul, 1959).

DURKHEIM, Émile — *Le suicide* (Paris, PUF, 1960), livre 2.

EYSENCK, H.J. — *Les dimensions de la personnalité* (Paris, PUF, 1950).

FUGUITT, Glenn V. — "A Typology of the Part-time Farmer", *Rural Sociology, 26* (1961), 1, 39-48.

GURVITCH, Georges — *La vocation actuelle de la sociologie,* Tome premier : *La sociologie différentielle* (Paris, PUF, 1957), chapitre 5.

GURVITCH, Georges — *Etudes sur les classes sociales* (Paris, Gonthier, 1966).

HALBWACHS, Maurice — *Esquisse d'une psychologie des classes sociales* (Paris, Marcel Rivière, 1964).

HALBWACHS, Maurice — *Classes sociales et morphologie* (Paris, Éditions de Minuit, 1972).

HESTER, James J. — "A Comparative Typology of New World Cultures", *American Anthropologist, 64* (1962), 1001-1015.

LE BRAS, Gabriel — *Etudes de sociologie religieuse,* Tome second : *De la morphologie à la typologie* (Paris, PUF, 1956).

LEROI-GOURHAN, André — *L'homme et la matière* (Paris, Albin Michel, 1943).

LEROI-GOURHAN, André — *Milieu et techniques* (Paris, Albin Michel, 1945).

MARX, Karl — *La guerre civile en France* (Paris, Éditions Sociales, 1953).

MARX, Karl — *Le Capital* (Paris, Éditions Sociales, 1976), en particulier livre 3, 7e section.

MERTON, Robert K. — *Social Theory and Social Structure* (The Free Press of Glencoe, 1957), principalement chapitres 4, 5, 9.

MICHELAT, Guy, THOMAS, Jean-Pierre "Contribution à l'étude du recrutement des écoles d'Officiers de la Marine", *Revue Française de Sociologie, 9* (1968), 1, 51-70.

MICHELAT, Guy, SIMON, Michel "Classe sociale objective, classe sociale subjective et comportement électoral", *Revue Française de Sociologie, 12* (1971), 4, 483-527.

MOLES, Abraham "Notes pour une typologie des événements", *Communications, 18* (1972), 90-96.

PROPP, Vladimir *Morphologie du conte* (Paris, Seuil, 1970).

RAMBAUD, Placide "Village et urbanisation", *Études rurales,* 49-50 (1973), 14-32.

SHELDON, W.H. *Les variétés de la constitution physique de l'homme* (Paris, PUF, 1950).

SHELDON, W.H. *Les variétés du tempérament* (Paris, PUF, 1951).

STONE, P.A. "The Economics of the Form and Organisation of Cities", *Urban Studies, 9* (1972), 3, 329-346.

4. Méthodes de classification automatique.

BARBUT, Marc, MONTJARDET, Bernard *Ordre et classification* (Paris, Hachette, 1970).

BELSON, W.A. "Matching and Prediction on the Principle of Biological Classification", *Applied Statistics, 8* (1959), 2, 65-75.

BENZECRI, Jean-Paul et collaborateurs *L'analyse des données,* Tome 1 : *La taxinomie* (Paris, Dunod, 1973).

BERGONIER, H., BOUCHARENC, L. "Une méthode de segmentation basée sur la théorie de l'information", *Prix Marcel Dassault*, 1966.

BERNARD, G., BESSON, M.L. "Douze méthodes d'analyse multicritère", *Revue d'Informatique et de Recherche Opérationnelle*, *5* (1971), V-3, 19-64.

BERTIER, Patrice, LABBÉ, Bernard "La méthode de Belson", Note interne SEMA, mars 1966.

BERTIER, Patrice, BOUROCHE, Jean-Marie *Analyse des données multidimensionnelles* (Paris, PUF, 1975), en particulier chapitres 4, 6, 11 et 14.

BOUROCHE, Jean-Marie, TENENHAUS, Michel "Quelques méthodes de segmentation", *Revue d'Informatique et de Recherche Opérationnelle*, *4* (1970), V-2, 29-42.

BOUROCHE, Jean-Marie, TENENHAUS, Michel *Méthode de typologie sur variables hétérogènes* (COREF, note de travail n° 10, 1976).

CAPECCHI, Vittorio "Une méthode de classification fondée sur l'entropie", *Revue Française de Sociologie*, *5* (1964), 290-306.

CELLARD, J.C., LABBÉ, B., SAVITSKY, G. "Le programme Elisée - Présentation et application", *Metra*, *6* (1967), 3, 503-520.

DIDAY, E. "An Introduction to the Dynamic Clusters Method", *Metra*, *11* (1972), 3, 505-519.

DIDAY, E. "Optimisation en classification automatique et reconnaissance des formes", *Revue Française d'Automatique, Informatique, Recherche Opérationnelle*, *6* (1972), V-3, 61-96.

FERNANDEZ DE LA VEGA, W. "Techniques de classification automatique utilisant un indice de ressemblance", *Revue Française de Sociologie*, *8* (1967), 4, 506-520.

FERNANDEZ DE LA VEGA, W. "Quelques aspects du problème de la détermination automatique des classifications", *Quality and Quantity, 4* (1970), 1, 117-152.

GILLET, Bernard "La méthode typologique de Mc Quitty", *Informatique en Sciences Humaines*, 12 (1971).

LERMAN, I.C. *Les bases de la classification automatique* (Paris, Gauthier-Villars, 1970).

Mc QUITTY, L. "Agreement Analysis : classifying persons by predominant pattern of responses", *British Journal of Statistical Psychology, 9*, 5-16.

MOITRY, Jérôme "Ordinateurs et traitement de données", *Revue internationale des Sciences Sociales, 10*, (1971), 4, 103-136, et 5, 101-131.

MORGAN, J.N., SONQUIST, J.A. "Problems in the Analysis of Survey Data and a Proposal", *Journal of the American Statistical Association* 1963.

SANDRI, G. "On the Logic of Classification", *Quality and Quantity, 2* (1968), 1, 80-124.

SOKAL, R.R., SNEATH, P.H.A. *Principles of Numerical Taxonomy* (San Francisco, Freeman, 1963).

SOKAL, R.R. "Numerical Taxonomy", *Scientific American, 215* (1966), 6, 106-111.