ANALYSE DES DONNÉES

Tome I

Publication de l'A.P.M.E.P.

(Association des Professeurs de Mathématiques
de l'Enseignement Public)
N° 28

Si vous voulez savoir ce qu'est

l'Association des Professeurs de Mathématiques de l'Enseignement Public

voyez page 244.

Si vous voulez adhérer à l'A.P.M.E.P., lui commander des brochures, écrivez à :

Secrétariat de l'A.P.M.E.P.
37 rue Jacob, 75006 PARIS

ANALYSE DES DONNÉES

Tome I

Publication de l'A.P.M.E.P.

(Association des Professeurs de Mathématiques
de l'Enseignement Public)
N° 28

TABLE DES MATIERES

Ont collaboré à l'élaboration de cette brochure, outre les auteurs de chaque article :

A. CARLIER, Laboratoire de Statistique, Université Sabatier, Toulouse.
L. CARTER, Université de Paris X, Nanterre.
M.C. DAUVISIS, I.R.E.M. de Toulouse.
C. DENIAU, Université R. Descartes, Paris.
F. DUBAIL, I.R.E.M. de Lyon.
L. DUVERT, A.P.M.E.P.
C. LAVILLE, Université de Paris X, Nanterre.
J.P. LEMOAN, Université R. Descartes, Paris.
G. LENEZET, I.R.E.M. de Rennes.
G. MISON, I.R.E.M. de Lyon.
G. OPPENHEIM, Université R. Descartes, Paris.
P. SUBTIL, I.R.E.M. de Lyon.

INDEX TERMINOLOGIQUE

Cet index rassemble la plupart des mots techniques utilisés dans cette brochure et donne pour chacun d'eux le numéro de l'article et la page où ils sont définis ou utilisés le plus souvent. L'astérisque indique une notion qui sera étudiée en détail dans le Tome II.

P.L. HENNEQUIN
IREM de Clermont

INTRODUCTION

En entreprenant la publication d'une brochure consacrée à l'analyse des données, l'A.P.M.E.P. se proposait un triple but : présenter des techniques de calcul issues des mathématiques mais développées surtout depuis l'apparition des ordinateurs, montrer comment l'analyse des données peut être pratiquée avec des élèves de différents niveaux et en relation avec la plupart des disciplines, indiquer enfin comment la recherche didactique l'utilise pour analyser et présenter ses observations. Reprenons plus en détail chacun de ces trois points :

— Notre époque se caractérise par l'interrelation de collectivités ou sociétés de plus en plus importantes et techniquement évoluées. Ceci amène aussi bien les gouvernements que les spécialistes des mass media à recueillir des informations chiffrées de plus en plus nombreuses et variées. Une fois rassemblés, ces nombres doivent être analysés et présentés, tant aux responsables politiques qu'aux citoyens. La manipulation d'un grand nombre de données en vue de représentations synthétiques est maintenant rendue possible grâce à l'utilisation généralisée d'ordinateurs. Bien entendu, l'ordinateur n'est que l'exécutant d'algorithmes de calcul élaborés par le mathématicien. Connaître ces algorithmes, qui utilisent quelques concepts simples d'algèbre ou d'algèbre linéaire, est essentiel pour le citoyen qui veut maîtriser l'analyse et non se laisser dominer par ceux qui la lui présentent : quel est le rôle des techniques de calcul dans l'utilisation qui est faite des données ?

Une présentation schématique n'induit-elle pas une interprétation abusive ? (cf par exemple [3]). L'ordinateur n'est-il pas, par sa seule intervention, considéré comme la baguette magique sacralisant certaines données par ailleurs fort banales, voire entachées de nombreuses erreurs au moment de leur recueil ?

— La méthodologie de l'analyse des données comporte plusieurs étapes qui ne sont pas toutes d'activité mathématique, mais dont le déroulement complet doit être perçu pour une bonne maîtrise : délimitation d'un cadre d'études, définition des situations et variables observées, recueil des données, codage, élaboration de questions auxquelles les données recueillies doivent permettre de répondre, calculs, représentations graphiques, réponse aux questions posées mais aussi nouvelles questions qui se posent, nouvelles observations à effectuer pour infirmer ou confirmer une impression, une hypothèse. Il s'agit là d'une démarche expérimentale développant les facultés d'analyse et de

synthèse et l'esprit critique et par conséquent à laquelle il convient de familiariser de bonne heure le futur citoyen. Si nos élèves ont l'occasion de la pratiquer en Sciences Physiques ou Naturelles, il en est moins souvent ainsi en Sciences Humaines ou Sociales où les horaires font moins de place aux "activités dirigées". De toutes façons, il est important que le professeur de mathématiques participe, sinon comme maître d'œuvre, au moins comme co-responsable, à de telles activités qui permettent de faire fonctionner les concepts introduits dans le cours de mathématiques au moment où l'élève en éprouve le besoin et dans un contexte naturellement motivant. Il se peut d'ailleurs que les projets d'études suggérés par les élèves amènent la nécessité d'introduire de nouveaux concepts. Notre brochure fait une place au compte rendu de telles activités, pratiqués avec les horaires et les programmes actuels, tant dans le premier que dans le second cycle. On peut espérer que les nouveaux programmes de second cycle permettront de les développer et de les généraliser.

Nous avons déjà évoqué l'utilisation de l'analyse des données, tant pour maîtriser l'évolution de nos sociétés que pour dégager des lois en Sciences expérimentales. Mais les Sciences Humaines constituent un des domaines privilégiés de l'analyse des données et en particulier des techniques de classification et d'analyse factorielle. En 1904, Spearman a introduit en Psychologie la notion de "facteur" pour dégager le concept d'intelligence d'une série d'aptitudes et de réussites observées, et si cette étape est aujourd'hui dépassée, elle reste importante historiquement (cf. [1]). Les Sciences Humaines n'ont pas en effet encore atteint le niveau de formalisation des Sciences "exactes" et avant de vouloir y dégager des lois, il faut, de la foule des observations accumulées, essayer d'extraire quelques variables "indépendantes" susceptibles de caractériser les individus ou les comportements. Travail gigantesque, que l'analyse des données ne saurait effectuer à elle seule, mais pour laquelle elle fournit des modes de représentation standardisés permettant la communication entre chercheurs et la comparaison des expériences,et apportant un éclairage nouveau. Son rôle est en cela comparable à celui du microscope en biologie : fabriquer un bon microscope, définir ses caractéristiques relèvent de l'opticien, utiliser ce microscope pour reconnaître telle ou telle pathologie cellulaire relève du biologiste ; analyser des données, c'est fournir une représentation de ces données à l'utilisateur ; c'est à lui de les interpréter sous sa seule responsabilité. Il nous a semblé intéressant, dans une brochure destinée avant tout à des enseignants de mathématiques, de montrer sur des exemples comment l'analyse des données était utilisée dans des recherches de didactique des mathématiques. L'utilisateur est alors un mathématicien (travaillant en général en équipe avec un psychologue). La didactique des mathématiques pose en effet au chercheur des problèmes d'évaluation quand l'évaluation n'est pas l'objet même de son étude. Les exemples traités concernent aussi bien des élèves du premier cycle que des élèves entrant à l'Université. On trouvera d'autres exemples relatifs au premier degré dans [5] et [6].

Cette brochure n'est pas un traité d'analyse des données car il en existe déjà de nombreux en langue française que nous avons cités dans la bibliographie générale à la fin de ce volume. L'ensemble des articles recueillis dépassant largement le volume admis pour une brochure de l'A.P.M.E.P., nous avons décidé de la publier en deux tomes. Plusieurs partis pouvaient être retenus pour ce découpage : traiter dans un premier volume les méthodes et la théorie, laissant au second le développement des exemples d'application, ou bien regrouper dans le Tome 1 toutes les méthodes purement algébriques de classification reposant sur le calcul d'un distance ultramétrique et conduisant à une hiérarchie de découpages de l'ensemble des individus ou des variables, laissant pour le Tome II les méthodes d'algèbre linéaire représentant les données par un nuage dans un espace euclidien puis celui-ci par ses projections sur des plans convenablement choisis. Nous avons pris un parti intermédiaire permettant au lecteur pressé de se contenter de la lecture du premier tome. Celle-ci ne suppose comme connaissances mathématiques que celles d'un bachelier scientifique, y compris le vocabulaire de la statistique descriptive et de l'initiation aux probabilités. Ce tome contient tous les exemples nécessaires à sa compréhension et même l'utilisation de certaines techniques qui ne seront développées que dans le Tome II : un élève de sixième utilise bien en Sciences Naturelles un microscope dont il ne comprendra le fonctionnement optique au mieux qu'en seconde ou première. Le Tome II développera d'une part la richesse du point de vue euclidien rattachant géométrie et statistique par l'utilisation d'outils issus de la mécanique tels que centre de gravité ou inertie, d'autre part le développement détaillé d'applications empruntées à la didactique des mathématiques ; les outils mathématiques utilisés dans le Tome II (diagonalisation des matrices symétriques) font partie du programme du DEUG.

On trouvera à la fin de ce volume une table des matières et un bulletin de souscription pour le Tome II.

Cette brochure résulte d'un travail d'équipe. Une vingtaine de collègues y ont travaillé pendant plusieurs années, les uns en écrivant un ou plusieurs articles qui la composent, les autres en lisant, critiquant, relisant les versions successives jusqu'à leur élaboration définitive. Telle qu'elle est, et devant la personnalité affirmée de ses auteurs, elle peut sembler une mosaïque, voire un manteau d'arlequin. Nous avons tenté d'unifier la présentation et le vocabulaire mais chaque article peut se lire indépendamment des autres et par conséquent, le lecteur choisira lui-même le mode de parcours qui répondra le mieux à son goût.

Donnons quelques indications pour faciliter ce cheminement. La brochure commence par un article sur "données et codages" qui nous semble fondamental pour qui veut s'initier à l'analyse des données : négliger les problèmes qu'il soulève, ignorer les embûches qui guettent le néophyte serait bâtir sur le sable, quelle que soit l'exactitude des calculs faits ensuite et la puissance de l'ordinateur utilisé pour les mener à bien.

Le second article donne un exemple de travaux effectués dans une recherche INRP-IREM, dans des classes du premier cycle ne bénéficiant d'aucunes conditions particulières. Le lecteur intéressé pourra consulter pour plus de détails le Compte rendu de la recherche, publié par l'I.N.R.P., ainsi que les 23 dossiers disponibles dans les IREM.

Le troisième article, écrit par un collègue géographe, est exemplaire d'un travail interdisciplinaire. Bien entendu, la mise en œuvre par des élèves de lycée, pour l'enseignement de la géographie, de techniques élaborées, suppose l'accès à l'ordinateur, qui n'est encore possible que dans les établissements privilégiés mais s'étendra rapidement, nous l'espérons.

Avec les articles **4** et **5**, on rentre dans le détail de l'exposé mathématique des méthodes de classification les plus classiques : classifications hiérarchiques avec tous les choix successifs qu'elles supposent et méthode des nuées dynamiques particulièrement performante pour dégrossir la classification d'ensembles très vastes.

L'article **9**, largement indépendant des autres, pose, sur un exemple simple de pesées, un problème fondamental : comment organiser une série de mesures pour obtenir le maximum d'informations sur la quantité mesurée pour un coût minimum ?

L'article **7** présente un exemple d'application de diverses techniques d'analyse des données : classification hiérarchique, analyse factorielle des correspondances, à l'étude des attitudes d'élèves du premier cycle face aux mathématiques (questionnaire utilisé pour l'expérience O.P.C.)

Avec l'article **8**, les techniques probabilistes s'introduisent pour résoudre les deux problèmes de l'analyse discriminante : comment classer les objets d'une population par la mesure de quelques variables seulement ? Comment rattacher un objet à une des classes définies dans la population ? Les exemples sont ici empruntés à la biologie.

L'article **9**, largement indépendant des autres, pose, sur un exemple simple de pesées, un problème fondamental : comment organiser une séries de mesures pour obtenir le maximum d'informations sur la quantité mesurée pour un coût minimum ?

Enfin, le dernier article (**10**), donne une liste de documents statistiques officiels et les adresse des organismes qui les diffusent, ce qui permettra aux collègues qui le souhaitent de les obtenir facilement.

Chaque article possède sa propre bibliographie mais le volume comporte in fine une bibliographie générale. Les références bibliographiques sont données entre crochets [ ] pour la bibliographie propre à l'article et avec la mention B.G. pour la bibliographie générale.

Un index terminologique placé en début d'ouvrage indique les principaux termes techniques rencontrés et, pour chacun d'eux, la page où il est défini.

Nous n'avons pas eu la place de faire ici une histoire détaillée de l'analyse des données ; nous renvoyons à [1] qui en fait un panorama complet. Nous signalons enfin [2] et [4] qui constituent deux bons articles d'introduction à notre sujet.

Ecrite pour les membres de l'Association par des membres de l'Association, cette brochure s'est voulue utile ; nous attacherons du prix à recevoir les critiques et les suggestions de tous ceux qui nous liront, fût-ce partiellement.

BIBLIOGRAPHIE

[1] BENZECRI (J.-P.) : *Histoire et préhistoire de l'analyse des données* Cahiers de l'analyse des données, DUNOD, 1976-II,2,3,4:III.

[2] BOUROCHE (J.M.) et SAPORTA (G) : *L'analyse des données.* Pour la science n° 5, mars 1978, p. 23.

[3] BOURSIN (J.L.) : *Sondages, indices, statistiques, la forme scientifique du mensonge ?* TCHOU, 1978.

[4] DIDAY (E) et LEBART (L) : *L'analyse des données.* La recherche, n° 74, janvier 1977, p. 15.

[5] I.R.E.M. de Bordeaux : *Regard sur la numération.* Bulletin A.P.M.E.P. n° 321, p. 788.

[6] VINRICH (G) : *Dépendances didactiques.* Bulletin A.P.M.E.P. n° 319, p. 377.

la collection MOTS

L'Association des Professeurs de Mathématiques de l'Enseignement Public a entrepris de publier une série de brochures, intitulées MOTS, contenant des réflexions sur quelques mots-clés utilisés en mathématique à l'Ecole Elémentaire :

égalité ; exemple et contre-exemple ; couple ; relation binaire ; nombre naturel ; entiers et rationnels ; nombre décimal, nombre à virgule ; fraction ; ensembles de nombres (Mots I, brochure 1974) ;

représentations graphiques ; application, fonction, bijection ; partition équivalenee ; partages ; divisibilité ; division euclidienne ; division (Mots II, brochure 1975) ;

numération ; opération et loi de composition ; propriétés des lois de composition ; congruences ; ordre ; préordre ; propriétés des relations binaires dans un ensemble ; dictionnaires, naturels, décimaux et ordres (Mots III, brochure 1976).

Chaque rubrique est détachable ; les feuilles, de format 15 X 21, sont perforées.

MOTS est une oeuvre collective ; l'équipe de rédaction, bénévole, constituée d'instituteurs, IDEN, professeurs (d'Ecole Normale, du Second Degré, du Supérieur) soumet ses projets à de nombreux instituteurs ; leurs avis lui sont précieux, surtout quand ils émanent de bacheliers littéraires qui n'ont pas eu l'occasion d'activité mathématique depuis leur sortie du lycée ou de l'école normale.

Sans être un manuel de mathématique, ni un lexique, MOTS permet au lecteur, à propos du vocabulaire rencontré dans les manuels scolaires ou les documents de formation permanente, de faire le point sur son évolution, sur les concepts et les idées qui s'y rattachent, et sur les notations utilisées.

Ces brochures, qui s'adressent aux enseignants, non aux élèves, sont vendues par l'APMEP aux prix suivants :

MOTS I : 100 pages - **10 F** (avec port : **14 F**)
MOTS II : 108 pages - **10 F** (avec port : **14 F**)
MOTS III : 136 pages - **12 F** (avec port : **16 F**)
MOTS IV : 152 pages - **12 F** (avec port : **16 F**)

F. PLUVINAGE
Université Louis Pasteur, Strasbourg

1. DONNEES ET CODAGES

1. Qu'est-ce que des données ?

DONNEES, adj. pris subst. terme de Mathématique, qui signifie certaines choses ou quantités, qu'on suppose être données ou connues, et dont on se sert pour en trouver d'autres qui sont inconnues, et que l'on cherche. Un problème, ou une question, renferme en général deux sortes de grandeurs, les données et les cherchées, data et quaesita. Voyez PROBLEME, etc. Euclide a fait un Traité exprès sur les données ; il se sert de ce mot pour désigner les espaces, les lignes et les angles qui sont donnés de grandeur, ou auxquels on peut assigner des espaces, des lignes ou des angles égaux.

Ce mot, après avoir d'abord été en usage dans les Mathématiques, a été ensuite transporté dans les autres Arts, comme la Philosophie, la Médecine, etc. on s'en sert dans ces sciences pour désigner les choses que l'on prend pour accordées, sans avoir la preuve immédiate de leur certitude, mais simplement pour servir de base aux raisonnements : c'est aussi pour cette raison que, dans les ouvrages de Physique, on appelle quelquefois data, données, les choses connues, par le moyen desquelles on parvient à la découverte des choses inconnues, soit dans la Philosophie naturelle, soit dans l'œconomie animale, soit dans l'opération des remèdes. Voyez DEMANDE.

La définition ci-dessus est celle donnée par d'Alembert dans l'Encyclopédie de Diderot. De nos jours, si un ouvrage comparable, comme l'Encyclopedia Universalis, consacre des articles à l'analyse des données, on n'y trouve en revanche pas de définition analogue. C'est que la réflexion sur la nature des données se situe bien souvent **en amont**

des préoccupations du praticien des analyses. Et pourtant l'analyse des données met fréquemment en relation ce praticien avec des spécialistes d'autres disciplines (médecine, botanique, linguistique, etc.). Un dialogue est donc nécessaire, qui demande quelques connaissances communes. Notamment, un accord sur la nature des données recueillies évite :

1° des illusions sur les possibilités des analyses,
2° des erreurs non rectifiables parce qu'antérieures à tout traitement.

Même dans le cas où recueil des données et traitements sont le fait de la même personne, il existe des possibilités de commettre des erreurs "classiques" (exemple : faire des moyennes sur des variables qui ne sont pas quantitatives), évitables par la connaissance des diverses natures de données, ou de négliger des traitements simples qui permettraient d'obtenir à bon compte certains des renseignements souhaités. C'est pourquoi les quelques pages qui suivent ont été rédigées pour présenter un contenu accessible et utile (modestement certes) à de nombreuses catégories d'utilisateurs de données, autrement dit vous et moi dans beaucoup de nos activités.

Un examen attentif de la définition de d'Alembert donne lieu à de premières réflexions. D'Alembert distingue les situations mathématiques, où "données" renvoie à "problème", et les situations expérimentales, où "données" renvoie à "demande", qui, à l'époque et dans ce contexte, signifiait postulat ou hypothèse. Le cas d'un problème mathématique est clair : les données font partie de l'énoncé. A propos de situations expérimentales, on peut remarquer l'emploi de deux mots par d'Alembert : le verbe *prendre* (... les choses que l'on prend pour accordées ...) et le pronom *on* , employé huit fois. Le verbe "prendre" nous invite à une attitude critique vis-à-vis de données du domaine expérimental. Elles ne tombent pas d'en haut, mais ont été prises, choisies, sélectionnées. C'est dans leur traitement qu'elles méritent vraiment leur nom, mais pas a priori. Et il peut être intéressant de se demander si les données qui ont été prises ne reflètent pas imparfaitement ou mal les phénomènes auxquels elles se rapportent. On connaît par exemple le cas des **artefacts** : il y a artefact lorsqu'un résultat apparent ne provient pas des phénomènes examinés, mais est "fabriqué" par l'observateur ou l'expérimentateur, soit à cause de la procédure d'investigation utilisée, soit à cause de la nature des données recueillies.

Exemple d'artefact.

Dans un article intitulé "A propos des évaluations en mathématiques : Du questionnement à l'interprétation et au diagnostic", J. Adda cite l'exemple suivant d'un artefact, dont les homologues ne sont nullements exceptionnels. Il s'agit d'études sur l'absentéisme. Statistiques à l'appui, on a pu accuser les femmes de pratiquer davantage d'absentéisme que les hommes. Ceci jusqu'au jour où le secrétariat d'Etat à la

condition féminine fit procéder à une étude prenant en compte le niveau de responsabilité. On s'aperçut alors qu'à niveau de responsabilité égal, il n'y avait pas d'écart entre l'absentéisme féminin et masculin. En revanche, il y a corrélation nette entre taux d'absentéisme et niveau de responsabilité. Le premier résultat apparent avait en fait sa source dans l'inégalité d'emploi entre les femmes et les hommes.

Le pronom "on" met l'accent sur le rôle humain, en indiquant tout à la fois :

1° que l'homme intervient là où il y a données,

2° que ce rôle humain n'est en principe pas individualisé.

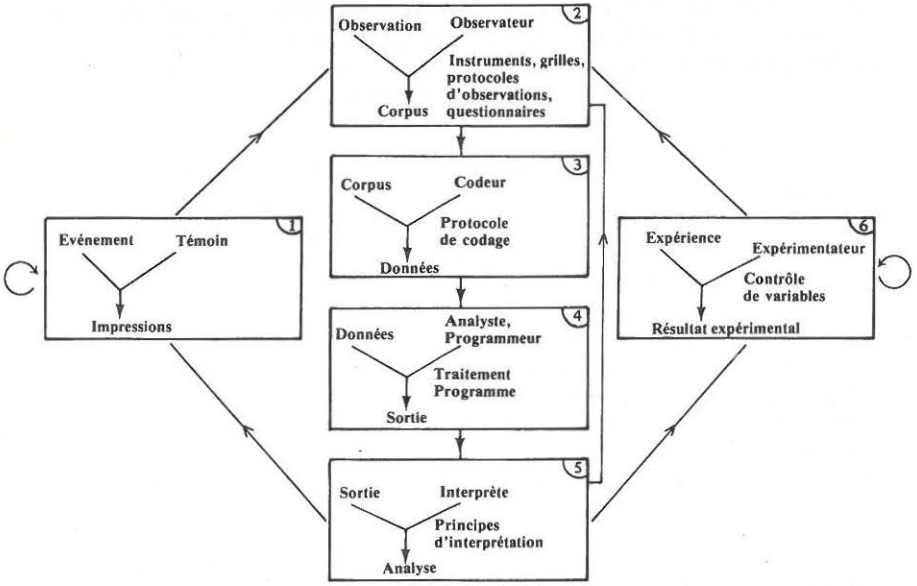
Un exemple élémentaire permet d'éclairer ce propos : Nous ne serions que médiocrement satisfaits d'observations de la température extérieure rédigées comme suit : "Ce matin, c'est monsieur G. Pachaud qui a fait l'observation de la température. Après être sorti quelques instants sur son balcon, il a déclaré supporter son pull-over". Pour que l'information sur la température soit plus pertinente, on a plutôt convenu à la fois d'un protocole d'observation (température relevée sous abri à telle heure) et d'une échelle de température (ce n'est d'ailleurs que très récemment que l'accord mondial s'est réalisé sur les degrés centigrades ou sur les degrés Kelvin selon les cas). Ainsi, ce qui est transmis à l'auditeur ou au lecteur n'est pas une information sur la sensation perçue par un individu particulier, mais une information que quiconque, placé dans les conditions ad-hoc, eût obtenue.

Note pour lecteurs minutieux : En fait, la pratique expérimentale actuelle est d'une prudence telle qu'elle n'ose pas aller aussi loin. Ce qui est véritablement **garanti** n'est en général pas l'obtention par quiconque, mais seulement une très forte probabilité d'obtention par quiconque d'un résultat tombant dans une fourchette donnée.

2. La collecte et la circulation d'informations en sciences expérimentales

Le schéma suivant résume une circulation de l'information. Dans chaque case, on voit apparaître : en haut à gauche le nom attribué au phénomène pris en compte, en haut à droite le nom attribué à la personne qui en rend compte, et en bas le nom attribué à l'information transmise. Le cas échéant, les règles à suivre sont mentionnées à côté des flèches.

COLLECTE ET TRAITEMENT D'INFORMATION
EN SCIENCES EXPERIMENTALES



Une même personne peut très bien, pour une étude déterminée, s'octroyer plusieurs des rôles figurant dans le schéma. Dans ce cas, elle aura, à un moment donné, à observer les règles de son rôle de l'instant. On voit qu'il existe de telles règles dans les cases autres que celle concernant les témoins d'événements : ce sont ces règles qui sont garantes de la qualité de l'information obtenue et transmise. Sans ces règles, les plus fermes intentions d'objectivité ne peuvent avoir que des effets très limités.

On peut remarquer, avant d'examiner de plus près les cases du schéma, que le désir de s'entourer de certaines garanties conduit, dans d'autres domaines que les sciences expérimentales, à appliquer des procédures voisines de celles indiquées par le schéma. Prenons, même si cet exemple n'est pas le plus sympathique, le déroulement d'un examen.

Notre examen comporte des commissions d'élaboration des sujets, une surveillance, des correcteurs, un jury qui peut se contenter de rassembler les résultats ou peut aller jusqu'à publier un rapport. Les commissions et la surveillance fonctionnent comme l'observateur de la case (2); le corpus est composé ici des copies des candidats ; correcteur est à rapprocher de codeur ; enfin le jury peut se contenter de faire "tourner" un traitement très simple : faire la somme des notes de chaque candidat, ou faire plus et être alors à rapprocher de l'interprète de la case (5). Toutes les similitudes que nous venons d'indiquer ne sont pas fortuites : les résultats d'un examen doivent apparaître comme entourés de garanties, tout comme des informations de nature scientifique (savoir si ceci

est toujours réalisé en pratique n'est pas notre propos). Lors d'un enseignement, un professeur peut souhaiter obtenir des informations qu'il n'oriente pas trop (même à son insu), il peut alors lui-même s'entourer de certaines des garanties indiquées ci-dessus. Très souvent, on ne pense pour ce faire qu'à l'évaluation du "niveau" des élèves, mais on néglige d'autres possibilités (évaluation formative, durées d'apprentissages, intérêts des élèves, ...).

En psychologie, la technique des entretiens est à ranger dans la case ② *(Observation)* : ce sont les contraintes que s'imposent les psychologues qui transforment des conversations en entretiens. Ces contraintes sont explicitement décrites dans les protocoles d'entretien. Notons que ces protocoles n'ont pas pour but de "déshumaniser" le praticien, mais simplement d'éviter que ses comportements soient l'objet de fluctuations trop grandes et surtout non repérables après coup. Pensons aussi au diagnostic d'un médecin : celui-ci peut être établi avec plus ou moins de finesse par l'un ou l'autre praticien, selon leurs qualités d'observation et de contact humain, et leur expérience personnelle ; mais les symptômes repérés par un praticien devraient en principe être **repérables** par tout autre se livrant au même examen. On voit qu'éviter l'individualisation dans l'observation n'est pas déshumaniser : par exemple, un observateur attentif, et peut-être aussi doué, remarquera des phénomènes qui auraient échappé à beaucoup (sinon à tout autre) ; mais **après** lui, il y aura accord sur l'existence des phénomènes indiqués.

Cet accord n'est toutefois pas sans limitations. Pour préciser, il convient d'indiquer les trois volets caractéristiques de la notion de phénomène scientifique :

1°) une problématique
2°) un protocole
3°) un résultat

LA PROBLEMATIQUE, c'est l'ensemble des questions que l'on se pose compte tenu des hypothèses que l'on admet ; elle fait intervenir les buts que l'on poursuit et comporte ce dont on accepte la mise en cause par rapport à ce que l'on considère comme intangible.

LE PROTOCOLE consiste en l'ensemble des outils et méthodes utilisés avec leur mode d'emploi, décrits de manière à rendre possible une reproduction ultérieure. Qu'une modification porte sur la problématique, le protocole ou bien entendu les résultats, c'est une modification du phénomène envisagé.

Après ces considérations, intéressons-nous de plus près aux contraintes qui régissent la collecte d'informations, concrétisée par l'obtention d'un corpus.

> ***Question** : Une bande de magnétophone peut être un exemple de corpus. Nous avons parlé ci-dessus de paquets de copies comme exemple de corpus. Citer d'autres exemples de corpus.*

3. Contraintes : cas des corpus

Le corpus est une **trace permanente** de l'observation. Celle-ci a souvent un caractère fugitif, et il s'agit alors de la **fixer**. On demande donc à un corpus de conserver fidèlement une **partie** du phénomène observé. Par exemple, une bande de magnétophone ne conserve aucun élément **visuel** ; c'est un bon corpus pour un entretien car, en revanche, elle restituera bien toute la partie sonore de l'observation ; en général c'est un mauvais corpus pour l'observation du déroulement d'une classe, à cause du phénomène de brouhaha. Pour ce qui est des copies ou des questionnaires remplis, ils constituent un corpus qui ne restitue pas l'**ordre** temporel de l'écriture ni les brouillons préalables ; là aussi il y a bien une perte. Mais les réponses finales sont restituées pourvu que les copies ne soient pas illisibles. La question à nous poser devant un corpus est :

"Quel est précisément le phénomène que le corpus restitue ?"

On s'aperçoit ainsi que si un corpus est constitué par le paquet des bulletins scolaires d'une population d'élèves, ce qui est contenu dans le corpus est le résultat d'une occurrence simultanée d'élèves, de professeurs et d'épreuves posées en situation scolaire. A partir de ce corpus, il est donc impossible de restituer les seules réactions des élèves devant les sujets scolaires qui leur ont été posés ; ainsi, comparer ces réactions entre des populations scolaires de même niveau espacées mettons de 10 ans (par exemple les élèves de seconde en 1965 et en 1975) serait impossible au vu de ces bulletins. Pour que ce soit possible, il faudrait que les professeurs se soient soumis à des contraintes assurant l'invariabilité de leurs critères à 10 ans d'écart : le corpus pourrait être constitué de réponses d'élèves à des épreuves normalisées. Remarquons qu'une fois faite une observation, le seul moyen d'accès ultérieur à cette observation est le corpus. Devant un corpus non satisfaisant, une seule conduite est possible : fabriquer un nouveau corpus, selon de nouvelles contraintes.

Exemple :

Dans certains ouvrages de psychopédagogie, on présente des "grilles" d'observation des échanges verbaux professeur-élèves. Ces "grilles" contiennent des questions du type :

Le professeur accepte-t-il les initiatives des élèves ?

☐	☐	☐	☐
souvent	parfois	rarement	jamais

Une "grille" ainsi constituée de telles questions, remplie par un "observateur", constitue un corpus qui renvoie à l'occurrence de "l'observateur" et du phénomène pris en compte (les échanges verbaux dans la classe). Ou bien donc on est renvoyé à la case ① *(impressions d'un témoin)*, ou bien on considère qu'il s'agit d'un corpus mais dans

lequel le soi-disant "observateur" est en fait partie intégrante du phénomène (par exemple s'il s'agit de confronter les impressions de plusieurs "observateurs" de la même classe). Le véritable observateur ne pourra donc être qu'une tierce personne.

Au contraire, on ne peut faire les mêmes remarques à propos du travail d'un groupe IREM sur la communication, présenté dans le compte rendu du colloque EVA *(Bulletin spécial Inter-IREM n° 12).* Pour la même observation, ce groupe propose de représenter chaque **phrase** dite par un rond *(professeur : ○ , élève : ●).* Ce rond est diversement barré selon que la phrase est une assertion, une question, un ordre ou un jugement de valeur. De plus, un rond est placé à droite du précédent si la phrase correspondante ne contient que des informations déjà dites et en-dessous s'il y apparaît un élément nouveau. Le corpus ainsi obtenu est, pour des raisons évidentes, appelé un chapelet. Certes, il existe quelques cas ambigus à l'écoute (faut-il à tel moment compter une ou deux phrases), mais ils sont l'exception. De plus les diverses catégories de phrases correspondent à des différences d'intonation (marquées aussi par la ponctuation dans les textes écrits), sauf ordres et jugements de valeur *(ponctuation : " ! ")*, mais l'un détermine un acte ultérieur *(exemple : "Viens !")* tandis que l'autre renvoie au passé *(exemple : "Bien !")* au moins momentané. Les différences sont donc assez nettes pour que l'on puisse ici estimer qu'un *"chapelet"* est un corpus produit par un véritable observateur. A titre d'essai, plusieurs membres du groupe avaient simultanément relevé de cette façon la même séquence scolaire ; la coïncidence des "chapelets" obtenus avait été tout à fait satisfaisante.

Le développement, peut-être un peu long pour ce propos, d'un exemple est justifié par le fait suivant : trop fréquemment, une étude que souhaiterait faire une personne ou un groupe est tout à fait irréalisable sur le corpus obtenu. Et alors aucun codage ni aucun traitement de données ne peut rétablir la situation. Répétons-le : Dans un tel cas, il n'y a qu'à recommencer à zéro.

> ***Question.*** *D'après ce qui précède, il semblerait que le défaut à craindre est de croire confectionner un corpus alors que l'on est en train de livrer les impressions d'un témoin. A contrario, citer des exemples d'objets qui constituent des corpus valables, alors qu'ils paraissent être des impressions de témoins défectueuses ou même tendancieuses.*

4. Contraintes : des corpus aux données

La plupart du temps, un corpus présente un défaut majeur pour qui voudrait l'étudier directement : il n'est accessible ni **rapidement**, ni **commodément**. De plus, dans ce qu'il présente, le tri n'est pas fait parmi les possibilités de **généralisation** du cas particulier envisagé.

C'est pourquoi les données vont fournir du corpus une **image organisée**. Parler d'images, c'est parler d'applications. Et en effet, ce qui est nommé PARAMETRE, ou VARIABLE, est une application dont l'ensemble de départ est formé d'objets mathématiques attachés à un corpus et l'ensemble d'arrivée est le plus souvent un ensemble numérique ({0 ; 1}, une partie de **N**, de **Z** ou de **R**). D'une façon générale, on peut considérer que des données sont une famille de telles applications.

[*Remarque* : Les fonctions **coordonnées** permettent de ramener à ce cas les applications dans des espaces de dimension supérieure à 1].

Une telle façon de voir est beaucoup plus contraignante qu'il n'y peut paraître si l'on n'y prend pas garde. En effet, il ne s'agit de rien d'autre que de la **modélisation**, au moins locale.

Exemple : Le cas le plus simple est celui où l'on distingue dans le corpus des entités distinctes. Ainsi dénombrer les voyageurs dans une étude sur les transports ne soulève pas de difficulté **de principe.** Mais pour certaines études, ce n'est pas la variable nombre de voyageurs qui est considérée comme pertinente, c'est la variable nombre de "voyageurs-kilomètres". Dans ce dernier cas, le rôle de la modélisation est évident. Pensons maintenant au transport des marchandises. Il n'y a plus d'entités aussi évidentes que les voyageurs. Laissons au lecteur le soin de chercher dans ses souvenirs de lecture ou de documentation les différentes variables auxquelles on a recours dans les études sur ce sujet.

Mais, dira-t-on, si les variables sont déjà tributaires de modèles (même implicites), tout le travail sur les données est déjà fait a priori. A cela, on peut répondre oui et non. Oui parce que les données sont dès le départ porteuses des informations auxquelles aboutira leur traitement. Et non, précisément parce qu'au départ on est dans l'ignorance de ces informations. N'oublions pas le caractère local des modèles dont sont issues les variables : l'architecture de la famille des variables relatives à un corpus est, elle, inconnue ou mise en doute. Repensons à la définition de d'Alembert, pour le cas mathématique : dans les données d'un problème mathématique est déjà "inscrite" la solution (ou les solutions, ou l'absence de solution), "y a qu'à" raisonner. Dans les données d'un corpus sont déjà inscrites leurs relations, "y a qu'à" traiter. Pas de miracle par conséquent : nous ne tirerons rien de plus des données que ce qui y est ; en revanche les traitements dont il est question dans cet ouvrage ont pour objectif idéal d'en tirer tout ce qui y est et rien que ce qui y est.

Exemple :

Considérons un questionnaire constitué d'un certain nombre d'exercices de calcul proposés à des élèves. Pour chaque exercice, nous supposons qu'une analyse de contenu **nous** a permis de privilégier un résultat que nous avons appelé le résultat juste. Nous pouvons alors (ce qui est souvent fait) décider que pour chaque question nous partagerons les réponses en deux classes, selon que sera fournie ou non la réponse

juste. Nous avons alors en tête un modèle très algébrique : une variable booléenne (à valeurs 0 ou 1) prend pour ensemble de départ chaque résultat donné, simplement considéré comme une suite de caractères d'écriture. Si la suite des caractères (en général des chiffres, mais aussi peut-être des signes $-$, des virgules, ...) présentée par une réponse est conforme à la suite de référence, elle aura pour image 1, et sinon (par exemple si cette suite est formée du seul "blanc", cas d'une absence de réponse), elle aura pour image 0.

Un modèle exactement du même type que le premier peut être mis en œuvre : au lieu de la référence à la suite de caractères correspondant à la réponse dite juste, on peut avoir recours à la référence à la suite dont nous venons de parler, à savoir la suite formée du seul caractère "blanc". Dans ce cas, on se propose l'analyse dite des non-réponses.

On voit que l'un ou l'autre des deux modèles décrits fonctionne parfaitement bien pour **chaque** question : expérimentalement, à quelques (rares) erreurs matérielles près, on peut attendre d'un "codeur" un tableau de données non sujet à caution. Mais cette "perfection" sur chaque question ne nous apprend absolument rien sur la structure du tableau issu de la passation d'un questionnaire donné auprès d'une population donnée. De plus, mais nous n'en discutons pas ici, la réduction d'information sur chaque réponse (dans les deux cas envisagés) est trop forte pour certaines études.

Remarques :

1° Malgré sa simplicité, l'exemple qui vient d'être indiqué est fondamental en ce qu'il peut servir de référence à une multitude de cas plus compliqués (parce que correspondant à des images plus nombreuses que les deux seuls 0 et 1 par réponse). Ceci est dû au fait que chaque réponse peut être codée par référence à deux modèles (celui de la réponse juste et celui de la non-réponse), entre lesquels il y a a priori des relations, puisque la réponse juste est une réponse (donc que la non-réponse n'est pas la réponse juste). Plusieurs solutions de traitement sont alors pratiquables : la référence à trois classes (réponse juste, réponses fausses, non-réponse), ou l'utilisation de techniques d'analyse faisant de certains objets des "éléments supplémentaires". Détailler les principes d'analyse serait nécessaire pour préciser le jeu des éléments supplémentaires, d'une très grande utilité dans nombre de cas ; ces précisions seraient trop "techniques" pour cet article. Dans notre exemple, on peut, dans une analyse traitant de l'obtention ou non de la réponse juste, placer en éléments supplémentaires le fait de donner ou non une réponse.

2° Dans les enquêtes que nous avons nous-mêmes effectuées, nous avons eu le plus souvent recours pour le codage à d'autres références que celles citées en exemple. C'est que les modèles indiqués

ci-dessus ne renvoient pas aux démarches de réponse, mais à la comparaison avec des réponses-types fixées au départ. Nous utilisons habituellement des modèles tenant compte de la façon dont une réponse a pu être obtenue.

Pour préciser davantage les contraintes du codage, nous allons examiner les différentes sortes de variables qui interviennent en pratique : ce sont les variables qui ***commandent*** *les choix de méthodes de traitement.*

5. Variables de différentes espèces

L'espace **R** des nombres réels possède des structures. Un sous-ensemble quelconque de **R** se voit donc affublé des relations induites par ces structures.

Exemple :
L'ensemble $\{-1 ; 0 ; +1\}$ est tel qu'il comporte l'opposé de chacun de ses éléments (relations de type algébrique). De plus l'écart entre -1 et 0 est le même qu'entre 0 et 1, et c'est la moitié de l'écart entre -1 et 1 (relations de type métrique). Enfin -1 est plus petit que 0 qui est lui-même plus petit que 1 (ordre). Considérons maintenant l'ensemble $\{0;1;2\}$. Pour la métrique et l'ordre, il ne diffère pas du précédent. Algébriquement, il n'en est pas de même (1 et 2 n'ont pas d'opposé dans l'ensemble, et 2 est le double de 1).

Nous l'avons dit, coder représente deux opérations successives :

1° Se rapporter à un **modèle**, c'est-à-dire un ensemble structuré E (lequel reste souvent implicite pour celui qui procède au codage).

2° Mettre en **bijection** E avec une partie de **R** (ou d'un $\mathbf{R}^n$, mais le recours aux coordonnées permet de ramener ce cas à celui de **R**, moyennant un passage éventuel au quotient sur E par une relation d'équivalence).

Dans la pratique, on distingue différentes espèces de variables, d'après les isomorphismes (compatibilités de structures) vérifiés par la bijection f inverse de celle de E sur $A \subset \mathbf{R}$, c'est-à-dire

$$f : A \to E \quad , \qquad \text{avec} \quad A \subset \mathbf{R} \ .$$

Autrement dit, nous nous poserons, pour un codage donné, la question:

"Parmi les propriétés de **R**, *lesquelles se transportent au modèle E ?"*

Nous allons donner une classification grossière dans le but de fixer les idées : selon les besoins on se réfère éventuellement à des classifications qui relèvent de distinctions plus fines.

VARIABLE CONTINUE : L'ensemble A est un intervalle, éventuellement illimité, et la bijection f est **continue** ainsi que sa réciproque (on dit que f est un homéomorphisme). Donc, pour une variable continue, f détermine un transport de la topologie des nombres réels.

Exemples :

A l'échelle macroscopique, il en est ainsi de grandeurs comme masse, temps, capacité (volume), intensité d'un courant, ... Pratiquement, une variable continue est associée aux idées de **reproduction** (ou report) de l'unité d'une part, et de **subdivision** quelconque de l'unité d'autre part.

VARIABLE QUANTITATIVE : La bijection f transporte sur E les opérations + et × des nombres.

Exemples :

Une variable quantitative peut être continue (voir exemples ci-dessus). Elle peut aussi être **discrète**, comme c'est le cas pour des dénombrements : l'unité ne peut alors pas être subdivisée. Ainsi : nombre d'élèves d'une classe, nombre de véhicules ayant franchi un poste de péage, ... C'est aussi le cas pour la monnaie : l'unité (le franc) n'est pas subdivisée au-delà du centime. Lorsque l'on procède à des mesures avec un instrument bien déterminé, la situation est analogue : ainsi, le double décimètre a une précision de l'ordre du millimètre, le pied-à-coulisse du $1/10^e$ de millimètre (c'est donc quand l'instrument de mesure varie que le modèle continu s'impose). Remarquons que des décisions administratives, juridiques ou techniques (normalisation) créent des objets, qui peuvent ensuite être dénombrés, dans ce qui nous apparaît au départ comme un continuum (ainsi : chevaux-fiscaux, départements français, catégories d'habitations, ...).

VARIABLE ORDONNEE : La bijection f transporte sur E l'ordre (plus petit, plus grand) des nombres.

Bien sûr une variable quantitative est toujours ordonnée : on peut définir par exemple des masses plus ou moins grandes, des durées plus ou moins longues, ... Mais la réciproque n'est pas vraie.

Exemples :

En principe, une **notation** de copies devrait être une variable ordonnée. Sur un sujet donné, une copie notée 14 est estimée plus conforme aux critères d'évaluation qu'une copie notée 6. Mais si nous fabriquons une copie en y mettant tout ce qui se trouve dans la copie notée 14 mais pas dans la copie notée 6, nous n'obtiendrons pas une copie (fictive) qui serait notée 8 (écart de 6 à 14). Comparer ce cas à la situation d'une pesée (variable quantitative).

D'une façon plus générale, toute question d'**appréciation** ou d'**opinion** ("très défavorable" à "très favorable") ne peut donner lieu qu'à une variable ordonnée. Pour de telles questions, on demande parfois de répondre en plaçant un point sur un segment (en général [0,1]). En aucune façon, un tel procédé ne permet d'obtenir une variable quantitative continue pour les opinions exprimées : ce serait un artefact.

VARIABLE QUALITATIVE : La seule propriété de f est d'être une bijection.

Notons qu'une variable continue n'est pas nécessairement quantitative. Pour s'en convaincre, un moyen simple est de se demander si une moyenne a un sens. Si ce n'est pas le cas, la variable considérée n'est pas quantitative, même si elle est continue. En voici deux exemples, pour lesquels les obstructions à la quantification sont de nature différente :

— Le premier exemple est celui d'une intensité sonore exprimée en décibels. On sait que par rapport à la pression, les décibels sont répartis sur une échelle logarithmique. Ceci rendrait injustifiée vis-à-vis du phénomène physique l'opération consistant à effectuer une moyenne sur des observations relevées en décibels. Ici intervient donc le fait qu'une échelle ne corresponde pas à une application affine.

— Le deuxième exemple est celui de la position de points sur un cercle. Moyennant une coupure, le choix d'une origine permet un repérage angulaire continu. Mais la "position angulaire moyenne" serait tributaire du choix de l'origine (où la coupure a été faite). Ici, il y a donc un obstacle de nature topologique. Pour considérer un cas précis, examinons trois points disposés sur notre cercle en triangle équilatéral. La moyenne angulaire désignera celui des trois points qui est le plus éloigné de l'origine que nous aurons choisie.

Exemples :

Tout ce qui n'est pas (ou pas encore) quantifié est de ce type. Ainsi posons une question de calcul à une population scolaire et relevons tout simplement le résultat numérique fourni par chaque élève, en supposant, pour simplifier, qu'il n'y a pas de non-réponse. Ici le codage s'est donc réduit à l'application identique. Par rapport au résultat exact, un résultat plus petit ou un résultat plus grand sont également incorrects. Ce simple fait suffit à montrer que la variable ainsi considérée n'est pas quantitative ni même ordonnée.

Comparons cet exemple avec celui d'une estimation de la longueur d'un même objet par un certain nombre de personnes. Ici l'application identique fournit une variable quantitative (la longueur estimée), en l'absence de référence à la même longueur mesurée. La référence à cette mesure demanderait simplement une translation sur cette variable (placement de l'origine des longueurs à la longueur mesurée) qui prendrait alors des valeurs positives ou négatives. Cette nouvelle variable serait, ainsi, quantitative.

Revenons au premier exemple. Supposons qu'il y ait cette fois-ci des non-réponses. La variable :

réponse exacte ⟶ 1
non-réponse ou réponse fausse ⟶ 0

qui ne prend que deux valeurs est de ce fait quantitative de façon évidente.

En revanche, ce n'est pas le cas pour la variable :

réponse juste ⟶ 2
réponse fausse ⟶ 1
non-réponse ⟶ 0

ni pour :

réponse juste ⟶ 1
non-réponse ⟶ 0
réponse fausse ⟶ −1

L'objection est ici que "non-réponse" et "réponse-fausse" ne sont pas hiérarchisées.

A l'utilisation près de certains termes, on retrouve couramment en statistique les distinctions indiquées ci-dessus (mais, par exemple, pour beaucoup de statisticiens, une variable continue est forcément quantitative ; par ailleurs on trouve souvent une catégorie distinguée pour "variable discrète" : on pourra trouver dans le livre de G. de Landsheere, *Introduction à la Recherche en Education*, chez Labor et F. Nathan, une classification plus fine portant sur les échelles et les niveaux de mesure, dont s'est inspirée M.C. Dauvisis (Thèse à paraître, Université de Toulouse) pour s'arrêter aux catégories suivantes :
Echelles nominale (⟵ variable qualitative),
partiellement ordonnée, ordinale (⟵ variable ordonnée),
métrique ordonnée, d'intervalles, et enfin de rapports (⟵ variable quantitative) —
variable ne prenant qu'un ensemble discret de valeurs (mais a priori éventuellement infini). En effet ces distinctions sont importantes pour les traitements numériques auxquels il est possible de soumettre une variable ou une famille de variables. Par ailleurs, la **présentation** d'information utilise, elle, des supports visuels ; les études les concernant sont désignées par le terme de **"la graphique"**. Pour la graphique, on distingue un type de variable qui n'est pas dans la liste ci-dessus : la

variable d'écart. Une variable d'écart est continue mais elle n'est pas quantitative par défaut d'origine : le modèle E est isomorphe à une partie de la droite affine réelle. Nous ne détaillons pas davantage ici, renvoyant le lecteur intéressé à un article de J. Bertin, dans l'Encyclopedia Universalis, article intitulé précisément "la graphique". Dans cet article, une variable d'écart est associée au symbole $\neq$ (repérage des différences, entendez : des écarts).

6. Moralité

Dans le schéma qui a été indiqué sur la circulation d'informations en sciences expérimentales, nous nous sommes arrêtés à l'examen de la case ③. C'est que le présent ouvrage se propose d'étudier en détail ce qui concerne les cases ④ et ⑤, et que la case ⑥ sort du champ de la présente étude.

La préoccupation principale de ce court article est d'éviter les erreurs bêtes et impossibles à corriger après coup qui entachent un grand nombre de comptes-rendus d'observations. Quelqu'un comme H. Freudenthal (ex-directeur de la revue internationale "Educational Studies in Mathematics") proclame, et écrit parfois, à bon droit que l'attitude habituelle du mathématicien, à savoir la confiance a priori dans les résultats annoncés, n'est souvent pas de mise. Une certaine méfiance devant les conclusions d'observations s'avère souvent mieux adaptée.

Lorsque soi-même on se livre à une observation, il est important de comprendre que c'est **dès le début** qu'il y a à envisager les différentes étapes ultérieures. On sait bien qu'aucune chaîne n'est plus solide que le plus fragile de ses maillons.

Un contrôle sérieux de chaque maillon vaut donc la peine. Par exemple, pour le codage, si nous attribuons à une variable plus de propriétés que n'en a notre modèle, les calculs ultérieurs risqueront de prendre en compte ces propriétés fictives, d'où des conclusions tout à fait douteuses (c'est une situation d'artefact). Mieux vaut ne pas exploiter peut-être toutes les propriétés du modèle, en ayant recours à une variable moins "riche" : les résultats seront sûrs. On peut toujours, si l'on est insuffisamment satisfait, **développer** une première observation, en construisant une deuxième observation orientée vers certains aspects seulement désignés par la première. C'est incomparablement plus fructueux que d'avoir à tout recommencer à zéro, ou apparemment à zéro. En fait, se mettre dans le cas de devoir recommencer arrive à tout le monde, y compris l'auteur de cet article, et l'expérience montre que l'acquis méthodologique fourni par un essai non concluant est d'une grande utilité pour effectuer une observation ultérieure avec toutes les précautions et toute l'attention souhaitables.

G. DENIAU - P. ERRECALDE - D. HAUGAZEAU
J. LEHALLE - J. PINCEMIN - J. UEBERSCHLAG
I.N.R.P., Paris

2. ANALYSE DE DONNEES PAR DES ELEVES DU PREMIER CYCLE

UN EXEMPLE : L'ETUDE DE L'EVOLUTION DES EFFECTIFS SCOLAIRES PAR DES ELEVES DE TROISIEME

La recherche, dont notre groupe fait partie*, se propose d'étudier les procédures développées par les élèves du 1er cycle pour analyser des données réelles et nombreuses. Le choix de cette situation répond à un certain nombre d'objectifs dont les deux principaux sont que la lecture critique de tableaux de données nous est apparue comme quelque chose de capital (les choix politiques et économiques sont quotidiennement justifiés par de tels tableaux) et que cela nous a semblé être une approche originale des probabilités et des statistiques (non pas directement dans leur aspect inférentiel, mais plutôt dans leur aspect descriptif). Sur ce dernier point, il nous a semblé que, si de nombreuses et intéressantes recherches ont pu montrer que les enfants (même très jeunes) trouvent intérêt à la pratique de l'aléatoire (sous forme de jeux de hasard : tirage au sort dans des urnes par exemple), cette pratique d'initiation au calcul probabiliste pouvait laisser de côté un aspect important de la démarche du statisticien qui est de ne pas privilégier a priori tel ou tel modèle probabiliste, mais d'observer des données recueillies dans une situation non artificielle, c'est-à-dire non fabriquée *ad hoc* pour illustrer un théorème.

* Recherche I.N.R.P.-I.R.E.M., inscrite au catalogue de l'I.N.R.P.
Code : 73-02.9.01
Titre : « *Evolution des critères de décision en situation aléatoire et approche de modèles probabilistes.* »
Elle concerne les I.R.E.M. de Bordeaux, Grenoble, Lyon, Orléans, Paris, Rennes, Rouen, Versailles.
Cet article a été rédigé par le groupe parisien et D. Haugazeau.
Le rapport de recherche a été publié dans le n° 101 de la revue « *Recherches Pédagogiques* » de l'I.N.R.P. sous le titre « *MATHEMATIQUES DU QUOTIDIEN DANS LES COLLEGES, Activités d'analyse de données dans le premier cycle* ».

Le principe des expérimentations que nous avons conduites consiste à présenter aux élèves des tableaux de données en leur demandant de rechercher à quelles questions on pouvait répondre à l'aide de ces données. Chaque groupe de travail choisissait ensuite une question qu'il traitait. Des observations ont été conduites dans les classes de manière à ce que nous puissions rendre compte du comportement des élèves.

Les expérimentations conduites par les différentes équipes régionales aboutissaient à la rédaction de « dossiers » (un par thème étudié), conçus primitivement comme des dossiers de recherche, mais pouvant cependant avoir une diffusion assez large. Chaque dossier est rédigé selon un plan défini en commun (objectifs généraux, objectifs particuliers au dossier, sujet de l'étude mathématique possible à travers le thème, analyse des observations, conclusion, annexes comportant des travaux d'élèves). Vingt-trois dossiers de ce type sont actuellement disponibles sur des sujets très divers, le plus souvent économiques (ainsi : consommation d'électricité, résultats sportifs, immatriculation de voitures, enquête sur la télévision, le port de Bordeaux, les niveaux de la Loire, etc.). Cf. en annexe la liste de ces dossiers qu'on peut consulter dans les bibliothèques des IREM.

De manière à préciser la démarche que nous avons suivie, il nous a semblé intéressant de choisir un dossier et d'en résumer les caractéristiques. Il s'agit du dossier **EFFECTIFS SCOLAIRES** réalisé par l'équipe parisienne. La suite de cet article est constituée essentiellement d'extraits de ce dossier.

I — Objectifs particuliers au dossier Choix du thème

Nous avons souhaité faire réfléchir les élèves à des problèmes qui les préoccupent beaucoup en classe de 3ème : leur orientation et le choix de leur carrière. En effet, ils ont à cet âge une vision très stéréotypée de « la bonne » profession, mais ont peu de connaissances quant aux débouchés qui leur sont offerts. En outre ils connaissent mal le système scolaire et ne savent pas comment l'utiliser pour conduire au mieux leurs études en fonction de leurs possibilités et de leurs motivations.

Nous n'avons pas trouvé, concernant les carrières, de statistiques nous convenant, c'est-à-dire indiquant la profession exercée en rapport avec les études faites. Les statistiques fournies par la Documentation Française et les services de l'I.N.S.E.E. concernaient des populations trop âgées ou des catégories socio-professionnelles trop larges. Par contre, les statistiques de l'Education Nationale fournissent de nombreux renseignements sur le système scolaire et les orientations des élèves à l'issue de la classe de 3ème. C'est donc sur ce thème que nous avons choisi de travailler.

Nous avons extrait des statistiques publiées par l'Education Nationale : 11 tableaux (cf. annexe) regroupés comme suit :

— Effectifs des élèves du premier cycle, du second cycle court, du second cycle long de l'enseignement secondaire pour les années scolaires 71-72, 72-73, 73-74 et 74-75. Répartition dans les classes en enseignement classique et moderne, transition ou terminal pratique, C.P.P.N. ou C.P.A. (tableaux 1, 2 et 2 bis).

— Répartition des élèves par année de naissance de 1951 à 1963, par niveaux d'enseignement dans le premier cycle et dans le second cycle court ou long (tableaux 3, 4 et 5).

— Répartition suivant l'origine scolaire des élèves du second degré dans le premier cycle et le second cycle long ou court (tableaux 6 et 7).

— Répartition en pourcentages suivant la catégorie socio-professionnelle du père :
  • par filière (tableau n° 8)
  • par option (tableau n° 9)
  • par type de préparation (C.A.P., C.E.P., ... etc.) (tableau n° 10)

— Résultats du baccalauréat : candidats présentés, admis (tableau n° 11) - candidats individuels (tableau n° 11 bis).

Ces deux derniers tableaux n'ont été distribués qu'à certains élèves sur leur demande.

Ces tableaux ont l'avantage d'offrir des données nombreuses, qui peuvent être, soit des nombres assez grands, soit des pourcentages. Ce sont des tableaux, la plupart du temps, à double entrée, d'une lecture relativement difficile, et dont l'interprétation n'est pas immédiate.

II — Déroulement de l'expérimentation, d'après l'analyse des observations

L'expérimentation s'est déroulée en quatre séquences dont nous allons successivement donner les principales caractéristiques.

II.1. Première séquence : PRESENTATION DES TABLEAUX, RECHERCHE DES QUESTIONS

Au cours de cette première séquence les élèves ont pris connaissance des tableaux de données (cf. annexe p. 42 à 53). Certains problèmes de lecture se sont posés, liés aux difficultés d'appréhender la structure des tableaux. Ainsi les tableaux 8, 9, 10 (p. 50-52) ont été interprétés dans un groupe comme donnant la profession exercée en fonction des études faites et non les études faites en fonction de la catégorie socio-professionnelle du père. Exemple : « Des mineurs, il y en a 31 %, pourtant ils ont fait des études au lycée. »

Les questions notées par les élèves sont nombreuses et diversifiées. Elles sont de quatre types :

— **Question de compréhension ou de lecture**
« Pourquoi dans le tableau 4 la case « effectifs de la population » est vide ? »

— **Opinions ou remarques sur les tableaux**
« Manque d'information sur les filières pour interpréter les données. » - « N'est-ce pas une faute de ne pas avoir fait redoubler les élèves de 10 ans qui passent en 6ème de transition ? »

— **Questions à réponses non fournies par les tableaux**
« Beaucoup d'enfants d'ouvriers vont en seconde AB. Pourquoi ? » - « On remarque que le C.A.P. en 3 ans décourage les élèves d'année en année. Pourquoi ? »

— **Questions solubles par une analyse de données**
« Est-ce que la classe sociale du père peut agir sur l'avenir des enfants ? » - « Où vont les élèves qui sortent de quatrième ? » - « Devenir après la classe de troisième, de seconde, de première, après la terminale ? »

II.2. Deuxième séquence : CHOIX D'UN THEME DE TRAVAIL POUR CHAQUE GROUPE ET RECHERCHE D'UN PLAN DE TRAITEMENT

L'enseignant commence par distribuer à tous la liste des questions posées par chaque groupe (cette liste a été entre temps polycopiée, les questions sont classées comme indiqué au *II.1.*).

Les élèves se mettent rapidement au travail et déterminent dans chaque groupe un ou plusieurs thèmes sur lesquels ils désirent travailler. Pour six groupes sur dix, le thème choisi est lié aux questions posées lors de la première séquence.

Un groupe a également, au cours de cette deuxième séquence, préparé un plan de traitement. Les autres se sont contenté de poser des questions à résoudre, sur le thème choisi, sans avoir une vue claire du traitement des données qu'ils devront effectuer.

Le dernier quart d'heure sera consacré à une synthèse, chaque groupe venant exposer, à ses camarades, le sujet sur lequel il va travailler. Voici un résumé des interventions :

Groupe I.a. : Devenir des élèves après les classes de troisième, seconde, et terminale et leur entrée dans la vie active ; pourcentage d'élèves quittant l'enseignement après chacune de ces classes et pourcentage d'élèves continuant à l'université après la classe de terminale.
Moyenne d'âge des élèves à chaque niveau.

Groupe I.b. : Comparaison classe ouvrière et cadres.

Groupe I.c. : Elèves des classes de transition et origine sociale.

Groupe I.d. : Evolution du nombre d'élèves de la sixième à la terminale. Etude des redoublements et des changements de section.

Groupe I.e. : Influence socio-professionnelle du père sur les études des enfants, en traitant quelques exemples : agriculteurs, cadres moyens...

Groupe II.a. : Etude des « déchets ». Passage du cycle long au cycle court, de l'enseignement classique vers l'enseignement de transition ou de terminal pratique.

Groupe II.b. : Influence des classes sociales sur les études des enfants.

Groupe II.c. : Influence des classes sociales et études des redoublements.

Groupe II.d. et e. : Effectifs des élèves en seconde, première et terminale, et leur répartition dans les différentes sections (A, AB, C, T). Combien d'élèves de troisième entrent en seconde A, AB, C ou T ?

Groupe II.f. : Influence du milieu social sur les études. Et les élèves de ce groupe proposent le plan de traitement suivant :

1) Elèves du premier cycle (classes de 4ème et 3ème) répartis en secteurs primaire, secondaire, tertiaire, et chômeurs (catégories de l'I.N.S.E.E.) et suivant l'enseignement reçu (classique, moderne ou transition).

2) Elèves du second cycle : étude similaire et répartition dans les différentes sections (A, AB, C, T, B.E.P.).

II.3. Troisième séquence : TRAITEMENT DES DONNEES

Les groupes, à l'issue de la deuxième séquence, avaient choisi un thème de travail, et certains d'entre eux avaient déjà élaboré un plan de traitement. Au cours de cette troisième séquence qui s'est étalée sur quatre heures, les groupes ont réalisé leur étude, constitué un dossier et préparé un exposé de leur travail pour leurs camarades.

D'après les observations, il y a eu une bonne entente dans les groupes. Les élèves ont eu conscience qu'il fallait s'organiser et ont su se répartir le travail. Il est à noter cependant que, dans un des groupes, un élève ne s'est intéressé au travail que de façon épisodique. Ce comportement peut être interprété comme un conflit de leadership.

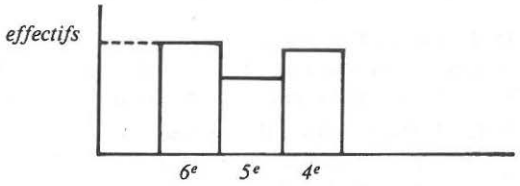
D'un point de vue mathématique, trois types de traitement sont à souligner :

A) Les graphiques

Tous les groupes ont construit des graphiques et nous n'avons pas observé de difficultés particulières pour leur élaboration. Ils sont de deux types : diagramme en bâtons et graphiques type courbe de température.

Voici le relevé de deux observations illustrant la perception que les élèves en ont.

Groupe I.d. : (En parlant d'un graphique « type courbe de température »)
L'enseignante : « Vous avez joint les points. Est-ce que cela a une signification ? »
François : « On a relevé les points pour bien se rendre compte. »
L'enseignante montre un point quelconque à l'intérieur d'un segment.
L'enseignante : « Est-ce que ce point a une signification ? »
François : « Non. »
Bruno : « Non. »
L'enseignante désigne les points construits à l'aide des valeurs extraites des tableaux.
L'enseignante : « Ce qui compte, c'est uniquement cela ? »
François : « Il faudrait pas faire comme ça, mais comme cela :



L'enseignante : « Etait-ce la seule façon de visualiser ? »
Bruno : « Tout le monde a fait comme ça dans les groupes. Ça présente mieux ! »

Groupe II.d. : *Manuelle : « Elle sera drôle notre courbe ! »*
.....
« Dis donc, elle sera drôlement drôle notre courbe ! Elle fait ça ! » (Elle dessine un zigzag avec son doigt)... *« Vraiment drôle, je trouve... »*
Manuelle en construisant son graphique a choisi de mettre en abscisses les effectifs, et en ordonnées les sections dans l'ordre A B C D G. Manuelle montre son graphique à l'enseignante.
Manuelle : « J'en ai oublié un... Madame, ça fait drôle ! »
L'enseignante : « Pourquoi ? »
Manuelle : « Elle monte comme ça ! » (Nouveau signe du doigt).
L'enseignante : « A cause de quoi ? »
Manuelle : « Avec G en bas et A en haut ? »
L'enseignante : « Ça ferait pareil ! »
Manuelle : « C'est parce que c'est à peu près pareil la répartition... Si on avait inversé les axes ? »...
Manuelle semble avoir une image mentale de ce que doit être un graphique type température. Elle est très surprise par ce qu'elle vient d'obtenir qui ne correspond pas à l'idée qu'elle en avait. Elle cherche à savoir ce qu'elle doit modifier pour obtenir un résultat qui cadre avec cette image, d'où finalement sa proposition d'inverser les axes. Ce problème réglé, elle s'en pose un autre, celui de l'adéquation du graphique avec ce qu'elle veut faire apparaître.
Manuelle : (désignant les élèves d'un autre groupe) *« Est-ce qu'on fait comme eux... des bâtons... ou une ligne ? On veut représenter une vue d'ensemble... proportionnellement,... donc c'est des bâtons. »*
Elle découvre donc que ce qui compte, c'est la longueur des bâtons qui doit être proportionnelle aux effectifs. Finalement, elle choisira de laisser en abscisses les effectifs et en ordonnées les sections, l'aspect du graphique lui étant apparu secondaire quand elle eut découvert pourquoi son graphique semblait « anormal ».

B) Calcul de pourcentages

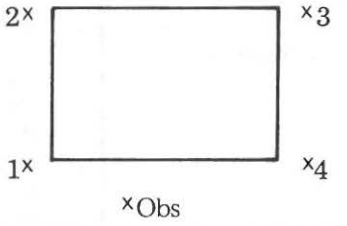
On a observé que tous les élèves avaient déjà entendu parler de pourcentages. Certains d'entre eux se souvenaient de la méthode de calcul. D'autres en ont une vague idée et retrouvent assez rapidement comment conduire le calcul : *(Valérie : « On pourrait quand même faire quelque chose, les pourcentages. Je ne sais plus comment... Tu prends les gros, les petits, les petits par les gros et tu multiplies par 100. »)*

Dans la plupart des groupes, une fois rappelée, la technique du calcul ne présente pas de difficultés. Certains élèves ont même très bien compris la signification d'un pourcentage : *(Olivier : « Justement, les pourcentages, ce sera pour mettre sur un plan d'égalité. »* Cependant, dans un groupe, les élèves n'ont peut-être pas une idée très claire de la signification des pourcentages car ils ont calculé des moyennes de pourcentages, sans se demander si ce calcul avait un sens.

D'autres difficultés sont apparues, qui nous semblent davantage liées aux approximations qu'aux pourcentages eux-mêmes. Ainsi le groupe II.b. a travaillé, à la séance précédente, sur une catégorisation qui regroupait certaines professions. Les élèves décident maintenant de relever, pour les trois catégories ainsi définies, les pourcentages donnés dans les tableaux 8, 9 et 10, d'en faire la somme et la moyenne par section (A, AB, ...). Ils font tous ces calculs sur une calculatrice HEWLETT-PACKARD 20 et s'amusent à déplacer la virgule pour avoir plus de précision (ils choisissent 5 chiffres après la virgule). Ils font la somme pour vérifier si « ça tombe sur 100 », mais n'obtenant jamais exactement 100, mais un nombre inférieur à 100, ils notent dans un coin de leur feuille la petite différence à 100, due pensent-ils aux approximations de calculs, c'est-à-dire aux calculs effectués avec un nombre insuffisant de décimales. Ils tombent tout à coup sur 100,117 : *(« Ça c'est embêtant ! On ne tombe pas sur 100 ! On s'est gourré ! »)* Alors qu'ils n'ont pas été étonnés par le fait que le total soit légèrement inférieur à 100, ils n'admettent pas une différence dans le sens inverse (supérieur à 100 !) et ils recommencent leur calcul. Ils arrivent à 100, 00100 et estiment toujours s'être trompés : *(« Y a pas de mecs en plus ! On refera les calculs chez nous ! »)*

Voici enfin l'observation complète réalisée auprès du **groupe II.d.e.** au moment du calcul de pourcentage. (On remarquera la disposition adoptée par les observateurs et qui sépare les dialogues des élèves — à droite — de leurs actions commentées — à gauche).

Groupe II.d.e. : *1.* — Séverine
2. — Manuelle
3. — Pascale
4. — Isabelle



2. — Il faut trouver le nombre d'élèves du privé qui entrent en A sachant qu'il y en a 16 %... Il faut savoir d'abord à combien ça correspond en nombre... 69 707 élèves dans le privé.

P. — Il faut prendre les 16 %.

2. — Il faut bien trouver 16 % de 69 707 ?

P. — Quelle est l'équation ?

4. — $\frac{16}{100} = \frac{100\,A}{B}$

P. — Comment ?

2. — $\frac{16}{100} = 69\,707$

P. — Et l'inconnue ?

1. — L'inconnue... $\frac{16}{100} = \frac{69\,707 - B}{69\,707}$

2. — Sur 100... Il faut multiplier...

1. — 69 707 − 16 % = ... Ça ne va pas.

4 a trouvé, mais il y a un malentendu.

4. — 69 707 × 16... Ça fait 11 153,12.

P. — Tu as multiplié par 16 ? Alors, tu trouves plus d'élèves en seconde A qu'au total ?... Il faut écrire l'équation.

2. — Sur 100 élèves, 16 en A. Sur 69 707, 16 × x.

P. — Pourquoi par x ?

2. — En multipliant 100 par combien, on trouve 69 707 ?

P. — Pourquoi en multipliant 100 ?... Ecris avec des flèches.

4. — Faut pas diviser 100 par 16 pour trouver la flèche inverse ? On divise ça par...

P. — Vous vous concentrez chacune de votre côté. Vous essayez de trouver.

Ils cherchent sur leurs feuilles, soit « avec des flèches », soit en faisant des multiplications.

1. — Il faut trouver combien de fois 100 dans 69 707... Diviser 69 707 par 100.

2. — La prof a dit « pourquoi diviser ? »

1. — Je le fais, on verra bien.

2 cherche l'équation.

1. — Y a un reste... 697 fois 100 dans... Hein ! Ça ne va pas du tout.

2. — Ben, tu divises par 100.

1. — Donc, 697 fois dans 69 707... Mais attention, reste le 7... Mais en gros, il y aura... sur... Ça fait que ça sera multiplié par 697... Bouge pas.

2 essaie de trouver l'équation.

2 va voir le prof.

4 appelle le prof.

4 expose sa solution au prof en termes de fonction (je n'ai pas pu noter).

1 fait ses calculs
2 aussi.
4 contrôle ses résultats.

Elles vérifient les chiffres.
2 trouve 11 251,12.
...
Après un long tâtonnement, elles redécouvrent la technique du calcul.

1. — On va passer l'heure là-dessus !

1. — Ah ! Non, viens voir ! Si il y a 697 fois 100, il va falloir multiplier 16 par 697. Attends.

4. — Je crois avoir trouvé : $\frac{16}{100} = \frac{x}{69\ 707}$

P. — Qu'est-ce que vous en pensez ?

2. — $\frac{69\ 707}{16} = \frac{100}{x}$

P. — Il faut prendre 16 % de 69 707.

4. — Eh ! Ben oui. C'est bon ce que j'ai fait !

2. — Je trouve la même chose que toi.

4. — Opération pour trouver les pourcentages :

$16\ \% = \frac{100\,x}{69\ 707}$

Donc, l'opération :

$\frac{16 \times 69\ 707}{x} = x$

1. — Moi, je dis que $\frac{\frac{16}{100}}{69\ 707} = x$

2. — Regarde... Je sais même plus comment s'appelle cette opération ; je l'ai fait au primaire... 69 707, c'est égal à quoi ?

2. — Je fais pas comme vous. En réduisant au même dénominateur, tu trouves...

4. — Mais moi, j'effectue directement.

1. — On va voir si je trouve la même chose que vous par mon truc... Calculez tout bas !

4. — Effectivement, je me suis trompée dans le résultat... J'ai oublié d'additionner. Je trouve 1 million... (Elle donne les chiffres, un par un) parce qu'un million...

1. — Combien tu trouves pour x ?

4. — 1 115 312, mais j'ai dû me tromper.

2. — Pourquoi 1 115 312 ?

2. — Tu oublies la virgule.

4. — Oui, j'ai pas divisé par 100. J'ai sauté une étape.

Cette dernière observation nous a paru significative des difficultés qui peuvent encore apparaître en classe de troisième, lorsque les élèves ont à utiliser des pourcentages. On peut noter en particulier comment ces difficultés se situent à la fois au niveau des procédures de calculs (les élèves se souviennent de morceaux d'algorithmes) et des concepts sous-jacents.

C) Comparaison par rapports

Il est intéressant de noter que certains élèves ont eu l'idée de faire des comparaisons en calculant des rapports.

Groupe I.c. :

Anne Christine : « Moi je fais des rapports de baisse ! »
Elle commente un tableau et dit :
« Baisse de plus de la moitié, une année plus tôt. »
Olivier : « Regarde, les élèves d'âge normal, en transition, 4 996, en classe normale 367 373. C'est incroyable, la différence est énorme ! »
Vincent : « Mais regarde à 12 ans. »
Anne : « Oui, c'est intéressant ! »
.....
Ils font quelques calculs, puis pensent à utiliser la calculatrice HP 20 qui est à leur disposition dans la classe.
Vincent : « Tu veux diviser ? »
Anne : « C'est ça, donne ton chiffre. »
Olivier : « Au total, dans les classes normales, les élèves d'âge normal sont 4,7 fois plus nombreux que dans les classes de transition. Mais en 4ème... (inaudible). *Donc, les élèves de transition sont plus vieux en proportion. »*

II.4. Quatrième séquence : SYNTHESE FINALE

Au cours de cette séquence, chaque groupe présente à la classe les travaux réalisés.

Les élèves se présentent à cette séance avec le dossier qu'ils ont réalisé dans chaque groupe et avec des affiches pour leurs camarades, qui visualisent les grandes lignes de leurs exposés.

La disposition spatiale de la classe a été modifiée de manière à faciliter la communication : les tables rassemblées au fond et les chaises disposées autour d'une aire de parole proche du tableau.

Pour cette séquence, nous disposions d'un temps limité (une heure par demi-groupe) en raison de la fin de l'année scolaire proche, ce qui a entraîné quelques difficultés : il fallait que tous les groupes puissent présenter leur travail et cela ne nous a pas permis de faire une discussion générale suffisamment longue après les exposés.

Les élèves ont développé leurs résultats sous forme de commentaires explicatifs qui n'apparaissent pas nécessairement dans leurs dossiers. Des discussions entre les groupes sont apparues, les élèves posant des questions de compréhension, mais il y eut moins d'échanges que ce que l'on aurait pu attendre.

III — Travaux réalisés

Chaque groupe de travail a réalisé un dossier qui a été remis à la fin de la quatrième séquence. Une analyse de ces dossiers permet de faire une description exhaustive des travaux réalisés. Ces travaux sont très diversifiés. Nous en donnerons deux exemples :

III.1. TRAVAIL DU GROUPE I.a.

Question : Pour l'année 1973-1974, répartition (par type d'enseignement et par classe) de tous les enfants nés en 1959, c'est-à-dire ayant **14 ans** à cette date.

Traitements : Les élèves ont travaillé sur les tableaux n° 3, 4 et 5. Pour la classe d'âge choisie (nés en 1959, c'est-à-dire 14 ans), ils ont regroupé les effectifs en établissant la classification suivante :

— C.E.S.	enseignement classique et moderne enseignement de transition et pratique
— Second cycle long	sections conduisant à un baccalauréat sections conduisant à un brevet de technicien
— Enseignement élémentaire	
— Second cycle court	sections conduisant à un C.A.P. sections conduisant à un B.E.P. sections conduisant à un C.E.P.
— Non scolarisés.	

Les élèves ont ensuite transformé les effectifs en pourcentages par rapport à la population totale (841 478), puis ils ont construit les diagrammes en bandes, correspondant aux pourcentages trouvés.

Une deuxième partie du travail a consisté à calculer le nombre d'élèves de 14 ans ayant : trois ans / deux ans / un an de retard / l'âge normal / un an / deux ans / trois ans d'avance. Puis les effectifs obtenus ont été transformés en pourcentages et ont été représentés sous forme de diagrammes en bâtons.

III.2. TRAVAIL DU GROUPE I.b.

Question : « La classe sociale du père agit-elle sur l'avenir des enfants ? »

Traitement : Les élèves de ce groupe se sont attachées uniquement à une comparaison de la classe ouvrière et des cadres supérieurs, aux niveaux de 4ème et de seconde. Pour effectuer cette comparaison, elles ont relevé dans les tableaux 8 et 9 les pourcentages d'enfants de : cadres supérieurs / manœuvres / ouvriers qualifiés / ouvriers spécialisés / mineurs / marins pêcheurs, fréquentant différentes filières.

Au niveau de la classe de 4ème, les élèves se sont intéressées aux 4èmes aménagées / 4èmes pratiques / 4èmes de type I et II.

Au niveau de la classe de seconde, les élèves se sont intéressées aux sections A, AB, C et T.

Par filière, elles ont fait la somme des pourcentages de chacune des cinq catégories professionnelles ouvrières (sans les cadres supérieurs) et calculé la moyenne correspondante. Cette moyenne est supposée caractériser la « classe ouvrière ».

Un graphique a été construit dont les points représentent les filières envisagées. Ces points sont repérés en abscisse par le pourcentage d'enfants de cadres supérieurs et en ordonnée par le pourcentage d'enfants d'ouvriers. Les élèves ont comparé la position des points par rapport à la première bissectrice et commenté leurs résultats.

IV — Conclusion

En conclusion, nous allons préciser, au travers des divers travaux effectués par les équipes concernées par la recherche, quel a été l'intérêt de ces études.

Les activités d'analyse des données ont été l'occasion pour les élèves d'utiliser des notions mathématiques qui n'étaient pas connues d'eux, de revoir et mieux comprendre certaines autres, mais aussi de découvrir des notions nouvelles qui étaient particulièrement pertinentes pour effectuer le travail choisi.

Parfois il s'est agi d'une simple utilisation de notions. On a pu remarquer un fait constant dans tous les dossiers : les élèves de sixième et cinquième sont heureux de faire des calculs et la longueur de ceux-ci ne les effraie pas.
« *Ce travail est passionnant pour les calculs qu'il y a à exécuter* » dit un élève.

Dans bien des cas le thème traité a permis aux élèves d'approfondir des notions mathématiques qu'ils connaissaient, mais n'utilisaient pas de façon habituelle et ne dominaient pas.
« *Je trouve que c'était intéressant car on a revu les graphiques et les pourcentages.* »

Quelquefois le thème a été à l'origine de la découverte de nouvelles notions par les élèves (par exemple les entiers relatifs lors de l'étude des immatriculations de voitures ; ou encore la notion de distance).

Enfin, les questions de prévisions se sont posées naturellement lors de l'étude de données chronologiques.

Un autre aspect qui nous paraît important a été de conduire les élèves à utiliser une démarche voisine de celle du statisticien. Ils ont eu des données réelles, ont formulé et choisi les travaux qu'ils désiraient effectuer, ont cherché à répondre à la question qu'ils s'étaient posée. Ils ont ainsi eu l'occasion de critiquer la méthode de traitement choisie.

Les élèves ont aussi comparé les diverses méthodes utilisées pour traiter une même question, en particulier lorsque des groupes différents avaient choisi le même travail. Enfin, ils ont su être critiques et prudents lors de la conclusion.

Le travail de groupe en situation ouverte a été pour beaucoup d'élèves une heureuse révélation (surtout pour ceux qui n'étaient pas les « bons élèves » dans le cadre traditionnel). Ils ont eu l'occasion de mieux connaître leurs camarades, de prendre une responsabilité vis-à-vis du groupe et de la classe ; ils ont écrit à ce propos :

« nous avons la responsabilité de faire un bon travail, de chercher avec ceux de notre groupe, de réfléchir ensemble ; en nous laissant plus libre cela nous donne plus envie de travailler. »

Ils ont, en général, ressenti la nécessité d'une organisation, d'une part au niveau du groupe, mais aussi dans leur propre travail :

« ce travail a été long mais nous nous sommes réparti les différents exercices ce qui prouve que tous ceux de mon groupe ont travaillé. »

Lors du travail de groupe, ou des synthèses, les élèves ont été conduits à exposer le résultat de leurs travaux et à défendre, le cas échéant, leurs opinions mises en cause par d'autres.

En majorité les élèves ont aimé ce travail et désirent recommencer de telles activités :

« là on comprend toujours, on n'apprend pas par cœur, parce que, comme c'est nous qui l'avons dit, on comprend mais on sait expliquer. »
« ce travail nous a permis de devenir un peu plus logiques ; je voudrais encore travailler ensemble pour raisonner et exposer nos idées et nos exemples de travail. »

Ainsi on constate à travers les divers dossiers qu'un enseignement par thèmes est, d'une part possible, d'autre part intéressant à la fois pour les élèves et pour les enseignants. En effet, les méthodes pédagogiques mises en place dans cette expérience ont amené ces derniers à se remettre en cause, à modifier leur comportement à l'égard de leurs élèves, ce que l'un d'entre eux a énoncé :

« *Je ne pourrai plus enseigner comme avant.* »

ANNEXES

I. LISTE DES DOSSIERS

		Equipe de
1	Climatologie (Bordeaux, Brest, Strasbourg)	**Bordeaux**
2	Climatologie (Brest, Châteauroux, Orléans, Strasbourg)	**Orléans**
3	Consommation d'eau à Grenoble	**Grenoble**
4	Consommation d'électricité dans le Sud-Est de 1971 à 1975	**Lyon**
5	Consommation d'électricité le 3e mercredi du mois de décembre dans le Sud-Est	**Lyon**
6	Effectifs scolaires	**Paris**
7	Etude de performances sportives	**Bordeaux**
8	Evolution de la population en Gironde et en Dordogne	**Bordeaux**
9	Football	**Paris**
10	Immatriculation de voitures dans le département du Rhône	**Lyon**
11	Immatriculation de voitures dans les pays de la C.E.E.	**Rennes**
12	Loire	**Orléans**
13	Loisirs - Vacances	**Orléans**
14	Mesures de la longueur d'une salle	**Bordeaux**
15	Port de Bordeaux	**Bordeaux**
16	Port de Rouen	**Rouen**
17	Résultats sportifs	**Grenoble**
18	Sondages	**Paris**
19	Taille d'élèves de 5ème d'un C.E.S. d'Orléans	**Orléans**
20	Taille et poids d'élèves de 5ème d'un C.E.S. de Valence	**Grenoble**
21	Taille, poids, envergure et résultats sportifs	**Rennes**
22	Taille et poids : croissance de la naissance à dix-huit ans	**Orléans**
23	Télévision	**Rouen**

TABLEAU N° 1

(Enseignement public + privé)

EFFECTIFS DES ELEVES DU 1er CYCLE DE L'ENSEIGNEMENT SECONDAIRE

Année scolaire	*Enseignement classique et moderne*				*Enseignement transition ou terminal pratique*				*Classes nouvelles CPPN ou CPA*
	6ème I et II	*5ème I et II*	*4ème I, II et II amén.*	*3ème I, II et II amén.*	*6ème III*	*5ème III*	*4ème pratique*	*3ème pratique*	
71 - 72	702 379	621 703	561 431	515 828	155 369	165 360	123 853	69 791	n'existaient pas à cette date
72 - 73	727 617	545 463	578 774	504 020	172 970	187 368	89 759	70 651	57 246
73 - 74	738 579	664 621	586 873	532 521	157 662	188 077	61 740	48 719	110 972
74 - 75	751 230	678 823	595 481	544 737	147 282	178 066	37 432	31 529	157 561

TABLEAU N° 2

(Enseignement public + privé)

EFFECTIFS DES ELEVES DU SECOND CYCLE COURT DE L'ENSEIGNEMENT SECONDAIRE

Année scolaire	*C.A.P. en 3 ans*			*C.A.P. en 2 ans*		*B.E.P.*		*C.E.P.*
	1ère année	*2ème année*	*3ème année*	*1ère année*	*2ème année*	*1ère année*	*2ème année*	
71 - 72	167 799	140 546	123 029	15 814	16 639	92 300	66 191	6 204
72 - 73	181 835	145 126	125 221	15 310	15 051	113 411	81 371	9 477
73 - 74	179 614	142 930	119 066	14 955	15 372	116 487	94 235	11 422
74 - 75	173 816	138 058	115 038	15 225	14 848	126 661	96 580	11 683

TABLEAU N° 2 BIS

(Enseignement public + privé)

EFFECTIFS DES ELEVES DU SECOND CYCLE LONG DE L'ENSEIGNEMENT SECONDAIRE

Année scolaire	*Baccalauréat ABCDE*			*Baccalauréat de technicien*		*Brevet de technicien*		
	2ndes	*1ères*	*Terminales*	*1ères*	*Terminales*	*2ndes*	*1ères*	*Terminales*
71 - 72	337 581	197 602	203 281	Les chiffres exacts ne sont pas connus				
72 - 73	349 712	203 971	214 378	76 811	67 201	9 614	6 538	6 367
73 - 74	331 060	208 201	215 540	84 505	72 640	8 964	6 713	6 072
74 - 75	336 055	196 083	217 633	90 220	80 908	9 232	6 774	6 245

TABLEAU N° 3

Année scolaire 1973-1974

(Enseignement public + privé)

EFFECTIFS DES ELEVES DU 1er CYCLE DE L'ENSEIGNEMENT SECONDAIRE

Année de naissance	*Age au 1.1.74*	*Effectifs de la pop. totale* (1)	*Enseign. Elém.*	*Enseignement classique et moderne*				*Enseignement transition ou terminal pratique*				*Classes nouvelles CPPN et CPA*
				6ème I et II	*5ème I et II*	*4ème I, II et II amén.*	*3ème I, II et II amén.*	*6ème III*	*5ème III*	*4ème pratique*	*3ème pratique*	
1963	10	862 940	795 673	42 789	855	24	—	132	6	—	—	—
1962	11	835 485	389 326	367 373	39 223	952	47	4 996	156	1	—	—
1961	12	848 298	77 738	266 396	306 987	40 376	1 181	119 678	5 790	65	6	—
1960	13	829 942	17 207	59 180	247 212	255 096	36 997	29 772	153 312	1 764	70	—
1959	14	841 478	4 263	2 470	67 053	224 717	220 434	2 554	26 568	51 163	3 161	64 397
1958	15	825 807	1 293	305	2 996	62 427	209 178	402	2 093	8 235	40 766	43 392
1957	16	831 216	393	54	233	3 046	60 530	79	138	465	4 459	2 754
1956	17	826 224	18(2)	8	56	193	3 719	43	12	41	222	429
1955	18	826 478	—	4	6	42	435	6	2	6	35	—

(1) Evaluation de l'INSEE
(2) Elèves de l'enseignement privé

TABLEAU N° 4

Année scolaire 1973-1974
(Enseignement public + privé)
EFFECTIFS DES ELEVES DU SECOND CYCLE LONG DE L'ENSEIGNEMENT SECONDAIRE

Année de naissance	*Age*	*Effectifs de la population totale* [1]	*Baccalauréat ABCDE*			*Baccalauréat de technicien*		*Brevet de technicien*		
			2ndes	*1ères*	*Terminales*	*1ères*	*Terminales*	*2ndes*	*1ères*	*Terminales*
1959	14	841 478	31 206	641	8	16	3	47	—	—
1958	15	825 807	150 620	23 568	554	1 415	8	1 366	54	1
1957	16	831 216	115 387	94 370	20 107	24 755	847	3 103	1 376	32
1956	17	826 224	30 977	67 434	81 009	36 901	16 030	3 069	3 855	771
1955	18	826 478	2 381	19 981	74 776	17 738	30 762	1 379	4 384	2 210
1954	19	834 938	251	1 868	33 155	3 230	19 662	196	1 599	2 210
1953	20	835 517	47	142	5 158	326	4 661	13	151	690
1952	21	865 905	51	65	465	43	453	16	36	83
1951 et avant	22	—	140	132	308	81	208	8	37	97

(1) Evaluation de l'INSEE

TABLEAU N° 5

Année scolaire 1973-1974
(Enseignement public + privé)
EFFECTIFS DES ELEVES DU SECOND CYCLE COURT DE L'ENSEIGNEMENT SECONDAIRE

Année de naissance	*Age*	*Effectifs de la population totale* (1)	*C.A.P. en 3 ans*			*C.A.P. en 2 ans*		*B.E.P.*		*C.E.P. Certificat d'enseign. profess.*
			1ère année	*2ème année*	*3ème année*	*1ère année*	*2ème année*	*1ère année*	*2ème année*	
1960	13	829 942	462	7	13	5	—	—	1	—
1959	14	841 478	93 407	965	13	192	13	304	25	331
1958	15	825 807	71 652	73 432	1 313	2 090	290	19 953	808	8 161
1957	16	831 216	12 059	56 269	56 365	6 140	2 476	59 038	18 149	2 505
1956	17	826 224	1 545	10 411	48 602	4 123	5 837	31 567	44 787	425
1955	18	826 478	310	1 505	10 713	1 269	4 276	4 636	25 430	—
1954 et avant	19	—	179	341	2 047	1 136	2 480	1 009	5 035	—

(1) Evaluation de l'INSEE

TABLEAU N° 6

Année 1973-1974
(Enseignement public + privé)
ORIGINE SCOLAIRE DES ELEVES DU SECOND DEGRE

Orientation 1973-1974 / *Origine 1972-1973*	*Premier cycle*				*Second cycle long*		
	6e	*5e*	*4e*	*3e*	*2e*	*1re*	*Termin.*
Elémentaire :							
C.M.1.	33 224						
C.M.2.	770 988						
Premier cycle :							
6e	90 607	774 672					
5e		59 096	591 935				
4e			52 108	536 702			
3e				44 125	290 451		
Second cycle long :							
2e					42 370	276 648	
1re						21 501	248 083
Termin.							45 818
Second cycle court :							
B.E.P.1.					312	299	
B.E.P.2.						5 226	72
C.A.P.1.							
C.A.P.2.							
C.A.P.3.					6 492		
C.E.P.							
Classes nouvelles :							
C.P.P.N.			40	55			
C.P.A.							
Autres origines ...	3 237	20 562	5 747	1 654			
TOTAL 1973-1974 ..	**898 065**	**854 330**	**649 830**	**582 536**	**339 625**	**303 674**	**293 973**

TABLEAU N° 7

Année 1973-1974
(Enseignement public + privé)
ORIGINE SCOLAIRE DES ELEVES DU SECOND DEGRE

Orientation 1973-1974 / *Origine 1972-1973*	*Second cycle court*					*Classes nouvelles*			*Sortie*
	B.E.P.1.	*B.E.P.2.*	*C.A.P.1.*	*C.A.P.2.*	*C.A.P.3.*	*C.E.P.*	*C.P.P.N.*	*C.P.A.*	
Premier cycle :									
6e									35 308
5e			104 528	2 368	1 652	1 502	54 967	6 630	10 153
4e			34 752	1 948	1 315	2 548	4 748	15 527	19 255
3e	107 571	2 671	2 099	193	107	1 160	434	913	124 947
Second cycle long :									
2e	13 168	4 090							23 083
1re									22 164
Terminales.....									242 109
Second cycle court :									
B.E.P.1.	6 120	97 646							24 344
B.E.P.2.		4 662							86 462
C.A.P.1.			9 488	131 268					41 279
C.A.P.2.				4 570	108 556				32 000
C.A.P.3.					5 092				113 637
C.E.P.						70			9 407
Classes nouvelles :									
C.P.P.N.			8 846	222	103	2 941	3 779	8 282	25 937
C.P.A.								15	7 025
Autres origines :	4 498	465	19 841	2 321	2 202	3 201	11 393	4 554	
TOTAL 1973-1974	**131 357**	**109 534**	**179 554**	**142 890**	**119 027**	**11 422**	**75 321**	**35 651**	

TABLEAU N° 8

Année 1973-1974

ORIGINE SOCIO-PROFESSIONNELLE DES ELEVES DU SECOND DEGRE PUBLIC

Classes de 4ème (tous établissements)

Répartition en pourcentages, par filière, pour chaque catégorie

Catégorie socio-professionnelle du père / *Sections*	*Type I*	*Type II*	*Type II amén.*	*4e pratique*	*Effectif total* = 100,0	*Effectif total* % du total général
Agriculteurs exploitants	27,4	51,4	10,2	11,0	39 755	7,6
Ouvriers agricoles	19,6	38,8	10,0	31,6	10 177	1,9
Patrons de l'industrie et du commerce :						
Industriels	60,7	31,0	6,3	2,0	3 475	0,7
Artisans	40,8	42,9	8,5	7,8	21 942	4,2
Patrons pêcheurs	38,7	45,7	7,6	8,0	473	0,1
Grands commerçants	60,8	30,6	6,4	2,2	3 533	0,7
Petits commerçants	44,5	40,8	8,6	6,1	20 541	3,9
Professions libérales - Cadres supérieurs	78,0	18,6	2,9	0,5	43 866	8,3
Cadres moyens	61,8	30,1	6,0	2,1	55 340	10,5
Employés	46,7	37,1	9,0	7,2	63 077	12,0
Ouvriers :						
Contremaîtres	48,4	36,5	9,2	5,9	21 683	4,1
Ouvriers qualifiés	36,7	38,5	10,6	14,2	80 562	15,3
Ouvriers spécialisés	31,1	38,1	11,0	19,8	57 507	10,9
Mineurs	31,8	39,7	10,0	18,5	5 884	1,1
Marins pêcheurs	33,6	39,8	10,4	16,2	1 964	0,4
Manœuvres	22,5	33,4	10,5	33,6	16 534	3,1
Personnel de service	35,6	38,2	11,4	14,8	15 940	3,0
Autres catégories	49,4	34,3	8,4	7,9	17 586	3,3
Sans profession - Retraités	30,5	33,8	10,0	25,7	23 164	4,4
Orphelins, Pupilles	17,4	32,2	11,9	38,5	3 286	0,6
TOTAL DEFINI	219 505	183 194	44 852	58 748	506 299	96,2
Indéterminés	9 929	6 564	1 896	1 555	19 944	3,8
TOTAL GENERAL	**229 434**	**189 758**	**46 748**	**60 303**	**526 243**	**100,0**
(%)	*43,6*	*36,0*	*8,9*	*11,5*	*100,0*	

TABLEAU N° 9

Année 1973-1974

ORIGINE SOCIO-PROFESSIONNELLE DES ELEVES DU SECOND DEGRE PUBLIC

Classes de secondes (tous établissements)

Répartition en pourcentages, par option, pour chaque catégorie

Catégorie Sections socio-profess. du père	*A*	*AB*	*C*	*T*	*Enseign. spécial (3)*	*T.I. T.H.*	*Effectif total* = *100,0*	*Effectif total* % *du total général*
Agriculteurs exploitants	12,1	30,5	35,7	18,8	1,2	1,7	16 884	6,5
Ouvriers agricoles ...	15,1	33,0	28,0	20,0	2,0	1,9	2 378	0,9
Patrons de l'industrie et du commerce :								
Industriels	19,1	20,0	44,7	12,7	0,9	2,6	2 815	1,1
Artisans	16,3	28,0	33,0	19,0	1,2	2,5	10 750	4,1
Patrons pêcheurs..	17,2	40,4	32,8	8,6	0,5	0,5	198	0,1
Grds commerçants	19,6	24,1	43,7	11,0	0,2	1,4	2 903	1,1
Petits commerçants	17,0	27,4	37,5	15,1	0,9	2,1	11 827	4,5
Professions libérales - Cadres supérieurs	16,4	14,9	59,2	8,4	0,2	0,9	38 032	14,6
Cadres moyens	17,3	22,8	43,9	13,8	0,7	1,5	38 448	14,7
Employés	16,4	28,8	34,5	17,5	1,1	1,7	33 362	12,8
Ouvriers :								
Contremaîtres	14,8	27,5	32,3	22,0	1,3	1,2	11 303	4,3
Ouvriers qualifiés .	14,6	31,9	26,4	22,8	1,7	2,6	26 782	10,3
Ouvriers spécialisés	15,4	32,7	26,7	21,6	1,4	2,2	17 897	6,8
Mineurs	14,3	32,1	24,9	24,6	2,0	2,1	1 919	0,7
Marins pêcheurs ..	14,7	33,8	32,2	15,6	2,0	1,7	868	0,3
Manœuvres	14,9	35,3	24,7	21,3	2,3	1,5	3 817	1,5
Personnel de service .	17,7	33,4	26,0	19,1	1,6	2,2	6 062	2,3
Autres catégories....	19,5	27,6	35,9	14,9	0,9	1,2	9 835	3,8
Sans profession - Retraités	17,9	32,0	28,7	18,0	1,4	2,0	10 229	3,9
Orphelins, Pupilles ..	19,7	30,5	24,8	18,7	2,9	3,4	685	0,3
TOTAL DEFINI..	39 995	65 579	93 559	40 944	2 608	4 349	247 034	94,5
Indéterminés	1 874	3 512	5 229	2 485	353	866	14 319	5,5
TOTAL	**41 869**	**69 091**	**98 788**	**43 429**	**2 961**	**5 215**	**261 353**	**100,0**
(%)	*16,0*	*26,5*	*37,8*	*16,6*	*1,1*	*2,0*	*100,0*	

(3) Sections accueillant des élèves ayant un handicap (mal entendants, mal voyants, infirmes...).

TABLEAU N° 10

Année 1973-1974

ORIGINE SOCIO-PROFESSIONNELLE DES ELEVES DU SECOND DEGRE PUBLIC

1re année d'enseignement professionnel court (tous établissements)

Répartition en pourcentages, par type de préparation, pour chaque catégorie.

Catégorie socio-professionnelle du père \ *Sections*	*1re année* C.A.P. B.E.P. 2 ans	*1re année* C.A.P. 3 ans	C.P.P.N.	C.P.A.	C.E.P.	*Effectif total* = 100,0	*Effectif total* % du total général
Agriculteurs exploitants	33,6	34,5	17,9	11,9	2,1	22 078	6,3
Ouvriers agricoles ...	20,7	36,1	23,9	16,6	2,7	7 968	2,3
Patrons de l'industrie et du commerce :							
Industriels	41,3	38,1	11,7	7,5	1,4	850	0,2
Artisans	30,1	40,9	16,5	10,8	1,7	11 266	3,2
Patrons pêcheurs..	29,2	35,8	15,5	18,6	0,9	226	0,1
Grds commerçants	44,1	33,1	13,4	8,3	1,1	904	0,3
Petits commerçants	34,0	37,2	15,7	11,3	1,8	9 040	2,6
Professions libérales - Cadres supérieurs	52,8	31,2	11,1	4,1	0,8	4 187	1,2
Cadres moyens	42,4	37,4	13,5	5,3	1,4	14 373	4,1
Employés	30,8	41,3	17,3	8,5	2,1	37 391	10,8
Ouvriers :							
Contremaîtres	32,9	41,6	15,9	7,5	2,1	12 120	3,5
Ouvriers qualifiés .	22,2	44,0	20,9	10,0	2,9	65 960	19,0
Ouvriers spécialisés	20,4	42,2	22,6	11,5	3,3	52 054	15,0
Mineurs	23,4	43,4	21,0	9,2	3,0	5 068	1,5
Marins pêcheurs ..	29,8	39,7	18,9	8,1	3,5	1 472	0,4
Manœuvres	14,4	39,6	27,6	14,5	3,9	18 910	5,4
Personnel de service .	22,1	44,3	20,6	10,1	2,9	14 775	4,2
Autres catégories	30,9	42,3	16,7	7,9	2,2	9 420	2,7
Sans professions - Retraités	22,5	39,5	23,3	11,3	3,4	23 482	6,8
Orphelins, Pupilles ..	11,1	45,7	26,2	12,5	4,6	4 671	1,3
TOTAL DEFINI..	81 721	129 570	63 626	32 778	8 520	316 215	90,9
Indéterminés	14 443	11 848	2 351	2 275	737	31 654	9,1
TOTAL	**96 164**	**141 418**	**65 977**	**35 053**	**9 257**	**347 869**	**100,0**
(%)	*27,6*	*40,6*	*19,0*	*10,1*	*2,7*	*100,0*	

TABLEAU N° 11

RESULTATS DU BACCALAUREAT

Session	A		B		C		D		D'		E		F		G		H	
	Prés.	Adm.	Prés.	Adm.	Prés.	Adm.	Prés.	Adm.	Prés.	Adm.	Prés.	Adm.	Prés.	Adm.	Prés.	Adm.	Prés.	Adm.
72	83 890	59 404	23 856	15 494	39 627	26 861	64 244	39 333	2 014	1 038	8 956	5 222	24 731	13 736	35 382	22 622	697	486
73	81 278	56 181	29 953	18 033	42 125	28 344	68 502	41 406	2 227	1 238	9 025	5 098	29 017	15 568	38 880	24 794	900	537
74	75 655	54 533	28 098	19 487	41 736	29 900	69 727	43 294	2 250	1 142	8 635	5 094	31 296	17 168	42 399	27 591	809	515
75	72 109	50 436	30 517	20 946	42 676	30 396	72 804	45 304	2 428	1 406	8 299	5 197	33 669	19 113	48 202	31 272	699	419
76*	63 535	44 909	31 384	21 503	43 993	32 313	73 105	44 368	2 123	1 159	8 084	5 178	35 118	19 523	50 218	30 636	707	479

Les effectifs du tableau n° 11 tiennent compte des candidats individuels

TABLEAU N° 11 BIS

RESULTATS AU BACCALAUREAT DES CANDIDATS INDIVIDUELS

Session	A		B		C		D		D'		E		F		G		H	
	Prés.	Adm.	Prés.	Adm.	Prés.	Adm.	Prés.	Adm.	Prés.	Adm.	Prés.	Adm.	Prés.	Adm.	Prés.	Adm.	Prés.	Adm.
72	6 629	3 105	1 200	478	1 358	391	2 727	906	132	40	371	143	922	367	531	258	5	1
73	7 381	3 003	1 416	493	1 719	326	3 057	797	153	52	383	135	1 078	334	1 346	515	27	8
74	5 680	2 746	1 371	504	1 345	368	2 820	860	149	36	325	125	1 346	337	1 279	487	33	7
75	5 458	2 391	1 327	519	1 253	269	2 603	700	201	47	297	117	1 315	411	1 545	572	24	7
76*																		

* Dans le tableau n° 11, il manque les résultats de la session de septembre. Et dans le tableau n° 11 bis, on ne dispose pas encore (début mai 1977) des résultats des candidats individuels pour les deux sessions juin et septembre.

Une brochure A.P.M.E.P. sur le caclul des probabilités et son enseignement :

HASARDONS-NOUS ...

220 pages. Prix : **25 F** (port compris : **31 F**).

Cette brochure est un ouvrage collectif.

S'adressant aux membres de l'A.P.M.E.P., elle se propose d'apporter un outil de réflexion et une aide pour l'enseignement des probabilités, de l'école élémentaire où de nombreuses expériences ont été poursuivies, à l'Université où il faut former de nombreux utilisateurs.

Elle doit donc intéresser à la fois les maîtres du premier degré qui y trouveront des relations d'expériences qui peuvent d'ailleurs être entreprises tout au long du premier cycle, les maîtres du second degré qui pourront approfondir leur réflexion dans une perspective différente de celle des manuels, les universitaires responsables d'un enseignement de probabilités et de statistiques, en particulier à des non-mathématiciens, les candidats aux concours de recrutement, C.A.P.E.S. et Agrégation.

Th. HATT
IREM de Strasbourg

3. INTRODUCTION DES METHODES D'ANALYSE DES DONNEES EN GEOGRAPHIE AU LYCEE

Nous présentons ici un travail d'introduction des méthodes de l'analyse des données dans des classes de secondes et premières de lycée classique. Ce travail, commencé en 1973 avec le soutien de l'IREM de Strasbourg (moyens de calculs en particulier) a été continué sur le matériel dont le lycée a été équipé en 1974 dans le cadre d'une expérience de sensibilisation des élèves à l'informatique à travers les disciplines traditionnelles. L'Institut de Recherche Pédagogique a organisé la formation de 500 enseignants, 58 établissements ont été équipés de matériel informatique. Cette expérience d'introduction de l'informatique ne concerne pas seulement les classes scientifiques.

Le but de l'expérience n'est pas de créer une discipline technique nouvelle mais de promouvoir des "applications" de l'informatique dans les disciplines traditionnelles. Cette optique est originale et ne se rencontre dans aucun autre pays lancé dans ce type d'expérience. Le statut expérimental et les excellentes conditions de travail dont nous disposons au lycée limitent les conclusions que l'on peut tirer d'un tel travail, néanmoins nous essaierons de montrer que l'analyse des données insuffle un **esprit nouveau** dans notre discipline et n'induit pas seulement l'utilisation de moyens informatiques plus ou moins puissants.

Nous nous attacherons à trois aspects essentiels de l'introduction de méthodes statistiques :

1) Pourquoi utiliser avec les élèves de telles méthodes ?
2) Quel est le contexte pédagogique d'introduction de ces méthodes ?
3) Quels sont les exemples effectivement traités avec les élèves (et non pas "qui pourraient l'être") ?

I. Pourquoi utiliser au lycée les méthodes de l'analyse des données ?

Les professeurs d'histoire et géographie ont toujours utilisé peu ou prou les statistiques élémentaires : graphiques pluviométriques et thermiques, pyramides des âges, valeurs relatives ... Ce que nous essayons de pratiquer en classe est plus ambitieux : traiter un problème géographique ou historique le plus complètement possible avec une **DEMARCHE** statistique scientifique. Pourquoi adopter une telle démarche ?

A. UNE EXIGENCE CRITIQUE EN FACE DE LA GEOGRAPHIE TRADITIONNELLE

J. Beaujeu-Garnier, une autorité en matière de géographie écrit dans [2] p. 292 : les géographes se seraient caractérisés dans le passé par "... une description souvent plus subjective et intuitive que chiffrée et logique, une complaisance pour l'attitude littéraire plus que scientifique ...", qui donnait à ceux qui les lisaient, "... l'impression parfois de talent, souvent de minutie, rarement d'efficacité ...".

B. UNE POSITION THEORIQUE QUI MET L'ACCENT SUR L'INTERDEPENDANCE DES FAITS GEOGRAPHIQUES

a) Cette position repose d'abord sur une définition de la géographie en tant que discipline de recherche où l'accent est mis sur les interdépendances simultanées entre les faits géographiques à la surface de la Terre. La "région", échelon spatial subordonné, est décrite comme un "ensemble organisé à variables multiples", donc un vecteur. Dans chacun des cas, sans exclure l'histoire, nous insistons sur une explication cybernétique de la géographie où la référence à la théorie des systèmes et à la notion de modèle est fondamentale. Chaque variable du système géographique est à la fois effet et facteur dans un ensemble maintenu cohérent par les interdépendances entre variables. On est donc amené à considérer le maximum de variables possibles, ce qui conduit naturellement à l'analyse des données.

b) Deux exigences scientifiques nous paraissent importantes :

— un souci d'"objectivité" de la démarche scientifique. Cette "objectivité" ne doit pas être entendue au sens d'A. Comte comme l'exacte reproduction du réel, que nous pensons inaccessible, mais comme "l'application correcte d'un instrument dans le cadre de conventions explicitées ..." (A. VOLLE, *Economie et Statistique,* n° 96, janvier 1978) ;

— un souci de "reproductibilité" du raisonnement. L'un des objectifs que nous nous donnons est d'aboutir à une typologie des objets étudiés. Comment, quel que soit le chercheur, sur des données identiques, aboutir à des résultats identiques ? Comment comparer

des régions ou des stations climatiques sans fluctuation des critères de comparaison ? Sans démarche "reproductible", voire automatisable, pas de comparaison possible, pas de généralisation possible. Les mathématiques, l'informatique et la graphique sont des moyens de formaliser la démarche du traitement scientifique. Il nous paraît très didactique de combiner les trois outils pour une approche de l'analyse des données.

C. METTRE LES ELEVES EN POSITION DE RECHERCHE SCIENTIFIQUE

a) Il y a place dans l'enseignement secondaire pour une telle démarche. Les élèves n'apprennent bien que ce qu'ils trouvent par eux-mêmes, mais d'autre part l'apprentissage par l'expérience est long, comment raccourcir le temps nécessaire ? Il paraît exclu de tout enseigner, l'évolution des connaissances est trop rapide. Il est donc légitime de faire des choix dans le programme et, sur des questions traitées de manière approfondie, mettre les élèves en situation de recherche personnelle active avec des moyens modernes de traitement.

b) Le mode interrogatif nous paraît plus didactique que le mode affirmatif : la géographie n'apporte pas un corpus figé de savoir immuable, c'est une discipline vivante qui change de méthodes, qui pose des questions. Le contenu des manuels n'est pas un dogme tombé du ciel mais un instrument de travail comme un autre pour les élèves, parfois plus critiquable qu'un autre dans la mesure où il n'expose jamais ses présupposés théoriques, ses sources qui permettraient de refaire la démarche avec d'autres méthodes ... Il semble beaucoup plus intéressant pour les élèves de ne pas supposer le problème des types régionaux résolu au départ mais d'élaborer des solutions avec eux. Comme le montre le tableau 1, la démarche statistique est faite de "choix multiples". Formaliser les problèmes oblige à expliciter toutes les étapes, surtout lorsqu'on souhaite programmer la machine, il est beaucoup moins facile que dans le discours littéraire d'escamoter les questions.

c) Les choix multiples de la démarche statistique :

TABLEAU 1

LES CHOIX MULTIPLES DE LA DEMARCHE STATISTIQUE

I QUEL PROBLEME GEOGRAPHIQUE ?
TYPOLOGIE RAISONNEE ET REPRODUCTIBLE D'OBJETS GEOGRAPHIQUES

POURQUOI ?	Un but de connaissance et d'explication scientifique de l'organisation spatiale actuelle.
	Un but de diagnostic prévisionnel et de décision.
	Un but pédagogique d'apprentissage des méthodes qui permettent de tirer parti d'une information chiffrée volumineuse.

II QUELS CHOIX INITIAUX ?

1. Quelle échelle spatiale ?
   Parcelles, communes (traitements par sondages), cantons, régions de programme, unités climatiques ou de végétation ... (traitements exhaustifs) ?
2. Quelle définition des unités spatiales ?
   Une combinaison originale de variables interdépendantes sur une portion d'espace.
3. Pourquoi des mesures de ces variables ?
   La mesure est nécessaire aux **comparaisons** entre objets géographiques.
4. Quelles variables ?
   Quelle définition statistique ? Quelles erreurs de mesure ? Quelles variables absentes ? (Non mesurables ou non mesurées). Quel sens implicite à ces absences ?

III QUEL CODAGE DE CES VARIABLES ?

5. Données brutes (effectifs ...) ?, prétraitées (%, rangs, centrées réduites ...) ?, ventilées en classes (selon quel principe ?), codées en disjonctif ou non ? ...

IV QUELLE MATRICE DES DONNEES ?

6. Faut-il introduire le temps (collection de tableaux) ?
7. Le tableau est-il homogène ? pertinent ? exhaustif ? vaste ? amorphe ?

ETAPE SEMANTIQUE

COMMENT TIRER LE MAXIMUM D'INFORMATION DE CE TABLEAU POUR UNE TYPOLOGIE RAISONNEE ?
Une matrice d'observations est un "super nombre" manipulable par les mathématiques.

V QUEL MODELE MATHEMATIQUE ?

8. Une fois le modèle choisi,quels sont les choix du modèle ?
9. Quelle métrique ?
10. Quelle forme du nuage ? (déterminé par le choix de la métrique). Nuage des points régions dans l'espace des variables et des points variables dans l'espace des régions.
11. Quelles méthodes de radiographie de l'espace multidimensionnel ?
12. Quelles méthodes de filtrage, de classement ?

ETAPE SYNTAXIQUE

VI QUELLES VALIDITE DES RESULTATS ?

13. Critique des variables et des objets. Quel est le rôle des exceptions ?
14. Quelles erreurs de saisie des données ? Quelles variables redondantes ou non significatives ? Quelles combinaisons et interactions de variables importantes ?
15. Quelle critique du classement peut-on faire ?

VII RETOUR CRITIQUE AU TABLEAU

Ne pas oublier que l'on n'a pas étudié des régions géographiques mais la structure d'une matrice de données. Le découpage obtenu n'a de sens que dans le cadre de ce tableau.

ETAPE SEMANTIQUE

Le tableau 1 montre l'importance des phases "amont" et "aval" qualifiées ici de "sémantiques" par rapport à la phase "syntaxique" de l'analyse des données. Quels sont les problèmes scientifiques à résoudre ? Quelle échelle régionale et quelles mesures statistiques choisit-on ? Quels sont les codages des données brutes qui conviennent ? Les méthodes de l'analyse des données s'appliquent à des "données" ; les résultats d'une procédure typologique, l'interprétation que l'on peut en faire dépendent avant tout de la qualité des données de départ ; aussi convient-il d'accorder un soin extrême à cette phase méthodologique initiale. L'utilisation du modèle mathématique suppose close une liste impressionnante de questions préalables, les choix dans cette étape sont pratiquement infinis. Les résultats du traitement seront aussi divers, l'important est d'expliciter et de justifier les choix à chaque étape. Le grand intérêt de cette démarche est d'être anti-dogmatique, elle permet de montrer aux élèves comment on arrive à certains résultats au lieu de les leur faire accepter comme parole divine.

d) Un moyen de tirer parti d'une information chiffrée massive. Grâce aux efforts de l'Institut National de la Statistique et des Etudes Economiques, les données numériques de l'économie tombent de plus en plus dans le domaine public. Il existe des méthodes très puissantes de traitement de cette information chiffrée, certaines sont déjà dans le grand public (Enquêtes du Nouvel Observateur utilisant l'analyse des correspondances). On peut penser qu'il est utile que les élèves soient initiés à l'utilisation de ce genre de méthodes.

Cet ensemble d'arguments prouve que l'adaptation à la géographie des méthodes de l'analyse des données correspond à un grand nombre d'objectifs. Il est évident que, pour espérer une généralisation de ces méthodes, il faudrait que la revendication de formation des enseignants soit prise en compte.

II. Contexte pédagogique d'introduction de l'analyse des données

A. MATERIEL ET LOGICIEL AU SERVICE DE L'EXPERIENCE(*)

a) un matériel adapté à l'enseignement

La configuration informatique étudiée pour l'Education Nationale est particulièrement bien adaptée aux applications pédagogiques. La seule lacune du point de vue du géographe est l'absence d'une imprimante rapide.

(*) Pour plus de détails on pourra consulter [9].

Le cœur est un mini-ordinateur à mémoire centrale de 8K(**) octets (4,2K octets pour le système d'exploitation en temps partagé et 3,8K octets utilisateur). Cette capacité est faible : sans aucun programme il est possible de stocker un millier de valeurs numériques. Cette faiblesse oblige, pour traiter les gros problèmes de fichiers géographiques, à une programmation particulière dite en "mémoire externe". Les données découpées logiquement sont cherchées par programme à la demande sur le disque magnétique à têtes fixes de 400K octets. Ceci augmente le temps de calcul mais le temps machine n'est pas facturé et, l'ordinateur pouvant tourner 24 h sur 24, la difficulté n'est pas grave. La capacité du disque à têtes fixes permet, par exemple, le stockage du système de traitement statistique SESAM, dont nous parlerons plus loin, 70000 octets de programmes, 205000 octets de fichiers, le reste étant occupé par le système d'exploitation. Le disque fixe peut être transféré à grande vitesse sur un disque souple ou "disquette" bon marché. Chaque discipline possède ainsi ses propres disquettes. Le télétype,à lecteur perforateur de ruban, est le point faible de l'installation, lent et difficile à régler.

La salle de travail des élèves est équipée de huit téléviseurs à clavier, ce qui permet de travailler avec des demi-classes de 16 élèves. Ecrans et claviers sont totalement silencieux, le matériel bruyant est dans la salle voisine, ceci procure un grand confort de travail. Ce matériel est peu sujet aux pannes, l'Education Nationale paie une maintenance-assurance qui garantit des dépannages rapides.

b) un logiciel interactif puissant

L'intérêt de ce matériel est renforcé par un logiciel interactif mis au point par l'équipe de J. Hebenstreit à l'Ecole Supérieure d'Electricité. Ce Langage Symbolique d'Enseignement (LSE) est commun à toutes les machines de l'expérience, ce qui permet les échanges de programmes. C'est un langage dont les caractéristiques sont très intéressantes pour le géographe-programmeur :

— facilité de programmation des calculs matriciels
— facilité de gestion des fichiers de données
— puissance des traitements des chaînes de caractères, ce qui se prête aussi bien aux tests des réponses d'élèves qu'aux applications graphiques
— dimensionnement dynamique des tableaux, variables logiques, procédures externes ...

Ce langage permet le dialogue de l'élève et du programme à la console, chaque élève pouvant progresser à son propre rythme.

(**) $K = 2^{10} = 1024$ ainsi $2K = 2048$ et $8K = 8192$ (K : abréviation de kilo).
1 octet $= 2^8$ unités binaires.

B. LE TRAVAIL D'EQUIPE ENTRE MATHEMATICIENS ET GEOGRAPHES

Le travail d'équipe est absolument indispensable. Le géographe n'a pas les compétences suffisantes pour utiliser sans appui extérieur les méthodes d'analyse des données. Le travail d'équipe au lycée est également indispensable pour d'autres raisons.

a) le travail d'équipe à l'IREM

L'IREM nous a toujours donné le meilleur accueil en nous recevant comme stagiaire puis comme animateur d'un groupe informatique. Les spécialistes d'analyse des données présents à Strasbourg nous ont permis de surmonter bien des obstacles techniques. Si cette collaboration pluridisciplinaire est possible, c'est en partie grâce aux structures ouvertes et non hiérarchiques de l'IREM.

b) au lycée

La situation est moins bonne. Depuis longtemps nous recherchons des collègues qui acceptent de traiter dans certaines classes communes des aspects statistiques. Le plus souvent les demandes que nous avons faites se sont soldées par un échec. Deux raisons sont invoquées :

— le programme chargé qui ne laisse aucune place pour ce qui n'y figure pas ou en a été enlevé comme ... les statistiques !

— le manque de formation initiale qui oblige à trop de travail pour élaborer un cours. L'exemple du premier collègue qui ait accepté n'est guère probant : particulièrement motivé, il est engagé dans l'expérience informatique.

Il est très regrettable qu'une collaboration même symbolique soit si difficile à obtenir. Les élèves sont très sensibles au fait de retrouver dans des matières différentes des notions identiques traitées de la même manière quoiqu'avec des points de vue différents. Cette possibilité de faire référence à des notions traitées ailleurs nous paraît très importante. Il ne s'agit pas de trouver chez le mathématicien une caution à une attitude nouvelle du géographe mais de montrer aux élèves la nécessité d'ouvrir, parfois pendant la même heure de cours, deux tiroirs différents de l'emploi du temps.

C. LA PRATIQUE PEDAGOGIQUE AVEC LES ELEVES

a) des méthodes actives

Les élèves prennent en charge l'élaboration complète d'un "dossier" tiré du programme. Le professeur choisit dans les sujets du programme des questions assez larges, il prépare le plan de travail, la liste des exposés et des documents. Les dossiers : "Oppositions régionales de l'agriculture ou de la démographie ou de l'industrie française", "la

révolution industrielle de 1800 à 1914", "les séismes, les volcans et la dérive des plaques lithosphériques", "les climats et les types de temps", etc., sont traités par périodes bloquées d'une vingtaine à une trentaine d'heures. Les cours magistraux (introductions, méthode de travail ...) sont réduits au minimum. Les travaux pratiques sur ordinateur et les travaux indépendants par équipes sont très nombreux. Les équipes, responsables de comptes rendus devant la classe, ont en main une grande variété de documents : articles du journal "Le Monde", de la "Recherche", "Science et Avenir", "Economie et Statistique", des ouvrages universitaires, des dictionnaires économiques, des manuels, etc. Les articles sont regroupés autour de thèmes. Le dossier "Oppositions démographiques des régions françaises" en comprend 5 :

1. Les moyens de connaissance de la population. Quelles sont les sources chiffrées ? Quelle est leur précision ? Quelles sont les définitions exactes des termes démographiques ? (6 articles).

2. Fécondité et population (6 articles).

3. Mortalité et population (8 articles).

4. La croissance démographique, simulation sur ordinateur (programme de simulation simple à deux paramètres), T.P. par équipe.

5. Les oppositions régionales démographiques : évolutions et structures ; analyse quantitative des structures démographiques des régions de programme et des cantons alsaciens.

Les paragraphes **1** à **4** constituent la préparation du traitement quantitatif du paragraphe **5**. Les élèves ont en main en abordant ce cinquième paragraphe les exposés des articles et un glossaire de définitions. La même méthode est suivie pour les dossiers de géographie générale et d'histoire. Les résultats des comptes rendus des équipes sont polycopiés et distribués à tous. A la fin du dossier les élèves ont en main un gros document de 40 à 100 pages selon le sujet traité.

b) les problèmes avec les élèves

Les élèves sont séduits par la nouveauté de la manipulation technique, par le calme de la demi-classe, l'accent mis sur la responsabilité et le travail d'équipe. Un certain nombre de problèmes se posent néanmoins : les élèves ont de la peine à s'insérer dans un dossier qui a été conçu en dehors d'eux, et pourtant seule une préparation soignée permet de mettre à leur disposition des documents originaux et adaptés. Beaucoup jugent difficile, abstraite, ingrate la démarche statistique, jamais ils n'ont abordé de cette manière les questions de géographie et le tiroir "mathématique" est refermé quand s'ouvre l'heure de géographie. Pourtant la satisfaction des élèves est grande lorsque, à l'issue du cours commun, ils rencontrent en géographie des notions abordées en mathématiques. Voici l'exemple de quelques notions utilisées en classe de première dans le dossier de géographie quantitative :

— vecteur, matrice, projection plane
— graphe orthonormé, échelle millimétrée, logarithmique, log-log, coordonnée, rapport
— histogramme, classe, amplitude, profil
— moyenne, variance, écart-type, variable centrée-réduite, corrélation
— distance euclidienne (comme cas particulier de distance de Minkovski *), similitude, arbre hiérarchique, partition, algorithme de classification
— termes techniques de l'analyse des données, ACP**, AFC***, nuées, etc.

Certaines notions paraissent difficiles à certains élèves : la distance euclidienne dans le cas général avec une écriture indicée, pour les élèves de seconde tout ce qui se rapporte au dessin d'un graphique cartésien cadré avec calcul de l'échelle, tout ce qui touche la notion de rapport et d'échelle. Il faut un entraînement intensif pour que les élèves arrivent à placer 15 points sur un graphique en une demi-heure. La notion d'arbre, qui paraît facile à expliquer avec l'arbre généalogique, leur semble difficile et ils placeront le fils au niveau du père ...

Nous avons l'impression qu'une certaine facilité de la géographie traditionnelle où "il suffit d'apprendre", où "il n'y a rien à comprendre", rend l'approche statistique plus difficile qu'elle n'est en réalité.

III. Quelques exemples d'application de l'analyse des données avec les classes

Nous pensons qu'il faut d'abord traiter avec les élèves des exemples qu'ils connaissent (les 22 régions de programme sont traitées en classe de troisième) et qui soient praticables à la main (30 cantons d'une région, 16 stations climatiques ...).

Nous mettons d'autre part l'accent sur les méthodes graphiques de façon à faire sentir aux élèves aux bouts de leurs doigts les algorithmes de classification dans les cas univariable et bivariable.

---

* Une distance de Minkovski entre deux objets i et j pour lesquels on dispose d'une mesure X sur chacune des n variables (k = 1,...,n) se définit comme suit :

$$d_{ij} = \left(\sum_{k=1}^{n} |X_{ik} - X_{jk}|^p \right)^{1/p} \quad \text{où} \quad p \geq 1 .$$

** ACP : Analyse en composantes principales.

*** AFC : Analyse factorielle des correspondances. Ces deux méthodes d'analyse factorielle seront décrites en détail dans le Tome II.

A. LES EXEMPLES TRAITES A LA MAIN

a) Analyser une ligne ou une colonne de la matrice des données.

On commence par ne fournir aux élèves qu'une ligne ou une colonne du tableau des données : 22 régions de programme décrites par la densité de 1975 ou par le pourcentage d'habitants en zone urbaine, une station décrite par douze températures et douze précipitations, seize stations décrites par les températures du mois de juillet ...

Les élèves tracent à la main sur un axe orienté les coordonnées de chaque objet. Les points s'organisent (figures 2 et 4 p. 74) sur cet axe en une variation continue. Les points sont parfois regroupés, "proches", et des zones de moindre densité séparent des zones plus denses. Comment analyser ces "distances" entre points ? La méthode classique, couramment pratiquée, consiste à diviser la variable en classes d'égales amplitudes (figure 5). Cette première approche permet de comparer plusieurs cartes (à condition que la variable ait été normalisée, par exemple centrée-réduite) grâce à une échelle graphique commune (on trouvera un exemple de ce type d'échelle figure 7). La comparaison de la carte et de l'axe montre souvent des "anomalies" : certaines régions proches sur l'axe sont pourtant placées dans des classes différentes simplement parce que les limites de classes les séparent aveuglément. Comment regrouper des régions "proches" en évitant ce genre d'écueil ?

b) Construction d'un arbre hiérarchique à la main.

L'arbre hiérarchique apporte une réponse souple et puissante à la question : combien de types ? La méthode est indépendante de l'opérateur, donc automatisable, elle permet une bien meilleure adaptation aux données. Il ne faut pas oublier néanmoins que toute méthode d'analyse des données fait deux choses à la fois : mettre en évidence une structure propre au tableau et mettre en lumière une structure propre à la procédure elle-même.

Trois règles d'agrégation sont courantes en matière d'analyse hiérarchique, on les trouvera exposées en détail dans l'article de J.P. Letourneux (**4**) : la règle de la distance moyenne ou lien moyen, du diamètre ou saut maximum, du lien minimum ou saut minimum. Les règles du diamètre et du lien minimum sont les seules à pouvoir être pratiquées par les élèves à la main car elles n'obligent à calculer que des différences.

ALGORITHME DU SAUT MINIMUM

Dans chaque regroupement de classe,

1. *Ordonner les objets en séquence croissante et calculer la valeur absolue de la différence de valeur de la variable entre deux objets successifs.*
   Tracer en abscisse une échelle de distance entre classes en retenant comme distance la plus petite des différences trouvées. Considérer au départ chaque objet comme une "classe" formée d'un seul individu.
2. *Tant que toutes les classes n'ont pas été fusionnées en une seule, examiner toutes les paires de classes adjacentes, repérer la différence courante la plus faible et fusionner les deux classes ainsi repérées.*

La figure 8 donne l'exemple du pourcentage de population 1975 en zone de population industrielle et urbaine traité de cette manière. Le tableau des données se trouve figure 6 (4e colonne). Les régions de la Loire et du Centre ont le plus faible niveau de différence (0,3 %) ainsi que l'Aquitaine et la Bourgogne, aussi ces régions sont-elles réunies d'abord. Au niveau 0,4 % la Provence-Côte d'Azur et le Nord sont réunies, puis Midi-Pyrénées et Auvergne au niveau 0,7, etc.

Dans chaque regroupement de classe, c'est la région la plus "proche" qui devient représentant de la classe : Roussillon pour la classe {ROU, FCT, CHA} regroupée à {PIC, HNO} ...

Comment utiliser cette représentation ?

Si l'arbre est coupé au niveau des "feuilles" (différence 0 %), aucune des régions n'est fusionnée à une autre, il y a autant de types que de régions différentes. Si l'arbre est coupé au niveau du "tronc" (différences entre classes supérieures à 6,3 %) tout l'arbre tombe, les régions sont toutes réunies en un type unique.

La typologie retenue se situe bien entendu quelque part entre ces deux extrêmes.

Le choix sera fonction des objectifs, du degré de perte de détail accepté, du degré de précision des données. *Il sera peu raisonnable de couper l'arbre à un niveau de différence trop faible pour la qualité des données : le sondage au 1/5e du recensement de 1975 entraîne des erreurs ; si une différence de 2 % est trouvée entre deux régions, alors que la population n'est connue qu'à 3 % près, il faudra couper l'arbre à un niveau de différence plus grossier. On trouve un exemple de résultats en 7 types figure 9 p. 79.*

c) Les graphes traditionnels de températures, de précipitations, les pyramides d'âges sont aussi demandés aux élèves.

L'attitude "analyse des données" consiste à les leur faire tracer tous à la même échelle et à les reporter sur calque dans le but de faire des comparaisons.

d) Analyse simultanée de 2 lignes ou 2 colonnes de la matrice des données.

Le dessin successif de chaque ligne ou de chaque colonne du tableau présente un inconvénient majeur : on perd chaque fois l'information des autres lignes ou colonnes. Le graphe bivariable permet de surmonter cet obstacle et d'étudier le nuage de dispersion des objets dans l'espace de deux variables ou des variables dans l'espace de deux objets ;de même apparaît la relation, la "corrélation" qui lie deux variables.

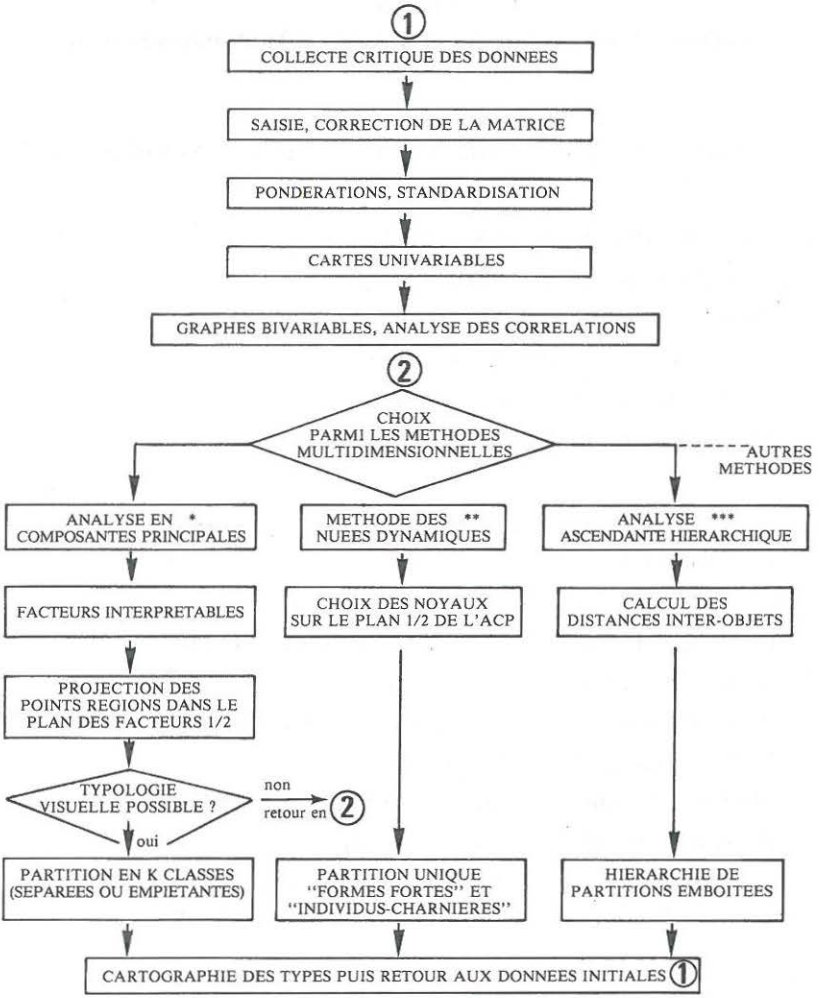
Les méthodes ne changent pas, que l'on s'intéresse aux lignes ou aux colonnes. Les techniques vues précédemment sont utilisables, la mesure des distances entre points sur le graphe, au lieu de se faire par les différences, se fait au compas. Les figures 10, 11, 23 et 24 présentent un exemple de résultats.

Une ou deux variables, il est patent que cela est tout à fait insuffisant pour caractériser les systèmes complexes que sont les climats ou les systèmes de culture. Il est impossible a priori de remplacer 12 ou 24 mesures par deux ou trois.

Le traitement des 22 régions de programme décrites par 12 variables démographiques (tableau des données figure 13) supposerait par des méthodes traditionnelles le dessin de 12 cartes univariées + 66 cartes bivariables = 78 documents à interpréter et comparer, tâche évidemment impossible. Il n'y aurait que 78 cartes à condition de négliger le traitement des variables ... La tâche devient rapidement inexécutable, c'est là qu'intervient l'ordinateur.

TABLEAU 2

DEMARCHE DE TRAITEMENT DE LA MATRICE GEOGRAPHIQUE



* cf. Tome II
** cf. 5
*** cf. 4

TABLEAU 3

LES PRINCIPALES FONCTIONS DE LA BANQUE DE DONNEES SESAM

SESAM

Système d'Etudes Spatiales et d'Analyse Multidimensionnelle

I SAISIE, EDITION, CORRECTION DE LA MATRICE DES DONNEES

II PROGRAMMES DE SERVICE :

Manipulations sur les matrices

Calcul de valeurs relatives

Calcul des valeurs centrées-réduites

Transformation en tableau de rangs

Dédoublement de tableau

III TRACE AUTOMATIQUE D'UN NUAGE DE POINTS ETIQUETES

IV PROGRAMMES DE TRAITEMENTS STATISTIQUES

Analyse en composantes principales

Analyse ascendante hiérarchique (lien minimum, moyen, maximum)

Méthode Iphigénie (M.L. BESSON)

Méthode des nuées dynamiques (BENZECRI-DIDAY)

Régression simple, analyse des résidus

Information significative d'un grand tableau (A. VOLLE)

Analyse factorielle des correspondances (BENZECRI)

V CARTOGRAPHIE AUTOMATIQUE POUR POINTS DISTRIBUES IRREGULIEREMENT

Interpolation linéaire

Ajustement d'une surface de tendance de degré n

Transformation de Fourier

B. LES EXEMPLES TRAITES AVEC L'ORDINATEUR

L'auteur a rédigé pour l'expérience INRP un ensemble de programmes informatiques qui permettent de faire tourner au lycée une banque de données régionales. Ce système publié sous le nom de SESAM (Système d'Etudes Spatiales et d'Analyse Multidimensionnelle) assure la saisie, la mise à jour, le traitement graphique et statistique de fichiers de données compatibles. Le système est facile à manier sur le plan informatique, affichant à la demande de l'utilisateur les fonctions disponibles ainsi que les noms des fichiers courants utilisables (tableau 3). Ce sont les élèves qui font tourner les programmes sauf lorsque le traitement suppose une cuisine technique trop complexe. Le tableau 2 donne un exemple de cheminement possible dans le système. L'ensemble des 100 modules actuels du système occupe 70 000 octets du disque à têtes fixes, les fichiers données, rentrés par les élèves en 4 ans, de l'ordre de 205 000 octets.

a) L'exemple des régions démographiques.

L'analyse multidimensionnelle systématise les comparaisons de cartes que le géographe intégrait traditionnellement par son intuition et sa connaissance du problème. Les premiers exemples donnés aux élèves sont volontairement bien typés, les classifications faciles à voir. La projection plane dans le plan factoriel (figures 14, 15) corroborée par l'analyse hiérarchique visualise ce que dissimulait l'opacité multidimensionnelle. Le travail de comparaison automatisé débouche sur une carte de synthèse (figure 16). Le deuxième exemple est volontairement plus difficile (figures 18, 19). Les cantons périphériques sont bien caractérisés comme le montrent les profils ; ce n'est pas le cas du noyau central. Les élèves cartographient directement les scores factoriels (figures 20, 21).

b) L'exemple des stations climatiques.

Le tableau 80 stations × 24 mesures est traité avec des élèves de seconde. Les équipes tapent les données aux consoles téléviseurs. Les programmes font le traitement inabordable à la main. Les sorties télétype sont simples au début : graphes bivariables croisant les températures de juillet et de janvier ou les précipitations de deux mois. Ces graphes montrent les problèmes d'échelle qui conduisent à tronquer la matrice des précipitations à 350 mm (22 mois sur 960 (80 × 12) dépassent 350 mm). Les résultats du traitement par l'analyse factorielle des correspondances se trouvent figure 25. Sur notre machine le programme d'AFC ne permet pas de traiter avec 1.9 Kmots plus de 30 variables. Or, si l'on veut traiter des croisements de mesures climatiques, on est amené à dédoubler les variables, en particulier pour donner même poids à

toutes les stations. Ceci signifie que si une variable x prend ses valeurs sur l'intervalle [0, a], on la double dans le tableau par la variable $a-x$. Ainsi notre tableau aurait 48 colonnes, à partir des 24 colonnes initiales. On dépasserait la limite matériellement imposée des 30 variables. Pour contourner cet obstacle, nous avons procédé en 2 temps : d'abord des traitements séparés des températures et des précipitations, puis un traitement global de l'information extraite de ces études séparées.

Précisément nous avons opéré comme suit :

Le tableau des températures contient, au départ, des nombres négatifs, ce qui est une contre-indication à l'emploi de l'analyse factorielle des correspondances. Nous avons donc opéré une translation du tableau des mesures, amenant le minimum à zéro. Le maximum est alors à 80. Chaque mesure de température t, en degrés, donne donc lieu à des valeurs $x = t + 40$ et $y = 80 - x$. C'est le tableau ainsi obtenu, comportant 24 variables, qui est soumis à l'analyse. On en extrait les deux premiers facteurs qui retiennent respectivement 88,6 % et 10,6 % de la variance, donc pratiquement toute l'information (99,2 % de la variance totale).

Le tableau des précipitations est tronqué à 350 mm : c'est-à-dire que les valeurs qui y figurent se trouvent toutes dans l'intervalle [0,350] ; les valeurs qui initialement dépassaient 350 mm sont ramenées à ce maximum de 350. La raison de cette limitation est le désir de ne pas "noyer" les petites valeurs sous les "déluges". D'autres techniques de codage étaient d'ailleurs possibles.

Du tableau, dédoublé, nous avons extrait 3 facteurs retenant 64 %, 22 % et 6,3 % de la variance soit 92,3 % du total. Les 5 facteurs (2 pour les températures et 3 pour les précipitations) ont été transférés dans un tableau qui à son tour est translaté et dédoublé. La normalisation par les valeurs propres n'a pas été jugée utile, les variances tirées des deux tableaux de départ étant du même ordre de grandeur. L'analyse finale retient 75,6 % de la variance. Sur la figure 25 :

— les stations s'organisent en continuum en fonction de critères que l'on devinait déjà dans les graphes bivariables, on trouve par importance décroissante :

1) un axe de précipitations croissantes dont le poids est le plus fort pour les stations les plus humides qui sont aussi les moins nombreuses ;
2) un axe de température croissante ; ces deux derniers axes sont indépendants ;
3) un axe d'opposition entre les stations dont les précipitations ont un maximum d'hiver et celles dont le maximum est en été.

— les stations se distribuent en variation continue en fonction de ces trois axes descriptifs. Certaines zones plus denses permettent de tracer une typologie à main levée.

— certaines curiosités apparaissent : la réunion de stations côtières et de stations de montagne ou de stations côtières et de stations continentales. L'AFC ne les a pas séparées car les critères d'altitude et de continentalité, importants en climatologie traditionnelle, ne figuraient pas dans la matrice des données. Toute typologie dépend étroitement des critères choisis pour l'élaborer. Il ne faudrait pas en rester là mais compléter l'étude par l'analyse hiérarchique comme on l'a vu précédemment ou par les nuées dynamiques comme dans l'exemple suivant.

c) Les systèmes de culture des cantons alsaciens.

Les combinaisons de culture sont un exemple typique d'interactions complexes qui relèvent de l'analyse des données. 31 cantons alsaciens sont ici décrits par 14 types de culture en % de la surface agricole utile. La projection factorielle permet de déterminer parmi les cantons bien représentés des centres plausibles pour initialiser l'algorithme des nuées dynamiques.

Conclusion

La démarche que nous avons exposée ici nous paraît passionnante par les questions qu'elle soulève à chaque pas. Le changement de langage, la possibilité de traiter de manière algorithmique des données en grandeur réelle avec les élèves introduisent à notre avis une rupture pédagogique et méthodologique. Pour certains, l'analyse des données représenterait une sorte de confort, de fuite devant la recherche théorique originale ; il nous semble au contraire que ces méthodes permettent d'aborder des thèmes qui n'étaient auparavant pas même imaginables. Le choc en retour est considérable dans la mesure où même sans équipement informatique l'attitude du chercheur devant les données géographiques est complètement modifiée.

Bibliographie

I. *Introduction à l'évolution récente de la géographie*

Le numéro 2 vol. XXVII, 1975, de la Revue Internationale des Sciences Humaines, Unesco : "L'utilité de la géographie" avec les contributions de :

[1] G. SAUTTER : *Quelques réflexions sur la géographie en* 1975, p. 245-263.

[2] J. BEAUJEU-GARNIER : *Les géographes au service de l'action,* p. 290-302.

[3] P. GOULD : *Les mathématiques en géographie : révolution théorique ou apparition d'un nouvel outil ?,* p. 319-347.

[3bis] OUVRAGE COLLECTIF : *L'espace rural français,* Masson 1978.

II. *Initiation statistique*

[4] GROUPE CHADULE : *Initiation aux méthodes statistiques en géographie.* Masson 1974, 191 p. Avec l'ouvrage de Racine et Reymond seule initiation en français.

[5] J.-B. RACINE, H. REYMOND : *L'analyse quantitative en géographie.* PUF, coll. SUP, 1973, 316 p. Met l'accent sur l'analyse factorielle ; important quant à la critique des méthodes même s'il n'est pas complet.

III. *Applications au second degré*

[6] Th. HATT : *Informatique, statistique et géographie quantitative au lycée.* Information géographique, mai-juin 1977, p. 131-148.

[7] Th. HATT : *SESAM, système d'étude spatiale et d'analyse multidimensionnelle.* Manuel d'utilisation. INRP, 1977, 69 p.

[8] Y. GUERMOND : *Jalons pour un renouvellement de l'enseignement de la géographie : l'utilisation de modèles.* CRDP, Rouen, 1974.

[9] *L'Informatique dans l'enseignement secondaire.* Bulletin de liaison, numéros 1 à 15. INRP, 1972-79.

IV. *Source statistique*

[10] INSEE : publication annuelle de SIRF : *Statistiques et indicateurs des régions françaises.*

[11] Ch. P. PEGUY. *Manuel de climatologie, 2e édition.* Masson, 1970.

Figures

FIGURE 1 : CORRESPONDANCE ENTRE LES REGIONS ET LES DEPARTEMENTS

CODE A TROIS LETTRES DES REGIONS DE PROGRAMME

Les départements ont été regroupés en 21 *circonscriptions d'action régionale* (décret n° 60-516 du 2 juin 1960) en vue de l'application des plans régionaux de développement économique et social et d'aménagement du territoire.
Le nombre des circonscriptions d'action régionale a été porté au début de l'année 1970 (décret n° 70-18 du 9 janvier 1970) à 22 en scindant la région de Provence -Côte d'Azur - Corse en Provence - Côte d'Azur d'une part, Corse d'autre part.
La loi n° 75-356 du 15 mai 1975 a créé sur le territoire de la Corse les deux départements de Corse du Sud (arrondissements d'Ajaccio et de Sartène) et de Haute-Corse (arrondissements de Bastia, de Calvi et de Corte) et supprimé celui de la Corse.
Depuis la loi n° 72-619 du 5 juillet 1972, le terme de *région* désigne l'établissement public créé dans chaque circonscription d'action régionale, à l'exception de la région parisienne.

FIGURES 2, 3, 4, 5 : QUELQUES MODES DE REPRESENTATION UNIDIMENSIONNELLE

Figure 2 :
Densité des 22 régions de programme en 1975 (tableau des données figure 6) sur un axe orienté. La différence des valeurs de la densité pour le Nord et la région parisienne est telle qu'il a fallu changer d'échelle.

Figure 3 :
Le profil est un autre mode de représentation couramment pratiqué. Les régions sont classées par densité décroissante. Le niveau français permet de faire une typologie grossière (au-dessus et en-dessous de la moyenne).

Figure 4 :
Part de la population en ZPIU sur un axe orienté.
Ces graphes sont destinés à mettre en valeur auprès des élèves le caractère continu de la variation des variables géographiques et la difficulté qu'il y a à trouver des critères "objectifs" de découpage de ce continuum.

Figure 5 :
La carte montre un exemple de découpage d'une variable. Cette dernière (% en ZPIU), centrée-réduite, a été découpée en 7 classes d'égale amplitude. La légende des classes graphiques se trouve figure 7. L'intérêt essentiel d'un tel découpage est de permettre les comparaisons de cartes.

FIGURE 2 : DENSITE TOTALE (1975)

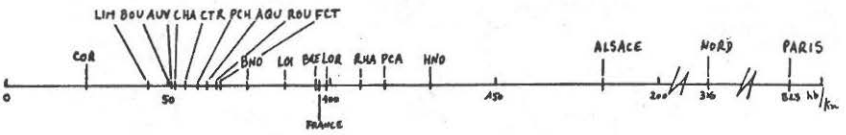


FIGURE 3 : DENSITE TOTALE (1975) profil de la variable

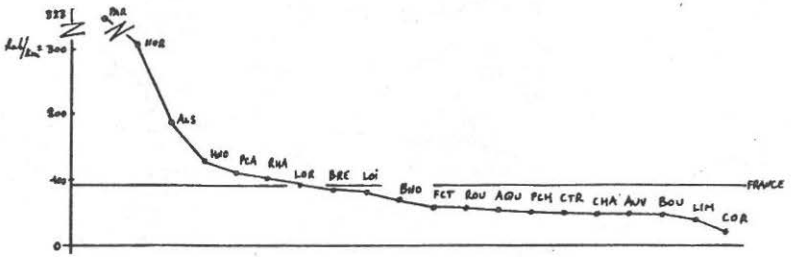


FIGURE 4 : PART DE LA POPULATION HABITANT DANS LES COMMUNES QUI APPARTIENNENT A UNE ZONE DE PEUPLEMENT INDUSTRIEL ET URBAIN (1975)

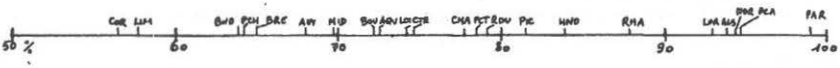


FIGURE 5 : LA CARTE UNIVARIEE

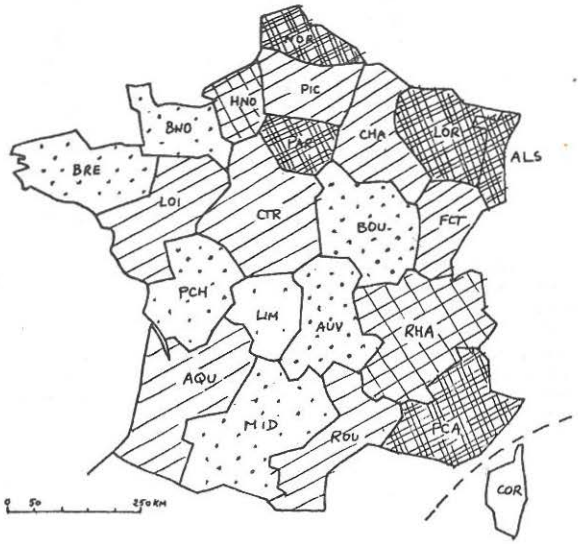


FIGURE 6 : 22 REGIONS DE PROGRAMME, 2 VARIABLES DEMOGRAPHIQUES

VARIABLES BRUTES ET CENTREES REDUITES

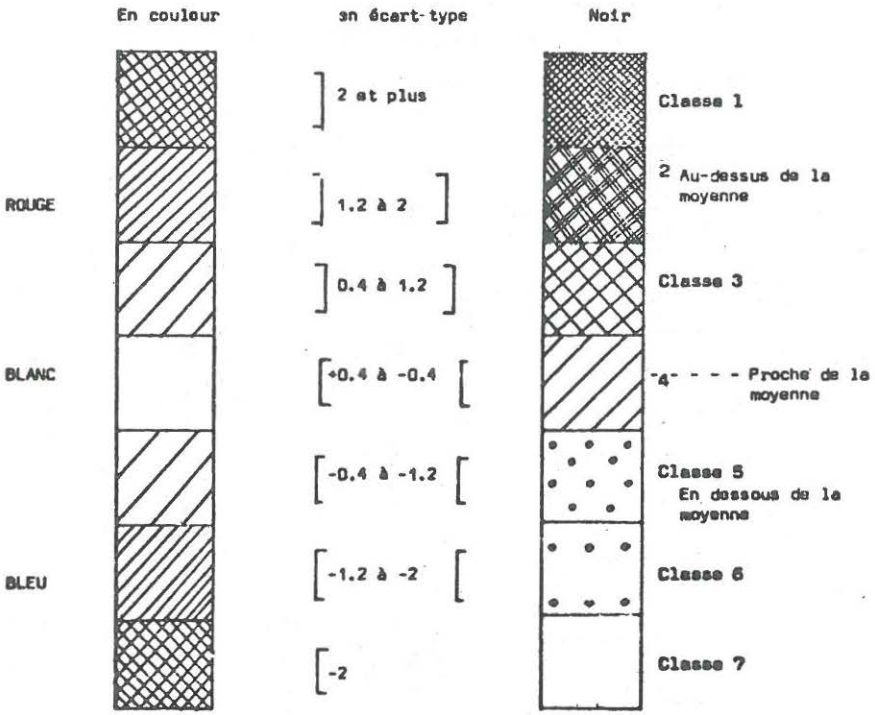
Variable 1 : densité au km 2 (1975)
Variable 2 : part de l'ensemble de la population des communes appartenant à une zone de peuplement industriel et urbain (1975)

1	PAR	823.0	99.3	4.17	1.75
2	CHA	52.0	77.8	−0.42	0.04
3	PIC	87.0	81.7	−0.21	0.35
4	HNO	130.0	83.8	0.04	0.52
5	CTR	55.0	74.4	−0.40	−0.23
6	BNO	74.0	63.9	−0.29	−1.07
7	BOU	50.0	72.1	−0.43	−0.42
8	NOR	316.0	94.3	1.15	1.13
9	LOR	99.0	92.8	−0.14	1.23
10	ALS	183.0	93.7	0.36	1.30
11	FCT	66.0	78.5	−0.34	0.09
12	LOI	86.0	74.1	−0.22	−0.26
13	BRE	96.0	65.1	−0.16	−0.97
14	PCH	59.0	64.2	−0.38	−1.04
15	AQU	62.0	72.4	−0.36	−0.39
16	MID	50.0	68.8	−0.43	−0.68
17	LIM	44.0	57.6	−0.47	−1.57
18	RHA	109.0	87.9	−0.08	0.84
19	AUV	51.0	68.1	−0.43	−0.73
20	ROU	65.0	79.2	−0.34	0.15
21	PCA	117.0	94.7	−0.03	1.38
22	COR	25.0	56.5	−0.58	−1.66

La ZPIU est un ensemble de communes caractérisées par :
— homogénéité de peuplement (peu d'agriculteurs)
— importance d'échanges de travail
— existence d'activités industrielles
Toute unité urbaine fait donc partie d'une ZPIU. 8,7 millions de personnes vivent en dehors d'une ZPIU. Toutes les communes hors ZPIU sont rurales.

FIGURE 7 : CLASSES GRAPHIQUES CARTOGRAPHIABLES D'UNE MATRICE QUANTITATIVE

Solution choisie : découpage arbitraire en 0,8 écart-type.
L'échelle graphique de droite est utilisée pour cartographier les classes issues d'un traitement par analyse hiérarchique.



FIGURES 8, 9 : TYPOLOGIE REGIONALE PAR ANALYSE HIERARCHIQUE D'UNE SEULE VARIABLE (% EN ZPIU)

Figure 8 :
L'arbre est construit graphiquement par les élèves à la main selon l'algorithme exposé dans le texte. Le procédé est celui du saut minimum. L'arbre permet une analyse détaillée très souple de la variable : écart entre Limousin, Corse et toutes les autres régions, similitude du groupe Provence, Nord, Alsace ... En coupant l'arbre à la hauteur du pointillé (niveau de différence de 3,5 % entre classes) on peut dessiner la carte en 7 classes de la figure 9.

Figure 9 :
La carte en 7 types d'urbanisation décroissante est une des exploitations possibles de l'arbre de la figure 8.

FIGURE 8 : L'ARBRE

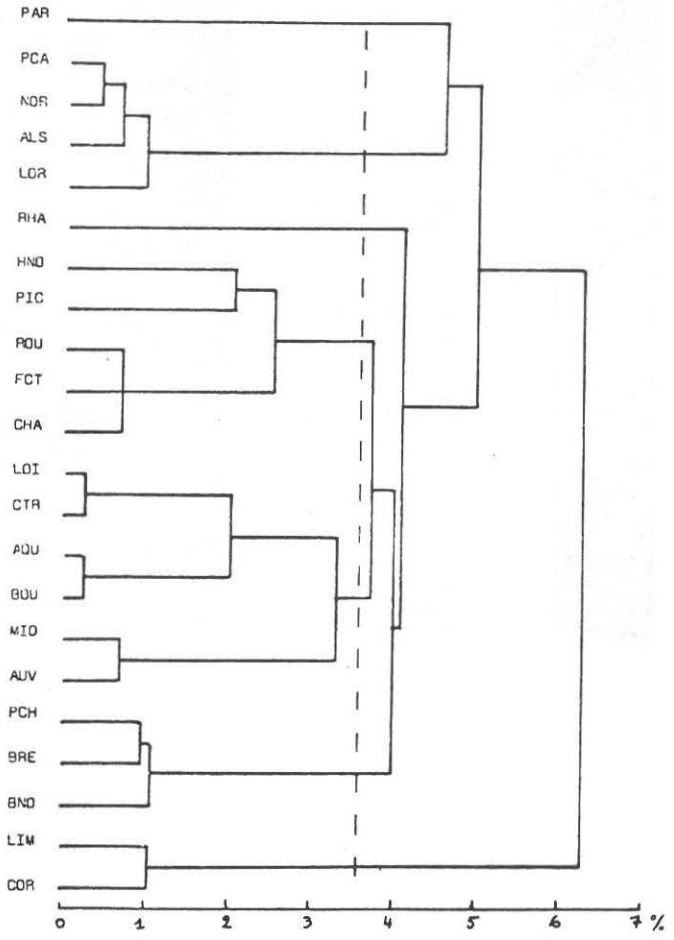
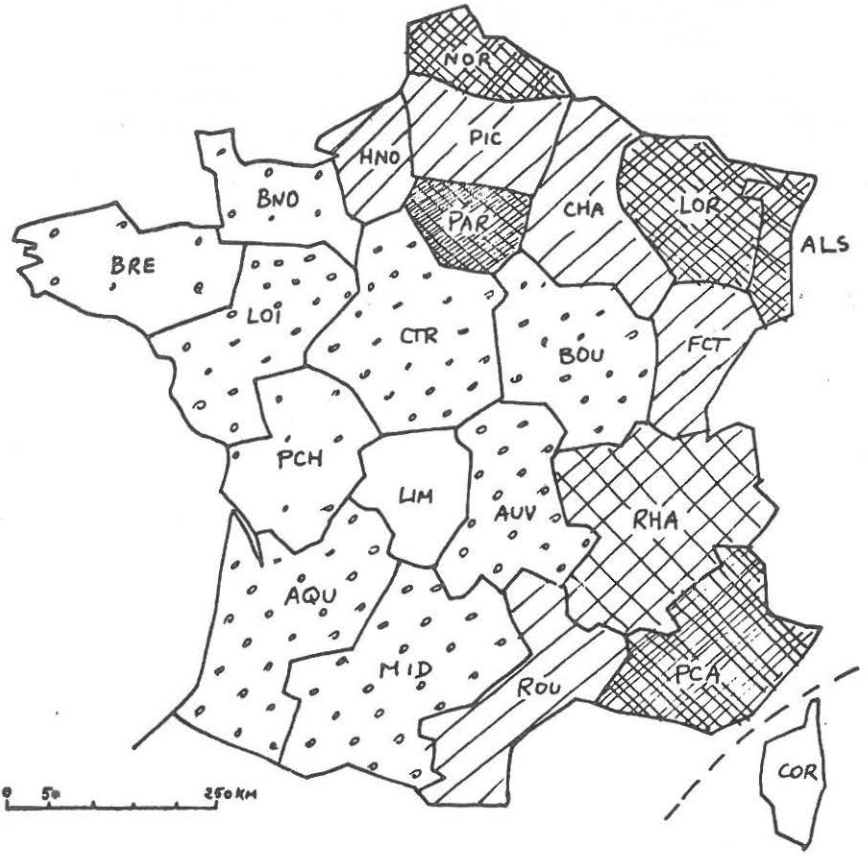


FIGURE 9 : LA CARTE



Légende des classes graphiques commune à toutes les cartes, figure 7.

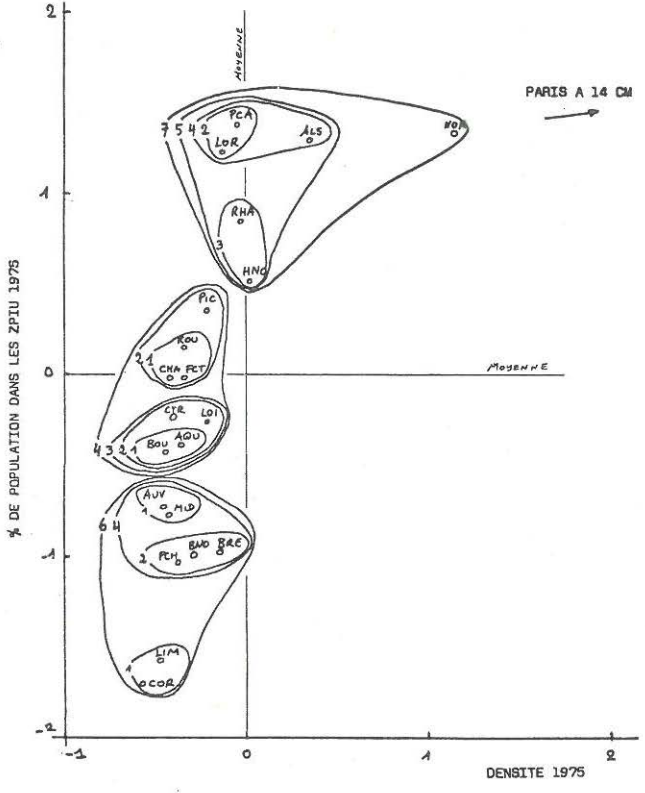
FIGURES 10, 11 : ANALYSE HIERARCHIQUE DE DEUX VARIABLES MISES EN CORRELATION

Figure 10 :
Sur chaque figure on retrouve les deux variables, densité et pourcentage de population en ZPIU (tableau des données figure 6). Les deux variables sont centrées-réduites. Les courbes de niveau numérotées indiquent les étapes successives du regroupement des classes. Seules les premières étapes ont été indiquées. La figure 10 donne les résultats du procédé de la distance moyenne.

Figure 11 :
Le graphique bivariable a été exploité avec l'algorithme du lien minimum. La méthode peut mettre en valeur des structures non sphériques avec l'inconvénient de l'effet de chaînage.

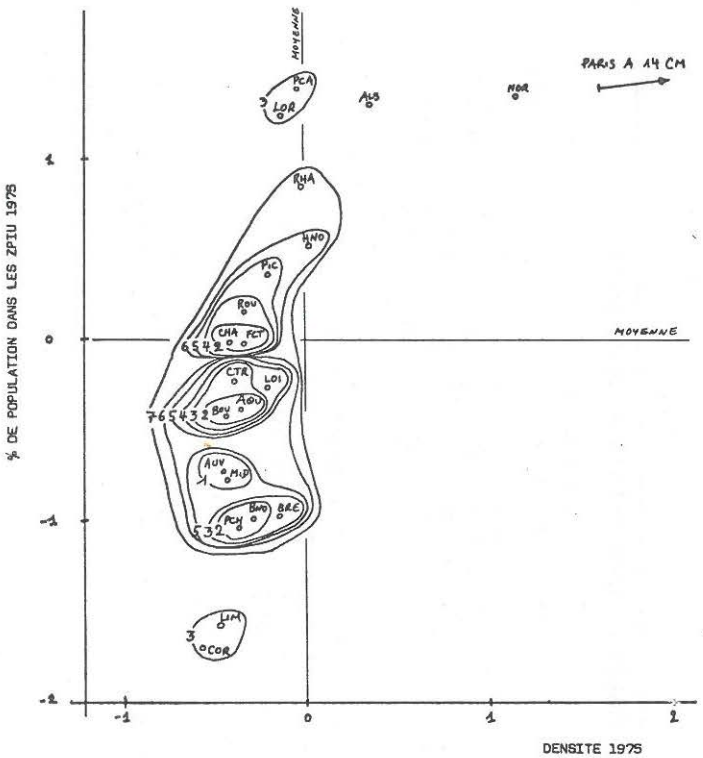
Les élèves traitent ces exemples graphiques à la main. Les rattachements successifs étant déterminés au compas, ils sont fonction de la qualité du dessin de départ. Les résultats sont donc rarement identiques quand ils sont faits par les élèves.

FIGURE 10 : METHODE DE LA DISTANCE MOYENNE



ZPIU:ZONE DE PEUPLEMENT INDUSTRIEL ET URBAIN (INSEE)

FIGURE 11 : METHODE DU SAUT MINIMUM



(ZPIU:ZONES INSEE DE PEUPLEMENT INDUSTRIEL ET URBAIN)

FIGURE 12 : VARIABLES DEMOGRAPHIQUES

I - **Les classes d'âge**

1	H00	Hommes 0-25 ans 1968 (% de la population totale)
2	F00	Femmes 0-25 ans 1968 (% de la population totale)
3	H25	Hommes 25-44 ans 1968 (% de la population totale)
4	F25	Femmes 25-44 ans 1968 (% de la population totale)
5	H65	Hommes 65 et plus 1968 (% de la population totale)
6	F65	Femmes 65 et plus 1968 (% de la population totale)
7	IVI	Indice de vieillesse $= \frac{100\ V\ [65\ et\ +]}{J\ [0\text{-}25]}$

II - **Fécondité - mortalité**

8	TFG	Taux de fécondité générale (67-69)
9	DMH	Durée moyenne de vie Hommes (67-69)
10	DMF	Durée moyenne de vie Femmes (67-69)

III - **Comportement global**

11	TRE	Taux net de reproduction
12	TAN	Taux d'accroissement naturel

FIGURE 13 : TYPOLOGIE DES REGIONS SUR LA BASE DE 12 VARIABLES DEMOGRAPHIQUES CODEES FIGURE 12

MATRICE DES DONNEES BRUTES

	H00	F00	H25	F25	H65	F65	IVI	TFG	DMH	DMF	TRE	TAN
1 PAR	18.9	18.3	15.1	14.2	4.4	7.6	23.3	64.6	68.6	75.8	10.4	85.0
2 CHA	22.3	21.2	12.8	11.8	5.0	7.6	22.4	82.0	67.2	74.6	13.9	85.0
3 PIC	22.4	21.5	12.4	11.6	5.3	7.8	23.7	83.5	66.7	74.0	14,2	78.0
4 HNO	22.3	21.4	12.7	12.3	4.4	7.2	19.7	80.3	66.7	74.5	13.5	92.0
5 CTR	20.4	19.4	12.2	11.6	6.2	9.2	30.4	75.3	68.6	75.5	12.9	56.0
6 BNO	22.3	21.6	12.0	12.0	4.2	7.7	18.8	77.5	65.4	74.7	13.8	77.0
7 BOU	20.2	19.2	12.0	11.3	6.4	9.6	31.7	73.7	68.1	75.5	13.0	43.0
8 NOR	22.6	21.8	12.4	12.0	4.5	7.1	19.9	81.4	65.0	72.8	14.2	85.0
9 LOR	22.6	21.6	13.6	12.5	4.2	6.5	18.6	78.4	66.1	73.8	13.4	84.0
10 ALS	20.9	19.9	13.6	12.8	4.8	7.6	23.0	74.3	65.2	73.2	12.8	63.0
11 FCT	21.8	20.9	12.9	12.1	4.9	7.6	22.5	80.1	67.5	74.8	13.5	86.0
12 LOI	22.0	21.3	11.8	12.0	4.7	8.1	21.4	79.4	67.5	75.0	14.0	86.0
13 BRE	20.7	20.0	12.0	11.8	4.8	8.1	23.2	75.0	64.8	74.0	13.8	58.0
14 PCH	20.5	19.6	11.6	11.4	6.1	9.1	29.8	72.7	69.4	75.9	12.9	49.0
15 AQU	19.1	18.4	11.9	12.0	6.1	9.7	31.9	64.4	68.3	75.7	11.5	30.0
16 MID	18.8	17.9	12.0	12.0	6.4	9.6	34.0	61.4	69.3	75.5	11.0	20.0
17 LIM	17.2	16.3	11.6	11.2	7.5	11.4	43.6	57.9	69.2	76.1	10.9	12.0
18 RHA	20.4	19.6	13.7	13.0	4.8	7.8	23.5	73.0	67.4	75.1	12.0	72.0
19 AUV	19.2	18.3	12.4	11.6	6.3	9.7	32.8	66.4	67.4	74.8	11.7	24.0
20 ROU	18.6	17.8	11.6	11.8	6.6	9.9	35.5	62.7	69.0	75.8	11.3	18.0
21 PCA	18.4	17.6	13.0	12.7	5.9	9.1	32.1	63.5	68.5	75.8	10.8	33.0
22 COR	17.2	16.3	15.0	12.1	6.3	9.8	36.6	62.7	69.0	76.9	11.2	17.0
23 FRA	20.3	19.5	13.0	12.5	5.2	8.3	25.6	71.5	67.5	75.0	12.2	64.0

MATRICE CENTREE REDUITE

	H00	F00	H25	F25	H65	F65	IVI	TFG	DMH	DMF	TRE	TAN
1 PAR	-0.9	-0.7	2.5	3.2	-1.1	-0.8	-0.6	-1.0	0.8	0.8	-1.7	1.0
2 CHA	1.1	1.0	0.1	-0.5	-0.5	-0.8	-0.7	1.3	-0.2	-0.4	1.1	1.0
3 PIC	1.2	1.2	-0.3	-0.8	-0.1	-0.6	-0.5	1.5	-0.6	-1.0	1.3	0.7
4 HNO	1.1	1.1	0.0	0.3	-1.1	-1.1	-1.1	1.0	-0.6	-0.5	0.8	1.2
5 CTR	0.0	-0.1	-0.5	-0.8	0.8	0.6	0.5	0.4	0.8	0.5	0.3	-0.0
6 BNO	1.1	1.2	-0.7	-0.2	-1.3	-0.7	-1.2	0.7	-0.8	-0.3	1.0	0.7
7 BOU	-0.1	-0.2	-0.7	-1.2	1.0	0.9	0.7	0.2	0.4	0.5	0.4	-0.5
8 NOR	1.3	1.3	-0.3	-0.2	-1.0	-1.2	-1.1	1.2	-1.8	-2.3	1.3	1.0
9 LOR	1.3	1.2	0.9	0.6	-1.3	-1.7	-1.3	0.8	-1.0	-1.2	0.7	0.9
10 ALS	0.3	0.2	0.9	1.1	-0.7	-0.8	-0.6	0.3	-1.7	-1.8	0.2	0.2
11 FCT	0.8	0.8	0.2	0.0	-0.6	-0.8	-0.7	1.0	-0.0	-0.2	0.8	1.0
12 LOI	0.9	1.0	-0.9	-0.2	-0.8	-0.4	-0.9	0.9	-0.0	0.0	1.2	1.0
13 BRE	0.2	0.3	-0.7	-0.5	-0.7	-0.4	-0.6	0.4	-2.0	-1.0	1.0	0.1
14 PCH	0.1	0.0	-1.1	-1.1	0.7	0.5	0.4	0.1	1.3	0.9	0.3	-0.2
15 AQU	-0.8	-0.7	-0.8	-0.2	0.7	1.0	0.7	-1.0	0.6	0.7	-0.9	-0.9
16 MID	-0.9	-1.0	-0.7	-0.2	1.0	0.9	1.0	-1.4	1.3	0.5	-1.3	-1.2
17 LIM	-1.9	-1.9	-1.1	-1.4	2.2	2.4	2.4	-1.9	1.2	1.1	-1.3	-2.3
18 RHA	0.0	0.0	1.0	1.4	-0.7	-0.6	-0.5	0.1	-0.1	0.1	-0.5	0.5
19 AUV	-0.7	-0.7	-0.3	-0.8	0.9	1.0	0.8	-0.8	-0.1	-0.2	-0.7	-1.1
20 ROU	-1.0	-1.0	-1.1	-0.5	1.3	1.1	1.2	-1.2	1.1	0.8	-1.0	-1.3
21 PCA	-1.2	-1.1	0.3	0.9	0.5	0.5	0.7	-1.1	0.7	0.8	-1.4	-0.8
22 COR	-1.9	-1.9	2.4	0.0	0.9	1.1	1.4	-1.2	1.1	2.0	-1.1	-1.3
23 FRA	-0.1	-0.0	0.3	0.6	-0.3	-0.2	-0.2	-0.1	-0.0	0.0	-0.3	0.3

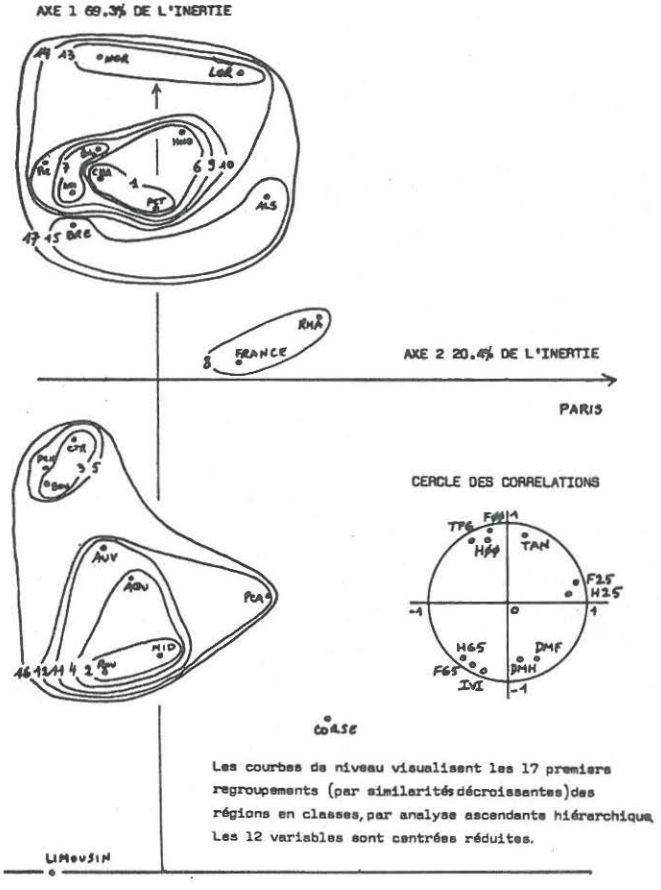
FIGURES 14, 15 : ANALYSE EN COMPOSANTES PRINCIPALES ET HIERARCHIQUE DU TABLEAU DEMOGRAPHIQUE

Figure 14 :
Le tableau 22 régions, 12 variables démographiques (données : figure 13) a été traité par l'analyse en composantes principales. Cette figure montre le premier plan factoriel, 89,7 % de la variance totale. Les courbes de niveau visualisent les proximités des points dans l'espace multidimensionnel déterminées par l'analyse ascendante hiérarchique. Les déformations locales (Nord-Lorraine, Alsace-Bretagne) mettent en valeur les "erreurs de perspective" liées à la projection.

Figure 15 :
Le premier plan factoriel est repris ici, les similarités régionales sont représentées par intensité décroissante. Nette opposition entre France du Nord et France du Sud, la région parisienne formant un type à part étant donnée l'importance des actifs.

ANALYSE EN COMPOSANTES PRINCIPALES ET HIERARCHIQUE DU TABLEAU 22 REGIONS, 12 VARIABLES

FIGURE 14



Les courbes de niveau visualisent les 17 premiers regroupements (par similarités décroissantes) des régions en classes, par analyse ascendante hiérarchique. Les 12 variables sont centrées réduites.

FIGURE 15

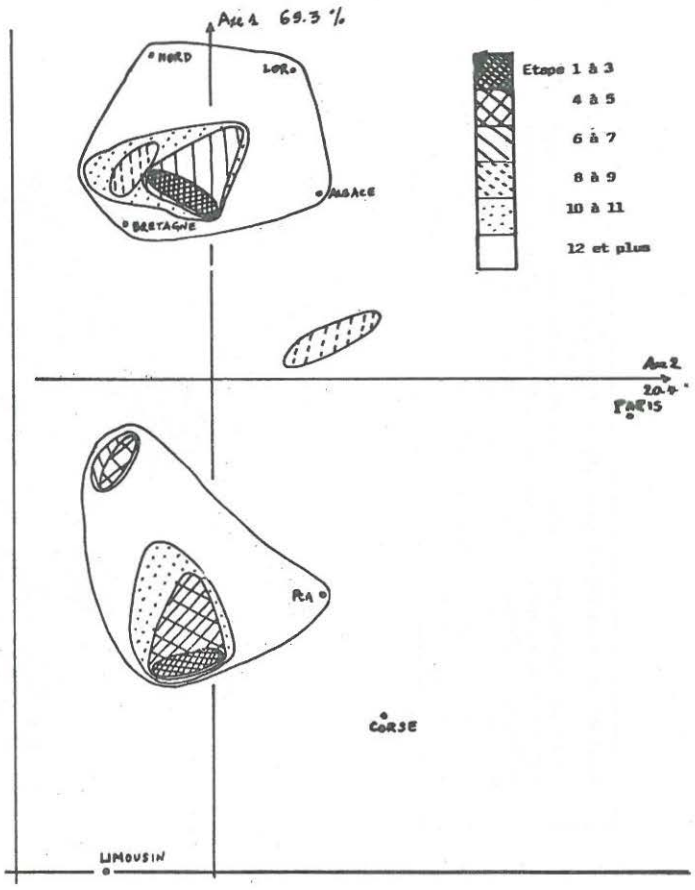


FIGURE 16 : CARTE DE SYNTHESE DEMOGRAPHIQUE

Le but de l'analyse des données est de traiter les tableaux les plus complets possibles. Le tableau traité ici par l'analyse ascendante hiérarchique (distance moyenne) est le tableau 22 régions, 12 variables de la figure 13. La carte de synthèse oppose les régions de la France du Nord à fort dynamisme démographique aux régions freinées sur le plan démographique. La légende des classes graphiques commune à toutes les cartes se trouve figure 7.

FIGURE 16 : CARTE DE SYNTHESE

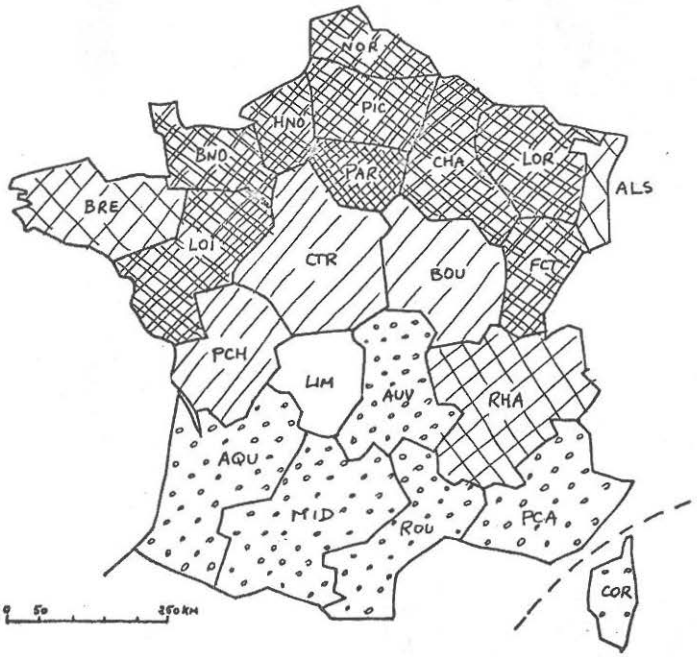
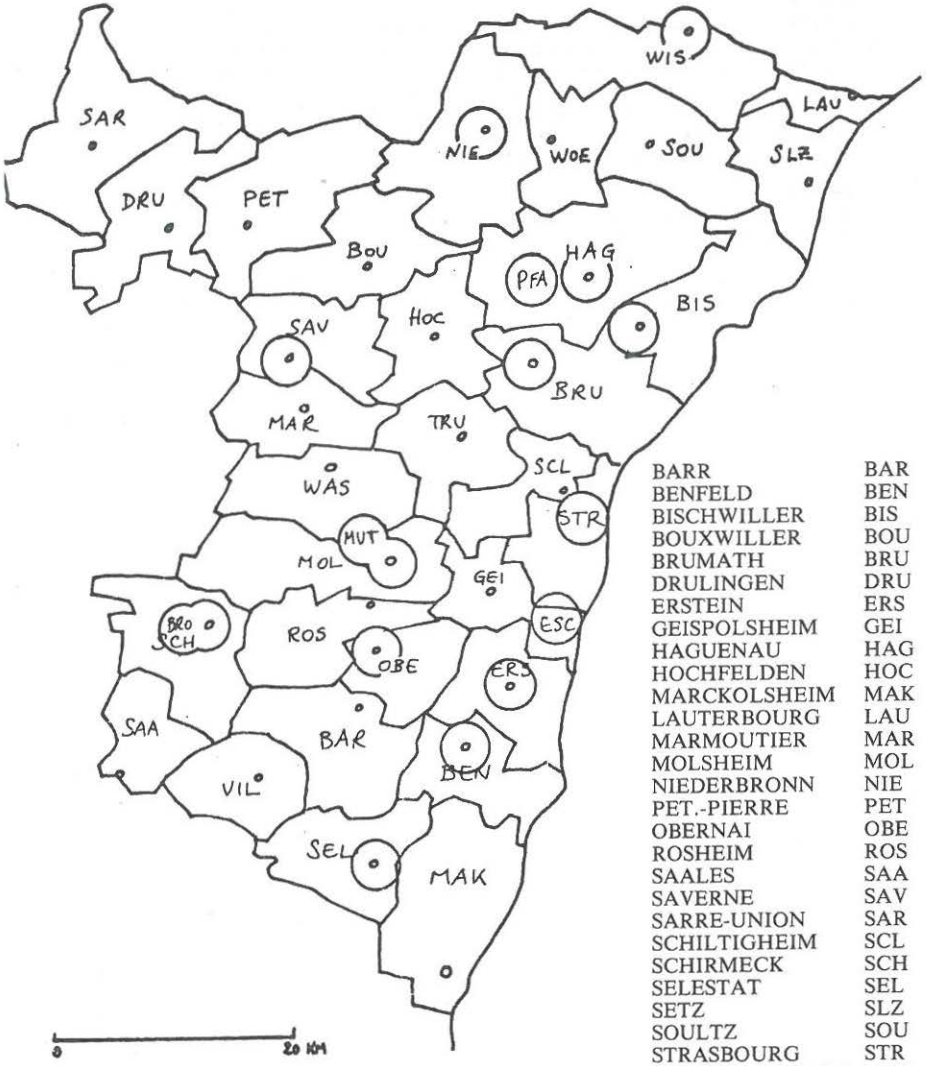


FIGURE 17 : 31 CANTONS ALSACIENS (BAS-RHIN) ET 17 UNITES URBAINES



BARR	BAR
BENFELD	BEN
BISCHWILLER	BIS
BOUXWILLER	BOU
BRUMATH	BRU
DRULINGEN	DRU
ERSTEIN	ERS
GEISPOLSHEIM	GEI
HAGUENAU	HAG
HOCHFELDEN	HOC
MARCKOLSHEIM	MAK
LAUTERBOURG	LAU
MARMOUTIER	MAR
MOLSHEIM	MOL
NIEDERBRONN	NIE
PET.-PIERRE	PET
OBERNAI	OBE
ROSHEIM	ROS
SAALES	SAA
SAVERNE	SAV
SARRE-UNION	SAR
SCHILTIGHEIM	SCL
SCHIRMECK	SCH
SELESTAT	SEL
SETZ	SLZ
SOULTZ	SOU
STRASBOURG	STR
TRUCHTERSHEIM	TRU
VILLE	VIL
WASSELONNE	WAS
WISSEMBOURG	WIS
WOERTH	WOE
PFAFFENHOFFEN	PFA
LA BROQUE	BRO
ESCHAU	ESC

FIGURES 18, 19 : ANALYSE EN COMPOSANTES PRINCIPALES ET HIERARCHIQUE DU TABLEAU 31 CANTONS × 7 CLASSES D'AGE

Figure 18 :
Le premier plan factoriel est représenté sur cette figure (58 % de la variance). Points-variables et points-cantons sont représentés sur le même graphe. Ce dernier a en outre été exploité par l'analyse ascendante hiérarchique (distance moyenne) : les niveaux successifs du regroupement visualisent la qualité de la projection. L'exemple des cantons est plus complexe que celui des 22 régions. L'opposition des cantons âgés aux autres est nette mais le noyau central est difficile à débrouiller. On peut se contenter dans ce cas d'étudier les structures typiques.

Figure 19 :
Le plan factoriel 1/2 est utilisé ici pour montrer que l'ACP effectue une classification des profils. On a dessiné les pyramides d'âge (renversées et centrées-réduites, l'ordre des variables est indiqué en haut à gauche. C'est celui du premier axe factoriel).

ANALYSE EN COMPOSANTES PRINCIPALES ET HIERARCHIQUE DU TABLEAU 31 CANTONS, 7 CLASSES D'AGE

FIGURE 18

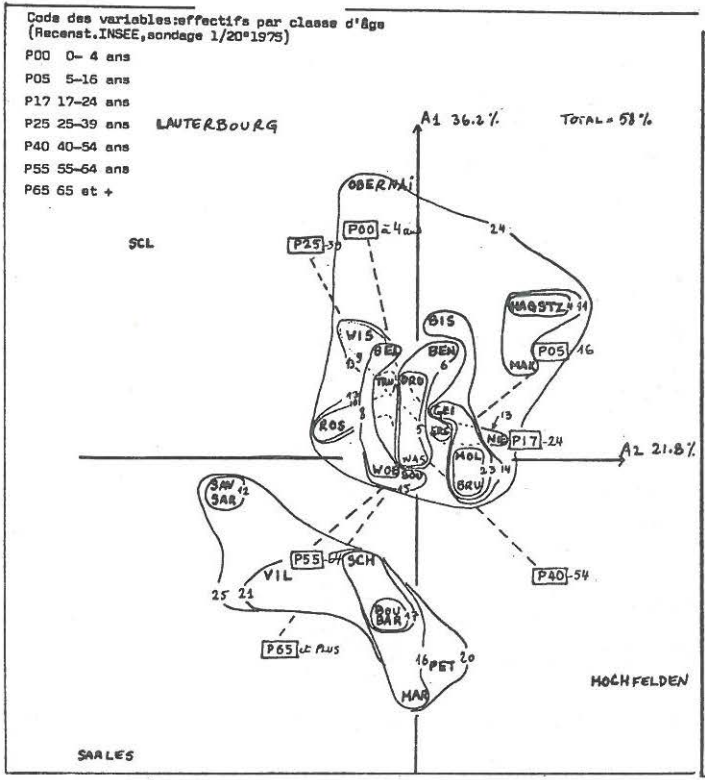
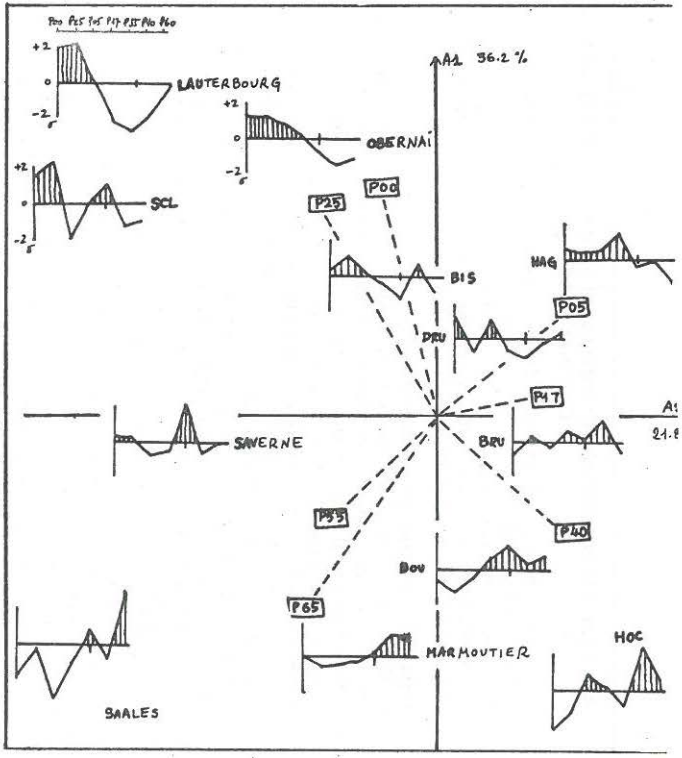


FIGURE 19



ANALYSE EN COMPOSANTES PRINCIPALES D'UN TABLEAU 48 UNITES, 7 CLASSES D'AGES

Le tableau, tiré de "Chiffres pour l'Alsace", juxtapose 31 cantons du Bas-Rhin et 17 unités urbaines de plus de 5 000 habitants décrits par 7 classes d'âge d'après le recensement de la population de 1975, sondage au 1/5e (0-4 ans, 5-16, 17-24, 25-39, 40-54, 55-64, 65 ans et plus). Le tableau a été traité par la méthode d'analyse en composantes principales (ACP dans le texte).

Figure 20 : Le graphe et son interprétation

L'ACP tient compte de toutes les interrelations simultanées des variables, et calcule les meilleures combinaisons linéaires successives de ces variables. Ces combinaisons remplacent les anciennes variables et sont en nombre plus faible qu'elles parce qu'elles recueillent une variabilité plus forte du nuage de points qu'une variable unique. Ces combinaisons s'appellent *"axes"* ou *"composantes principales"*. Les axes 1 et 2 représentés ici sauvegardent 59 % de la variabilité totale du nuage alors que deux variables originelles ne pourraient en représenter au mieux que 28 %. L'ACP calcule ensuite les coordonnées des points cantons et des points variables dans ce nouvel espace de représentation. Le caractère multidimensionnel de la matrice est ainsi surmonté par une généralisation de la méthode du graphe à deux variables.

Le graphe factoriel visualise dans le plan des composantes principales choisies (ici les deux premières) les positions relatives des points cantons et des points variables en respectant le mieux possible (à la perte d'information près liée à une erreur de perspective) leurs distances réciproques dans l'espace initial à 7 dimensions.

La position sur le graphe des points variables permet de savoir pourquoi les points cantons sont à tel ou tel emplacement : l'axe 1 (39 % de la variation totale) caractérise l'opposition entre jeunes et vieux : 0-4 ans et 25-39 ans en haut, 55-64 ans, 65 et plus, en bas. Ainsi Pfaffenhoffen est un canton plutôt *"jeune"* (par rapport à tous les autres) et Marmoutier un canton plutôt *"vieux"*. Le deuxième axe caractérise une opposition de plus faible importance (20 % de la variabilité) entre les 5-16 ans et les 55-64 ans. La méthode permet de respecter le caractère continu de la variation des profils démographiques.

Figure 21 : La carte

Si le géographe veut étudier de manière traditionnelle la distribution spatiale du phénomène démographique en Alsace, il lui faut réaliser d'abord autant de cartes que de variables (7 ici), puis autant de cartes que de combinaisons de variables prises deux à deux soit dans notre cas 28 cartes qu'il faut ensuite interpréter et synthétiser.

L'ACP réalise automatiquement la synthèse de ces 28 cartes possibles. La carte montre, découpée en 7 classes graphiques d'égale amplitude la position des cantons sur l'axe de *"jeunesse"* du graphe. Plus la trame est foncée, plus la région est *"jeune"* au sens de la combinaison des 7 variables de la matrice qui rassemble 39 % de la variabilité totale.

FIGURE 21

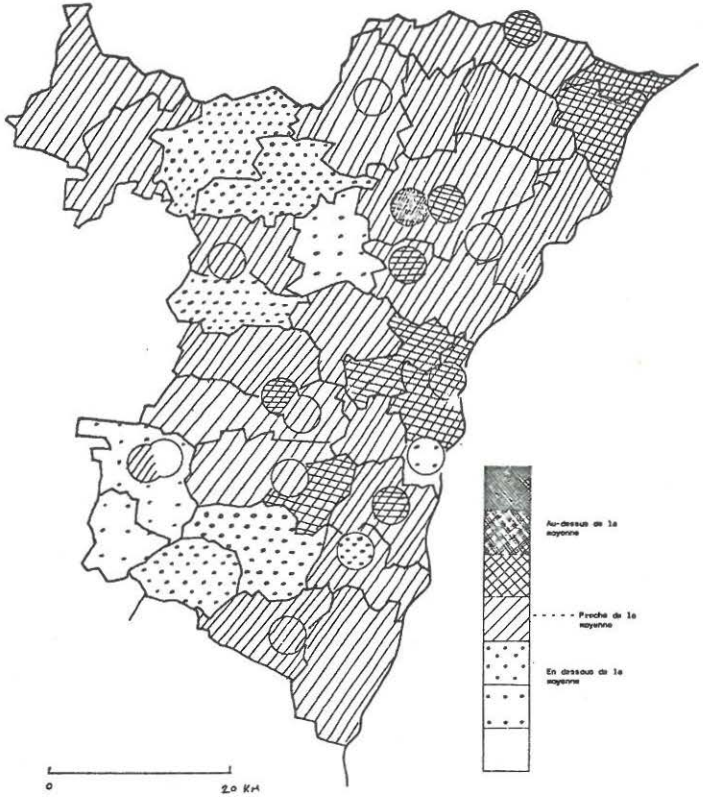


FIGURE 20

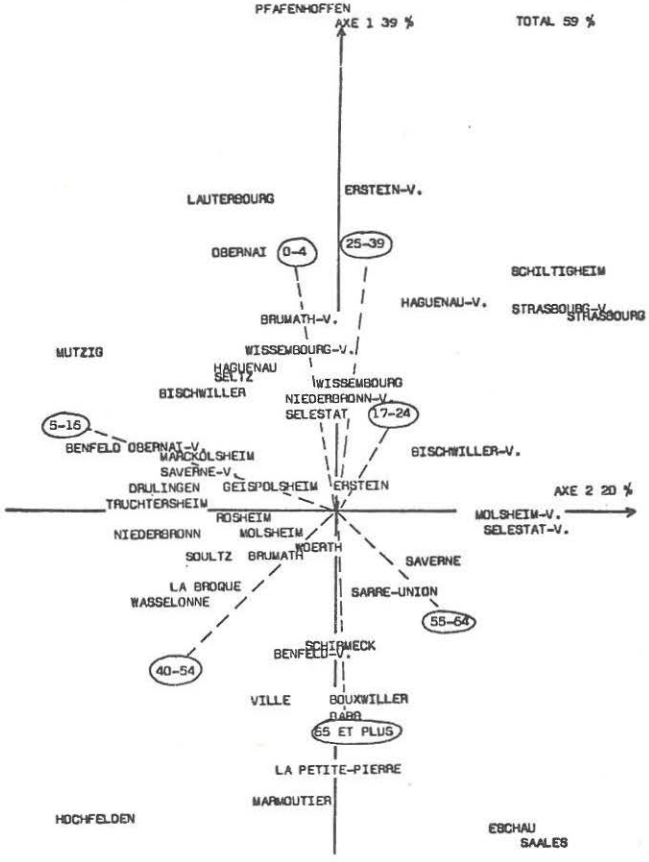
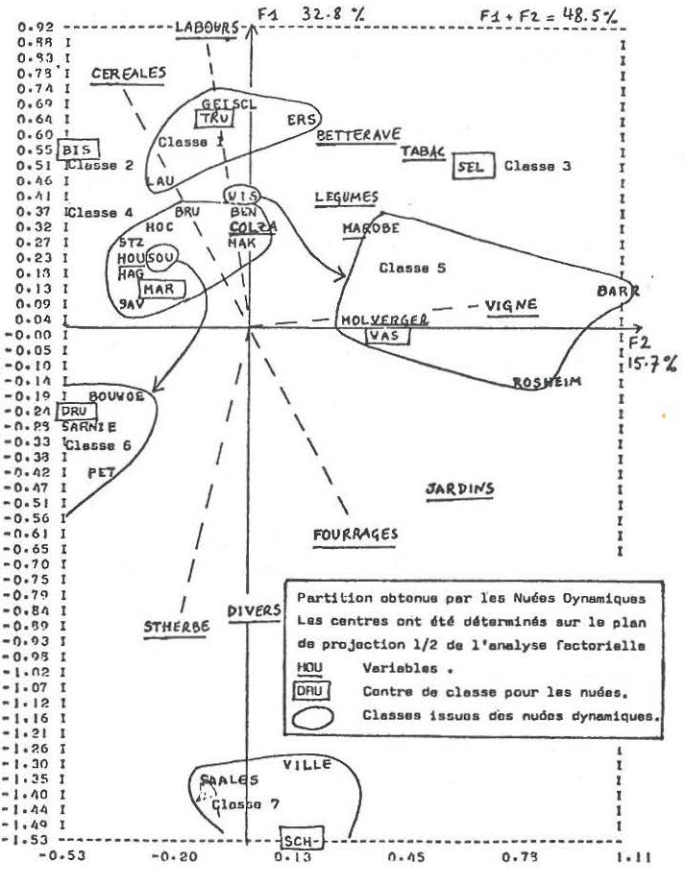


FIGURE 22 : ANALYSE EN COMPOSANTES PRINCIPALES ET NUEES DYNAMIQUES

Les 31 cantons alsaciens sont décrits par 14 variables agricoles (utilisation de la surface agricole utile en %). Points-variables et points-cantons sont dessinés sur le plan. Le premier axe oppose système à labours prédominants aux systèmes à herbe, l'axe 2 les cantons viticoles à tous les autres. C'est à partir de ce premier plan factoriel que l'on a choisi des centres plausibles pour initialiser la procédure des nuées dynamiques. On voit que certains cantons sont affectés par la méthode à des classes assez éloignées de leur position sur le graphe. Ces points sont par ailleurs mal représentés sur ce premier plan.



CODE DES VARIABLES (en % de surface agricole utile)

LAB	Labours	COL	Colza
MAR	Maraîchages	TAB	Tabac
VIG	Vignes	BET	Betteraves industrielles
VER	Vergers	FOU	Cultures fourragères
STH	Surfaces toujours en herbe	LEG	Légumes frais
JAR	Jardins familiaux et divers	HOU	Houblon
CER	Céréales	DIV	Divers

FIGURES 23, 24 : ANALYSE HIERARCHIQUE DE DEUX VARIABLES MISES EN CORRELATION

Figure 23 :
Les élèves traitent un petit tableau décrivant 16 stations climatiques. Ils tracent le graphe bivariable des points stations dans l'espace de deux précipitations. La méthode du saut minimum permet de regrouper les stations à des niveaux donnés de ressemblance.

Figure 24 :
L'arbre décrit la même corrélation graphique d'une manière différente. L'intérêt de la représentation arborescente est de n'être pas limitée comme le graphique de la figure 23 à deux ou trois dimensions au maximum. L'arbre permet la représentation des distances entre objets même pour des espaces de dimension supérieure à trois.

FIGURE 24 : L'ARBRE

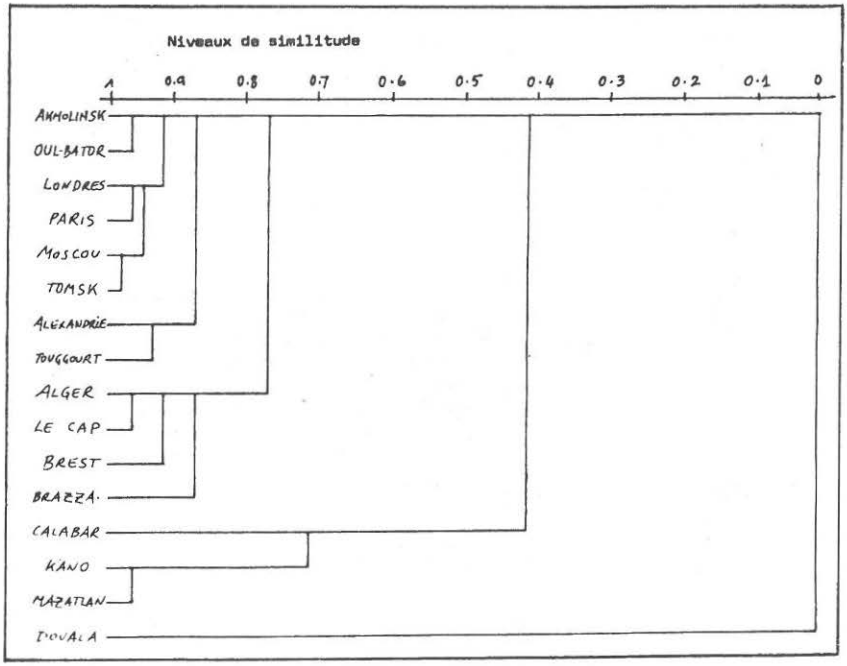


FIGURE 23 : METHODE DU SAUT MINIMUM SUR LE GRAPHE BIVARIE

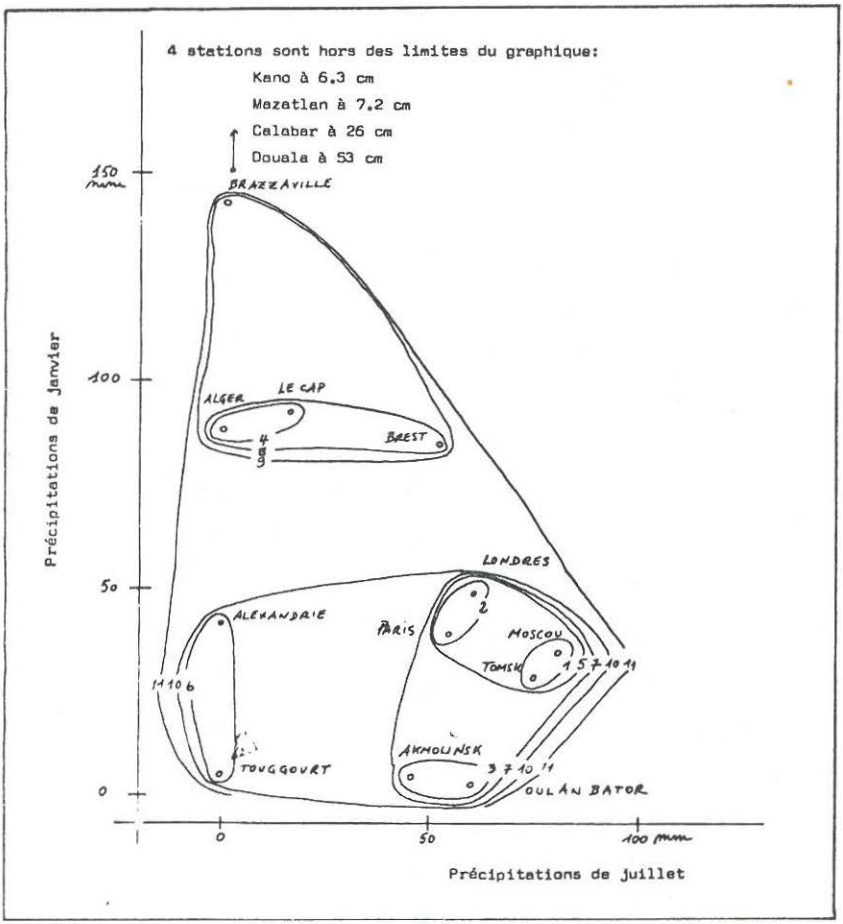
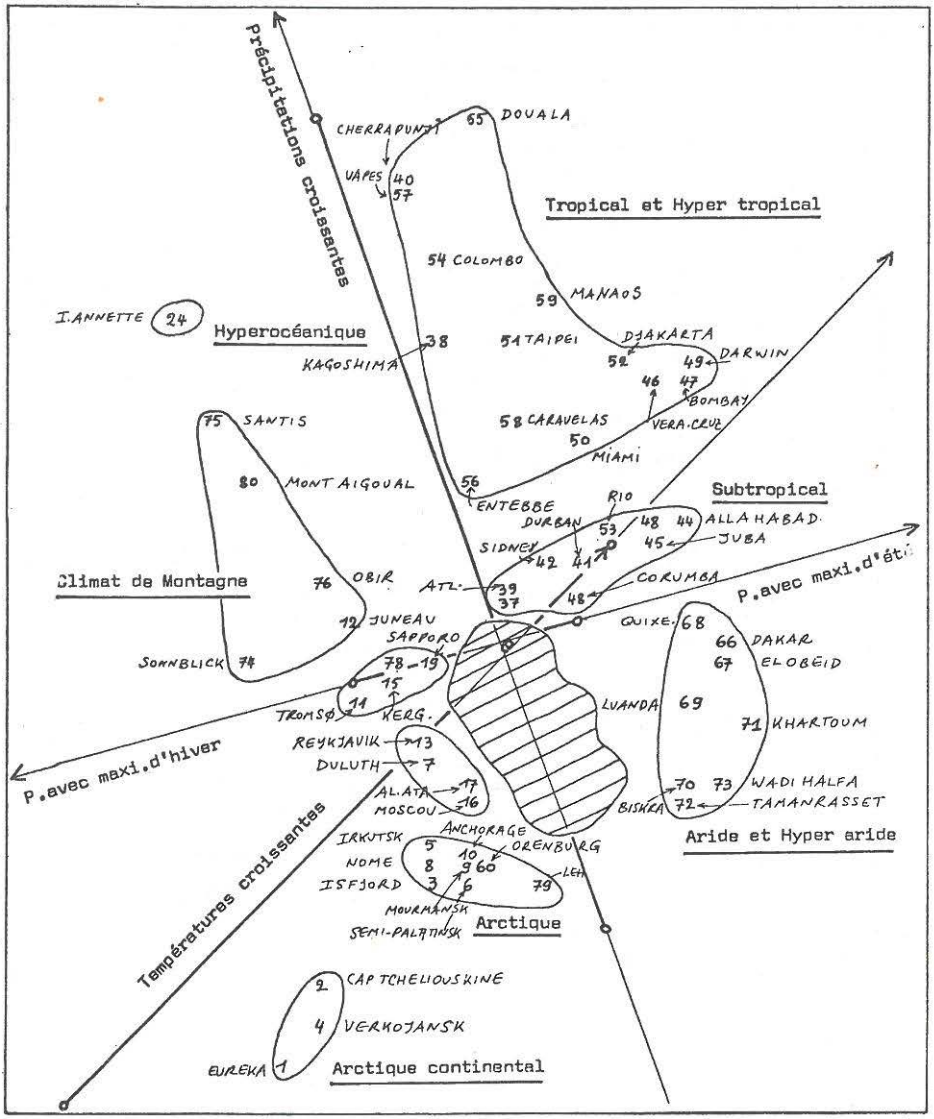


FIGURE 25 : 80 STATIONS CLIMATIQUES DECRITES PAR 12 TEMPERATURES ET 12 PRECIPITATIONS MENSUELLES

Analyse multidimensionnelle : premier plan de projection de l'analyse factorielle des correspondances, axes un et deux, 75,6 % de la variance totale. On a tenté une typologie "à la main". Le noyau central regroupe 24 stations de type "tempéré", où la structure du nuage, trop peu différenciée, n'a pas été débrouillée.



J.P. LETOURNEUX
IREM de Nantes

4. CLASSIFICATION HIERARCHIQUE ASCENDANTE

Faute de moyen de calcul (avant l'avènement des ordinateurs), la statistique descriptive s'est longtemps réfugiée dans une vision unidimensionnelle du monde : un individu est repéré par un seul caractère (par exemple son salaire). En tentant de trouver des « explications », une vision bidimensionnelle s'impose peu à peu : on cherche par exemple à « expliquer » le salaire par le nombre d'années d'études. Peu à peu l'insuffisance de l'« explication » unicausale apparaît, on cherche des modèles tri puis multidimensionnels : par exemple on expliquera la formation du salaire par l'âge de l'individu et le nombre d'années d'études.

La statistique multidimensionnelle devient une illustration de l'algèbre linéaire. Un individu est un point dans un espace vectoriel de dimension finie. Un tableau de données devient un nuage de points. Mieux une variable devient un point dans un espace vectoriel ayant pour dimension le nombre des individus. On est alors amené à chercher un sous-espace de faible dimension dans lequel on peut « résumer au mieux » le nuage des individus (le nuage des variables). Les diverses variantes de l'analyse factorielle (Analyse factorielle, Analyse en composantes principales, Analyse des correspondances...) fournissent la théorie (c'est-à-dire permettent de donner un sens mathématique à « résumer au mieux »), l'ordinateur étant le moyen le plus souvent indispensable.

Envisageant toujours un tableau de données comme un nuage de points, on peut privilégier les relations de « proximité » existant entre les points eux-mêmes. La question est alors : Comment « partager » le nuage de points individus (resp. variables) en sous-ensembles « stables ». La Classification Hiérarchique Ascendante (C H A) est une des théories permettant de donner un sens mathématique à la notion de « proximité », à celle de « partage » en classes stables. Les diverses méthodes de l'analyse discriminante se rattachent également à cette problématique. Là encore le recours à l'ordinateur est souvent indispensable.

Un choix cornélien

Le statisticien est placé devant le dilemme suivant :

— Réduire les données, les ordonner, les classer en vue de mettre en évidence la structure du phénomène.

— Perdre le moins d'information possible et, donc classifier et réduire le moins possible car toute transformation induit une perte d'information.

La méditation du tableau des précipitations, jour par jour sur l'année, recueillies dans les différentes stations météo ne risque pas de faire jaillir comme une évidence une typologie (classification) des climats (surtout si les stations sont données dans l'ordre alphabétique !). Mais sa réduction à la précipitation annuelle risque de faire conclure à l'observateur pressé que Brest et Nice ont à peu près le même climat !

Comme l'écrit J.P. BENZECRI : « La tâche du statisticien est de « représenter les données avec un minimum de perte d'information... et un maximum d'explication ».

Jusqu'à une époque récente les statisticiens résolvaient ce dilemme par l'ajustement des données empiriques à un modèle théorique connu (loi normale par exemple). Bien souvent l'ajustement était plus ou moins justifié. Cela a conduit à un usage abusif de la loi normale dont, malheureusement, la plupart des livres de statistique souffrent encore aujourd'hui.

La vision moderne des statistiques descriptives consiste à traiter les données elles-mêmes en se donnant un critère de minimisation de la déformation et une mesure de cette déformation.

La classification hiérarchique ascendante

Pour le statisticien d'aujourd'hui, un individu est essentiellement un n-uple. L'ensemble E des individus est un nuage de points de $\mathbf{R}^n$ (ou de $(\mathbf{Z}/2\mathbf{Z})^n$). Bien que plongeable dans un $\mathbf{R}^n$, la structure induite par celle de $\mathbf{R}^n$ n'a pas nécessairement une signification pour le problème posé.

On voit donc apparaître un premier choix :

• **Choix des données elles-mêmes et de leur codification.**

Ce premier choix n'est pas spécifique à la C.H.A. Il concerne la structuration des données sous forme d'un tableau. Il se décompose en fait en deux :

— choix des variables : Etudiant la réussite à une batterie de tests d'un groupe d'élèves, on pourra retenir ou non des variables comme l'âge des élèves, leurs caractéristiques sociologiques, etc.

Ce choix est essentiel ; c'est lui qui détermine le reste. Inutile de mettre en œuvre un outillage sophistiqué sur des données de médiocre intérêt.

— choix d'une codification : dans le cas d'un test, attribuera-t-on une note de 0 à 100 ou adoptera-t-on un codage binaire : réussite-échec ? Conditionnant la suite, ce premier choix est très important, il concerne en fait beaucoup plus chaque discipline que les mathématiques elles-mêmes. C'est au niveau de la codification que l'analyse booléenne peut être une aide à l'utilisateur dans le codage des questionnaires notamment.

Ce premier choix étant fait, l'ensemble des individus est un nuage de points dans $\mathbf{R}^n$. Il s'agit, pour le statisticien et l'utilisateur, de définir une notion de proximité entre deux individus signifiante par rapport au problème posé. C'est le second choix :

- **Choix d'un indice de distance (ou de similarité).**

La notion **d'indice de distance** est un affaiblissement de la notion de distance.
De façon précise :

Soit E un ensemble, on appelle **indice de distance** sur E une application d de E × E dans $\mathbf{R}^+$ telle que :

1) $\forall x \in E \quad d(x,x) = 0$
2) $\forall (x,y) \in E \quad d(x,y) = d(y,x)$
3) $\forall (x,y) \in E \quad x \neq y \quad d(x,y) > 0$

Cette notion extrêmement faible permet notamment dans le cas de codages binaires, de choisir un indice adapté au problème posé. C'est là un avantage des méthodes de classification sur les méthodes linéaires (Analyse en composantes principales, Analyse des correspondances) qui utilisent une métrique euclidienne. Dans bon nombre de questions, en sociologie notamment, l'on s'intéresse moins à une mesure qu'à la possibilité de donner un sens à l'expression :

« *Les individus x_1 et x_2 ont plus de ressemblances entre eux qu'en ont les individus x_3, x_4* ». (on notera cette relation $\{x_1, x_2\}$ R $\{x_3, x_4\}$).

Si E est muni d'un indice de distance d, la relation :

$$\{x_1, x_2\} \text{ R } \{x_3, x_4\} \Longleftrightarrow d(x_1, x_2) \leqslant d(x_3, x_4)$$

est une relation de **préordre total** sur l'ensemble des paires d'éléments de E.

Le but de la C.H.A. est de passer de la partition constituée par tous les singletons de $\mathcal{P}$ (E) à la partition réduite à E en rapprochant à chaque étape les individus ou les sous-ensembles d'individus les moins distants (c'est-à-dire les plus proches ou encore les plus semblables).

Nous arrivons ainsi au 3ème choix :

- **Choix d'une stratégie d'agrégation.**

Il s'agit, en fait, de prolonger l'application d en une application ∂ définie sur ($\mathcal{P}$ (E) \ ϕ) × ($\mathcal{P}$ (E) \ ϕ) ou au minimum entre les élé-

ments d'une partition de E. Nous reviendrons plus précisément sur ce choix, mais amorçons pour le lecteur la discussion dans le cas de $\mathbf{R}^2$ muni de la distance euclidienne d (par exemple).

On peut pour prolonger d définir ∂ par exemple par :

$$\partial_1(\{A, B\}, C) = \inf (d(A, C), d(B, C))$$

$$\text{ou } \partial_2(\{A, B\}, C) = \sup (d(A, C), d(B, C))$$

$$\text{ou } \partial_3(\{A, B\}, C) = \frac{d(A, C) + d(B, C)}{2}$$

$$\text{ou } \partial_4(\{A,B\}, C) = d(C, I) \text{ (I milieu de AB)}$$

Le choix d'un prolongement (c'est-à-dire d'une stratégie d'agrégation) dépend du problème étudié. Dans bon nombre de problèmes sociologiques ∂_3 et ∂_4 n'ont aucun sens, même si parfois on les utilise !

Comme nous allons le préciser dans la suite, le choix d'un indice de distance et d'une stratégie d'agrégation permettent de définir une hiérarchie de parties sur E.

$\mathcal{H}$ est une hiérarchie de parties sur E si, et seulement si :

1) $E \in \mathcal{H}$
2) $\forall x \in E \qquad \{x\} \in \mathcal{H}$
3) $(\forall H \in \mathcal{H}) \quad (\forall K \in \mathcal{H}) \quad H \cap K \in \{H, K, \emptyset\}$

Sur une hiérarchie de partie obtenue par une C.H.A., on peut définir grâce à une stratégie d'agrégation, une application croissante (pour la relation d'inclusion dans $\mathcal{H}$).

$$f : \mathcal{H} \longrightarrow \mathbf{R}^+ \text{par :}$$

$$f(\{x_i\}) = 0 \quad \text{pour } x_i \in E$$

$$f(A \cup B) = \partial(A, B) \quad \text{pour } A \in \mathcal{H} \quad \text{et } B \in \mathcal{H}$$

On obtient ainsi une **hiérarchie indiciée** (∂ étant le prolongement à $\mathcal{H}$ de l'indice de distance d).

($\mathcal{H}$, f) permet de définir sur $E \times E$ une application d* par : $d^*(x_i, x_j) = f(A_{ij})$ où A_{ij} est le plus petit élément de $\mathcal{H}$ (au sens de l'inclusion) contenant $\{x_i, x_j\}$.

Par construction d* prend ses valeurs dans $\mathbf{R}^+$ et

1) $d^*(x_i, x_j) = 0 \Rightarrow x_i = x_j$
2) $d^*(x_i, x_j) = d^*(x_j, x_i)$

De plus, pour x_i, x_j, x_k dans E on a :

$$A_{ij} \cap A_{ik} \neq \phi \text{ car } x_i \in A_{ij} \cap A_{ik} ;$$

comme $\mathcal{H}$ est une hiérarchie

$$A_{ij} \cap A_{ik} = A_{ik},$$

ou

$$A_{ij} \cap A_{ik} = A_{ij}.$$

• Si $A_{ij} \cap A_{ik} = A_{ik}$, alors $A_{ik} \subset A_{ij}$; comme f est croissante,

$$d^*(x_i, x_k) = f(A_{ik}) \leqslant d^*(x_i, x_j) = f(A_{ij}),$$

d'où a fortiori

$$d^*(x_i, x_k) \leqslant \sup (d^*(x_i, x_j), d^*(x_j, x_k))$$

• Si $A_{ij} \cap A_{ik} = A_{ij}$, alors $A_{ij} \subset A_{ik}$, et $A_{jk} \subset A_{ik}$ (par définition de A_{jk}) mais on a également

$A_{ij} \cap A_{jk} \neq \phi$ d'où $A_{ij} \cap A_{jk} = A_{ij}$, c'est-à-dire $A_{ij} \subset A_{jk}$

ou

$A_{ij} \cap A_{jk} = A_{jk}$, c'est-à-dire $A_{jk} \subset A_{ij}$

si $A_{ij} \subset A_{jk}$ alors $A_{ik} \subset A_{jk}$ d'où $A_{ik} = A_{jk}$

si $A_{jk} \subset A_{ij}$ alors $A_{ik} \subset A_{ij}$ d'où $A_{ik} = A_{ij}$

Il en résulte que dans tous les cas d* vérifie :

$$3)\ d^*(x_i, x_k) \leqslant \sup (d^*(x_i, x_j), d^*(x_j, x_k))$$

Au terme d'une procédure de C.H.A., nous avons donc muni E d'une distance **ultramétrique** (c'est-à-dire vérifiant 1, 2, 3) d*. Elle sert à mesurer la qualité de la classification obtenue, c'est-à-dire sa capacité à rendre intelligible les données initiales sans trop les déformer. Nous sommes confrontés là au difficile problème de la reconstitution des données expérimentales à partir du modèle obtenu, ici, par une procédure de C.H.A. (ailleurs, mais n'anticipons pas sur le tome II,par une procédure d'analyse en composantes principales par exemple).

Il s'agit d'un 4ème choix. Ce choix correspond à la tolérance de déformation que l'on accepte, c'est le :

• **Choix d'un critère de déformation.**

Il s'agit de mesurer la déformation (l'erreur systématique) que l'on fait en remplaçant les données expérimentales (E, d) par les données calculées (E, d*). Si pour le problème étudié c'est l'indice de distance d qui a une signification profonde, c'est un critère de déformation entre d et la distance ultramétrique d* que nous chercherons. Si au contraire, seule la relation R entre les paires déduites de d :

$$\{x_1, x_2\}\ R\ \{x_3, x_4\} \iff d(x_1, x_2) \leqslant d(x_3, x_4)$$

est pertinente pour le problème étudié ; c'est la déformation existant entre R et R* (déduite de d* par

$$\{x_1, x_2\}\ R^*\ \{x_3, x_4\} \iff d^*(x_1, x_2) \leqslant d^*(x_3, x_4))$$

que nous retiendrons. N'ayant pas voulu alourdir cet article par des considérations sur les critères de comparaison entre ordre ou préordre, nous n'aborderons le 4ème choix (critère de déformation) que sous l'aspect déformation entre d et d*. C'est à l'utilisateur en fonction de son problème de décider de la forme et de l'intensité de la déformation acceptable, comme ce sera à lui de décider en analyse en composantes principales ou en analyse des correspondances du nombre de facteurs à extraire.

Un exemple pour commencer

(Cet exemple est également traité par R. GRAS, Tome II à l'aide d'une analyse des correspondances).

La mise en route d'une C.H.A. ne se justifie bien entendue que lorsque nous avons affaire à un nombre d'individus suffisant, d'où le caractère artificiel de l'exemple qui va suivre.

6 élèves (individus) répondant à 4 questions (variables), si la réponse est correcte on note 1, sinon 0. On obtient le tableau suivant (1er choix).

	Q_1	Q_2	Q_3	Q_4
E_1	1	0	1	1
E_2	1	1	0	0
E_3	0	1	0	0
E_4	0	1	1	1
E_5	1	1	1	1
E_6	1	1	1	0

Considérons l'ensemble $Q = \{Q_1, Q_2, Q_3, Q_4\}$ des 4 questions. Un élève est identifié à l'ensemble des questions auxquelles il a répondu correctement.

Ainsi E_1 s'identifie à $E_1 = \{Q_1, Q_3, Q_4\}$

La différence symétrique sur l'ensemble des parties de Q permet de définir une distance sur l'ensemble

$$E = \{E_1, E_2, E_3, E_4, E_5, E_6\} \text{ des élèves.}$$

On pose $d_1(E_i, E_j) = \text{card}(E_i \triangle E_j)$ où $\text{card}(E_i \triangle E_j)$ est le nombre d'éléments de l'ensemble $E_i \triangle E_j$ différence symétrique de E_i et E_j.

Rappelons que si A, B, C sont 3 parties d'un ensemble

$$A \triangle C \subset (A \triangle B) \cup (B \triangle C)$$

il en résulte que :

$$(3)\ d_1(E_i, E_k) \leqslant d_1(E_i, E_j) + d_1(E_j, E_k)$$

De plus on a, d'après les propriétés de la différence symétrique

$$(2)\ d_1(E_i, E_j) = d(E_j, E_i)$$
$$(1)\ d_1(E_i, E_j) = 0 \iff E_i = E_j$$

Afin d'obtenir une distance variant de 0 à 1, on pose (2ème choix) :

$$d(E_i, E_j) = \frac{1}{4}\, d_1(E_i, E_j).$$

Compte tenu de cette distance, on obtient la matrice de distances entre les points de E :

$$\begin{array}{c} \\ E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \\ E_6 \end{array} \begin{array}{c} \begin{array}{cccccc} E_1 & E_2 & E_3 & E_4 & E_5 & E_6 \end{array} \\ \begin{pmatrix} 0 & 3/4 & 1 & 1/2 & 1/4 & 1/2 \\ 3/4 & 0 & 1/4 & 3/4 & 1/2 & 1/4 \\ 1 & 1/4 & 0 & 1/2 & 3/4 & 1/2 \\ 1/2 & 3/4 & 1/2 & 0 & 1/4 & 1/2 \\ 1/4 & 1/2 & 3/4 & 1/4 & 0 & 1/4 \\ 1/2 & 1/4 & 1/2 & 1/2 & 1/4 & 0 \end{pmatrix} \end{array} = D_0$$

La partition de départ de l'ensemble E est :

$$\{ \{E_1\} , \{E_2\} , \{E_3\} , \{E_4\} , \{E_5\} , \{E_6\} \} = \mathcal{T}_0$$

La mise en route de la procédure de C.H.A. se fait de la manière suivante. On réunit, dans un même ensemble, 2 élèves dont la distance est minimale. Le minimum de la distance entre 2 éléments distincts est 1/4. Il est réalisé par le couple $\{E_1, E_5\}$. Notons que d'autres couples réalisent ce minimum, par exemple $\{E_4, E_5\}$; ceci montre que l'algorithme de C.H.A. n'assure pas l'unicité du passage d'une partition à une autre. Au terme de cette première étape nous avons la partition

$$\{ \{E_1, E_5\} , \{E_2\} , \{E_3\} , \{E_4\} , \{E_6\} \} = \mathcal{T}_1$$

Il faut alors choisir un procédé permettant de calculer la « proximité » entre les éléments de cette nouvelle partition. C'est ce qu'on appelle une stratégie d'agrégation (3ème choix).

Faisons le choix suivant : si A et B sont deux sous-ensembles non vides de E

$$\partial(A, B) = \sup \{d(E_i, E_j) \ ; \ E_i \in A, E_j \in B\}$$

On calcule ainsi une nouvelle matrice de « proximité » entre les éléments de la partition $\mathcal{T}_1$.

$$\begin{array}{c} \\ \{E_1,E_5\} \\ \{E_2\} \\ \{E_3\} \\ \{E_4\} \\ \{E_6\} \end{array} \begin{array}{c} \begin{array}{ccccc} \{E_1,E_5\} & \{E_2\} & \{E_3\} & \{E_4\} & \{E_6\} \end{array} \\ \begin{pmatrix} 1/4 & 3/4 & 1 & 1/2 & 1/2 \\ 3/4 & 0 & 1/4 & 3/4 & 1/4 \\ 1 & 1/4 & 0 & 1/2 & 1/2 \\ 1/2 & 3/4 & 1/2 & 0 & 1/2 \\ 1/2 & 1/4 & 1/2 & 1/2 & 0 \end{pmatrix} \end{array} = D_1$$

Le choix de la stratégie d'agrégation (proximité entre deux ensembles) nous conduit à réunir 2 éléments distincts de $\mathcal{T}_1$ de telle sorte que le diamètre de la réunion (c'est-à-dire Sup $d(E_i, E_j)$ pour $E_i \in A \cup B$, $E_j \in A \cup B$) soit minimum.

Le minimum de la proximité entre 2 éléments distincts de $\mathcal{T}_1$ est 1/4. Il est atteint par le couple $\{E_2, E_3\}$. On obtient la nouvelle partition :

$$\{ \{E_1, E_5\}, \{E_2, E_3\}, \{E_4\}, \{E_6\} \} = \mathcal{T}_2$$

Calculons la matrice D_2 des « proximités » entre les éléments de $\mathcal{T}_2$

$$\begin{array}{c} \\ \{E_1,E_5\} \\ \{E_2,E_3\} \\ \{E_4\} \\ \{E_6\} \end{array} \begin{array}{c} \begin{array}{cccc} \{E_1,E_5\} & \{E_2,E_3\} & \{E_4\} & \{E_6\} \end{array} \\ \begin{pmatrix} 1/4 & 1 & 1/2 & 1/2 \\ 1 & 1/4 & 3/4 & 1/2 \\ 1/2 & 3/4 & 0 & 1/2 \\ 1/2 & 1/2 & 1/2 & 0 \end{pmatrix} \end{array} = D_2$$

Le minimum de la « proximité » entre 2 éléments de $\mathcal{T}_2$ est 1/2. Il est réalisé par les éléments $\{E_1, E_5\}$, $\{E_4\}$. On obtient la nouvelle partition :

$$\{ \{E_1, E_5, E_4\} , \{E_2, E_3\} , \{E_6\} \} = \mathcal{T}_3$$

La nouvelle matrice de « proximité » D_3 entre les éléments de $\mathcal{T}_3$ est :

$$\begin{array}{c} \\ \{E_1,E_5,E_4\} \\ \{E_2,E_3\} \\ \{E_6\} \end{array} \begin{array}{c} \begin{array}{ccc} \{E_1,E_5,E_4\} & \{E_2,E_3\} & \{E_6\} \end{array} \\ \begin{pmatrix} 1/2 & 1 & 3/4 \\ 1 & 1/4 & 1/2 \\ 3/4 & 1/2 & 0 \end{pmatrix} \end{array} = D_3$$

Le minimum de la « proximité » entre 2 éléments distincts de $\mathcal{T}_3$ est 1/2. Il est réalisé par les éléments $\{E_2, E_3\}$ et $\{E_6\}$. On obtient la nouvelle partition :

$$\{ \{E_1, E_5, E_4\} , \{E_2, E_3, E_6\} \} = \mathcal{T}_4$$

La matrice D_4 des « proximités » entre les éléments de $\mathcal{T}_4$ est :

$$\begin{array}{c} \\ \{E_1,E_5,E_4\} \\ \{E_2,E_3,E_6\} \end{array} \begin{array}{c} \begin{array}{cc} \{E_1, E_5, E_4\} & \{E_2, E_3, E_6\} \end{array} \\ \begin{pmatrix} 1/2 & 1 \\ 1 & 1/2 \end{pmatrix} \end{array} = D_4$$

La procédure s'arrête, à l'étape suivante on obtient la partition pleine :

$$\{ \{E_1, E_2, E_3, E_4, E_5, E_6\} \} = \mathcal{T}_5$$

Considérons $\mathcal{H} = \bigcup_{i=1}^{5} \mathcal{T}_i$

$$= \{\{E_1\}, \{E_2\}, \{E_3\}, \{E_4\}, \{E_5\}, \{E_6\}, \{E_1,E_5\}, \{E_2,E_3\}, \{E_1,E_5,E_4\}, \{E_2,E_3,E_6\}, E\}$$

Les choix (2) (indice de distance, ici distance déduite de la différence symétrique) et (3) (stratégie d'agrégation, ici le diamètre minimum) permettent de définir une application f : $\mathcal{H} \longrightarrow \mathbf{R}$

$f(\{E_i\}) = d(E_i, E_i) = 0 \qquad$ pour i = 1... 6

$f(\{E_1, E_5\}) = d(E_1, E_5) = 1/4$

$f(\{E_2, E_3\}) = d(E_2, E_3) = 1/4$

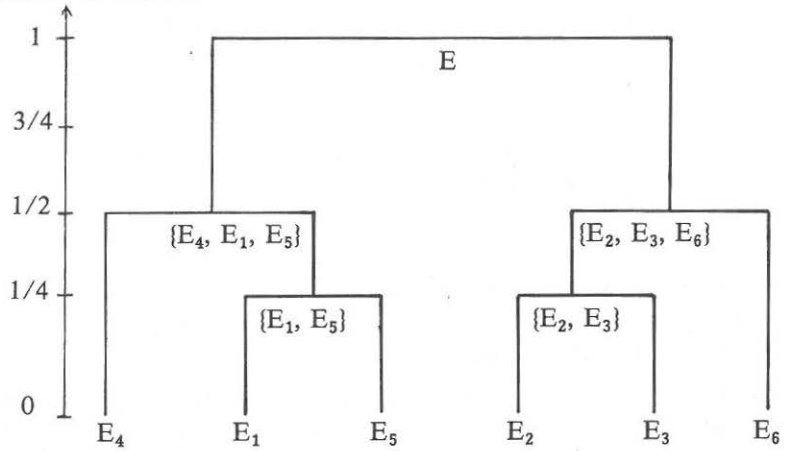
$f(\{E_1, E_5, E_4\}) = d(\{E_1, E_5\}, E_4) = 1/2$

$f(\{E_2, E_3, E_6\}) = d(\{E_2, E_3\}, E_6) = 1/2$

$f(E) = d(\{E_1, E_5, E_4\}, \{E_2, E_3, E_6\}) = 1$

Pour $A \in \mathcal{H}$, f(A) est égal au minimum de la matrice de « proximité » ayant servi à construire A, c'est-à-dire que pour A et $B \in \mathcal{H}$ on a $f(A \cup B) = \partial(A, B)$ (où ∂ est le prolongement à $\mathcal{T}$ (E), ici par la stratégie d'agrégation du saut minimum, de d indice de distance déduit de la différence symétrique). L'application f est croissante et ($\mathcal{H}$, f) est une hiérarchie indiciée.

Cela nous conduit à la représentation graphique suivante de la procédure de C.H.A.



f permet de définir sur E une distance d* par la formule

$$d^*(E_i, E_j) = f(A_{ij})$$

où A_{ij} est le plus petit élément de $\mathcal{H}$ contenant $\{E_i, E_j\}$
(le plus petit élément s'entend au sens de l'inclusion)

Comme nous l'avons montré dans le paragraphe précédent d* est une distance ultramétrique sur E. C'est-à-dire vérifie :

1) $d^*(E_i, E_j) = 0 \iff E_i = E_j$
2) $d^*(E_i, E_j) = d^*(E_j, E_i)$
3) $d^*(E_i, E_j) \leqslant \sup (d^*(E_i, E_k), d^*(E_k, E_j))$

On calcule maintenant une matrice D* de distance entre les éléments de E muni de la distance d*. La « comparaison » de D* et de D_0 va permettre de mesurer la déformation due à la procédure de C.H.A. Il convient de donner un critère de comparaison entre les distances d et d*. C'est le choix d'un **critère de déformation.** (4ème choix).

La matrice D* est :

$$\begin{array}{c} \\ E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \\ E_6 \end{array} \begin{array}{c} \begin{array}{cccccc} E_1 & E_2 & E_3 & E_4 & E_5 & E_6 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1/2 & 1/4 & 1 \\ 1 & 0 & 1/4 & 1 & 1 & 1/2 \\ 1 & 1/4 & 0 & 1 & 1 & 1/2 \\ 1/2 & 1 & 1 & 0 & 1/2 & 1 \\ 1/4 & 1 & 1 & 1/2 & 0 & 1 \\ 1 & 1/2 & 1/2 & 1 & 1 & 0 \end{pmatrix} \end{array} = D^*$$

Pour mesurer la déformation on peut prendre

$$DEF(d, d^*) = \sup_{\substack{i, j \\ i = 1...5 \\ j = 2...6}} |d^*(E_i, E_j) - d(E_i, E_j)|$$

On obtient dans l'exemple

$$DEF(d, d^*) = \frac{1}{2}$$

Cela correspond par exemple à $d(E_1, E_6) = 1/2$ $\qquad$ $d^*(E_1, E_6) = 1$

Bien entendu, d'autres critères de déformation sont possibles, nous en donnerons quelques exemples dans la suite.

Notons cependant, dès maintenant, que la déformation est importante ; en effet d et d* varient de 0 à 1. Cela provient du fait que les élèves « moyens » E_1, E_4, E_6 sont à « mi-chemin » du « bon élève » E_5 et des « mauvais » élèves E_2 et E_3. Notre stratégie d'agrégation nous conduit à exacerber les différences entre les « bons » et les « mauvais » alors que le rattachement des moyens à l'un des deux groupes correspond à un choix relativement arbitraire. En effet, seules les partitions $\mathcal{P}_4$ et $\mathcal{P}_5$ ont été imposées. On voit apparaître ici des déformations liées à l'effet de chaîne et qui sont un des inconvénients de cette procédure.

Nous allons dans les paragraphes suivants préciser les différents choix possibles et les éléments mathématiques fondant la généralisation de la méthode.

Les différents types de tableaux

Il s'agit de préciser ici la discussion concernant le 1er choix. Nous supposerons résolu le difficile problème du choix des variables. Cela nous échappe en tant que statisticien, c'est le problème essentiel du spécialiste. Voyons un exemple.

Si un sociologue veut mettre en évidence les différences de consommation selon la place sociale de l'individu : c'est à lui de décider s'il utilisera la nomenclature de l'INSEE (avec 1 ou 2 ou 3 chiffres) ou s'il construira une classification qu'il justifiera. Ce qui est certain, c'est que certaines agrégations trop artificielles rendront toute analyse ultérieure vaine.

En ce qui concerne la forme du tableau on distingue généralement :
— les tableaux de fréquences
— les tableaux de mesures
— les tableaux de description logique.

LES TABLEAUX DE FREQUENCES

Un tableau de fréquences est une loi de probabilité sur un ensemble produit I × J. Soient m le nombre d'éléments de I, n celui de J. On note :

Pij la probabilité de la case d'indice (i, j)

$$Pi. = \sum_{j} Pij \qquad i = 1...m$$

$$P.j = \sum_{i} Pij \qquad j = 1...n$$

(Pi.) i = 1...n est la loi de probabilité marginale sur I
(P.j) j = 1...m est la loi de probabilité marginale sur J

$Pi(j) = \frac{Pij}{Pi.}$ lorsque j varie de 1 à n, Pi(j) est la loi de probabilité conditionnelle sur J pour i fixé

$Pj(i) = \frac{Pij}{P.j}$ lorsque i varie de 1 à m, Pj(i) est la loi de probabilité conditionnelle sur I pour j fixé.

Les colonnes et les lignes du tableau apparaissent donc essentiellement comme (à un facteur près) des distributions de probabilités. Il y a parfaite symétrie entre lignes et colonnes.

On a donc une telle situation si on considère une matrice de mobilité sociale : les lignes représentent la catégorie professionnelle des pères, les colonnes celles des fils, la case (i, j) contient le pourcentage de fils de catégorie j ayant un père de catégorie i.

LES TABLEAUX DE MESURES

I représente un ensemble d'individus, J les variables observées, la case (i, j) représente l'intensité du caractère j pour l'individu i. On note :

x_{ij} l'intensité du caractère j pour l'individu i

$$x_{i.} = \sum_{j=1}^{n} x_{ij} \qquad i = 1...m$$

$$x_{.j} = \sum_{i=1}^{m} x_{ij} \qquad j = 1...n$$

$$x = \sum_{i,j} x_{ij}$$

Les colonnes de la matrice (x_{ij}) sont des distributions statistiques des variables sur la population I. On calcule :

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{m} x_{ij} = \frac{x.j}{n} \qquad \text{(moyenne de la variable j)}$$

$$\sigma_j^2 = \frac{1}{n} \sum_{i=1}^{m} (x_{ij} - \bar{x}_j)^2 \qquad \text{(variance de la variable j)}$$

Les lignes de la matrice (x_{ij}) donnent les caractéristiques de l'individu i par rapport aux caractères. Si les variables sont homogènes, on définit

$$\bar{x}_i = \frac{1}{m} \sum_{j=1}^{n} x_{ij} = \frac{x_{i.}}{m}$$

$$\sigma_i^2 = \frac{1}{m} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2$$

Une situation correspondant à un tableau de mesure est par exemple :

I : ensemble d'élèves
J : ensembles de disciplines scolaires
x_{ij} : « note » de l'élève i dans la discipline j.

Faute de pouvoir réaliser cette homogénéité, on est parfois amené à transformer un tableau de mesures en un tableau de présence-absence, c'est-à-dire un tableau de description logique.

LES TABLEAUX DE DESCRIPTION LOGIQUE

I est une population, J un ensemble de caractéristiques.
$x_{ij} = 1$ si l'individu i a la caractéristique j.
$x_{ij} = 0$ si l'individu i n'a pas la caractéristique j.

Les lignes du tableau décrivent les individus par rapport aux caractères. Dans le cas d'un questionnaire (oui-non) on dit qu'ils constituent les patrons de réponses observés. Pour deux individus i et i′ on note :

00(i, i′) = le nombre de zéros communs aux patrons correspondant à i et i′ (c'est-à-dire le nombre de caractéristiques absentes chez les individus i et i′ simultanément)

11(i, i′) = le nombre de caractéristiques présentes simultanément chez les individus i et i′

01(i, i′) = le nombre de caractéristiques absentes chez i mais présentes chez i′

10(i, i′) = le nombre de caractéristiques présentes chez i mais absentes chez i′

Les colonnes du tableau décrivent le comportement des individus par rapport à chaque caractéristique.

En inversant les rôles des individus et des variables, on définit de même :

00(j, j′) ; 11(j, j′) ; 01(j, j′) ; 10(j, j′)

Comme dans le cas des tableaux de mesures, on définit $x_{i.}$ et $x_{.j}$ qui représentent ici le nombre de caractéristiques observées sur l'individu i et le nombre d'individus ayant la caractéristique j.

Comme exemple on peut prendre :

I un ensemble de régions géographiques

J un ensemble de plantes

x_{ij} note la présence ou l'absence de la plante j dans la région i.

Principaux indices de distances (2ème Choix)

Pour un éventail plus détaillé des différents indices de distances, on pourra se reporter à l'article de Michel JAMBU [5].

On appelle indice de distance sur un ensemble E une application d de $E \times E$ dans R^+ vérifiant :

1) $(d(x, y) = 0) \iff (x = y)$ pour tout x et y dans E

2) $d(x, y) = d(y, x)$ pour tout x et y dans E.

Une distance est un indice de distance vérifiant en plus l'inégalité triangulaire

3) $d(x, z) \leqslant d(x, y) + d(z, y)$ pour tout x, y, z dans E.

L'objet de ce second choix est de définir, à l'aide d'un indice de distance, une relation de proximité entre les lignes ou les colonnes d'un tableau de données.

INDICE DE DISTANCE DANS LES TABLEAUX DE FREQUENCES [1]

Les lignes et les colonnes étant essentiellement des distributions de probabilités, c'est la distance du χ^2 (khi deux) qu'il apparaît souhaitable d'employer. Pour une justification théorique complète de ce choix, on pourra se reporter à J.P. BENZECRI (B.G. [1]) LEBART et FENELON (B.G. [8]).

La distance entre les deux lignes s'écrit :

$$d_{11}(i, i') = \left[\sum_{j=1}^{m} \frac{1}{P_{.j}} \left(\frac{P_{ij}}{P_{i.}} - \frac{P_{i'j}}{P_{i'.}} \right)^2 \right]^{1/2}$$

INDICE DE DISTANCE DANS LES TABLEAUX DE MESURES[1]

Nous ne retiendrons ici que 3 distances :

- la distance euclidienne usuelle
- la distance euclidienne réduite
- la distance du χ^2 (khi deux).

Ces 3 distances correspondent respectivement à celles utilisées en analyse en composantes principales, en analyse en composantes principales normées et en analyse des correspondances, 3 techniques d'analyse des données d'origine géométrique.

On a :

$$d_{21}(i, i') = \left[\sum_{j=1}^{n} (x_{ij} - x_{i'j})^2 \right]^{1/2}$$

Elle correspond à la vision du nuage de points des individus dans l'espace R^n des variables muni de la distance euclidienne canonique.

Du point de vue de l'analyse des données, la principale critique est qu'elle favorise les variables ayant une grande dispersion absolue (σ_j écart type) mais éventuellement faible relativement à $\bar{x}_j$ ($\frac{\sigma_j}{\bar{x}_j}$ coefficient de variation).

Dans une enquête de consommation cela avantage le pain au détriment du poivre. Pour cela, on est amené à changer d'unité sur chaque axe. On prend alors :

$$d_{22}(i, i') = \left[\sum_{j=1}^{n} \frac{1}{\sigma_j^2} (x_{ij} - x_{i'j})^2 \right]^{1/2}$$

(1) Les distances proposées dans ce cas, dans cet article, sont euclidiennes. En fait, cela n'est absolument pas nécessaire. Par exemple, quelqu'un étudiant les classements faits par des individus (acheteurs potentiels) à propos des propriétés souhaitées d'un produit (voiture) par exemple : sécurité, vitesse... pourra définir la distance entre deux individus par le nombre d'inversions existant entre les deux classements.

Enfin, en divisant chaque x_{ij} par x on peut passer d'un tableau de mesures à un tableau de fréquences. Cela conduit à prendre la distance du χ^2

$$d_{23}(i, i') = \left[\sum_{j=1}^{m} \frac{x}{x_{.j}} \left(\frac{x_{ij}}{x_{i.}} - \frac{x_{i'j}}{x_{i'.}} \right)^2 \right]^{1/2}$$

INDICES DE DISTANCES DANS LES TABLEAUX DE DESCRIPTION LOGIQUE

Dans ce cas les choix sont multiples, ils consistent à combiner entre eux les nombres 00(i, i'), 11(i, i'), 10(i, i'), 01(i, i').

Nous ne retiendrons dans cette classe d'indices de distances que la distance de la différence symétrique

$$d_{31}(i, i') = \frac{10(i, i') + 01(i, i')}{n}$$

Interprétant un tableau de présence absence comme un tableau de mesure on peut encore prendre la distance du χ^2

$$d_{32}(i, i') = \left[\sum_{j=1}^{m} \frac{x}{x_{.j}} \left(\frac{x_{ij}}{x_{i.}} - \frac{x_{i'j}}{x_{i'.}} \right)^2 \right]^{1/2}$$

Bien entendu, pour avoir la distance entre deux lignes, il suffit d'échanger les rôles de i et de j. Ceci dans tous les cas. Avant d'exhiber quelques stratégies d'agrégation, donnons la mise en route du processus de classification hiérarchique ascendante après le choix d'un indice de distance sur I (resp. sur J).

On calcule la matrice des distances D_0 sur les éléments de I (resp. de J). Soit h le plus petit élément non nul de D_0. Ce minimum est atteint pour deux éléments i et i' distincts. On obtient ainsi une partition :

$$\mathcal{T}_1 = \{\{i, i'\}\} \cup \{\{i''\} \quad ; \quad i'' \neq i \text{ et } i'' \neq i'\}$$

Pour passer à l'étape suivante, il est nécessaire de définir une relation de proximité entre les éléments de $\mathcal{T}_1$. Pour poursuivre le processus, il faut pour tout $k(1 \leqslant k \leqslant m-2)$ savoir construire une partition $\mathcal{T}_{k+1}$ à partir de la partition $\mathcal{T}_k$. C'est cela une stratégie d'agrégation.

Quelques stratégies d'agrégation (3ème choix)

Si nous savons construire, à partir de l'indice de distance d défini sur I, un indice sur $\mathcal{P}_k$ prolongeant d, il suffit d'itérer la 1ère étape. C'est ce que nous avons dans la stratégie du saut minimum.

STRATEGIE DU SAUT MINIMUM

Pour A et B dans $\mathcal{P}(I) - \phi$ on définit

$$\partial(A, B) = \inf_{\substack{a \in A \\ b \in B}} d(a, b)$$

Ce n'est pas un indice de distance sur $\mathcal{P}(I) - \phi$. C'est seulement un semi-indice de distance, en effet, si $A \cap B \neq \phi$ alors $\partial(A, B) = 0$. C'est un indice de distance sur toute partition de E. Connaissant la partition $\mathcal{P}_k$ obtenue à l'étape $k(k \leqslant n-2)$ on calcule la matrice D_k des distances entre les éléments de $\mathcal{P}_k$. Soit h_{k+1} le plus petit élément non nul de D_k. Ce minimum est atteint au moins pour deux éléments A et B de $\mathcal{P}_k$. On prend pour partition $\mathcal{P}_{k+1}$

$$\mathcal{P}_{k+1} = \{A \cup B\} \cup (\mathcal{P}_k \setminus \{A, B\})$$

Le principe d'une telle stratégie est de réunir deux parties de I dont le « saut » est minimum (i.e, il suffit d'une paire d'éléments proches pour décider de réunir deux parties).

Une autre possibilité est de définir sur les partitions issues de $\mathcal{P}_k$ par réunion de deux parties, une application et de choisir $\mathcal{P}_{k+1}$ parmi ces partitions celle qui minimise ou maximise cette application. La stratégie du diamètre minimum procède de cette idée.

STRATEGIE DU DIAMETRE MINIMUM

Pour A et B dans $\mathcal{P}(I) - \phi$ on définit :

$$\partial(A, B) = \sup_{\substack{a \in A \\ b \in B}} d(a, b)$$

Ce n'est plus un semi-indice de distance. En effet, si A n'est pas un singleton, $d(A, A) \neq 0$.

Connaissant la partition $\mathcal{P}_k$ obtenue à l'étape $k(k \leqslant m-2)$, on calcule la matrice D_k qui est la matrice des diamètres des éléments de $\mathcal{P}_k$ réunis deux à deux. Soit h_{k+1} le plus petit élément non diagonal de D_k. Ce minimum est atteint pour deux éléments A et B de $\mathcal{P}_k$. On prend pour partition $\mathcal{P}_{k+1}$

$$\mathcal{P}_{k+1} = \{A \cup B\} \cup (\mathcal{P}_k \setminus \{A, B\})$$

Le principe d'une telle stratégie est de réunir deux parties de I en minimisant le diamètre de la réunion.

Remarquons que si A et B sont des singletons, on a bien la distance initiale sur les éléments de $\mathcal{T}_0$.

STRATEGIE DE LA DISTANCE MOYENNE

Les deux stratégies que nous venons de décrire ne tiennent en définitive compte que des éléments les plus proches ou au contraire des éléments extrêmes. La stratégie que nous allons décrire tient compte de tous les éléments.

Nous noterons Card A le nombre d'éléments d'un ensemble fini A.

Pour A et B appartenant à $\mathcal{T}(I) - \phi$ on définit :

$$\partial(A, B) = \frac{\sum_{a \in A,\, b \in B} d(a, b)}{\text{Card } A \times \text{Card } B}$$

Ce n'est pas un semi-indice de distance car $\partial(A, A) \neq 0$ si A n'est pas un singleton. Si A et B sont des singletons, ∂ coïncide avec l'indice de distance d défini sur E.

Connaissant la partition $\mathcal{T}_k$ obtenue à l'étape $k(k \leqslant m-2)$ on calcule la matrice D_k des moyennes des distances entre les éléments de $\mathcal{T}_k$.

Soit h_{k+1} le plus petit élément non diagonal de D_k. Ce minimum est atteint pour deux éléments A et B de $\mathcal{T}_k$. On prend pour partition $\mathcal{T}_{k+1}$

$$\mathcal{T}_{k+1} = \{A \cup B\} \cup (\mathcal{T}_k \setminus \{A, B\})$$

Le principe de cette stratégie est de réunir deux parties de I en minimisant la moyenne des distances de la réunion.

Les 3 stratégies d'agrégation que nous venons de présenter n'utilisent pas le fait que le nuage de points I peut être considéré dans un espace vectoriel.

Pour l'utilisateur elles correspondent à des choix bien précis :

- s'il veut atténuer les différences entre les groupes, il choisit la stratégie du saut minimum (les amis de mes amis sont mes amis).
- s'il souhaite opposer les groupes, il choisira la stratégie du diamètre moyen. Il valorise les éléments extrêmes.
- s'il essaie de tenir compte de tous les éléments, il choisira la stratégie de la distance moyenne.

Pour les deux premières stratégies, le calcul de la matrice D_{k+1} à partir de la matrice D_k est évident. Dans le cas de la stratégie de la distance moyenne, explicitons ce passage. Il faut calculer :

$\partial(C, A \cup B)$ pour C, A, B appartenant à $\mathcal{T}_k$

Pour A, B, C distincts appartenants à $\mathcal{T}_k$, $A \cap B = \phi$ d'où :

$$\sum_{\substack{x \in A \cup B \\ c \in C}} d(c, x) = \sum_{\substack{a \in A \\ c \in C}} d(c, a) + \sum_{\substack{b \in B \\ c \in C}} d(c, b)$$

$$= \text{card } C \times \text{card } A \, \partial(C, A) + \text{card } C \times \text{card } B \, \partial(C, B)$$

il en résulte, comme card $A \cup B$ = card A + card B

$$\partial(C, A \cup B) = \frac{\text{card } A \times \partial(C, A) + \text{card } B \, \partial(C, B)}{\text{card } A + \text{card } B}$$

Il existe d'autres stratégies d'agrégation utilisant notamment la structure euclidienne de l'espace ambiant au nuage de points. On pourra se reporter aux articles de M. JAMBU pour une présentation de ces stratégies. Dans cet article, la simplicité nous a conduit à limiter à ces 3 stratégies notre présentation.

Déformation

Nous allons montrer qu'une procédure de C.H.A. sur l'ensemble des individus I permet de définir sur I une nouvelle distance d*. Soit m le nombre des individus. La procédure de C.H.A. comporte m étapes numérotées de 0 à m − 1. Dans l'ensemble $\mathcal{T}(I) - \phi$ considérons :

$$\mathcal{H} = \bigcup_{k=0}^{m-1} \mathcal{T}_k$$

$\mathcal{T}_0 = \{\{i\} ; i \in I\}$ et $\mathcal{T}_{m-1} = \{I\}$. Le passage de $\mathcal{T}_k$ à $\mathcal{T}_{k+1}$ étant assuré par la réunion de deux parties les plus proches de $\mathcal{T}_k$. Par construction $\mathcal{H}$ vérifie les propriétés suivantes :

$$\forall A, B \in \mathcal{H}, \quad A \cap B \in \{A, B, \phi\}$$

$$\forall A \in \mathcal{H} \quad U = \{B \in \mathcal{H} ; B \neq A \text{ et } B \subset A\} = \{\phi\}$$

$$\forall A \in \mathcal{H} \ \forall i \in A \quad \{i\} \in \mathcal{H}$$

$\mathcal{H}$ est une hiérarchie totale de parties de I, donc une hiérarchie sur I. On peut en outre définir une application $f : \mathcal{H} \longrightarrow \mathbf{R}^+$ par le formulaire suivant :

$$f(\{i\}) = 0 \text{ pour tout } i \in I$$

Pour $k \geqslant 1$ $f(A \cup B) = h_{k+1}$ si $A \in \mathcal{P}_k$, $B \in \mathcal{P}_k$, $A \cup B \in \mathcal{P}_{k+1}$ et $\partial(A,B) = h_{k+1}$.

Par construction f est croissante et $(\mathcal{H}, f)$ est une hiérarchie indicée. Comme nous l'avons montré précédemment,

$d^*(i,j) = f(A_{ij})$ où A_{ij} est le plus petit élément de $\mathcal{H}$ (au sens de l'inclusion) contenant {i, j} est une distance ultramétrique.

Nous pouvons alors calculer la matrice D* des distances entre les éléments de I muni de la distance d*. Pour mesurer la déformation, il faut se donner un critère de déformation permettant de mesurer l'écart entre D_0 et D_*.

Quelques critères de déformations (4ème choix)

Nous n'aborderons ici que quelques critères de comparaison de type numérique. Nous ne dirons rien de ceux de type ordinal, fondés sur la possibilité de définir une relation de préordonnance à partir d'une distance. Le lecteur désireux d'en connaître davantage pourra se reporter aux articles de M. JAMBU par exemple.

Un critère de déformation D E F (d, d*) devra vérifier :

$$D\,E\,F\,(d, d^*) = 0 \text{ si } d = d^*$$

Cela nous conduit à proposer les critères suivants :

$$D\,E\,F_1\,(d, d^*) = \sup_{(i,i')\in I^*} |d(i, i') - d^*(i, i')|$$

où $I^* = \{(i, i') \in I \times I \text{ avec } i > i'\}$ (par abus de langage on identifie l'individu et son numéro)

$$\text{card } I^* = \frac{m(m-1)}{2}$$

On définit également

$$D\,E\,F_2\,(d, d^*) = \frac{1}{\text{card } I^*} \sum_{(i,i')\in I^*} |d(i, i') - d^*(i, i')|$$

$$D\,E\,F_3\,(d, d^*) = \frac{1}{\text{card } I^*} \sum_{(i,i')\in I^*} (d(i, i') - d^*(i, i'))^2 \quad 1/2$$

Remarquons que les 3 critères que nous venons de définir vérifient

$$D\,E\,F\,(d, d^*) = 0 \iff d = d^*.$$

Géométriquement d et d* sont deux points dans un espace de dimension card I* : mesurer la déformation, c'est mesurer une distance entre les deux points représentant d et d*. Une fois encore l'analyse des données apparaît comme une illustration de l'algèbre linéaire.

Vertige

Devant le nombre de choix à faire, l'angoisse et le doute risquent d'envahir l'esprit du mathématicien. Le matheux souhaiterait trouver dans ce dédale un choix optimal. En fait, cette recherche est illusoire et il faut apprendre à vivre sans cette sécurité. Il faudra à l'utilisateur prendre l'habitude de justifier ses choix du propre point de vue de sa discipline, rechercher les similitudes entre les partitions obtenues en faisant des choix différents ; enfin dégager de cela une partition en classes stables suivant les critères d'agrégation et facilement caractérisables.

Eléments pour une comparaison

Nous avons vu qu'une procédure de C.H.A. associe aux différents choix une distance ultramétrique d*. Pour mesurer la ressemblance ou la dissemblance entre deux procédures de C.H.A. on peut prendre :

$$D\ E\ F\ (d_1^*, d_2^*)$$

où D E F est un des critères de déformation.

Une autre possibilité consiste à mesurer la distance entre deux hiérarchies de parties $\mathcal{H}^1$ et $\mathcal{H}^2$ provenant de 2 procédures de C.H.A.

$$d^0(\mathcal{H}^1, \mathcal{H}^2) = \sum_{k=1}^{m-1} \text{card}\,(A_k^1 \triangle A_k^2)$$

où A_k^1 : est l'ensemble formé à l'étape k dans la hiérarchie $\mathcal{H}^1$

A_k^2 : est l'ensemble formé à l'étape k dans la hiérarchie $\mathcal{H}^2$

$A_k^1 \triangle A_k^2$: est la différence symétrique entre les deux ensembles.

Une critique que nous pouvons faire à cette distance est qu'elle donne le même poids aux différentes étapes alors que l'information contenue dans les étapes successives croît avec k.

Cela nous conduit à pondérer la distance d^0 ; on obtient la distance

$$d^{00}(\mathcal{H}^1, \mathcal{H}^2) = \sum_{k=1}^{m-1} k\ \text{card}\,(A_k^1 \triangle A_k^2)$$

Donnons pour terminer un critère de comparaison entre deux partitions q^1 et q^2 d'un ensemble I.

Soient q_1 et q_2 le nombre d'éléments de chacune d'elles. On fabrique un tableau à q_1 lignes et q_2 colonnes

$$K(A_i^1, A_j^2) = \text{card } A_i^1 \cap A_j^2 \text{ pour } \begin{matrix} i = 1 \ldots q_1 \\ j = 1 \ldots q_2 \end{matrix}$$

où A_i^1 sont les éléments de q^1 et A_j^2 ceux de (q^2)

$$K(A_i^1) = \sum_j K(A_i^1, A_j^2)$$

$$K(A_j^2) = \sum_i K(A_i^1, A_j^2)$$

$$K = \sum_{i,j} K(A_i^1, A_j^2)$$

on peut alors faire un test de χ^2 à la quantité

$$x^2(q^1,q^2) = K \left[\sum_{\substack{i=1\ldots q_1 \\ j=1\ldots q_2}} \frac{K^2(A_i^1, A_j^2)}{K(A_i^1)\, K(A_j^2)} - 1 \right]$$

On prend un χ^2 à $(q_1 - 1)(q_2 - 1)$ degrés de liberté.

- si $x^2(q^1, q^2) > \chi^2_\alpha$ au seuil α les deux partitions ne sont pas étrangères l'une à l'autre.
- si $x^2(q^1, q^2) < \chi^2_\alpha$ au seuil α les deux partitions sont indépendantes l'une de l'autre.

Ce critère permet de tester la stabilité des classes obtenues en « sciant » deux hiérarchies à une certaine étape.

Classification et modèles factoriels

La méthode de classification hiérarchique ascendante, telle que nous l'avons brièvement décrite, apparaît comme une technique relativement simple d'analyse de données. Les liens l'unissant aux modèles factoriels n'apparaissent pas immédiatement. Cela provient du fait que les stratégies d'agrégation proposées n'utilisent pas la structure euclidienne de l'espace ambiant du nuage de points. Sans entrer dans une discussion complète nos voudrions dans le cas de la distance du χ^2 illustrer les liens entre les deux domaines.

Avec les notations introduites à propos des tableaux de fréquences notons :

$$N_J(I) = \{(P_i(j)_{j = 1..n})\ i = 1 \ldots m\}, \text{ avec } P_i(j) = \frac{P_{ij}}{P_{i.}},$$

le nuage des profils d'individus M_i, de coordonnées $P_i(j)$, affectés des masses $P_{i.}$.

De même on définit le nuage

$$N_I(J) = \{(P_j(i)_{i=1..m})\ j = 1..\ n\}$$ dont chaque point

$$M_j = (P_j(i)_{i=1..m})$$ est muni de la masse $P_{.j}$

Soit A un sous-ensemble de I. Notons :

$$P_{Aj} = \sum_{i \in A} P_{ij}$$

$$P_{A.} = \sum_{i \in A} P_{i.}$$

Le centre de gravité de A est :

$$g(A) = \sum_{i \in A} \frac{P_{i.}}{P_{A.}} M_i$$

où M_i a pour coordonnée $(P_i(j))_{j=1..n}$

Les coordonnées de g(A) sont donc $(\frac{P_{Aj}}{P_{A.}})\ j = 1..n$

Ainsi une classe peut être représentée par son centre de gravité affecté de la masse $P_{A.}$.

Si Q est une partition de I on lui associe un nuage de points

$$N_J(Q) = \{g(A)\ ;\ A \in Q\}$$

où g(A) est le centre de gravité de A affecté de la masse $P_{A.}$.

Nous pouvons donc considérer une hiérarchie comme une suite de points, il devient alors licite de regarder comment ces points se situent par rapport aux axes factoriels.

Si nous disposons d'une partition K de J on définit de la même façon $N_{(I)}(K) = \{g(A)\ ;\ A \in K\}$ nuage des points affectés des masses $P_{.A}$.

La représentation simultanée permet alors de préciser, notamment par le calcul, les liens entre les deux hiérarchies, sur I et J.

Une présentation complète de ceci dépasserait la simple présentation des méthodes de C.H.A. objet du présent article.

Conclusion

La présentation élémentaire que nous venons de faire de la C.H.A. appelle plusieurs remarques :

1) Il existe d'autres techniques de classifications : l'analyse des liens (J.B. RACINE 1972), les nuées dynamiques (DIDAY, cf. **5**) ; la classification de I.C. LERMANN (cf. **6**) etc.

2) L'algorithme de la C.H.A. débouche, non sur la construction d'une partition, mais sur celle d'une famille de partitions (une hiérarchie de parties). Une telle procédure présente deux inconvénients :

— Elle nécessite la mise en mémoire de $\frac{\text{card E(card E} - 1)}{2}$ distances.

Dès que nous dépassons quelques centaines d'individus, une telle nécessité devient insurmontable.

— Il faut avoir recours à la mesure de l'inertie retenue, par exemple, pour décider de la hauteur à laquelle on scie l'arbre.

En fait, dans de nombreux cas, on souhaite seulement obtenir une partition de l'ensemble de départ et mesurer le caractère significatif de celle-ci.

L'algorithme des nuées dynamiques de DIDAY et l'algorithme de la vraisemblance du lien de I.C. LERMANN répondent à cette préoccupation.

Il n'est pas question de présenter ici, ces deux algorithmes. Celui de DIDAY est présenté dans **5** ; dans **6**, I.C. LERMANN utilise une classification obtenue par cette procédure et la présente.

Bibliographie

(Cf. aussi [1], [2], [7], [8] et [9] de la bibliographie générale).

[1] CEHESSAT : Exercices de statistique et d'informatique appliquées. DUNOD 1976 (418 p.)

[2] CHADULE : Analyse de données géographiques. MASSON, 1974 (190 p.).

[3] M. JAMBU : Programme de calcul des contributions mutuelles entre classes d'une hiérarchie et facteurs de correspondance (*Cahiers de l'Analyse des données* n° 1, 1976).

[4] M. JAMBU : Introduction à l'Analyse des données. Les méthodes de classification automatique. *Consommation* n° 3, 1973.

[5] M. JAMBU : Sur les indices de distances en vue d'une classification ascendante hiérarchique. *Consommation* n° 2, 1974.

[6] M. JAMBU : Sur les critères d'Agrégations utilisées en classification automatique. *Consommation* n° 4, 1974.

E. DIDAY
Université Paris IX Dauphine

5. INTRODUCTION A LA METHODE DES NUEES DYNAMIQUES

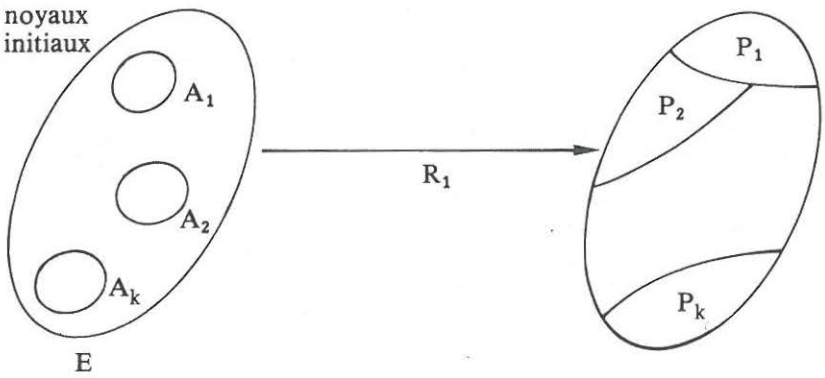
INTRODUCTION ET RESUME DE LA METHODE

1. La méthode des **nuées dynamiques** se situe dans le groupe des méthodes de classification **non hiérarchiques**. Cela signifie que l'objet de la méthode n'est pas de construire une suite de partitions d'un ensemble E : arbre hiérarchique (LETOURNEUX, **4** LERMAN **6**), mais une simple partition. En fait, la perte constituée par l'absence d'arbre hiérarchique, est avantageusement compensée par le fait que l'on peut travailler sur des ensembles E de cardinaux bien plus grands avec des temps de calcul (sur ordinateur) très raisonnables.

2. Donnons un bref résumé de la méthode :

Soit E un ensemble fini muni d'une distance d, on se propose de construire une partition de E en k (fixé) classes. Pour cela on se donne :

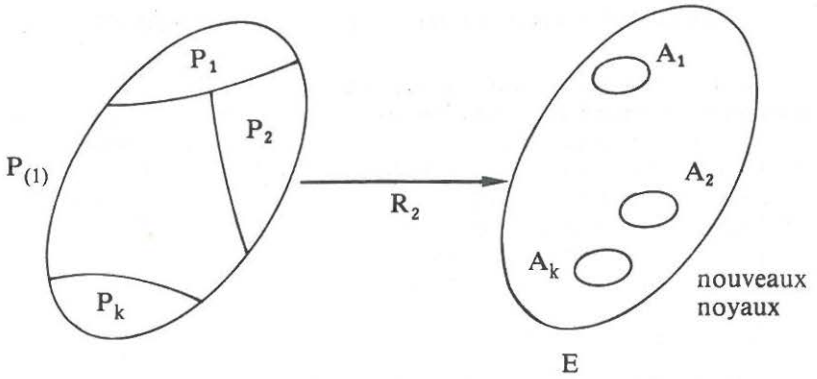
- une famille $(A_i)_{1 \leqslant i \leqslant k}$ de parties de E (les noyaux) de cardinaux (*) $(n_i)_{1 \leqslant i \leqslant k}$ respectivement et tels que $\sum_{i=1}^{k} n_i$ soit bien plus petit que Card E.
- une règle R_1 qui permet de déterminer, pour chaque A_i, l'ensemble des éléments de E qui sont plus proches (pour la distance d) de lui que des autres noyaux.



(*) En pratique, souvent : $n_i = n$; $1 \leqslant i \leqslant k$

On a ainsi construit une partition $P_{(1)}$ de E en k classes (*). Une fonction critère W, à valeurs réelles, fournit un indice de qualité de la partition, soit $W(P_{(1)})$ (c'est un indice de dispersion moyenne des classes autour de leur noyau).

• Une deuxième règle R_2 permet de construire k nouveaux noyaux (éléments les plus représentatifs de chacune des k classes de $P_{(1)}$ dans un sens à préciser) de cardinaux $(n_i)_{1 \leqslant i \leqslant k}$. Chaque classe possède alors son noyau.



A ces nouveaux noyaux on applique à nouveau R_1 et à la nouvelle partition de E ainsi construite on applique R_2, etc. ; à chaque étape on calcule $W(P_{(j)})$. La fonction W est, sous les hypothèses régissant R_1 et R_2, **strictement décroissante** (c'est-à-dire qu'à chaque étape on améliore la qualité de la partition), et Card E $< \infty$, donc la suite $W(P_{(j)})$ converge en un nombre **fini** d'étapes (l'algorithme converge). On dispose alors :

— d'une partition de E en k classes ;

— d'éléments centraux de ces classes (les noyaux).

3. Un des intérêts essentiels de la méthode est la notion de noyau :

— l'interprétation des résultats est facilitée : on peut dans certains cas se contenter d'examiner les noyaux de la partition finale (résumé des classes ; reconnaissance de forme) ;

— les noyaux initiaux peuvent être, soit donnés par la connaissance que l'on a des objets étudiés, soit tirés au hasard. Il est évident que la partition finale et ses noyaux dépendent du choix des noyaux initiaux. A l'aide de plusieurs tirages au hasard on peut obtenir plusieurs partitions ; l'intersection de ces partitions (**formes fortes**) fournira dans certains cas des parties stables d'une classification.

(*) En fait $P_{(1)}$ est une partition en $k' \leqslant k$ classes (DIDAY 1972, b).

1. PARTITIONNEMENT :

Il peut se trouver des cas d'espèce où l'on soit effectivement intéressé par l'obtention d'une hiérarchie ou d'un arbre ; en fait, c'est généralement la détection de classes bien significatives qui est intéressante. Pour obtenir ces classes à partir d'une hiérarchie ou d'un arbre, il faut utiliser des seuils de découpage dont la justification est difficile. De plus, ces méthodes se heurtent à deux difficultés :

— pour construire une hiérarchie par les méthodes classiques, il faut mettre en mémoire le tableau des distances de taille $\frac{\text{card E (card E} - 1)}{2}$, d'où la difficulté de traiter des tableaux de plus de 300 individus ;

— "l'effet de chaîne" [1] est souvent gênant pour ces méthodes.

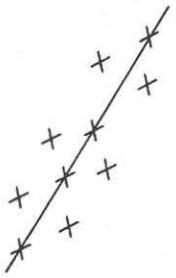
Les techniques de partitionnement dont nous allons maintenant parler ne sont pas sujettes à ces défauts. Leur but est de fournir une **partition** en k classes **bien séparées** d'objets bien agrèges (k donné à priori). Elles présentent l'intérêt d'être rapides et de permettre le traitement de très grands tableaux (on a classifié, par exemple, 40 000 personnes ayant répondu à une enquête pour une entreprise de vente par correspondance, afin d'obtenir des profils types de clientèle). Comment fonctionnent ces méthodes ? Décrivons à titre d'exemple un algorithme de type "Nuées dynamiques".

2. UN ALGORITHME DE TYPE "NUÉES DYNAMIQUES"

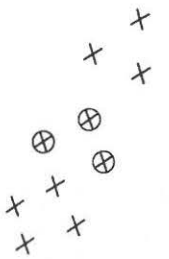
Cet algorithme nécessite tout d'abord la définition d'un mode de représentation symbolique de tout groupe d'objets. Un groupe d'objets étant donné, cette représentation symbolique appelée "noyau" peut être par exemple (voir Figure 1) :

— Une droite
— Un groupe de points de la population
— Un centre de gravité.

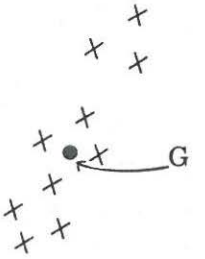
Figure 1



Le noyau est une droite



Le noyau est un groupe de points pris dans la population

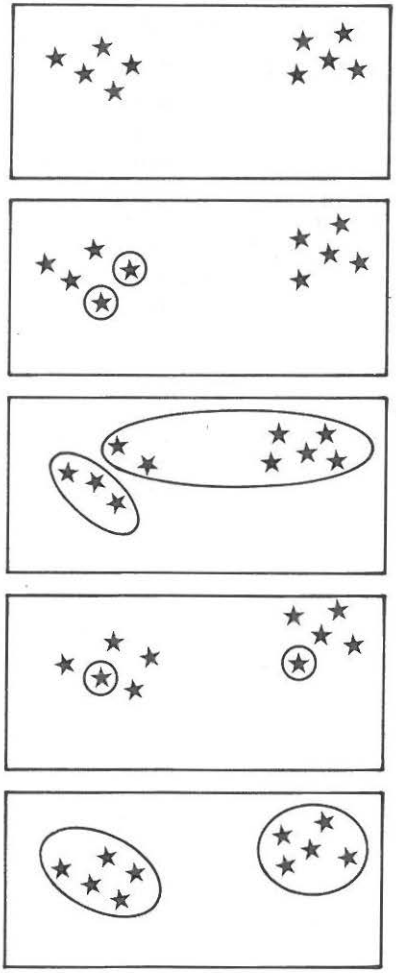


Le noyau est un centre de gravité

(1) Voir LETOURNEUX, 4.

Comment se déroule l'algorithme ?

On part de la donnée de k noyaux, ces noyaux étant choisis ou tirés au hasard ; chaque élément de la population est ensuite affecté au noyau dont il est le plus proche. On obtient ainsi une partition en k classes dont on calcule les noyaux. On recommence le procédé avec les nouveaux noyaux et ainsi de suite (voir Figure 2). On démontre que sous certaines conditions l'algorithme converge vers une position stable, en améliorant à chaque itération une fonction critère.



Trouver une bonne partition en deux classes.

Deux points A et B sont tirés au hasard et appelés *noyaux*.

On associe chaque point au noyau le plus proche.

Deux nouveaux noyaux C et D sont calculés en prenant le point le plus proche du centre de chaque classe de l'étape précédente.

On associe chaque point au noyau le plus proche : les classes naturelles ont donc été détectées par un procédé automatique.

Figure 2

3. *UN EXEMPLE DE CRITÈRE A OPTIMISER*

Plaçons-nous dans le cas où les noyaux sont des groupes de points de même cardinal n , pris dans la population et considérons le critère suivant :

$$W(L,P) = \sum_{i=1}^{k} D(A_i, P_i)$$

où $L = (A_1, \ldots, A_k)$, A_i est le $i^{ième}$ noyau contenant n objets de la population. $P = (P_1, \ldots, P_k)$ est une partition en k classes des objets.

$D(A_i, P_i)$ mesure la "dissemblance" du noyau A_i à la classe P_i. Le problème consiste donc à chercher L* et P* qui minimisent W.

Soit d(x,y) une mesure de dissemblance (une distance par exemple) entre couples d'objets.

Faisons l'hypothèse suivante :

$D(X,Y) = \sum_{x \in X} \sum_{y \in Y} d(x,y)$; $X,Y \subset E$ où E est l'ensemble des objets dont on cherche à construire une partition.

On peut maintenant préciser l'algorithme des nuées dynamiques ; les noyaux $L = (A_1, \ldots, A_k)$ étant donnés, la partition $P = (P_1, \ldots, P_k)$ qui s'en déduit est définie comme suit :

$$P_i = \{ x \in E / \forall j \ D(A_i, x) \leqslant D(A_j, x) \} ;$$

en cas d'égalité on affecte x à la classe du plus petit indice.

La partition P ainsi déduite de L est notée $P = f(L)$.

La partition P étant donnée, on construit les noyaux $L = (A_1, \ldots, A_k)$ comme suit et par abus de langage :
$A_i = \{ x \in E / x$ fait partie de l'ensemble des n éléments qui rendent $D(x, P_i)$ le plus petit possible $\}$.

On note $L = g(P)$.

On peut maintenant énoncer le résultat suivant :

Proposition :

Sous l'hypothèse $D(X,Y) = \sum_{x \in X} \sum_{y \in Y} d(x,y)$, l'algorithme des nuées dynamiques fait décroître W à chaque itération.

Démonstration :

Il suffit de montrer successivement que $W(L,P) \geqslant W(L,f(L))$ quelle que soit la partition P de E ; puis que $W(L,P) \geqslant W(g(P),P)$ quels que soient les noyaux définis par $L = (A_1, \ldots, A_k)$.

En effet, si cela est vrai on aura forcément $W(L,P) \geqslant W(L,f(L)) \geqslant W(g(f(L)),f(L))$ et ainsi de suite.

Montrons donc que $W(L,P) \geqslant W(L,f(L))$;

soit $f(L) = Q = (Q_1, \ldots, Q_k)$,

on a :

$$W(L,P) = \sum_{i=1}^{k} \sum_{x \in P_i} d(A_i, x), \quad W(L,Q) = \sum_{i=1}^{k} \sum_{y \in Q_i} d(A_i, y).$$

Prenons un objet z quelconque dans E ; s'il appartient à une classe de même indice dans les partitions P et Q, sa contribution dans les deux sommes W(L,P) et W(L,Q) est la même ; par contre s'il apparaît sous la forme $d(A_i, z)$ dans W(L,P) et $d(A_j, z)$ dans W(L,Q), cela signifie
que $d(A_j, z) \leqslant d(A_i, z)$ par construction même de Q. On peut faire le même raisonnement pour tous les éléments de E et donc $W(L,P) \geqslant W(L,Q)$.

Reste à montrer que $W(L,P) \geqslant W(g(P),P)$, autrement dit que

$$\sum_{i=1}^{k} \sum_{x \in A_i} d(x,P_i) \geqslant \sum_{i=1}^{k} \sum_{y \in B_i} d(y,P_i) \text{ où } g(P) = (B_1, \ldots, B_k) .$$

Ce résultat est immédiat car, par construction même de g(P), on a $\sum_{x \in A_i} d(x,P_i) \geqslant \sum_{y \in B_i} d(y,P_i)$ puisque A_i et B_i ont le même nombre d'éléments et B_i est formé des n éléments z qui minimisent $d(z,P_i)$.

C.Q.F.D.

4. AUTRES ALGORITHMES DE PARTITIONNEMENT ET DE TYPOLOGIE

Signalons deux autres approches couramment utilisées : "l'algorithme d'échange" de Regnier consiste à définir un critère à optimiser et à changer les objets de classes tant qu'il s'améliore. Les "méthodes à seuils" consistent à chercher les lieux à forte densité en utilisant un seuil à priori et en construisant autour de chaque objet des classes ayant pour rayon ce seuil ; on retient ensuite les sphères les plus denses.

On voit que cette dernière méthode donne des groupes types et non une partition ; c'est ce que l'on peut appeler une typologie.

5. STABILITÉ ET FORMES FORTES

Les algorithmes de types "Nuées dynamiques" ou "Algorithme d'échange" conduisent à des optimum locaux du critère et la solution obtenue peut différer suivant la partition ou les noyaux choisis au

départ de l'algorithme. Ayant obtenu plusieurs solutions à partir de différents tirages, on peut se poser le problème de la recherche de classes stables ; autrement dit de groupes d'objets qui restent groupés quel que soit le tirage de départ. D'où la méthode des "**formes fortes**" qui est souvent utilisée dans la pratique.

On construit le tableau dit des formes fortes de la façon suivante :

a) On retient les m meilleurs tirages (m = 5 ou 10, par exemple), autrement dit les tirages dont les solutions obtenues à la convergence de l'algorithme donnent les plus faibles valeurs à la fonction critère W.

b) A chaque ligne du tableau correspond un objet de E et à chaque colonne un tirage. Dans la case (i,j) on met le numéro de la classe dans laquelle est apparu l'objet i pour le tirage j .

c) Par un algorithme simple on regroupe ensemble des objets dont les lignes sont identiques ; autrement dit ceux qui, quel que soit le tirage, restent dans la même classe. Ces groupes s'appellent "formes fortes". On appelle **formes faibles** les groupes d'objets qui, dans un tirage au moins, sortent dans la même classe.

N.B. En fait, la partition de E en formes fortes n'est autre que la partition intersection des partitions obtenues au cours des différents tirages. De même, la partition en formes faibles n'est autre que la plus fine des partitions qui sont moins fines que celles obtenues au cours des différents tirages.

6. EXEMPLE

Le tableau 1 constitue un exemple de données où les 75 objets sont représentés par des points dans le plan et où l'on cherche 6 classes, la distance d étant la distance euclidienne usuelle dans le plan vectoriel.

Ces résultats obtenus après cinq tirages au hasard par la méthode des nuées dynamiques sont schématisés (figure 4). Le tableau des formes fortes est donné ensuite (figure 5).

Le lecteur remarquera qu'une partition en 6 classes est obtenue au cinquième tirage des noyaux, alors que des partitions en 5 classes sont obtenues aux quatre premiers tirages des noyaux ; en fait l'algorithme peut fournir des "classes vides".

TABLEAU I

Numéro des objets	*1ère coordonnée*	*2ème coordonnée*	*Numéro des objets*	*1ère coordonnée*	*2ème coordonnée*
1	45.	17.	39	127.	62.
2	53.	18.	40	130.	64.
3	50.	23.	41	103.	63.
4	65.	23.	42	113.	68.
5	41.	26.	43	116.	72.
6	57.	26.	44	109.	90.
7	74.	26.	45	94.	88.
8	63.	28.	46	78.	88.
9	51.	30.	47	77.	82.
10	54.	31.	48	78.	79.
11	62.	34.	49	100.	98.
12	61.	38.	50	94.	99.
13	63.	39.	51	102.	100.
14	48.	35.	52	97.	100.
15	46.	37.	53	105.	101.
16	49.	38.	54	93.	102.
17	43.	40.	55	90.	107.
18	49.	41.	56	92.	109.
19	51.	42.	57	93.	108.
20	59.	46.	58	103.	111.
21	125.	42.	59	95.	113.
22	126.	45.	60	93.	115.
23	127.	48.	61	3.	77.
24	129.	47.	62	8.	78.
25	127.	51.	63	9.	62.
26	124.	53.	64	11.	65.
27	122.	51.	65	11.	71.
28	120.	50.	66	11.	79.
29	120.	45.	67	13.	69.
30	118.	47.	68	14.	83.
31	131.	56.	69	15.	68.
32	124.	56.	70	15.	72.
33	119.	55.	71	17.	71.
34	118.	57.	72	19.	66.
35	122.	57.	73	20.	78.
36	124.	58.	74	23.	75.
37	118.	61.	75	23.	68.
38	120.	63.			

Tableau des données illustré par les
Figures 4 et 5

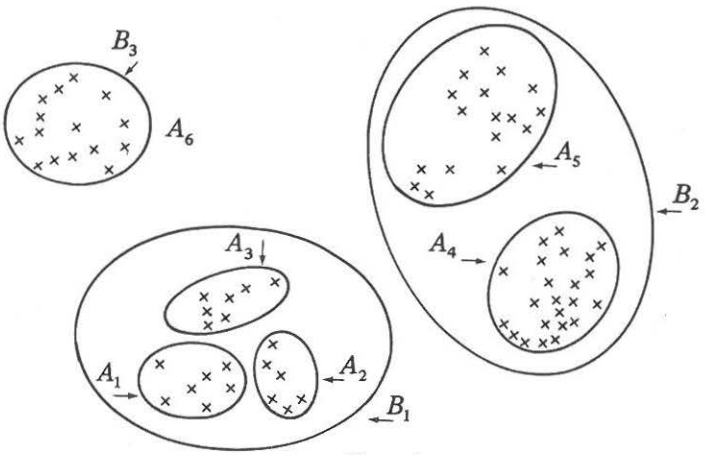


Figure 4

N° des objets	N° des solutions obtenues						
	1	2	3	4	5		
1	1	3	3	1	4		
2	1	3	3	1	4		
3	1	3	3	1	4		
5	1	3	3	1	4	A_1	
6	1	3	3	1	4		
9	1	3	3	1	4		
10	1	3	3	1	4		
8	1	3	3	1	6		
4	1	3	3	1	6		
7	1	3	3	1	6		
11	1	3	3	1	6	A_2	B_1
12	1	3	3	1	6		
13	1	3	3	1	6		
14	1	3	3	1	5		
15	1	3	3	1	5		
16	1	3	3	1	5		
17	1	3	3	1	5	A_3	
18	1	3	3	1	5		
19	1	3	3	1	5		
20	1	3	3	1	5		
21	2	2	1	2	2		
22	2	2	1	2	2		
23	2	2	1	2	2		
24	2	2	1	2	2		
25	2	2	1	2	2		
26	2	2	1	2	2		
27	2	2	1	2	2		
28	2	2	1	2	2		
29	2	2	1	2	2		
30	2	2	1	2	2	A_4	B_2
31	2	2	1	2	2		
32	2	2	1	2	2		
33	2	2	1	2	2		
34	2	2	1	2	2		
35	2	2	1	2	2		
36	2	2	1	2	2		
37	2	2	1	2	2		
38	2	2	1	2	2		
39	2	2	1	2	2		
40	2	2	1	2	2		
41	2	2	1	2	2	A_4	
42	2	2	1	2	2		
43	2	2	1	2	2		
44	3	4	1	4	1		
45	3	4	1	4	1		
46	3	4	1	4	1		
47	3	4	1	4	1		
48	3	4	1	4	1		B_2
49	3	4	1	4	1		
50	3	4	1	4	1		
51	3	4	1	4	1		
52	3	4	1	4	1	A_5	
53	3	4	1	4	1		
54	3	4	1	4	1		
55	3	4	1	4	1		
56	3	4	1	4	1		
57	3	4	1	4	1		
58	3	4	1	4	1		
59	3	4	1	4	1		
60	3	4	1	4	1		
61	4	1	2	3	3		
62	4	1	2	3	3		
63	4	1	2	3	3		
64	4	1	2	3	3		
65	4	1	2	3	3		
66	4	1	2	3	3		
67	4	1	2	3	3		
68	4	1	2	3	3	A_6	B_3
69	4	1	2	3	3		
70	4	1	2	3	3		
71	4	1	2	3	3		
72	4	1	2	3	3		
73	4	1	2	3	3		
74	4	1	2	3	3		
75	4	1	2	3	3		

Figure 5

7. *PARTITION OPTIMALE D'UNE POPULATION CARACTERISEE PAR* UNE SEULE *VARIABLE QUANTITATIVE : L'ALGORITHME DE FISHER*

Contrairement à l'algorithme des nuées dynamiques, l'algorithme de Fischer permet d'obtenir une partition optimale ; cependant il est basé sur la propriété de contiguité et donc il ne peut s'appliquer que dans le cas où la population est caractérisée par une seule variable.

Il s'agit de minimiser le critère :

$$W(P) = \sum_{i=1}^{k} \text{card } P_i \text{ var } P_i = \sum_{i=1}^{k} \sum_{x \in P_i} (x - G(P_i))^2$$

où P est une partition $P = (P_1, \ldots, P_k)$ de l'ensemble des objets et $G(P_i)$ est le centre de gravité de la classe P_i.

L'algorithme est basé sur les propriétés suivantes :

1°) La solution optimale P^* minimisant W est formée de classes contiguës au sens de la variable y.

2°) Si une partition $P^* = (P_1^*, \ldots, P_k^*)$ est optimale, alors $P_2^*, \ldots, P_k^*$ est une partition optimale à $k-1$ classes de $E \setminus P_1^*$ (l'ensemble des objets moins ceux de la classe P_1^*).

L'algorithme se déroule de la façon suivante :

Etape 1 : Enumération complète pour trouver la partition en deux classes de E.

Etapes suivantes : Soit $\{ y_1, \ldots, y_n \}$ l'ensemble des valeurs prises par la variable y sur l'ensemble des objets E. Comment se fait le passage de l'étape $k-1$ à l'étape k ?

On a enregistré à l'étape $k-1$ (à partir de $k=3$) l'ensemble des partitions optimales à $k-2$ classes des valeurs $[y_i, \ldots, y_m]$ pour $i = 1,2,\ldots, n-k+1$.

A l'étape k, on fixe y_{i_1} et on se sert de l'étape précédente pour calculer la partition optimale à $k-1$ classes en faisant varier un indice i_2 : à chaque valeur de $i_1 + i_2$ on peut faire correspondre la partition optimale à $k-2$ classes de $\{y_{i_1+i_2}, \ldots, y_n\}$. Parmi toutes les valeurs possibles de i_2 (i_2 varie de 1 à $n-k+1-i_1$) on retient celle qui donne la meilleure partition à $k-1$ classes de $\{y_{i_1}, \ldots, y_n\}$. Pour chaque valeur de i_1, on enregistre la partition ainsi obtenue (cela servira à l'étape suivante). On retient enfin parmi toutes les valeurs possibles de i_1 (i_1 varie de 1 à $n-k+1$) celle qui donne les meilleures partitions à k classes.

CONCLUSION

La méthode et ses divers prolongements ont été appliqués dans de nombreux domaines : sciences médicales, météorologie, géographie, économie, géologie, etc...

Il existe d'autres méthodes non hiérarchiques et construites antérieurement, suivant des principes voisins, mais où les noyaux sont remplacés par les centres de gravité des classes ; on pourra se reporter à FORGY [7], ou MAC QUEEN [8], ou BENZECRI [B.G.1].

BIBLIOGRAPHIE

(Cf. aussi [1], t1, de la biblio. générale)

[1] **DIDAY E.** Une nouvelle méthode en classification automatique et reconnaissance des formes : La méthode des nuées dynamiques, *R.S.A.* vol. XIX, n° 2, 1971, 19-34

[2] **DIDAY E.** Thèse d'Etat : Nouvelle méthode et nouveaux concepts en classification automatique et reconnaissance des formes, *Univ. P.VI.* 1970.

[3] **DIDAY E.** Optimisation en classification automatique et reconnaissances des formes. Note Scient. *IRIA* n° 6 et *RIRO* 1972, 61-95.

[4] **DIDAY E.** Classification automatique sequentielle pour grands tableaux *RIRO* 1975, 1-29

[5] **FISHER W.D.** Clustering and aggregation in economics.

[6] **SANDOR G., DIDAY E., LECHEVALLIER, BARRE** Une étude informatique des corrélations entre les modifications des protéines sériques en pathologie humaine, *C.R.A.S.*, Paris, t. 294, 1972, pp. 464-467

[7] **FORGY E.W.** Cluster Analysis of Multivariate Data : Efficiency versus interpretability of classifications, in Biometrics, vol. 21, 1965, n°3.

[8] **MAC QUEEN J.B.** Some Methods for Classification and Analysis of Multivariate Observations, 5th Berkeley Symposium on Mathematical Statistic and Probability, 1967, vol. 1 n° 1.

[9] **REGNIER S.** Sur quelques aspects mathématiques des problèmes de classification automatique, *I.C.C. Bulletin*, vol. 4, 1965, pp. 175-191.

Instituteurs, Professeurs du Premier Cycle ...
Pouvez-vous ignorer plus longtemps le phénomène CALCULATRICES ?

Regardez autour de vous ; ouvrez les yeux dans votre classe ...
La calculatrice est là, elle fait partie de l'univers quotidien de l'enfant.
Une récente circulaire ministérielle en recommande l'usage tout au long de la scolarité, l'impose en 4e et 3e, l'autorise dès cette année aux examens.

Vous qui devez enseigner l'art du calcul,
n'attendez plus !

lisez la nouvelle brochure A.P.M.E.P. n° 31

CALCULATRICES
QUATRE OPERATIONS
(ELEMENTAIRE ET PREMIER CYCLE)

Un ouvrage écrit pour vous, par des collègues, à un prix modique.
Un ouvrage qui apporte des réponses aux questions que vous vous posez ... et qui entend faire le point sur ce sujet d'actualité :

1 **Les calculatrices dans la classe** : Pour ou contre ?
Ce qu'en pensent enseignants, parents, élèves ... en France ou à l'étranger.

2 **Calculatrices et pédagogie** : les caractéristiques des calculatrices actuelles ; leur fonctionnement ; leurs possibilités mais aussi leurs limites, leurs contraintes ... et leurs dangers.
Quelques approches possibles de la machine dans la classe.

3 **Calculatrices et mathématique** : des calculatrices pour quoi faire ? du C.P. à la 3e, des thèmes à exploiter, des comptes rendus de travaux avec les élèves, des idées à développer, des objectifs simples mais précis...

4 **Calculatrices, autres disciplines ... et vie quotidienne.**

5 **Calculatrices et informatique** : outil à calculer bien sûr, la calculatrice permet aussi et surtout d'**expliciter les algorithmes** de nombreux savoir-faire ;
une façon intelligente, vivante et dynamique d'apprendre à calculer.

6 Et pour finir, une bibliographie fournie, et de bonnes adresses ...

176 pages — 15 F (port compris : **19 F**)

I.C. LERMAN
Université de Rennes

6. ANALYSE ORDINALE D'UNE CLASSE D'ECHELLES OU ANALYSE HIERARCHIQUE

I. INTRODUCTION

L'objet de cet article consiste à exposer dans ses grandes lignes et à illustrer sur un exemple réel notre méthode d'analyse hiérarchique.

Cette méthode a pris naissance en 1966, elle a été programmée puis appliquée à différents jeux de données réelles (cf. [8], [9], [10], [11 chap. 8] et [4])

C'est **improprement** que le terme d'analyse hiérarchique a pu désigner une analyse des données résultant d'une hiérarchie de classifications ; il s'agit, comme le titre l'indique, d'une analyse ordinale d'une classe d'échelles. Chaque échelle d'attitude qui représente un caractère dont l'ensemble des modalités est totalement ordonné, que nous appellerons également ci-dessous "item total", définit un préordre total sur l'ensemble défini par la population étudiée. Relativement à une même classe d'échelles recouvrant une même "dimension" sous-jacente, le problème de l'analyse hiérarchique unidimensionnelle consiste, à partir du comportement de la population, à ordonner totalement l'ensemble de toutes les modalités non minimales des diverses variables, formant ainsi une échelle fine d'attitude.

Le Psycho-Sociologue posait ce problème à propos d'une classe d'échelles établie a priori, pour laquelle il postulait de l'existence d'une même "variable sous-jacente" aux différentes échelles.

Il est moins contestable de poser le problème de l'analyse hiérarchique unidimensionnelle, relativement à une classe d'échelles reconnue comme définissant une forte tendance comportementale de la population étudiée.

Nous découvrons une telle classe comme sous-tendue par un nœud "significatif" de notre arbre des classifications qui organise par proximité en une structure en classes et sous-classes un vaste ensemble d'items observés sur un échantillon de la population concernée. Par conséquent, dans la pratique, notre méthode de classification hiérarchique intervient de façon préalable à une analyse hiérarchique proprement dite. C'est pour cette raison et au risque de déplacer le centre d'intérêt de l'article que nous chercherons à l'exprimer schématiquement avant l'application sur des données réelles. La méthode de classification hiérarchique est d'ailleurs illustrée par R. Gras **7** ; elle est également illustrée dans [12] où il s'agit d'un exemple réel de pédagogie mathématique. Dans ces deux cas les variables de description peuvent être assimilées à des attributs ; alors qu'il s'agit dans notre problème de la situation où les variables descriptives sont des échelles d'attitude. I. Cohen (cf. [4]) a précisément considéré ce cas dans le cadre d'une vaste enquête de Psycho-Pédagogie de l'enfant, élaborée par le Docteur A. Berge et sa collaboratrice G. Denjean (cf. [2]). Nous reprendrons ci-dessous un aspect des résultats qui concerne la méthode présentée ici.

Les spécialistes de l'analyse hiérarchique se sont longtemps et surtout préoccupés du cas des items dichotomiques (i.e. échelles à deux modalités). Ce travail se situe d'emblée dans le cas d'échelles à nombre quelconque de modalités : chaque échelle étant représentée par un ordre total sur l'ensemble des modalités, le support de l'information est un produit direct d'ordres totaux finis. Un même sommet de ce treillis sera appelé, lorsque nous emprunterons le langage des praticiens de ces méthodes, "patron de réponse" ; patron étant la traduction adaptée du mot anglais "pattern".

La possibilité d'erreur dans la réponse d'un sujet à un item pose le problème du choix de la "bonne" échelle hiérarchique permettant au "mieux" de ranger totalement l'ensemble des divers comportements élémentaires proposés à travers les différents items. Une première partie de notre travail aboutira à la proposition d'un algorithme ayant un caractère optimal, qui, en un "petit" nombre de pas, détermine l'échelle hiérarchique qui s'ajuste le "mieux" à l'ensemble des patrons de réponse observés. Cet algorithme repose naturellement sur une certaine idée de la "distance" entre deux patrons de réponse et entre un patron de réponse et une échelle hiérarchique ; cette idée de la distance se traduisant mathématiquement au niveau du treillis de représentation. Nous avons proposé trois distances et adapté l'algorithme à deux d'entre elles ; l'adaption à la distance que nous désignons par d_3 (la plus classique) a été effectuée par M. Barbut (cf. [1]). Nous montrons ainsi, par rapport à l'état de la question tel que rapporté dans l'ouvrage de B. Matalon (cf. [13]), comment remplacer le tâtonnement par la recherche systématique et comment distinguer un degré d'erreur dans la réponse d'un sujet à un item par le choix judicieux, tenant compte du problème particulier traité, de l'une des trois distances.

II. Définitions et représentation mathématique

1. DEFINITIONS GENERALES

Nous allons présenter ici, en le précisant, le vocabulaire de l'analyse hiérarchique.

On appelle "item" la donnée d'un ensemble fini de comportements proposés. Une modalité de l'item est un élément de l'ensemble ; c'est-à-dire un comportement proposé. On distingue habituellement les items dichotomiques des items non dichotomiques. Un item est dichotomique, si le nombre de ses modalités est deux, et ne l'est pas si le nombre de ses modalités est strictement supérieur à deux. L'item est dit **présentant k modalités** si le nombre de ses modalités est k.

Nous supposons qu'est définie sur l'item, c'est-à-dire sur l'ensemble des comportements proposés, une relation d'ordre ; cette relation d'ordre est en général une relation de préférence par rapport à un but donné de caractère psychologique social ou autre. Si l'item est totalement ordonné par la relation d'ordre définie, il sera dit **total**.

Nous nous intéressons seulement aux items totaux ; nous ne restreignons pas ainsi la généralité de notre étude parce qu'on peut toujours se ramener à ce cas en réalisant une partition de l'item non total en sous items totaux. Notons $(a_0, a_1, a_2, \ldots, a_{k-1})$ la suite des modalités de l'item total a, où on a

$$a_0 < a_1 < \ldots < a_j < \ldots < a_{k-1},$$

la modalité a_j est dite avoir le "code" j.

Exemples :

Dans le cadre de l'enquête de Psycho-Pédagogie à laquelle nous faisions allusion ci-dessus, la question : "L'enfant a-t-il eu des vomissements motivés par des causes extérieures à l'alimentation : contrariété, émotion,..." définit un item dichotomique dont les deux modalités sont :

$$a_0 = \text{non et } a_1 = \text{oui}.$$

D'autre part, la question : "L'enfant parle-t-il de l'école à la maison ?"

		Code
Oui, avec plaisir	☐	0
Oui, incidemment	☐	1
Oui, avec hostilité	☐	2
Non	☐	3

définit un item total à quatre modalités.

Pour cette question qui vise l'affection qu'a un enfant pour son école, une réponse "Non" est jugée par le Psychologue la plus défavorable, car elle reflète un aspect de démission de l'enfant vis-à-vis de son école.

2. *VARIABLE SOUS JACENTE A UN ITEM TOTAL*

Soit un item total a présentant k modalités : $a_0 < a_1 < \ldots < a_{k-1}$. On admet l'existence d'une échelle dense (*) de valeurs α que nous représenterons par un demi-axe orienté de gauche à droite. La variable α est liée à l'item a de la façon suivante : Un sujet donné répond à l'item a par la modalité a_j si et seulement si la valeur de la variable, mesurée sur le sujet, est comprise entre deux bornes α_j et α_{j+1}, **indépendantes du sujet** ; $\alpha_j \leqslant \alpha < \alpha_{j+1}$.

On pose $\alpha_0 = 0$ et $\alpha_k = +\infty$; d'où

$$0 = \alpha_0 < \alpha_1 < \ldots < \alpha_j < \ldots < \alpha_k = \infty$$

L'item a définit ainsi sur l'axe de α une subdivision $(\alpha_0, \alpha_1, \ldots, \alpha_{k-1}, \alpha_k)$

La variable α est dite sous-jacente à l'item a.

Bien qu'elles nous soient inconnues, les bornes α_j sont supposées fixées par la donnée de l'item. Par la suite nous donnerons un rapide aperçu de la manière de déterminer statistiquement les nombres α_j dans le cadre d'un modèle probabiliste qui suppose *continue* l'échelle α.

3. *ENSEMBLE D'ITEMS TOTAUX SE REFERANT A UNE MEME VARIABLE*

3.1. Introduction

Pour fixer les idées considérons un ensemble de deux items totaux a et b présentant respectivement h et k modalités :

$a_0 < a_1 < \ldots < a_i < \ldots < a_{(h-1)}$ et $b_0 < b_1 < \ldots < b_j < \ldots < b_{(k-1)}$.

Désignons a priori par α la variable sous-jacente à l'item a et par β, celle sous-jacente à l'item b. On a, avec des notations que l'on comprend à partir du paragraphe précédent,

$$0 = \alpha_0 < \alpha_1 < \ldots < \alpha_i < \ldots < \alpha_{h-1} < \alpha_h = \infty \quad \text{et}$$
$$0 = \beta_0 < \beta_1 < \ldots < \beta_j < \ldots < \beta_{k-1} < \beta_k = \infty$$

(*) Une échelle est dite ici dense dans le sens que Cantor donnait à ce mot pour les types d'ordres : entre deux éléments distincts quelconques x et y de l'échelle, on peut toujours en insérer un troisième z (et par suite une infinité).
Exemples d'échelles denses : l'ensemble ordonné Q des rationnels, l'ensemble ordonné D des nombres décimaux.

Définition

Les deux items a et b sont dits relatifs à une même variable sous-jacente si on a pu définir une relation d'ordre total R sur la réunion des deux items, compatible avec les relations d'ordre définies respectivement sur chacun des items et pour laquelle : a_i R $b_j \iff \alpha_i \leqslant \beta_j$.

Pratiquement, la relation R est une relation de préférence vis-à-vis d'un but donné ; celui pour lequel ont précisément été établis chacun des deux items a et b.

Chacun des deux items définit, sur le demi-axe de la variable commune que nous noterons alors γ, une subdivision. Celle définie par l'item a est $(0, \alpha_1, \alpha_2, \ldots, \alpha_{h-1}, \infty)$ et celle définie par b est $(0, \beta_1, \beta_2, \ldots, \beta_{k-1}, \infty)$.

L'item a sera dit plus *fin* que l'item b si la subdivision définie par a est plus fine que celle définie par b ; en d'autres termes si pour tout j, il existe i tel que $\alpha_i = \beta_j$.

3.2. Composition d'items relatifs à une même variable sous-jacente

Soient deux items a et b ; reprenons ceux que nous venons de considérer ci-dessus. Désignons par $(0, \gamma_1, \gamma_2, \ldots, \gamma_{h+k-2}, \infty)$ la suite ordonnée croissante des valeurs α_i et β_j (voir figure 1).

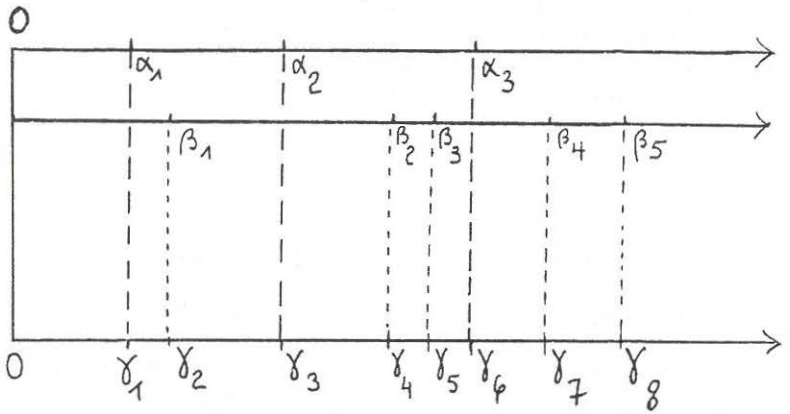


Figure 1

Au couple d'items (a,b) faisons correspondre l'item c = a∗b défini comme suit : si $\gamma_\ell = \alpha_i$ (resp. β_j), la modalité de code ℓ de l'item c est a_i (resp. b_j). L'item c est dit **composé** des items a et b. La loi de composition ainsi définie entre items totaux se référant à une même variable sous-jacente est associative et commutative.

En n'admettant pas d'erreur possible de la part d'un individu donné, la condition nécessaire et suffisante pour que cet individu réponde à l'item c par la modalité c_ℓ est que la valeur de la variable sous-jacente γ

mesurée sur le sujet soit comprise entre γ_ℓ et $\gamma_{\ell+1}$. Dans le cas où on a pour cette valeur mesurée $\alpha_i \leqslant \gamma < \alpha_{i+1}$ et $\beta_j \leqslant \gamma < \beta_{j+1}$, alors $\gamma_\ell = \sup(\alpha_i, \beta_j)$ et $\gamma_{\ell+1} = \inf(\alpha_{i+1}, \beta_{j+1})$ car, les deux intervalles $[\alpha_i, \alpha_{i+1}[$ et $[\beta_j, \beta_{j+1}[$ ayant une intersection non vide, il n'existe pas de points α_ℓ ou β_m entre $\sup(\alpha_i, \beta_j)$ et $\inf(\alpha_{i+1}, \beta_{j+1})$.

Considérons le cas particulier important de la composition d'items dichotomiques. Soient $b^1, b^2, \ldots, b^{m-1}$, $(m-1)$ items dichotomiques se référant à une même variable sous-jacente. L'item $c = b^1 * b^2 * \ldots * b^{(m-1)}$ est un item total à m modalités en général ; la subdivision définie par l'item b_j étant $(0, \beta_j, \infty)$, celle définie par c est $(0, \beta_1, \beta_2, \ldots, \beta_j, \ldots, \beta_{m-1}, \infty)$

Exemple :

$b^1, b^2, \ldots, b^{m-1}$ peuvent être définis à partir de $(m-1)$ questions d'un problème de mathématique s'échelonnant par difficulté croissante : la j-ème question fait intervenir toute l'aptitude requise pour la solution de la question $(j-1)$. La modalité codée 0 de l'item b_j est une réponse incorrecte à la question n° j. L'item $c = b^1 * b^2 * \ldots * b^{(m-1)}$ contient la suite totalement ordonnée des modalités suivantes :

	Code
réponse incorrecte à la question n° 1	0
réponse correcte à la question n° 1 et incorrecte à la question n° 2	1
réponse correcte à la question n° 2 et incorrecte à la question n° 3	2
.	
.	
.	
.	
.	
réponse correcte à la question $(m-1)$	$(m-1)$

Réciproquement, on peut aisément définir une décomposition d'un item total présentant m modalités en $(m-1)$ items dichotomiques totaux. En effet, à partir de l'item a présentant m modalités totalement ordonnées : $a_0 < a_1 < \ldots < a_j < \ldots < a_{m-1}$, on définit la suite $(b^1, b^2, \ldots, b^j, \ldots, b^{m-1})$ des items dichotomiques où l'item b^j consiste en la partition en deux classes R_j et S_j de l'ensemble des comportements proposés par l'item a, où $R_j = \{a_0, a_1, \ldots, a_{j-1}\}$ et $S_j = \{a_j, a_{j+1}, \ldots, a_{m-1}\}$. La modalité de code 0 (resp. 1) pour l'item dichotomique b_j correspond à une réponse à l'item a par un comportement appartenant à R_j (resp. S_j). Si $(0, \alpha_1, \ldots, \alpha_i, \ldots, \alpha_{m-1}, \infty)$ est la subdivision définie par l'item a, celle définie par l'item b_j est $(0, \alpha_j, \infty)$.

Nous allons abandonner maintenant cette approche avant de la reprendre plus richement mais très brièvement (cf. § III.3) dans le cadre d'une hypothèse générale de la distribution de la variable sous-jacente

sur la population étudiée et d'un modèle probabiliste cohérent de l'erreur dans la réponse d'un sujet à un item. Cette construction nous permettra alors de déterminer statistiquement les points de la subdivision sur l'axe de la variable sous-jacente à l'item.

4. TREILLIS DE REPRESENTATION

En vue de l'analyse du comportement d'une population donnée de sujets relativement à un but fixé, on suppose établi un ensemble fini d'items totaux

$$\{a^j / 1 \leqslant j \leqslant m\}.$$

Chaque a^j est un **ensemble fini** totalement ordonné ; à $\{a^j / 1 \leqslant j \leqslant m\}$, associons le produit fini d'ensembles finis totalement ordonnés

$$\prod_{1 \leqslant j \leqslant m} a^j = a^1 \times a^2 \times \ldots \times a^m, \qquad (1)$$

qui sera l'ensemble de représentation.

A a^j associons l'ensemble $\Omega_j = \{0,1,2,\ldots;r_j\}$ des codes de ses modalités ($\mathrm{card}(a^j) = r_j + 1$) que nous représentons géométriquement par l'intervalle commençant $[0, r_j]$ de **N** dont nous ne retiendrons que la structure ordinale.

A l'ensemble $\prod_{1 \leqslant j \leqslant m} a^j$ se trouve associé l'ensemble produit

$$\Omega = \prod_{1 \leqslant j \leqslant m} \Omega_j = \Omega_1 \times \Omega_2 \times \ldots \times \Omega_m , \qquad (2)$$

dont les éléments sont les points

$$\omega = (\omega_1, \omega_2, \ldots, \omega_j, \ldots, \omega_m)$$

où ω_j est le code d'une modalité de l'item a^j ; $\omega_j \in \Omega_j$, $0 \leqslant \omega_j \leqslant r_j$.

Ω sera dans ces conditions représenté géométriquement par le pavé de $\mathbf{N}^m$: $[0, r_1] \times [0, r_2] \times \ldots \times [0, r_m]$.

Sur Ω, on définit la relation d'ordre produit $\omega \leqslant \omega'$:

$$(\omega_1,\ldots,\omega_j,\ldots,\omega_m) \leqslant (\omega'_1,\omega'_2,\ldots,\omega'_j,\ldots,\omega'_m) \Leftrightarrow (\forall_j),\ \omega_j \leqslant \omega'_j \qquad (3)$$

ω' **succède** à ω si $\omega \leqslant \omega'$ et si, pour tout j , $\omega_j = \omega'_j$, sauf pour un seul j_0 où $\omega'_{j_0} = \omega_{j_0} + 1$.

Les deux lois de composition interne $\vee$ et $\wedge$ (supremum et infimum)

$$\omega \vee \omega' = \sup(\omega, \omega') \quad \text{et} \quad \omega \wedge \omega' = \inf(\omega, \omega') ,$$

confèrent à Ω une structure de treillis distributif dont le minorant universel est le point $(0,0,\ldots,0)$ et le majorant universel, le point $(r_1,r_2,\ldots,r_m)$.

Une **chaîne** du treillis est une suite de somments $(\omega^{(1)},\omega^{(2)},\ldots,\omega^{(k)})$ de Ω totalement ordonnée par la relation d'ordre (3) :

$$\omega^{(1)} \leqslant \omega^{(2)} \leqslant \ldots \leqslant \omega^{(k)} .$$

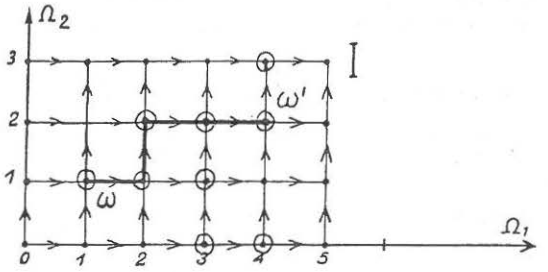
La chaîne est **maximale** si pour tout i , $1 \leqslant i \leqslant (k-1)$, $\omega^{(i+1)}$ succède à $\omega^{(i)}$; **la longueur** d'une chaîne maximale ayant k sommets est $(k-1)$. Toutes les chaînes maximales reliant deux sommets fixés de Ω ont même longueur. La longueur commune de toutes les chaînes maximales joignant le point origine $(0,0,\ldots,0)$ au sommet ω sera appelée (pour nous conformer notamment au vocabulaire des praticiens de l'Analyse hiérarchique) "score" de ω. La fonction "score" **gradue** le treillis.

On peut naturellement associer à Ω muni de la relation d'ordre (3) un graphe, au sens de Cl. Berge (cf. [3]), $G = (\Omega, U)$ dont l'ensemble des sommets est Ω et l'ensemble des arcs U ; du point ω sont issus les arcs qui le relient à l'ensemble $\Gamma(\omega)$ de ses successeurs :

$$\Gamma(\omega) = \{(\omega_1,\ldots,\omega_{(j-1)},\omega_j+1,\omega_{(j+1)},\ldots,\omega_m),\ \omega_j < r_j \ /\ 1 \leqslant j \leqslant m\} \qquad (4)$$

où on a noté $\omega = (\omega_1, \omega_2, \ldots, \omega_m)$.

Dans notre représentation géométrique, nous matérialiserons l'arc joignant $(\omega_1,\ldots,\omega_j,\ldots,\omega_m)$ à $(\omega_1,\ldots,\omega_{(j-1)}, \omega_j+1, \omega_{(j+1)},\ldots, \omega_m)$, par le vecteur v_j.



Représentation géométrique du graphe associé à $a^1 \times a^2$ où a_1 et a_2 sont des items totaux présentant respectivement 6 et 4 modalités.

Figure 2

La notion de chaîne maximale dans le treillis est celle de **chemin** dans le graphe. Les deux sommets O et I de scores 0 et $r_1 + r_2 + \ldots + r_m$ seront appelés extrémités du graphe et tout chemin les reliant sera dit **extrémal**.

Bien que le langage des treillis suffise, nous utiliserons également le vocabulaire, parfois plus souple, des graphes ; le contexte sera assez clair pour savoir de quoi nous parlons.

Chaque "patron de réponse" $\omega = (\omega_1,\ldots,\omega_m)$ étant représenté par le sommet correspondant du treillis, la représentation de la population de sujets étudiée est une mesure entière positive sur Ω : $\{n_\omega \ /\ \omega \in \Omega\}$ où n_ω est le nombre d'individus de la population dont le patron de réponse est ω. Un sommet ω du treillis est dit "patron de réponse **observé**" si la mesure n_ω qui lui est affectée est différente de 0.

Une **échelle relative à l'ensemble des items totaux** $\{a^j \ /\ 1 \leqslant j \leqslant m\}$ est représentée par une chaîne du treillis ; il s'agit d'un ensemble de sommets que peut recouvrir un chemin du graphe $G = (\Omega, U)$.

On dit que "le comportement d'une population donnée de sujets est **unidimensionnel** par rapport à l'ensemble des items totaux $\{a^j / 1 \leqslant j \leqslant m\}$" si l'ensemble des patrons de réponse **observés** forme une échelle.

On montre aisément qu'une condition suffisante d'unidimensionnalité du comportement de **toute** population est que l'ensemble des items $\{a^j/1 \leqslant j \leqslant m\}$ se refère à une même variable sous-jacente (cf. §3 ci-dessus). Désignons par $(\alpha^j) = (0 = \alpha_0^j, \alpha_1^j, \ldots, \alpha_{r_j}^j)$ la subdivision sur l'axe de la variable sous-jacente définie par l'item a^j, $1 \leqslant j \leqslant m$; et soit $(0 = \gamma_0, \gamma_1, \ldots, \gamma_\ell, \ldots, \gamma_{r_1 + r_2 + \ldots + r_m})$ la subdivision associée à la composition $c = a^1 * a^2 * \ldots * a^m$ des différents items a^j. En considérant qu'un même point ne peut appartenir à plus d'une subdivision (α^j), considérons la chaîne maximale reliant les extrémités du treillis $(\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(\ell)}, \ldots, \omega^{(r_1 + \ldots + r_m)})$, définie de la façon suivante : $\omega^{(0)}$ est le sommet dont toutes les coordonnées sont nulles ; si le ℓ-ème point γ_ℓ de la subdivision définie par c est un point provenant de la subdivision définie par le j-ème item a^j, alors le ℓ-ème sommet $\omega^{(\ell)}$ de la chaîne se déduit du précédent $\omega^{(\ell-1)}$ en ajoutant 1 à la j-ème composante :

$$\omega_h^{(\ell)} = \omega_h^{(\ell-1)} \text{ (resp. } = \omega_h^{(\ell-1)} + 1) \text{ si } h \neq j \text{ (resp. si } h = j).$$

Cette chaîne sera dite l'"échelle **définie** par la suite des items totaux $\{a^j/1 \leqslant j \leqslant m\}$", laquelle est sensée supporter toute la mesure définie sur Ω lorsqu'on soumet à une population donnée le questionnaire formé par la suite de ces items.

Nous avons pris soin de définir l'unidimensionnalité du comportement d'une population donnée par rapport à un ensemble d'items indépendamment de l'hypothèse d'une même variable sous-jacente. En effet, par rapport à un même ensemble d'items, une population donnée peut se comporter de façon unidimensionnelle alors que ce ne sera pas le cas pour une autre, peut-être d'ailleurs plus nombreuse. Inversement, relativement à un ensemble d'items se référant à une même variable (on dit également «formant une échelle»), l'ensemble des patrons de réponse observés d'une population donnée peut, en raison de réponses erronées, ne pas former une échelle ; d'où le problème de l'analyse hiérarchique unidimensionnelle.

Le support de la mesure $\{n_\omega / \omega \in \Omega\}$ sur l'ensemble Ω des sommets du treillis, c'est-à-dire $\{\omega \in \Omega / n_\omega \neq 0\}$, est une partie d'un produit fini d'ordres totaux finis ; il s'agit d'un ordre partiel sur un ensemble fini. Or tout ensemble fini partiellement ordonné Λ peut apparaître comme une partie d'un produit fini d'ensembles finis totalement ordonnés. Relativement à Λ on définit, fondamentalement, deux nombres entiers. Le nombre δ, **dimension de l'ordre**, est le nombre minimum d'ordres totaux dont Λ peut être considéré comme une partie de leur produit ; **le nombre de stabilité** β qui est le nombre minimum de chaînes totalement ordonnées pouvant recouvrir Λ.

Ainsi, en se référant à l'exemple fourni par la figure 2 précédente, considérons l'ensemble Λ des sommets marqués $\odot$:

$$\Lambda = \{(0,3), (0,4), (1,1), (1,2), (1,3), (2,2), (2,3), (2,4), (3,4)\}$$

muni de l'ordre produit défini par (3) ci-dessus. Pour cet ensemble ordonné, la dimension est 2 et le nombre de stabilité est 3.

La notion de multidimensionnalité en Analyse hiérarchique recouvre la seconde notion et non la première.

Le comportement de la population étudiée par rapport à l'ensemble des items est **k-dimensionnel** si le nombre minimum d'**échelles** recouvrant l'ensemble $\{\omega \in \Omega / n_\omega \neq 0\}$ des patrons de réponse observés est k.

Il peut y avoir plus d'un système de k **chaînes** (échelles) recouvrant le support de la mesure ; le problème se pose de reconnaître celui des systèmes dont chaque élément représente une échelle d'attitude pour une partie de la population étudiée. Nous proposerons un algorithme qui tient compte de la présence éventuelle de patrons de réponse erronés.

III. Analyse hiérarchique unidimensionnelle

1. POSITION DU PROBLEME

La question principale que se pose l'Analyse hiérarchique est la suivante : "Peut-on admettre que les items totaux a^j, $1 \leqslant j \leqslant m$, se réfèrent à une même variable ?" Or nous avons défini l'unidimensionnalité d'un ensemble d'items en supposant qu'il n'existe pas de possibilité d'erreur dans le choix par le sujet de l'une des modalités d'un item (cf. § II.2.). On se rend compte du caractère déterministe d'une telle hypothèse qui ne tient pas compte de la présence éventuelle de patrons de réponse « erronés ». Il y a lieu de remplacer cette hypothèse par une autre, à caractère probabiliste, dans le cadre de laquelle la probabilité d'un patron erroné est petite sans être exactement nulle.

Dans le cas d'une réponse positive à la question posée, l'ensemble des patrons observés non erronés est une échelle, qu'il faut alors déterminer. Pour l'estimation de l'échelle nous proposerons deux approches. La première, qui suppose la définition d'une métrique sur Ω, est algorithmique, elle consiste à déterminer celle des échelles qui "s'ajuste le mieux" à l'ensemble pondéré des patrons de réponse observés ; en d'autres termes, à « distance » minimum d'une mesure supportée par une partie du treillis de représentation. Il peut exister plus d'une solution ; d'autre part, une telle estimation est indépendante de l'existence d'une même variable sous-jacente à l'ensemble des items.

La seconde approche consiste à proposer un modèle probabiliste de l'erreur dans le choix par un sujet d'une modalité d'un item. Un tel modèle, dans l'hypothèse d'une certaine distribution (qu'on peut se fixer sans restreindre la généralité) de la variable sous-jacente sur la

population étudiée, permet d'une part de déterminer l'échelle, d'autre part, de définir les probabilités des différents patrons de réponse possibles. D'où la possibilité de tester le modèle.

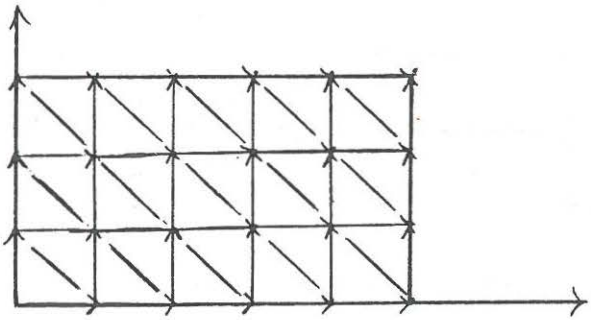
2. ALGORITHME DE CONSTRUCTION D'UNE ECHELLE

2.1. Distances adoptées sur Ω

En vue de la définition de l'"**écart" d'un patron de réponse observé à une échelle**, nous allons introduire trois distances sur Ω en utilisant le graphe G = (Ω, U) de représentation. Considérons pour cela les deux graphes suivants liés à G = (Ω, U).

1) Le graphe G', dit symétrisé de G, obtenu en adjoignant à tout arc de G, l'arc inverse ; les deux sommets extrémités sont alors reliés par une **arête.**

2) Le graphe $\overline{G}$, dit complété de G, obtenu à partir de G en reliant d'une arête tout couple de sommets ayant un même prédécesseur, c'est-à-dire tels que $\omega = (\omega_1,\ldots,\omega_m)$, $\omega' = (\omega'_1,\ldots,\omega'_m)$ où $\omega_j = \omega'_j$ pour tout j sauf pour deux indices h et h' pour lesquels $\omega_h = \omega'_h + 1$ et $\omega_{h'} = \omega'_{h'} - 1$.



Graphe G associé à un couple d'items présentant respectivement 6 et 4 modalités.

Figure 3

Nous adaptons sur Ω l'une des distances

a) d_1 ; $d_1(\omega,\omega') = 1$ (resp. $= 0$) si $\omega' \neq \omega$ (resp. si $\omega' = \omega$)

b) d_2 ; $d_2(\omega,\omega')$ est la **longueur** du plus court chemin de $\overline{G}$ joignant ω à ω'.

c) d_3 ; $d_3(\omega,\omega')$ est la **longueur** du plus court chemin de G' joignant ω à ω'.

On rappelle que la longueur d'un chemin est le nombre d'arcs qu'il contient.

2.2. "Ecart" d'un patron de réponse à une échelle

Ayant adopté, à partir de considérations concrètes, l'une des trois distances d_1, d_2 ou d_3, la manière qui nous semble la plus « naturelle » pour définir l'écart d'un patron de réponse observé représenté par un

sommet ω du graphe à une échelle représentée par un chemin $\mathcal{E}$, est de prendre la distance au sens topologique du sommet ω au sous-ensemble de points $\mathcal{E}$; soit

$$e_i(\omega, \mathcal{E}) = \min_{\bar{\omega} \in \mathcal{E}} d_i(\omega, \bar{\omega}), \qquad (1)$$

où i = 1, 2 ou 3.

Toutefois, pour pouvoir adapter l'algorithme, que nous présenterons bientôt, au cas de la distance d_3, il est nécessaire (cf. [1]) de définir la distance de ω à $\mathcal{E}$ comme étant la somme pondérée des distances d_3 des différents sommets de $\mathcal{E}$ à ω ; soit

$$e'_3(\omega, \mathcal{E}) = \sum_{\bar{\omega} \in \mathcal{E}} n_{\bar{\omega}} \, d_3(\omega, \bar{\omega}) \qquad (2)$$

Exemple :

Considérons le cas du système représentatif $a^1 \times a^2$, où a^1 et a^2 présentent respectivement 6 et 4 modalités, et l'échelle $\mathcal{E}$ dont les points sont :

(0,0), (1,0), (2,0), (2,1), (3,1), (3,2), (4,2), (5,2) et (5,3).

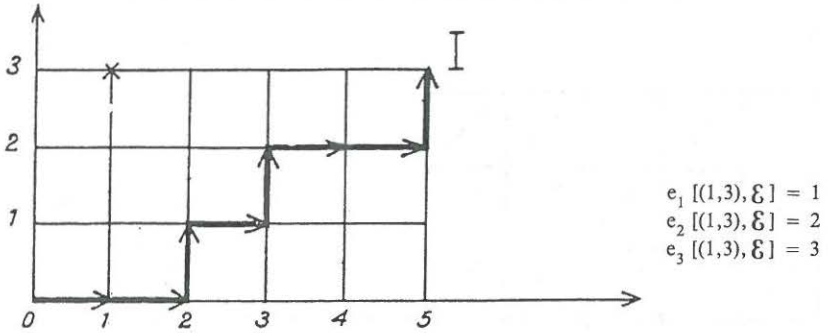


Figure 4

Supposons qu'à partir des réponses de la population étudiée, la suite des pondérations affectée à la suite des sommets de l'échelle est

7, 5, 12, 11, 23, 15, 10, 8, 2 ; on a

$$e'_3[(1,3), \mathcal{E}] =$$
$$7\times4 + 5\times3 + 12\times4 + 11\times3 + 23\times4 + 15\times3 + 10\times4 + 8\times5 + 2\times4.$$

Cherchons quelque indication intuitive sur le choix de l'une des distances. Si on admet que les différents patrons de réponse observés erronés ont le même "degré d'erreur", on adoptera la distance d_1.

Si on admet que le "degré d'erreur" du patron de réponse erroné est proportionnel à son « éloignement » du sommet de l'échelle de même score (i.e. situé sur le même niveau du treillis), on adoptera la distance

d_2, laquelle donne lieu à la forme analytique suivante de l'écart du sommet ω à l'échelle $\mathcal{E}$;

$$e_2(\omega, \mathcal{E}) = \frac{1}{m} \sum_{1 \leqslant j \leqslant m} | \omega_j - \overline{\omega}_j |, \qquad (3)$$

où $\overline{\omega} = (\overline{\omega}_1, \ldots, \overline{\omega}_m)$ est le sommet de l'échelle de même score que $\omega = (\omega_1, \ldots, \omega_m)$ qui représente le patron de réponse erroné. Cette expression se réduit, dans le cas particulier de deux items, à $(|\omega_1 - \overline{\omega}_1| + |\omega_2 - \overline{\omega}_2|)$ qu'il est naturel de considérer ; en effet, un sujet donné, normalement situé au point $(\overline{\omega}_1, \overline{\omega}_2)$ de l'échelle, et répondant trop fortement à l'un des items, a tendance pour récupérer, à répondre faiblement à l'autre item.

La distance mathématiquement la plus classique, celle la plus conforme à la structure du treillis de représentation, est la distance d_3, laquelle donne lieu à la forme analytique suivante de l'écart du sommet ω à l'échelle $\mathcal{E}$;

$$e_3(\omega, \mathcal{E}) = \min_{\overline{\omega} \in \mathcal{E}} \sum_{1 \leqslant j \leqslant m} |\omega_j - \overline{\omega}_j|, \qquad (4)$$

Ici, l'erreur dans la réponse d'un sujet à un item est considérée indépendamment, de l'erreur dans la réponse du sujet à un autre item.

Les deux critères d'adéquation d'une échelle obtenus respectivement à partir de l'une des distances d_2 ou d_3 doivent conduire sensiblement au même résultat et ce, d'autant que l'échelle $\mathcal{E}$ s'impose statistiquement.

2.3. Critère d'adéquation d'une échelle

Une échelle $\mathcal{E}$ est d'autant « meilleure » que la somme **pondérée** des écarts des sommets de Ω (i.e. des patrons de réponse) à $\mathcal{E}$ est plus petite.

Ayant adopté l'une des distances d_1, d_2 ou d_3, si on considère la forme (1), la plus naturelle de l'écart d'un patron de réponse observé à une échelle, la quantité critère s'écrit

$$\sum_{\omega \in \Omega} n_\omega \, e_i(\omega, \mathcal{E}) = \sum_{\omega \in \Omega} n_\omega \, \{ \min_{\overline{\omega} \in \mathcal{E}} d_i(\omega, \overline{\omega}\}, \qquad (5)$$

Si on considère la forme (2) de l'écart, seulement adaptée lorsque la distance choisie sur Ω est d_3, la quantité critère se met sous la forme

$$\sum_{\omega \in \Omega} n_\omega \, e'_3(\omega, \mathcal{E}) = \sum_{\omega \in \Omega} \sum_{\overline{\omega} \in \mathcal{E}} n_\omega \, n_{\overline{\omega}} \, d_3(\omega, \overline{\omega}), \qquad (6)$$

La quantité critère représente une « mesure » du désaccord entre les résultats observés et l'hypothèse d'unidimensionnalité. Notons avec intérêt que le critère d'adéquation d'un système de « noyaux » à une partition, utilisé par la suite par E. Diday dans sa méthode des « nuées dynamiques » (cf.**5**) est de même nature que ceux que nous présentons.

Remarquons enfin que pour $i = 1$, (5) représente ce que les praticiens de l'analyse hiérarchique désignent sous le nom de « coefficient de reproductibilité ».

2.4. Algorithme

Il s'agit de déterminer l'échelle, représentée par une chaîne maximale reliant les extrémités du treillis de représentation, autrement dit, par un chemin du graphe G reliant ses extrémités O et I, qui rend minimum le critère adopté. Pour cela, la méthode la plus directe consiste à affecter à chaque chemin de G joignant O et I, la valeur du critère d'adéquation défini et à retenir celui ou ceux des chemins pour lesquels cette valeur est minimum (on peut trouver plus d'une échelle répondant à la question). Dans ces conditions, pour chacun des chemins extrémaux, on aura à calculer la somme pondérée des écarts (5) ou (6) ; or, le nombre total de tels chemins est $(r_1+r_2+\ldots+r_m)\,!/r_1\,!r_2\,!\ldots r_m\,!$. On voit bien que le volume des calculs que nécessite une telle méthode est, en général, particulièrement important, même pour un puissant ordinateur ; d'où le caractère crucial d'un algorithme optimal dont le nombre de pas soit de l'ordre du nombre de sommets du treillis. Cet algorithme, que nous allons exprimer, ne peut s'adapter au cas de la distance d_3 avec la quantité critère (5). On peut dans ces conditions proposer l'algorithme sous-optimal dont la première étape consiste, par une recherche directe, à déterminer l'échelle $b^{(2)}$ relative au couple d'items (a^1, a^2) ; et la i-ème étape, à former l'échelle $b^{(i+1)}$ associée au couple d'échelles $(b^{(i)}, a^{i+1})$; $1<i<m$. L'échelle ainsi obtenue, relativement à la suite des items $(a^1, a^2,\ldots,a^m)$, peut dépendre de l'ordre dans lequel ont été rangés les items ; elle n'en dépend pas et est optimale si sa restriction à tout couple d'échelles $(a^j, a^{j'})$ est l'échelle optimale résumant a^j et $a^{j'}$.

Nous allons à présent exposer l'algorithme qui en un « petit » nombre de pas permet de déterminer les échelles optimales. Cet algorithme revient à construire, à partir d'une pondération positive des sommets du graphe $G = (\Omega, U)$:

$$\{p_\omega / p_\omega \in \mathbf{R}^+,\ \omega \in \Omega\},$$

le ou les chemins extrémaux les moins pesants ($\sum_{\omega \in \mathcal{E}} p_\omega$ minimum).

Pour fixer les idées nous nous plaçons dans le cas où $\Omega = \Omega_1 \times \Omega_2$ est un produit de deux ordres totaux finis (associé à un couple (a,b) d'items). Chaque sommet ω de Ω est défini, dans la représentation géométrique du treillis ou du graphe G, par ses coordonnées ; $\omega = (\alpha,\beta)$. A ω est attaché la masse $p(\alpha,\beta)$: nombre réel positif dépendant de ω et de n_ω.

La première étape de l'algorithme consiste à affecter aux sommets de Ω une nouvelle pondération $\{r_\omega / \omega \in \Omega\}$ où $r(\alpha,\beta)$ sera le poids du chemin le moins pesant reliant l'origine (0,0) au point $\omega = (\alpha,\beta)$. L'algorithme démarre à partir de l'origine où $r(0,0) = p(0,0)$.

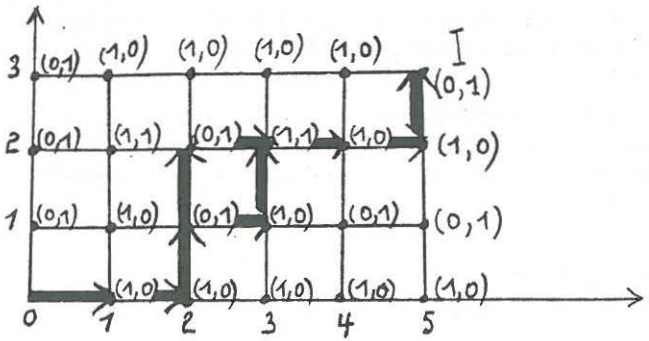


Figure 5

Il s'agit de la représentation d'un couple d'items (a,b) présentant 6 et 4 modalités respectivement. Chaque sommet (α,β) est affecté de son indicateur de précédence $i(\alpha,\beta)$. On trouve dans ce cas deux échelles représentées par deux chaînes qui ont une section commençante (resp. finissante) commune. Ces deux chaînes sont

(0,0), (1,0), (2,0), (2,1), (3,1), (3,2), (4,2), (5,2), (5,3) et
(0,0), (1,0), (2,0), (2,1), (2,2), (3,2), (4,2), (5,2), (5,3).

La première correspond à l'ordre suivant sur l'ensemble des modalités des deux items

$a_0 \sim b_0 < a_1 < a_2 < b_1 < a_3 < b_2 < a_4 < a_5 < b_3$

et la seconde, à

$a_0 \sim b_0 < a_1 < a_2 < b_1 < b_2 < a_3 < a_4 < a_5 < b_3$

Chaque sommet a un ou deux prédécesseurs. Le nouveau poids $r(\alpha,\beta)$ de (α,β) est la somme de $p(\alpha,\beta)$ et du poids ϱ de son prédécesseur le moins lourd ; soit :

$$\begin{aligned} r(\alpha,\beta) &= p(\alpha,\beta) + \min\{r(\alpha-1,\beta), \text{ si } (\alpha,\beta) \text{ a deux prédécesseurs} \\ &= p(\alpha,\beta) + r(\alpha-1,\beta), \text{ si le seul prédecesseur de } (\alpha,\beta) \text{ est } (\alpha-1,\beta) \\ &= p(\alpha,\beta) + r(\alpha,\beta-1), \text{ si le seul prédecesseur de } (\alpha,\beta) \text{ est } (\alpha,\beta-1) \end{aligned}$$

De plus, à (α,β) nous attacherons l'indicateur $i(\alpha,\beta)$ qui est un vecteur logique à deux composantes 0 ou 1 ; la composante égale à 1 indiquant le sommet le moins lourd au sens de r, précédant (α,β) ; soit

$$\begin{aligned} i(\alpha,\beta) &= (1,0) \text{ si } r(\alpha-1,\beta) < r(\alpha,\beta-1) \text{ ou si le seul prédécesseur de } (\alpha,\beta) \text{ est } (\alpha-1,\beta) \\ &= (0,1) \text{ si } r(\alpha-1,\beta) > r(\alpha,\beta-1) \text{ ou si le seul prédecesseur de } (\alpha,\beta) \text{ est } (\alpha,\beta-1) \\ &= (1,1) \text{ si } r(\alpha-1,\beta) = r(\alpha,\beta-1). \end{aligned}$$

De la sorte, le point I, extrémité du treillis ou du graphe, sera affecté du poids du chemin le moins pesant reliant les extrémités du graphe ; d'autre part la suite, à partir de I, des indicateurs permettra, de proche en proche, de retrouver le ou les chemins extrémaux de pondération minimum.

La généralisation de ce qui précède au cas d'un produit de plus de deux ordres totaux ne présente aucune difficulté théorique. Cependant, l'application de l'algorithme devient délicate s'il y a plusieurs échelles optimales.

Nous allons à présent pour chacune des distances d_1, d_2 ou d_3 définir la pondération $\{p_\omega / \omega \in \Omega\}$ de telle sorte que le poids r du chemin extrémal représente la valeur du critère d'adéquation de l'échelle définie par ce chemin.

a) Cas de la distance d_1

Dans ce cas $p_\omega = (n - n_\omega)$ où $n = \sum_{\omega \in \Omega} n_\omega$ est l'effectif de la population étudiée.

En effet ; de la sorte, l'algorithme minimise $\sum_{\omega \in \mathcal{E}} (n - n_\omega)$, où $\mathcal{E}$ est un chemin extrémal définissant une échelle ; cette quantité est égale, à une constante additive près, à $\sum_{\omega \in \Omega} n_\omega \, e_1(\omega, \Omega)$ (cf. § 2.3.).

b) Cas de la distance d_2

Désignons par $s(\omega)$ la longueur de la chaîne maximale d'extrémité ω et d'origine, celle du treillis et soit T(s) le niveau du treillis où la fonction "score" prend la valeur s ; en utilisant la représentation géométrique, on a

$$T(s) = \{\omega \in \Omega \,/\, \omega_1 + \omega_2 + \ldots + \omega_m = s\}$$

où $(\omega_1, \ldots, \omega_m)$ est la représentation du sommet ω.

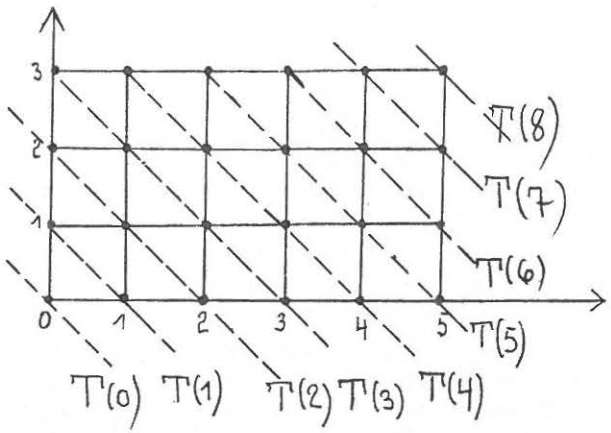


Figure 6

A $\omega \in \Omega$, on affectera le poids

$$p_\omega = \sum_{\varpi \in T(s(\omega))} n_{\varpi} \sum_{1 \leqslant j \leqslant m} |\varpi_j - \omega_j|$$

car de la sorte une échelle optimale minimisera $\sum_{\omega \in \Omega} p_\omega$; c'est-à-dire, $\sum_{\omega \in \Omega} n_\omega e_2(\omega, \mathcal{E})$, (cf. § 2.3.).

c) Cas de la distance d_3

Ici il s'agit d'affecter la pondération suivante des sommets de Ω (cf. [1])

$$p_\omega = \sum_{\varpi \in \Omega} n_{\varpi} d_3(\varpi, \omega)$$

On montre d'ailleurs, moyennant la relation caractérisant une mesure μ

$$\mu(\omega) + \mu(\omega') = \mu(\omega \vee \omega') + \mu(\omega \wedge \omega'),$$

qu'il suffit de calculer la valeur de p sur chacun des axes représentant Ω_i, pour la déterminer en tout point de Ω.

Avec une telle pondération une échelle optimale minimise $\sum_{\omega \in \Omega} n_\omega e'_3(\omega, \mathcal{E})$, (cf. § 2.3.).

Les méthodes, que nous venons d'exposer, de détermination de l'échelle sont essentiellement métriques et combinatoires.

Dans la pratique, il arrive assez souvent, surtout lorsque la tendance comportementale sous-jacente n'est pas très forte, que l'algorithme extraie plus d'une seule échelle optimale. Toutefois, les différentes échelles optimales restent très « voisines », au sens par exemple d'une distance entre ordres totaux finis, définie par le "nombre d'inversions".

Considérons l'ensemble des sommets du treillis qui se trouvent sur la réunion des chaînes définies par les différentes échelles optimales. En saturant cet ensemble au moyen des opération $\wedge$ et $\vee$; on trouve de façon opérationnelle une généralisation au cas d'un ensemble d'items totaux (où le nombre de modalités par item est quelconque et non nécessairement le même d'un item à l'autre) de la notion de **tresse** introduite par Cl. Flament (cf. [6]) dans le cas d'un ensemble d'items dichotomiques. C'est seulement dans ce dernier cas que la caractérisation donnée par Y. Kergall (cf. [7]) de la notion de tresse à partir d'une correspondance de Galois s'avère, à notre avis, pertinente.

3. ECHELLE A PARTIR D'UN MODELE PROBABILISTE

Nous avons associé à un item total une subdivision d'une variable sous-jacente qu'on supposera ici continue (cf. § II.2. dont on reprend ici les notations). On peut dans ces conditions introduire un modèle probabiliste qui tient compte d'une possibilité d'erreur dans le choix d'une

modalité de l'item par le sujet et qui permettra la détermination statistique des points de la subdivision. L'application de ce modèle à chacun des éléments d'une classe d'items totaux se référant à une même variable sous-jacente donnera la subdivision associée à la composition des différents items de la classe (cf. § II.3.2.) et par conséquent, l'échelle hiérarchique ordonnant totalement l'ensemble des modalités non minimales des différents items.

Ce modèle suppose d'une part la donnée de la fonction de répartition $G(\xi)$ de la variable sous-jacente sur la population étudiée et d'autre part, la donnée, pour un individu quelconque dont la valeur de la variable sous-jacente est ξ, de la probabilité $F_\xi(\alpha_j)$ d'observer une valeur inférieure à α_j ; c'est-à-dire d'avoir une réponse de l'individu par l'une des modalités $a_0, a_1, \ldots, a_{(j-1)}$.

Des considérations intuitives et techniques nous ont conduit à proposer le modèle où $G(\xi)$ et une fonction de répartition eulérienne :

$$G(\xi) = \frac{1}{\Gamma(\lambda)} \int_0^{\xi} e^{-x} x^{\lambda - 1} \, dx$$

et où

$$F_\xi(\alpha_j) = \exp[-k\xi / \alpha_j]$$

avec $k = \mathrm{Log}_e 2$.

λ est un paramètre d'échelle qu'on peut se fixer une fois pour toutes. Quant au paramètre k, on peut le déterminer par la condition "naturelle" suivante : "la probabilité d'observer sur un sujet dont la valeur de la variable sous-jacente est $\xi = \alpha_j$, une valeur inférieure à α_j est égale à 1/2".

Ce modèle permet la détermination en fonction des α_j $(1 \leqslant j \leqslant r)$ de la fréquence théorique des différentes modalités de l'item a ; ce qui permet d'identifier les abscisses α_j des points de la subdivision.

L'approche esquissée ici montre une voie par laquelle peut se faire le raccord entre analyse hiérarchique au sens de L. Guttmann et analyse de la structure latente au sens de P. Lazarfeld (voir dans [13] et dans [16]). Toutefois, nous n'allons pas nous y attarder car nous n'avons pas eu à appliquer cette technique dans les données dont il sera bientôt question.

Terminons ce paragraphe en signalant que l'algorithme du paragraphe 2 précédent peut être utilisé dans le cas de l'analyse hiérarchique multidimensionnelle (cf. [11], chap. 8) à laquelle nous avons fait allusion à la fin du paragraphe II.4. Une telle utilisation fait appel à des propriétés fondamentales de la théorie des ordres finis (cf. [5], [14], [15] et [17]).

IV. Présentation schématique de la méthode de classification hiérarchique

Comme nous le signalions dans l'introduction, dans la pratique, nous ne nous permettons de poser le problème de l'analyse hiérarchique unidimensionnelle que par rapport à une classe d'échelles reconnue de bonne cohésion dans une classification hiérarchique sur un ensemble plus vaste d'échelles pouvant recouvrir la dimension recherchée. On ne s'étonnera donc pas de voir ci-dessous (cf. § V.) les schémas d'organisation des classes d'échelles qu'on analysera. Ces derniers sont extraits de l'arbre des classifications portant sur l'ensemble des items du questionnaire. L'application des algorithmes d'analyse hiérarchique permet d'une part, une analyse plus complète de la classe d'échelles et surtout (c'est dans cette étude l'objectif central), de fournir au spécialiste une échelle ordinale de "mesure" fine.

Nous allons par conséquent rappeler ici très brièvement les grandes lignes de notre méthode de classification hiérarchique qui est basée sur une notion générale de vraisemblance du lien par rapport à une hypothèse d'absence de relation statistique.

Face à un tableau de données variables descriptives × Individus ou Objets, la méthode permet aussi bien une classification de l'ensemble des variables V que de l'ensemble E des objets et ceci quelle que soit la nature des variables. Si la classification des variables permet la découverte des principales tendances de comportement de la population étudiée, à travers un échantillon de cette dernière, la classification de E permet de découvrir une typologie de l'ensemble des objets conformément aux ressemblances perçues à travers V.

Si L est l'ensemble à classifier (L est soit l'ensemble V, soit l'ensemble E), le point de départ de la méthode est la définition d'un indice de proximité entre éléments de L. Cet indice qui se réfère à une échelle de probabilité est basé sur la vraisemblance du lien mesuré par un indice « brut » de proximité qui s'impose en général.

Il s'agit dans notre problème du cas de la classification d'un ensemble V de variables descriptives où chaque variable définit un item total (cf. [4]). Si ces différents items sont dichotomiques, en associant à la modalité haute (i.e. codée 1) de chaque item un attribut descriptif, l'ensemble V peut être assimilé à un ensemble A d'attributs de description. C'est la situation envisagée par Gras dans 7 et celle considérée dans [12] par rapport à laquelle nous allons nous exprimer.

Relativement à un couple (a,b) d'attributs, on introduit l'indice brut $s(a, b) = \text{card}(E_a \cap E_b)$ où E_a (resp. E_b) est l'ensemble des sujets possédant a (resp. b). Conformément au principe ci-dessus énoncé, l'indice définitif prend la forme

$$P(a, b) = \text{Prob.}^N \{S < s(a, b)\}, \qquad (1)$$

où S est la variable aléatoire associée à s dans une hypothèse N d'absence de lien, adéquate : qui tient compte des caractéristiques de cardinalité des structures à comparer ; soit, ici $n_a = \text{card}(E_a)$ et $n_b = \text{card}(E_b)$. L'indice P(a, b) s'obtient en fait à partir de la formule

$$P(a, b) = \phi[Q(a, b)], \qquad (2)$$

où ϕ est la fonction de répartition de la loi normale centrée réduite et où Q(a, b) s'obtient en « centrant » et en « réduisant » s par rapport à l'hypothèse N

$$Q(a, b) = \frac{s - \mu_{ab}}{\sigma_{ab}} , \qquad (3)$$

Cependant il y a trois formes « voisines» de l'hypothèse d'absence de lien ayant un caractère symétrique par rapport à a et b, qui conduisent à trois expressions, d'ailleurs proches, de l'indice Q(a, b). Mais à travers les nombreuses expériences menées, l'hypothèse N_1 qui a fourni les résultats les plus cohérents dans leurs nuances conduit à l'expression suivante de l'indice :

$$Q_1(a, b) = \frac{s - \dfrac{n_a n_b}{n}}{\sqrt{\dfrac{n_a n_b}{n}}} \qquad (4)$$

Cette notion de proximité entre deux attributs descriptifs est généralisée dans la méthode à la notion de proximité entre deux variables descriptives de même type (au sens de la structure induite sur E), quel que soit ce dernier.

La notion de proximité entre deux éléments est alors étendue à celle entre deux classes C et D d'éléments où le rôle de l'indice brut de proximité sera joué par

$$p(C, D) = \max \{P(c, d)/(c, d) \in C \times D\} \ ; \qquad (5)$$

celui, définitif, où on se réfère à une hypothèse d'absence de lien adéquate, prend la forme suivante

$$P(C, D) = (p(C, D))^{\ell m} , \qquad (6)$$

où $\ell = \text{card}(C)$ et $m = \text{card}(D)$.

L'"algorithme de la vraisemblance du lien" (A.V.L.) établit à partir de cette proximité entre classes, l'arbre détaillé des classifications où à chaque niveau les paires de classes les plus voisines sont réunies. Cet arbre est en général quasiment binaire et comporte presque autant de niveaux que l'ensemble à classifier a d'éléments.

Une étape décisive de la méthode consiste à condenser l'arbre aux niveaux où se produit un « nœud significatif » et ce, au moyen d'une statistique de proximité, obéissant au principe ci-dessus énoncé, entre

une forme adéquate de l'information quant aux ressemblances des éléments de l'ensemble à classifier et l'association correspondant au nœud. On introduit en fait, relativement à un même niveau i de l'arbre des classifications, deux statistiques : l'une « globale » Σ_i, qui rend compte de l'adéquation de la partition obtenue au niveau i, et l'autre, « locale » τ_i qui « mesure » le degré de signification de l'association qui se produit au niveau i, par rapport à l'ensemble des paires restant séparées à ce niveau. D'ailleurs le comportement de τ_i le long de la suite des niveaux est quasiment le même que celui du taux d'accroissement $\theta_i = (\Sigma_i - \Sigma_{i-1})$ de la statistique globale. Les nœuds significatifs correspondent aux associations qu'accompagnent des maxima locaux de τ_i (resp. θ_i).

Terminons ce point en précisant les expressions de Σ et de τ. Pour se ramener à la comparaison de deux structures de même type dans la recherche d'une mesure d'adéquation entre une partition π (éventuellement produite à un niveau de l'arbre) et l'information quant aux ressemblances de l'ensemble A à classifier, on ne retient de la définition de l'indice de proximité sur A que le préordre total associé sur l'ensemble B des paires d'éléments distincts de A (i.e. des parties à deux éléments de A), appelé préordonnance sur A et défini de la façon suivante :

$$(\forall (p,q) \in B \times B),\ p < q \iff Q(p) < Q(q), \qquad (7)$$

Pour l'indice Q que nous adoptons (cf. formule (3)), la préordonnance associée $\omega(A)$, se réduit le plus souvent dans la pratique à une ordonnance, ordre total sur B. $\omega(A)$ sera représentée par la partie $gr(\omega)$ de $B \times B$ définie par

$$gr(\omega) = \{(p, q) \in B \times B / p < q \text{ et non } q < p \text{ pour } \omega(A)\} \qquad (8)$$

La partition π sur A est regardée comme définissant un préordre total sur B à deux classes $S(\pi)$ et $R(\pi)$ où$S(\pi)$ est l'ensemble des paires séparées et où $R(\pi)$ est celui des paires réunies par la partition π. $S(\pi) < R(\pi)$ pour l'ordre quotient ; ainsi π sera représentée dans $B \times B$ par le rectangle

$$S(\pi) \times R(\pi) \qquad (9)$$

L'indice brut de proximité entre la partition et la préordonnance sera dans ces conditions

$$s(\omega,\pi) = \text{card}\{gr(\omega) \cap (S(\pi) \times R(\pi))\}, \qquad (10)$$

l'indice définitif prend la forme

$$\Sigma = [s(\omega,\pi) - r(\pi) \times s(\pi)/2] \ / \ \sqrt{r(\pi) \times s(\pi)\ (b+1)}/12, \quad (11)$$

où $r(\pi) = \text{card}(R(\pi))$, $s(\pi) = \text{card}(S(\pi))$ et $b = r(\pi) + s(\pi) = \text{card}\ (B)$.

Σ est obtenu en centrant et en réduisant $s(\omega,\pi)$ par rapport à l'hypothèse d'absence de lien où π est un élément aléatoire dans l'ensemble, muni d'une probabilité uniformément répartie, de toutes les partitions d'un même type. Dans le cadre d'une telle hypothèse, nous démontrons que Σ peut être considérée comme une réalisation d'une variable aléatoire $\mathcal{N}(0,1)$.

De la même façon, l'indice brut de proximité qui conduit à τ_i se met sous la forme

$$\sigma_i = \text{card}\{\text{gr}(\omega) \cap (S(\pi) \times R'(\pi_i))\} \qquad (12)$$

où $R'(\pi_i)$ est l'ensemble des paires réunies pour la première fois au niveau i.

Pour mieux contrôler l'interprétation, nous associons à chaque variable son degré de « neutralité » par rapport à une visée classificatoire, qu'on « mesure » par la petitesse de sa variance des proximités aux autres variables selon la formule

$$V(a) = \frac{1}{[\text{card}(A) - 1]} \sum_{c \in A - \{a\}} (Q(a,c) - Q(a))^2, \quad (13)$$

où Q(a) est la moyenne des $\{Q(a,c)/c \in A - \{a\}\}$.

La présence d'attributs neutres dans une classe peut expliquer la difficulté de son interprétation.

V. Application à des données réelles

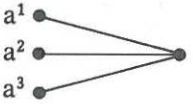
Les données proviennent d'une vaste et très riche enquête sur la Psycho-Pédagogie de l'enfant à laquelle nous avons fait allusion dans l'introduction (cf. [2] et [4]). Le questionnaire comprend une centaine d'items totaux où le nombre de modalités par item varie entre 2 et 10 ; le nombre total de modalités est de 250. Quatre mille parents ont répondu sur différents aspects du développement de leur enfant en relation avec le milieu psychologique où il évolue. Les deux principaux aspects visés sont la "Scolarité" et le "Comportement Alimentaire". Nous nous limiterons ici à l'analyse des tendances associées à l'**utilisation des connaissances.** Rappelons que chacune des tendances nous apparaît sous-tendue par un nœud significatif de l'arbre des classifications organisant par proximité l'ensemble des items et que l'algorithme mis en œuvre pour la recherche de l'échelle hiérarchique est celui du paragraphe III.2.4., avec usage de la distance d_2.

La première classe regroupe les items suivants :

a^1 : *Est-il plutôt*
a_0^1 : consciencieux et méthodique
a_1^1 : normalement appliqué pour son âge
a_2^1 : tendance à bâcler son travail
a^2 : *Dans son travail scolaire, est-il*
a_0^2 : ordonné
a_1^2 : non ordonné
a_2^2 : brouillon

a^3 : *Pour son travail scolaire, est-il*
a_0^3 : méticuleux
a_1^3 : soigneux
a_2^3 : assez soigneux
a_3^3 : peu soigneux

L'organisation de l'ensemble des trois précédents items dans l'arbre des classifications est la suivante



L'échelle hiérarchique qui représente la "concordance entre l'ordre et le soin" est la suivante :

e_0 : consciencieux et méthodique, travail scolaire ordonné, travail scolaire méticuleux.

$e_1 = a_1^3$: travail scolaire soigneux
$e_2 = a_1^1$: normalement appliqué pour son âge
$e_3 = a_2^3$: assez soigneux
$e_4 = a_1^2$: non ordonné
$e_5 = a_2^1$: tendance à bâcler son travail
$e_7 = a_3^3$: peu soigneux
$e_8 = a_2^2$: brouillon.

La deuxième classe regroupe les items suivants :

$\acute{a}^1$: *Votre enfant, à la maison, parle-t-il de l'école ou du lycée ?*
$\acute{a}_0^1$: oui, avec plaisir
$\acute{a}_1^1$: oui, incidemment
$\acute{a}_2^1$: oui, avec hostilité
$\acute{a}_3^1$: non.

$\acute{a}^2$: *Est-il intéressé par son travail scolaire ?*
$\acute{a}_0^2$: oui
$\acute{a}_1^2$: partiellement
$\acute{a}_2^2$: non.

$\acute{a}^3$: *Ses résultats scolaires actuels sont-ils ?*
$\acute{a}_0^3$: bons
$\acute{a}_1^3$: assez bons
$\acute{a}_2^3$: moyens
$\acute{a}_3^3$: mauvais.

$\acute{a}^4$: *Ses résultats ont-ils été ?*

$\acute{a}^4_0$: satisfaisants jusqu'à présent.

$\acute{a}^4_1$: satisfaisants pour certains cycles d'étude seulement

$\acute{a}^4_2$: toujours insuffisants.

$\acute{a}^5$: *Pensez-vous ?*

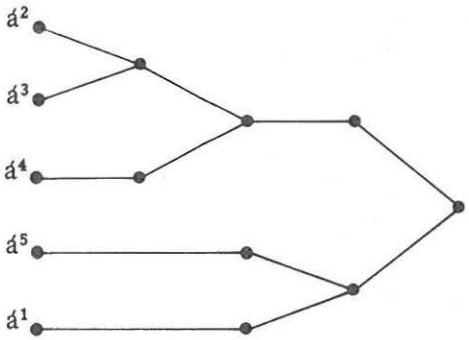
$\acute{a}^5_0$: qu'il sait bien tirer parti de ce qu'il a appris

$\acute{a}^5_1$: qu'il retient ce qu'il apprend

$\acute{a}^5_2$: qu'il apprend mais oublie

$\acute{a}^5_3$: qu'il a une mauvaise mémoire.

L'organisation de l'ensemble des cinq précédents items dans l'arbre des classifications est la suivante :



L'échelle hiérarchique obtenue qui représente "la relation entre la mémoire, le travail et les résultats scolaires" est la suivante :

e'_0 : intéressé par son travail scolaire, résultats toujours satisfaisants, résultats actuels bons, tire parti de ce qu'il apprend, parle de l'école avec plaisir.

$e'_1 = \acute{a}^5_1$: retient ce qu'il apprend

$e'_2 = \acute{a}^3_1$: résultats actuels assez bons

$e'_3 = \acute{a}^4_1$: résultats satisfaisants selon certains cycles

$e'_4 = \acute{a}^3_2$: résultats actuels moyens

$e'_5 = \acute{a}^1_1$: parle de l'école incidemment

$e'_6 = \acute{a}^2_1$: partiellement intéressé par son travail scolaire

$e'_7 = \acute{a}^5_2$: apprend mais oublie

$e'_8 = \acute{a}^4_2$: résultats scolaires toujours insuffisants

$e'_9 = \acute{a}^5_3$: a mauvaise mémoire

$e'_{10} = \acute{a}^3_3$: résultats actuels mauvais
$e'_{11} = \acute{a}^1_2$: parle de l'école avec hostilité
$e'_{12} = \acute{a}^1_3$: ne parle pas de l'école
$e'_{13} = \acute{a}^2_2$: non intéressé par son travail scolaire.

La troisième et dernière sous-classe de variables considérée est formée des items suivants :

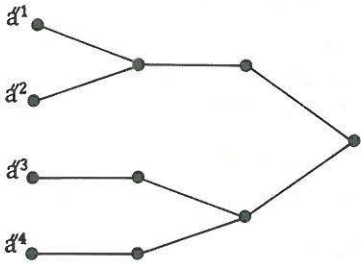
$\acute{a}^1$: *Est-il intéressé par d'autres occupations que le travail scolaire ?*
$\acute{a}^1_0$: oui
$\acute{a}^1_1$: parfois
$\acute{a}^1_2$: non.

$\acute{a}^2$: *Réussit-il mieux dans les travaux nécessitant une initiative personnelle ?*
$\acute{a}^2_0$: oui, $\acute{a}^2_1$: non.

$\acute{a}^3$: *Réussit-il mieux dans les travaux nécessitant une réflexion personnelle ?*
$\acute{a}^3_0$: oui, $\acute{a}^3_1$: non.

$\acute{a}^4$: *A-t-il des réussites dans des activités pratiques hors de l'école ?*
$\acute{a}^4_0$: oui, $\acute{a}^4_1$: non.

L'organisation de l'ensemble des quatre items précédents dans l'arbre des classifications est la suivante :



L'échelle hiérarchique obtenue qui représente "l'intérêt pour les activités extra-scolaires" est la suivante :

e''_0 : intéressé par des activités extra-scolaires, réussit mieux dans les travaux nécessitant une initiative personnelle, réussit mieux dans les travaux nécessitant une réflexion personnelle, a des réussites dans des activités pratiques hors de l'école.

$e''_1 = \acute{a}^3_1$: ne réussit pas mieux dans les travaux nécessitant une réflexion personnelle

$e_2'' = \mathcal{A}_1^2$: ne réussit pas mieux dans les travaux nécessitant une initiative personnelle

$e_3'' = \mathcal{A}_1^4$: n'a pas de réussites dans des activités pratiques hors de l'école

$e_4'' = \mathcal{A}_1^1$: parfois intéressé par d'autres occupations que le travail scolaire

$e_5'' = \mathcal{A}_2^1$: non intéressé par des activités extra-scolaires.

Les différentes échelles extraites ont parfaitement satisfait le spécialiste.

A partir de ces méthodes le pédagogue - mathématicien peut ordonner par complexité croissante une suite de tests mathématiques, dont chacun définit un item total dichotomique, relatifs tous à l'appropriation d'un concept donné. Cet ordre fournit une échelle ordinale fine de « mesure » permettant de situer de façon unidimensionnelle tout sujet examiné.

Bibliographie

(cf. aussi [10] de la biblio. générale)

[1] BARBUT M. - "*Echelles à distances minimum d'une partie donnée d'un treillis distributif fini*" ; dans "*Ordres totaux finis*", (Aix-en-Provence, juillet 67) ; Gauthier-Villars, Mouton ; Paris, 1971.

[2] BERGE A. et DENJEAN G. - "*Comportement digestif et fonctionnement intellectuel*" ; *Revue de Neuropsychiatrie Infantile,* 1974.

[3] BERGE Cl. - "*Graphes et Hypergraphes*", Dunod, Paris, 1970.

[4] COHEN HALLALEH I. - "Classification d'une famille d'échelles au moyen d'un nouvel indice. Comparaison avec le traitement par l'Analyse des correspondances. Application à des données en Psycho-Pédagogie et en Sociologie rurale. *Thèse de 3ème cycle soutenue le 7/02/77* à l'Université Paris VI.

[5] DILWORTH R.P. - "*A decomposition theorem for partially ordered sets*", *Ann. of Math.*, 60, 2, 1954 (359-364).

[6] FLAMENT Cl. - "*Tresses de Gutmann*" ; dans "*Ordres totaux finis*" (Aix-en-Provence, juillet 67) ; Gauthier-Villars, Mouton ; Paris, 1971.

[7] KERGALL Y. - *"Etude des tresses de Gutmann en Algèbre à p valeurs" : Rev. Math. et Sc. Hum.* n° 46, pp. 5-19, Paris, 1974.
[8] LERMAN I.C. - *"Essai sur l'Analyse Hiérarchique"* ; rap. int. Maison des Sciences de l'Homme, Centre de Calcul, juin 1966, Paris.
[9] LERMAN I.C. - *"Essai sur l'Analyse hiérarchique"* ; *Rev. Math. et Sc. Hum.* n° 17, Paris 1967.
[10] LERMAN I.C. et RISO-LEVY Cl. - *"Analyse Hiérarchique"*, rap. int. Maison des Sciences de l'Homme, Centre de Mathématiques Appliquées et de Calcul ; décembre 1969, Paris.
[11] LERMAN I.C. - *"Reconnaissance et Classification des structures finies en Analyse des données"* ; rapport 70, I.R.I.S.A., Université de Rennes ; Rennes, 1977.
[12] LERMAN I.C. - *"Formes d'aptitude et taxinomie d'objectifs cognitifs en mathématiques"*, Revue Française de Pédagogie, n° 44, Paris, 1978.
[13] MATALON B. - *"L'Analyse Hiérarchique"* ; Gauthier-Villars, Paris 1965.
[14] MONJARDET B. - *"Problèmes de transversalité dans les hypergraphes, les ensembles ordonnés et en théorie de la décision collective"* ; Thèse de doctorat ès Sciences Mathématiques, Université Paris VI, juin 1974.
[15] SPERNER E. - *"Ein Satz über Untermengen einer endlichen Menge"*, Math. Z., 27, 1928 (544-548).
[16] TORGERSON W.S. - *"Theory and methods of scaling"*, Wiley, 1958.
[17] TVERBERG - *"On Dilworth's decomposition theorem for partially ordered sets"*, Annals of Math., 51, 1950, (161-166).

BROCHURE A.P.M.E.P.

QUELQUES APPORTS DE L'INFORMATIQUE A L'ENSEIGNEMENT DES MATHEMATIQUES

Cette brochure est essentiellement écrite pour les professeurs de mathématiques

- qui ont entendu parler d'informatique ;
- qui voudraient en savoir plus sans être noyés ;
- qui se demandent ce qui a été fait par d'autres collègues et ce qu'ils pourraient eux-mêmes réaliser ;
- bref ! *tous ceux qui voudraient avoir rapidement mais sérieusement une vue globale sur le sujet.*

En voici les différents chapitres :

* Une introduction fait traditionnellement le point général.

* Le chapitre 1 (RENOUVEAU DE L'ART DU CALCUL) développe quelques exemples simples d'activités montrant que, loin de reléguer au musée le calcul et ses techniques, les calculatrices en font un centre d'attraction particulièrement vivant.

* Le chapitre 2 (QUELQUES DEVELOPPEMENTS EN SITUATIONS PEDAGOGIQUES) montre le parti que le professeur de mathématiques peut tirer, pour sa classe, de la démarche et de l'emploi du matériel informatique.

* Le chapitre 3 (LANGAGES ET METHODES) explicite la nature spécifique du discours informatique et les idées fondamentales qui sous-tendent son utilisation.

* Le chapitre 4 (AIDE DE L'INFORMATIQUE A L'ENSEIGNEMENT) fait le point sur les possibilités actuelles.

* Enfin, le chapitre 5 (INFORMATIONS DIVERSES) donne des indications sur les matériels, les livres, les équipes de recherche, les lycées équipés en ordinateurs, les bonnes adresses ...

280 pages. **25 F** (port compris : **31 F**).

R. GRAS
IREM de Rennes

7. ANALYSE D'UN QUESTIONNAIRE D'ATTITUDES A L'EGARD DES MATHEMATIQUES A L'ENTREE DE 4e EXPERIMENTALE

Cet article reprend l'essentiel d'une publication que nous avons diffusée sous le même titre en juin 1976. L'analyse développée dans cette publication porte sur un questionnaire, donné en annexe 1, sensiblement analogue à celui déjà utilisé lors de deux précédentes enquêtes (cf. bibliographie [3] et [4]). Le traitement des données recueillies à travers ces deux enquêtes utilisait les simples outils de la statistique élémentaire (fréquences, moyennes). Le traitement du dernier questionnaire s'appuie, par contre, sur deux méthodes d'analyse décrites respectivement dans ce tome de notre brochure et dans le tome II à paraître prochainement.

Il s'agit de :

- l'analyse en classification hiérarchique de I.C. Lerman (cf. articles **6** de I.C. Lerman et **4** de J.P. Letourneux de cette brochure, ainsi que [6] et [7]) ;
- et de l'analyse factorielle des correspondances (cf. [2] et l'article de R. Gras du tome II).

1. Objectifs de l'enquête

Elle a tout d'abord pour but de rendre compte de **l'évolution éventuelle de l'attitude** des élèves à l'égard des mathématiques à la fin d'un cycle expérimental (1) de 2 ans. Pour cette raison, le questionnaire est présenté à l'entrée du cycle à des classes dites témoins et à des classes expérimentales de la région de Vannes. C'est son résultat que nous analysons ici. Il devait être présenté à la sortie du cycle, dans les mêmes classes, mais l'imperfection de l'outil de sondage nous a fait différer son

(1) Il s'agit de l'expérience dite O.P.C. (Offre publique de collaboration) portant sur les programmes de mathématiques en vigueur en 1975 dans les classes de quatrième et troisième. Elle vise à rendre l'enseignement moins formel, plus centré sur l'enfant, sur ses propres attentes, sur son environnement et à travers de réelles activités mathématiques.

emploi. Nous songeons plutôt à utiliser un questionnaire, mis au point dans le laboratoire de M. Postic, professeur de Psychologie à l'Université de Haute-Bretagne de Rennes. Ce questionnaire semble permettre de mieux déceler l'attitude profonde des enfants. Pour ce faire, on placera ceux-ci en juges d'affirmations considérées par d'autres juges comme révélatrices de l'intérêt ou du désintérêt à l'égard des mathématiques.

Un deuxième objectif de l'enquête consiste à **faire apparaître une typologie des attitudes** des élèves à l'entrée en quatrième, sur la base des réponses au questionnaire par un échantillon d'enfants de 11 à 14 ans. On vise ainsi une meilleure connaissance des opinions des élèves, de leurs liaisons mutuelles, non seulement au sujet des mathématiques mais aussi sur la place, dans le cursus scolaire, de leurs professeurs, de leur famille, de leurs camarades, etc., sous l'influence éventuelle de certaines variables, comme leur environnement socio-culturel.

Signalons enfin, que des tests de connaissances et de capacités complètent l'évaluation de l'impact expérimental. Ils conduisent à des analyses publiées par l'IREM de Rennes [5] et par l'A.P.M.E.P. [8]. Nous disposons ainsi sur le plan théorique, d'instruments de mesure permettant de cerner, plus précisément que par de simples impressions, l'atteinte ou non des objectifs socio-affectifs et des objectifs cognitifs visés dans le projet expérimental.

2. Données numériques et résultats statistiques globaux

Le questionnaire est présenté, au début du mois d'octobre 1975, par R. Bessière, conseiller d'orientation à Vannes, en dehors de l'heure de mathématiques et, a fortiori, en dehors de la présence de l'enseignant de cette discipline.

2.1. DONNEES

209 élèves consultés sont répartis en huit classes de 3 établissements :

- 4 classes du C.E.S. Jules Simon de Vannes
- 2 classes du C.E.S. St Exupéry de Vannes
- 2 classes du C.E.G. de Questembert

La répartition des variables descriptives est donnée ci-dessous par leur pourcentage moyen. Nous utiliserons souvent les codes suivants :

- A_1, A_2, A_3, A_4 pour les âges respectifs 11, 12, 13 ou 14 ans
- F ou G pour fille ou garçon
- P_0, P_1, ... P_9 pour les catégories socio-professionnelles du chef de famille suivant la classification I.N.S.E.E.(1)
- OR et NR pour redoublant et non redoublant.

La ventilation de ces variables est assez uniforme d'un établissement à l'autre. Cependant, le C.E.S. St-Exupéry et le C.E.G. de Questembert admettent une plus forte densité d'enfants d'agriculteurs.

Caractère	A_1	A_2	A_3	A_4	F	G	OR	NR		
Pourcentage	1,5%	21%	57%	21%	51%	49%	8%	92%		
Caractère	P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
Pourcentage	7%	0%	13%	12%	21%	16%	16%	6%	7%	0%

(1) P_0 : Agriculteurs exploitants
P_1 : Salariés agricoles
P_2 : Patrons de l'industrie et du commerce
P_3 : Professions libérales et cadres supérieurs
P_4 : Cadres moyens
P_5 : Employés
P_6 : Ouvriers
P_7 : Personnel de service
P_8 : Autres catégories (armée, police)
P_9 : Personnes non actives

Par exemple, un instituteur relève de la catégorie P_4. Un professeur certifié appartient à la catégorie P_3.

2.2. RESULTATS STATISTIQUES BRUTS

On trouvera en annexe 1 les pourcentages de réponses aux items du questionnaire. Signalons qu'au sujet :

— des professions souhaitées (notées S_M et S_i, $i = 0$ à 9 selon la catégorisation I.N.S.E.E. des P_i) ou refusées (notées R_M et R_i, $i = 0$ à 9), nous accordons un statut particulier à celle d'enseignant de mathématiques symbolisé suivant le cas respectivement S_M ou R_M;

— des disciplines préférées, nous en retenons huit et distinguons le classement en première ou deuxième place choisi par l'élève ; ce sont les matières suivantes, quelquefois regroupées :

Education Physique
Français - Latin
Langues vivantes (Anglais, Allemand, Italien, Espagnol)
Mathématiques
Sciences Naturelles
Technologie
Travaux manuels - Dessin - Musique

— de la question ouverte n° 3 réduite à 11 et 9 modalités (respectivement positives ou négatives), nous notons sur les listings : GOUT ou REFUS suivi de la modalité explicitant la raison du goût ou du refus.

Nous constatons, malheureusement, que de nombreuses questions restent sans réponse. C'est le cas, en particulier, des questions sur la profession souhaitée ou refusée où nous n'enregistrons respectivement que 62 % et 50 % de prises de position. De plus, nous observons des aberrations aux items (16), (17) et (18). A chaque fois, la modalité RIEN excluait le choix d'autres modalités : certains enfants, assez contradictoirement, obéissant à la consigne générale, lui ont associé deux autres modalités. La faute est imputable au questionnaire, à ses consignes et nous le regrettons.

3. Typologie des attitudes données par la classification hiérarchique.

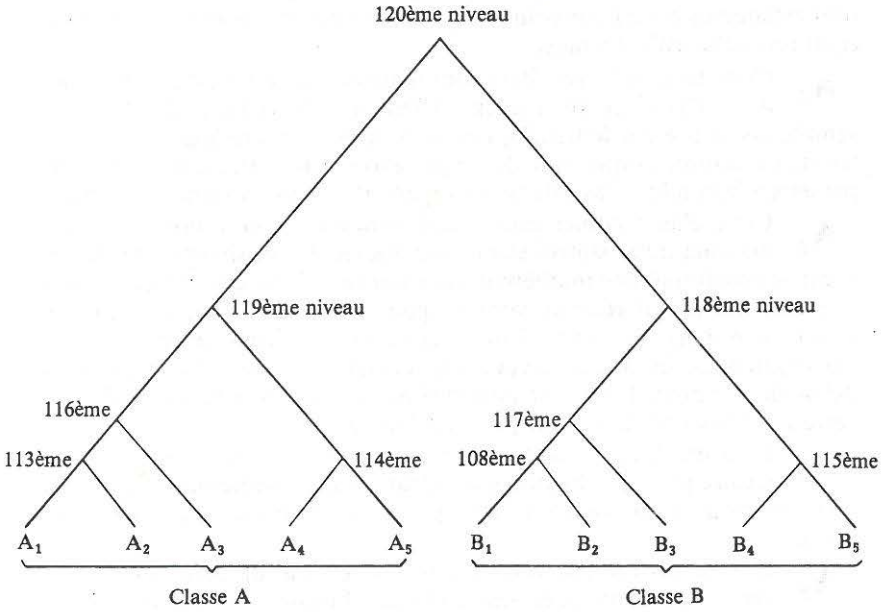
3.1. CONSTRUCTION DE L'ARBRE DE CLASSIFICATION

Pour ce type d'analyse, en raison des capacités (en 1975) du programme, nous ne retenons, dans l'étude, que les items suivants : (3), (4), (7), (8), (10), (11), (12), (14), (15), (16), (17) et (18), le choix étant guidé par la pertinence des réponses pour la recherche entreprise. La question 3 sera assortie des 20 modalités supplémentaires. Des classements en 1, 2 ou 3 des dernières questions, nous ne retenons que les deux premières places.

Les 129 modalités seront traduites en 1 (présence) ou en 0 (absence) pour chaque enfant et traitées par la méthode de I.C. Lerman, c'est-à-dire selon l'algorithme de la vraisemblance du lien.

Nous fournissons en annexe 2, l'arbre condensé de la classification obtenue par la méthode précédente. Cet arbre ne garde de tous les niveaux de formation de classes que ceux où les agrégations sont les plus significatives, c'est-à-dire ceux correspondant à la formation des classes les plus cohérentes.

Après que nous ayons retenu une partition des items en dix classes, cet arbre se présente, à grands traits, ainsi :



Par synthèse des modalités les composant, aux dix classes d'items correspondent dix types d'attitudes auxquels nous donnons une signification. Mais ces types ont peu de chances d'être représentés à l'état pur. Ils ne définissent qu'une typologie des attitudes de la population scolaire, en relation avec l'enseignement des mathématiques. Par conséquent, il serait absurde, dangereux et étranger à notre interprétation, d'utiliser cette typologie pour étiqueter un quelconque de nos élèves. L'attitude d'un enfant est un complexe de certaines de ces attitudes, apparaissant comme composantes plus ou moins intenses, plus ou moins évolutives, de son attitude personnelle.

On pourra se reporter à l'annexe 2 pour suivre, comprendre et contester les interprétations qui vont suivre.

3.2. INTERPRETATION DE LA CLASSE A

Dans A_1 et A_2, l'attitude n'exprime aucune antipathie à l'égard des mathématiques. Ce sont l'"absence de dons" ou le manque de travail qui semblent les facteurs de difficultés rencontrées dans la discipline.

A_1 Les modalités de cette classe permettent de camper un type d'élève dont les résultats seraient inférieurs à la moyenne, dans la zone D d'où il désespère de ne pouvoir jamais sortir. Les "dons" lui paraissent nécessaires : il ne les a pas, pense-t-il. Les mathématiques lui sont étrangères et ne lui procurent que de mauvaises notes : c'est le principal reproche qu'il formule.

A_2 Cette fois, si les résultats sont mauvais, la cause est à rechercher dans l'absence de travail. Même si l'élève-type de la classe semble peu attiré par le travail, l'école le sécurise : il redoute de la quitter et n'a comme crainte que de ne pas avoir le B.E.P.C., ni le droit de passer en Seconde. L'avenir ne s'exprime donc qu'en termes scolaires.

A_3 Cette classe admet une bonne cohésion : on y note plusieurs maxima de la statistique locale, indice de consistance de classe. C'est le passionné des mathématiques qui profile la classe ; et comme souvent, passion et réussite vont de pair. C'est le cas ici. Notons que l'on trouve dans A_3 non seulement ceux qui préfèrent notre discipline mais également les langues vivantes (2e langue, en général l'allemand, au début de l'année). Le plaisir annoncé de la construction, de la découverte est à associer à ce goût pour les langues.

A_4 L'antithèse de la classe précédente ; on trouve ici le profil du littéraire pur, non hostile mais étranger aux mathématiques : c'est une science abstraite où tout va trop vite, où le professeur a toujours raison.

A_5 L'élève profilé par cette classe a une attitude complexe ; sans doute veut-il plaire et peut-être faire illusion : son comportement apparaît, par conséquent, contradictoire, incohérent : partagé entre le besoin de vérifier, de représenter, le refus de manipuler, les hautes vertus de cohérence, etc., il se satisfait avant tout des bons résultats.

3.3. INTERPRETATION DE LA CLASSE B

B_1 C'est la classe des opinions révoltées, plus peut-être à l'égard de l'école et du professeur qu'à l'égard des mathématiques. Elles-mêmes manquent d'intérêt et sont difficiles à comprendre. Mais un alibi se présente : elles sont inutiles dans le futur.

B_2 Encore l'hostilité ... Mais cette fois, l'expression est plus atténuée et plus franche. Peut-être le travail est-il insuffisant ? Cependant, le professeur demeure un facteur important dans l'échec, ainsi que les "dons" que l'enfant ne pense pas avoir.

B_3 Pour l'enfant-type de cette classe, dont la réussite paraît la meilleure du groupe B, les mathématiques sont cependant inintéressantes ; la part du professeur, si elle apparaît sous des formes positives, est fortement corrélée à l'avenir : métier, vie, B.E.P.C., seconde. C'est dans cette classe que les modalités d'attente sont les plus nombreuses. Sans doute est-elle définie par les attitudes des enfants anxieux et inquiets, satisfaits de l'amalgame discipline - enseignant en sixième et cinquième, plus mécontents de cette association cette année.

B_4 L'élève-type de cette classe attend plus du groupe, du camarade donc que de lui-même et du professeur. Il est concret, a les pieds sur terre : éducation physique, sciences de la nature, histoire-géographie sont ses disciplines préférées. Les mathématiques doivent être un outil (comprendre plus vite) et lui apporter le plaisir de découvrir. Rien d'étonnant à ce que son attitude apparaisse incohérente : c'est lui qui répond *"rien"* aux deux items (17) et (18) tout en retenant d'autres modalités. Mais quelles sont-elles : les hautes valeurs intellectuelles que les mathématiques doivent développer (rigueur, cohérence, ...) et l'attente : *"calculer"* ? Y a-t-il vraiment incohérence pour qui connaît les premières semaines des cours de quatrième ?

B_5 La personnalité-type est moins tranchée : autorité des parents, démonstration du professeur le convainquent. Cette personnalité semble être celle d'un artiste (discipline préférée : travaux manuels, dessin ou musique), d'un tendre (qualités humaines du professeur) ; s'il s'intéresse plus que n'importe quel autre à son avenir en termes d'utilité future, de connaissances de la vie, c'est sans doute pour satisfaire l'adulte qui le presse, qui lui rappelle que *"les mathématiques modernes ne servent à rien"*. Peut-être est-ce pour cela et pour son goût du dessin qu'il attend beaucoup de la géométrie.

4. Facteurs d'attitude donnés par l'analyse des correspondances

On pourra éventuellement se reporter à la présentation théorique de cette méthode dans le tome II de la brochure. Disons seulement, en quelques mots, qu'elle consiste à extraire les lignes de force (les facteurs) d'un tableau de données. Pour prendre une image mécanique, ces lignes de force sont représentées par les directions massiques et géométriques, les axes principaux d'inertie du nuage des variables principales, les items du questionnaire. Ces variables sont munies des pondérations de leur fréquence et séparées par une distance prenant en compte les comportements discriminants des enfants à leur égard. Les espaces engendrés par les axes principaux, dans l'ordre de leur importance, restituent une image d'autant plus fidèle du nuage que le nombre des axes pris en considération augmente.

4.1. GRANDES LIGNES

Le programme Tabet pouvant accepter plus de modalités que le précédent, nous avons introduit cette fois 198 variables principales représentant chacune une modalité. C'est ainsi que nous pouvons traiter les réponses aux questions sur les professions souhaitées ou refusées,

ainsi qu'aux questions ③ (avec ses 20 modalités supplémentaires), ④, ⑤, ⑦, ⑧, ⑩, ⑪, ⑫, ⑬, ⑭, ⑮, ⑯, ⑰ et ⑱ en intégrant pour ces quatre dernières les classements 1, 2 et 3 proposés par les enfants.

En variables supplémentaires, nous admettons les âges, les sexes, le caractère redoublant, les 8 classes de C.E.S. et la profession des parents; au total 27 variables supplémentaires.

Nous demandons au programme les directions des axes principaux d'inertie, les composantes des (198 + 27) variables sur les quatre premiers axes, les projections de ces points dans les plans 1-2, 1-3 et 2-3.

Les trois premières valeurs propres, en dépit du nombre important des variables, représentent des pourcentages intéressants de l'inertie donc de l'information globale : respectivement 3,12 %, 2,25 % et 2 %. Nous interpréterons donc les trois facteurs associés et tenterons de dégager les appuis respectifs des deux analyses.

A grands traits, les trois facteurs peuvent être définis ainsi :

Le facteur 1 oppose les enfants selon leur intérêt pour les mathématiques. On verra la part du professeur dans ce facteur.

Le facteur 2 est celui de la neutralité ; il oppose les élèves qui n'apprécient rien à ceux qui ne reprochent rien. Cet axe sépare de plus les variables définissant les raisons du refus des mathématiques de celles définissant l'intérêt porté à elles.

Le facteur 3 oppose les goûts extrêmes à l'égard des mathématiques où le professeur intervient nettement, au goût plus neutre ; ce sont cette fois les extrêmes s'opposant au centre.

Cette désignation des facteurs n'est pas sans nous rappeler fortement celle mise en évidence dans l'analyse du référendum et des élections présidentielles françaises de 1969, effectuées dans le laboratoire de J.P. Benzecri par B. Alcantar, G. Bordier et J. Obadia. Les trois premiers facteurs représentaient respectivement :

— une opposition gauche-droite
— une opposition vote blanc (opposant actif) et non-vote (opposant passif)
— une opposition extrêmes-centre.

4.2. PREMIER FACTEUR

Dans le plan 1-2, on obtient une projection du nuage ayant approximativement les formes suivantes :

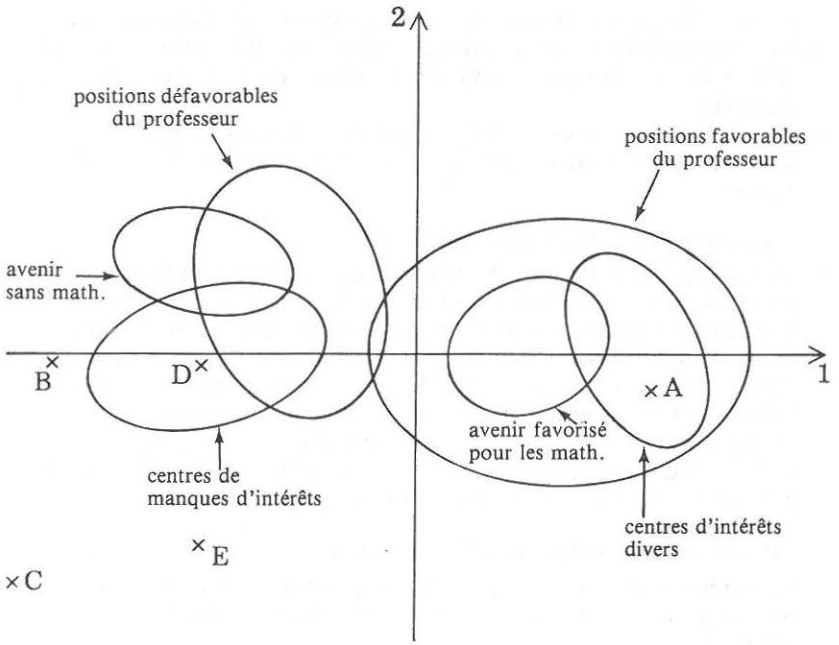


Figure 1
Les regroupements en "patate" sont issus d'interprétation d'un critère classifiant

a) Variables principales

La nature des cinq points ayant les plus fortes contributions à ce facteur permet de le définir, sans ambiguïté semble-t-il :

— amour des mathématiques : beaucoup (A)
— amour des mathématiques : pas du tout (B)
— attente de l'enseignement des mathématiques : rien (en 1er choix) (C)
— appréciation des mathématiques : rien (en 1er choix) (D)
— intérêt depuis l'école primaire : a régressé (E).

Corrélativement, on trouve les trois "nimbus" représentés sur la figure 1 :
— centres d'intérêt
— place du professeur
— avenir favorisé ou non par les mathématiques.

Bien entendu, sur la partie droite de l'axe, on rencontre les points qui en confirment la nature :

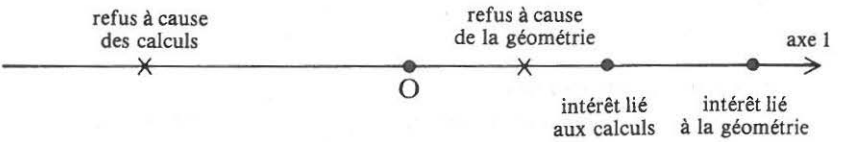
— les mathématiques constituent la discipline préférée (en 1er)
— les mathématiques sont aimées en raison de bons résultats scolaires qu'elles apportent (significatif de la liaison : intérêt - résultats)
— on souhaite exercer la profession d'enseignant de mathématiques
— les mathématiques sont aimées parce qu'elles sont originales, objectives, rigoureuses, amusantes, parce qu'il y faut chercher, découvrir
— une restriction : excès de manipulations et de dessins (les manipulations ne correspondent pas à l'image traditionnelle des mathématiques).

A gauche, on retrouve les antithèses :

— les résultats sont faibles, les mathématiques sont peu intéressantes, les explications sont trop rapides, il faut trop réfléchir
— le travail personnel n'amène qu'une légère amélioration, d'ailleurs les mathématiques donnent trop de travail et il faut être doué pour réussir. Le travail de groupe a, par contre, de grandes vertus.
— l'histoire et la géographie, le français-latin ou les travaux manuels - musique - dessin sont les disciplines préférées en premier
— la profession d'enseignant de mathématiques n'est pas souhaitée ; par contre, celle de commerçant-artisan paraît s'y opposer.

Remarquons un élément assez étonnant :

— les points représentatifs des facteurs d'intérêt ou de désintérêt, relatifs d'une part à la géométrie, d'autre part aux calculs, se placent ainsi sur le 1er axe :



Il n'y a donc pas au niveau de la géométrie l'homogénéité attendue : est-ce dû à l'impact expérimental ou témoin au bout de trois semaines de cours ? Quoi qu'il en soit, l'attitude à l'égard des mathématiques est plus confuse lorsque la géométrie est évoquée.

Ces différentes remarques permettent d'établir un portrait du satisfait-type (resp. du mécontent-type). Retenons la nature des disciplines préférées, la place du travail personnel (resp. du groupe) et celle des résultats dans les définitions de ces portraits. On va maintenant préciser ces profils par les variables descriptives des enfants et on appréciera la plus grande netteté de leurs contours.

b) Examinons donc la position des variables supplémentaires sur cet axe

La figure ci-dessous représente les points âges, sexes, professions des parents(1), caractère redoublant - non redoublant dans le plan 1-2. On peut superposer mentalement les deux figures 1 et 2 pour replacer les variables secondaires par rapport aux variables principales.

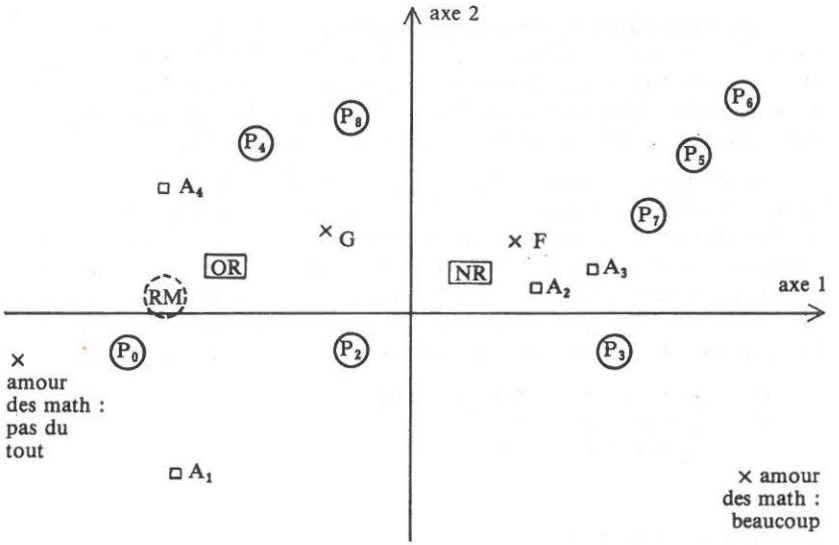


Figure 2

On peut constater que les filles sont plus favorables aux mathématiques que les garçons, de même que les non-redoublants le sont très naturellement plus que les redoublants.

(1) Rappelons le codage socio-professionnel INSEE et notre codage le prolongeant :
Profession d'agriculteur exercée (P_0), souhaitée (S_0), refusée (R_0) par l'élève.
Profession de patron d'industrie ou du commerce exercée (P_2), souhaitée (S_2), refusée (R_2).
Profession libérale (+ cadres supérieurs) exercée (P_3), souhaitée (S_3), refusée (R_3).
Profession de cadre moyen exercée (P_4), souhaitée (S_4), refusée (R_4).
Profession d'employé exercée (P_5), souhaitée (S_5), refusée (R_5).
Profession d'ouvrier exercée (P_6), souhaitée (S_6), refusée (R_6).
Profession de personnel de service exercée (P_7), souhaitée (S_7), refusée (R_7).
Profession d'armée, de police exercée (P_8), souhaitée (S_8), refusée (R_8).
Personne non active (P_9), souhait (S_9), refus (R_9).

On remarquera :

— que les âges *"normaux"* (12 et 13 ans) occupent une position positive sur le premier axe ;
— que les plus jeunes (11 ans) se rapprochent des disciplines suivantes: sciences naturelles, travaux manuels, musique, dessin et apprécient le travail de groupe (actif ...) ;
— que les plus âgés (14 ans) pensent à la classe de seconde et soutiennent qu'il faut être doué pour réussir.

Les professions P_3 sont naturellement à droite de l'axe mais également s'y trouvent les milieux plus modestes : P_5, P_6 et P_7. Par contre, on est surpris de trouver les cadres moyens (P_4) à gauche de l'axe, dans une position moins marquée peut-être que P_0 mais plus que P_2 et P_8. Une étude psycho-sociologique serait très intéressante à mener.

A ce sujet, on notera les positions respectives des professions souhaitées (S_i), refusées (R_j) et exercées par le chef de famille (P_k), tout en se méfiant d'une interprétation hâtive, apparemment contradictoire, d'une proximité d'un point S_i et d'un point R_i : en effet, un enfant peut souhaiter exercer une profession du groupe i et en même temps refuser une profession déterminée du même groupe i . De plus, d'autres variables entrent en jeu pour définir les proximités. Relevons cependant que :

P_3 *"s'oppose"* nettement à S_2 et S_7
P_4 *"s'oppose"* à S_3(!) et S_5
P_6 *"s'oppose"* à S_2 et S_7.

4.3. DEUXIEME FACTEUR

Les nuances sont ici beaucoup plus subtiles et nous ne nous aventurerons pas dans des hypothèses trop contestables. Signalons pour commencer que les variables contribuant le plus à définir l'axe 2 sont peu significatives, car faiblement représentées : la distance utilisée les sépare fortement des autres :

à droite
- — réussite par aide des parents
- — appréciation dans l'enseignement : rien

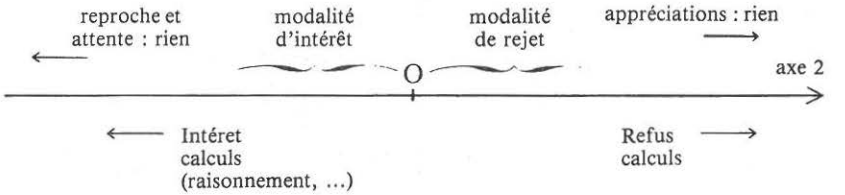
à gauche
- — attente de l'enseignement : rien
- — conviction par l'affirmation du camarade le plus doué
- — reproche à l'égard de l'enseignement des mathématiques : rien
- — profession souhaitée : personnel de service

Evoquant l'analyse du référendum de 1969, nous comparerons un refus actif *("on n'apprécie rien")* au vote blanc qui nécessite le déplacement vers les urnes pour exprimer un mécontentement. Il s'oppose alors à l'attitude passive, indifférence totale et presque contradictoire : on n'attend rien des mathématiques mais on ne leur reproche rien. Elles

représentent une discipline étrangère, très extérieure à la vie scolaire et affective. On notera alors les places des parents et du camarade le plus doué :

— les parents obligent à mener une vie scolaire normale
— le camarade peut représenter un écran intellectuel confortable.

Autour de ces variables on observe, à droite, presque toutes celles de l'item 3 qui expriment les raisons du rejet des mathématiques, en particulier le calcul déplaisant. A gauche, on trouve celles qui ont dû conduire à des *"oui"* tièdes et en particulier, le calcul. Ces variables ont de faibles contributions à l'axe mais, on retiendra ... leur contribution au sens que nous lui avons donné. Soulignons la place du calcul dans la détermination des enfants. Il s'accompagne, à gauche, de choix très nobles dans les attentes ou les éléments de conviction : rigueur, cohérence, beauté du raisonnement, etc.



Rappelons la position de certaines variables supplémentaires sur cet axe :

— à gauche : 11 ans, à droite : 14 ans, au centre : 12 et 13 ans
— à gauche : P_3, à droite essentiellement : P_4, P_8.

4.4. TROISIEME FACTEUR

De nombreuses variables contribuent à définir le demi-axe positif (partie supérieure de la figure 3) :

— aucune attente (en premier)
— résultats en cinquième faibles
— discipline préférée : travaux manuels, musique, dessin
— amour des mathématiques : pas du tout
— amour des mathématiques : beaucoup.

On trouve sur le demi-axe négatif essentiellement :

— amour des mathématiques : un peu
— attente : passer en seconde
— reproches : rien.

Le sens que nous avons donné à ce troisième facteur n'est donc pas démenti par ces variables. Les voici représentées dans le plan 2-3.

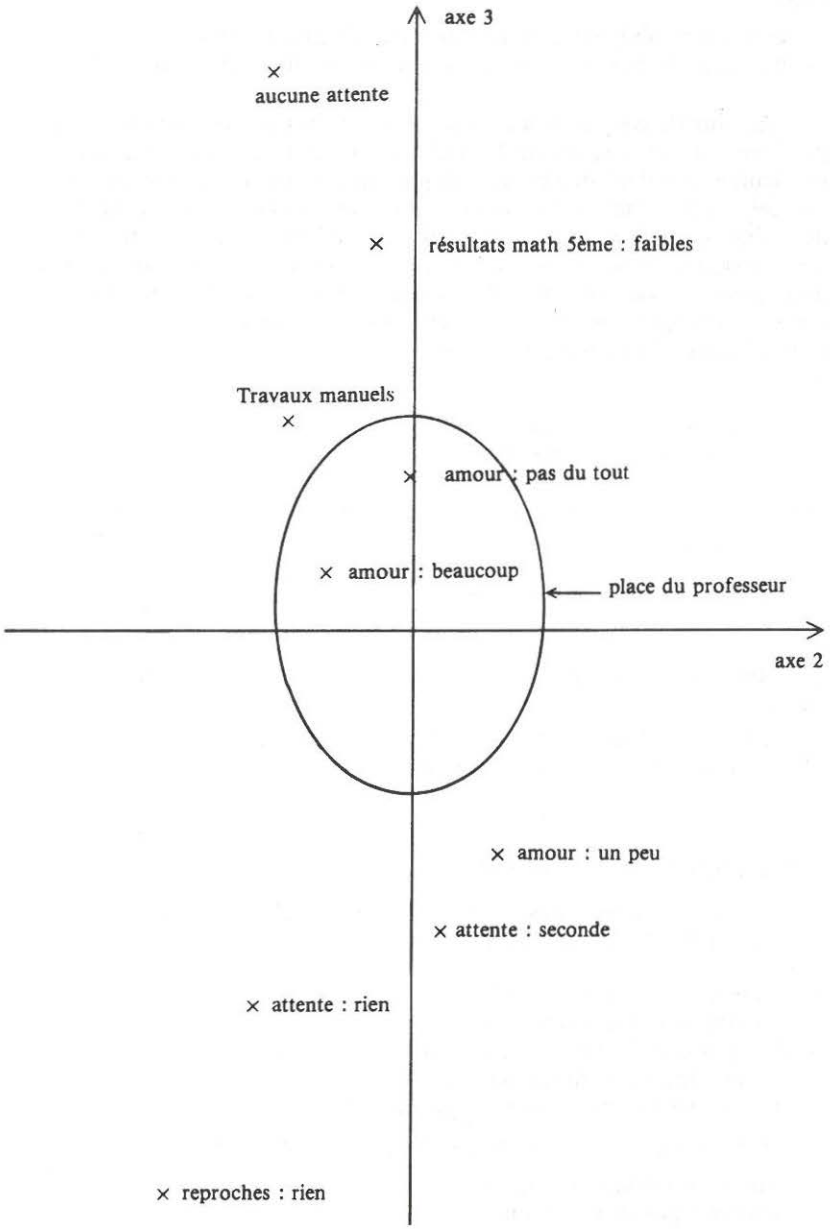


Figure 3

En consolidation de notre hypothèse d'opposition *"extrêmes-centre"* citons les quelques variables suivantes :

à droite
- — temps agréable que fait passer le professeur
- — intérêt dû au professeur, aux résultats, à l'originalité, à l'ouverture des idées, à la rapidité de la compréhension des mathématiques.
- — rejet par trop de travail
- — l'intérêt a décru nettement depuis le primaire

à gauche
- — refus de la profession d'enseignant de mathématiques
- — *"affirmation du camarade"* et *"aide des parents"*
- — raisonner, calculer, être rigoureux ...
- — et curieusement : *"les mathématiques de quatrième paraissent plus faciles que celles de cinquième"*.

Les points relatifs aux variables géométrie et calcul se placent nettement, le premier à droite (attitude très marquée des enfants), le second à gauche (ébauche d'une certaine indifférence).

Remarquons la présence des quelques variables supplémentaires significatives sur l'axe 3 :

— les enfants de 11 ans par opposition à ceux de 14 ans ont une attitude marquée à l'égard des mathématiques

— c'est le cas également des enfants des groupes socio-professionnels
P_3 et P_6 à droite
P_7 et P_2 à gauche

5. Quelques conclusions

Les deux types d'analyse de questionnaire font ressortir que l'élément discriminant majeur, dans une telle enquête, est l'intérêt porté à la matière abordée par l'élève. Suspicion à l'égard de la sincérité des enfants mise à part, retenons que cet intérêt est fondé principalement sur deux variables et fortement corrélé à elles :

— le professeur en tant qu'être humain et en tant que didacticien
— les résultats obtenus dans la discipline.

Ces derniers sont eux-mêmes très liés tout naturellement au travail mais aussi au professeur (responsable de la bonne ou de la mauvaise note) et aux dons, surtout lorsque les résultats sont mauvais.

Enfin, il apparaît bien souvent que la citation du calcul en facteur d'*"intérêt"* est accompagné d'un manque ... d'intérêt à l'égard de la discipline : son côté terne (associé aux hautes vertus intellectuelles supposées atteintes par les mathématiques) lui enlève tout rapport avec la vie et les joies scolaires. Quant à la géométrie, elle paraît avoir déjà provoqué des différences entre classes expérimentales et témoins. La seule différence dans cette analyse, peut-être ...

Bibliographie

(cf. aussi [1] de la bibliographie générale)

[1] V. ALEXANDRE. *Echelles d'attitude.* (Editions Universitaires 1971).

[2] R. GRAS. *Analyse statistique succincte d'une enquête auprès d'élèves de quatrième sur l'enseignement des mathématiques* (IREM de Rennes, mars 1975).

[3] R. GRAS. *Analyse d'un questionnaire sur l'enseignement des mathématiques en classes témoins et expérimentales en fin de quatrième* (IREM de Rennes, août 1975).

[4] R. GRAS. *Analyses factorielle et classificatoire d'un test national d'entrée en quatrième en mathématique.* (IREM de Rennes, avril 1976).

[5] I.C. LERMAN. *Cours d'analyse des données.* (Université de Rennes).

[6] I.C. LERMAN. *Introduction à une méthode de classification automatique illustrée par la recherche d'une typologie de personnages enfants à travers la littérature enfantine.* (Revue de Statistique appliquée 1973, volume XXI, n° 3).

[7] I.C. LERMAN. *Reconnaissance et classification des structures finies en analyse des données,* (I.R.I.S.A. Université de Rennes, rapport n° 70).

[8] Brochure A.P.M.E.P. n° 34, *"Recherche inter-IREM 1973-78, en géométrie de 4ème et 3ème dite O.P.C. : réflexion critique et évaluation"*, 1979.

Annexe 1

Université de Rennes
I.R.E.M. - C.I.O.
VANNES

Date : .....

ENQUETE O.P.C. - VANNES

NOM : ..... Prénom : .....
Date de naissance : ..... Sexe : .....
Profession du chef de famille : .....
Quelle profession aimeriez-vous exercer ? .....
Quelle est la profession que vous n'aimeriez pas exercer ? .....
Redoublant : OUI ☐ NON ☐

Ce questionnaire doit nous permettre de mieux vous connaître, et notamment de savoir ce que vous pensez des mathématiques.

Pour répondre, il vous suffira de faire une croix dans la case qui correspond à ce que vous pensez.

① Fréquentez-vous un ☐ C.E.S. un ☐ C.E.G.

② Etes-vous un élève de ☐ 4e I ou II ☐ 4e aménagée

③ Aimez-vous les mathématiques ?
☐ un peu (65 %) ☐ beaucoup (27 %) ☐ pas du tout (8 %)
Dites pourquoi : .....
.....

④ Quels étaient vos résultats en mathématiques en classe de cinquième ?
(8%)☐ très bons (49%)☐ bons (34%)☐ moyens
(7%)☐ médiocres (2%)☐ faibles

⑤ Pensez-vous, actuellement, que les mathématiques de quatrième sont, par rapport aux mathématiques de cinquième :
(1%)☐ plus faciles (64%)☐ plus difficiles (23%)☐ de même difficulté
(12%)☐ je ne sais pas

⑥ Pensez-vous que la manière de vous enseigner les mathématiques est cette année, par rapport aux années précédentes :
☐ différente ☐ semblable ☐ je ne sais pas

⑦ Pensez-vous que pour réussir en mathématiques, il soit nécessaire d'être doué
(27%)☐ OUI (55%)☐ NON (18%)☐ Je ne sais pas

⑧ Croyez-vous que le travail personnel puisse apporter une amélioration des résultats en mathématiques :
(50%)☐ importante (21%)☐ plutôt moyenne (14%)☐ légère
(15%)☐ je ne sais pas

⑨ En ce qui vous concerne, en classe de quatrième, le travail personnel a-t-il amélioré vos résultats en mathématiques :
☐ oui, nettement ☐ oui, un peu ☐ non
☐ je n'ai pas fourni un travail suffisant

⑩ Depuis l'école primaire, votre intérêt pour les mathématiques s'est-il accru :
(35%)☐ oui, nettement (39%)☐ oui, un peu (16%)☐ non
(10%)☐ non, au contraire, il a régressé

⑪ Quelle part attribuez-vous au professeur dans vos résultats en mathématiques et votre intérêt pour cette discipline ?
(2%)☐ nulle (19%)☐ légère (52%)☐ grande (23%)☐ très grande

⑫ Parmi les disciplines enseignées en quatrième, quelles sont celles que vous préférez ?

	E.P.	F.L.	H.G.	L.V.	M.	S.N.	T.	T.M.
En 1er	5 %	19 %	13 %	26 %	10 %	8 %	5 %	4 %
En 2e	4 %	13 %	14 %	25 %	18 %	6 %	4 %	4 %

⑬ Actuellement, envisagez-vous une profession où les "mathématiques" ont un rôle important ?
(29%)☐ OUI (33%)☐ NON (38%)☐ Je ne sais pas

⑭ A votre avis, quel est le facteur le plus important pour réussir en mathématiques ?
(3%)☐ une bonne connaissance de la langue française
(7%)☐ les dons
(1%)☐ l'aide des parents
(29%)☐ les qualités professionnelles du professeur
(10%)☐ le travail en petit groupe
(49%)☐ le travail
(1%)☐ l'autorité des parents
(9%)☐ les qualités humaines du professeur

Pour les questions qui suivent, choisissez trois réponses que vous *classez* par ordre préférentiel, en mettant devant elles les chiffres 1, 2 ou 3.

⑮ Etes-vous convaincu d'un résultat ou d'une propriété géométrique par (1)
(16%)☐ une constatation visuelle sur une figure dessinée avec soin
(35%)☐ une démonstration rigoureuse du professeur
(12%)☐ une affirmation du professeur
(22%)☐ une démonstration que vous avez trouvée vous-même
(2%)☐ la cohérence (ou accord) avec des résultats antérieurs
(6%)☐ une vérification par des calculs complémentaires
(9%)☐ l'utilisation d'un instrument de dessin (compas, équerre, ...)
(1%)☐ l'affirmation du camarade qui vous paraît le plus doué

⑯ Qu'attendez-vous de l'enseignement des mathématiques ?
(14%)☐ une bonne note au B.E.P.C.
(32%)☐ un niveau suffisant pour entrer en seconde
(4%)☐ apprendre à calculer
(14%)☐ apprendre à raisonner
(17%)☐ la préparation d'un métier
(18%)☐ des connaissances utiles dans la vie
(2%)☐ rien

⑰ Que reprochez-vous aux mathématiques ?
(4%)☐ de donner trop de travail
(31%)☐ d'être difficiles à comprendre
(5%)☐ la rareté des constructions et des représentations
(5%)☐ d'être peu intéressantes
(10%)☐ des formulations trop complexes
(1%)☐ l'excès des manipulations et des dessins
(4%)☐ l'excès du calcul
(20%)☐ la trop grande vitesse des explications
(14%)☐ rien

(1) *Les pourcentages indiqués entre parenthèses sont uniquement relatifs au premier choix des enfants.*

⑱ Qu'appréciez-vous dans l'enseignement des mathématiques ?
(2%)☐ la rigueur
(6%)☐ la beauté du raisonnement
(14%)☐ leur originalité par rapport aux autres matières
(17%)☐ des résultats vérifiables dans la réalité
(27%)☐ le plaisir de chercher, de découvrir
(5%)☐ des résultats imprévisibles

(1%)☐ leur objectivité
(9%)☐ le fait de parler et d'écrire avec un minimum de mots
(3%)☐ le temps agréable que vous fait passer le professeur
(13%)☐ rien

Modalités de réponses observées à la question 3 ouverte

Raisons d'intérêt

— géométrie
— calculs
— plaisir de rechercher et découvrir
— apprendre à raisonner et réfléchir
— le plaisir d'en faire grâce à une compréhension rapide
— élargir les idées
— résultats obtenus
— les mathématiques sont intéressantes
— les mathématiques sont originales et amusantes
— le professeur
— utilité future

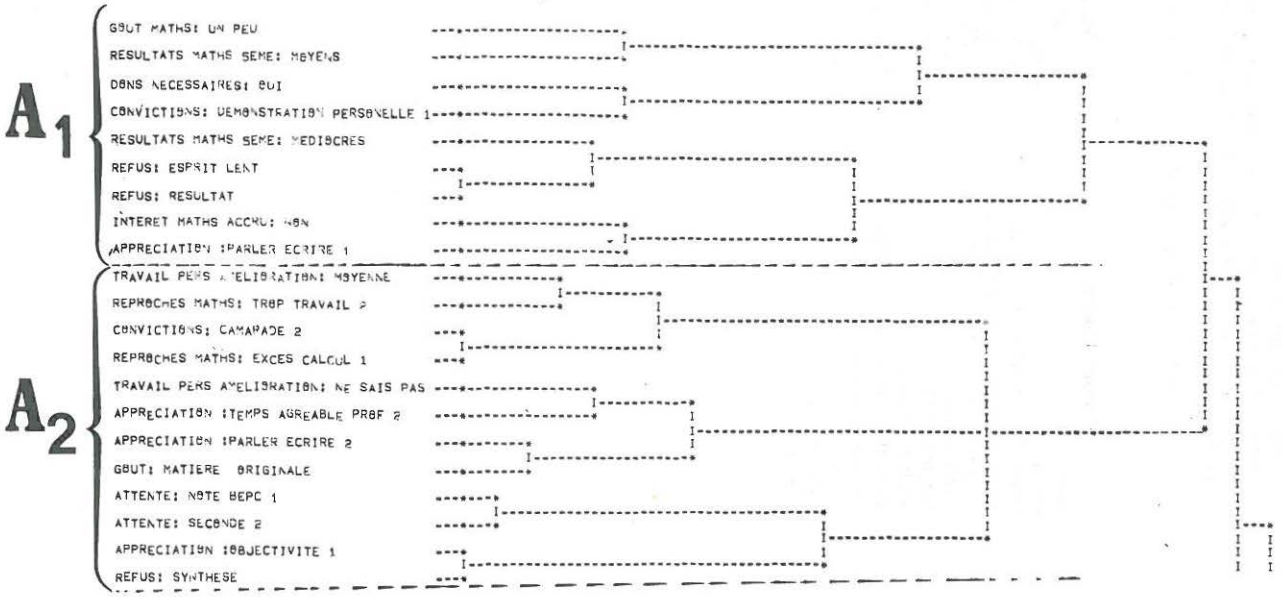
Raisons de refus des math.

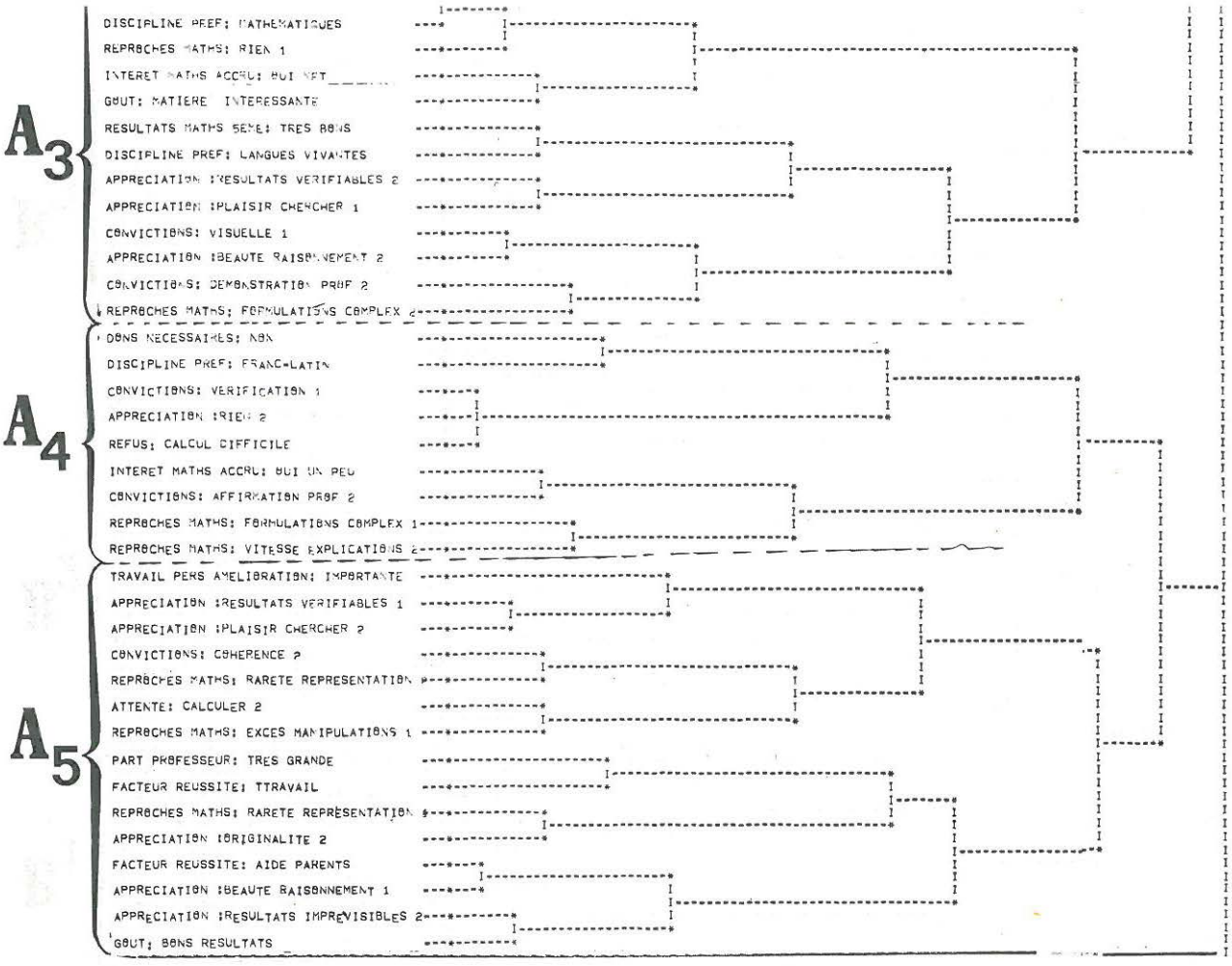
— géométrie
— calculs
— le désagrément d'en faire à cause d'un esprit lent
— le désagrément à cause du raisonnement et de la réflexion
— compréhension difficile
— la difficulté de synthèse
— les résultats obtenus
— les mathématiques sont inintéressantes
— le professeur

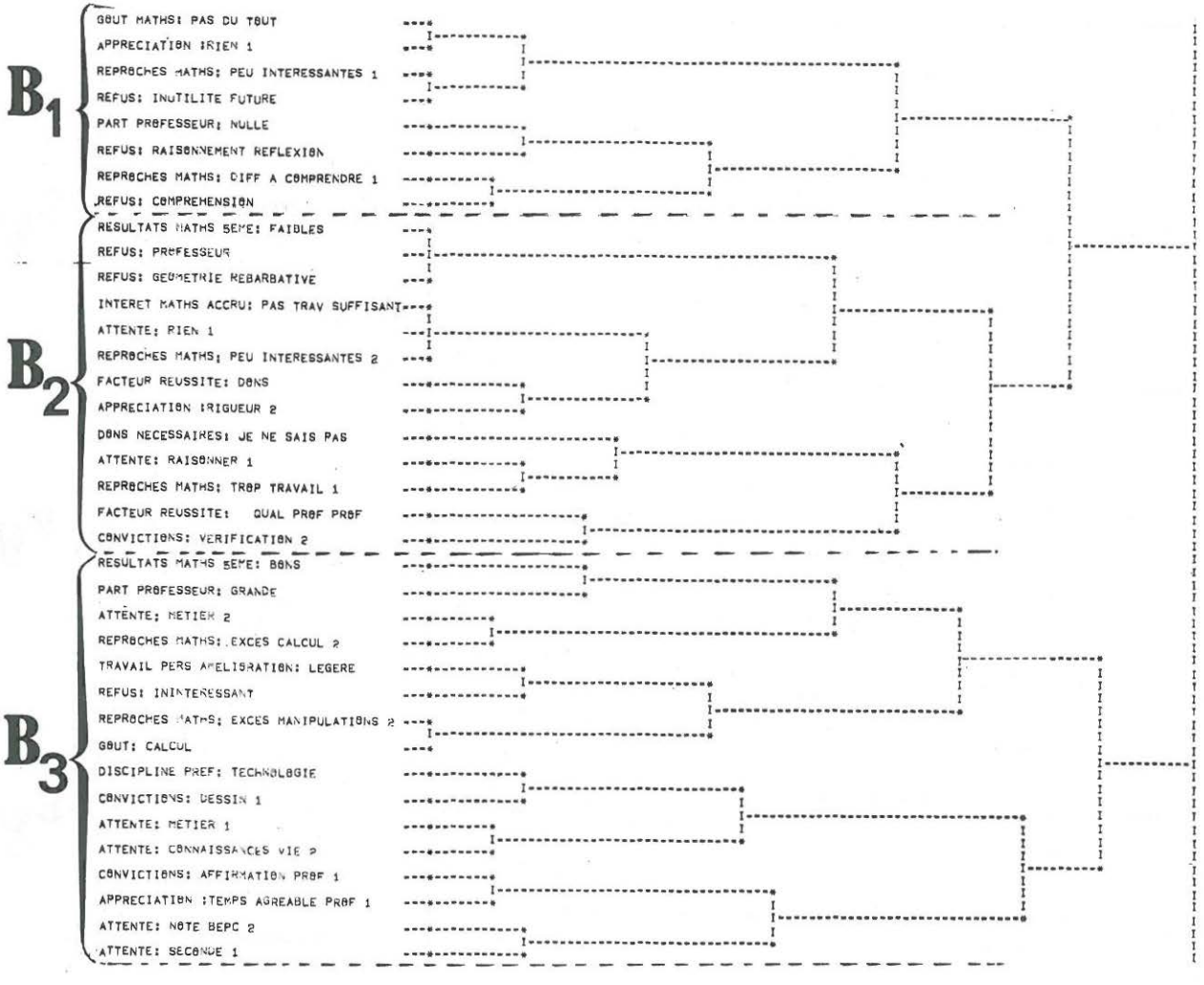
Annexe 2

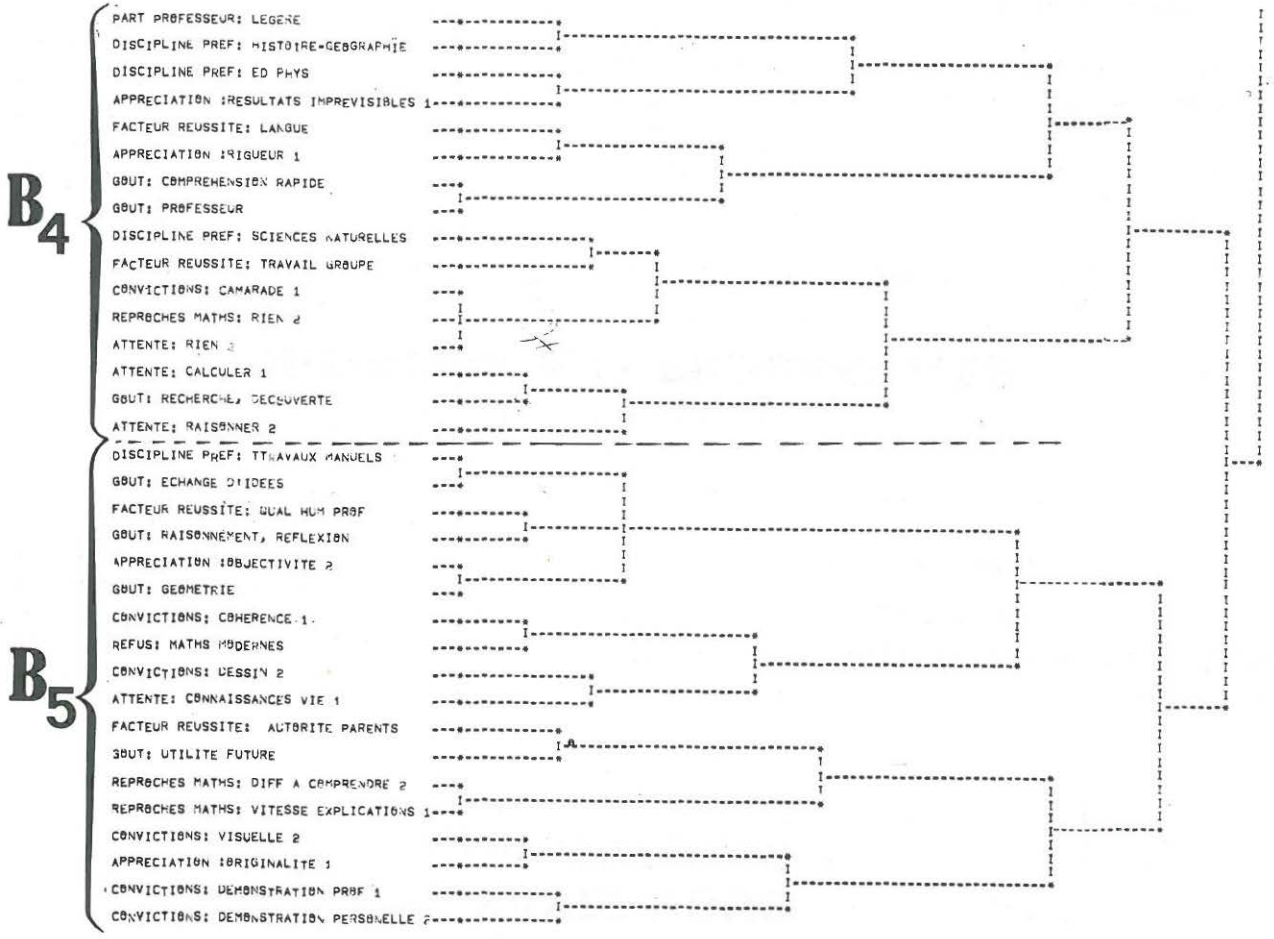
REPRESENTATION DE L'ARBRE CONDENSE A SES NIVEAUX LES PLUS SIGNIFICATIFS

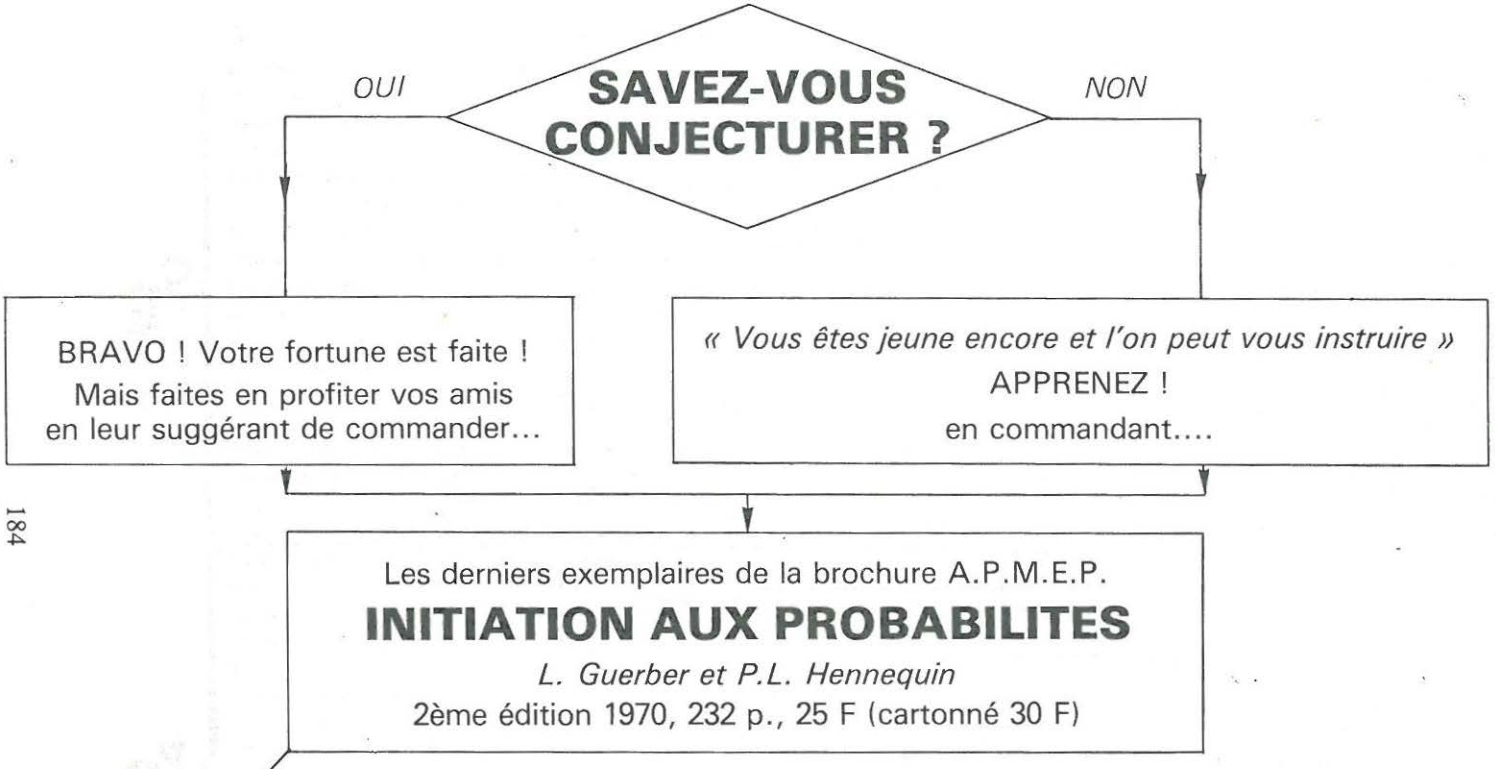
ARBRE DE REPARTITION DES ITEMS
CLASSIFICATION LERMAN, OPC VANNES, ENTREE QUATRIEME, 75-76
NIVEAUX 1 à 21




A3
DISCIPLINE PREF: MATHEMATIQUES
REPROCHES MATHS: RIEN 1
INTERET MATHS ACCRU: OUI BEAUCOUP
GOUT: MATIERE INTERESSANTE
RESULTATS MATHS 5EME: TRES BONS
DISCIPLINE PREF: LANGUES VIVANTES
APPRECIATION :RESULTATS VERIFIABLES 2
APPRECIATION :PLAISIR CHERCHER 1
CONVICTIONS: VISUELLE 1
APPRECIATION :BEAUTE RAISONNEMENT 2
CONVICTIONS: DEMONSTRATION PREF 2
REPROCHES MATHS: FORMULATIONS COMPLEX 2
A4
BONS NECESSAIRES: NON
DISCIPLINE PREF: FRANC-LATIN
CONVICTIONS: VERIFICATION 1
APPRECIATION :RIEN 2
REFUS: CALCUL DIFFICILE
INTERET MATHS ACCRU: OUI UN PEU
CONVICTIONS: AFFIRMATION PREF 2
REPROCHES MATHS: FORMULATIONS COMPLEX 1
REPROCHES MATHS: VITESSE EXPLICATIONS 2
A5
TRAVAIL PERS AMELIORATION: IMPORTANTE
APPRECIATION :RESULTATS VERIFIABLES 1
APPRECIATION :PLAISIR CHERCHER 2
CONVICTIONS: COHERENCE 2
REPROCHES MATHS: RARETE REPRESENTATION
ATTENTE: CALCULER 2
REPROCHES MATHS: EXCES MANIPULATIONS 1
PART PROFESSEUR: TRES GRANDE
FACTEUR REUSSITE: TTRAVAIL
REPROCHES MATHS: RARETE REPRESENTATION
APPRECIATION :ORIGINALITE 2
FACTEUR REUSSITE: AIDE PARENTS
APPRECIATION :BEAUTE RAISONNEMENT 1
APPRECIATION :RESULTATS IMPREVISIBLES 2
GOUT: BONS RESULTATS


B1
B2
B3
GOUT MATHS: PAS DU TOUT
APPRECIATION :RIEN 1
REPROCHES MATHS: PEU INTERESSANTES 1
REFUS: INUTILITE FUTURE
PART PROFESSEUR: NULLE
REFUS: RAISONNEMENT REFLEXION
REPROCHES MATHS: DIFF A COMPRENDRE 1
REFUS: COMPREHENSION
RESULTATS MATHS 5EME: FAIBLES
REFUS: PROFESSEUR
REFUS: GEOMETRIE REBARBATIVE
INTERET MATHS ACCRU: PAS TRAV SUFFISANT
ATTENTE: RIEN 1
REPROCHES MATHS: PEU INTERESSANTES 2
FACTEUR REUSSITE: DONS
APPRECIATION :RIGUEUR 2
DONS NECESSAIRES: JE NE SAIS PAS
ATTENTE: RAISONNER 1
REPROCHES MATHS: TROP TRAVAIL 1
FACTEUR REUSSITE: QUAL PROF PROF
CONVICTIONS: VERIFICATION 2
RESULTATS MATHS 5EME: BONS
PART PROFESSEUR: GRANDE
ATTENTE: METIER 2
REPROCHES MATHS: EXCES CALCUL 2
TRAVAIL PERS AMELIORATION: LEGERE
REFUS: ININTERESSANT
REPROCHES MATHS: EXCES MANIPULATIONS 2
GOUT: CALCUL
DISCIPLINE PREF: TECHNOLOGIE
CONVICTIONS: DESSIN 1
ATTENTE: METIER 1
ATTENTE: CONNAISSANCES VIE 2
CONVICTIONS: AFFIRMATION PROF 1
APPRECIATION :TEMPS AGREABLE PROF 1
ATTENTE: NOTE BEPC 2
ATTENTE: SECONDE 1


B4
B5
PART PROFESSEUR: LEGERE
DISCIPLINE PREF: HISTOIRE-GEOGRAPHIE
DISCIPLINE PREF: ED PHYS
APPRECIATION :RESULTATS IMPREVISIBLES 1
FACTEUR REUSSITE: LANGUE
APPRECIATION :RIGUEUR 1
GOUT: COMPREHENSION RAPIDE
GOUT: PROFESSEUR
DISCIPLINE PREF: SCIENCES NATURELLES
FACTEUR REUSSITE: TRAVAIL GROUPE
CONVICTIONS: CAMARADE 1
REPROCHES MATHS: RIEN 2
ATTENTE: RIEN 2
ATTENTE: CALCULER 1
GOUT: RECHERCHE, DECOUVERTE
ATTENTE: RAISONNER 2
DISCIPLINE PREF: TRAVAUX MANUELS
GOUT: ECHANGE D'IDEES
FACTEUR REUSSITE: QUAL HUM PROF
GOUT: RAISONNEMENT, REFLEXION
APPRECIATION :OBJECTIVITE 2
GOUT: GEOMETRIE
CONVICTIONS: COHERENCE 1
REFUS: MATHS MODERNES
CONVICTIONS: DESSIN 2
ATTENTE: CONNAISSANCES VIE 1
FACTEUR REUSSITE: AUTORITE PARENTS
GOUT: UTILITE FUTURE
REPROCHES MATHS: DIFF A COMPRENDRE 2
REPROCHES MATHS: VITESSE EXPLICATIONS 1
CONVICTIONS: VISUELLE 2
APPRECIATION :ORIGINALITE 1
CONVICTIONS: DEMONSTRATION PROF 1
CONVICTIONS: DEMONSTRATION PERSONNELLE 2


OUI
SAVEZ-VOUS CONJECTURER ?
NON
BRAVO ! Votre fortune est faite !
Mais faites en profiter vos amis
en leur suggérant de commander...
« Vous êtes jeune encore et l'on peut vous instruire »
APPRENEZ !
en commandant....
Les derniers exemplaires de la brochure A.P.M.E.P.
INITIATION AUX PROBABILITES
L. Guerber et P.L. Hennequin
2ème édition 1970, 232 p., 25 F (cartonné 30 F)
votre Régionale
ou
A.P.M.E.P.
29, rue d'Ulm, Paris 5e
CCP. Paris 5708.21N
1. Analyse Combinatoire.
2. Probabilités et variables aléatoires sur un ensemble fini.
3. Dépendance et indépendance, espaces produits et suites de variables aléatoires.
4. Jugement sur un sous-ensemble ou sur un échantillon.

J. PONTIER
Université Claude Bernard, Lyon I

8. ANALYSE DISCRIMINANTE

Contrairement aux précédents, la lecture de cet article suppose quelques connaissances préalables de calcul des Probabilités que le lecteur pourra s'il le souhaite trouver dans [4] et [6] de la bibliographie générale.

L'exposé qui suit donne un aperçu des problèmes que se proposent de résoudre les méthodes d'analyse discriminante, et décrit certaines de ces méthodes dans des cas où leur application ne nécessite pas de moyens de calcul trop importants. Si l'on veut discriminer deux groupes, sur la base d'une variable, ou de deux variables, les calculs pourront être faits « à la main » ; l'aide d'une simple minicalculatrice est évidemment souhaitable, et peut même rendre attrayant pour l'élève ce type de calcul. Mais bien entendu il est illusoire de vouloir étudier par une analyse discriminante (comme d'ailleurs par toute méthode d'analyse multivariée) des situations relativement compliquées (plus de deux groupes, plus de deux variables), sans avoir une bonne connaissance du calcul matriciel (y compris des éléments propres d'une matrice), et sans avoir recours à l'ordinateur pour la mise en œuvre des calculs numériques.

0 Index terminologique et notations

(La liste ci-dessous est présentée dans l'ordre d'apparition des termes dans le texte)

- Modèle d'urne
- Règle de décision
- Risque d'erreur
- Loi de Gauss
- Moyenne
- Dispersion
- Espérance mathématique
- Ecart-type
- Degré de liberté
- Variation
- Variation intra-groupes
- Variance intra-groupes
- Variation inter-groupes
- Degré de liberté intra-groupes
- Variance inter-groupes
- Degré de liberté inter-groupes

— Densité de probabilité
— Vraisemblance
— Seuil de décision
— Fonction de répartition de Gauss
— Groupes
— Sous-populations
— Classes
— Catégories
— Echantillon
— Variance
— Estimation
— Pouvoir discriminant
— Théorème de Bayes
— Variable F de Fisher-Snedecor
— Discrimination quadratique
— Discrimination linéaire
— Inégalité de Tchébycheff
— Fonction discriminante
— Covariation
— Covariance
— Covariation inter-groupes
— Covariation intra-groupes

μ = moyenne théorique d'une variable aléatoire x
$\overline{x}$ = estimation de μ à partir d'un échantillon
$\sigma^2(x) = \sigma^2$ = variance théorique de la variable aléatoire x
$\hat{s}^2(x) = \hat{s}^2$ = estimation de σ^2 à partir d'un échantillon
$\sigma(x) = \sigma$ = écart-type théorique
$\hat{s}(x) = \hat{s}$ = estimation de σ
$\sigma(x, y)$ = covariance théorique de x et y
$\hat{s}(x, y)$ = estimation de $\sigma(x, y)$

1. Exercices introductifs :

Les exercices qui suivent sont proposés au lecteur dans le but de le mettre dans l'« ambiance » de l'analyse discriminante, c'est-à-dire dans des situations où il est impossible de classer à coup sûr des objets, donc en prenant un risque d'erreur de classement, risque que l'on peut calculer. De tels exercices, éventuellement adaptés, peuvent donner lieu à des travaux pratiques « d'éveil » avec les élèves de terminales ou de propédeutique. Un lecteur déjà averti du type de problème que se propose de résoudre l'analyse discriminante peut aller directement au **§ 2.**

Dans ces exercices nous avons choisi délibérément le « *modèle d'urne* », car c'est un modèle de référence susceptible d'être utilisé pour représenter un grand nombre de phénomènes réels, modèle d'autre part susceptible d'une réalisation concrète simple, peu coûteuse, et facilement manipulable.

A) PREMIERE SITUATION.

Une urne contient 43 billes, dont chacune appartient à l'une ou l'autre de deux catégories P_1, P_2 (par exemple : billes rouges et billes vertes, ou bien billes en bois et billes en terre, etc.). Dans chaque catégorie, on distingue diverses grosseurs : les billes « petites », les « moyennes », les « grosses ». La répartition des 43 billes est donnée dans le tableau 1.

	Catégorie P_1	P_2	Totaux
petites	6	12	18
moyennes	9	7	16
grosses	3	6	9
Totaux	18	25	43

Tableau 1. Répartition des billes dans l'urne (première situation):

Pour une raison quelconque, il se trouve que nous ne pouvons pas distinguer les billes de la catégorie P_1 de celles de la catégorie P_2 (par exemple : les billes sont de couleurs différentes, mais l'observateur est daltonien ; ou encore les billes en bois et les billes en terre ont été peintes de la même couleur, de sorte que leur aspect extérieur seul ne permet pas de les distinguer).

Sortons de l'urne une bille prise au hasard. Nous pouvons constater qu'elle est petite, moyenne ou grosse, mais nous ne savons pas si elle appartient à P_1 ou à P_2. Or nous nous trouvons dans l'obligation de la classer soit en P_1, soit en P_2. Dans ces conditions, nous ne pouvons pas espérer opérer ce classement à coup sûr : il y a forcément un risque non nul de classer en P_1 une bille appartenant en fait à la catégorie P_2, ou de classer en P_2 une bille appartenant à P_1. Nous pourrons donc :
— adopter une *règle de décision,*
— et calculer le *risque d'erreur* de classement lié à cette règle de décision.

Le choix d'une règle de décision est relativement arbitraire (divers critères peuvent présider à ce choix). Une règle étant choisie, le risque d'erreur qui lui est lié peut être calculé, de sorte que, en présence de plusieurs règles possibles, nous pourrons par exemple retenir celle correspondant au plus petit risque.

Exemple n° 1 : nous décidons d'ignorer délibérément l'information dont nous disposons (connaissance des proportions dans l'urne, constatation de la grosseur de la bille sortie), et nous effectuons un classement au hasard, c'est-à-dire que quelle que soit la bille, nous avons une chance sur deux de la classer en P_1, et une chance sur deux de la classer en P_2. Une telle règle de décision nous permet de calculer ainsi le risque d'erreur :

$$\begin{aligned}\text{Risque d'erreur} &= \text{Prob(bille } P_1 \text{ classée } P_2) + \text{Prob(bille } P_2 \text{ classée } P_1) \\ &= \text{Prob}(P_1)\,\text{Prob}(P_2/P_1) + \text{Prob}(P_2)\,\text{Prob}(P_1/P_2) \\ &= \frac{18}{43} \times \frac{1}{2} + \frac{25}{43} \times \frac{1}{2} = \frac{1}{2} = 0{,}50\end{aligned}$$

Exemple n° 2 : nous décidons maintenant de tenir compte de toute l'information disponible (proportions dans l'urne et grosseur de la bille sortie). Constatons d'abord que, pour chaque grosseur de bille, la répartition entre les deux catégories n'est pas uniforme. Nous décidons alors d'attribuer la bille à celle des deux catégories où cette grosseur est la plus fréquente.

Explicitons cette règle :

— si nous sortons une petite bille, la probabilité qu'elle provienne de P_1 est $\frac{6}{18}$; la probabilité qu'elle provienne de P_2 est $\frac{12}{18}$; nous décidons de la classer en P_2.

— si nous sortons une bille moyenne, la probabilité qu'elle provienne de P_1 est $\frac{9}{16}$; la probabilité qu'elle provienne de P_2 est $\frac{7}{16}$; nous décidons de la classer en P_1.

— enfin si nous sortons une grosse bille, la probabilité qu'elle provienne de P_1 est $\frac{3}{9}$; la probabilité qu'elle provienne de P_2 est $\frac{6}{9}$; nous décidons de la classer en P_2.

En résumé, la règle de décision adoptée se traduit ainsi :
Classer en P_1 une bille moyenne ; classer en P_2 une bille petite ou une bille grosse.

Quel est le risque d'erreur lié à cette règle de décision ?

$$\text{Risque d'erreur} = \text{Prob(petite bille } P_1) + \text{Prob(bille moyenne } P_2) + \text{Prob(grosse bille } P_1)$$

$$= \frac{6}{43} + \frac{7}{43} + \frac{3}{43} = \frac{16}{43} = 0{,}37$$

Ce risque étant inférieur au risque lié à la règle de décision de l'exemple n° 1, nous choisirons de préférence la règle de décision de l'exemple n° 2 :

B) DEUXIEME SITUATION.

Dans une urne il y a des billes en bois (catégorie P_1) en proportion $\pi_1 = \frac{2}{3}$, et des billes en matière plastique (catégorie P_2), en proportion $\pi_2 = \frac{1}{3}$. Toutes les billes ont même aspect extérieur, et rien ne permet de savoir sans casser la bille, si celle-ci est en bois ou en plastique. Nous disposons d'une balance de précision, et nous savons que, dans chaque catégorie, la masse d'une bille est une variable aléatoire distribuée selon une *loi de Gauss.*

Rappelons que, lorsque nous disons qu'une variable aléatoire continue X est distribuée selon une « loi de Gauss », cela signifie que, connaissant sa « moyenne » (mesurée par son *espérance mathématique* μ) et sa « dispersion » (mesurée par son *écart-type* σ), nous pouvons exprimer la *densité de probabilité* correspondant à toute valeur numérique réelle x par la quantité :

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \quad e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

La fonction $x \rightarrow p(x)$, dont l'étude est facile, admet une représentation graphique du type suivant (voir figure 1).

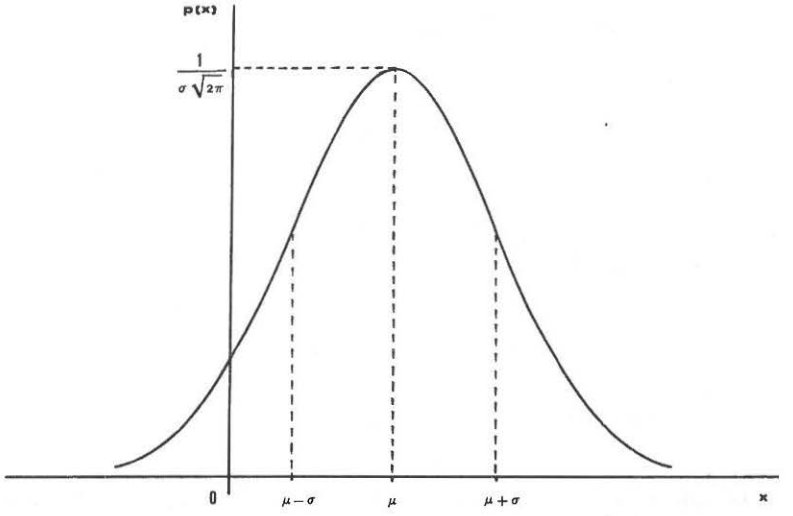


FIG. 1 : Représentation graphique de la densité de probabilité d'une distribution de Gauss

La courbe (dite « courbe en cloche »), symétrique par rapport à un axe vertical d'abscisse μ, admet l'axe des abscisses pour asymptote à gauche et à droite. Elle a une forme d'autant plus « pointue » que la valeur de σ est faible.

Revenons à nos billes, en supposant que par exemple la masse moyenne des billes de la catégorie P_1 est $\mu_1 = 10$ g, et l'écart-type $\sigma_1 = 1$ g. Pour la catégorie P_2, les valeurs de ces paramètres sont respectivement $\mu_2 = 12$ g et $\sigma_2 = 2$ g. Par conséquent, les densités de probabilité ne sont pas identiques dans les deux groupes :

— dans le groupe P_1 : $p_1(x) = \frac{1}{\sqrt{2\pi}} \; e^{-\frac{1}{2}(x-10)^2}$

— dans le groupe P_2 : $p_2(x) = \frac{1}{2\sqrt{2\pi}} \; e^{-\frac{1}{2}(\frac{x-12}{2})^2}$

Ces deux fonctions sont représentées simultanément dans la figure 2.

Comme dans la première situation, nous cherchons à établir une règle de décision tenant compte, partiellement ou totalement, de l'information disponible (c'est-à-dire : les proportions de P_1 et P_2 dans l'urne, la loi de distribution de la variable X, masse d'une bille, dans chaque catégorie, et la masse x de la bille effectivement sortie de l'urne). Là encore nous aurons plusieurs possibilités de règles de décision.

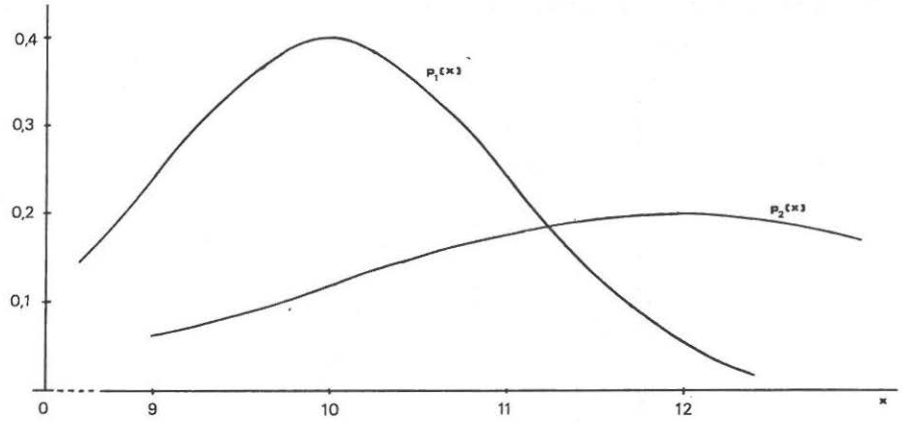


FIG. 2 : Représentation graphique de $p_1(x)$ et de $p_2(x)$

Exemple n° 1 : Nous voulons apprécier la *vraisemblance* de l'appartenance de la bille à l'un des groupes plutôt qu'à l'autre, en tenant compte de tous les éléments d'appréciation dont nous disposons : la densité de probabilité de X dans chaque groupe et la proportion, le « poids », du groupe. Intuitivement, la « vraisemblance » d'un groupe, pour une valeur x observée, est fonction directe (croissante) du poids de ce groupe, soit π, et de la densité de probabilité en x, soit p(x). Nous pouvons donc utiliser, pour mesurer cette vraisemblance, la quantité $\pi p(x)$; ainsi, la vraisemblance de P_1 est $\pi_1 p_1(x)$, celle de P_2 est $\pi_2 p_2(x)$. La règle de décision est naturellement : la bille est attribuée au groupe de plus grande vraisemblance. Ainsi, dans le cas numérique proposé, les vraisemblances sont, pour une bille de masse 11 :

— pour P_1 : $\pi_1 p_1(11) = \frac{2}{3} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(11-10)^2} = 0{,}1613$

— pour P_2 : $\pi_2 p_2(11) = \frac{1}{3} \times \frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{11-12}{2}\right)^2} = 0{,}0587$

(Ici l'usage d'une minicalculatrice facilite évidemment les calculs). Notre règle de décision nous conduit donc à classer cette bille dans la catégorie P_1 .

Si nous multipliions les exemples numériques, nous serions conduits chaque fois à un calcul comme ci-dessus, suivi d'une décision de classement. Il sera plus commode d'essayer de définir des zones dans lesquelles on a l'inégalité $\pi_1 p_1(x) \geqslant \pi_2 p_2(x)$, et des zones dans lesquelles on a l'inégalité contraire, autrement dit :

$$\Omega_1 = \{x \mid \pi_1 p_1(x) \geqslant \pi_2 p_2(x)\}$$
$$\Omega_2 = \{x \mid \pi_1 p_1(x) < \pi_2 p_2(x)\}$$

Caractérisons la zone Ω_1 :

$$\frac{2}{3} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-10)^2} \geqslant \frac{1}{3} \frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-12}{2})^2}$$

d'où :

$$4 \geqslant e^{-\frac{1}{2}[(\frac{x-12}{2})^2 - (x-10)^2]}$$

soit finalement :

$$-\frac{3}{4} x^2 + 14\,x - 64 + 2 \operatorname{Log} 4 \geqslant 0$$

Il s'agit d'un problème classique d'étude du signe d'un trinome du second degré, soit $y = -0{,}75x^2 + 14\,x - 61{,}2274$. Ce trinome possède deux racines réelles : $x_0 = 11{,}673$ et $x'_0 = 6{,}994$. Donc :

$$\Omega_1 = \{x \mid 6{,}994 \leqslant x \leqslant 11{,}673\} ;$$

et

$$\Omega_2 = \{x \mid x < 6{,}994 \text{ ou } x > 11{,}673\}.$$

Remarquons que l'une des racines, soit $x_0 = 11{,}673$, est de l'ordre de grandeur des masses moyennes des billes contenues dans l'urne. Au contraire l'autre racine, soit $x'_0 = 6{,}994$, est très en-dessous de cet ordre de grandeur ; le calcul des probabilités, gaussiennes, dans chacune des deux catégories, nous apprend que dans P_1, la probabilité d'observer une masse inférieure à 7 grammes est 2×10^{-3}, et dans P_2 cette probabilité est $6{,}2 \times 10^{-3}$. Dans les deux cas elle est très faible, négligeable. Donc dans la pratique nous sommes assurés d'observer des valeurs de x « autour » de x_0, et la règle de décision choisie plus haut s'énoncera simplement :

— si $x \leqslant 11{,}673$, la bille est classée en P_1 ;
— si $x > 11{,}673$, la bille est classée en P_2.

Cette valeur $x_0 = 11{,}673$, frontière entre les deux zones Ω_1 et Ω_2, est un *seuil de décision*. Le risque d'erreur correspondant à cette règle de décision peut être calculé :

Risque d'erreur

$$= \text{Prob}\,(P_1 \text{ classée } P_2) + \text{Prob}\,(P_2 \text{ classée } P_1)$$
$$= \text{Prob}(P_1) \times \text{Prob}(x > x_0 | P_1) + \text{Prob}(P_2) \times \text{Prob}(x \leqslant x_0 | P_2)$$
$$= \frac{2}{3} \times 0{,}0475 + \frac{1}{3} \times 0{,}4364 = 0{,}1771$$

(dans les calculs ci-dessus, les probabilités gaussiennes ont été déterminées par lecture dans une table numérique de la *fonction de répartition de Gauss ;* les heureux possesseurs d'une minicalculatrice programmable ou d'un ordinateur pourront rendre automatique cette détermination grâce à l'une des formules de calcul approché disponibles).

Exemple n° 2 : Nous remarquons que les deux probabilités d'erreur

$$\text{Prob}(x > x_0 | P_1) = 0{,}0475 \quad \text{et} \quad \text{Prob}(x \leqslant x_0 | P_2) = 0{,}6364$$

sont très disproportionnées. Nous nous posons alors la question suivante : tout en retenant le principe d'une règle de décision simple, basée sur l'existence d'une valeur-seuil x_0, nous souhaiterions que ces deux risques soient égaux ; comment alors choisir x_0 ?

La relation $\text{Prob}(X > x_0 | P_1) = \text{Prob}(X \leqslant x_0 | P_2)$ s'écrit ici :

$$\frac{1}{\sqrt{2\pi}} \int_{x_0}^{+\infty} e^{-\frac{1}{2}(x-10)^2} dx = \frac{1}{2\sqrt{2\pi}} \int_{-\infty}^{x_0} e^{-\frac{1}{2}\left(\frac{x-12}{2}\right)^2} dx$$

soit, sous la forme standardisée :

$$\frac{1}{\sqrt{2\pi}} \int_{t_1}^{+\infty} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t^2} e^{-\frac{t^2}{2}} dt$$

avec $t_1 = x_0 - 10$ et $t_2 = \dfrac{x_0 - 12}{2}$.

Compte tenu des propriétés de la distribution de Gauss standard, l'égalité des deux probabilités nous permet d'affirmer que $t_1 = -t_2$; d'autre part, les relations entre t_1, t_2 et x_0 permettent d'écrire :

$$x_0 = t_1 + 10 = 2t_2 + 12$$

D'où : $t_1 + 10 = -2t_1 + 12$, soit $t_1 = \dfrac{2}{3} = 0{,}667$.

De sorte que : $x_0 = 10{,}667$. La règle de décision correspondant à ce nouveau seuil est :

— si $x \leqslant 10{,}667$, la bille est classée en P_1 ;
— si $x > 10{,}667$, la bille est classée en P_2.

Le risque d'erreur correspondant à cette règle est :

Risque d'erreur

$$= \text{Prob}(P_1) \times \text{Prob}(X > x_0 | P_1) + \text{Prob}(P_2) \times \text{Prob}(X \leqslant x_0 | P_2)$$

$$= \frac{2}{3} \times 0{,}2514 + \frac{1}{3} \times 0{,}2514 = 0{,}2514$$

Ce risque est plus élevé que pour la règle de décision précédente.

En résumé, dans cette deuxième situation, si nous tenons à avoir le plus petit risque « global », nous choisirons de préférence la règle de décision étudiée à l'exemple n° 1. Si nous souhaitons plutôt « équilibrer » les risques, il vaudra mieux utiliser la règle de décision vue à l'exemple n° 2.

C) CONCLUSION

Des quelques exemples qui précèdent — et qui sont loin de représenter toutes les possibilités — retenons :

— qu'il est possible, à condition d'accepter de prendre certains risques, de trouver des règles de décision permettant le classement d'un individu inconnu dans tel ou tel groupe, le risque d'erreur étant éventuellement calculable ;

— qu'il n'existe pas de règle de décision unique, universelle, permettant de faire face à toute situation ;

— mais que l'on peut en forger, à la demande, selon le but poursuivi, selon l'information disponible, selon la stratégie de distribution des risques, comme on forge un outil pour travailler tel matériau dans des circonstances bien déterminées.

2 Lignes directrices de l'analyse discriminante statistique

Une situation dans laquelle nous pouvons utiliser une méthode d'analyse discriminante peut se résumer ainsi. Une population possède des subdivisions que nous appellerons des *groupes* (ce sont des *sous-populations*, des *classes* ou *catégories* diverses d'individus), notées P_1, P_2, ..., P_k, de sorte que tout individu de la population appartient forcément à un et un seul de ces groupes. En tant qu'observateur, nous examinons un individu de la population ; deux situations peuvent alors se présenter :

— ou bien nous possédons sur chaque groupe d'une part, et sur cet individu d'autre part, les informations nécessaires et suffisantes pour nous permettre d'attribuer *avec certitude* cet individu au groupe auquel il appartient effectivement ;

— ou bien nous ne possédons pas ces informations, auquel cas, si nous décidons malgré tout d'attribuer cet individu à un groupe, nous savons qu'il existe un certain *risque d'erreur de classement.*

Les méthodes d'analyse discriminante se proposent de définir des règles de conduite à suivre par un observateur placé dans la deuxième situation : comment l'observateur peut-il utiliser "au mieux" les informations dont il dispose, pour classer des individus dans des groupes, en maîtrisant dans une certaine mesure le risque de commettre des erreurs de classement. Pour fixer les idées, voici quelques exemples de situations susceptibles d'être analysées au moyen d'une méthode d'analyse discriminante.

DISCRIMINATION A BUT DESCRIPTIF.

La définition des groupes P_1, P_2, ..., P_k est imprécise, floue, voire seulement intuitive (races, familles politiques, ...). Par contre, il est relativement aisé de classer la plupart des individus dans tel ou tel groupe, sauf éventuellement certains cas "marginaux" pour lesquels le classement est rendu impossible par l'absence de définition précise. Nous souhaitons trouver un critère simple de définition de ces groupes, basé sur l'observation d'un petit nombre de caractères, critère qui nous permette de retrouver avec le minimum d'erreurs la connaissance, informelle au départ, que nous avons de cette subdivision en groupes.

DISCRIMINATION A BUT DECISIONNEL.

Ici les groupes sont bien définis. Nous nous intéressons au bon classement d'un individu sur lequel nous sommes incomplètement informés. Cette absence d'information complète peut être due à diverses causes :

— l'information existe mais est difficile (ou coûteuse) à obtenir (en biologie par exemple, il arrive que l'information exacte ne puisse être obtenue que moyennant le sacrifice de l'individu) ;

— l'information a existé mais est définitivement perdue (individus incomplets : objets endommagés, restes fossiles, etc.) ;

— l'information n'existe pas encore : prédiction (le succès ou l'échec à l'examen de fin d'année peut-il être pronostiqué en cours d'année à partir de résultats partiels ; le succès ou l'échec d'une thérapeutique peut-il être pronostiqué dès l'entrée à l'hôpital, selon les conditions physiques, l'origine, etc... du malade à ce moment-là ?).

Les exercices introductifs du paragraphe précédent ont montré que, dans les situations où l'on connaît la distribution des probabilités de l'information dont on dispose (le "profil" de l'individu, constitué d'un ou plusieurs caractère(s) qualitatif(s) ou quantitatif(s)), la discrimination peut être basée sur une règle de décision appliquée au résultat d'un calcul de probabilité. Cependant ce cas se présente rarement dans la pratique : il est presque exceptionnel de connaître ces distributions de probabilités. Beaucoup plus fréquemment nous ne pouvons avoir de ces distributions qu'une connaissance "statistique", obtenue à partir d'*échantillons* extraits des divers groupes P_1, P_2, ..., P_k. C'est pourquoi dans la suite de ce texte nous n'envisageons que cette situation, sans pour autant épuiser le sujet.

Dans le but d'illustrer numériquement les diverses techniques que nous allons décrire, nous donnons ci-dessous (tableau 2) un exemple concret de cas susceptible d'occasionner une analyse discriminante. Il s'agit de mensurations faites sur des crânes de petits mammifères fossiles (Oreodontes) d'Amérique. Les données concernent six genres d'Oreodontes, aujourd'hui disparus. Sur chaque individu on a mesuré deux variables X et Y (voir légende du tableau). Nous analyserons des sous-tableaux de ce tableau initial.

Les techniques de l'analyse discriminante reposeront toujours sur la recherche de bons moyens pour, à la fois :

— homogénéiser le plus possible chaque groupe pris séparément ;
— distinguer au maximum les groupes les uns des autres.

Une notion essentielle est donc celle de *dispersion* : la dispersion propre à chaque groupe (c'est-à-dire propre à l'ensemble des positions des individus d'un groupe les uns par rapport aux autres) doit être la plus faible possible, et en tout cas faible par rapport à la dispersion entre les groupes. Il nous faudra donc être en mesure d'apprécier de manière quantitative la dispersion en général, et de pouvoir comparer la dispersion interne aux divers groupes et la dispersion entre les groupes. Les méthodes statistiques font le plus souvent appel, pour mesurer la dispersion d'un ensemble de données, à la *variance* de cet ensemble de données. Nous nous attarderons ci-dessous sur cette notion, qui sera fondamentale par la suite.

E_1		E_2		E_3		E_4		E_5		E_6	
X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
45	16	37	17	47	26	78	35	42	15	58	26
46	17	43	14	42	26	77	37	48	18	52	27
48	19	43	19	40	22	65	30	45	18	50	28
46	19	42	17	46	22	74	31	48	17	52	29
45	15	39	12	46	24	65	31	46	16	60	33
51	19	39	15	42	26	70	34	51	21	61	28
47	16	40	18	43	23	69	28	46	17	54	30
48	18	34	16	44	23	67	31	50	18	65	32
47	18	35	15	44	25	65	34	46	16	55	32
50	17	45	17	47	27	64	28	48	15	64	26
48	19	33	15	47	27	68	32	47	17	56	28
49	18	42	13					49	18		
49	17							43	15		
49	19							47	19		
								46	18		

Tableau 2. Mensuration sur des Oreodontes fossiles.

Variables : X = largeur de la boîte crânienne à la suture pariétale squamosale (mm)
Y = longueur maximum de la bulle tympanique (mm)

Genres : E_1 = *Merychoidodon culbertsoni* ; E_2 = *Prodesmatochoerus meeki* ;
E_3 = *Subdesmatochoerus sp.* ; E_4 = *Megoreodon gigas loomisi* ;
E_5 = *Oreodon osborni* ; E_6 = *Desmatochoerus hatcheri.*

Un ensemble de données numériques $\{x_1, x_2, \ldots, x_n\}$ peut être assez bien résumé par la connaissance de deux indices statistiques principaux, qui sont :

— la **moyenne** $\overline{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$, qui est un indice « de position » donnant l'ordre de grandeur des données ;

— la **variance** $\hat{s}^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x})^2$, qui est un indice « de dispersion » mesurant la plus ou moins grande dispersion des données x_j de part et d'autre de leur moyenne : plus la variance est faible, plus les données sont regroupées au voisinage de la valeur moyenne $\overline{x}$ (dans la notation $\hat{s}^2$, l'accent circonflexe sur le s pourra se lire « chapeau »).

Ces deux indices $\overline{x}$ et $\hat{s}^2$, calculés à l'aide de données provenant d'un échantillon, sont des « estimations» des indices correspondants concernant l'ensemble de la population, indices qualifiés de « théoriques », notés respectivement μ(moyenne théorique, ou espérance mathématique) et σ^2(variance théorique), dont la connaissance numérique est le plus souvent impossible.

Le lecteur pointilleux notera que parfois la variance de l'échantillon est définie ainsi :

$$s^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2$$

la seule différence étant le dénominateur n au lieu de $n-1$. Il est certain que ces deux « variances » sont numériquement différentes, et néanmoins liées par une relation simple :

$$(n-1)\,\hat{s}^2 = n\,s^2$$

Elles constituent toutes les deux un moyen de mesurer la dispersion d'un ensemble fini de données (d'un échantillon). Mais, en tant qu'estimations de la variance théorique σ^2, des raisons théoriques que nous ne développerons pas ici font que $\hat{s}^2$ est une « meilleure » estimation de σ^2 que s^2.

Notons d'autre part que la racine carrée (positive) de la variance est souvent utilisée aussi pour mesurer la dispersion de la variable ; ce nouvel indice est appelé *l'écart-type*. Nous aurons par la suite l'occasion de l'utiliser :

$$\text{écart-type} = \hat{s} = \sqrt{\hat{s}^2} = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})^2}$$

Remarque : selon sa définition, la variance apparaît comme une moyenne particulière : la moyenne des carrés des écarts entre les données individuelles et la moyenne du groupe. L'utilisation du dénominateur $n-1$ au lieu de n peut se justifier par le fait que nous mesurons la dispersion non pas autour d'un point fixe indépendant du groupe, mais autour d'un point (le point moyen d'abscisse $\bar{x}$) dépendant du groupe. On traduit cette dépendance en utilisant le vocable *degré de liberté* pour désigner la quantité $n-1$.

Dans ce qui suit, nous utiliserons exclusivement la première définition donnée ci-dessus de la variance. Dans l'expression de cette dernière, le numérateur, soit $\sum_{j=1}^{n} (x_j - \bar{x})^2$, sera appelé *la variation* de l'ensemble des données (le dénominateur étant, comme dit plus haut, *le degré de liberté* de cet ensemble de données).

Dans la pratique du calcul numérique, on aura souvent avantage à utiliser une formule équivalente à la formule de définition, et facile à établir (développer le carré de la différence $x_j - \bar{x}$, puis effectuer la sommation) :

$$\hat{s}^2 = \frac{1}{n-1} \sum_{j=1}^{n} x_j^2 - \frac{n}{n-1} \bar{x}^2$$

$$\hat{s}^2 = \frac{1}{n-1} \left[\sum_{j=1}^{n} x_j^2 - \frac{1}{n} \left(\sum_{j=1}^{n} x_j \right)^2 \right]$$

Lorsque nous aurons affaire à une seule variable (X), une situation où l'analyse discriminante pourra être utilisée nous conduira à présenter le calcul des indices statistiques nécessaires (moyennes, variances, écarts-types) sous la forme d'un tableau du type du tableau 3.

Groupe	E_1	...	E_i	...	E_k	Ensemble
Effectif	n_1		n_i		n_k	$n = n_1 + n_2 + \ldots + n_k$
Somme des données			$\sum_{j=1}^{n_i} x_{ij}$			$\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$
Somme des carrés			$\sum_{j=1}^{n_i} x_{ij}^2$			$\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}^2$
Variation			$\sum_{j=1}^{n_i} x_{ij}^2 - \frac{1}{n_i} (\sum_{j=1}^{n_i} x_{ij})^2$			$\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}^2 - \frac{1}{n} (\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij})^2$
Moyenne			$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$			$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$
Variance			$\hat{s}_i^2 = \frac{1}{n_i - 1} (\sum_{j=1}^{n_i} x_{ij}^2 - n_i \bar{x}_i^2)$			$\hat{s}^2 = \frac{1}{n-1} (\sum_i \sum_j x_{ij}^2 - n\bar{x}^2)$
Ecart-type			$\hat{s}_i = \sqrt{\hat{s}_i^2}$			$\hat{s} = \sqrt{\hat{s}^2}$

Tableau 3 : Présentation du calcul des indices statistiques concernant les groupes E_1, $E_2, \ldots, E_k$ et l'ensemble $E_1 \cup E_2 \cup \ldots \cup E_k$.

La variance générale $\hat{s}^2$ ne peut pas être calculée directement à partir des variances des divers échantillons (alors que la moyenne générale est calculable à partir des moyennes des échantillons : $\bar{x} = n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_k\bar{x}_k$). Mais nous pourrons utiliser une relation intéressante concernant la variation générale soit $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$.

Remarquons en effet que

$$(x_{ij} - \bar{x})^2 = (x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x})^2$$

$$(x_{ij} - \bar{x})^2 = (x_{ij} - \bar{x}_i)^2 + 2(x_{ij} - \bar{x}_i)(\bar{x}_i - \bar{x}) + (\bar{x}_i - \bar{x})^2$$

Donc :

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2+2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)(\overline{x}_i-\overline{x})$$

$$+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{x}_i-\overline{x})^2$$

Or :

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{x}_i-\overline{x})^2=\sum_{i=1}^{k}n_i\,(\overline{x}_i-\overline{x})^2\quad;$$

et

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)(\overline{x}_i-\overline{x})=\sum_{i=1}^{k}[(\overline{x}_i-\overline{x})\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)]=0$$

$$\text{car chaque }\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)\text{ est nul.}$$

Par conséquent :

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2+\sum_{i=1}^{k}n_i(\overline{x}_i-\overline{x})^2$$

Au second membre, les deux termes peuvent être interprétés comme des variations. Le premier de ces deux termes est la somme des variations propres aux divers échantillons :

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2=\sum_{i=1}^{k}(n_i-1)\,\hat{s}_i^2\quad;$$

si nous voulons calculer une « moyenne » des variances des divers échantillons, nous pouvons exprimer cette moyenne sous la forme :

$$\Big(\sum_{i=1}^{k}(n_i-1)\,\hat{s}_i^2\Big)\Big/\sum_{i=1}^{k}(n_i-1)=\frac{1}{n-k}\sum_{i=1}^{k}(n_i-1)\,\hat{s}_i^2$$

Cette dernière expression, qui est donc une « moyenne » des variances des k échantillons, est appelée *la variance intra-groupes,* notée $\hat{s}^2_w$

$$\hat{s}^2_w = \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1)\, \hat{s}^2_i$$

Le numérateur $\sum_{i=1}^{k} (n_i - 1)\, \hat{s}^2_i$ ou $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2$ est la *variation intra-groupes ;* le dénominateur $n-k$ est le *degré de liberté intra-groupes.*

Quant au second terme $\sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2$, de la variation générale, il peut être interprété ainsi : $\overline{x}$, qui est la moyenne de l'ensemble des données, est aussi la moyenne des moyennes des échantillons ; il s'agit là d'une moyenne pondérée (chaque $\overline{x}_i$ étant affecté d'un poids proportionnel à l'effectif n_i). Donc $\sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2$ peut être considérée comme la variation entre les moyennes des divers groupes ; nous l'appellerons la *variation inter-groupes.* La variance correspondante (variance entre les moyennes) est :

$$\hat{s}^2_B = \frac{1}{k-1} \sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2$$

appelée *variance inter-groupes ;* le dénominateur $k-1$ est le *degré de liberté inter-groupes.*

Nous résumerons ainsi ce qui vient d'être dit, en remarquant que $n-1 = (n-k) + (k-1)$:

— la variation générale est la somme de la variation inter-groupes et de la variation intra-groupes ;

— le degré de liberté général est la somme du degré de liberté inter-groupes et du degré de liberté intra-groupes.

Exemple : Pour illustrer numériquement les calculs décrits ci-dessus, travaillons sur le sous-tableau du tableau 2 constitué uniquement de la variable X sur les groupes E_2 et E_3. A partir des données de ces deux groupes, nous avons établi le tableau 4, sur le modèle du tableau 3.

groupe	E_2	E_3	ensemble
effectif	12	11	23
somme des données	472	488	960
somme des carrés	18732	21708	40440
variation	166,666670	58,545450	370,434780
moyenne	39,333333	44,363636	41,739130
variance	15,151515	5,854545	16,837945
écart-type	3,8925	2,4196	4,1034

Tableau 4. Exemple d'application du schéma du tableau 3 aux groupes E_2 et E_3.

Revenons à l'analyse discriminante. Nous pourrons d'autant mieux discriminer les groupes entre eux que les conditions suivantes seront réalisées :

— les groupes sont chacun le moins dispersé possible ;
— les moyennes des groupes sont le plus dispersées possible.

Nous pouvons maintenant traduire facilement ces conditions en termes de variances :

— la variance intra-groupes doit être la plus faible possible ;
— la variance inter-groupes doit être la plus élevée possible.

Ainsi les groupes seront d'autant mieux discriminés que le rapport

$$f = \frac{\text{variance inter-groupes}}{\text{variance intra-groupes}} = \frac{\hat{s}^2_B}{\hat{s}^2_W}$$

est élevé. Cet indice f peut éventuellement être utilisé comme mesure du "pouvoir discriminant" d'une variable. Certains auteurs utilisent plutôt dans ce but le rapport $g = f/(1+f)$, qui est toujours compris entre 0 et 1 : plus g est proche de 1, plus le pouvoir discriminant est élevé.

APPLICATION NUMÉRIQUE.

A partir des résultats contenus dans le tableau 4, il est facile de vérifier que, si nous voulons discriminer les groupes E_2 et E_3, nous nous baserons sur les valeurs suivantes :

variance inter-groupes $= \hat{s}^2_B = 145,22$;

variance intra-groupes $= \hat{s}^2_W = 10,72$;

$f = 13,54$; $g = 0,93$.

Exercice : Faire les mêmes calculs entre les groupes E_1 et E_5, toujours sur la variable X . Vérifier que l'on trouve :

$$\hat{s}^2_B = 6{,}05 \quad ; \quad \hat{s}^2_W = 4{,}56 \quad ; \quad f = 1{,}33 \quad ; \quad g = 0{,}57 \ .$$

La variable X discrimine beaucoup mieux les groupes E_2 et E_3 que les groupes E_1 et E_5.

Pour conclure, disons que nous aurons un double souci en analyse discriminante basée sur un ou plusieurs caractères quantitatifs :

— d'une part, nous assurer que le pouvoir discriminant des variables est suffisamment grand ;

— d'autre part, trouver une règle de décision (d'attribution d'un individu à un groupe) pour laquelle nous saurons maîtriser le risque d'erreur de classement.

3 Discrimination statistique entre deux groupes sur la base d'une variable

La population dont on examine les individus est subdivisée en k groupes $P_1, P_2, \ldots, P_k$. Sur chaque individu appartenant à cette population, nous nous intéressons à la valeur prise par une variable aléatoire X . Pour un individu dont le groupe d'appartenance ne nous est pas connu, X prend une certaine valeur numérique x . Si nous connaissons :

— les proportions respectives $\omega_1, \omega_2, \ldots, \omega_k$ des k groupes dans l'ensemble de la population,

— la distribution des probabilités de X dans chaque groupe,

alors il nous est possible de déterminer (théorème de Bayes), pour chaque groupe P_i , la probabilité que cet individu appartienne à P_i . La règle de décision que nous adoptons, pour attribuer l'individu à tel ou tel groupe, est alors la suivante : on attribue l'individu d'origine inconnue à celui des groupes pour lequel la probabilité ainsi calculée est la plus grande (règle assortie d'une convention en cas d'égalité entre deux probabilités).

Mais dans la plupart des situations pratiques, nous ne connaissons rien de ces informations théoriques. Nous devons nous baser uniquement sur des informations statistiques, obtenues par échantillonnage. Pour chaque groupe P_i, nous disposons d'un échantillon E_i, de taille n_i,

à partir duquel nous pouvons estimer μ_i et σ_i^2, respectivement moyenne et variance de X dans le groupe P_i :

μ_i est estimé par $\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$

σ^2 est estimé par $\hat{s}_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$

Sur un individu d'origine inconnue, la valeur mesurée de la variable X est un nombre réel x. Nous pourrions chiffrer l'"éloignement" de cette valeur x, par rapport au groupe E_i, par exemple au moyen de l'écart absolu entre x et la moyenne du groupe, soit $\bar{x}_i$. En fait, cet écart $|x - \bar{x}_i|$ a un défaut : il ne tient pas compte de la dispersion du groupe E_i. Ainsi, sachant que j'habite à 5 km du centre d'une agglomération, cette distance absolue 5 n'aura pas la même signification si cette agglomération est une grande ville (dans ce cas il y a de fortes chances que j'habite dans cette agglomération), ou si c'est un village (dans ce cas, la distance 5 signifierait plutôt que j'habite un village voisin).

Nous pondèrerons donc l'écart absolu par une fonction inverse de la dispersion : en l'occurrence, par l'inverse de l'écart-type. Nous chiffrerons donc l'"éloignement" d'une valeur x par rapport au groupe E_i de moyenne $\bar{x}_i$, d'écart-type $\hat{s}_i$, au moyen de la quantité $\frac{|x - \bar{x}_i|}{\hat{s}_i}$. Dans la pratique des calculs, l'utilisation du carré de cette quantité, soit $\frac{(x - \bar{x}_i)^2}{\hat{s}_i^2}$, sera plus commode ; comme il s'agit d'une fonction directe de la quantité précédente, la hiérarchie des éloignements sera respectée.

Nous nous étendrons ci-après uniquement sur le cas de deux groupes P_1, P_2, ce qui va nous permettre de montrer l'existence de "valeurs-seuils" rendant commode la séparation entre les groupes. Notons d'abord que la discrimination n'aura de sens que si μ_1 et μ_2 sont distinctes, c'est-à-dire si $\bar{x}_1$ et $\bar{x}_2$ sont suffisamment éloignées l'une de l'autre pour enlever tout doute au sujet de la différence entre μ_1 et μ_2. Pour apprécier l'"éloignement" entre $\bar{x}_1$ et $\bar{x}_2$, nous utiliserons un indice "symétrique", c'est-à-dire qui reste égal à lui-même quand on intervertit $\bar{x}_1$ et $\bar{x}_2$. Ainsi, la quantité f dont nous avons parlé plus haut peut être choisie comme un tel indice :

$$f = \frac{\text{variance inter-groupes}}{\text{variance intra-groupes}} = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2}{\hat{s}^2}$$

$$\text{avec } \bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} \text{ et } \hat{s}^2 = \frac{(n_1 - 1)\,\hat{s}_1^2 + (n_2 - 1)\,\hat{s}_2^2}{n_1 + n_2 - 2}$$

La valeur de f , qui permet d'apprécier le "pouvoir discriminant" de la variable X , doit être la plus grande possible. Le lecteur habitué de la statistique inductive pourra préciser cette notion de "plus grande possible" en se référant, lorsque les conditions voulues seront réunies, à la *variable F de Fisher-Snedecor.*

La règle de décision que nous avons adoptée plus haut, basée sur l'utilisation de l'indice de distance $\frac{(x-\overline{x}_i)^2}{\hat{s}_i^2}$, nous conduit *dans le cas de deux groupes* à nous intéresser au signe de la quantité :

$$y = \frac{(x-\overline{x}_2)^2}{\hat{s}_2^2} - \frac{(x-\overline{x}_1)^2}{\hat{s}_1^2}$$

Si $y > 0$, l'individu est attribué au groupe P_2 ;
si $y \leqslant 0$, il est attribué au groupe P_1 .

Cette quantité y étant une fonction trinome du second degré de x , nous pouvons prévoir que son signe dépendra de la position de x par rapport à ses racines éventuelles.

Ce polynome y s'écrit, sous forme développée :

$$y = \frac{1}{\hat{s}_1^2 \hat{s}_2^2} \left[(\hat{s}_1^2 - \hat{s}_2^2)\, x^2 - 2\,(\hat{s}_1^2 \overline{x}_2 - \hat{s}_2^2 \overline{x}_1)\, x + (\hat{s}_1^2 \overline{x}_2^2 - \hat{s}_2^2 \overline{x}_1^2) \right]$$

Nous étudierons séparément deux cas :

— le cas où y est un vrai polynome du second degré ($\hat{s}_1^2 \neq \hat{s}_2^2$) : discrimination dite "quadratique" ;

— le cas où y est dégénéré en un polynome du premier degré ($\hat{s}_1^2 = \hat{s}_2^2$) : discrimination dite "linéaire".

A) DISCRIMINATION QUADRATIQUE : $\hat{s}_1^2 \neq \hat{s}_2^2$.

Le discriminant réduit de y est $\frac{(\overline{x}_1 - \overline{x}_2)^2}{\hat{s}_1^2 \hat{s}_2^2}$;

nous sommes donc assurés de l'existence de deux racines réelles distinctes dès lors que $\overline{x}_1 \neq \overline{x}_2$ (ce qui est en principe toujours le cas dans ce type d'analyse : voir ci-dessous). Les deux racines sont :

$$x_0 = \frac{\hat{s}_1 \overline{x}_2 + \hat{s}_2 \overline{x}_1}{\hat{s}_1 + \hat{s}_2}$$

$$x_0' = \frac{\hat{s}_1 \overline{x}_2 - \hat{s}_2 \overline{x}_1}{\hat{s}_1 - \hat{s}_2}$$

La première, x_0, est entre $\overline{x}_1$ et $\overline{x}_2$; elle partage le segment joignant ces deux moyennes dans le rapport $-\hat{s}_1/\hat{s}_2$.

La seconde, x_0', est à l'extérieur du segment joignant les deux moyennes ; elle partage ce segment dans le rapport $+ \hat{s}_1/\hat{s}_2$.

Ces deux racines partagent la droite réelle en trois zones :

— l'une est l'intervalle entre x_0' et x_0 ; dans cette zone le trinôme y est du signe de $\hat{s}_2^2 - \hat{s}_1^2$;

— les deux autres sont les demi-droites à gauche et à droite de cet intervalle ; dans chacune de ces zones le trinôme y est du signe de $\hat{s}_1^2 - \hat{s}_2^2$.

La connaissance des deux quantités x_0 et x_0', et du signe de $\hat{s}_2^2 - \hat{s}_1^2$, suffit donc pour conclure et attribuer l'individu x soit à P_1, soit à P_2.

Remarque : dans la pratique, seule la valeur x_0 jouera effectivement ce rôle de frontière entre deux zones de décision ; en effet, la valeur x_0' étant trop éloignée aussi bien de $\overline{x}_1$ que de $\overline{x}_2$, il sera "improbable" d'observer des valeurs situées au-delà de x_0' (voir l'inégalité de Tchébycheff).

B) DISCRIMINATION LINÉAIRE : $\hat{s}_1^2 = \hat{s}_2^2 = \hat{s}^2$.

Le polynome y est du premier degré :

$$y = 2 \frac{\overline{x}_1 - \overline{x}_2}{\hat{s}^2} \left(x - \frac{\overline{x}_1 + \overline{x}_2}{2}\right)$$

Sa racine est $x_0 = \dfrac{\overline{x}_1 + \overline{x}_2}{2}$, milieu de l'intervalle des moyennes. Nous pouvons dire encore ici que x_0 partage l'intervalle des moyennes dans le rapport $- \hat{s}_1/\hat{s}_2$ (ici égal à -1).

C) CONCLUSION. Si nous voulons discriminer deux groupes P_1, P_2, sur la base d'une seule variable X , à partir d'échantillons E_1, E_2 (les deux moyennes de ces échantillons étant distinctes : $\overline{x}_1 \neq \overline{x}_2$), nous procèderons ainsi :

1°) Eventuellement, à titre indicatif nous calculerons le rapport f et l'indice $g = \dfrac{f}{1+f}$, pour apprécier le "pouvoir discriminant" de la variable X . Signalons au passage que, *dans le cas de deux groupes*, l'expression de f est équivalente à l'expression suivante, dont le calcul est plus simple :

$$f = \frac{n_1 n_2 (n_1+n_2-2)}{n_1 + n_2} \quad \frac{(\overline{x}_1-\overline{x}_2)^2}{(n_1-1)\,\hat{s}_1^2 + (n_2-1)\,\hat{s}_2^2}$$

2°) Nous déterminerons aussi les abscisses des points qui partagent l'intervalle des deux moyennes $\overline{x}_1$, $\overline{x}_2$, dans le rapport $\pm\ \hat{s}_1/\hat{s}_2$, c'est-à-dire les valeurs x_0 et x_0' telles que :

— si $\hat{s}_1^2 \neq \hat{s}_2^2$

$$\begin{cases} x_0 = \dfrac{\hat{s}_1\overline{x}_2 + \hat{s}_2\overline{x}_1}{\hat{s}_1 + \hat{s}_2} \\ x_0' = \dfrac{\hat{s}_1\overline{x}_2 - \hat{s}_2\overline{x}_1}{\hat{s}_1 - \hat{s}_2} \end{cases}$$

— si $\hat{s}_1^2 = \hat{s}_2^2$

$$\begin{cases} x_0 = \dfrac{\overline{x}_1 + \overline{x}_2}{2} \\ x_0' \text{ infini} \end{cases}$$

A partir de x_0 et x_0', nous délimiterons les zones d'attribution respectivement à P_1 et P_2.

3°) Pour estimer les deux risques de mauvais classement, nous appliquerons ces zones d'attribution à tous les individus provenant d'un échantillon E_1' extrait de P_1, et d'un échantillon E_2' extrait de P_2. Dans cette opération, il est déconseillé d'utiliser les échantillons E_1 et E_2 dont les données ont servi à calculer les valeurs-frontières x_0 et x_0'.

Application numérique. Reprenons la discrimination entre les genres P_2 (*Prodesmatochoerus meeki*) et P_3 (*Subdesmatochoerus sp.*). Nous avons déterminé plus haut $f = 13{,}54$ et $g = 0{,}93$, qui dénotent un pouvoir discriminant très élevé. A partir des moyennes et des variances contenues dans le tableau 4, nous calculons x_0 et x_0' :

$$x_0 = 42{,}4353 \simeq 42{,}4 \quad ; \quad x_0' = 52{,}6272 \simeq 52{,}6$$

La zone d'attribution au genre P_2 est :

$$\Omega_2 = \{x \mid x \leqslant 42{,}4 \text{ ou } x > 52{,}6\} \quad ;$$

la zone d'attribution au genre P_3 est :

$$\Omega_3 = \{x \mid 42{,}4 < x \leqslant 52{,}6\}$$

Dans la pratique l'observation d'un individu tel que $x > 52{,}6$, appartenant cependant à P_2 ou P_3, sera forcément un événement exceptionnel, un "monstre". De sorte que la seule frontière vraiment opérationnelle est $x_0 = 42{,}4$; ainsi nous considèrerons que

$$\Omega_2 = \{x \mid x \leqslant 42{,}4\} \quad \text{et} \quad \Omega_3 = \{x \mid x > 42{,}4\} .$$

4. Discrimination de deux groupes à partir de deux variables

Le principe de la discrimination entre k groupes E_1, E_2, ..., E_k, basée sur p variables X_1, X_2, ..., X_p, consistera à se ramener à la discrimination basée sur *une* variable, cette dernière variable étant fonction des p variables mesurées. On l'appellera une *fonction discriminante*. On la construira de façon à répondre à certains critères qui seront généralement les suivants :

— cette fonction doit être simple à exprimer et à calculer (d'où le choix le plus courant d'une fonction *linéaire*) ;
— cette fonction doit avoir le pouvoir discriminant le plus élevé possible (ce qui nous amènera à rechercher une fonction maximisant le rapport $f = \hat{s}^2_B / \hat{s}^2_W$).

Nous n'étudierons pas ici le cas général (k groupes, p variables), dont l'exposé serait trop long et dépasserait le cadre de l'initiation, auquel nous nous limitons ici. Nous renvoyons le lecteur intéressé, à l'ouvrage de ROMEDER (B.G. [12]).

Nous exposerons une méthode consistant à discriminer les groupes deux à deux, sur la base de deux variables X et Y . La restriction à deux variables aura le mérite d'une représentation géométrique facile. Partant d'un tableau de données du type du tableau 2, nous établissons un tableau ressemblant au tableau 3, avec quelques différences dues à la considération simultanée de deux variables, au lieu d'une seule : à chaque groupe E_i correspondent deux colonnes, une pour la variable X , une pour la variable Y ; de plus, seront adjointes au tableau trois lignes qui contiendront respectivement la *somme des produits,* la *covariation*, et la *covariance* des deux variables, dans le groupe considéré (voir tableau 5). La définition et l'usage de ces concepts sont rappelés ci-dessous.

Si nous voulons calculer la moyenne et la variance d'une combinaison linéaire de deux variables aléatoires X et Y , soit $Z = aX + bY$, nous pouvons appliquer les formules suivantes :

$$\bar{z} = a\bar{x} + b\bar{y} \quad ; \quad \hat{s}^2(z) = a^2\hat{s}^2(x) + 2ab\,\hat{s}(x,y) + b^2\,\hat{s}^2(y)$$

où $\bar{x}$, $\bar{y}$, $\bar{z}$ sont les moyennes de X, Y, Z dans l'échantillon considéré ; $\hat{s}^2(x)$, $\hat{s}^2(y)$, $\hat{s}^2(z)$ sont les variances ; $\hat{s}(x,y)$ est la *covariance* de X et Y dans l'échantillon, définie ainsi :

$$\hat{s}(x,y) = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})\,(y_j - \bar{y})$$

On obtient facilement une formule équivalente :

$$\hat{s}(x,y) = \frac{1}{n-1} \left(\sum_{j=1}^{n} x_j y_j - n \bar{x} \bar{y} \right)$$

qui fait ressortir le rôle important joué, dans le calcul numérique, par la *somme des produits* :

$$\sum_{j=1}^{n} x_j y_j .$$

La quantité $\sum_{j=1}^{n} (x_j - \bar{x})(y_j - \bar{y}) = \sum_{j=1}^{n} x_j y_j - n \bar{x} \bar{y}$ est appelée la *covariation* de X et Y dans l'échantillon.

Groupe	E_1		...	E_i		...	E_k	
Variable	X	Y		X	Y		X	Y
Effectif	n_1			n_i			n_k	
Somme des données				$\sum_{j=1}^{n_i} x_{ij}$	$\sum_{j=1}^{n_i} y_{ij}$			
Somme des carrés				$\sum_{j=1}^{n_i} x_{ij}^2$	$\sum_{j=1}^{n_i} y_{ij}^2$			
Somme des produits				$\sum_{j=1}^{n_i} x_{ij}\, y_{ij}$				
Variation				$\sum_{j=1}^{n_i} x_{ij}^2 - \frac{1}{n_i}\left(\sum_{j=1}^{n_i} x_{ij}\right)^2$	$\sum_{j=1}^{n_i} y_{ij}^2 - \frac{1}{n_i}\left(\sum_{j=1}^{n_i} y_{ij}\right)^2$			
Covariation				$\sum_{j=1}^{n_i} x_{ij}y_{ij} - \frac{1}{n_i}\left(\sum_{j=1}^{n_i} x_{ij}\right)\left(\sum_{j=1}^{n_i} y_{ij}\right)$				
Moyenne				$\bar{x}_i$	$\bar{y}_i$			
Variance				$\hat{s}_i^2(x)$	$\hat{s}_i^2(y)$			
Ecart-type				$\hat{s}_i(x)$	$\hat{s}_i(y)$			
Covariance				$\hat{s}_i(x,y)$				

Tableau 5

Présentation du calcul des indices statistiques concernant les groupes E_1, E_2, ..., E_k dans le cas de deux variables X, Y. Une colonne ''ensemble'' peut éventuellement être adjointe à ce tableau.

Nous cherchons à construire une troisième variable, Z, fonction simple de X et de Y, et dont le pouvoir discriminant soit le plus élevé possible.

La simplicité requise de Z, comme fonction de X et de Y, nous oriente vers la famille des fonctions linéaires :

$$(X, Y) \longmapsto Z = aX + bY$$

Il s'agira donc d'attribuer aux coefficients a , b des valeurs telles que le rapport f soit maximum.

INTERPRETATION GEOMETRIQUE

Une telle variable Z étant donnée, son utilisation dans la discrimination entre E_1 et E_2 conduit, comme cela a été exposé plus haut, à déterminer une valeur-seuil z_0 telle que, si pour un individu i on a $z_i \leqslant z_0$, cet individu est attribué au groupe E_1, et si $z_i > z_0$, il est attribué au groupe E_2.

Dans le plan muni d'un repère (Ox, Oy), la droite d'équation $ax + by = z_0$ partage le plan en deux demi-plans :

P_1 pour lequel on a $ax + by \leqslant z_0$
P_2 pour lequel on a $ax + by > z_0$

La règle de décision énoncée ci-dessus, quant à l'attribution de l'individu i au groupe E_1 ou au groupe E_2, peut donc s'énoncer ainsi, par référence à cette représentation géométrique : si le point de coordonnées (x, y), représentant un individu, est situé dans le demi-plan P_1, cet individu est attribué au groupe E_1 ; si le point est dans P_2, l'individu est attribué à E_2.

Ainsi la droite d'équation $ax + by = z_0$ est une droite-seuil entre deux zones (deux demi-plans). Elle joue dans le cas de deux variables le rôle de frontière que joue la valeur-seuil x_0 dans le cas d'une variable.

L'analyse discriminante consistera donc à rechercher la "meilleure" droite séparant les deux groupes.

RECHERCHE DE LA FONCTION LINEAIRE DISCRIMINANTE

La maximisation du pouvoir discriminant de Z nous fait rechercher le système de deux coefficients a , b tel que le rapport f_Z (concernant la variable Z) soit le plus élevé possible.
Ce rapport s'écrit

$$f_Z = \frac{\hat{s}^2_{B,Z}}{\hat{s}^2_{W,Z}}$$

où $\hat{s}^2_{B,Z}$ désigne la variance intergroupes relative à la variable Z, soit :

$$\hat{s}^2_{B,Z} = n_1 (\bar{z}_1 - \bar{z})^2 + n_2 (\bar{z}_2 - \bar{z})^2$$

et où $\hat{s}^2_{W,Z}$ désigne la variance intra-groupes relative à Z, soit :

$$\hat{s}^2_{W,Z} = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_i)^2$$

En remplaçant dans ces expressions z par $ax + by$, nous obtenons finalement :

$$\hat{s}^2_{B,Z} = a^2 B^2_X + 2ab\, B_{XY} + b^2 B^2_Y$$

et

$$\hat{s}^2_{W,Z} = \frac{1}{n_1 + n_2 - 2} (a^2 W^2_X + 2ab\, W_{XY} + b^2 W^2_Y)$$

où nous avons noté :

$B^2_X = \sum_{i=1}^{2} n_i (\bar{x}_i - \bar{x})^2$ = la variation inter-groupes de X ;

$B^2_Y = \sum_{i=1}^{2} n_i (\bar{y}_i - \bar{y})^2$ = la variation inter-groupes de Y ;

$B_{XY} = \sum_{i=1}^{2} n_i (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})$ = une quantité que nous appellerons désormais la *covariation inter-groupes* de X et Y.

$W^2_X = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$ = la variation intra-groupes de X ;

$W^2_Y = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ = la variation intra-groupes de Y ;

$W_{XY} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i)$ = une quantité que nous appellerons désormais la *covariation intra-groupes* de X et Y.

Le rapport f_Z dépend donc des deux paramètres a et b . Quelles valeurs devons-nous attribuer à a et b pour que f_Z soit maximum ? Il s'agit d'un problème classique de recherche d'extremum d'une fonction dérivable de plusieurs variables. Ici, comme le nombre d'inconnues, soit 2, est peu élevé, nous pouvons utiliser la méthode ordinaire, consistant à rechercher les valeurs des paramètres a et b qui annulent les dérivées partielles $\frac{\partial f_Z}{\partial a}$ et $\frac{\partial f_Z}{\partial b}$, ensuite à discuter sur la nature du point trouvé (maximum, minimum, col).

APPLICATION NUMERIQUE

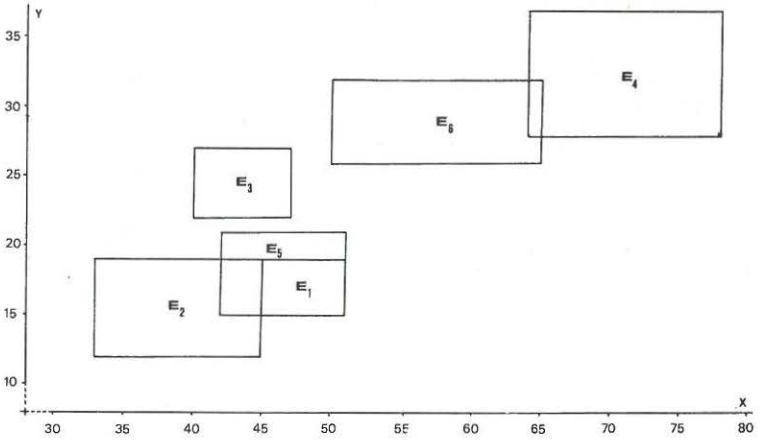


FIG. 3 : "Plan de situation" des six groupes E_1 à E_6, dans le plan de coordonnées X, Y

La figure 3 est un "plan de situation" des groupes E_1 à E_6, dans le plan de coordonnées X, Y. A chaque groupe on a associé un rectangle délimité, pour chacune des variables, par la plus petite et la plus grande valeurs prises dans l'échantillon disponible. Ainsi, d'un simple coup d'œil, ce plan permet de se rendre compte de la situation des groupes les uns par rapport aux autres : E_3 est isolé (donc pour lui la discrimination est immédiate) ; E_4 et E_6 sont isolés des autres, mais doivent être discriminés entre eux ; E_1 est "inclus" dans E_5 (on peut se poser la question : le genre E_1 est-il biologiquement un sous-ensemble du genre E_5 ? mais la discussion à ce sujet est hors de notre propos actuel) ; enfin E_1, E_2, E_5 s'interpénètrent fortement, c'est sur eux que nous allons concentrer notre effort de discrimination.

Dans la figure 4, nous avons repris, à plus grande échelle que dans la figure 3, ces trois groupes E_1, E_2, E_5, en y représentant chaque individu. Le tableau 6 nous donne les résultats numériques des calculs concernant ces trois groupes, selon le schéma du tableau 5. A partir de ces résultats numériques, nous pouvons appliquer la méthode décrite plus haut, à chacun des couples (E_1;E_2), (E_2;E_5), (E_1;E_5).

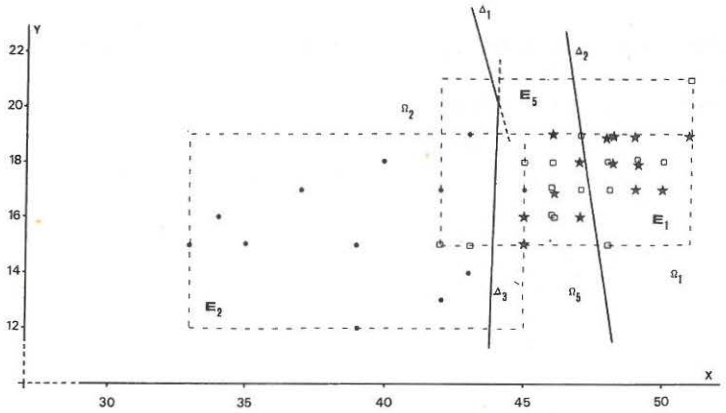


FIG. 4 : Représentation des données individuelles appartenant aux groupes E_1 (★), E_2 (•), E_5 (□), et des droites de discrimination : $\triangle_1$ (entre E_1 et E_2), $\triangle_2$ (entre E_1 et E_5), $\triangle_3$ (entre E_2 et E_5). Autres explications dans le texte.

Groupe	E_1		E_2		E_5	
Variable	X	Y	X	Y	X	Y
Effectif	14		12		15	
Somme des données	668	247	472	188	702	258
Somme des carrés	31 916	4 381	18 732	2 992	32 934	4 476
Somme des produits	11 802		7 410		12 112	
Variation	42,857140	23,214286	166,666670	46,666667	80,400000	38,400000
Covariation	16,571430		15,333333		37,600000	
Moyenne	47,714286	17,642857	39,333333	15,666667	46,800000	17,200000
Variance	3,296703	1,785714	15,151515	4,242424	5,742857	2,742857
Ecart-type	1,815683	1,336306	3,892495	2,059715	2,396426	1,656157
Covariance	1,274725		1,393939		2,685714	

Tableau 6

Calcul des indices statistiques concernant les groupes E_1, E_2, E_5 pour les deux variables X, Y.

Nota : l'apparition d'un grand nombre de chiffres "significatifs" dans certains résultats peut étonner le lecteur. En fait, dans la pratique des calculs statistiques, la règle est de conserver, dans tous les résultats intermédiaires, le maximum de chiffres permis par le moyen de calcul utilisé. Le non-respect de cette règle, c'est-à-dire la pratique de l'arrondi à chaque étape, risque par accumulation de faire aboutir à un résultat complètement erroné.

Développons complètement les calculs concernant les groupes E_2 et E_5. A partir du tableau 6 nous calculons les variations et covariations inter-groupes et intra-groupes :

$$B_X^2 = 371{,}674070 \quad ; \quad W_X^2 = 247{,}066670 \quad ;$$
$$B_{XY} = 76{,}325930 \quad ; \quad W_{XY} = 52{,}933333 \quad ;$$
$$B_Y^2 = 15{,}674074 \quad ; \quad W_Y^2 = 85{,}066667 \quad .$$

La détermination du pouvoir discriminant de chacune des deux variables prise séparément nous donne :

— pour X : $f_X = \dfrac{B_X^2}{W_X^2} \times 25 = 37{,}609$ d'où $g_X = 0{,}974$;

— pour Y : $f_Y = \dfrac{B_Y^2}{W_Y^2} \times 25 = 4{,}606$ d'où $g_Y = 0{,}822$.

Quant au pouvoir discriminant f_Z de la variable $Z = aX + bY$ dont nous cherchons à déterminer les coefficients a et b , il s'écrira :

$$f_Z = 25 \times \frac{371{,}674070\ a^2 + 2 \times 76{,}325930\ ab + 15{,}674074\ b^2}{247{,}066670\ a^2 + 2 \times 52{,}933333\ ab + 85{,}066667\ b^2}$$

Le calcul de $\dfrac{\partial f_Z}{\partial a}$ et $\dfrac{\partial f_Z}{\partial b}$, et l'annulation de ces deux expressions, conduit finalement à un système de deux équations en a et b , qui sont en réalité deux équations identiques entre elles, de la forme :

$$a^2(B_X^2 W_{XY} - B_X W_X^2) + ab(B_X^2 W_Y^2 - B_Y^2 W_X^2) + b^2(B_{XY} W_Y^2 - B_Y^2 W_{XY}) = 0$$

équation "homogène" en a et b , ce qui signifie qu'il y a une infinité de solutions, seul le *rapport* $\dfrac{a}{b} = r$ devant nécessairement prendre une valeur adéquate pour que f_Z soit extrêmum. Cette valeur r satisfait l'équation du second degré :

$$r^2(B_X^2 W_{XY} - B_{XY} W_X^2) + r(B_X^2 W_Y^2 - B_Y^2 W_X^2) + (B_{XY} W_Y^2 - B_Y^2 W_{XY}) = 0$$

qui s'écrit, dans le cas numérique étudié :

$$816{,}353950\ r^2 + 27\,744{,}53308\ r + 5\,663{,}0371 = 0$$

Cette équation possède deux racines :

$$r_1 = -0{,}205357 \quad ; \quad r_2 = -33{,}780554$$

Les valeurs correspondantes de f_Z sont :

$$f_Z(r_1) = 0{,}0000 \quad ; \quad f_Z(r_2) = 37{,}6186$$

C'est la racine r_2 qui correspond au maximum de f_Z. Les conséquences de ce calcul sont les suivantes :

— pour exprimer la variable $Z = aX + bY$, nous choisissons b arbitrairement, et le coefficient a tel que $\frac{a}{b} = r_2 = -33{,}780554$; par exemple pour $b = 1$, nous avons $a = -33{,}7805$, soit $Z = -33{,}8X + Y$.

— nous discriminons ensuite les deux groupes sur la base de cette variable Z, dont le pouvoir discriminant est apprécié par

$$f_Z = f_Z(r_2) = 37{,}619 \quad \text{et} \quad g_Z = 0{,}974$$

(notons une très légère augmentation du pouvoir discriminant lorsqu'on remplace X par Z) ; calculons le seuil de discrimination z_0 selon la procédure indiquée au § 3 :

$$\bar{z}_2 = -33{,}780554\,\bar{x}_2 + \bar{y}_2 = -1\,313{,}035124$$
$$\bar{z}_5 = -33{,}780554\,\bar{x}_5 + \bar{y}_5 = -1\,563{,}729927$$

$$\hat{s}_2^2(z) = (-33{,}780554^2 \times 15{,}151515 - 2 \times 33{,}780554 \times 1{,}393939 + 4{,}242424$$
$$= 17\,199{,}85197 \quad \text{d'où} \quad \hat{s}_2(z) = 131{,}148206$$

$$\hat{s}_5^2(z) = (-33{,}780554)^2 \times 5{,}742857 - 2 \times 33{,}780554 \times 2{,}685714 + 2{,}742857$$
$$= 6\,374{,}615642 \quad \text{d'où} \quad \hat{s}_5(z) = 79{,}841190$$

Ainsi :

$$z_0 = \frac{\bar{z}_2 \hat{s}_5(z) + \bar{z}_5 \hat{s}_2(z)}{\hat{s}_2(z) + \hat{s}_5(z)} = -1\,468{,}863684$$

En conclusion, la discrimination entre les deux groupes E_2 et E_5 sera pratiquée par l'utilisation de la fonction $Z = -33{,}8X + Y$, avec la valeur-seuil $z_0 = -1\,468{,}9$; ou encore, plus commodément, par l'utilisation de la fonction $Z - z_0 = -33{,}8X + Y + 1\,468{,}9$, avec la valeur-seuil 0. Pour les deux autres couples de groupes, le lecteur pourra reconstituer aisément, à partir du tableau 6, les calculs conduisant aux résultats suivants :

Discrimination entre E_1 et E_5 :

$$B_X^2 = 6{,}053200 \quad ; \quad W_X^2 = 123{,}257140$$
$$B_{XY} = 2{,}932020 \quad ; \quad W_{XY} = 54{,}171430$$
$$B_Y^2 = 1{,}420197 \quad ; \quad W_Y^2 = 61{,}614286$$

d'où la fonction discriminante :

$$Z - z_0 = 6{,}4X + Y - 320{,}1$$

dont le pouvoir discriminant n'est tout de même pas très élevé :

$$f = 1{,}33 \quad ; \quad g = 0{,}57$$

Discrimination entre E_1 et E_2 :

$$B_X^2 = 453{,}860810 \quad ; \quad W_X^2 = 209{,}523810$$
$$B_{XY} = 107{,}018320 \quad ; \quad W_{XY} = 31{,}904763$$
$$B_Y^2 = 25{,}234432 \quad ; \quad W_Y^2 = 69{,}880953$$

d'où la fonction discriminante :

$$Z - z_0 = 3{,}6X + Y - 177{,}0$$

dont le pouvoir discriminant est meilleur :

$$f = 53{,}15 \quad ; \quad g = 0{,}98$$

Nous avons représenté sur la même figure 4 les trois droites d'équations respectives : $\triangle_1 : 3{,}6X + Y - 177{,}0 = 0$; $\triangle_2 : 6{,}4X + Y - 320{,}1 = 0$; $\triangle_3 : -33{,}8X + Y + 1\,468{,}9 = 0$.

Pratiquement le plan — ou plus exactement la portion de plan compatible avec les données observées — se trouve subdivisé en trois zones Ω_1, Ω_2, Ω_5 d'attribution des trois groupes E_1, E_2, E_5, zones délimitées par des portions de ces droites-seuils (voir figure 4).

D. FENEUILLE
I.U.T. d'Aix-en-Provence

9. PLANIFICATION D'EXPÉRIENCES (*)

Faut-il planifier une démarche expérimentale ? c'est-à-dire faut-il réfléchir *avant expérimentation* à l'organisation des expériences ? Voyons sur un exemple :

> Comment déterminer les masses de k objets en N pesées sous les contraintes suivantes :
> — minimum de pesées (N = k + 1)
> — maximum d'efficacité.

Précisons les deux contraintes ci-dessus :

Pour conduire son expérience l'expérimentateur dispose d'une balance et de masses marquées. Il s'agit (voir dessin plus bas) d'une balance à deux plateaux **égaux**. Celle-ci possède *toutes les qualités* d'une balance. Une aiguille traduit par son déplacement (droite ou gauche) l'inclinaison vers le plateau le plus "lourd". Une marque (que nous appellerons le "zéro") sur une sorte de cadran (non forcément gradué) permet à l'expérimentateur de "voir" la position d'équilibre ; c'est-à-dire qu'à l'aide de masses *marquées* il ramène l'aiguille (déviée par les objets sur les plateaux) en coïncidence avec ce "zéro" pour obtenir "l'équilibre". Nous admettrons que l'expérimentateur (même doué d'une excellente vue) fait des *erreurs de lecture* (ce sont les *seules* erreurs de l'expérience) ; c'est-à-dire que l'aiguille n'est pas toujours ramenée *exactement* sur le "zéro". Cette "erreur", même petite devant les masses théoriques des objets, répercute une "autre erreur" sur l'estimation de ces masses (voir plus loin). Etre efficace, c'est rendre minimum, non pas l'erreur de lecture, mais la "répercussion" en question. En d'autres termes, être efficace, c'est cerner "au plus près" les masses vraies des objets à peser.

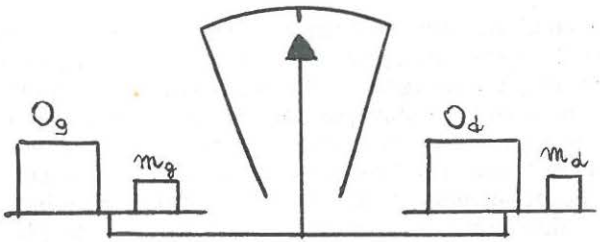
(*) Le début de cet article a été publié sous une forme un peu différente dans le Bulletin APM n° 304 : D. FENEUILLE ; D. MATHIEU ; R. PHAN-TAN-LUU (Invitation à la méthodologie de la recherche expérimentale : Notion de Plan d'Expériences).

Nous admettrons également que l'aiguille de la balance en l'absence d'objets et de masses sur les plateaux ne pointe pas forcément sur le "zéro". Il ne s'agit pas d'un défaut de la balance mais par exemple du "plan" sur lequel elle repose (car cette balance est transportable). On peut alors "tarer" la balance ou (ce qui sera équivalent) procéder à une pesée supplémentaire (ce qui explique : $N = k + 1$ ci-dessus).

Balance à l'équilibre avec

- 0_g (respectivement 0_d) l'ensemble des objets (éventuellement vide) à peser placés à gauche (respectivement à droite)
- m_g (respectivement m_d) valeur des masses marquées (éventuellement nulle) placées à gauche (respectivement à droite) pour rétablir l'équilibre.

Effectuer une pesée c'est : mettre *aucun, un* ou *des* objets sur *un* ou les *deux* plateaux, puis à l'aide de masses marquées établir l'équilibre en ramenant l'aiguille sur le "zéro".



Nous noterons y_i *la mesure de la pesée n° i* avec la *convention*

$$y_i = m_d - m_g$$

Développons les concepts sur un cas particulier :

Quatre expérimentateurs ont à déterminer les masses de trois objets 0_1, 0_2, 0_3 sous les deux contraintes du problème : ils s'accordent pour convenir du nombre minimum de pesées égal à quatre (le "zéro" étant à déterminer : par exemple au prix d'une pesée fictive). Par contre ils diffèrent dans le choix de la stratégie à suivre pour répondre au critère : "maximum d'efficacité".

Ci-dessous nous avons ordonné les stratégies, proposées par les quatre expérimentateurs, par efficacité croissante ; on invite le lecteur à imaginer d'autres *stratégies ou plans d'expériences* (limités de toute manière au nombre de façons de placer les objets sur les plateaux).

EXPÉRIMENTATEUR N° 1

Il n'utilise qu'un seul plateau de la balance pour mettre les objets.

N° pesée	Etat du système	Résultat $(m_d - m_g)$
1	0_1 seul sur le plateau	y_1
2	0_2 seul sur le plateau	y_2
3	0_3 seul sur le plateau	y_3
4	aucun objet sur le plateau	y_4

Cette façon de procéder est la plus traditionnelle (c'est la pesée pratiquée dans la vie courante). Nous allons sur ces expériences modéliser notre problème ; nous reprendrons les mêmes notations et hypothèses tout au long de cet article.

L'expérimentateur n° 1 annonce au vu des résultats :

j'estime la masse de l'objet 0_1 par $y_1 - y_4$
j'estime la masse de l'objet 0_2 par $y_2 - y_4$
j'estime la masse de l'objet 0_3 par $y_3 - y_4$

Dire "j'estime la masse de 0_i par $y_i - y_4$" c'est aussi souhaiter que le résultat expérimental soit "proche" de la masse réelle.

Modélisons notre problème.

Soit :

η_i le résultat *théorique* de la pesée n° i
y_i le résultat *expérimental* de la pesée n° i
$\hat{p}_j$ l'estimateur de p_j, masse de 0_j .

Pour différencier p_j, masse inconnue, de son estimation, masse calculée, nous mettons un chapeau sur l'estimateur $\hat{p}_j$ (qui se dit : p chapeau indice j).

Nous allons faire quatre hypothèses simplificatrices mais cohérentes.

① à chaque expérience on a :

$$y_i = \eta_i + e_i$$

c'est-à-dire que la mesure expérimentale de la pesée diffère du résultat théorique d'un "bruit additif" e_i (l'erreur de lecture dont il est question au début).

② le bruit est une variable aléatoire telle que :

$$E(e_i) = 0 \quad ; \text{espérance mathématique nulle.}$$

ceci signifie que, sur un grand nombre de répétitions de la pesée n° i, en "moyenne", l'erreur de lecture est nulle (ou insignifiante) ou encore que, en moyenne, la mesure expérimentale tend vers le résultat théorique (si l'on répète de nombreuses fois la pesée n° i), c'est-à-dire :

$$E(y_i) = \eta_i \quad ; y_i \text{ est aussi une variable aléatoire.}$$

③ $var(y_i) = \sigma^2$ pour toute pesée i ; σ^2 est une constante liée à la balance et à l'observateur. σ^2 mesure la dispersion de y_i à η_i. On admettra que cette dispersion est la même pour toutes les stratégies de pesée d'un même manipulateur. (En d'autres termes : plus σ^2 est petit, plus l'individu est précis)

④ $cov(y_i, y_j) = 0$ pour $i \neq j$; il n'y a pas corrélation entre deux mesures de deux pesées différentes.

Enfin, pour schématiser les quatre pesées nous utiliserons la représentation matricielle suivante que nous appelons *matrice d'expériences.*

$$D = \begin{array}{c} \begin{matrix} 0_1 & 0_2 & 0_3 \end{matrix} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \end{array} = (d_{ij}) \qquad \begin{matrix} i : \text{n° ligne} \\ j : \text{n° colonne} \end{matrix}$$

d_{ij} illustrant l'état de l'objet j au cours de la pesée i soit :

$$\begin{cases} 0 & \text{absent de la pesée i} \\ 1 & \text{sur le plateau (de droite pour 2 plateaux)} \\ -1 & \text{sur le plateau de gauche (cas à 2 plateaux)} \end{cases}$$

Cette matrice D modélise le *plan d'expériences :*

n° pesée	0_1	0_2	0_3	résultat
1	droite	non	non	y_1
2	non	droite	non	y_2
3	non	non	droite	y_3
4	non	non	non	y_4

Si p_j est la masse de 0_j (masse à estimer), on a les équations *théoriques* à l'équilibre :.

$$\left.\begin{aligned} p_0 + p_1 &= \eta_1 \\ p_0 + p_2 &= \eta_2 \\ p_0 + p_3 &= \eta_3 \\ p_0 &= \eta_4 \end{aligned}\right\}$$ en notant p_0 la masse de l'objet fictif 0_0 et η_i la mesure théorique de la pesée i ;

comme $p_j = \eta_j - \eta_4 \qquad j = 1, 2, 3$

on pose donc :

$$\begin{cases} \hat{p}_j = y_j - y_4 & j = 1, 2, 3 \\ \hat{p}_0 = y_4 \end{cases}$$

Pour évaluer l'efficacité de la stratégie proposée par l'expérimentateur, il faut :

— que pour tout j, l'estimateur $\hat{p}_j$ de p_j (masse de 0_j) soit centré sur p_j

$$E(\hat{p}_j) = p_j$$

— déterminer la variance de cet estimateur

L'estimateur $\hat{p}_j$ est une variable aléatoire

$\hat{p}_j = y_j - y_4$ (différence de 2 variables aléatoires).

D'où :

$$\begin{aligned} E(\hat{p}_j) &= E(y_j - y_4) = E(y_j) - E(y_4) \\ &= \eta_j - \eta_4 \\ &= p_j + p_0 - p_0 \\ E(\hat{p}_j) &= p_j, \text{ ce qui se dit :} \end{aligned}$$

$\hat{p}_j$ est un estimateur non biaisé de p_j.

$$\begin{aligned} \text{var}(\hat{p}_j) &= \text{var}(y_j - y_4) \\ &= \text{var}(y_j) + \text{var}(y_4) - 2\,\text{cov}(y_j, y_4) \\ &= \sigma^2 + \sigma^2 + 0 = 2\sigma^2 \end{aligned}$$

En résumé : en adoptant le plan d'expérience modélisé par la matrice D définie ci-dessus l'expérimentateur n° 1 estime les masses p_j des objets 0_j grâce à des estimateurs $\hat{p}_j$: non biaisés, de variance $2\sigma^2$.

$$\begin{cases} \text{var}(\hat{p}_j) = 2\sigma^2 & j = 1, 2, 3 \\ \text{var}(\hat{p}_0) = \sigma^2 \end{cases}$$

EXPERIMENTATEUR N° 2

Il n'utilise aussi qu'un seul plateau de la balance pour les objets. Envisageons la stratégie modélisée par la matrice d'expériences suivante :

$$D = \begin{array}{c} \begin{matrix} 0_1 & 0_2 & 0_3 \end{matrix} \\ \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \end{array}$$

C'est-à-dire que l'on pèse 2 objets à chaque fois après avoir déterminé le « zéro » par une pesée fictive.

En conservant les mêmes notations, c'est-à-dire :

$$\begin{cases} \hat{p}_j \text{ estimateur de } p_j, \text{ masse de } 0_j \\ \eta_i \text{ résultat théorique de la pesée i} \\ y_i \text{ résultat expérimental de la pesée i,} \end{cases}$$

On a les équations théoriques d'équilibre :

$$\begin{cases} p_0 = \eta_1 \\ p_0 + p_1 + p_2 = \eta_2 \\ p_0 + p_1 + p_3 = \eta_3 \\ p_0 + p_2 + p_3 = \eta_4 \end{cases}$$

d'où les estimateurs :

$$\begin{cases} \hat{p}_1 = (y_2 + y_3 - y_1 - y_4) / 2 \\ \hat{p}_2 = (y_2 + y_4 - y_1 - y_3) / 2 \\ \hat{p}_3 = (y_3 + y_4 - y_1 - y_2) / 2 \end{cases}$$

On montre facilement que :

$$E(\hat{p}_j) = p_j \text{ et } \text{var}(\hat{p}_j) = \sigma^2 \qquad j = 0, 1, 2, 3$$

Donc l'expérimentateur n° 2 a **doublé l'efficacité** obtenue par le premier expérimentateur (puisqu'il a réduit de moitié la dispersion de $\hat{p}_j$ à p_j) et ceci simplement en pesant simultanément deux objets au lieu d'un !

EXPERIMENTATEUR N° 3

Il utilise les deux plateaux de la balance. Soit la matrice d'expériences :

$$D = \begin{array}{c} \begin{matrix} 0_1 & 0_2 & 0_3 \end{matrix} \\ \begin{pmatrix} -1 & -1 & -1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix} \end{array}$$

$d_{ij} = -1$ schématise la présence de l'objet 0_j sur le plateau de gauche au cours de la pesée i (voir la description de la balance en début d'article).

Nous avons :
équations théoriques :

$$\begin{cases} p_0 - p_1 - p_2 - p_3 = \eta_1 \\ p_0 + p_1 - p_2 - p_3 = \eta_2 \\ p_0 - p_1 + p_2 - p_3 = \eta_3 \\ p_0 - p_1 - p_2 + p_3 = \eta_4 \end{cases}$$

d'où les estimateurs :

$$\begin{cases} \hat{p}_1 = (y_2 - y_1) / 2 \\ \hat{p}_2 = (y_3 - y_1) / 2 \\ \hat{p}_3 = (y_4 - y_1) / 2 \\ \hat{p}_0 = (-y_1 + y_2 + y_3 + y_4) / 2 \end{cases}$$

On montre facilement que :

$$\boxed{\begin{array}{lll} E(\hat{p}_j) = p_j & & j = 0, 1, 2, 3 \\ var(\hat{p}_j) = \sigma^2 / 2 & & j = 1, 2, 3 \\ var(\hat{p}_0) = \sigma^2 & & \end{array}}$$

L'efficacité est encore meilleure !

EXPERIMENTATEUR N° 4

Reprenons la matrice d'expériences proposée par l'expérimentateur n° 3 et considérons la première pesée

$$\begin{matrix} 0_1 & 0_2 & 0_3 \end{matrix}$$
$$[-1 \quad -0 \quad -1]$$

Pour estimer la correction du « zéro », l'expérimentateur n° 3 met les trois objets sur le plateau de gauche. Ceci est totalement arbitraire (les deux plateaux sont identiques) et nous pouvons nous poser la question de savoir ce qu'il se passerait si nous décidions, pour cette pesée, de mettre les trois objets sur le plateau de droite.

Expérimentateur 3			*Expérimentateur 4*		
0_1	0_2	0_3	0_1	0_2	0_3
$[-1$	-1	$-1]$	$[1$	1	$1]$

Dans une démarche classique, dans laquelle les expériences sont faites suivant l'intuition ou le flair, ce choix ne devrait pas avoir d'influence sur les résultats obtenus.

Est-ce bien vrai ?
Nous obtenons la matrice d'expériences

$$D = \begin{array}{c} \begin{matrix} 0_1 & 0_2 & 0_3 \end{matrix} \\ \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix} \end{array}$$

Cette matrice ne diffère de la matrice proposée par l'expérimentateur n° 3 que par la 1re ligne.

Nous avons :
Equations théoriques :

$$\begin{cases} p_0+p_1+p_2+p_3 = \eta_1 \\ p_0+p_1-p_2-p_3 = \eta_2 \\ p_0-p_1+p_2-p_3 = \eta_3 \\ p_0-p_1-p_2+p_3 = \eta_4 \end{cases}$$

d'où les estimateurs :

$$\begin{cases} \hat{p}_0 = (y_1+y_2+y_3+y_4) / 4 \\ \hat{p}_1 = (y_1+y_2-y_3-y_4) / 4 \\ \hat{p}_2 = (y_1-y_2+y_3-y_4) / 4 \\ \hat{p}_3 = (y_1-y_2-y_3+y_4) / 4 \end{cases}$$

On trouve :

$$\boxed{\begin{aligned} E(\hat{p}_j) &= p_j \\ \text{var}(\hat{p}_j) &= \sigma^2 / 4 \end{aligned}} \qquad j = 0, 1, 2, 3$$

Le résultat obtenu met en évidence l'importance du choix de chaque expérience.
L'expérimentateur n° 3 est passé bien près d'un plan meilleur .

Manque de flair, d'intuition ou de méthode ?
A ce jeu on peut se demander s'il n'est pas possible de faire mieux. Remarquons tout de même qu'entre les efficacités des plans n° 1 et 4 il existe un rapport 8, ce qui est pour le moins important !

On démontre le :

Théorème[1]

* Pour n'importe quel plan d'expériences de N pesées utilisant une balance à deux plateaux, la variance de l'estimateur $\hat{p}_j$ est telle que :

$$\text{var}(\hat{p}_j) \geqslant \sigma^2/N$$

$$j = 0, 1, \ldots, K \quad (K : \text{nombre d'objets à peser})$$

* La condition nécessaire et suffisante pour que :

$$\text{var}(\hat{p}_j) = \sigma^2 / N$$

est que l'on ait :

$$^tXX = NI_{K+1} \qquad (1)$$

où

N : nombre de pesées ($N \geqslant K+1$)
K : nombre d'objets
X : matrice des coefficients des p_j dans les équations théoriques ; ^tX : matrice transposée de X
I_{K+1} : matrice unité de rang $K+1$

REMARQUE

Certains expérimentateurs, faisant fi de méthode, préfèrent répéter un grand nombre de fois leur plan d'expérience et prendre comme estimateur la moyenne arithmétique des estimateurs calculés à chaque plan : par exemple l'expérimentateur n° 1 répètera t fois le plan modélisé par D.

---

(1) Les démonstrations peuvent être demandées au Laboratoire I.U.T. Aix-en-Provence, elles n'offrent pas d'intérêt ici.

A la fin des manipulations il pose :

$$\hat{p}_j = \frac{1}{t} \sum_{h=1}^{h=t} \hat{p}_j^{(h)} \qquad (j = 1, 2, 3)$$

où $\hat{p}_j^{(h)}$ est l'estimateur obtenu au cours de la h-ième **répétition du plan.**

On montre facilement (théorie des échantillons) que :

$$E(\hat{p}_j) = p_j \quad \text{et} \quad \text{var}(\hat{p}_j) = \frac{1}{t} \text{var}(\hat{p}_j^{(h)})$$

Pour l'expérimentateur 1 on a : $\text{var}\, \hat{p}_j^{(h)} = 2\sigma^2$

$$\text{var}(\hat{p}_j) = \frac{2\sigma^2}{t} \quad (j = 1, 2, 3) \text{ (après t répétitions du plan)}$$

$\hat{p}_j$, ainsi construit, est un estimateur très efficace puisque :

$$\lim_{t \to \infty} \frac{2\sigma^2}{t} = 0$$

Mais c'est un estimateur coûteux !

On a remarqué aussi que l'expérimentateur n° 1 est obligé de répéter 8 fois son plan pour égaler celui de l'expérimentateur 4 :

$$\frac{2\sigma^2}{t} = \frac{\sigma^2}{4} \Rightarrow t = 8$$

On imagine facilement le coût d'une telle attitude dans un problème de sciences expérimentales autres que de banales pesées !

Nous avons établi que pour un problème aussi simple que celui des pesées, il existe une stratégie « optimale ».

Que se passe-t-il si nous introduisons des contraintes nouvelles et diverses ?

Existe-t-il alors une stratégie « optimale » ?

Quelques exemples (susceptibles d'applications dans nos classes) :

PROBLEME A

Plaçons-nous sous les critères suivants :
— nombre de pesées égal à 8
— efficacité maximale
— nombre de manipulations minimum

} ordonnés ainsi

On démontre que sous la première contrainte (N = 8) et si nous utilisons une balance à deux plateaux, le critère d'efficacité est réalisé par un plan modélisé par la matrice D, tel que :

$${}^tXX = 8\, I_4 \text{ (voir théorème plus haut)}$$

La matrice X est par définition la matrice des coefficients des p_i (dite matrice du modèle) mais c'est en fait la matrice d'expériences D à laquelle on ajoute une première colonne dont tous les éléments valent 1 (coefficient de p_0 dans les équations d'équilibre).

$$X = [1\,D]$$

Dans ce cas nous avons :

$$\text{var}(\hat{p}_j) = \sigma^2 / 8$$

Il existe de nombreuses constructions de D. Une famille de telles matrices est caractérisée par :

$$D_1 = \begin{array}{c} \begin{matrix} 0_1 & 0_2 & 0_3 \end{matrix} \\ \begin{pmatrix} -1 & -1 & -1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \\ 1 & -1 & 1 \\ -1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \end{array}$$

L'ordre dans lequel sont écrites les expériences est dit **ordre standard*.** Mais il est clair que les matrices $({}^tXX)$ et $({}^tXX)^{-1}$ sont inchangées par des permutations sur les lignes (donc en permutant l'ordre des pesées).

Etudions le dernier critère.

Nombre minimum d'objets à manipuler entre les pesées

On peut supposer que les objets sont lourds (économie d'énergie) ou fragiles (risque de rupture). On peut imaginer facilement des situations réelles (changement de conditions expérimentales...).

Pour illustrer ce critère supposons que tout déplacement d'objet s'accompagne d'une dépense de 10 U.

Calculons le nombre de manipulations pour la matrice ci-dessus (on excepte le dépôt des 3 objets sur les plateaux au départ et leur retrait à la fin : identiques dans tous les plans).

---

* *Première colonne : alternance de − et de + avec un pas de 1 ; deuxième colonne : − − + + etc.*

$$D_1 = \begin{array}{c} \begin{array}{cccc} 0_1 & 0_2 & 0_3 & \text{coût} \end{array} \\ \left(\begin{array}{ccc|c} -1 & -1 & -1 & \\ 1 & -1 & -1 & 10 \\ -1 & 1 & -1 & 20 \\ 1 & 1 & -1 & 10 \\ -1 & -1 & 1 & 30 \\ 1 & -1 & 1 & 10 \\ -1 & 1 & 1 & 20 \\ 1 & 1 & 1 & 10 \end{array}\right) \\ \hline 110 \text{ U} \end{array}$$

par exemple pour passer de la 2ème pesée à la 3ème, il faut :
— déplacer l'objet 0_1 du plateau de droite sur le plateau de gauche ..... dépense 10 U
— déplacer l'objet 0_2 du plateau de gauche sur le plateau de droite ..... dépense 10 U
Total ..... 20 U

Il est évident aussi que pour les matrices de cette famille, le nombre de manipulations ne peut être inférieur à 7 : une manipulation seulement entre deux pesées !

Les deux plans suivants (parmi d'autres) répondent au critère de manipulations minimales.

$$D_2 = \begin{array}{c} \begin{array}{cccc} 0_1 & 0_2 & 0_3 & \text{coût} \end{array} \\ \left(\begin{array}{ccc|c} -1 & -1 & -1 & \\ 1 & -1 & -1 & 10 \\ 1 & 1 & -1 & 10 \\ -1 & 1 & -1 & 10 \\ -1 & 1 & 1 & 10 \\ 1 & 1 & 1 & 10 \\ 1 & -1 & 1 & 10 \\ -1 & -1 & 1 & 10 \end{array}\right) \\ \hline 70 \text{ U} \end{array} \qquad D_3 = \begin{array}{c} \begin{array}{cccc} 0_1 & 0_2 & 0_3 & \text{coût} \end{array} \\ \left(\begin{array}{ccc|c} -1 & -1 & -1 & \\ 1 & -1 & -1 & 10 \\ 1 & 1 & -1 & 10 \\ 1 & 1 & 1 & 10 \\ 1 & -1 & 1 & 10 \\ -1 & -1 & 1 & 10 \\ -1 & 1 & 1 & 10 \\ -1 & 1 & -1 & 10 \end{array}\right) \\ \hline 70 \text{ U} \end{array}$$

Ils sont obtenus à partir de la matrice d'ordre standard D_1.

Ils seront choisis de préférence à tout autre car dans le cas où on pense refaire le plan, une nouvelle fois, au vu des résultats du premier plan (exemple de démarche séquentielle), le passage de la dernière pesée du premier plan à la première pesée du deuxième plan s'effectue par une seule manipulation.

REMARQUE :

Il existe une famille de matrices résolvant le problème avec 6 manipulations seulement ; par exemple :

$$D = \begin{array}{c} \begin{array}{cccc} 0_1 & 0_2 & 0_3 & \text{coût} \end{array} \\ \left(\begin{array}{ccc} 1 & 1 & 1 & \\ 1 & 1 & 1 & 0 \\ 1 & -1 & -1 & 20 \\ 1 & -1 & -1 & 0 \\ -1 & 1 & -1 & 20 \\ -1 & 1 & -1 & 0 \\ -1 & -1 & 1 & 20 \\ -1 & -1 & 1 & 0 \end{array}\right) \\ \hline 60\ U \end{array}$$

Le principe est de faire deux fois un plan d'expérience de pesées optima mais en répétant la même pesée deux fois de suite.

Cette pratique est à rejeter au moins pour deux raisons :

— dans le cas de démarche séquentielle demandant éventuellement de dupliquer le plan elle perd de son intérêt (passage du premier plan au deuxième),

— mais surtout le fait de mesurer deux fois de suite la même opération **rompt l'hypothèse d'indépendance des erreurs de mesure à chaque pesée.**

PROBLEME B

Ce modèle de pesée étant une source de situations intéressantes, en voici une dernière :

Nous supposerons que "l'aiguille" de la balance entre deux manipulations subit une dérive, par exemple se déplace toujours dans le même sens et cumulativement d'une quantité petite mais constante : C.

On laisse au lecteur le soin de valider une telle hypothèse (par exemple : usure du matériel, vieillissement d'un catalyseur).

Plaçons-nous dans le cas N = 8 (nombre de pesées).

Un plan d'efficacité optimale est par exemple celui modélisé par la matrice du problème A (matrice d'ordre standard D_1).

Si l'on note :

Y_{iD} la valeur observée dans le cas de dérive au cours de la manipulation n° i (i = 1, ..., 8)

on a :

$Y_{iD} = y_i + (i-1)\,C$
où y_i est la valeur de la mesure sans dérive,
C le "pas" de la dérive.

0_1	0_2	0_3	Y_{iD}	noté
−1	−1	−1	y_1	y'_1
1	−1	−1	$y_2 + C$	y'_2
−1	1	−1	$y_3 + 2C$	y'_3
1	1	−1	$y_4 + 3C$	y'_4
−1	−1	1	$y_5 + 4C$	y'_5
1	−1	1	$y_6 + 5C$	y'_6
−1	1	1	$y_7 + 6C$	y'_7
1	1	1	$y_8 + 7C$	y'_8

ce qui donne :

$$\begin{aligned}\hat{p}_1 &= (-y'_1+y'_2-y'_3+y'_4-y'_5+y'_6-y'_7+y'_8)\,/\,8\\ &= (-y_1+y_2+c-y_3-2c+y_4+3c-y_5-4c+y_6+5c-y_7-6c+y_8+7c)\,/\,8\\ &= (-y_1+y_2-y_3+y_4-y_5+y_6-y_7+y_8)\,/\,8 + c/2\end{aligned}$$

donc

$$E(\hat{p}_1) = p_1 + c/2$$

De même, nous trouvons

$$\begin{cases} E(\hat{p}_2) = p_2 + c \\ E(\hat{p}_3) = p_3 + 2c \end{cases}$$

$\hat{p}_j$ est devenu un estimateur **biaisé** de p_j.

Le biais dépend de la matrice D, par exemple si l'on examine les résultats des deux plans donnant des déplacements d'objets minimum (D_2 et D_3 du problème A).

On montre respectivement que :

$$\begin{cases} E(\hat{p}_1) = 0 + p_1 \\ E(\hat{p}_2) = 0 + p_2 \\ E(\hat{p}_3) = 2c + p_3 \end{cases}$$

avec D_2

$$\begin{cases} E(\hat{p}_1) = c + p_1 \\ E(\hat{p}_2) = c + p_2 \\ E(\hat{p}_3) = c + p_3 \end{cases}$$

avec D_3

Il existe cependant au moins trois plans appartenant à la famille engendrée par D_1, pour lesquels il n'existe pas de biais pour les estimateurs.

$$D_4 = \begin{array}{c} \begin{matrix} 0_1 & 0_2 & 0_3 \end{matrix} \\ \begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \\ -1 & 1 & 1 \\ -1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & 1 & 1 \end{pmatrix} \end{array} \qquad D_5 = \begin{array}{c} \begin{matrix} 0_1 & 0_2 & 0_3 \end{matrix} \\ \begin{pmatrix} 1 & 1 & -1 \\ -1 & -1 & 1 \\ -1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & 1 \\ -1 & 1 & -1 \\ -1 & -1 & -1 \\ 1 & 1 & 1 \end{pmatrix} \end{array}$$

$$D_6 = \begin{array}{c} \begin{matrix} 0_1 & 0_2 & 0_3 \end{matrix} \\ \begin{pmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ -1 & -1 & -1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \end{array}$$

On trouve pour ces trois plans

$$\boxed{\begin{array}{c} E(\hat{p}_j) = p_j \\ \text{var}\,(\hat{p}_j) = \sigma^2 / 8 \end{array}}$$

$\hat{p}_j$ est un estimateur **non biaisé** et d'efficacité maximale.

Si l'on donne enfin une priorité moindre aux économies d'énergie on prendra D_4 plutôt que D_5 ou D_6 car le nombre de manipulations est respectivement de 11, 15, 13.

Il appartient bien sûr à l'expérimentateur de **classer** les objectifs avant d'effectuer les manipulations. De ce classement dépendra son plan d'expérience, donc ses manipulations (dans le problème B nous avions le classement :

— efficacité : $\frac{\sigma^2}{8}$

— pas de biais si dérive

— manipulation minimum d'objets).

Conclusion

On voit sur ces exemples simples qu'il était possible **a priori de bâtir une stratégie optimale** compte tenu des contraintes du problème.

L'un des objectifs de la *méthodologie de la recherche expérimentale* est d'inviter tout expérimentateur à réfléchir *avant l'expérience* sur un modèle de cette expérience (plan d'expérience) et à rechercher parmi les plans *raisonnables* celui ou ceux qui optimisent certains critères (notamment coût et efficacité).

Il est clair cependant que dès qu'il est question d'illustrer une méthodologie par un exemple (élémentaire de surcroît) ou par quelques principes (c'est ce que nous avons fait ci-dessus), on risque de paraître mettre en lumière des "évidences". Les "consultations" fréquentes où nous cotoyons des praticiens de l'expérimentation montrent qu'il n'en est rien ; nous trouvons parfois des chercheurs plus soucieux de s'exprimer en termes de moyens que de méthode, de résultats que d'objectifs ; quand enfin, devant leurs expériences réalisées, nous contestons leur procédure, et en proposons (démonstrations à l'appui) une "plus performante" qui aurait fait souvent l'économie de manipulations, c'est l'incrédulité ou la stupéfaction.

B. SALGUES
INSEE

10. QUELQUES SOURCES STATISTIQUES

A. Statistique officielle

1. L'INSEE

représente la principale source de données officielles. Celles-ci peuvent être obtenues dans les observatoires économiques de l'INSEE (voir annexe). On peut aussi lire les revues de l'INSEE (Statistique et Etudes économiques, Bulletin mensuel de la Statistique).

• Les principales études statistiques sont :

— le recensement de la population française, ainsi que de nombreuses études démographiques dont beaucoup sont complétées par l'INED,

— les résultats extraits des grands répertoires nationaux (SIREN : répertoire des entreprises),

— les enquêtes de conjonctures. En particulier celle réalisée auprès des industriels. A différentes questions, les réponses du type +, −, = sont soldées par l'INSEE puis corrigées des variations saisonnières,

— les indices de prix et leurs découpages plus fins et en particulier l'indice des prix de détail. Il s'agit d'un sondage sur lequel nous allons donner quelques précisions :

son champ : ensemble des ménages urbains vivant,sur le territoire français dont le chef est employé ou ouvrier

son année de référence : 1970

les méthodes employées : sondage ou double tirage, chacun à plusieurs degrés (agglomération, point de vente et sous-ensembles de biens et services, variétés, produits), données corrigées des variations saisonnières.

• Conjointement avec d'autres ministères, d'autres travaux sont réalisés :

— l'enquête annuelle d'entreprise (STISI) permet de connaître les principales données de la vie des entreprises,

— le bilan emploi-formation avec l'aide du ministère de l'Education et des universités. Un attaché de l'INSEE est en place dans certains rectorats (Créteil, Versailles, Strasbourg) afin d'élaborer et de diffuser des données sur le domaine scolaire,

— l'INSERM diffuse des données médicales.

• Il existe un certain nombre d'organismes d'information statistique de caractère parapublic, travaillant sur un secteur particulier de l'économie.

Citons le CREDOC qui étudie la consommation. Le CREP observe l'épargne. Le CERC analyse les revenus des Français. Le CEREN (Charbonnage de France, EDF, GDF) se consacre aux études sur l'énergie. Le bien-être et la qualité de la vie sont auscultés par le CEREBE tandis que le CEPREMAP est spécialisé dans l'étude des fonctions de la puissance publique.

Ils réalisent de nombreuses études intéressantes ; souvent les données originales proviennent de l'INSEE.

2. LES PUBLICATIONS OFFICIELLES

Dans le *Journal Officiel de la République Française* on trouve en particulier des statistiques mensuelles sur les mouvements des vins et des alcools, les bilans des entreprises nationalisées et parfois la gestion de leur trésorerie.

Dans les textes édités par *la Documentation Française* abondent de nombreux rapports chiffrés et détaillés.

3. LA CENTRALE DES BILANS DE LA BANQUE DE FRANCE

Sous ce titre, la Banque de France agrège des données chiffrées extraites des documents comptables annuels des entreprises faisant ou ayant fait appel au crédit public (Banque de France, Crédit National ...). Deux remarques s'imposent : les documents ne concernent en aucun cas toutes les entreprises françaises, de nombreuses cases des tableaux sont laissées en blanc afin de respecter le secret statistique.

Ces documents peuvent être obtenus (en petit nombre) gratuitement auprès de la Banque de France.

4. AU NIVEAU INTERNATIONAL

Nous nous limiterons à quelques remarques.

• Les statistiques de l'O.N.U., sont vendues assez cher par le canal des librairies. On y découvre essentiellement des données démographiques de qualités diverses (surtout hétérogènes).

• La Communauté européenne réalise quelques études au sein d'Eurostat-OSCE. Elle publie le courrier de la CEE (gratuit) ; les principaux thèmes sont l'énergie, la sidérurgie, l'agriculture, les gains et les

salaires. Le prochain recensement de la population devrait avoir lieu au niveau européen, ce qui pourra permettre de comparer les habitants des neufs pays sans trop de difficultés.

• L'OCDE publie d'intéressantes données économiques sur les pays membres par l'intermédiaire de plusieurs publications (perspectives de l'OCDE, etc.). Cet organisme explicite parfaitement les méthodes statistiques employées, ce qui permet de faire des vérifications immédiates.

B. Statistique privée

5. LA DAFSA

Société privée, filiale des grandes banques, elle réalise un travail analogue (mais plus complet) à celui de la centrale des bilans de la Banque de France. Cette source présente donc les mêmes inconvénients. Elle édite un important annuaire : *liaisons financières*, qui détaille l'actionnariat des grandes sociétés françaises.

6. SECODIP

SECODIP utilise trois "panels". De plus elle note les campagnes publicitaires dans les journaux et sur les ondes.

Le panel des consommateurs représente 4500 ménages qui sont représentatifs de 16.600.000 ménages français. Par des relevés hebdomadaires, on obtient des chiffres sur les prix de vente, les quantités achetées d'un produit, le volume des achats et les lieux d'achats. Par regroupement, on obtient des résultats régionaux ou nationaux (indices, quantités).

Le panel des détaillants est un échantillon de 2200 points de vente. Ce panel permet d'obtenir des prix de vente et des quantités vendues.

Le panel des collectivités regroupe 950 établissements (écoles, restaurants universitaires, hôpitaux, restaurants d'entreprises). Il permet d'obtenir un niveau des quantités achetées et des prix pratiqués.

7. SOFRES

SOFRES utilise des enquêtes statistiques périodiques (E.S.P.) de trois types :

• les E.S.P. "grand public" qui regroupent 2000 personnes de plus de 15 ans représentatives de l'ensemble de la population,

• les E.S.P. "postales" réalisées auprès d'un panel de 10 000 consommateurs (sondage),

• les E.S.P. "spéciales" touchent des catégories particulières (exemple : mère de bébé, agriculteur, possesseurs de jardin d'agrément, ...).

On obtient alors différentes données en prix, volumes ou motivations d'achat. Il faut remarquer qu'il s'agit de sondages, donc il serait intéressant d'avoir une évaluation du biais mais celui-ci est rarement fourni.

8. NIELSEN

Nielsen dispose d'un échantillon statistique permanent de points de vente de détail et de gros représentatifs à un moment précis d'un univers commercial donné. Cet échantillon permet de suivre à un instant donné l'écoulement des produits parmi les canaux de distribution.

Les résultats obtenus sont des ventes aux consommateurs, des achats et des stocks des commerçants ainsi que des ventes à confrère, soit sur le plan national ou régional, soit par type d'agglomérations ou de magasins. Un certain nombre d'études spéciales sont réalisées à partir de ces données. Il faut mentionner le fait suivant : cette méthodologie s'apparente à la technique photographique (l'écoulement précédent n'est pas dynamique et ressemble plutôt à une carte postale de la Seine).

9. PRAXIS

A partir d'un échantillon de 200 magasins de meubles dont 100 de plus de 1000 m² et 100 de moins de 1000 m² répartis dans neuf régions d'enquêtes, PRAXIS effectue des enquêtes dans ces magasins auprès des responsables. Pour une vingtaine de types de produit répartis en 8 familles, on obtient une description qui comporte le style, la forme, les coloris et les matériaux du produit, son prix de vente, son origine et les quantités vendues. Ces résultats permettent en général d'obtenir d'intéressantes analyses des données.

10. LE CNPF, LES CHAMBRES SYNDICALES, LES SYNDICATS PROFESSIONNELS, LES CHAMBRES DES METIERS, DE COMMERCE

Ces chambres disposent de statistiques variées en quantité et en qualité. On trouve surtout des productions, des prix de gros, des prix de détail. Souvent, elles auscultent les grands marchés nationaux (Rungis). Il convient dans tous les cas de se méfier (prix modulés à leur guise, quantités disparues, pertes déraisonnables[1] existent de temps à autres), par contre les chambres de commerce et de métier possèdent en général des centres de documentation complets d'accès libre et gratuit.

11. LES SYNDICATS OUVRIERS

Les syndicats ouvriers produisent et diffusent de nombreux résultats parfois (souvent) de bonne qualité. Citons en particulier la C.G.T.

---

(1) Années de sècheresse par exemple ou violents orages.

qui calcule un indice de prix de détail dont le champ est un ménage ouvrier de la région parisienne ainsi que des statistiques conjoncturelles (salaires, chomage). Le tout étant publié dans la revue "Liaisons sociales".

12. DEUX AUTRES METHODES

• Des données issues de journaux et de revues ("Que choisir", etc...) permettent de faire d'excellentes analyses de données mais aucun contrôle de celles-ci n'est en général possible.

• On peut commander des données à un institut de sondage ou à une entreprise spécialisée dans les études économiques ; cette solution présente l'inconvénient majeur de dépasser en valeur l'épaisseur du portefeuille d'un établissement.

Annexes : adresses

Cette liste ne saurait être exhaustive. La "Documentation Française" publie un index des bibliothèques et centres de documentation officiellement ouverts au public.

INSEE

OBSERVATOIRES ECONOMIQUES REGIONAUX DE L'INSEE

Ajaccio (CORSE : Corse du Sud, Haute-Corse). — Résidence du Parc Belvédère — B.P. 40 — 20176 AJACCIO CEDEX. Tél. 21.28.35

Amiens (PICARDIE : Aisne, Oise, Somme). — 2, rue Robert-de-Luzarches, 80026 AMIENS. Tél. 91.31.87

Besançon (FRANCHE-COMTE : Doubs, Jura, Haute-Saône, Territoire de Belfort). — 2, rue de l'Industrie, 25042 BESANÇON CEDEX. Tél. 80.19.34

Bordeaux (AQUITAINE : Dordogne, Gironde, Landes, Lot-et-Garonne, Pyrénées-Atlantiques). — 349, boulevard du Président-Wilson, 33200 BORDEAUX. Tél. 08.58.17

Caen (BASSE-NORMANDIE : Calvados, Manche, Orne). — 13, rue Paul-doumer, 14037 CAEN CEDEX. Tél. 81.71.11

Clermont-Ferrand (AUVERGNE : Allier, Cantal, Haute-Loire, Puy-de-Dôme). — 52, avenue de Royat, 63400 CHAMALIERES. Tél. 93.87.50

Dijon (BOURGOGNE : Côte-d'Or, Nièvre, Saône-et-Loire, Yonne). — Immeuble Mercure, avenue Albert-1er, 21033 DIJON CEDEX. Tél. 05.31.45

Lille (NORD : Nord, Pas-de-Calais). — 12, boulevard Vauban, 59000 LILLE. Tél. 54.39.36

Limoges (LIMOUSIN : Corrèze, Creuse, Haute-Vienne). — 38, rue François-Chénieux — B.P. 1553, 87031 LIMOGES CEDEX. Tél. 77.16.11

Lyon (RHONE-ALPES : Ain, Ardèche, Drôme, Isère, Loire, Rhône, Savoie, Haute-Savoie). — 84, rue du 1er-Mars 1943, 69625 VILLEURBANNE. Tél. 84.66.81

Marseille (PROVENCE - COTE D'AZUR : Alpes-de-Haute-Provence, Hautes-Alpes, Alpes-Maritimes, Bouches-du-Rhône, Var, Vaucluse). — 10, rue Léon-Paulet (8e), 13285 MARSEILLE CEDEX 2. Tél. 76.42.20

Montpellier (LANGUEDOC-ROUSSILLON : Aude, Gard, Hérauld, Lozère, Pyrénées-Orientales). — Cité administrative, ex-caserne Joffre, 34064 MONTPELLIER CEDEX. Tél. 72.98.67

Nancy (PAYS DE LA LOIRE : Loire-Atlantique, Maine-et-Loire, Mayenne, Sarthe, Vendée). — 5, boulevard Louis-Barthou, 44037 NANTES CEDEX. Tél. 73.02.60

Orléans (CENTRE : Cher, Eure-et-Loir, Indre, Indre-et-Loire, Loir-et-Cher, Loiret). — 43, avenue de Paris, 45018 ORLEANS CEDEX. Tél. 87.71.08

Paris (REGION PARISIENNE : Paris, Essonne, Hauts-de-Seine, Seine-Saint-Denis, Seine-et-Marne, Val-de-Marne, Val d'Oise, Yvelines). — OEP, Tour Gamma A, 195, rue de Bercy, 75582 PARIS CEDEX 12. Tél. 345.70.75

Poitiers (POITOU-CHARENTES : Charente, Charente-Maritime, Deux-Sèvres, Vienne). 1, place Aristide-Briand, 86000 POITIERS. Tél. 88.31.69

Reims (CHAMPAGNE-ARDENNES : Ardennes, Aube, Marne, Haute-Marne). — 1, rue de l'Arbalète, 51084 REIMS CEDEX. Tél. 88.24.12

Rennes (BRETAGNE : Côtes-du-Nord, Finistère, Ille-et-Vilaine, Morbihan). — Le Colbert B.P. 17, 36, place du Colombier, 35031 RENNES CEDEX. Tél. 30.91.90

Rouen (HAUTE-NORMANDIE : Eure, Seine-Maritime). — 8, quai de la Bourse, 76043 ROUEN CEDEX. Tél. 98.43.50

Strasbourg (ALSACE : Bas-Rhin, Haut-Rhin). — 14, rue Adolphe-Seyboth, 67084 STRASBOURG CEDEX. Tél. 32.03.18

Toulouse (MIDI-PYRENEES : Ariège, Aveyron, Haute-Garonne, Gers, Lot, Hautes-Pyrénées, Tarn, Tarn-et-Garonne). — 34, rue des 36-Ponts, 31054 TOULOUSE CEDEX. Tél. 53.36.36

JOURNAL OFFICIEL, 26, rue Desaix, 75732 PARIS CEDEX 15

DOCUMENTATION FRANÇAISE, 29-31, quai Voltaire, 75340 PARIS CEDEX 07

CENTRALE DES BILANS DE LA BANQUE DE FRANCE, 39, rue Croix des Petits Champs, 75001 PARIS. Tél. 508.23.45, poste 6211

EUROSTAT auprès du Journal Officiel ou 5, rue du Commerce, B.P. 1003, LUXEMBOURG. Tél. 49.00.81

O.C.D.E., Direction de l'Information, 2, rue André-Pascal,75775 PARIS CEDEX 16

DAFSA, 125, rue Montmartre, 75002 PARIS. Tél. 233.21.23

SECODIP, rue François-Pedron, 78241 CHAMBOURCY. Tél. 965.56.56

SOFRES, 16-20, rue Barbès, 92129 MONTROUGE. Tél. 657.13.00

PRAXIS, 116ter, rfue Marietton, 69130 ECULLY. Tél. 83.70.52

NIELSEN, Plaza Northbook Illinois, 6002 U.S.A. — 28, boulevard de Grenelle, 75732 PARIS CEDEX 15. Tél. 578.61.20

C.N.P.F., 31, avenue Pierre 1er de Serbie, 75016 PARIS. Tél. 732.61.61 - 723.61.58 723.61.69

C.G.T., 213, rue Lafayette, 75010 PARIS

BIBLIOGRAPHIE GENERALE

(On s'est limité ici aux ouvrages et revues de langue française).

Ouvrages

[1] BENZECRI (J.P.) et ses collaborateurs. *Analyse des données.* Tome 1 : *La taxinomie,* (615 p.) - Tome 2 : *Analyse des correspondances,* (619 p.). Dunod 1973.

[1bis] BENZECRI (J.P.) et ses collaborateurs. *Pratique de l'analyse des données.* Tome 1 : *Analyse des correspondances. Exposé élémentaire* (432 p.) - Tome 2 : *Abrégé théorique. Etude de cas modèle,* (480 p.) - Tome 3 : *Linguistique et lexicologie.* Dunod 1980.

[2] BERTIER (P) et BOUROCHE (J.M.). *Analyse des données multidimensionnelles.* P.U.F., 1975 (270 p.).

[3] CAILLEZ (F) et PAGES (J.P.). *Introduction à l'analyse des données.* Smash, 1976 (616 p.).

[4] CALOT (G). *Cours de calcul des probabilités.* Dunod, 1967 (500 p.).

[5] CALOT (G). *Cours de statistique descriptive.* Dunod, 1969 (519p.).

[6] GUERBER (L) et HENNEQUIN (P.L.). *Initiation aux probabilités.* A.P.M.E.P., 1970 (228 p.).

[7] JAMBU (M). *Classification automatique pour l'analyse des données.* Tome 1 : *Méthodes et algorithmes* (310 p.) — Tome 2 : *Logiciels* (400 p.). Dunod, 1978.

[8] LEBART (L) et FENELON (J.P.). *Statistique et Informatique appliquée.* Dunod, 3e édition, 1975 (442 p.).

[9] LEBART (L), MORINEAU (A) et TABARD (N). *Techniques de la description statistique.* Dunod, 1977 (351 p.).

[10] LERMAN (I.C.). Les bases de la classification automatique. Gauthier-Villard, 1970 (136 p.).

[11] MARCOTORCHINO (J.F.) et MICHAUT (P). *Optimisation en analyse ordinale des données.* Masson, 1979 (211 p.).

[12] ROMEDER (J.M.). *Méthodes et programmes d'analyse discriminante.* Dunod, 1973 (274 p.).

[13] SAPORTA (G). *Théorie et méthodes de la statistique.* Technip, 1979 (388 p.).

[14] VOLLE (M). *Analyse des données.* Economica, 1978 (267 p.).

Revues

On se borne ici aux revues spécialisées ; de nombreuses revues de Sciences Humaines ou Naturelles publient des articles utilisant l'analyse des données.

Cahiers de l'analyse des données (4 numéros par an depuis 1976). Dunod.

Mathématiques et Sciences Humaines (4 numéros par an depuis 1962). Centre de Mathématique Sociale, 54, bd Raspail. Dunod.

Publications de l'Institut de Statistique de l'Université de Paris (4 numéros par an depuis 1952). I.S.U.P., 4, place Jussieu (tour 45-55 E3), 75230 Paris Cedex 05.

Revue de Statistique appliquée (4 numéros par an depuis 1953). C.E.R.S.A., 4, place Jussieu (tour 45-55 E2), 75230 Paris Cedex 05

Statistique et Analyse des données (3 numéros par an depuis 1976). Association des Statisticiens Universitaires, J.M. Bouroche, 4 rue G. Millandy, 92360 Meudon-la-Forêt.

ACHAT EN SOUSCRIPTION
DE LA BROCHURE A.P.M.E.P.

ANALYSE DES DONNÉES

TOME II

- Brochure disponible le 1er octobre 1980
  Souscription ouverte jusqu'au 1er octobre 1980
- Nombre de pages : 320.
  PRIX DE SOUSCRIPTION : 25 F (avec port : 31 F)
- Prix ultérieur : 33 F sans port, 39 F port inclus.

Le Tome 2 complète le Tome 1, par une description détaillée des méthodes d'analyse factorielle (analyse en composantes principales et analyse factorielle des correspondances) et par l'étude d'exemples empruntés à la didactique des mathématiques. Elle permettra donc aux enseignants de mathématiques de mieux maîtriser des techniques d'analyse employées dans les domaines les plus divers.

Un index et une bibliographie générale faciliteront l'utilisation de la brochure.

TABLE DES MATIÈRES

POUR SOUSCRIRE,

Veuillez vous conformer strictement aux indications suivantes (en respectant le compte à rebours) :

3. Remplir complètement et lisiblement le présent bulletin et **l'étiquette** qui servira à l'expédition de l'ouvrage souscrit.

2. Remplir les trois volets d'un chèque postal au compte de l'A.P.M.E.P. Paris 5708-21 N en y faisant figurer le montant correspondant à votre souscription.

1. Envoyer le tout, bulletin de souscription et les trois volets du virement postal, sous enveloppe affranchie, au secrétaire administratif de l'A.P.M.E.P. : M. André BLONDEL, 154, avenue Marcel Cachin, 92320 Châtillon-sous-Bagneux.

SOUSCRIPTION ANALYSE DES DONNÉES - Tome 2

NOM : ..... Prénom : .....

Adresse : .....

Nombre d'exemplaires souscrits : a =

Montant du virement postal : a × =

Date :

ETIQUETTE indiquant très lisiblement l'adresse où vous désirez recevoir l'ouvrage souscrit :

ASSOCIATION DES PROFESSEURS DE MATHÉMATIQUES DE L'ENSEIGNEMENT PUBLIC, 37 RUE JACOB, 75006 PARIS

M .....

.....

.....

□□□□□ .....

ASSOCIATION DES PROFESSEURS DE MATHEMATIQUES DE L'ENSEIGNEMENT PUBLIC

Secrétariat : 37, rue Jacob 75006 Paris

Qu'est-ce que l'A.P.M.E.P. ?

L'A.P.M.E.P. est une association qui regroupe tous les enseignants concernés par l'enseignement des mathématiques "de la Maternelle jusqu'à l'Université". Fondée en 1909, elle regroupe aujourd'hui près de 13 000 enseignants. L'A.P.M.E.P. est un lieu d'échanges, pédagogiques et scientifiques, pour tous les enseignants de mathématiques.

Les Régionales

Dans chaque académie, il existe une section régionale de l'A.P.M.E.P. avec, très souvent, des sections départementales, voire locales. En effet, à la dispersion géographique de ses adhérents, l'A.P.M.E.P. propose un remède : la constitution d'équipes de maîtres, qui enseignent des mathématiques "de la Maternelle jusqu'à l'Université", en dehors de toute hiérarchie administrative, par-dessus les barrières officielles des divers degrés d'enseignement.

Les Journées Nationales

L'A.P.M.E.P. organise chaque année des Journées Nationales qui sont, pour les membres de l'Association, l'occasion de se retrouver. Elles ont, ces dernières années, regroupé de 500 à 800 participants autour de : Pluridisciplinarité [Orléans, 1975]. Problèmes de comportement [Rennes, 1976]. Formation Permanente [Limoges, 1977]. Problèmes, évaluation, erreur [Reims, 1978]. Enseignement, innovation, recherche [Grenoble, 1979]. En septembre 1980 (4 au 7 septembre), le thème sera : Quelle formation pour les enseignants de mathématiques ? [Bordeaux].

Les Publications

L'A.P.M.E.P. édite un bulletin (5 numéros par an) qui réunit des articles de documentation mathématique et pédagogique, et qui rapporte la vie de l'association, tant régionale que nationale. On y trouve notamment les rubriques suivantes : études, études didactiques, dans nos classes, mathématiques et société, examens et concours, manuels scolaires, évaluation, interdisciplinarité, formation des maîtres, informatique, audio-visuel, problèmes, jeux et maths, matériaux pour une documentation, un coin du ciel ...

De plus, l'A.P.M.E.P. publie toute une série de brochures. Ces brochures permettent de répondre à des demandes plus spécifiques de telle ou telle catégorie d'adhérents.
Parmi les dernières brochures parues :
Elem-Math 5 (1979) : Aides pédagogiques pour le Cours Elémentaire.
Activités mathématiques en 4e-3e, tome 1 (1979) : Ouvrage de base, avec ses textes de réflexions générales (assorties d'exemples), et la présentation de 29 activités, référencées à 2 index.
Les manuels scolaires de mathématiques (1979) : Pièce maîtresse d'une réflexion indispensable. Exemples pris dans le premier cycle... mais aisément transposables.
Pour une mathématique vivante en Seconde (1979) : 21 exemples, très variés,... et à suivre !
Pavés et bulles (1978) : Met en évidence l'efficacité d'outils mathématiques. Etablit de beaux résultats (post-bac surtout).
Calculatrices quatre opérations (1979) : Elémentaire et premier cycle.
Du quotidien à la mathématique (1979) : Une expérience en formation d'adultes (fiches de travail commentées, également utilisables dans le premier cycle).

Le Présent

L'A.P.M.E.P., association représentative des enseignants de mathématiques, agit comme telle vis-à-vis des syndicats, des associations d'enseignants, d'autres disciplines, des associations de parents d'élèves, ainsi que des Ministères de l'Education et de l'Université. Par exemple, actions à propos des programmes, ... ; intervention de novembre 1979 auprès du Ministère de l'Education (ce qui a permis d'obtenir une heure de travaux dirigés pour toutes les Secondes "Indifférenciées" de la rentrée 1981, alors qu'aucune n'était prévue).

L'Avenir

Après avoir obtenu la création des IREM (puis lutté pour leur maintien), l'A.P.M.E.P. est à la pointe du combat pour une véritable formation permanente, dont elle a défini les principes dans son Texte d'Orientation 1978 (caractère non obligatoire ; formation intégrée dans le service des enseignants ; large indépendance vis-à-vis de la hiérarchie ; ...).

Texte d'Orientation

Après les Chartes de Chambéry (avril 1968) et de Caen (mai 1972), l'A.P.M.E.P. a actualisé ses positions fondamentales par son Texte d'Orientation (1978). Les principales préoccupations des enseignants de Mathématiques y sont abordées et de nombreuses propositions, à court et à long terme, sont faites, permettant une réforme en profondeur de l'enseignement des mathématiques. [On peut se le procurer gratuitement, en écrivant au Secrétariat de l'A.P.M.E.P. (adresse ci-dessus)]

L'A.P.M.E.P. s'intéresse donc à toutes les questions qui concernent l'enseignement des mathématiques, depuis les premières initiations jusqu'aux études supérieures, sans oublier la formation permanente des non-enseignants et des enseignants. Aussi ne pouvez-vous vous désintéresser de l'A.P.M.E.P. et des possibilités d'action qu'elle vous offre.

L'A.P.M.E.P. a besoin des forces, de l'expérience et de l'action du plus grand nombre d'enseignants de mathématiques. Son efficacité, les services qu'elle vous rend ou pourrait vous rendre, tiennent au nombre et au dynamisme de ses membres. Si vous ne les avez pas encore rejoints, faites-le donc sans tarder.

Juin 1980

RECHERCHE INTER-IREM 1973-78, EN GEOMETRIE DE 4e - 3e, DITE « O.P.C. » : REFLEXION CRITIQUE ET EVALUATION

brochure éditée par l'A.P.M.E.P.

- Nombre de pages : environ 200.
- Prix : **F. 28** (sans port) ; **F. 33** (frais de port inclus)

Parmi toutes les recherches suscitées, en géométrie premier cycle, par les difficultés surgies à l'occasion de la mise en œuvre des programmes de 1971, la recherche dite « O.P.C. » a, de 1973 à 1978, sous la direction de Charles PEROL, IREM de Clermont, intéressé plusieurs IREM, cinq au départ, puis davantage.

La plupart des équipes O.P.C. refusaient la dichotomie affine-métrique, et avaient alors des programmes différents de ceux de 1971 (qui n'ont pas été pour autant exactement ceux de 1978).

Toutes mettaient l'accent sur les démarches expérimentales et s'intéressaient à des objectifs prioritairement attentifs aux capacités des élèves (expérimenter, conjecturer, s'auto-contrôler, etc.).

Après une « Introduction », la présente brochure comporte d'abord les études suivantes : l'O.P.C. et les programmes 1978 — Une équipe O.P.C. et l'axiomatique — Usage des transformateurs articulés — La continuité dans l'enseignement mathématique — Les thèmes pris comme centres d'activités — Le géométrique et le numérique — Repérage dans le plan.

Elle comporte ensuite une « Evaluation de l'expérience O.P.C. », très fouillée, à partir de tests d'entrée et de sortie méthodiquement et scientifiquement analysés par Régis GRAS.

BROCHURE A.P.M.E.P.

CALCULATEURS PROGRAMMABLES ET ALGEBRE DE QUATRIEME

(Une recherche -inter-I.R.E.M.)

Cette brochure présente un compte rendu complet de l'expérience menée, depuis 1974, par un groupe de recherche inter-IREM.

Pour cette expérience, une vingtaine de professeurs ont accepté d'intégrer totalement des calculateurs programmables de différents types dans leur cours d'algèbre de quatrième. Parallèlement, un nombre égal de professeurs a dispensé un enseignement de même nature, mais sans calculateur.

Une évaluation, mise en place à l'aide de psychologues, et par le biais d'un traitement informatique, a permis d'estimer de manière assez fine l'apport du calculateur quant aux objectifs choisis.

Les maîtres du premier cycle trouveront dans cette publication :

— des extraits du matériel pédagogique élaboré à l'usage soit des professeurs, soit des élèves ;

— un exemple de méthode d'évaluation, analysée de façon critique;

— une première approche explicative des changements que peut apporter ce nouvel outil pédagogique en classe de mathématique.

120 pages - 20 F (avec port : 24 F).

Imprimerie VAUDREY - LYON

ISBN 2-902680-04-X