LES CAHIERS DE L'ANALYSE DES DONNÉES

J. P. BENZÉCRI

Histoire et préhistoire de l'analyse des données. Partie V L'analyse des correspondances

Les cahiers de l'analyse des données, tome 2, n° 1 (1977), p. 9-40

<http://www.numdam.org/item?id=CAD_1977__2_1_9_0>

Les Cahiers de l'Analyse des Données
Vol. II - 1977 - n° 1 - p. 9-40

HISTOIRE ET PRÉHISTOIRE DE L'ANALYSE DES DONNÉES

Partie V - L'analyse des correspondances

par J. P. Benzécri (1)

Avertissement : dans le présent article, moins encore que dans ceux qui l'ont précédé, nous ne pouvons prétendre à l'exhaustivité. Les recherches contemporaines qui recourent à la statistique multidimensionnelle -- analyse factorielle, analyse discriminante, classification automatique, régression etc... - se multiplient à perte de vue. De la préhistoire et de l'histoire de la statistique, nous n'avons voulu retenir que ce qui, selon nous, sert à l'analyse des données ; de l'analyse des données elle-même nous n'exposerons ici que ce qui nous a personnellement servi. C'est pourquoi cette dernière partie est placée sous le titre de l'analyse des correspondances, méthode qui bien mieux que tout autre nous a permis de découvrir les faits de structure que recèle un tableau de données quel qu'il soit.

3. *L'analyse des correspondances*

3.1 *Convergence* :

Le terme même d'analyse des correspondances remonte à l'automne de 1962, et le premier exposé de la méthode sous ce titre fut donné par J. P. Benzécri au Collège de France dans une leçon du cours Peccot de l'hiver 1963. En nous référant au terme même, nous évitons de nous prononcer d'abord, quant à la définition des facteurs issus d'un tableau rectangulaire de nombre positifs, sur des questions de priorité qu'un article récent(*) pourrait soulever, mais que nous préférons réduire à leur juste proportion sinon à leur solution définitive par un exposé chronologique, où seront scrupuleusement notées les rencontres successives de l'analyse des correspondances avec les travaux d'autres écoles (cf §§ 3.4 & 3.5.2).

L'analyse des correspondances telle qu'on la pratique en 1977 ne se borne pas à extraire des facteurs de tout tableau de nombres positifs. Elle donne pour la préparation des données, des règles, telles que le codage sous forme disjonctive complète (§ 3.7.3) ; aide à critiquer la validité des résultats, principalement par des calculs de contribution (§ 3.8.4) ; fournit des procédés efficaces de discrimination et de régression (§ 3.8.2) ; se conjugue harmonieusement avec la classification automatique (§ 3.8.3). Ainsi une méthode unique dont le formulaire reste simple est parvenue à s'incorporer des idées et des problèmes nombreux apparus d'abord séparément, certains depuis plusieurs décennies. Nous expliquerons ce succès par deux causes : d'une part, la formule initiale de la distance distributionnelle permet à elle seule de donner à un tableau de nombres positifs une structure mathématique compensant, autant que possible, l'arbitraire dans le choix des pondérations et subdivisions des faits ; d'autre part, de nombreux

(1) *Professeur : Laboratoire de Statistique ; Université Pierre & Marie Curie ; Paris.*

(*) *M. O. Hill : Correspondence Analysis : A neglected Multivariate Method in Appl. Statist. T. 23 pp 340-354 (1974).*

chercheurs (les tomes I et II du traité de l'*Analyse des données* comptent 70 auteurs ; qui ne sont pas les seuls à avoir contribué aux progrès de l'analyse des correspondances) ont eu pour programme non d'inventer chacun une variante nouvelle de quelque méthode statistique en cours, mais de réduire à l'unité le traitement des problèmes posés par les données les plus diverses. Il serait sans doute vain d'exposer jour après jour l'histoire détaillée de tous ces efforts ; mais nous croyons utile de distinguer les principales étapes en les illustrant d'exemples, sans prétendre citer tous les auteurs en exacte proportion de leurs mérites. Des étudiants qui entreprennent aujourd'hui l'étude de l'Analyse des données trouveront dans cette esquisse chronologique la raison des perfectionnements progressifs qui leur sont enseignés en un seul cours ; les statisticiens déjà instruits ailleurs se reconnaîtront mieux dans les travaux d'une école certes indépendante mais qui ne saurait prétendre être isolée. Ceux du dedans et du dehors verront dans la relative lenteur de ces progrès où interviennent pourtant presque uniquement des idées assez simples, naturelles et bien connues, une nouvelle confirmation de ce que, comme aimaient à le rappeler les docteurs médiévaux dans leur latin traduit d'Aristote : *mens humana se habet ad manifestissimas sicut oculus noctuae ad lumen solis* : l'esprit humain est, devant l'évidence comme l'oeil de la chouette exposé à la lumière du soleil!

3.2 *Tableaux de contingence et distance distributionnelle* :

3.2.1 La méthode inductive en linguistique : L'analyse des correspondances a été initialement proposée comme une méthode inductive d'analyse(*) des données linguistiques. Expliquons en quelles circonstances. Vers 1960 la traduction automatique semblait un objectif assez rapidement accessible (on sait qu'au contraire en 1977, les difficultés qu'on avait d'abord sous estimées sont jugées par beaucoup insurmontables ; jugement auquel nous ne souscrirons pas cependant). L'Association pour la Traduction Automatique A.T.A.L.A. fondée à l'initiative de E. Delavenay, aidait efficacement aux rencontres des chercheurs français ; auxquels le regretté Pr. J. Favard ouvrit bientôt un séminaire spécialisé. La linguistique mathématique, à laquelle nous avait invité Y. Lecerf alors détaché à l'Euratom, était dominée par le renom de N.Chomsky dont le petit volume *Syntactic Structures* s'imposait à tous. Parmi d'autres thèses qu'il n'y a pas lieu d'exposer ici(**), N. Chomsky affirme là qu'il ne peut exister de procédure systématique pour déterminer la grammaire d'une langue ou plus généralement les structures linguistiques, à partir d'un ensemble de données telles qu'un recueil de textes que les linguistes nomment *corpus*. Donc, pour Chomsky la linguistique ne peut être inductive (s'élever par une méthode explicitement formulée des faits aux lois qui les régissent) ; elle doit être déductive (en ce sens que partant d'axiomes elle engendre des modèles des langues réelles). Cette thèse(idéaliste ; en ce qu'elle tendait à séparer le jeu de l'esprit, des faits qui en sont l'inspiration et l'objet) nous déplaisait ; et à défaut d'un algorithme universel pour passer de 10.000 pages de texte d'une langue à une syntaxe doublée d'une sémantique, nous prétendions par la statistique offrir au linguiste une méthode inductive efficace pour traiter utilement des tableaux de données qu'on pouvait immédiatement recueillir, avec à l'horizon

---

(*) *Pour la philosophie l'analyse des correspondances qui traite simultanément de grands ensembles de faits et les confronte afin d'en découvrir l'ordre global relève plutôt de la synthèse (étymologiquement synthétiser veut dire mettre ensemble) et de l'induction que de l'analyse et de la déduction (distinguer les éléments d'un tout ; et considérer les propriétés des combinaisons dont ceux-ci sont susceptibles) cf § 2.3.6 ; mais les termes d'analyse factorielle et d'analyse des données ayant pris racine, nous les conservons.*

(**) *Pour une critique de certaines de ces thèses, cf Linguistique et Mathématique in Revue Philosophique pp 309-374 (1966).*

l'ambitieux étagement des recherches successives ne laissant rien dans l'ombre, des formes, des sens et du style.

Sans entreprendre une leçon de linguistique(*) montrons comment ce programme qui requerrait l'analyse de tableaux de contingence conduisit à définir la distance distributionnelle - plus communément appelée aujourd'hui distance du χ^2 - qui est à la base de l'analyse des correspondances.

3.2.2 *Les données distributionnelles* : Pour édifier inductivement la linguistique à partir de données non préalablement élaborées selon les vues *a priori* du linguiste, on doit tendre à regarder les mots, les phrases et les discours comme des suites d'éléments (suites de lettres ou de syllabes pour les mots, suites de mots pour les phrases) dont il faut découvrir suivant quelles règles certaines combinaisons sont seules permises parmi un bien plus grand nombre de combinaisons interdites (mots imprononçables ; phrases incorrectes ou absurdes). Dès lors un mot (ou un segment, suite de plusieurs mots) sera caractérisé par l'ensemble de tous les contextes dans lequel il est permis de l'insérer pour obtenir une phrase correcte. Pour définir les phrases correctes on pourra soit recourir à des juges (auxquels on demandera : cette phrase est-elle correcte, est-elle absurde ?) soit s'attacher au traitement d'un corpus clos (l'ensemble des phrases correctes étant par définition l'ensemble des phrases données ; auquel il convient d'adjoindre celles qui en un certain sens en diffèrent le moins ; nous reviendront sur ce dernier point). Pratiquement il convient d'assigner à la recherche des étapes successives dont les premières soient immédiatement accessibles et dont la progression semble assez douce pour ne devoir jamais s'interrompre ! On considérera donc d'abord les phrases les plus courtes qui suivent on le sait, des modèles simples tels que sujet-verbe ; sujet-verbe-complément etc. Ainsi la notion indéfiniment extensible de *contexte permis* pour un mot se trouve réduite aux contextes ne comportant qu'un ou deux mots. On aboutit à donner pour base à la statistique linguistique des tableaux tels que le suivant. Soit I un ensemble fini de noms (les lignes du tableau), J un ensemble de verbes (les colonnes du tableau) : à l'intersection de la ligne i et de la colonne j on inscrit le nombre k(i,j) de fois que dans un certain corpus le nom i a été trouvé sujet du verbe j. (Ou encore J est un ensemble d'adjectifs ; et k(i,j) est le nombre de fois que le nom i a été trouvé qualifié par l'épithète j ; etc..). Dans un tel tableau le contexte du nom i est réduit au verbe j ; et réciproquement le contexte du verbe j est réduit au nom i ; si $k(i,j) \neq 0$, j est un contexte admissible pour i (et i pour j) ; l'association de i avec j est d'autant plus licite que k(i,j) est plus élevé. Plus précisément si l'on considère un nom i il convient de mesurer l'importance relative, pour ce nom, du contexte j, par le quotient $k(i,j)/k(i) = f^i_j$ du nombre k(i,j) des emplois de i avec j au nombre total des emplois de i(k(i) = total de la ligne $i = \Sigma\{k(i,j) \mid j \in J\}$). La suite des nombres f^i_j caractérisant l'affinité d'un nom i donné avec tous les verbes j,j',j" pourra être appelée *profil* du nom i et notée $f^i_J = \{f^i_j \mid j \in J\}$. Deux noms i et i' seront synonymes (du point de vue de leur association avec les verbes) si ils ont même profil, i.e. si quel que soit j : $f^i_j = f^{i'}_j$; cette synonymie est acceptable en ce sens que deux êtres qui *courent, poussent, chantent,* etc. avec la même fréquence ne peuvent que se ressembler. (De même pour un verbe j on définit $f^j_i = k(i,j)/k(j)$ où k(j) est le total de la colonne j ; et un profil : $f^j_I = \{f^j_i \mid i \in I\}$). Pratiquement, il est peu vraisemblable que deux noms i et i' (ou deux verbes j et j') aient exactement le même

(*) *Pour une introduction à ces recherches, simple mais centrée sur le problème linguistique, cf : "Combattre pour la linguistique" in Mathématiques et Sciences Humaines n° 35 (1971).*

profil ; mais la similitude des profils peut être plus ou moins grande ; ce qui pose le problème fondamental d'une représentation spatiale de l'ensemble des profils. Incidemment notons que s'ouvre ici une voie pour étendre un corpus fini donné : on y adjoindra les phrases obtenues en substituant aux mots, d'autres mots de profil non identique, mais voisin.

Le premier linguiste que nous entretînmes de ces spéculations fut notre collègue de l'Université de Rennes, J. Gagnepain, qui ne fut ni surpris ni enthousiasmé mais nous répondit en substance : "Ce sont là les idées de Harris ; mais ce linguiste est le seul à croire aux méthodes purement inductives que vous prétendez appliquer à grand renfort de statistique". Nullement découragé par ce verdict, nous nous hâtâmes de rechercher les travaux de Z. S. Harris, où brillait cette définition digne d'être retenue : "On appelle *distribution* d'un mot l'ensemble de ses environnements possibles".

3.2.3 *L'espace des profils* : Restait à définir mathématiquement l'espace des profils. Il était naturel de penser à l'analyse factorielle (§ 2.4) ; méthode dont la pratique était en 1962 réservée aux psychologues, voire aux biométriciens, mais dont pour un mathématicien les principes sont clairs. A un ensemble I d'individus i, chacun décrit par une série J de mesures (m(i,j) étant le résultat de la mesure j effectuée sur l'individu i) est associé un nuage de points i de R^J (un point par individu ; une coordonnée par variable) ; on introduit dans R^J de nouvelles coordonnées appelées facteurs (combinaisons linéaires de mesures primaires j) comptées sur des axes orientés suivant les directions principales du nuage ; quelques facteurs suffisant à exprimer la diversité des individus en résumant de multiples variables. Pour le praticien de 1960 la réduction des variables apparaît généralement liée à des hypothèses de normalité (loi de Laplace Gauss multidimensionnelle) et repose sur des calculs traditionnels de corrélation ; tandis que l'ajustement à un nuage d'un système d'axes principaux d'allongement, d'une *échelle multidimensionnelle* est un problème de géométrie relevant de méthodes qui semblent nouvelles, appelées justement par les psychologues américains : *multidimensional scaling* (cf § 2.5). Pour le mathématicien instruit de l'algèbre linéaire, de la géométrie euclidienne multidimensionnelle, du calcul tensoriel (dans l'enseignement français, G. Bouligand avait fait oeuvre de pionnier ; A. Lichnerowicz développait avec élégance les théories d'Einstein ; et N. Bourbaki sans s'adonner au calcul tensoriel donnait des produits tensoriels une définition à méditer) il n'y a là qu'un problème unique : les individus sont des points ou vecteurs d'un espace ; les variables sont des formes linéaires, ou vecteurs de l'espace dual ; (les coefficients de corrélation des statisticiens s'identifient aux produits scalaires des géomètres) espace et dual sont isomorphes si on a fixé une métrique euclidienne (ou formule de distance). Pour appliquer les formules classiques il n'y a qu'une question à résoudre : fixer judicieusement cette métrique.

Ici intervient le *principe d'équivalence distributionnelle* : la distance d(i,i') entre deux noms i et i' ne doit pas être modifiée si on identifie deux verbes j et j' qui sont des synonymes distributionnels (ont même profil) ; i.e. si on remplace les colonnes j et j' (qui sont proportionnelles l'une à l'autre) par une nouvelle colonne j" somme des deux précédentes (et donc également proportionnelle à celles-ci). Disons pour faire image que si *parler* et *dire* admettent les mêmes sujets dans les mêmes proportions, on peut identifier ces deux verbes. Le principe d'équivalence distributionnelle, complété par l'exigence mathématique que la formule de distance soit quadratique (comporte une somme de carrés avec des coefficients) suffit à fixer la *distance distributionnelle*(*)

$$d^2(i,i') = \Sigma\{(f_j^i - f_j^{i'})^2/f_j \mid j \in J\}.$$

(*) *Le terme équivalent de distance du χ^2 nous fut suggéré par une remarque de K. Krickeberg ; qui sans prendre intérêt à nos recherches y reconnut incidemment une formule classique dans l'épreuve du χ^2 (cf § 2.2.6)*

Exiger une formule quadratique peut sembler arbitraire au non-mathématicien : il est vrai que la formule ci-dessous :

$$\delta(i,i') = \Sigma\{|f_j^i - f_j^{i'}| \mid j \in J\}$$

satisfait également au principe d'équivalence distributionnelle ; elle semblera même plus simple au profane ; mais elle ne permet pas d'utiliser la géométrie euclidienne multidimensionnelle ; elle donnera des résultats qui qualitativement ressembleront à ceux obtenus par la distance distributionnelle quadratique ; mais au prix de calculs plus compliqués et sous une forme moins commode. Sans permettre à l'outil mathématique de défigurer le réel, on doit lui concéder que la transmission à l'esprit humain d'un vaste ensemble de données synthétisé (résumé ; rendu perceptible par le calcul) ait ses lois propres. (On se souvient que le primat de la géométrie euclidienne est admis par Torgerson, cf § 2.5.2).

3.2.4 Premier état de l'analyse des correspondances : Résumons donc le premier état de l'analyse des correspondances (Hiver 1963). Comme données, un tableau rectangulaire I x J : si les k(i,j) inscrits dans le tableau sont entiers (nombre de fois que i a été trouvé associé à j) on parle de *correspondance statistique* (ou de tableau de contingence) ; si les k(i,j) sont astreints à valoir 0 ou 1 (0 si l'association de i à j est impossible ; 1 si elle est possible) on a une *correspondance ensembliste* ; si les k(i,j) sont des probabilités (valeurs limites de fréquences observées ; ou valeurs postulées selon un modèle hypothétique) c'est une *correspondance probabiliste*. Il n'est pas question de tableaux de mesures variant continûment (e.g. mensurations sur un crâne) ; le domaine visé est la linguistique ; mais les tableaux de données linguistiques publiées sont rares(*) ; au contraire les psychologues recensent communément en un tableau de contingence les résultats de leurs expériences : tableau S x R ; k(s,r) est le nombre de fois que le stimulus s a évoqué la réponse r. Vers 1960, les tableaux de contingence étaient la donnée de prédilection du *Multidimensional Scaling* ; tandis que les tableaux de données où les rôles des ensembles I et J sont nettement dissymétriques (I individus ; J variables, par exemple des notes à des épreuves psychotechniques) offraient matière à analyse factorielle (e.g. en composantes principales ; cf §§ 2.4 & 2.5). Le premier tableau traité par nous (analysé sans le secours de l'ordinateur moyennant des simplifications assez hasardeuses) fut le tableau de correspondance logique qui figure dans tout manuel moderne du chinois : en ligne les consonnes (ou préphonèmes) ; en colonnes les finales vocaliques ; à la croisée d'une ligne c et d'une colonne v, non le 0,1 du mathématicien, mais l'orthographe (e.g. dans la transcription de l'*Academia Sinica)* du monosyllabe cv s'il existe (si l'association n'est pas permise, la case est blanche). Le terme lui-même de correspondance paraîtra naturel pour désigner le système des associations entre les éléments de deux ensembles I et J. Son choix n'est toutefois pas étranger à des soucis philosophiques moins apparents. Aux structuralistes qui affirment que les objets n'existent pas, seules existent les relations, nous voulons répondre que les objets existent, mais ne nous sont révélés que par les relations : il faut ici citer Aristote (de l'âme L. III Ch. 1)

ζητήσειε δ' ἄν τίς τίνος ἕνεκα πλείους ἔχομεν αἰσθήσεις ἀλλ' οὐ μίαν μόνην....

(*) *On trouvera dans [Ana. Ling.] une revue des analyses de correspondances réalisées en linguistique jusqu'en 1974. Les efforts des chercheurs de Nancy, Saint-Cloud, Vincennes, Montpellier coordonnés par A. Salem, commencent seulement à porter leurs fruits. Il n'est pas exagéré de dire que nos espérances n'ont pas été déçues, bien au contraire ; mais la collecte des données est oeuvre de bénédictins qui requiert des congrégations de chercheurs, armés d'ordinateurs! Nous préparons la publication d'un recueil rendant compte de l'ensemble des analyses réalisées jusqu'à ce jour en linguistique.*

"On se demandera pourquoi nous avons plusieurs sens et non un seul! N'est-ce pas pour révéler les [réalités sensibles] dérivées et communes comme mouvement grandeur et nombre ; car s'il n'y avait que la vue et la vue du blanc [seulement]... tout semblerait ne faire qu'un pour être toujours ensemble, par exemple couleur et grandeur. Mais comme ces sensibles communs [i.e. mouvement grandeur et nombre] se retrouvent dans un autre sens [que la vue] il est clair qu'ils sont quelque chose en propre". Nous dirons que la connaissance est à la rencontre de plusieurs voies. (Sur ce thème, on a au § 2.2.7 exposé les vues de K. Pearson).

A partir du tableau I x J, on construit deux nuages $N(I)$ et $N(J)$ (dans notre exemple principal : nuage des noms et nuage des verbes), i.e. deux ensembles de points munis de masses et distances ; chacun de ces nuages est naturellement placé dans un espace ambiant euclidien ; dans cet espace on recherche (par un calcul classique pour les axes principaux d'inertie) les droites ou axes qui en un certain sens (exactement au sens des moindres carrés cf § 2.4.4) s'ajustent le mieux au nuage. Afin de voir le nuage dans un espace accessible à nos sens, on projette celui-ci sur un plan engendré par deux de ses axes principaux. Pour l'heure, les deux nuages $N(I)$ et $N(J)$ bien que construits symétriquement d'après un même tableau de données ne sont pas unis : ils flottent dans deux espaces différents ; et l'on ne songe pas à identifier les axes de l'un et ceux de l'autre.

3.3 Représentation simultanée de deux ensembles en correspondance :

3.3.1 Une construction géométrique : Une pareille identification est cependant suggérée par l'intelligence des données. Si un axe du nuage $N(I)$ (ou le facteur, coordonnée mesurée sur cet axe) figure une gradation dans une qualité des noms (par exemple l'activité : depuis l'inanimé, extrémité négative ; jusqu'à l'animé, extrémité positive), et que cette qualité est révélée par les affinités entre noms et verbes, n'est-il pas naturel qu'on la retrouve dans les verbes? Un artifice géométrique permet de définir des axes sur lesquels se projettent simultanément les nuages I et J. Considérons les espaces ambiants des nuages $N(I)$ et $N(J)$ comme deux sous-espaces perpendiculaires d'un même espace euclidien (leur somme directe(*) ; dont la dimension sera la somme de celles des deux premiers) ; à tout couple (i,j) (d'un nom i et d'un verbe j) on peut associer le milieu (dans l'espace somme directe) du segment joignant le point i du nuage $N(I)$ au point j du nuage $N(J)$; en attribuant à ce milieu la masse k(i,j) (nombre des associations de i avec j, inscrit au tableau de correspondance analysé) on a dans l'espace somme un *nuage des couples* $N(I \times J)$. Le nuage a des axes principaux sur lesquels on projettera les deux nuages $N(I)$ et $N(J)$ (qui sont inclus dans l'espace somme). Un exemple très simple suggère que cette construction peut donner des résultats significatifs : I et J sont réduits à deux éléments : $I = \{i_1 , i_2\}$; $J = \{j_1 , j_2\}$; les associations se font uniquement entre i_1 et j_1 ou entre i_2 et j_2 : les couples (i_1 , j_2) et (i_2 , j_1) ont masse nulle. On a alors dans l'espace somme le dessin suivant (fig 3-1) : l'axe principal relie les deux couples lourds (i_1 , j_1) et (i_2 , j_2) ; sur cet axe, i_1 et j_1 se projettent confondus ainsi que i_2 et j_2 ; ce qui est agréable et donne à espérer qu'en général la représentation simultanée placera un point i et un point j d'autant plus près l'un de l'autre que le taux d'association entre ces deux éléments est plus élevé relativement à leurs masses. Cependant rien ne laisse deviner que les facteurs ainsi mesurés dans l'espace somme en projetant $N(I)$ et $N(J)$ sur les axes du nuage des couples $N(I \times J)$ soient ceux mêmes définis d'autre part en considérant séparément $N(I)$ et $N(J)$. L'expérience allait montrer que tel est le cas.

(*) *On se souviendra de distinguer de la réunion ensembliste, la somme directe de deux espaces vectoriels ; somme directe qui en tant qu'ensemble coïncide avec le produit ensembliste.*

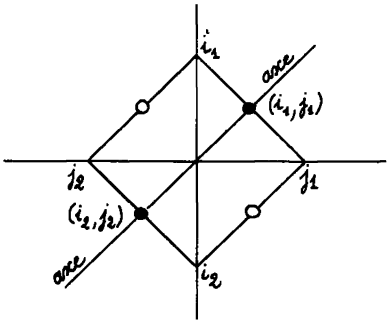

Figure 3-1: Un exemple très simple d'analyse du nuage des couples.

3.3.2 *Expériences et démonstration* : A l'automne de 1963 fut créé à Rennes sur l'initiative du Doyen Y. Martin un laboratoire de calcul équipé d'un 1620 IBM (modeste ordinateur dont le sigle a peu de chances de passer à la postérité!). B. Cordier (depuis madame J. P. Escofier) écrivit rapidement le premier programme d'analyse de correspondance : pour l'analyse du nuage $N(I)$ ou $N(J)$ (associé chacun à un seul des deux ensembles en correspondance) ; puis un deuxième programme pour l'analyse du nuage des couples et la représentation simultanée de I et J. Quant à l'interprétation des facteurs, les premières analyses (recensées pour la plupart dans la thèse de B. Cordier) sont aussi satisfaisantes qu'il est possible (compte tenu de la petite taille des tableaux ; et de la structure simple des données). Mais le résultat le plus précieux fut d'abord que les deux nuages $N(I)$ et $N(J)$, analysés séparément avaient les mêmes moments principaux d'inertie (on parle encore de valeurs propres λ_1, λ_2, etc... associées aux axes principaux successifs) coïncidence qui déjà surprenait ; et ensuite que les facteurs sur I et J issus de l'analyse du nuage des couples $N(I \times J)$ étaient les mêmes que ceux obtenus séparément par analyse de $N(I)$ et $N(J)$.

Dès lors il s'imposait de démontrer ces résultats d'expérience : ce que fit rapidement B. Cordier. Dans un exposé de 1977, le nuage des couples n'est que l'occasion d'un exercice de calcul matriciel ; et le lecteur non averti s'étonne qu'un paragraphe lui soit consacré dans la leçon [Dis. χ^2 Corr.] du traité d'Analyse des données. Mais à l'origine ce fut l'occasion d'entreprendre une démonstration très utile dont les calculs auraient découragé si l'issue n'en avait été assurée. Dans la thèse de B. Cordier on trouve l'équivalence des facteurs issus des trois nuages $N(I)$, $N(J)$, $N(I \times J)$; avec la formule de transition qui permet de n'analyser qu'un seul nuage (celui qui requiert le moins de calculs) puis de passer très simplement des facteurs trouvés sur un ensemble, aux facteurs sur l'autre ; et qui, de plus, répond à notre désir que dans la représentation simultanée un élément i (resp. j) soit entouré des j (resp. i) avec lesquels il s'associe le plus (de façon précise, à un coefficient près égal à la racine carrée de la valeur propre λ_α, i se projette sur l'axe α au barycentre du système des j affectés des masses f_j^i ; d'où le nom de formule barycentrique donné encore à la formule de transition ; cf infra § 3.5). A ces formules B. Cordier adjoignit bientôt celle de reconstitution du tableau des données à partir des facteurs (reconstitution approchée très utile si on se borne aux premiers facteurs seuls interprétables ; i.e. si l'on réduit la dimension des nuages par ajustement à un sous-espace de dimension 2 ou 3 ; en négligeant les autres dimensions qui relèvent de fluctuations d'échantillonnage, plutôt que de structure, cf *infra* § 3.4). Tout cela au

niveau de la science mathématique enseignée aujourd'hui à l'Université peut être qualifié d'évident ; mais on ne l'a trouvé qu'au cours du traitement statistique des données sur ordinateur.

L'isomorphisme familier au mathématicien entre espace vectoriel et espace dual, ou autrement dit entre vecteur et forme linéaire (cf *supra* § 3.2.3), joint à la parfaite symétrie de rôle entre les deux ensembles I et J (dans le cas des données que nous avions d'abord visées lignes = noms ; colonnes = verbes) avait permis d'unifier les deux points de vue de l'analyse factorielle (d'un ensemble de *variables*) et des échelles multidimensionnelles (*multidimensional scaling* : représentation d'un nuage de *points*). Chaque i (*nom*), nous l'avions vu dès le départ, est à la fois un point caractérisé par ses associations avec les éléments j de l'autre ensemble (*verbes*) et une variable qui réciproquement caractérise ceux-ci (*les verbes*) ; et de même pour les j : il était maintenant prouvé que les deux points de vue conduisent aux mêmes facteurs (à condition de choisir judicieusement les coefficients de pondération).

3.4 *Rencontre avec l'école américaine* :

3.4.1 *R. N. Shepard* : Enseignant à Rennes de 1960 à 1965 nous eûmes le plaisir de collaborer avec notre ami le psychologue J.F. Richard (présentement à l'université de Vincennes). Par J.F. Richard nous bénéficiâmes des avis de H. Rouanet, chercheur parisien très averti des travaux de l'Ecole américaine. Ainsi notre analyse des correspondances se trouva mise en parallèle avec l'*analyse des proximités* que développait alors R.N. Shepard. Les données traitées par R.N. Shepard - des matrices de contingence issues d'expériences d'associations entre stimulus et réponse - étaient exactement du format que nous recherchions ; et comme nous le psychologue américain entendait représenter dans un espace de faible dimension (généralement dans une carte plane) les relations de proximité entre éléments d'un ensemble I ; mais au lieu de traiter les distances, d(i,i'), il se bornait à considérer les inégalités entre celles-ci (e.g. i est plus proche de i' que i" ne l'est de i'") ; et au prix de ces informations très affaiblies parvenait pourtant à déterminer une figure euclidienne de dimension p choisie (et ce de façon à peu près univoque si toutefois les données se prêtaient à une représentation de dimension p). Mathématiquement parlant,c'est la méthode de Shepard qui - de beaucoup - pose les problèmes les plus profonds ; mais quant aux résultats statistiques, nous sommes convaincus que l'avantage est à l'analyse des correspondances. Celle-ci permet la représentation simultanée de deux ensembles ; traite plus rapidement des données beaucoup plus amples (la convergence de l'algorithme de Shepard est au contraire souvent hasardeuse), parvient à extraire des représentations significatives de dimension plus élevée ; et est enrichie de règles de codage (§ 3.7) et d'interprétation (§ 3.8).

Quoi qu'il en soit de ce jugement, nous entrâmes en correspondance avec R.N. Shepard qui eu la bienveillance de nous inviter à Murray Hill, au laboratoire de la Compagnie des téléphones Bell. Notre hôte, statisticien seulement pour servir la psychologie, et psychologue par amour de la philosophie, ne portait plus alors sur l'analyse des données que les regards souriants d'un sage ; mais il nous présenta à D. Caroll qui peu soucieux, au contraire, de stimuler les réponse des rats ou des hommes, employait joyeusement toute son ingéniosité - qui est grande - à agiter des données dans un ordinateur comme on ferait de perles dans un kaléïdoscope.

3.4.2 *De D. Caroll à Eckart & Young* : Remarquablement secondé par Madame J.J. Chang, D. Caroll a conçu d'intéressants algorithmes dont l'un destiné à l'analyse des tableaux à plus de deux entrées (e.g. tableaux cubiques ou parallépipédiques : k(i,j,t) = association entre i et j pendant l'année t) est communément utilisé au Etats-Unis. Il a quant aux doctrines de l'analyse factorielle de fortes convictions dont

il ne me fit pas mystère : "Faisandé, leur bazar de "rotations" et de "communautés" !(cf § 2.4). Il n'y a derrière tout cela qu'un problème pur et dur ; résolu depuis trente ans par Eckart et Young(*) : la recherche du tableau de rang p (p fixé *a priori*) le plus proche d'un tableau n x n donné (k(i,j)) , au sens des moindres carrés. La solution est simple et unique ; on a des fonctions (ou facteurs) φ_α et ψ_α telles que:

$$k(i,j) = \Sigma_\alpha \varphi_\alpha(i) \quad \psi_\alpha(j) \ ;$$

cette somme, arrêtée aux facteurs α de rang 1 à p, donne l'approximation de rang p c'est tout".

Cette formule d'approximation vieille de trente ans, n'était autre que la *formule de reconstitution* des données à partir des facteurs démontrée par B. Cordier pour l'analyse des correspondances ; à une différence importante près toutefois : la présence ici des coefficients de la métrique distributionnelle (ou métrique du χ^2). Considérée dès l'abord sous deux aspects différents d'une part comme l'*analyse factorielle* d'un ensemble de variables, et d'autre part comme l'ajustement d'une *échelle multidimensionnelle* à un nuage *(multidimensional scaling)*, l'analyse des correspondances nous apparaissait maintenant d'un troisième point de vue, comme la recherche de la meilleure approximation de rang fixé d'un tenseur (tableau rectangulaire) donné, au sens des moindres carrés pour la métrique du χ^2 (approximation n'a de sens que si l'on a arrêté ce qu'on entend par proche ; et présuppose donc le choix d'une distance). Insistons sur ce que le rang fixé (nombre des facteurs extraits) n'a aucune incidence sur les résultats en ce sens que par exemple la meilleure approximation de rang 4 s'obtient avec 4 facteurs dont les 3 premiers ne sont autres que ceux qui ont fourni la meilleure approximation de rang 3 : c'est là un des mérites qui impose de choisir les formules quadratiques de distance et la géométrie euclidienne (cf § 3.2.3) ; (Mais pour la réduction des tableaux à plus de deux entrées pareille unicité n'existe pas ; et le programme de Caroll et Chang cité plus haut est en butte à cette difficulté quand il cherche une approximation :

$$k(i,j,t) \approx \Sigma_\alpha \ \varphi_\alpha(i) \ \psi_\alpha(j) \ \theta_\alpha(t)$$

les facteurs $\varphi_\alpha, \psi_\alpha, \theta_\alpha$ dépendent du rang p - i.e. du nombre de facteurs demandés ; et la décomposition cherchée n'est même pas unique pour p fixé...).

3.4.3 <u>*L. Guttman*</u> : Dans la bibliothèque du laboratoire de R. N. Shepard et D. Caroll, nous trouvâmes bientôt à l'analyse des correspondances une quatrième interprétation chez un prédécesseur beaucoup plus proche qu'Eckart et Young : L. Guttman. A un tableau rectangulaire I x J de nombres positifs (e.g. un tableau de contingence) ce maître de l'analyse des données avait dès 1941(**) proposé d'associer des facteurs définis par la condition d'être des couples de fonctions F(i), G(j) définies sur les deux ensembles I et J et le plus corrélées entre elles sur I x J en un certain sens (en bref, les fonctions F(i) et G(j) d'une seule variable sont des cas particuliers de fonctions de deux variables H(i,j) ; en donnant à (i,j) le poids k(i,j) on peut calculer sur I x J un coefficient de corrélation entre F et G). De ce point de vue qui est, on l'a dit, (cf § 2.4.6 et P. Cazes, thèse 3° cycle, Paris 1970) celui de l'analyse canonique de Hotelling, L. Guttman avait défini les facteurs mêmes calculés par l'analyse des correspondances. Il ne les

(*) *The approximation of one matrix by another of lower rank ; Psychometrika, 1936 ; T 1 ; pp 211-218.*

(**) *Louis Guttman : The quantification of a class of attributes, in P. Horst et coll, The Prediction of Personal Adjustement, Social Science Research (council N.-Y. ; 1941).*

avait toutefois pas calculés ; pour la seule raison qu'en 1941 les moyens de calcul requis (ordinateurs) n'existaient pas (cf § 2.5.3). Mais le modèle bien connu des échelles de Guttman (en bref analyse d'un tableau I x J par permutation de ses lignes et colonnes jusqu'à faire apparaître une bande centrale de forme parallélogrammatique aussi parfaite que possible et bordée de zéros), avec les composantes principales qui y sont associées rentrait dans le cadre général d'abord conçu par cet auteur et retrouvé par nous. Par le fait était posé un cinquième problème d'interprétation ; celui des rapports de l'analyse factorielle des correspondances avec des modèles de structure (cf 3.4.5).

Auparavant, notons que, bien que nous ne sachions pas que Guttman lui-même soit jamais retourné aux idées proposées par lui en 1941 (il a en revanche travaillé à perfectionner la méthode d'analyse des proximités de R.N. Shepard) son projet n'a pas été sans suite : nous avons appris de J. Faverge (cf. *Cours* Bruxelles 1970-71) que dès 1952 un auteur japonais C. Hayashi(*) avait proposé de calculer les facteurs définis comme couple de fonctions ayant corrélation extrémale sur deux ensembles en correspondance ; et que cette méthode avait été dans la suite appliquée au Japon à des enquêtes d'opinion. La priorité de ces auteurs est donc certaine : la seule originalité que puissent revendiquer les chercheurs français est d'avoir conjugué avec une méthode de calcul découverte indépendamment par plusieurs auteurs, des idées et des problèmes multiples dont la synthèse n'était pas faite ; et d'avoir élaboré une philosophie statistique nouvelle. Quant à remonter dans le temps avant Guttman (1941) comme Hill (1974, cf § 3.1) y invite, nous serons plus réservés ; il est vrai que l'école anglaise (Fisher en 1940 et avant lui Hirsfeld en 1935) a proposé la première (sous réserve de découvertes bibliographiques encore possibles) de calculer les valeurs propres et aussi les facteurs qui sont ceux de l'analyse des correspondances. Mais chez ces auteurs (qui n'ont traité que des tableaux de données de très petite taille) le problème n'est pas l'analyse des données telle qu'elle est pratiquée par Guttman : c'est la mesure de corrélation entre deux variables qualitatives ayant respectivement I et J pour ensembles de modalités, à partir du tableau rectangulaire I x J donnant les probabilités p_{ij} qu'à la modalité i de la première variable soit associée la modalité j de la deuxième variable (par exemple les deux variables sont la couleur des yeux et la couleur des cheveux ; l'ensemble I = {foncés, moyens, clairs, bleus} ; l'ensemble J = {noirs, foncés, moyens, roux, blonds} ; et p_{ij} est la probabilité qu'au sein d'un certain groupe un sujet ait à la fois i pour couleur d'yeux et j pour couleur de cheveux). On sait (§ 2.2.7) que l'épreuve classique du χ^2 permet de confronter à un échantillon l'hypothèse d'indépendance de i et j : $p_{ij} = p_i \times p_j$ (e.g. dans notre exemple du § 3.2.2. les associations des verbes et noms se feraient au hasard, sans affinité particulière entre ceux-ci) ; or la quantité critère (mesure de l'écart entre le tableau des p_{ij} et celui des $p_i\ p_j$) est justement la somme, $\Sigma\lambda_\alpha$, des valeurs propres extraites de l'analyse des correspondances ou encore de l'inertie (dispersion) du nuage $N(I)$ égale à celle de $N(J)$. Les facteurs eux-mêmes sont pour les auteurs de l'Ecole anglaise des mesures numériques permettant de calculer un coefficient de corrélation entre les qualités exprimées par i et j : nous reviendrons sur leurs travaux au § 3.5.2.

3.4.4 Les tests : Nous avions dès le départ considéré ce critère classique du χ^2 (cf § 2.2.6) afin de décider à quel rang arrêter l'interprétation des facteurs. En bref pour un échantillon d'effectif donné (dans l'analyse d'un tableau de contingence cet effectif est le nombre

(*) *C. Hayashi ; in Ann. of the Inst. of Stat. Math. T. 3 n° 2 Tokyo 1952.*

total $\Sigma k(i,j)$ des paires recensées ; sous la réserve essentielle que ces paires puissent être considérées comme des manifestations aléatoires indépendantes, de l'affinité entre éléments i et j des deux ensembles ; condition qui n'est jamais réalisée qu'approximativement dans la pratique ; et ne l'est aucunement pour un tableau de correspondance "ensembliste", i.e. empli de 1 et de 0, cf § 3.2) on sait l'ordre de grandeur de la somme $\Sigma\lambda_\alpha$ (appelée trace) sous l'hypothèse que les affinités apparentes entre i et j (différences entre les fréquences des paires f_{ij} et les produits $f_i \times f_j$ des fréquences marginales) soient seulement dues aux fluctuations d'échantillonnage ; les dernières valeurs propres, dont la somme est comprise dans cet ordre de grandeur, correspondent à des facteurs de bruit. On notera ici l'optique propre à l'analyse des données (cf 3.8.5) : tandis que la statistique pratiquée vers 1950 multiplie les "tests" (épreuves) pour protéger l'acceptation d'une hypothèse (ou d'un modèle) posée *a priori* (e.g. la normalité d'une distribution), l'analyse des correspondances se réfère à l'hypothèse d'indépendance entre i et j, mais n'a d'objet que pour autant que celle-ci n'est pas vérifiée, qu'il y a entre éléments des deux ensembles des affinités inégales dont on représentera spatialement la structure.

3.4.5 *Les modèles* : Cette rencontre avec la statistique des tests nous ramène aux modèles dont l'échelle de Guttman est un exemple. L'analyse des correspondances peut par le calcul des facteurs et la représentation graphique associée, découvrir sans l'avoir postulé *a priori* qu'un tableau de données est conforme (en général cette conformité ne sera qu'approximative) au modèle d'échelle. Elle peut aussi révéler d'autres modèles : e.g. une partition des ensembles I et J en classes $\{I_1, I_2, I_3 \ldots ; J_1, J_2, J_3 \ldots\}$; les éléments d'une classe I_2 s'associant exclusivement (ou quasi-exclusivement) avec ceux de la classe J_2 de même rang ; ou encore une variable normale sous-jacente aux associations (cf infra § 3.5.2). De tels résultats démontrés à partir de 1965 par B. Cordier, Ch. Rousse-Lacordaire (Mme Bourgarit) etc., quand notre expérience s'est étendue aux données les plus diverses, permettent d'atteindre des modèles typiques non *a priori* mais *a posteriori*, au terme d'un traitement commun à tous les tableaux sans hypothèse restrictive ; et de conjuguer inductivement ces modèles. Ainsi se réalise le projet initial suscité par la thèse chomskienne : donner à l'induction une méthode formalisée ; projet à la vérité bien ancien, car il est dans Bacon ($\approx$ 1600) (*).

3.5 *Le calcul des transitions* :

Dès l'exposé de la formule de probabilité des causes (au § 1.4.2) sont apparues les probabilités conditionnelles p_j^i ou p_i^j ; cette même notion se retrouve dans la formule de transition de l'analyse des correspondances (§ 3.3.2) qui est à la base de la définition des facteurs proposés par Guttman (§ 3.4.3) et aussi d'interprétations voisines connues antérieurement de l'Ecole anglaise (*infra* § 3.2.5). Pour nous, le calcul des transitions probabilistes est une variante du calcul tensoriel, adaptée aux espaces probabilisables : là est le principe des notations utilisées en analyse des données principalement pour les ensembles finis qui sont l'objet propre de la statistique (cf § 1.7.6). Dans ce § nous ferons l'histoire de ces notations ; puis nous exposerons les travaux de l'Ecole anglaise sur lesquels un récent article de M.O. Hill (cf § 3.1) a appelé notre attention.

(*) *En hommage au Novum Organum de Bacon, il nous semble permis d'appeler l'analyse des données Novius Organum ; et nous avons écrit sous ce titre un exposé des méthodes inductives de la statistique publié dans le volume Organum de l'Encyclopaedia Universalis.*

3.5.1 *Le calcul tensoriel des mesures et des fonctions* : Dans les premiers exposés de l'analyse des correspondances (cours de 1963 et thèse de B. Cordier en 1965) les notations à indice de calcul tensoriel ne sont pas utilisées. Les probabilités de la paire (i,j), de l'élément i, de l'élément j sont notées p(i,j), p(i), p(j) et non p_{ij} , p_i , p_j , comme aujourd'hui ; de même la probabilité conditionnelle de j quand i s'écrit p(j/i) = p(i,j)/p(i) et non p_j^i. Le nuage $N(I)$ n'est pas décrit en associant à chaque point i la loi conditionnelle $p_J^i = \{p_j^i \mid j \in J\}$ qui est une mesure sur J ; mais on considère le système $\{p(i,j)/(p(i) \times p(j)) \mid j \in J\}$, c'est à dire une fonction sur J qui n'est autre que la densité de la loi conditionnelle p_J^i par rapport à la loi marginale p_J (car p(i,j)/(p(i) x p(j)) = p(j/i)/p(j)) . Le changement intervenu dans les notations au cours de l'année 1965 a de multiples causes.

1°) Le modèle du calcul tensoriel sous la forme adoptée par A. Lichnérowicz pour exposer les théories d'Einstein incitait à faire apparaître dans les notations mêmes la distinction entre un espace vectoriel et son espace dual ainsi que les divers isomorphismes entre produits tensoriels et espaces d'applications linéaires clairement enseignés dans l'algèbre de N. Bourbaki. Comme le notait Laplace, une langue bien faite va d'elle-même au vrai ; car, dirons-nous, les défauts du raisonnement y apparaissent comme des fautes de syntaxe. Le calcul matriciel communément utilisé par les statisticiens, note sans autre distinction, par un tableau carré n x n une application linéaire d'un espace E de dimension n dans un autre espace F de dimension n, ou élément du produit tensoriel $E \otimes F$; donc aussi une application linéaire de E dans lui-même, et un tenseur de $E \otimes E$ (par exemple la forme quadratique d'inertie d'un nuage de points de E). Avec de telles notations le calcul du produit a x b de deux matrices apparaît possible, pourvu que les lignes de a aient même longueur que les colonnes de b. On sait cependant que le produit de composition f ∘ g de deux applications linéaires f et g n'a de sens que si l'espace-but de g coïncide avec l'espace source de f. Avec les notations tensorielles adaptées d'Einstein, la composition ou contraction se fait en sommant par rapport à un indice apparaissant deux fois l'une en position haute, l'autre en position basse, e.g. : $h_i^k = \Sigma\{f_i^j \; g_j^k \mid j \in J\}$; en sorte que la règle compatibilité de f avec g apparaît comme une sorte de règle d'accord grammatical. En calcul des probabilités si d'abord on se borne à des ensembles finis I, J, K d'éventualités, on considérera l'espace vectoriel R^I (ou R^J, R^K) des fonctions sur I (ou J, K) qu'il faut distinguer de l'espace des mesures qu'on notera R_I (ou R_J , R_K). Une transition de I vers J est un élément du produit tensoriel $R_J \otimes R^I$ (c'est une fonction sur I, mais à valeur dans les mesures sur J) : on écrira $\tau_J^I = \{\tau_j^i\} \in R_J \otimes R^I$; $\tau_j^i = \tau(j/i)$ étant la probabilité conditionnelle de j quand i. Une telle transition sert à la fois à transporter les systèmes de masses (mesures ; lois de probabilités) de I vers J ; et à associer à toute fonction sur J une fonction sur I etc. Sans recourir à toutes les ressources des notations tensorielles, le calcul est grandement soutenu si l'on respecte la distinction entre indice haut et indice bas, expression typographique de celle entre espace et dual, entre fonction et mesure (cf TII B n° 1).

2°) Dans les modèles probabilistes, les transitions apparaissent comme une généralisation de la notion d'application ensembliste. Soit φ une telle application de I dans J : on lui associera la transition φ_J^I qui à tout point i donne pour image le profil φ_J^i concentré au point

$\varphi(i)$ (i.e. : φ_j^i vaut 1 si $j = \varphi(i)$ et zéro sinon). Avec une transition τ_J^I, le point i aura une image dans J non ponctuelle, mais étalée suivant la loi τ_J^i. L'importance de ce point de vue est grande en statistique classique : mais ici on rencontre des transitions non seulement entre ensembles finis mais entre espaces, (ce qui sera précisé en 3°). Rappelons par exemple comment l'estimation d'une grandeur $\theta \in \Theta$ (généralement une grandeur multidimensionnelle : e.g. le système des paramètres d'une loi ; cf § 2.3.3) se fait à partir d'un système de données aléatoires $y \in Y$ (e.g. un échantillon fini issu de cette loi). Du véritable θ (inconnu) on passe à y par une transition p_Y^Θ qui a θ associe la loi de y, p_Y^θ (image diffuse de θ dans Y) ; l'estimateur est une fonction certaine ϵ_Θ^Y de Y vers Θ (on a vu que ϵ est un type particulier de transition, nous dirons une transition déterministe, sans aléa) ; finalement entre le θ vrai et le θ' estimé le passage se fait par une transition composée $\tau_\Theta^\Theta = \epsilon_\Theta^Y \circ p_Y^\Theta$: et la théorie de l'estimation, n'est que le choix d'un estimateur ϵ tel que pour tout θ, la loi τ_Θ^Θ(du θ' estimé) soit aussi concentrée que possible autour de θ. Autre exemple : quand on tente d'appliquer la formule des probabilités des causes (cf § 1.4.2) on a une transition θ_Y^X de X (espace des causes) vers Y (espace des effets) qui décrit bien la loi de y à partir de x ; mais on doit y adjoindre la loi de x (loi dite *a priori*) pour construire une transition Y vers X, χ_X^Y qui à tout associe la loi χ_X^Y de sa cause ; etc.

3°) Les formules de la théorie des processus que nous faisaient connaître les exposés de M. Métivier ou le traité (alors tout récent) de J. Neveu, et que Ph. Courrège et H. Rouanet(*) entendaient incorporer à des modèles psychologiques, sont à la fois simples dans leur structure et obscurs dans leur écriture intégrale : aussi les spécialistes les devinent-ils parfois plutôt qu'ils ne les lisent ; témoin ce commentaire de J. Neveu à l'énoncé du théorème de Ionescu-Tulcea : "La formule de définition de P_{x_0} (dont le 2° membre doit être lu à rebours) est intuitive malgré sa complication apparente". Ces formules s'écrivent bien si l'on remarque que les espaces probabilistes, avec pour morphismes les probabilités de transition, forment une catégorie. Ainsi est précisée en termes mathématiques notre intuition que (cf 2°) les transitions sont comme des fonctions mais entachées d'aléas (l'image d'un point x n'est pas un point $\varphi(x)$ mais une loi étalée τ_Y^x) ; la composition des morphismes généralisant celle des fonctions. Et les notations peuvent être allégées : tandis que communément un espace probabilisable est désigné d'une autre lettre que la tribu de ses parties mesurables, on pourra se borner à une seule lettre, comme on note les espaces topologiques sans en spécifier les ouverts ; la tribu des mesurables étant au besoin désignée par la lettre soulignée (ou grasse). Pour une transition τ_Y^X on a la notation en composantes : $\tau_Y^X = \{\tau_{\underline{Y}}^x \mid x \in X \ ; \ \underline{Y} \in \underline{Y}\}$; $\tau_{\underline{Y}}^x$ désigne la masse de la partie mesurable $\underline{Y}$ de Y pour la loi τ_Y^x image du point x de X ; considérée comme fonction de deux arguments, le point x de X et la partie mesurable $\underline{Y} \in \underline{Y}$, τ doit

(*) *Nos notations de calcul tensoriel des transitions apparaissent pour la première fois dans : Analyse statistique et modèle probabilistes en psychologie, in Revue de l'Inst. Intern. de Stat. V. 34, pp 139-155 (1966). Ce travail suggère d'utiliser l'analyse factorielle des correspondances pour extraire des données statistiques expérimentales des structures modèles proposées par H. Rouanet comme une généralisation de nombreux modèles classiques, en psychologie.*

être, on le sait, fonction mesurable en x, et mesure en y (au § 1.4.2, avec le problème de Bayes sur la probabilité inverse, on a un exemple de transition vers I = (0,1)). Ce n'est pas le lieu d'exposer le calcul tensoriel des transitions entre espaces probabilisables ; mais voici ce que devient la formule d'Ionescu-Tulcea objet du commentaire de Neveu. Il s'agit, en bref, de définir un processus aléatoire (suite d'événements e_1 , e_2 , ..., e_t ,...) par la donnée des transitions $\tau_{E_t}^{E_0 \ldots E_{t-1}}$ (ou lois conditionnelles de e_t , quand e_0 , e_1 ,..., e_{t-1} sont connus) : on aboutit à une transition $\rho_E^{E_0}$ (où $E = E_0 \times E_1 \ldots$ Produit infini des E_t) qui donne à partir de e_0 la loi de toute la suite des e_t ; ρE^{E_0} est définie comme limite projective par composition d'une suite infinie de transitions :

$$\rho_E^{E_0} = \ldots \circ (\delta_{E_0 \ldots E_{t-1}}^{E_0 \ldots E_{t-1}} \times \tau_{E_t}^{E_0 \ldots E_{t-1}}) \circ \ldots \circ (\delta_{E_0 E_1}^{E_0 E_1} \times \tau_{E_2}^{E_0 E_1}) \circ (\delta_{E_0}^{E_0} \times \tau_{E_1}^{E_0}) ;$$

les intégrales ont disparu (remplacées par le signe de composition des transitions) ; et la suite des transitions qu'il faut composer ne doit plus "être lue à rebours".

Avouons- le, ce formalisme dans toute sa généralité n'est pas indispensable à la pratique de la statistique (n'avons-nous pas affirmé que la théorie des probabilités elle-même est pour l'analyse des données une source d'inspiration plutôt qu'une méthode ; cf § 1.7.6) ; mais en est résulté un système de notations qui marque explicitement toutes les distinctions conceptuelles importantes et attribue à la notion de transition probabiliste le rôle central qui lui revient. Rôle dont témoignent les travaux britanniques que nous avons pour cette raison placés ci-dessous en § 3.5.2.

3.5.2 *Quelques travaux de l'Ecole anglaise sur l'analyse des matrices de contingence* :

Pour référence princeps à l'analyse des correspondances, M.O. Hill (1974) donne H.O. Hirschfeld (1935 : A connection between correlation and contingency ; in *Proc. Camb. Phil. Soc*, 31, pp 520-524) ; puis R.A. Fisher (1940 : The precision of discriminant functions ; in *Ann. Eugen. Lond.*, 10, pp 422-429) avec une application par K. Maung (1941 : Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colours of Scottish school children ; in *Ann. Eugen. Lond.*,11, pp 189-223). Après l'article de Hill, nous avons lu ces références dont voici le contenu exposé, pour plus de brièveté avec les notations de nos cours.

Hirschfeld pour étudier la corrélation entre deux variables qualitatives part (comme il est classique depuis K. Pearson ; cf § 2.2.7) de la matrice de contingence des $\{p_{ij}\}$, probabilités qu'à la modalité i de la première variable soit associée la modalité j de la deuxième variable. Or un calcul de corrélation requiert classiquement qu'aux modalités i et j soient associées des valeurs numériques φ^i , φ^j : d'où la question : "introduire ces variables de telle sorte que les *deux* régressions entre elles soient linéaires" ; par quoi Hirschfeld demande que :

$$\Sigma_i \ \varphi^i \ p_i^j = \rho \ \varphi^j \quad ; \quad \Sigma_j \ \varphi^j \ p_j^i = \rho' \ \varphi^i$$

(où il apparaît que si les φ^i et les φ^j ont variance 1, on a $\rho = \rho' = \lambda^{1/2}$). Ainsi se trouve posée l'équation des facteurs normalisés définie par les formules de transition $\varphi^I \ p_I^J = \lambda^{1/2} \ \varphi^J$; $\varphi^J \ p_J^I = \lambda^{1/2} \ \varphi^I$. Hirschfeld trouve d'abord le facteur trivial constant et égal à 1 qui ne répond

pas à son propos (*). Il s'intéresse donc au facteur de moyenne nulle relatif à la plus forte valeur propre possible. Il est clair pour lui que λ (qu'il interprète comme le carré d'un coefficient de corrélation) ne peut dépasser 1 : il affirme qu'une valeur propre 1 (pour un facteur de moyenne nulle) correspond à "une dépendance parfaite dans la distribution" des p_{ij} ; mais il ne semble pas avoir vu que cette dépendance correspond précisément à une matrice de partition de la matrice des p_{ij} en deux blocs diagonaux suivant le modèle rappelé au § 3.4.5. Il sait que la trace (ou somme des valeurs propres $\lambda = \rho^2$ relatives aux facteurs non-triviaux) n'est autre que le χ^2 calculé pour éprouver l'indépendance entre les deux variables i et j (cf §§ 2.2.7 & 3.4.4). Il aperçoit la possibilité de généraliser son problème à l'étude des correspondances entre variables continues (cas où I et J sont des espaces) : mais recule devant la complexité des calculs. Cependant au terme d'une note si riche en résultats ou en suggestion, Hirschfeld n'envisage pas de faire la synthèse des liens entre les i et les j d'après une carte plane réunissant deux facteurs (il ne songe d'ailleurs pas aux facteurs de variance λ, plus appropriés à une telle carte que les φ de variance 1) : son propos reste la mesure de la corrélation.

Dans sa note de 1940, Fisher part lui aussi de la loi de probabilité conjointe des modalités de deux variables qualitatives : il prend pour exemple I, ensemble de quatre couleurs d'yeux ; J, ensemble de cinq couleurs de cheveux (p_{ij} est e.g. la probabilité d'avoir les yeux bleus, i, et d'être roux, j). Son propos est la discrimination : il cherche, disons, une fonction φ^I définie sur l'ensemble des couleurs d'yeux, telle que si tout individu reçoit pour abscisse la valeur φ^i correspondant à la couleur de ses yeux, alors les cinq classes de sujets ayant une couleur j de cheveux déterminée (les blonds, les roux, etc.) soient aussi bien regroupés que possible ; le critère précis étant : maximisation de la variance interclasse (variance du nuage des cinq centres des classes définies par les couleurs j) relativement à la variance intraclasse (variances ajoutées des cinq classes, chacune rapportée à son centre) ; (cf § 2.3.5)(**). Il aboutit à l'équation des facteurs par transition, équation qu'il propose avec sa solution itérative : partir d'une fonction quelconque sur I, soit θ_1^I ; passer sur J à $\theta_1^I \circ p_I^J$; revenir sur I par $\theta_1^I \circ p_I^J \circ p_J^I$ etc... (en d'autres termes d'une fonction sur les couleurs d'yeux i, on passe à une fonction sur les couleurs de cheveux j, en faisant pour chaque classe j la moyenne des notes données à la couleur d'yeux de ses sujets ; etc...); et par va-et-vient à condition de normaliser les fonctions on parvient à la stabilisation. Fisher se demande incidemment si les valeurs de la fonction ainsi calculée (le premier facteur non trivial) diffèrent significativement pour les yeux clairs et les yeux bleus. Il suffit de lire l'analyse que Fisher lui-même donne de son travail dans le recueil *Contributions to math. stat,* pour voir qu'il ne pensait nullement avoir fait là une analyse factorielle.

Maung reprend en détail l'exposé des idées de Fisher et précise l'étude des données relatives aux couleurs d'yeux et de cheveux. Avec

(*) *Ce facteur trivial est évidemment éliminé des résultats de l'analyse des correspondances ; il apparaît quand on définit les facteurs par un calcul de transition ; mais non quand on recherche explicitement les axes principaux d'inertie.*

(**) *Ce critère où l'on peut voir une $(n+1)^e$ caractérisation des facteurs issus d'un tableau de correspondance (valable seulement pour une interprétation particulière du tableau des données) se trouve en TII B n° 7 § 1.5.*

l'interprétation donnée explicitement par Fisher - maximisation du rapport de la variance interclasse à la variance intraclasse - il en donne deux autres qui aboutissent à la même équation des facteurs : recherche de fonctions qui se reproduisent mutuellement par transition (c'est le problème résolu par Hirschfeld ; et vu déjà par Fisher, sinon exposé nettement dans sa solution itérative ; Maung fait référence aux corrélations canoniques de Hotteling, cf § 2.4.6) ; et calcul de la corrélation entre φ^I et φ^J considérées toutes deux comme des fonctions sur un même ensemble support, celui des sujets dont on a noté les caractères i et j, couleurs d'yeux et de cheveux (il revient au même de considérer les corrélations sur I x J ; cf supra § 3.4.3). Comme Hirchfeld, Maung voit que la trace n'est autre qu'un χ^2. Il entreprend, de plus, d'avoir une épreuve de signification pour la première valeur propre (non triviale) d'après l'effectif de l'échantillon étudié ; en fait ce problème que L. Lebart a étudié par une simulation (cf TII B n° 8 § 3) n'est pas au clair chez Maung. La formule de reconstitution des données en fonction des facteurs est attribuée sans démonstration à Fisher (à qui une formule d'analyse quadratique aurait difficilement échappé!) "Prof. Fisher *has pointed out that...*" Un § de Maung est consacré à ce que nous appelons correspondances normales (cf § 3.4.5) c'est à dire à l'analyse des tableaux de contingence définis comme suit. Soient x et y deux variables aléatoires dont la loi conjointe est normale et entre lesquelles le coefficient de corrélation est r ; soient I et J deux partitions de la droite réelle en une suite d'intervalles assez resserrée relativement à la dispersion de x et y ; on note p_{ij} la probabilité que x soit dans l'intervalle i et y dans l'intervalle j. On sait depuis Pearson (1904) que la trace (ou le χ^2) du tableau p_{IJ} (tableau qui, on l'a vu, joua un rôle historique dans les recherches biométriques de F. Galton puis de K. Pearson ; cf §§ 2.2.2. & 2.2.7) tend pour des partitions infiniment fines vers $r^2/(1-r^2)$; Maung souligne que la corrélation entre le premier couple de facteurs φ_1^I, φ_1^J est r (i.e. avec nos notations, que $\lambda_1 = r^2$; cf TII B n° 7 § 4.1). Quant aux données concrètes Maung conclut à une association positive entre les pigmentations des yeux et des cheveux ; il trouve les filles plus claires de cheveux (mais non d'yeux) que les garçons : ce qui a été attribué à ce que celles-là coupent moins leurs cheveux que ceux-ci... Mais pas plus que Hirschfeld et Fisher, Maung n'envisage un déploiement multidimensionnel des résultats (il sait que le calcul fournit une suite de facteurs ; mais ne regarde que le premier, en tant que solution optimale aux trois problèmes qu'il a posés) ; ni une généralisation à des données d'autre format. Et, somme toute, s'il fallait à l'analyse des correspondance un patronage britannique, c'est au grand K. Pearson qu'il nous plairait de le demander (cf § 2.2.7).

3.6 *Extension du domaine de l'analyse des correspondances* :

Les analyses effectuées à Rennes en deux années de 1963 à 1965 portaient exclusivement sur des tableaux de contingence d'assez petite taille (de 8 x 8 à 30 x 30 environ) issus d'expériences de psychologie, ou plus rarement de relevés linguistiques. Le laboratoire de statistique fondé à Paris en 1965 sous l'égide du Doyen M. Zamansky et du Pr. D. Dugué, Directeur de l'I.S.U.P., allait grandement élargir ce programme.

Dans ce § nous décrivons superficiellement ce progrès ; en jalonnant quelques étapes et rappelant les épisodes les plus marquants (cf § 3.6.3) au travers desquels le laboratoire a bénéficié des travaux de chercheurs de toute discipline. S'il est vrai, comme nous aimons à l'affirmer, que la statistique est une science expérimentale, il fallait évoquer cette fructueuse collaboration avant les développements des méthodes (§§ 3.7 & 3.8) et des programmes (§ 3.9).

3.6.1 Ecologie et biosystématique : Dès 1966, M. Roux s'appliqua aux données écologiques : I, ensemble de parcelles (de terrains) x J, ensemble d'espèces végétales ; k (i,j) = 1 si j est présent dans i, zéro sinon (ou encore, k(i,j) est un coefficient d'abondance). Ce fut le début d'une collaboration ininterrompue depuis, avec le laboratoire du Pr. M. Guinochet (Orsay). Dans sa Phytosociologie (Masson ; Paris ; 1973) M. Guinochet a bien voulu faire une place de choix aux méthodes d'analyse statistique multidimensionnelle. Pour le statisticien, les données phytosociologiques et biosystématiques ont joué un rôle essentiel dans le progrès des méthodes : confrontation de l'analyse factorielle avec la classification automatique (cf § 3.8.3) et l'analyse discriminante (§§ 2.3.5 & 3.8.2), traitement de tableaux hétérogènes (mêlant variables quantitatives - dimensions, températures...; qualitatives - couleur, rugosité...; logiques - présence ou absence etc...) (cf § 3.7.1) ont grandement progressé à l'épreuve de ces données. A partir de 1969, le traitement du monumental corpus d'observations rassemblé par L. Bellier en Côte d'Ivoire a introduit l'analyse des correspondances dans le domaine de la biosystématique et de l'écologie animales. Le sommaire à la partie C du tome I de l'*Analyse des données* suffit à découvrir la diversité de ces sortes de recherches.

3.6.2 Le colloque d'Honolulu : A l'invitation du Pr. M. S. Watanabe fut écrit au début de 1968 le rapport de J. P. Benzécri "*Statistical analysis as a tool to make patterns emerge from data*". Ce rapport présenté par Ch. Masson au colloque de Honolulu sur la reconnaissance des formes offre un panorama de ce qu'était alors chez nous la pratique de l'analyse des données (*). Quant aux méthodes, à côté de l'analyse des correspondances et de la classification automatique (partitions seulement alors ; et non hiérarchie de classes) on trouve l'analyse des proximités (traitée par un algorithme original très simple) et l'analyse des préférences : aujourd'hui nous traiterions plutôt les données de préférence (I ensemble de sujets ; J ensemble d'objets ; chaque sujet i range dans l'ordre de ses préférences les objets j de J) non par un programme spécifique, mais par le programme d'analyse des correspondances, en affectant à chaque objet deux colonnes, l'une contenant son rang et l'autre le complémentaire de son rang (cf. J. P. Fénelon, thèse; et ses travaux en collaboration avec madame Y. Bernard ; et *infra* §§ 3.7.1 & 3.7.4 : dédoublement). Quant au domaine, à la psychologie et à la linguistique s'adjoint la médecine (analyse de M. Kerbaol sur les données de l'Hôtel Dieu de Rennes ; Prs M. Bourel et P. Lenoir, bientôt rejoints par le Pr. G. Sandor de l'Institut Pasteur de Paris). Le titre même du colloque "*Methodologies of Pattern Recognition*" pose le problème de la reconnaissance des formes : en 1977 comme en 1968 il ne fait pas de doute pour nous que l'analyse des correspondances est toute désignée pour réduire à un petit nombre de traits significatifs la description primaire des objets dont il s'agit de reconnaître la forme (cf § 2.5.6) ; les expériences de Ph. Marano (*Ann. des Télécom.* T 27 pp 163-172 ; 1972), P. Graillot (C.N.E.T. 1972), P. Chaumereuil et J. P. Villard (Stage D.E.A. 1970) confirment notre conception du problème, mais l'étude systématique des images mobiles et de la chaîne sonore n'a pas encore été faite avec toute l'ampleur convenable. Cependant, à fréquenter des chercheurs intéressés par la reconnaissance des formes (R. Guedj ; T. Dao) le laboratoire a gagné d'être initié à l'approximation stochastique : méthode suivant laquelle la solution d'un problème d'analyse est atteinte comme la limite d'un processus aléatoire convenable. Sur ce principe on a conçu un algorithme d'analyse de correspondance fort simple requérant très peu d'espace en mémoire centrale (cf § 3.9).

3.6.3 Institution des stages : Le printemps de 1968 aura vu crouler plus d'une colonne d'argile qu'on avait crue de bronze, et s'élever

(*) *Le texte [Honolulu] publié dans l'Analyse des données diffère du rapport de 1968 ; dont l'original en langue anglaise se trouve aux actes mêmes du colloque (cf Methodologies of Pattern Recognition, ed. Watanabe ; Acad. Press. N. Y., 1969).*

plus d'un château de cartes qu'on prit alors pour un Colisée. Pour l'analyse des données aussi, ce fut un passage historique. Empreints du lyrisme prudent qui s'impose aux heures chaudes, tels sont les documents d'époque. Voici un alinéa d'un rapport destiné à la D R M E, organisme alors prodigue en contrats et qui nous aidait libéralement. "Comme on l'imagine, le laboratoire en tant que tel a cessé d'exister pendant quatre mois... Il est heureux que les chercheurs dispersés aient pu continuer leurs travaux : mais certaines de nos recherches ont été entravées. Les étudiants, désireux de nous faire bien estimer leur génération, ont dans l'ensemble fait un effort considérable ; beaucoup de travaux de recherche entrepris cet été n'auraient sans doute jamais vu le jour sans la crise morale de Mai...". En effet, notre invitation aux recherches appliquées avait d'abord suscité peu d'échos ; témoin ce préambule à la circulaire qu'au début de l'été, après six semaines de silencieuse absence, le professeur adressait aux étudiants : "Comme nous l'avons fait en 1966 et 1967, nous invitons cette année les candidats au D.E.A. de Statistique, à faire de la session d'examen une fructueuse expérience de travail pratique et de recherche. Pareille méthode nous avons pu le constater demande de tous beaucoup plus qu'ils ne sont habitués à donner..." Mais dans les projets fiévreux de réforme des examens, chacun se trouva pris à ses propres paroles : les stages demandés par tous, s'imposèrent à tous : Dieu Merci, la vague était franchie sans naufrage. Cette autre circulaire sonne comme un appel à la levée en masse : "Je vous communique ci-joint le sujet d'une recherche statistique historique qui pourrait occuper une équipe d'une douzaine de chercheurs. Le travail s'accomplirait dans les conditions suivantes : 1° Constitution de l'équipe : les étudiants susceptibles de travailler dans la région parisienne pendant 2 à 3 semaines au cours de l'été (i. e. du 10 Juillet au 10 Septembre) et s'intéressant à la recherche historique se feront connaître en écrivant au secrétariat du laboratoire, etc..." Il y eut pour ce projet une équipe franco-iranienne de 4 volontaires (et non 12!) : ce fut le début de notre collaboration avec l'historien A. Prost (cf TII n° 2).

Depuis 1968 grâce au dévouement de nombreux chercheurs (au premier rang desquels il faut citer P. Cazes, J. P. Fénelon, M. Jambu, M.O. Lebeaux, M. Roux, S. Stépan, Y. Grelet...) les élèves du laboratoire ont produit des centaines de rapports de stage et des dizaines de thèses de 3° cycle. Par la collaboration avec de très nombreux laboratoires et autres institutions, notre expérience s'est étendue aux données les plus diverses : Géologie (P. Cazes avec F. Leroy d'ELF-ERAP puis P. Soléty du B R G M ; J. P. Bordet, J. M. Monget et P. Roux à l'Ecole des Mines) ; Géographie (Ph. Massonie et ses collègues de l'Université de Besançon ; le laboratoire de géomorphologie, dirigé par F. Verger à l'E.P.H.E.) ; Sociologie (J. P. Fénelon et Madame Y. Bernard en esthétique expérimentale ; M. de Virville et les élèves du Pr. Cuisenier ; D. Kalogéropoulos et ses confrères criminologistes ; M. O. Lebeaux et l'équipe d'Economie et Humanisme ; L. Lebart et N.Tabard au C R E D O C); Economie (M. Volle à l'I.N.S.E.E. ; J. L. Guigou en Faculté ; A. W. Hamrouni avec M. Lenco au ministère de l'Agriculture) ; Phénomènes physiques (haute atmosphère avec J. P. Bordet chez le Pr. Barliet à l'Observatoire de Meudon ; fiabilité des composants mécaniques, L. F. Pau et M. Bichara à Air-France ; ou électroniques : P. Graillot et G. Vasserot au C.N.E.T.) ; sans oublier la psychologie (M. O. Lebeaux avec M. Zlotowicz) et la linguistique (A. Salem avec le centre de lexicologie de Saint-Cloud et G. E. Weil à Nancy ; V. Huynh à l'Université de Vincennes) cultivées dès les débuts du laboratoire, ni la médecine.

Nous suivrons les progrès dans la diversité de ces travaux en feuilletant les publications du laboratoire (§ 3.6.4) et les programmes des colloques qu'il a organisés (§ 3.6.5) avant d'en faire le bilan méthodologique (§§ 3.7 & 3.8).

3.6.4 : Publications du laboratoire : Sans reproduire un catalogue, signalons brièvement les exposés successifs de l'analyse des correspondances, ainsi que les progrès des recueils d'exemples d'analyses. Le premier exposé oral de l'analyse des correspondances (cf §§ 3.1 & 3.2.4) fut donné en hiver 1963 au cours de six leçons du Cours Peccot professées par J. P. Benzécri sous le titre "Statistique et structure des langues naturelles ; essai de synthèse mathématique". Au début de 1964 une rédaction de ce cours fut publiée à Rennes en cinq parties intitulées *leçons* . Les deux premières de ces leçons traitent de la linguistique générale ; et la deuxième plus spécialement de la sémantique. La troisième leçon est une théorie algébrique des grammaires de constituants non-connexes (en bref penser au latin : *romanam condere gentem*, romaine fonder la nation ; où le constituant "la nation romaine", est interrompu par le verbe fonder dont il est l'objet). La quatrième leçon rappelle les principes de l'ajustement d'un systèmes d'axes à un nuage de points d'un espace euclidien ; et donne selon ces principes un algorithme simplifié d'analyse des proximités (cf §§ 3.4.1 & 3.6.2). La cinquième leçon est un premier exposé de l'analyse des correspondances, comportant la représentation simultanée des deux ensembles I et J, mais sans théorème (cf 3.3.1) ni notation tensorielle (cf § 3.5.1). Il était prévu une sixième leçon d'applications linguistiques de l'analyse des correspondances : le texte [Ana. Ling.] écrit en 1974 peut être regardé comme la réalisation différée de ce projet, mais il ne reçut pas le titre de sixième leçon, car en 1974 les diverses parties du cours de 1963-1964 s'étaient séparées après avoir connu d'inégales fortunes ; la cinquième leçon ayant, quant à elle disparu depuis 1965, à la parution de la thèse de B. Cordier (cf § 3.3.2)!

Le cours de 1967 d'analyse factorielle, écrit dans les notations du calcul tensoriel comporte deux parties : 1° la représentation approchée d'un nuage dans un espace de faible dimension (cf [Repr. Eucl.] TII B n° 2) ; 2° l'analyse des correspondances (cf[Dis χ^2 Corr.] TII B n° 5). Le texte Réduction d'un élément du produit tensoriel de deux espaces euclidiens, écrit en 1968 (cf [Red. Tens.] TII n° 6) reprend l'analyse des correspondances du point de vue d'Eckart et Young (cf § 2.4.2).

Cependant les exemples d'analyses factorielles ne formaient encore que de brèves notes indépendantes, ou des paragraphes insérés dans un recueil dont le thème central n'était pas l'analyse des données. En 1968, on reprit l'analyse de tous les tableaux de contingence issus d'expériences psychologiques que nous ayons pu rassembler : cet ensembles d'analyses, coordonnées et commentées forme l'article "Sur l'analyse des matrices de confusion", achevé en 1969, et publié en 1970 par la *Revue de Statistique Appliquée*. Ces données sont pour une méthode statistique, ce que sont les diatomées pour un objectif de microscope : la structure en étant bien connue (ce sont le plus souvent des ensembles de stimuli admettant dans le plan une représentation par un segment, un arc, ou un cercle, imposée par leurs propriétés physiques) il ne s'agit pas pour le statisticien de découvrir du neuf, mais de s'assurer de la fidélité de l'outil qu'il propose. Disons ici que cette épreuve nous semble indispensable : il est imprudent d'entreprendre de mettre au jour des dimensions cachées (e.g. de faire de la psychométrie) par une méthode qui ne distingue pas avec aisance les dimensions directement accessibles à nous (e.g. les variables d'une étude de psychophysique ; cf § 2.4.1).

Après les matrices de confusion, on entendait traiter aussi systématiquement et non seulement pour l'illustration d'un exposé (tel que celui envoyé au colloque de Honolulu ; cf § 3.6.2) toutes les données les plus diverses. A ce grand inventaire aida d'une part l'institution des stages (cf § 3.6.3), d'autre part le passage de M. O. Lebeaux au laboratoire de Cl. Picard (du C.N.R.S.) alors installé rue du Maroc. Et en 1970 nous pûmes sous le titre "l'analyse des données" constituer un recueil d'une douzaine d'analyses différentes touchant à la

psychologie, la sociologie, la linguistique. De ce recueil la moitié des chapitres dépassaient une vingtaine de pages : le codage, la critique et l'interprétation des résultats s'étant perfectionnés (cf § 3.7 & 3.8) au fur et à mesure que se diversifiaient les données.

Désormais, le progrès des exposés théoriques et de la systématisation des applications allait aboutir au recueil en 2 tomes publié chez Dunod en 1973 et réédité en 1976. Depuis, les publications d'exemples se poursuivent ; ainsi que celles de compléments méthodologiques et de programmes qui les mettent en oeuvre.

3.6.5 *Les colloques sur l'analyse des données* : Les salons des précieuses ont vécu. Les congrès internationaux, aux salles tantôt encombrées tantôt délaissées, n'attirent que par l'espérance des trop brefs *a parte* de couloirs. Les séminaires hebdomadaires sont une ligne de plus à l'agenda des chercheurs. Voici le temps des colloques : à défaut d'une salle capitulaire ou d'un rendez-vous de chasse tout abri dans une autre ville où les appels, les sonneries et les horaires sont pour les autres, suffit à recréer la société des savants. Le premier colloque du laboratoire se tint à Besançon les 14 et 15 Avril 1970 : il n'y en eut jamais plus de semblables. Ph. Massonie avait rassemblé dans un palais universitaire des collègues à l'esprit aussi cultivé que non prévenu ; et les exposés se succédaient à un rythme rapide (trois ou quatre par heure) presque tous élémentaires, dans un chatoiement de sujets divers : questionnaires, écologie, méthodes, linguistique, psychologie, taxinomie. Dans la suite il fallut descendre de ces délices encyclopédiques pour approfondir un domaine particulier déjà connu des auditeurs. Dès l'automne de 1970 ce fut le colloque de Marseille (chez le Pr. Sarles) couplé avec celui de Nice (U.E.R. sur le domaine méditerranéen), celui-ci consacré à l'écologie et à la systématique, celui-là aux données médicales. Puis à l'Arbresle (auprès d'un véritable couvent, sinon d'une antique abbaye) le colloque sur l'analyse des données appliquées aux sciences humaines, organisé en collaboration avec le laboratoire L. J. Lebret (E.R.A. 122 du C.N.R.S.). Puis Rennes (retour aux sources) ; et Besançon 2, Orléans, Rennes, Grenoble, Montpellier ... successivement géographie et économie ; écologie et botanique, etc... linguistique enfin. Recevant ainsi de multiples disciplines des problèmes de plus en plus complexes, la statistique peut progresser en offrant à tous une méthode unique.

3.7 *Perfectionnements apportés à la méthode* :

Quand à la fin de 1965 débutèrent les travaux du laboratoire de Paris nous avions les formules et théorèmes de l'analyse des correspondances (§ 3.3.2) écrits dans les notations du calcul des transitions (§ 3.5.1) ; et les principales interprétations des facteurs, géométriques ou probabilistes avaient été rassemblées (§ 3.4). Mais la pratique de l'analyse des correspondances telle que nous la connaissons aujourd'hui n'existait pas encore. Un programme permettait de calculer facteurs et valeurs propres ; les accès à ce programme, l'entrée comme la sortie, manquaient. A l'entrée en effet, se place le codage : c'est à dire la représentation d'un ensemble de données, d'un ensemble de faits par un tableau rectangulaire de nombres positifs apte à être soumis à l'analyse des correspondances ; en 1965 n'avaient guère été traités que des tableaux de contingence directement analysables (cf § 3.2.4). A la sortie, sont l'interprétation et la critique des résultats : en 1965 l'ordinateur fournissait seulement la liste des valeurs des facteurs ; les graphiques étaient tracés à la main et la seule aide statistique à l'interprétation était le critère du χ^2 (qui suggère quels sont les facteurs significatifs ; mais est applicable seulement à un tableau de contingence, sous l'hypothèse que les données résultent de tirages indépendants ; cf § 3.4.4). Dans la suite nous considérerons donc les progrès de l'analyse des correspondances dans le codage des données (§ 3.7) ; l'interprétations des facteurs (§ 3.8) ; l'organisation des

programmes (§ 3.9). Sans les multiples perfectionnements ainsi apportés à la méthode, l'extension du domaine de l'analyse des correspondances dont le § 3.6 offre le panorama, n'eût pas été possible : mais réciproquement, nous verrons les perfectionnements eux-mêmes souvent suscités par les exigences quotidiennes du traitement des données.

L'essence du codage des données, est de traduire fidèlement les relations observées entre des choses, par des relations entre des êtres mathématiques ; de telle sorte qu'en réduisant par le calcul (§ 3.9) la structure mathématique choisie pour image du réel, on ait de celui-ci un dessin simplifié accessible à l'intuition et à la réflexion avec la garantie d'une critique mathématique (§ 3.8). De ce point de vue l'analyse des correspondances, même appliquée aux tableaux de contingence pour le traitement desquels elle a été créée, comporte un codage qui est la représentation géométrique des ensembles I et J par des nuages euclidiens, avec identification des axes des deux nuages (cf §§ 2.2.3 & § 3.3). Dans le présent § ce codage géométrique est accepté pour tout tableau de nombres positifs : la question est donc d'une part de passer, si nécessaire, des données à un tel tableau ; d'autre part de justifier pour de multiples classes de données la fidélité au réel du codage géométrique appliqué à ce tableau.

3.7.1 Homogénéité et exhaustivité : Le tableau lui-même est d'autant plus fidèle au réel qu'il résulte du relevé exhaustif d'un champ homogène. Pratiquement, l'exhaustivité n'est souvent qu'approchée par échantillonnage ; et le détail du relevé d'un continuum est arrêté à une partition. L'étude des dépenses privées des Français se fera sur un échantillon de la population ; selon une nomenclature distinguant e.g. le tabac des conserves alimentaires, mais non une marque de cigarettes d'une autre. Ici le principe d'équivalence distributionnelle a le mérite d'assurer que le codage géométrique du réel est peu sensible au choix de la nomenclature (partition des dépenses). Mais que signifie exhaustivité pour un tableau de mensurations somatiques, de dosages biochimiques, un questionnaire d'opinion ? On doit au départ se fier au spécialiste, admettre que les données qu'il a recensées sont comme un échantillonnage du champ réel qu'il vise (échantillonnage dont la densité correspond à la redondance des données) ; puis par les résultats d'analyse, critiquer la composition des données et s'il se peut améliorer celles-ci. Quant à l'homogénéité, il semble facile de la respecter au moins approximativement; cependant certaines études requièrent la confrontation de deux ou plusieurs groupes de variables de nature différente : par exemple un végétal i sera décrit à la fois par un ensemble J_1 de mensurations (longueur et largeur de la feuille, longueur de tige entre deux noeuds ; etc..) et par un exemple J_2 de variables logiques ou qualitatives (couleur des pétales, pilosité des feuilles etc..) ; on a donc deux tableaux juxtaposés $I \times J_1$ et $I \times J_2$ (lignes du second, au bout des lignes du premier). Pour analyser ces données hétérogènes, on peut soit les réduire à l'homogénéité en les codant toutes sous forme logique (forme disjonctive complète : § 3.7.4) ; soit appliquer à chaque groupe le codage qui lui est propre, mais multiplier le deuxième tableau par un coefficient numérique de pondération afin que dans l'analyse du tableau global $I \times (J_1 + J_2)$, les contributions des deux groupes de colonnes J_1 et J_2 s'équilibrent (en un sens qui sera précisé au § 3.8.4, avec la définition des contributions). Le principe de pondération nous était connu depuis 1968, mais la méthode n'a été appliquée qu'en 1972 par A.W. Hamrouni, qui a donné dans sa thèse un programme de calcul des coefficients de pondération relative de deux tableaux (ou de plusieurs $I \times J_1$, $I \times J_2$,..., $I \times J_n$).

(Les deux textes [Pondération] et [Pond. Pr.] donnant le principe et le programme FORTRAN de cette méthode, doivent être publiés prochainement dans les cahiers).

Nous avons rencontré plusieurs exemples de données : nombres entiers dans les tableaux de fréquence, poids et valeurs dans les bilans ; mensurations ; dosages biochimiques ; qualités et variables logiques (en Oui ou Non). Efforçons-nous de ranger toutes les données sous quelques types dont nous considérerons successivement le codage. La théorie des grandeurs élaborée par les physiciens et les psychophysiciens nous servira de guide.

3.7.2 *Grandeurs additives extensives* : Il est classique en physique d'opposer les grandeurs extensives (masse, volume...) aux grandeurs intensives (température) d'après le critère suivant : si on sépare en deux une quantité de liquide homogène, les deux parties ont chacune même température que le tout ; mais elles s'en partagent la masse et le volume. Ici nous appellerons grandeurs additives extensives les grandeurs numériques positives pour lesquelles l'opération mathématique d'addition correspond à une manipulation réelle, à une réunion. D'abord les grandeurs entières recensées dans les tableaux de fréquence (k(i,j) = nombre de fois que le nom i a été trouvé sujet du verbe j, cf § 3.2.2). Ensuite les poids et valeurs des bilans : toutes les données d'un tel tableau peuvent être mesurées en une même unité (dont le choix importe peu : gramme ou once ; franc ou dollar) : pour de telles données, additionner deux colonnes revient à fusionner deux postes du bilan (e.g. les dépenses en riz et pâtes avec celles en légumes secs). Egalement les tableaux de mensurations prises sur un végétal, un sujet vivant, un crâne...: ici l'addition ne correspond à une opération réelle que s'il s'agit de mesurer deux segments qui se prolongent l'un l'autre : e.g. le premier et le deuxième entre-noeud sur une tige ; ou la longueur du bras et celle de l'avant-bras. Ne contenant que des nombres positifs, les tableaux de grandeurs additives extensives sont directement traitables par l'analyse des correspondances ; ce traitement a l'avantage d'être, de par le principe d'équivalence distributionnelle, insensible aux regroupements ou subdivisions éventuels de colonnes. Il n'est donc pas besoin ici de codage. De plus, en traitant des profils l'analyse de correspondance permet d'étudier la dispersion des formes indépendamment de celle des tailles ; tandis que l'analyse en composantes principales usuelle extrait pour premier facteur un facteur de taille, puis des facteurs non corrélés à celui-ci et appelés pour cette raison facteurs de formes (cf e.g. Kendall, *A course in multivariate analysis* ; Griffin ; Londres, (1957) ; p 151). Mais conformément à l'expérience du naturaliste, l'analyse de correspondance révèle un premier facteur de forme (facteur de forme parce que l'analyse ne traite que des profils) fortement corrélé à la taille (cf TI C n°s 5, 6 & 7).

3.7.3 *Variables logiques et qualités discrètes* : Tout à l'opposé des grandeurs additives extensives se trouvent d'autres tableaux parfaitement traitables tels quels par l'analyse des correspondances : ce sont les tableaux en (0,1), mis sous forme disjonctive complète. Voici le modèle commun à ces tableaux : soit I un ensemble d'individus ; Q un ensemble de questions ; J_q l'ensemble discret (i.e. discontinu, fini) des réponses possibles à la question $q \in Q$; $J = \cup \{J_q | \ q \in Q\}$, i.e. J est l'ensemble des modalités de réponse à toutes les questions, à chaque question q est affecté un bloc J_q de colonnes ; la ligne afférente à chaque sujet i comporte dans chaque bloc J_q un 1 dans la colonne correspondant à la modalité de réponse choisie pour i à la question q, et des 0 ailleurs. On dit qu'un tel tableau est mis sous forme disjonctive complète parce que à chaque question toutes les modalités sont explicitement prévues et distinguées. D'un même format sont les tableaux de description par des qualités discrètes : Q ensemble de qualités ; J_q ensemble des modalités de la qualité q : par exemple si q est la couleur des pétales, J_q sera l'ensemble {jaune, bleu, rouge}. Quand J_q ne comporte que deux modalités : {Oui, Non}, {présence, absence}, etc... il est commode de noter $J_q = \{q^+, q^-\}$: on parle alors de dédoublement (cf *infra* § 3.7.4). Le dédoublement attribue des rôles symétriques à

une qualité et à sa contraire ; ce qui est souvent indispensable, mais est parfois inopportun (on a rencontré au § 3.2.4 sous le nom de tableau de correspondance logique le cas des tableaux I x Q : k(i,q) = 1 si i possède la propriété q, zéro sinon ; sans dédoublement). Dans le tableau des votes d'une assemblée (I ensemble des députés ; Q ensemble des scrutins) on doit outre les *Oui* et les *Non* recenser les abstentions et les absences ; problème rencontré d'abord en 1968 lors de notre collaboration avec l'historien A. Prost (cf§ 3.6.3) et dont une étude très complète a été faite en 1973 à propos de l'analyse par A.W. Hamrouni des votes à l'O.N.U. (études publiées dans ces cahiers : Ca I n° 2 pp 161-195 & n° 3 pp 259-286): le codage peut alors s'écarter de la forme disjonctive complète (e.g. présence de (1/2) dans les colonnes q^+ et q^- etc.).

L'analyse des tableaux logiques dédoublés (dont les échelles de Guttman, cf § 3.4.3, sont un exemple très classique) et plus généralement des tableaux sous forme disjonctive complète (cf *infra* § 3.7.4) se pratiquait chez nous depuis plusieurs années quand en 1972 L. Lebart en apporta la meilleure justification : les facteurs sur J issus de l'analyse d'un tel tableau I x J ne sont autres (à un coefficient constant près) que ceux issus de l'analyse du véritable tableau de contingence J x J suivant(*) : k(j,j') = nombre des individus i ayant à la fois la modalité j et la modalité j'. Dès lors on rejoint le format original pour lequel a été conçue l'analyse des correspondances. De plus on a analysé des sous-tableaux rectangulaires du tableau J x J : tableaux $J_1 \times J_2$, où J_1 est l'ensemble des modalités des qualités q d'une partie Q_1 de Q ; et de même J_2 pour une autre partie Q_2 : ainsi on peut étudier la correspondance entre un ensemble d'opinions (réponses des sujets aux questions Q_1) et des caractères socioéconomiques (réponses à Q_2) ; et c'est par un tel tableau que l'analyse de correspondance résoud le problème de la régression (cf § 3.8.2) après avoir codé sous forme disjonctive complète (§ 3.7.4) variable à expliquer et variables explicatives.

3.7.4 Grandeurs intensives : Le succès maintenant bien compris des analyses de tableaux en 0,1 mis sous forme disjonctive complète invite à rapprocher de cette forme, par un codage approprié, les données les plus diverses.

Grandeurs intensives bipolaires : les résultats de nombreuses enquêtes sont comme ceux des examens scolaires et des épreuves psychotechniques, exprimés par des notes comprises entre deux bornes, qu'on peut après changement linéaire d'échelle supposer être 0 ou 1. Soit donc un tableau de notes I x Q : k(i,q) = note de l'individu i à l'épreuve q ; on créera pour chaque épreuve q un couple de colonne $\{q^+, q^-\}$ $(k(i,q^+) = k(i,q)$; $k(i,q^-) = 1 - k(,q)$; ou plus généralement : $k(i,q^-) = M_q - k(i,q^+)$; M_q étant la note maxima à l'épreuve q). L'analyse de tableaux ainsi dédoublés est pratiquée depuis 1968 par M. O. Lebeaux sur des données psychologiques (cf § 3.8.4) puis sur les enquêtes socioéconomiques de l'IRFED (cf TII C n^{os} 4, 5 & 6).

Grandeurs qualitatives ordinales : beaucoup de mesures numériques doivent être comprises non comme des quantités, mais comme des qualités susceptibles d'avoir une intensité plus ou moins grande repérée sur un axe ; nous dirons que ce sont des qualités ordinales. Faire l'analyse d'une roche en ses éléments (ou composés chimiques) majeurs ; doser l'argile, les carbonates, etc... dont la masse totale sera celle de toute la roche ; c'est faire un véritable bilan au sens considéré ci-dessus (§ 3.7.2) mais doser dans le sérum sanguin une suite d'enzymes,

(*) *Ce tableau avait déjà été considéré par C. Burt cf supra § 2.4.6.*

catalyseurs très actifs mais de masse infime, c'est plutôt situer des qualités sur une échelle ordinale où sont marqués quelques repères : moyenne normale, seuils pathologiques etc. On peut rejoindre le modèle bipolaire grâce au codage par rang étudié par L. Lebart : soit I un ensemble d'individus (constituant un échantillon satisfaisant pour l'étude en vue) ; Card I (nombre des individus) = N ; Q, un ensemble de qualités ordinales ; on notera $k(i,q^+)$ = rang de l'individu i au sein de I sur l'échelle de la qualité q ; $k(i,q^-) = N - k(i,q^+)$. On peut encore partager l'intervalle de variation de chaque qualité ordinale en autant d'intervalles que le spécialiste estime devoir distinguer de niveaux ; par exemple cinq : très fort, fort, moyen, faible, très faible : et l'on rejoint alors strictement la forme disjonctive complète. Le premier exemple d'un tel codage fut présenté par J. P. Nakache au colloque de Marseille (Septembre 1970) pour l'analyse de données biologiques. Cette représentation des données nous parut d'abord abusive : selon nous, il eut été préférable de donner au moins des valeurs continues aux nombres inscrits dans les colonnes affectées à une seule qualité. Par exemple lorsqu'un individu se trouve entre moyen et fort, lui donner des zéros dans les colonnes des autres modalités ; mais partager sa note entre celles-là : 0,4 dans moyen, et 0,6 dans fort s'il est plutôt fort, etc.. Arrondir ainsi les angles augmente certes la précision du codage ; mais écarte de la forme disjonctive complète, dont l'étude par L. Lebart s'est révélée si féconde (cf [Bin. Mult.], ce cahier pp 55 sqq). L'initiative de Nakache fut d'autant plus heureuse qu'en 1970 les analyses de questionnaires débutaient seulement. Depuis lors les données les plus diverses, les plus hétérogènes ont reçu grâce au codage sous forme disjonctive complète un format acceptable pour l'analyse. Ainsi nous nous trouvons analyser efficacement des tableaux de données qu'en toute rigueur méthodologique nous préférerions voir brûlés parce qu'ils manquent à la règle d'homogénéité et d'exhaustivité règle que nous répéterons ainsi (cf § 3.7.1) : faire du réel une coupe bien choisie, et y regarder comme en un miroir, toute la structure.

3.7.5 Grandeurs algébriques : Le programme d'analyse des correspondances requiert un tableau de nombres positifs (quelques nombres négatifs n'interdisent toutefois pas le calcul des facteurs, pourvu que la somme de toute ligne et de toute colonne reste positive) : que faire des grandeurs affectées d'un signe ? Dans la pratique, le cas n'est pas si fréquent qu'on le croirait *a priori*. Les grandeurs additives extensives (dénombrements, pesées etc..., cf § 3.7.2) sont essentiellement positives ; quand, en un certain sens, deux quantités s'opposent, par exemple les exportations et les importations, il ne convient pas d'en faire la somme algébrique : on doit les compter sur des colonnes distinctes : laisser apparaître tous les postes du bilan. Les variables logiques (cf § 3.7.3) sont le mieux codées par les deux nombres 0 et 1 dont aucun n'est négatif. Les grandeurs intensives repérées sur un intervalle bipolaire (cf § 3.7.4) ne sont pas rapportées à un centre et mesurées par des nombres algébriques (positifs à droite de l'origine, négatifs à gauche) mais rapportées aux deux extrémités, aux deux pôles ; les distances de ceux-ci fournissant les deux notes complémentaires $k(i,q^+)$, $k(i,q^-)$; il est également commun de traiter ces sortes de données sans dédoublement, comme des grandeurs centrées, par l'analyse en composantes principales : nous reviendrons au § 3.8.4 sur la comparaison des deux méthodes : disons tout de suite que les résultats diffèrent assez peu : F. Nakhlé (Thèse 1973 ; publiée dans les *Cahiers* Ca I n°s 3 & 4) a montré que les facteurs issus de l'analyse de correspondance d'un tableau à n variables dédoublées (tableau donc à 2 n colonnes) peuvent être calculés par diagonalisation d'une matrice n x n (comme pour le tableau à n colonnes que considère l'analyse en composantes principales ; la distance carrée entre deux individus restant une combinaison des carrés des différences de leurs notes mais affectés de coefficients) ; et il a écrit à cet effet un programme spécial. Cependant le codage logique sous forme disjonctive complète reste

toujours possible, et il est le meilleur si parmi un ensemble de variables soit extensives, soit qualitatives, il se rencontre une seule grandeur vectorielle ; par exemple, pour le vecteur *vitesse du vent* on découpera le plan en cinq zônes : vent faible, fort vent du nord, fort vent d'est etc...; ou en un plus grand nombre de zônes choisies après examen de l'histogramme bidimensionnel du vecteur vent. De même, si les trajectoires des particules produites dans des réactions à haute énergie sont repérées non par leurs impacts dans une suite de plans parallèles, mais par un système de détecteurs dont la configuration est complexe, il faudra analyser les données après un codage qui traite chaque détecteur comme une question q avec pour ensemble J_q des réponses d'une part l'absence d'impact (si la trajectoire n'a pas rencontré de détecteur) d'autre part des cellules se partageant la surface du détecteur.

3.7.6 Données manquantes : Les tableaux proposés aux statisticiens présentent souvent des lacunes : si celles-ci ne sont ni fréquentes ni systématiques, on pourra les combler avec une précision suffisante pour que l'analyse soit fructueuse. Cette complétion des données est une sorte de codage, c'est pourquoi nous la présentons ici. Une méthode bien connue aujourd'hui après les travaux de F. Mutombo, Ch. Nora, B. Tallur, consiste en bref à utiliser la formule de reconstitution des données en fonction des facteurs pour des approximations successives : les facteurs obtenus par analyse du n^e tableau servant à combler les vides pour obtenir le tableau de rang n + 1.

3.8 L'interprétation :

Le point de vue original de l'analyse des correspondances est l'étude d'un nuage de points $N(I)$ (resp. $N(J)$) dans un espace euclidien : au centre des masses du nuage $N(I)$ des profils f_J^i (des divers éléments i de I) est le profil moyen f_J (ou profil marginal). L'analyse factorielle construit un système ordonné d'axes orthonormés (les axes factoriels) issus du centre f_J. Les facteurs $F_\alpha(i)$ sont les coordonnées du point i (des profils f_J^i) projetés sur ces nouveaux axes. Dans ce cadre géométrique, il est facile de définir de nouvelles notions qui aident à l'interprétation (§§ 3.8.1 & 3.8.4) ; de corroborer celle-ci par la classification automatique (§ 3.8.3) ; de faire servir la méthode inductive à des problèmes qui comme ceux de la régression et de la discrimination (§ 3.8.2) ont été initialement résolus par ajustement aux données d'une structure *a priori*.

3.8.1 Eléments supplémentaires : Tout autre profil f_J^s que ceux des éléments i de I peut aussi être projeté sur les axes factoriels ; on peut donc calculer des facteurs pour un individu s qui n'a pas été d'abord pris en compte dans la détermination des axes : c'est ce qu'on appelle un élément supplémentaire. L'introduction des éléments supplémentaires permet de placer sur les graphiques issus de l'analyse d'un échantillon I représentatif de la population à laquelle on s'intéresse (e.g. les malades atteints d'une affection hépatique), un sujet nouveau s qui se trouvera entouré d'individus i qui lui ressemblent, et d'après lesquels le cas de s pourra être mieux compris : c'est là le principe d'une nouvelle méthode de régression (cf § 3.8.2). On traitera encore en éléments supplémentaires les centres de gravité de certaines classes d'individus ; ainsi qu'un individu, une variable dont les mesures (ligne ou colonne) semblent soit entachées d'erreurs, soit quelque peu excentriques relativement au domaine principal de l'étude et menacent de perturber l'analyse, ou l'ont effectivement perturbée dans un premier essai. La mise en élément supplémentaire est très simple dans son principe ; mais elle n'est entrée dans le programme d'analyse des correspondances qu'en 1967 par un sous-programme dû à Fr. Friant.

3.8.2 *Régression et discrimination* : En régression les données sont scindées en deux blocs : d'une part la variable à expliquer ; de l'autre les variables explicatives ; et l'on cherche une formule (d'un type algébrique plus ou moins clairement fixé *a priori*) exprimant la première en fonction des dernières. La discrimination (§ 2.3.5) n'est qu'un cas particulier (de la régression) où la variable à expliquer prend ses valeurs dans un ensemble essentiellement fini (e.g. un ensemble de trois affections hépatiques ; qu'on doit distinguer d'après les variables biologiques explicatives). Méthode inductive (cf § 3.2.1), l'analyse des correspondances vise au contraire à extraire des facteurs qui révèlent et expriment mathématiquement des qualités non directement mesurables ; elle reçoit sa confirmation en retrouvant au passage des variables absentes du tableau des données mais explicitement connues par ailleurs. Souvent l'analyse d'un tableau de correspondance offrant à un certain niveau une représentation exhaustive et homogène d'un domaine naturel (cf § 3.7.1) a fourni directement en facteur une variable à expliquer (ainsi l'analyse d'une matrice de confusion entre signaux du code Morse range sur le 1° axe ces signaux dans l'ordre de leur durée) ; ou séparé dans le plan des axes 1 x 2 (ou l'espace 1x2 x 3) deux sous-nuages qu'il fallait distinguer (cf Danech-Pajouh, 1972 ; T. Moussa, 1972 ; et au § 2.3.5 l'exemple du genre Iris). Mais l'analyse d'un tableau de correspondance croisant variable à expliquer et variables explicatives (mises sous forme disjonctive complète) s'est révélée très utile . Soit donc I l'ensemble des modalités de la variable à expliquer (pour une variable continue y , ces modalités pourront être dix intervalles successifs en lesquels est partagé son intervalle global de variation) ; $J = \cup \{J_q | q \in Q\}$, l'ensemble des modalités de toutes les variables explicatives (Q désigne l'ensemble de ces variables ; et J_q est l'ensemble des modalités de la variable q ; cf §3.7.3); C l'ensemble des cas : par cas, c, on entend un individu, (ou une situation individuelle) pour lequel on a simultanément déterminé la variable à expliquer, et l'ensemble Q des variables explicatives. On soumet d'abord à l'analyse des correspondances le tableau k_{IJ} de cooccurrences des modalités : k(i,j) = nombre de cas où ont été associées la modalité i de la variable à expliquer et la modalité j d'une des variables explicatives. Au tableau k_{IJ} , chaque cas c fournit une ligne supplémentaire : k(c,j) = 1 si la modalité j appartient à la description de c, zéro sinon ; ce qui permet d'étendre à C les facteurs f_α issus de k_{IJ} , et de placer C avec I et J dans les diagrammes plans (e.g. plan des axes 1 et 2), ou les sous-espaces propres (e.g. sous-espace engendré par les axes 1,2,3 ; etc.).Ceci fait, dans la mesure où les variables explicatives apportent l'information nécessaire, on aura généralement une bonne approximation de la variable à expliquer (variable continue y) par une combinaison linéaire des facteurs $F_\alpha(c)$ (ce qui revient à une régression linéaire usuelle avec, pour variables explicatives, ces facteurs). Au prix d'un temps de calcul plus long(mais praticable) on aura des résultats plus précis grâce à ce qu'on appelle *la régression par boule*. Soit s un cas nouveau pour lequel ne sont connues que les variables explicatives, d'où une ligne $\{k(s,j) | j \in J\}$, permettant de calculer les facteurs $F_\alpha(s)$; plaçons s dans le plan 1 x 2 (ou dans l'espace 1 x 2 x 3...) on peut trier les 20 (ou 10) cas c (de C) qui dans ce plan sont les plus proches de s (sont contenus dans le voisinage ou *boule* de centre s) et calculer sur l'ensemble de ces cas la moyenne et l'écart-type de la variable y(c), d'où à la fois une estimation de y(s) et un ordre de grandeur de l'erreur commise. Si comme dans les problèmes de discrimination la variable à expliquer est une variable discrète, e.g. a trois modalités i_1 , i_2 , i_3, on comptera dans la *boule* p_1 cas c relevant de la modalité i_1 , p_2 de i_2 et p_3 de i_3 (avec $p_1 + p_2 + p_3 = 20$) et on dira que s peut être rattaché aux classes i_1 , i_2 , i_3 avec les probabilités respectives $(p_1/20)$,

$(p_2/20),(p_3/20)$. Ainsi la régression ou discrimination par boule fournit une estimation de la probabilité des causes ; c'est pourquoi on parle encore parfois de régression Bayesienne (cf § 2.3.3). La première analyse d'un tableau k_{IJ} croisant les modalités de la variable à expliquer et celles des variables explicatives fut faite par M. G. Caraux dans une étude agronomique (rendement de la culture du riz en Casamance - Sénégal ; 1971) ; la régression par boule a été appliquée d'abord par J. P. Bordet dans l'étude de la densité de la très haute atmosphère à l'altitude où circulent les satellites artificiels (Thèse:1973) ; le programme en usage au laboratoire est dû à M. O. Lebeaux (Thèse:1974 ; la notice de ce programme sera publiée dans le prochain cahier et une application en est donnée par C. Sabaton : cf ce cahier pp 79-96 et le cahier suivant).

3.8.3 *Classification automatique* : Revenons aux méthodes inductives : nous avons dit que dans un sous-espace propre (plan 1 x 2 ; espace 1 x 2 x 3 etc) issu de l'analyse d'un tableau de correspondance k_{IJ}, l'ensemble I pouvait apparaître partagé en des classes connues avant l'analyse, mais dont la composition n'était pas explicitement notée au tableau k_{IJ}. Ici il apparaît utile de conjuguer l'analyse de correspondance avec une autre méthode inductive visant à fournir non une représentation spatiale mais une classification. Le programme de classification que nous utilisons communément aujourd'hui est dû à M. Jambu : son principe, la méthode ascendante hiérarchique (en bref : réunir d'abord les deux individus les plus proches ; puis s'élever en constituant des classes nouvelles par réunion, on dit encore agrégation, de deux classes ou individus préexistants) est bien connue des taxinomistes (cf e.g. Sokal & Sneath : *Numerical Taxonomy*; signalons toutefois qu'un perfectionnement récent dû à M. Bruynooghe a grandement accéléré l'algorithme) ; il admet de multiples variantes différant par le critère d'agrégation choisi. La plus utilisée est l'agrégation suivant la variance avec pour distance celle du χ^2 ; le critère très classique n'est autre que la maximisation de la variance interclasse (de la dispersion du nuage des centres des classes), avec minimisation simultanée de la variance intraclasse (i.e. intérieure aux classes). Du fait de la distance choisie (cf § 3.2.3) la méthode se conjugue bien avec l'analyse de correspondance : il est notamment possible de donner de la variance totale du nuage $N(I)$ (représentant l'ensemble I à classer) une double décomposition suivant les noeuds de la classification et les axes de l'analyse factorielle, qui permet de conjuguer interprétation des axes et interprétation des facteurs (cf M. Sadaka Thèse 1974 ; et M. Jambu, programme version 1975 publié dans ces *Cahiers* : Ca I n° 1 pp 77-93). C'est là une généralisation de la notion de contribution, utilisée depuis 1969 en analyse de correspondance, et que nous exposons ci-dessous. Ainsi les formules de décomposition de l'inertie associées au grand nom de Huyghens, et entrées dans la statistique par l'analyse de la variance (§ 2.3.4) servent à l'analyse inductive des données.

3.8.4 *Calculs de contribution* : Notons $\rho^2(i)$ le carré de la distance (distance du χ^2) du profil f_J^i de l'élément i, au centre f_J du nuage $N(I)$: $\rho^2(i) = \|f_J^i - f_J\|^2$. On sait que l'inertie totale du nuage, ou trace $Tr = \lambda_1 + \lambda_2 + \ldots\ldots$ est la somme $\Sigma\{f_i\ \rho^2(i) \mid i \in I\}$: donc dans l'inertie totale du nuage l'élément i a une part $f_i\ \rho^2(i)/Tr$. De même sur un axe on a : $\lambda_\alpha = \Sigma\{f_i\ F_\alpha^2(i) \mid i \in I\} : f_i\ F_\alpha^2(i)$ est la contribution de i à la valeur propre λ_α (ou en bref à l'axe α). On sait encore que $\rho^2(i) = F_1^2(i) + \ldots + F_\alpha^2(i) + \ldots$: l'écart de i au centre, au profil moyen, est une somme de termes afférents aux facteurs successifs et dont l'importance relative est $F_\alpha^2(i)/\rho^2(i)$, quotient qui n'est autre

que le carré du cosinus de l'angle formé par l'axe α avec le vecteur joignant f_J^i au centre f_J. Il est essentiel d'avoir en vue ces diverses proportions quand on interprète et critique les résultats d'une analyse factorielle. Si par exemple $f_i F^2(i) \simeq \lambda_\alpha/3$ l'élément i fait à lui seul un tiers du facteur α : il est vraisemblable que ce facteur est instable, qu'il disparaîtra ou sera grandement perturbé si i est supprimé du tableau (ou mis en élément supplémentaire ; ce qu'on devra expérimenter). Au contraire si $F_\alpha(i)$ est très élevé mais que le produit $f_i F^2_\alpha(i)$ est petit relativement à λ_α , i bien que très en vue sur l'axe α ne joue aucun rôle dans la constitution de celui-ci.

De plus si $F_\alpha(i)^2$ est élevé relativement aux valeurs prises par ce même facteur α pour les autres éléments de I, mais faible relativement à $\rho^2(i)$ (parce que la part prépondérante de $\rho^2(i)$ appartient à un autre facteur F_β) le caractère principal de l'élément i ne sera pas exprimé par le facteur α (mais par le facteur β). Comme la mise en élément supplémentaire, le calcul des contributions repose sur des principes géométriques bien connus et le programme en est simple ; mais l'usage ne s'en est introduit que vers 1969 ; voici comment.

Depuis 1968, M. O. Lebeaux analysait les données recueillies par Madame L. de Bonis pour une thèse de psychologie. Des tableaux de grandeurs intensives bipolaires (comme nous les avons appelées au § 3.7.4) étaient simultanément analysés sans dédoublement, par l'analyse en composantes principales et avec dédoublement par l'analyse des correspondances (aujourd'hui nous préférerions analyser ces données après codage par classe ; en attribuant, par exemple 4 colonnes, 4 niveaux à chacune des variables : les diagrammes obtenus ainsi révèlent plus de nuances que ne le peuvent faire l'analyse en composantes principales, ou l'analyse du tableau dédoublé ; mais pour qu'il soit permis de multiplier les colonnes, il faut que l'échantillon des individus ait un effectif assez élevé ; e.g. 100). Dans l'ensemble les résultats concordaient (une comparaison précise des deux méthodes peut se faire d'après la thèse de F. Nakhlé : cf Ca I n° 3 pp 243 sqq) ; mais sur un axe issu de l'analyse de correspondance apparaissait parfois en position excentrique des variables au profil très contrasté (c'est à dire des colonnes dont les notes allaient du minimum au maximum possible ; disons de 0 à 1) dont pourtant la corrélation avec l'axe n'était pas des plus grandes. L'analyse en composante principale (cf § 2.4.4) ne présentait pas ce phénomène parce que dans cette analyse les variables sont toutes ramenées à avoir pour variance 1, et que par conséquent la caractéristique du lien entre une variable et un facteur appelée saturation n'est autre qu'un coefficient de corrélation (un cosinus, en terme géométrique) dont la valeur absolue ne peut dépasser 1. En analyse de correspondance on pouvait distinguer trois notions : le facteur $G_\alpha(j)$ (il est d'usage de prendre la lettre G et non F pour un facteur sur le deuxième ensemble : on écrit $F_\alpha(i)$, $G_\alpha(j)$) ; le coefficient de corrélation $G_\alpha(j)/\rho(j)$ (ou même son carré : le $\cos^2$, contribution relative de l'axe α à l'élément j) ; et $f_j G^2_\alpha(j)$, contribution absolue de l'élément j à l'axe α (à la valeur propre λ_α). Notions qui toutes trois ont leur rôle propre dans l'interprétation des résultats et la critique de leur validité.

Les calculs de contribution ont en effet permis non seulement de critiquer la stabilité et l'importance relative des résultats d'analyse, mais encore de pousser l'interprétation des facteurs au delà du 5ème ; ce qui fut fait pour la première fois dans le dépouillement d'une *Etude sur les conditions du développement de la Colombie* (TII C n° 6 § 5.6) en prenant pour indicateur du sixième facteur les éléments (questions) les plus corrélés avec celui-ci et, de plus, peu corrélés avec les

facteurs précédents. On conçoit que pour chercher ces indicateurs parmi un ensemble J qui peut compter plusieurs centaines d'éléments, il soit indispensable d'avoir une liste de toutes les contributions, élément par élément ceux-ci étant rangés dans l'ordre de leur projection sur l'axe qu'on considère (cf § 3.9).

3.8.5 *Stabilité et validité* : Dans les études statistiques conçues pour répondre à une hypothèse explicite, les épreuves de validité ont un rôle essentiel (cf §§ 2.2.3 & 2.2.6). En analyse des correspondances, il n'y a d'autre hypothèse *a priori* que l'existence entre les deux ensembles I et J d'une liaison dont on cherche la structure. On a vu (cf § 3.4.4) que l'épreuve classique du χ^2 fournit une estimation de la part significative de la trace ; donc un critère pour arrêter le nombre des facteurs significatifs. Mais d'une part cette épreuve ne vaut que sous la condition très restrictive que le tableau analysé dénombre des faits indépendants entre eux ; d'autre part la pratique de l'analyse des correspondances nous a convaincus qu'ordinairement l'interprétation s'enlise ou s'égare avant qu'on ait épuisé la part manifestement significative (i.e. non liée aux fluctuations) de la trace. C'est qu'il ne suffit pas d'affirmer que la disposition de I et J dans l'espace rapporté aux 3 premiers axes n'est pas due au hasard : il faut en comprendre le sens; faire dans ce qu'on observe sur les graphiques (positions relatives des points entre eux et avec les axes) la part de l'essentiel et celle du contingent ; ne s'attacher qu'à ce qui est stable. Or en analyse des données, les fluctuations d'échantillonnage affectent non seulement les nombres eux-mêmes recencés dans le tableau, mais surtout le choix du tableau lui-même (cf § 3.7.1) ; d'où le rôle essentiel des calculs de contribution (§ 3.8.4) et des mises en éléments supplémentaires (§ 3.8.1) pour critiquer la stabilité des résultats qu'on a remarqués;parfois l'effet de la suppression d'un élément peut être majoré efficacement sans refaire l'analyse : cf B. Le Roux et B. Escofier Ca I n° 3 pp 297 sqq. De plus, diverses épreuves de simulation (modification ou permutation aléatoire de certaines données etc.) ont pu être utilisées avec fruit. Pour ces recherches dues principalement à L. Lebart (voir aussi T. Moussa, thèse 1972) nous renvoyons à la leçon [Epr. Val.] (TII B n° 12).

3.9 *Organisation des programmes* :

Le premier programme écrit par B. Cordier (Mme J. P. Escofier) suivait les formules de la quatrième leçon du cours de 1964 (cf § 3.6.4). L'analyse de correspondance est un cas particulier de recherche des axes principaux d'inertie d'un nuage de points d'un espace euclidien ; les coordonnées sur ces axes sont définies comme vecteurs propres de l'application linéaire $m \circ \sigma$ (où m et σ sont des tenseurs d'ordre 2, ou matrices carrées : m est la métrique euclidienne ; et σ est la forme quadratique d'inertie du nuage). Le calcul des vecteurs propres se faisait par itération pour le premier, itération et orthogonalisation pour les suivants. La matrice m des coefficients de la métrique étant diagonale, et σ étant symétrique, le calcul des vecteurs propres aurait pu être réduit à la diagonalisation d'une matrice carrée symétrique (cf TII B n° 2 § 7.2), effectuée par un sous-programme de bibliothèque : mais cette réduction n'était pas faite. Dans la thèse de B. Cordier, le calcul des facteurs est fait d'abord pour l'un des ensembles (celui qui requiert le moins de calculs) ; puis la formule de transition donne les facteurs sur l'autre ensemble (cf § 3.3.2).

De 1965 à 1969, F. Friant et P. Leroy[+] perfectionnèrent le programme de B. Cordier, notamment par le tracé du nuage sur imprimante (graphique plan où chaque point désigné par trois caractères, a pour abscisse et ordonnée deux facteurs choisis ; e.g. le premier et le troisième), et par un sous-programme de calcul des facteurs pour les éléments supplémentaires (traités comme ayant masse nulle : cf § 3.8.1). En 1969 (cf § 3.8.4) M. O. Lebeaux ajouta les calculs de contributions et l'impression pour chaque axe α de la liste des individus des deux ensembles I et J rangés dans l'ordre du facteur α, avec sur la ligne afférente à

chaque individu tous les facteurs et contributions (et non seulement ceux de rang α).

Cependant la recherche des vecteurs propres effectuait de spectaculaires progrès, grâce à Givens, Golub, Householder, Reinsch etc. (cf TII B n° 12 § 4). Dès 1971 J. Robert constate qu'un algorithme dû à Golub & Reinsch fournit l'ensemble des facteurs issus d'un tableau rectangulaire en 10 fois moins de temps qu'il n'en fallait pour calculer les cinq premiers facteurs par des méthodes usuelles d'itération et orthogonalisation ! Le programme de J. Robert, complété et perfectionné par F. Nicolau, est celui publié dans la première édition du tome II du Traité. Simultanément M. Reinert utilise le programme SYMQR, distribué par IBM, pour calculer très rapidement l'ensemble des facteurs par diagonalisation d'un tableau carré symétrique. Puis J. P. Bordet conçoit, toujours autour de SYMQR, un programme d'analyse de correspondance qui calcule la matrice à diagonaliser sans qu'il soit nécessaire de tenir en mémoire centrale l'ensemble des données, celles-ci pouvant être présentées successivement par paquets à partir d'une mémoire auxiliaire d'accès rapide : ce programme peut traiter des tableaux dont le nombre des lignes (cardinal de I) est arbitrairement grand (plusieurs milliers si nécessaire), tandis que la longueur de chaque ligne (le cardinal de l'ensemble J) peut atteindre e.g. 200. Depuis 1974, le laboratoire utilise un programme de ce type (i.e. avec diagonalisation par SYMQR ; et introduction séquentielle des données) écrit par N. Tabet, dont les performances sont remarquables ; ce programme comporte de plus des sorties graphiques perfectionnées (notamment pour déplacer les points qui, ayant des coordonnées voisines, seraient si l'on n'y prenait garde, perdues ou imprimées en surcharge !) . Il est dans la deuxième édition du Traité, substitué au programme fondé sur l'algorithme de Golub & Reinsch.

Désormais grâce aux progrès effectués depuis dix ans par le calcul numérique et le calcul électronique, un laboratoire qui a accès à un grand centre de calcul peut traiter par l'analyse des correspondances en un temps acceptable (e.g. quelques minutes) les plus grands tableaux auxquels il soit raisonnable de s'intéresser (e.g. 200x5000) compte tenu des difficultés de la collecte des données très étendues et aussi de leur interprétation, même après réduction par le calcul. Pour la plupart, les praticiens ne sont pas encore avertis de cette puissance du calcul : dans la laborieuse collecte des données, ils mutilent souvent celles-ci par des simplifications ou des omissions souvent irrémédiables, en vue de renfermer le tableau dans un cadre étroit qu'ils croient être exigé par le calcul. Il faut toutefois reconnaître que certains questionnaires après codage sous forme disjonctive complète peuvent comporter quelque 500 colonnes ; ce qui sans être prohibitif est fort lourd à traiter ! Aussi est-il très intéressant que la voie soit ouverte à l'analyse des plus grands tableaux même sur ordinateur de petite capacité, grâce à l'approximation stochastique.

Cette méthode proposée en 1967 (cf J. P. BENZECRI : Approximation stochastique dans une algèbre normée non commutative ; *in Bull. Soc. Math. France* ; T 97 ; pp 225-241 ; 1969) calcule les premiers facteurs jusqu'à un rang choisi (d'abord sur l'ensemble J de plus faible effectif ; puis sur l'autre par transition) sans requérir le calcul explicite d'une matrice à diagonaliser. Les lignes du tableau, introduites d'une mémoire auxiliaire, une à une (ou plutôt par paquets aussi gros que le permet l'espace libre en mémoire centrale) sont successivement utilisées pour retoucher l'approximation des facteurs ; généralement le tableau doit être lu plusieurs fois avant stabilisation des résultats. Le programme étant très simple seul le tableau des facteurs en cours de calcul (tableau en deux états ; afin qu'on puisse suivre les fluctuations et noter la stabilisation) occupe en permanence la mémoire centrale : une mémoire de 32K suffit donc aux plus grands tableaux analysés jusqu'ici. Après un premier programme dû à J. P. Fénelon et expérimenté par M. Roux, la méthode fait l'objet de recherches de L. Lebart et N. Tabet ; les calculs sont très rapides et

généralement satisfaisants ; un point délicat est d'assurer le contrôle de la convergence (qui se réalise toujours, de par la théorie ; mais dans la pratique requiert un nombre variable de lectures du tableau) pour les tableaux de tout type. Présentement (en 1975) L. Lebart réalise les analyses de questionnaire mis sous forme disjonctive complète (cf 6 3.7.3) en cinq lectures du tableau *brut* des réponses ; l'éclatement logique des variables étant fait avec l'approximation stochastique ; ce qui accélère encore le programme. Outre que la méthode est très économique, il est frappant que dans l'ordinateur on aboutisse aux facteurs après des fluctuations successives, chaque ligne nouvelle apportant une retouche ; comme dans notre esprit, chaque fait nouveau corrige la vision synthétique que nous nous étions formés d'un domaine. A nous, cette analogie ne suggère pas que l'ordinateur soit pourvu d'intelligence ni que nous en soyons dépourvus ; c'est plutôt que l'ordinateur conduit par notre intelligence, est un *outil mental(*)* qui la sert après notre cerveau.

Enfin certains algorithmes servant à l'analyse des correspondances font l'objet de programmes séparés. Ainsi nous avons cité le programme de A.W. Hamrouni (modifié par Y. Grelet) pour calculer les pondérations relatives de deux tableaux $I \times J_1$ et $I \times J_2$ juxtaposés (§ 3.7.1), l'analyse faite par F. Nakhlé des tableaux dédoublés (§ 3.7.5) ; la reconstitution des données manquantes (§ 3.7.6). Citons encore l'ajustement d'ellipses de garde à un sous-nuage (J. P. Bordet) ; diverses épreuves de validité réalisées par simulation (L. Lebart ; cf [Epr. Val.] TII C n° 8). Et rappelons que le laboratoire utilise en classification hiérarchique (cf § 3.8.3) un programme de M. Jambu, riche en nombreuses variantes quant au calcul des distances et au critère d'agrégation des classes.

De 1963 à 1965 les premières analyses de correspondance furent faites à Rennes par B. Cordier : et l'on n'en faisait point ailleurs. A partir de 1965, des paquets de cartes déposés dans tel laboratoire de calcul accueillant (nous pensons particulièrement à celui du Professeur Laudet à Toulouse) ou confiés à des étudiants chaque année plus nombreux, ont servi sans que nous sachions à qui ou à quoi. Nous mêmes, bien que convaincus que seuls méritent analyse les tableaux de contingence recueillis sur une base homogène et exhaustive (cf § 3.7.1), en sommes venus à traiter des données qu'en toute rigueur méthodologique, nous préférerions voir brûlées (cf § 3.7.4) ! L'analyse des correspondances est une méthode ; elle est aussi un outil. A la philosophie de la méthode l'outil doit son efficacité ; mais, marteau sans maître, celui-ci frappe désormais librement. En nous appliquant à instruire des statisticiens philosophes, nous espérons au moins servir ceux qui saisissent l'outil pour dégager de la gangue des données le pur diamant de la véridique Nature.

(*) Le terme nous vient d'un auteur russe G.N. Povarov dont nous avons traduit la remarquable préface à la traduction publiée à Moscou d'un ouvrage de E.C. Berkeley : Symbolic logic and intelligent machines.

Bibliographie Générale

Cette bibliographie ne comprend pas tous les livres et articles que nous avons cités ; mais seulement quelques références à des sources assez facilement accessibles pour l'histoire des probabilités et des statistiques.

J. BERTRAND : Calcul des probabilités ; GAUTHIER-VILLARS ; Paris ; 1° éd. 1889 ; 2° éd., conforme à la première, (1907). Une troisième édition est actuellement disponible chez Chelsea, N.-Y. .

E. BOREL : Le Hasard ; 1° éd. 1914 ; nouvelle édition refondue : P.U.F. Paris 1948 ;

E. BOREL : Oeuvres : quatre volumes ; C.N.R.S. Paris (1972)

R.A. FISHER : *Constributions to Mathematical Statistics* : anthologie d'articles ; J. WILEY & SONS, INC ; CHAPMAN & HALL Ltd. (Londres) ; N.-Y. (1950).

R.A. FISHER : *Collected Papers : edited by* J.H. BENETT ; *The University of Adélaïde* ; T 1 (1971) ; T 2 (1972) ; ... .

B.V. GNEDENKO : Курс Теории Вероятностей ; Moscou, Léningrad ; (1950) ; cet ouvrage a été traduit en plusieurs langues.

P.S. LAPLACE : Essai philosophique sur les probabilités; d'après une leçon professée aux écoles normales ; sert d'introduction au suivant volume ; a été réédité seul par GAUTHIER-VILLARS ; Paris (1921).

P.S. LAPLACE : Théorie analytique des probabilités ; 1° éd. 1812 ; 3° éd. 1820 (avec supplément de 1825) ; la dernière faite par l'Auteur ; et Oeuvres complètes, Tome 7 ; GAUTHIER-VILLARS ; Paris ; (1886).

E.S. PEARSON & M.G. KENDALL : *Studies in the History of Statistics and Probability* ; recueil rassemblé par E.S. P. & M.G. K. d'articles historiques de divers auteurs parus de 1906 à 1968 dans *Biometrika* ; Charles GRIFFIN & Co Ltd ; Londres ; (1970).

K. PEARSON : A notre commaissance, il n'a pas été publié de recueil de ses oeuvres ; on pourra consulter divers périodiques, principalement *Biometrika* dont K. PEARSON a dirigé la rédaction de 1901 (fondation) jusqu'à 1936 (mort de K. P.).

H. POINCARE : Calcul des Probabilités ; leçons professées pendant le deuxième semestre 1893-1894 ; rédigées par A. QUIQUET ; GAUTHIER-VILLARS ; Paris ; (1896).

I. TODHUNTER : *History of the Mathematical Theory of Probability* ; 1° éd. ; (1865).