L'analyse des données

Enquêtes, sondages et statistiques accumulent un nombre grandissant d'informations sur des échantillons toujours plus importants. Pour dégager les paramètres caractéristiques de ces données, il s'est créé une nouvelle discipline mathématique que les ordinateurs ont rendue opérationnelle.

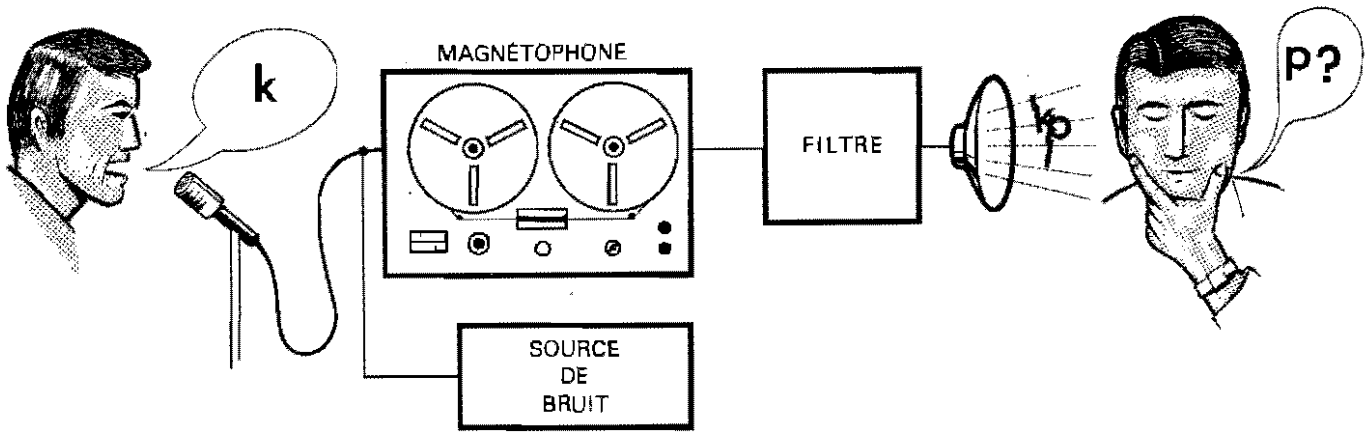
par J. M. Bouroche et G. Saporta

Pour prévoir l'avenir et ainsi optimiser les décisions de tous ordres, il faut aujourd'hui réunir un nombre toujours plus important de données (on utilise couramment des ensembles comprenant des centaines de milliers d'informations). Il est alors indispensable de traiter ces données selon des procédés rigoureux pour en dégager les paramètres importants.

Depuis une dizaine d'années, la statistique s'est enrichie d'une nouvelle famille de méthodes : l'analyse des données. Ces méthodes, développées pour la plupart avant l'apparition des ordinateurs, n'ont pu être appliquées que grâce à l'importante vitesse de calcul de ceux-ci et à leur faculté de mettre en mémoire de grandes quantités d'informations. Lorsqu'on dispose d'un fichier de 10 000 salariés sur lesquels ont été relevées plusieurs dizaines de caractéristiques tous les ans, il est difficile d'appréhender globalement l'information contenue. De plus les méthodes statistiques classiques sont insuffisantes. La statistique descriptive nous enseigne seulement comment représenter des pourcentages et comment construire des diagrammes qui mesurent l'intensité d'une caractéristique en fonction d'un paramètre, par exemple, le salaire en fonction de l'âge, de l'ancienneté, etc. La statistique mathématique, quant à elle, permet d'estimer des paramètres de distribution (moyenne, variance...) et de vérifier la validité d'hypothèses, notamment si deux séries de données correspondant aux variations de deux caractères sont corrélées. Ces méthodes ne sont pas suffisantes lorsqu'il s'agit d'extraire les informations d'un très grand fichier, d'où l'on veut, en dépistant des concomittances répétées, déduire les relations statistiques entre diverses caractéristiques : l'analyse des données autorise des études globales incluant toutes les caractéristiques de ces mêmes données; ces études ont pour but de mettre en lumière les phénomènes importants en faisant le minimum d'hypothèses *a priori* sur les importances relatives des informations.

Pour mieux comprendre ce dont il s'agit, prenous un exemple : les ressemblances entre consonnes, étudiées expérimentalement par G. A. Miller et P. E. Nicely et théoriquement par R. N. Shepard, J. D. Caroll et M. Wish. Il s'agit de déterminer les ressemblances entre seize consonnes, telles qu'elles sont perçues par l'oreille, ceci sans faire intervenir aucune idée préconçue sur la similarité de leur forme, sur leur spectre de fréquence etc. Les données sont rassemblées au cours de séances d'expérimentation, où un individu prononce une consonne au hasard et le son émis est dégradé de différentes manières : en lui superposant un bruit blanc, en le faisant passer à travers un filtre de fréquence, etc. Des auditeurs notent le son qu'ils perçoivent, et confondent certaines consonnes, ce qui fait qu'on peut reporter sur un tableau (une matrice de confusion), la fréquence des confusions, c'est-à-dire la fréquence avec laquelle elles sont prises l'une pour l'autre (voir figure 1) : on voit par exemple que les consonnes *b* et *d* sont confondues dans 5,8 % des cas; lorsque deux consonnes sont « proches », l'indice de confusion est élevé.

Une méthode d'analyse des données permet d'assigner aux différentes consonnes des positions sur un plan (voir figure 2). On peut vérifier que deux consonnes sont « proches » l'une de l'autre dans ce plan lorsqu'elles sont souvent confondues et inversement. La manière dont on les dispose sur le plan sera détaillée plus loin. Selon R. N. Shepard, les consonnes sonnantes (telles que *z*, *d*, *b*) forment un groupe séparé, le long de l'axe horizontal, des consonnes sourdes (telles que *t*, *k*, *p*). L'axe vertical sépare les consonnes en fonction de la nasalité (il s'agit d'une expérience réalisée aux États-Unis). Pour établir l'arbre (partie de droite de la figure 2), Shepard utilise une méthode de classification automatique; cette autre méthode d'analyse dont le résultat est également représenté sur la figure 2 permet de regrouper en classes (ou en types homogènes), les consonnes qui sont les plus proches (*k* et *p* par exemple); on considère ensuite cette classe comme une consonne fictive dont l'indice de proximité avec toutes les autres consonnes est recalculé : l'indice de la classe (*k*, *p*) avec une autre consonne, *s* par exemple, est le plus petit des deux indices de *k* avec *s* et de *p* avec *s*, c'est-à-dire le minimum du couple (0,063; 0,052), soit 0,052. Ce processus répété permet de construire l'arbre de classification. Sur cet arbre, le niveau d'un nœud est le plus petit indice entre les consonnes d'une même classe. Le tableau représentant les proximités et l'arbre de classification donnent le même résultat; on y distingue trois grandes classes de consonnes : les nasales, les sourdes non nasales et les sonnantes non nasales. Les deux analyses menées parallèlement sur les mêmes données donnent des interprétations compatibles qui s'enrichissent mutuellement : les groupements et les critères de ressemblance qui s'en dégagent en sont d'autant plus crédibles. Le tableau des données analysé dans cet exemple est de dimension très modeste; il arrive qu'il faille traiter avec ces mêmes techniques des tableaux bien plus grands. Mais avant d'exposer ces techniques, il nous faut considérer les données à analyser et les problèmes à résoudre. Puis nous examinerons les méthodes statistiques qui ont donné naissance à l'analyse des données. En 1960, on comptait environ 500 ordinateurs en Europe. En 1978, on en compte environ 100 000 de la même capacité. Le développement de l'ana-



p	–															
t	0,229	–														
k	0,432	0,241	–													
f	0,101	0,057	0,077	–												
Θ	0,124	0,079	0,084	0,423	–											
s	0,052	0,050	0,063	0,066	0,157	–										
ʃ	0,038	0,050	0,047	0,030	0,048	0,115	–									
b	0,022	0,013	0,018	0,046	0,045	0,024	0,012	–								
d	0,025	0,022	0,020	0,025	0,041	0,031	0,033	0,058	–							
g	0,013	0,016	0,030	0,015	0,039	0,033	0,021	0,069	0,342	–						
v	0,016	0,022	0,020	0,035	0,040	0,023	0,020	0,210	0,059	0,054	–					
ð	0,028	0,016	0,018	0,032	0,031	0,026	0,018	0,145	0,094	0,120	0,338	–				
z	0,025	0,023	0,026	0,018	0,033	0,035	0,017	0,055	0,106	0,139	0,080	0,161	–			
ʒ	0,019	0,017	0,019	0,007	0,017	0,022	0,012	0,027	0,089	0,125	0,029	0,033	0,138	–		
m	0,025	0,022	0,021	0,016	0,019	0,017	0,012	0,036	0,024	0,032	0,030	0,034	0,121	0,016	–	
n	0,017	0,018	0,020	0,012	0,018	0,013	0,011	0,024	0,032	0,030	0,022	0,028	0,016	0,030	0,151	–
	p	t	k	f	Θ	s	ʃ	b	d	g	v	ð	z	ʒ	m	n

1. CE TABLEAU DE CONFUSION, indiquant avec quelle fréquence une consonne est prise pour une autre, a été établi de façon expérimentale. Une personne émet une consonne, laquelle est enregistrée et déformée par une source de bruit et un filtre. Lorsque cette consonne est restituée, elle a été profondément transformée et peut être confondue avec une autre. Les deux consonnes les plus confondues et par conséquent les plus proches sont le *k* et le *p*. L'analyse des données précise cette notion de proximité.

lyse des données a été parallèle à celui des moyens de calculs; c'est probablement à travers les applications de l'analyse des données que l'ordinateur a le plus transformé nos vues statistiques du monde.

La nature des données.

On distingue généralement deux ensembles : les *individus* et les *caractères* relatifs à ces individus. Le terme « individu » peut désigner, selon les cas, un interviewé, l'employé d'une entreprise, un client, un animal, un lieu géographique, un pays, etc. L'ensemble des individus observés peut provenir d'un échantillonnage dans une population (il s'agit alors d'un sondage) ou il peut s'agir de la population tout entière.

Sur ces individus, on relève un certain nombre de caractères. Par exemple, si l'on considère une enquête par sondage, les caractères seront les questions; s'il s'agit des employés d'une entreprise, les caractères sont : le salaire, l'ancienneté, le diplôme, etc. Les caractères observés peuvent être quantitatifs ou qualitatifs. Un caractère est quantitatif lorsqu'il prend ses valeurs sur une échelle numérique : salaire, âge, chiffre d'affaire, taille, poids... Un caractère est qualitatif lorsqu'il prend des modalités non numériques; sexe, profession, région, couleur, niveau hiérarchique... Les modalités d'un caractère qualitatif peuvent être ordonnées (niveau hiérarchique); on dit alors que le caractère est ordinal. Sinon, on dit qu'il est nominal (profession, couleur).

Les données ainsi collectées peuvent être représentées dans un tableau explicitant les caractères des individus (voir figure 3) par exemple, le diplôme (caractère $\vec{x}^j$) du i^{eme} individu est représenté par la modalité x_i^j; son salaire est égal à x_i^2.

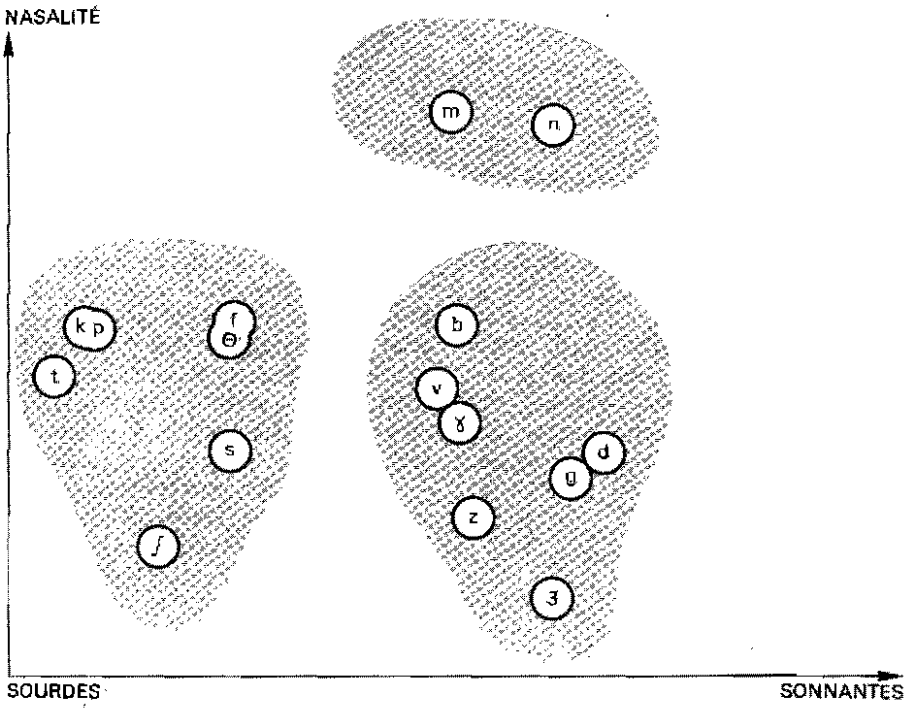
Bien entendu, on peut construire de nombreux autres types de tableaux de données. Citons par exemple, les tableaux de contingence et les tableaux de proximité. Un tableau de contingence contient les fréquences d'association entre les modalités de deux caractères nominaux. Si, par exemple, au cours d'un recensement, on relève sur les Français la classe d'âge et la région où ils habitent, on peut croiser ces deux caractères et construire le tableau des implantations géographiques en fonction de la classe d'âge. Cet exemple a été choisi pour les développements futurs qu'il va permettre : ici, la classe d'âge est considérée comme une modalité en dépit de son caractère numérique évident. On aurait pu également considérer le tableau des professions en fonction des quartiers de Paris qui relie cette fois deux véritables caractères nominaux.

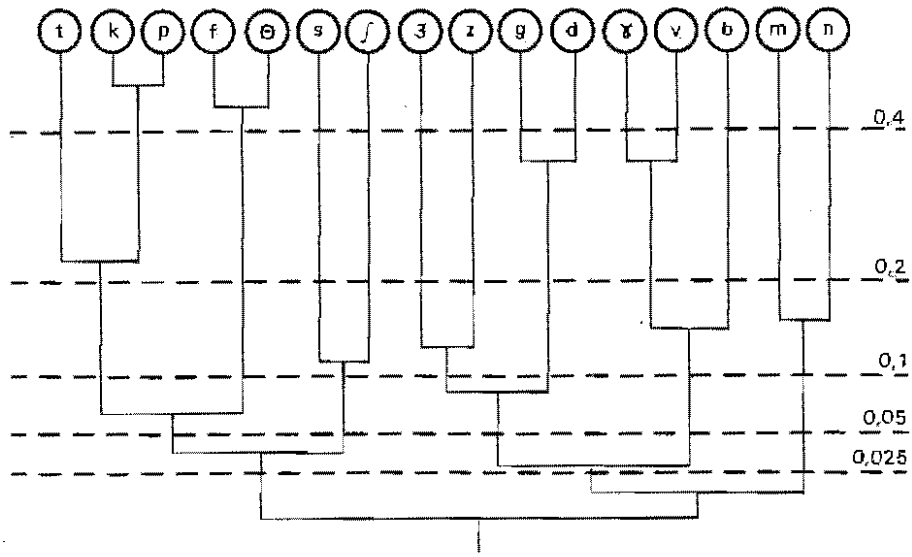
Dans le tableau de la figure 3, k_i^j représente le nombre de personnes dans la classe d'âge j habitant la i^{eme} région. Les

individus ont été agrégés et, contrairement au tableau précédent, ne peuvent plus être distingués. Le tableau de la figure 1 (association entre consonnes) contient des données de proximité. Le tableau des distances entre les principales villes de France est également un tableau de proximité *stricto sensu*.

Suivant la connaissance que l'on a du phénomène étudié, on peut chercher à analyser les données selon plusieurs points de vue. La recherche des ressemblances ou des différences entre les individus peut être un des objets de l'analyse. Par exemple, un économiste peut s'intéresser aux différents pays de l'OCDE. Ces pays sont représentés par les valeurs numériques prises par une batterie d'indicateurs économiques (voir figure 4). On considère que deux pays se ressemblent lorsque les profils caractérisant ces deux pays sont voisins. Il est possible, à l'aide d'une méthode d'analyse factorielle, de représenter les proximités entre pays; un exemple de ce genre est développé plus loin. Une autre méthode de classification automatique permettrait de regrouper les pays les plus proches relativement aux indicateurs économiques, de même que nous avons regroupé les consonnes les plus semblables. La description des relations entre caractères peut être un autre objet de l'analyse : deux caractères sont considérés comme liés, ou corrélés, lorsqu'ils varient de la même façon sur les différents individus; on peut, par exemple, privilégier un (ou plusieurs) caractère et chercher à expliciter ses variations en fonction des autres. Par exemple, un industriel peut chercher à expliquer son chiffre de vente auprès de différents clients à l'aide des caractéristiques de la clientèle. Une fois la formule établie, il pourra véritablement s'adapter à cette clientèle.

Lorsque tous les caractères jouent un rôle identique, on cherche simplement à mettre en évidence les groupes de caractères, soit corrélés, soit indépendants. Là encore, on peut utiliser l'analyse factorielle ou la classification. Selon le type de problèmes et selon la nature des données, on choisit la méthode appropriée.





2. L'ANALYSE DES PROXIMITÉS, permet de placer les consonnes sur un plan; leur proximité, traduisant leur fréquence de confusions, est établie expérimentalement. A partir du même tableau des données, c'est-à-dire le tableau de confusion (voir la figure précédente), on établit l'arbre où les consonnes sont groupées deux à deux par ordre de confusion croissante. Différentes sections de l'arbre correspondent à différents regroupements des consonnes. Pour la cinquième section, c'est-à-dire le 5e potutillé coloré, on différencie trois groupes de consonnes, les sourdes, les sonnantes et les nasales, regroupées sur le plan. La classification et l'analyse en composantes principales donnent les mêmes regroupements, ce qui est une confirmation de la validité des deux méthodes et de leur bonne interprétation.

Historique.

Les fondements mathématiques de l'analyse des données remontent au début du siècle : ils sont associés aux noms de K. Pearson et de H. Hotelling qui inventèrent l'analyse en composantes principales et l'analyse canonique. Les débuts véritables de l'analyse des données remontent aux travaux de C. Spearman et de l'école psychométrique anglo-saxonne. Les travaux de C. Spearman avaient pour but une mesure de l'intelligence; pour mesurer l'intelligence, on recourt à des batteries de tests auxquels on soumet de nombreux sujets : des données recueillies, multidimensionnelles, C. Spearman prétendait isoler un « facteur général » unique mesurant l'intelligence. Selon J. P. Benzecri (Histoire et préhistoire de l'analyse des données) « de ce que depuis des millénaires les hommes ont appelé intelligence, mémoire, imagination, patience..., aucune mesure immédiate n'est possible d'où la nécessité d'une construction statistique ».

Dans les années 30, L. L. Thurstone généralise et précise le problème en proposant un modèle moins contraignant et moins idéaliste. Il existe plusieurs types de facteurs et L. L. Thurstone différencie, dans son analyse, facteurs communs et facteurs spécifiques.

L'hypothèse fondamentale de l'analyse factorielle est qu'il existe un petit nombre de caractères numériques indépendants, non directement observables appelés « facteurs communs » et qui rendent compte des dépendances entre les quantités mesurées. Ces caractères s'expriment comme des sommes pondérées de ces facteurs à un terme près, le « facteur spécifique » du caractère, terme correcteur qui rend compte de l'aptitude particulière à exécuter un certain test. Le problème est donc de trouver pour chaque individu la valeur des différents facteurs, communs et spécifiques, et d'en déterminer les coefficients de pondération. L'analyse factorielle est donc un modèle *a priori* car on postule l'existence de facteurs communs et on cherche à vérifier le modèle à partir des données expérimentales. Les conceptions plus modernes visent à extraire l'information utilisable sans hypothèses *a priori* : cependant, la démarche qui fait apparaître des caractères cachés plus significatifs et en plus petit nombre que les caractères initiaux reste à la base des méthodes plus modernes d'analyse des données que l'on qualifie par le terme générique d'analyse factorielle, « l'analyse en composantes principales » en étant peut-être le plus bel exemple.

D'ailleurs, l'analyse factorielle en facteurs communs et spécifiques et l'analyse en composantes principales fournissent souvent des résultats voisins : les facteurs communs calculés par la première technique apparaissent naturellement sous forme de composantes principales dans la seconde qui a l'avantage d'être plus simple et plus sûre.

A l'opposé des méthodes issues de l'analyse factorielle, qui visent à élucider des relations entre caractères, les méthodes de classification et de typologie visent à faire apparaître des groupes homogènes d'individus ou de caractères. Elles tirent leur origine dans les travaux des naturalistes. La classification du règne animal de Linné est une des premières tentatives de classification hiérarchique. Depuis cette époque, naturalistes, botanistes et zoologistes n'ont cessé de manipuler d'immenses recueils de données, par des méthodes que la statistique a reprises et enrichies à l'aide des ordinateurs.

L'analyse en composantes principales.

Si à un ensemble d'individus on n'associait qu'au plus deux ou trois caractères, il serait facile de représenter les individus par un ensemble de points appelé « le nuage », dans un graphique cartésien où chaque coordonnée représente la mesure d'un des caractères. Une simple inspection visuelle apporte, dans ce cas, une foule de renseignements concernant la dépeudance entre les caractères, le repérage d'individus exceptionnels, la séparation entre d'éventuels groupes d'individus.

Lorsque le nombre des caractères est p, un individu i est un point, dans un espace à p dimensions, de coordonnées x_i^1, x_i^2 ... x_i^p . Cet espace est appelé l'espace des individus. Mais si p dépasse 3, il nous est impossible de représenter la figure formée par le nuage de points et d'en tirer « par inspection » des conclusions.

L'analyse en composantes principales va nous aider à obtenir une représentation dans un espace de dimension familière, en réduisant le nombre de caractères descriptifs : cette réduction sera d'autant plus facile que les caractères présenteront entre eux une corrélation importante. Les caractères obtenus grâce à cette analyse ne constituent pas une simple sélection des caractères de départ : ce sont de nouveaux caractères appelés caractères principaux (ou composantes principales) réalisant la synthèse de plusieurs caractères initiaux au moyen d'une formule linéaire (établie par le calcul) du type :

$\hat{c} = 0{,}3\,\hat{x}^1 + 0{,}2\,\hat{x}^2 - 1{,}5\,\hat{x}^3 + \ldots$

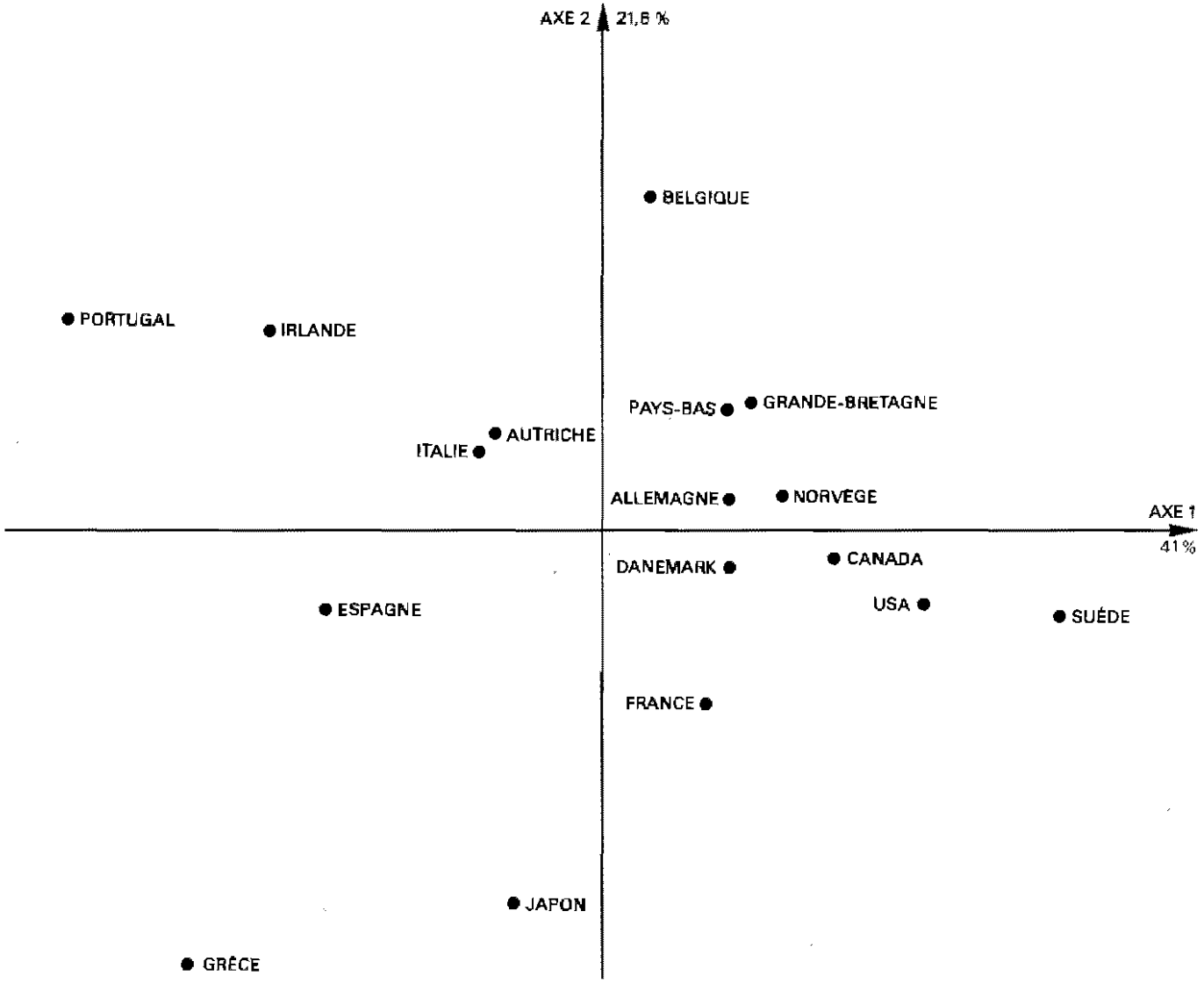
Il conviendra ensuite et dans la mesure du possible, de donner un sens concret aux composantes principales : c'est le difficile problème de l'interprétation.

Mais voyons d'abord le principe de la méthode : si on connaît les coordonnées

INDIVIDUS	CARACTÈRES					
	x^1 Age	x^2 Salaire	.....	x^j Diplôme	.....	x^p Ancienneté
1	x_1^1	x_1^2	.....	x_1^j	.....	x_1^p
2	x_2^1	x_2^2	.....	x_2^j	.....	x_2^p
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
i	x_i^1	x_i^2	.....	x_i^j	.....	x_i^p
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
n	x_n^1	x_n^2	.....	x_n^j	.....	x_n^p

RÉGIONS	CLASSE D'AGE					
	1 0-4 ans	2 4-9 ans	.....	j	.....	P
1	k_1^1	k_1^2	.....	k_1^j	.....	k_1^p
2	k_2^1	k_2^2	.....	k_2^j	.....	k_2^p
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
i	k_i^1	k_i^2	.....	k_i^j	.....	k_2^p
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
n	k_n^1	k_n^2	.....	k_n^j	.....	k_n^p

3. LES DONNÉES COLLECTÉES peuvent être représentées sur différentes sortes de tableau. Le tableau de gauche dénombre les divers caractères en fonction des individus. Ces caractères peuvent être quantitatifs ou qualitatifs. Le tableau de droite est un tableau de contingence où on a relevé le nombre d'habitants d'une classe d'âge déterminée en fonction des régions. Ces tableaux de contingence contiennent les fréquences d'association entre les modalités des deux caractères nominaux considérés dans l'analyse.



CARACTÈRES	COEFFICIENTS DE CORRÉLATION AVEC LES COMPOSANTES PRINCIPALES	
POURCENTAGE DE LA POPULATION ACTIVE TOTALE TRAVAILLANT DANS L'AGRICULTURE, LA SYLVICULTURE ET LA PÊCHE	– 0,8726	– 0,3428
POURCENTAGE DE LA POPULATION ACTIVE TOTALE TRAVAILLANT DANS L'INDUSTRIE	0,3541	0,5063
PRODUIT NATIONAL BRUT PAR HABITANT	0,9269	– 0,0536
POURCENTAGE DU PRODUIT INTÉRIEUR BRUT PROVENANT DE L'AGRICULTURE	– 0,8877	– 0,1563
FORMATION BRUTE DU CAPITAL FIXE en % du PNB	– 0,2152	– 0,5889
RECETTES COURANTES DE L'ÉTAT en % du PNB	0,7662	0,1855
RÉSERVES OFFICIELLES D'OR ET DE DEVISES en % du PNB	– 0,7005	0,5003
TAUX D'ESCOMPTE OFFICIEL	0,0685	– 0,3757
IMPORTATIONS TOTALES (CAF) en % du PNB	0,3944	– 0,7921
EXPORTATIONS TOTALES (FOB) en % du PNB	0,6818	– 0,0285
NOMBRE DE LOGEMENTS ACHEVÉS POUR MILLE HABITANTS	0,9166	0,0278
CONSOMMATION NETTE D'ÉLECTRICITE EN KWH PAR PERSONNE ET PAR AN (PERTES EN LIGNES DÉDUITES)	– 0,1349	0,6183
NOMBRE DE RÉCEPTEURS DE TÉLÉVISION POUR MILLE HABITANTS	0,2038	0,6889

4. REPRÉSENTATION PLANE des pays de l'O.C.D.E. L'analyse en composantes principales selon un plan donne cette répartition des différents pays. A partir des treize indicateurs dont la liste apparaît sur le tableau, on peut, par cette analyse, déguger deux composantes qui, après interprétation, représentent l'une le développement économique (*en abscisse*), l'autre les investissements et le commerce extérieur (*en ordonnée*). Le pourcentage d'information conservé par cette simplification selon deux dimensions, alors que l'espace initial en avait treize, est de 62,59 %. On remarquera la place du Japon qui investit beaucoup et exporte relativement peu. Sous le graphique, on a indiqué les corrélations des différents caractères avec les composantes principales, en coloriant les plus importantes. Ce sont ces corrélations qui permettent de donner un sens économique aux composantes principales dont le sens n'est pas connu *a priori*.

des n points représentatifs des individus, on peut calculer les distances entre tous les points pris deux à deux.

Notre préoccupation va être de trouver un sous-espace ayant un petit nombre de dimensions dans lequel on puisse faire une représentation de points-individus sans trop déformer les distances initiales entre les points; supposons que l'espace des individus soit à trois dimensions ($p = 3$) et que l'on cherche à en faire une représentation à deux dimensions, c'est-à-dire sur un plan (voir figure 5). En projetant les points-individus sur un plan, les distances entre les points projetés ne peuvent être qu'inférieures aux distances initiales. Le problème est : comment choisir au mieux ce plan?

Avant de résoudre ce problème, il faut introduire l'idée de dispersion et par conséquent refaire un peu de statistique : pour caractériser un ensemble de n nombres désignés par x_1, x_2 ... x_n ($\tilde{x}$ est une variable dont les valeurs sur n individus sont les x_i). La moyenne $\bar{x}$ est égale à :

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_n \ldots}{n}$$

Ainsi, les dix valeurs suivantes 3 100, 2 500, 2 800, 3 200, 4 000, 2 500, 3 000, 2 700, 3 000, 2 900 représentant les revenus mensuels de dix individus ont pour moyenne 2 970 francs. Caractériser un ensemble de nombres par sa moyenne est insuffisant : les dix revenus suivants 1 800, 2 000, 1 900, 4 500, 6 000, 5 000, 1 600, 2 400, 2 500, 2 000 ont aussi pour moyenne 2 970 francs mais il est clair que l'échelle des revenus n'est pas semblable : les valeurs sont plus dispersées. Pour quantifier la dispersion des valeurs, on utilise la variance notée s^2 :

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots (x_n - \bar{x})^2}{n}$$

Cette variance est d'autant plus forte que les valeurs de $\tilde{x}$ présentent de grands écarts entre elles. Si $s^2 = 0$, toutes les valeurs sont égales à la moyenne et le caractère est, en fait, une constante. La racine carrée s de la variance est mieux interprétable que la variance elle-même car elle est exprimée avec la même unité que le caractère : c'est l'écart-type. Ainsi, dans nos échelles de salaires, la première série a pour variance $s^2 = 168\,100$ et pour écart type $s = 410$ F; la deuxième série $s^2 = 2\,246\,100$ et $s = 1\,498{,}70$ F. La seconde série est 3,7 fois plus dispersée que la première.

Revenons au problème de choisir le meilleur plan possible pour observer les caractéristiques d'un nuage de points à trois dimensions. Le meilleur plan de projection est celui pour lequel les distances entre tontes les projections sont aussi grandes que possible, c'est-à-dire aussi semblables que possible aux distances initiales. Le critère ainsi retenu dans l'analyse en composantes principales est de rendre maximale la somme des carrés des distances entre les points projetés, c'est-à-dire de choisir le plan pour lequel les projections sont le plus dispersées. On peut montrer que ce plan passe toujours par le centre de gravité G des points représentant les individus; ce point a pour coordonnées les p moyennes des n coordonnées des caractères; pour continuer l'identification de ce plan, on cherche la droite (sous-espace de dimension 1) sur laquelle les distances entre points projetés sont les plus grandes possibles : c'est sur cette droite que les points sont le plus dispersés, ou en d'autres termes, que la variance de l'ensemble des n coordonnées sur cet axe (le premier axe principal) est la plus grande possible (voir figure 6). Cet axe choisi, on en détermine un autre, qui vérifie la même propriété avec la contrainte d'être perpendiculaire au premier.

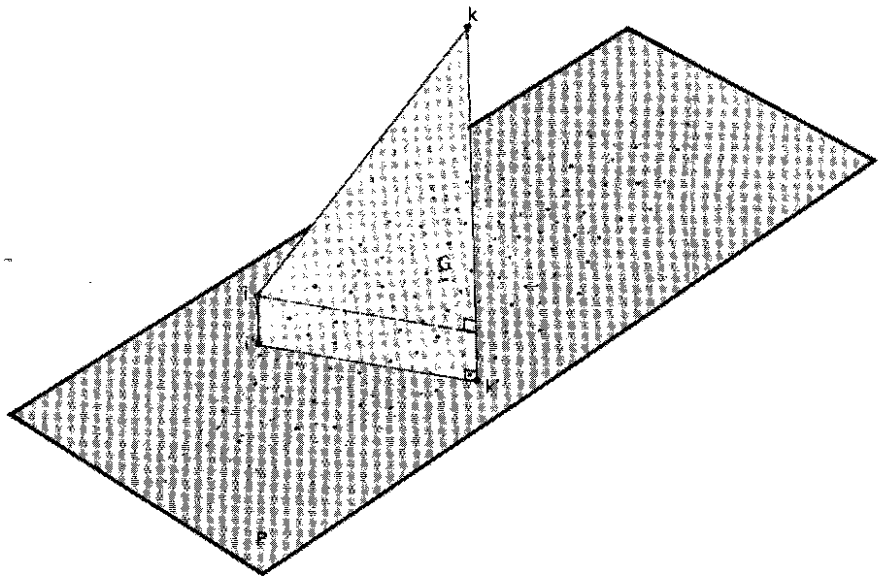
La « qualité » d'un axe est mesurée par le rapport entre la somme des carrés des distances entre points projetés et la somme des carrés des distances entre points individus. Ce rapport est appelé pourcentage d'inertie ou encore part de variance expliquée.

Pour choisir le meilleur plan, il suffit de déterminer deux axes principaux. Le pourcentage d'inertie attaché au plan contenant ces deux axes est égal à la somme des pourcentages d'inertie sur chaque axe, ceci parce que les axes sont perpendiculaires (cette affirmation se déduit immédiatement du théorème de Pythagore). Comme les pourcentages sont additifs, on arrête les recherches des axes principaux lorsque la somme des pourcentages d'inertie est suffisante, compte tenu du nombre des dimensions initiales de l'espace des individus. Les coordonnées $c_1 c_2$ ..... c_n des individus sur un axe principal représentent alors les différentes valeurs du caractère $\tilde{c}$, la composante principale. Nous avons déjà mentionné que celle-ci est obtenue (car c'est en fait une formule de changements d'axe) par une combinaison linéaire des caractères de départ,

$\tilde{c} = a_1 \tilde{x}^1 + a_2 \tilde{x}^2 + \ldots + a_p \tilde{x}^p$.

L'ensemble des coefficients (a_1, a_2 ... a_p) constitue ce qu'on appelle le facteur principal.

L'analyse en composantes principales



5. L'ANALYSE EN COMPOSANTES PRINCIPALES consiste à réduire le nombre de dimensions et l'espace des individus et, de façon pratique à grouper les paramètres pour ne dégager que ceux qui sont les plus importants. En effet le nombre des dimensions initiales de l'espace des individus est généralement élevé, égal au nombre des caractères initiaux auxquels on s'intéresse pour chaque individu. A titre d'illustration, on a ici un espace des individus à trois dimensions qu'on désire représenter par un plan (espace à deux dimensions) tout en gardant le maximum des informations contenues dans l'espace de départ. Pour cela, on choisit le plan tel que la somme des carrés des distances D' entre les projections des différents points-individus soit maximale. Cette somme est toujours inférieure à la somme D des carrés des distances entre les points-individus dans l'espace à trois dimensions. Le rapport D'/D mesure le pourcentage d'information conservée : géométriquement, c'est une mesure de l'aplatissement du nuage de points au voisinage du plan.

repose sur l'hypothèse de départ suivante : seul un nombre limité de caractères sont indépendants et les autres peuvent s'en déduire. Pour préciser cette notion de dépendance, nous allons introduire le coefficient de corrélation linéaire qui mesure l'intensité de la liaison entre deux caractères quantitatifs. On a relevé pour $n = 10$ appartements, deux caractères qui sont le prix de vente en milliers de francs et la surface en mètres carrés :

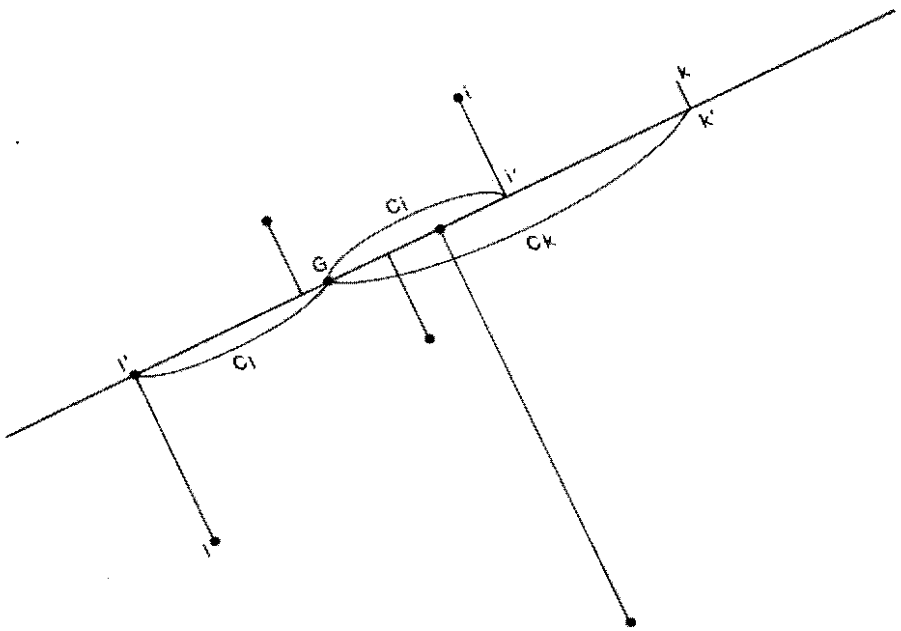
surface : x

28; 50; 55; 60; 48; 35; 86; 65; 32; 52

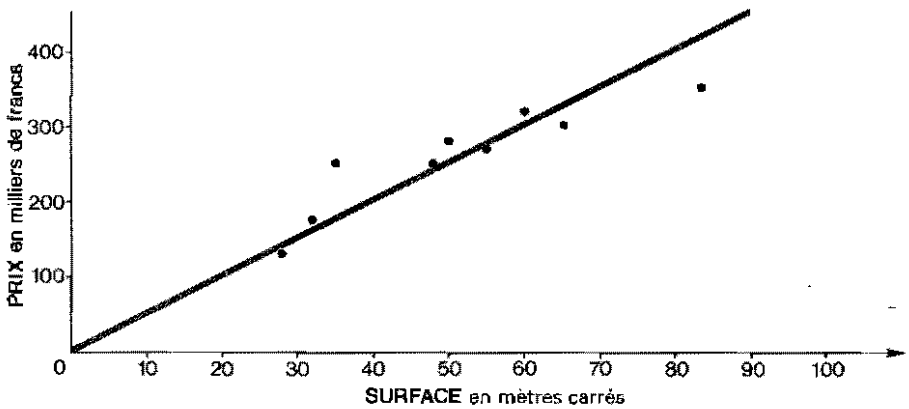
prix : y

130; 280; 268; 320; 250; 250; 350; 300; 155; 245

Le nuage de 10 points (voir figure 7), semble effilé le long d'une droite et il paraît raisonnable, si l'on veut prévoir le prix en fonction de la surface, de poser une formule $y = ax + b + e$ où e est une variable d'erreur. Les coefficients a et b sont obtenus par une méthode des moindres carrés, c'est-à-dire choisis de façon à rendre minimale la somme $(e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2)$. La droite d'équation $y^* = ax + b$ passe toujours par le centre de gravité du nuage de coordonnées $\bar{x}$ et $\bar{y}$. Dans le cas mentionné précédemment, $\bar{x} = 51{,}1$; $\bar{y} = 254{,}8$ et pour cet exemple $y^* = 3{,}524\,x + 74{,}707$. Le rapport sans dimension $(e_1^2 + e_2^2 + \ldots + e_n^2)/[(y_1 - y_2)^2 + \ldots + (y_n - y)^2]$ est toujours inférieur à 1. On pose ce rapport égal à $1 - r^2$ et r est le coefficient de corrélation linéaire. Si $r = 0$, la droite est horizontale, autrement dit, la valeur de x ne joue aucun rôle pour prévoir y. Si $r = \pm 1$, la prévision est parfaite car les écarts e_i sont nuls; le coefficient de corrélation r est d'autant plus grand que la valeur d'un caractère implique celle de l'autre, à condition que la relation entre ces caractères soit linéaire. Dans l'exemple précédent r valait 0,89.



6. L'AXE PRINCIPAL est tel que la somme des carrés des abscisses mesurés sur cet axe par rapport au centre de gravité est maximale. En effet, on démontre que tous les axes principaux passent par le centre de gravité G et que la condition mentionnée précédemment est équivalente à d'autres formulations (somme des carrés des distances entre points projetés maximale par exemple). Une fois obtenu ce premier axe principal, on choisit le second axe qui satisfasse la même propriété, avec une contrainte : le second axe doit être perpendiculaire au premier. Ces deux axes définissent le plan principal. On peut avoir ensuite besoin d'un troisième axe, etc.



7. LE PRIX DES APPARTEMENTS est une fonction quasi linéaire de leurs surfaces. Une méthode des moindres carrés détermine la droite (*en couleur*) qui représente le plus exactement cette relation. Le coefficient de corrélation compris entre — 1 et + 1 est d'autant plus grand que les points se rapprochent de la droite, c'est-à-dire que la somme des carrés des distances e_i est petite. Dans cet exemple, le coefficient de corrélation est égal à 0,89. La surface est par conséquent un facteur essentiel du prix des appartements à Paris.

Une autre interprétation du coefficient de corrélation linéaire, plus abstraite mais féconde, est la suivante : $\vec{x}$ et $\vec{y}$ sont des éléments d'un espace à n dimensions puisque les coordonnées de $\vec{x}$ et $\vec{y}$ sont $(x_1, x_2 \ldots\ldots x_n)$ et $(y_1, y_2 \ldots\ldots y_n)$. $\vec{x}$ et $\vec{y}$ sont ramenés à une moyenne nulle $(\bar{x} = 0, \bar{y} = 0)$; si ce n'est pas le cas, on considérera le vecteur $(x_1 - \bar{x}; x_2 - \bar{x} : \ldots; x_n - \bar{x})$; l'angle formé par les vecteurs $\vec{x}$ et $\vec{y}$ dans l'espace à n dimensions (espace des variables) a pour cosinus le coefficient de corrélation linéaire r.

Nous avons jusqu'ici présenté les composantes principales et les axes principaux, mais sans indiquer de procédé explicite d'obtention. On peut démontrer, mais c'est en dehors des limites de cet exposé, que ces axes principaux sont les vecteurs propres de la matrice des coefficients de corrélation linéaire r_i^j entre tous les caractères pris deux à deux et que les pourcentages d'inertie sont égaux aux quotients des valeurs propres par la trace de la matrice. Le processus de diagonalisation correspondant à l'obtention des vecteurs propres montre que ces vecteurs propres, donc les axes principaux, sont perpendiculaires et que deux composantes principales ont toujours un coefficient de corrélation nul.

En 1970, on a ainsi relevé sur 18 pays de l'OCDE (les *individus*) les valeurs de 13 caractères économiques. L'espace des individus a ainsi 13 dimensions. L'analyse en composantes principales permet d'obtenir le plan (projection de l'espace à 13 dimensions sur un espace

à deux dimensions) de la figure 4 et le tableau suivant montre les variations du pourcentage d'inertie avec le nombre d'axes principaux retenus.

numéro de l'axe principal	1	2	3	4
% d'inertie	41,04	21,55	16,66	7,78
% d'inertie cumulée	41,04	62,59	79,27	87,05

En se contentant de la représentation plane, le plan principal (deux axes principaux) permet de reconstituer les carrés des distances pour 62,59 %. C'est une mesure de l'information extraite, ou du point de vue géométrique, de l'aplatissement du nuage des points individus au voisinage de ce plan. Si tous les points étaient dans le plan, le pourcentage d'inertie serait égal à 100 % : il serait alors évidemment inutile d'observer autre chose qu'un plan.

En examinant maintenant les caractères les plus corrélés à une composante principale, on peut donner un sens aux composantes principales donc aux axes du plan de projection.

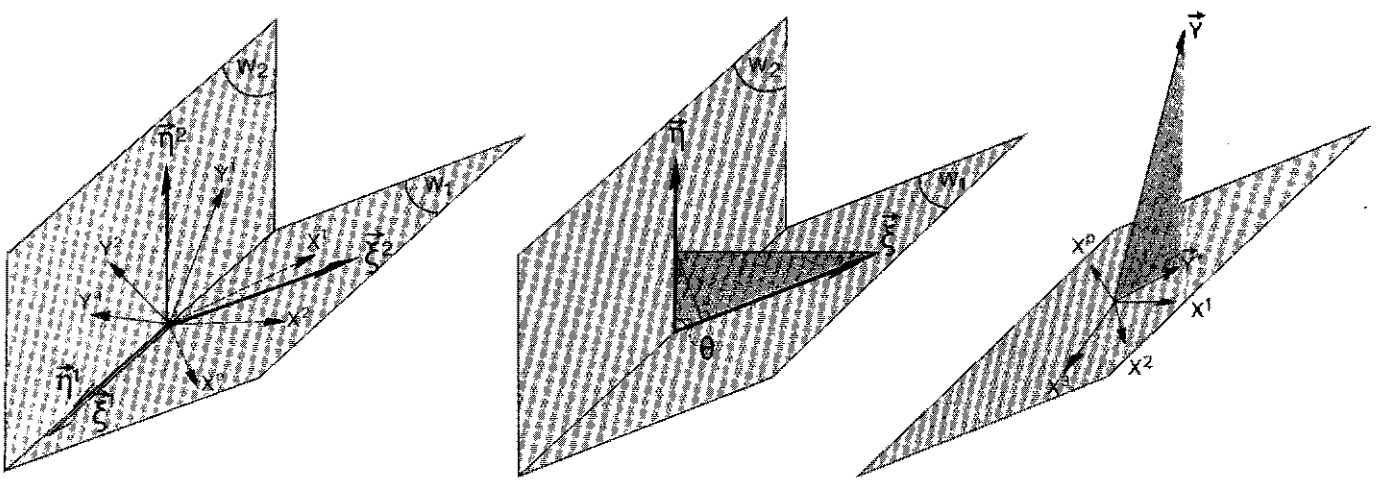
Ainsi la première composante principale de notre exemple, oppose les pays les plus développés (fort PNB, forte consommation d'électricité et fort équipement en télévision) à droite du graphique, aux pays moins développés à gauche du graphique, qui sont aussi ceux où l'agriculture pèse d'un poids important, dans la population comme dans le PIB. Les 18 pays s'échelonnent donc selon le premier axe, suivant leur niveau, général de développement, depuis le Portugal et la Grèce jusqu'aux États-Unis et à la Suède. L'origine des axes représente un pays fictif ayant pour caractéristique la moyenne de celle des 18 pays considérés. La première composante principale permettrait, si on le désirait, de définir un indice de développement économique.

La deuxième composante principale est corrélée positivement aux variables traduisant l'importance du commerce extérieur et négativement aux variables d'investissement (formation brute du capital fixe et construction de logements) : elle oppose les pays à fort taux d'investissement en bas du graphique à ceux qui investissent peu (Belgique) qui sont en même temps ceux dont l'économie est la plus tributaire des échanges avec l'extérieur. Certains lecteurs pourront être surpris de trouver le Japon en bas du graphique, parmi les pays à faible commerce extérieur : la raison est que, si les exportations (et les importations) du Japon sont importantes en valeur absolue, elles sont très faibles relativement au PNB : la part des exportations (et celle des importations) dans le PNB n'est que de 9 % pour le Japon tandis qu'elle atteint 41 % pour la Belgique (donnée de 1970).

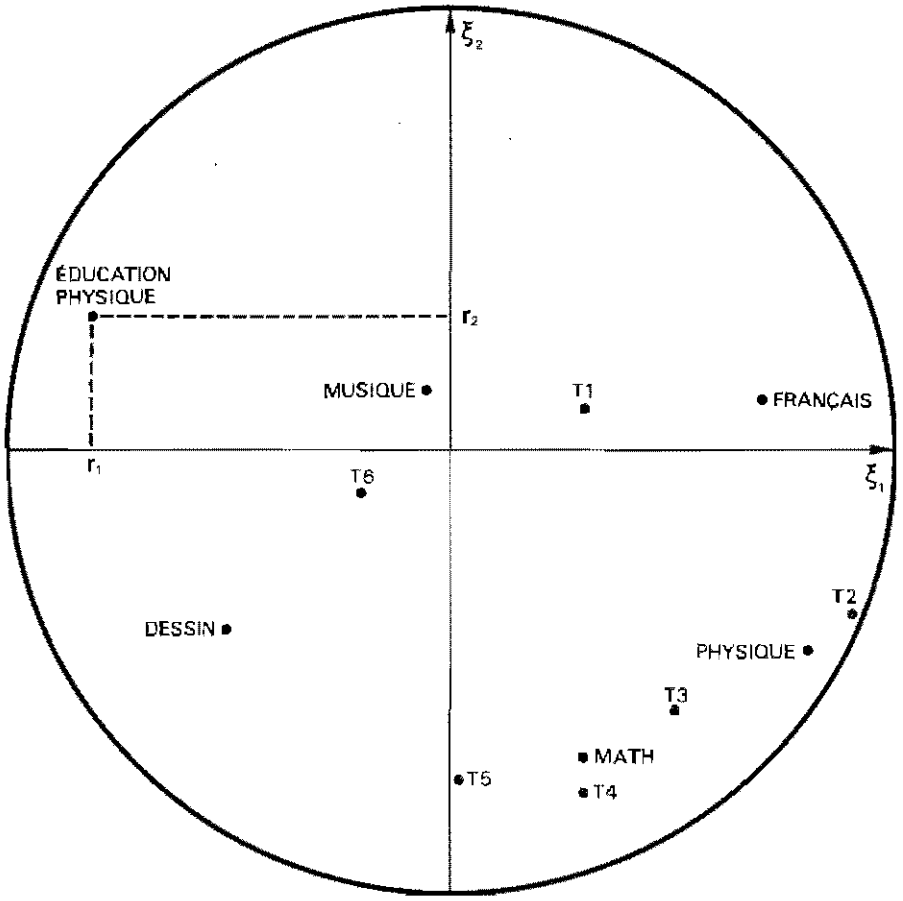
L'exemple de J. D. Carrol et M. Wish concernant les ressemblances entre consonnes, étudié au début de cet article, relève lui aussi de l'analyse en composantes principales appliquée au cas particulier où on ne connaît pas les coordonnées des individus, c'est-à-dire les caractères, mais seulement leurs distances entre eux. On parle alors d'analyse factorielle sur un tableau de distance. Ce genre de méthode permet de placer des points sur une carte connaissant simplement leurs distances mutuelles. Lorsque les données ne sont pas exactement des distances euclidiennes mais des indices de ressemblance ou de dissemblance, les méthodes utilisées sont légèrement différentes (on les qualifie d'analyses de proximités). Leur principe consiste à choisir des points dans un espace de dimension fixée (par exemple deux si l'on veut obtenir une figure plane) de telle sorte que leurs distances respectent l'ordre défini par les indices de dissemblance : ainsi, entre quatre points a, b, c, d tels que $\delta(a, b) < \delta(c, d)$ on essaiera d'obtenir $d(a, b) \leqslant d(c, d)$, où δ représente l'indice de dissemblance entre deux points, qui est la donnée de départ et d la distance géométrique entre les points figuratifs des individus.

L'analyse canonique.

Cette analyse est d'un intérêt théorique essentiel car elle englobe comme cas particulier la régression multiple déjà connue des spécialistes de la statistique, mais aussi l'analyse discriminante et l'analyse des correspondances. Le but de l'analyse canonique est d'étudier les relations existant entre deux groupes de p et q caractères quantitatifs et de comparer globalement les enseignements



8. L'ANALYSE CANONIQUE a pour but de comparer deux ensembles de caractères ($\vec{x}^1, \ldots \vec{x}^p$) et ($\vec{y}^1, \ldots \vec{y}^q$) appartenant à des individus. Pour ce faire, on détermine d'abord les caractères communs à l'intersection de deux ensembles W_1 et W_2, constitués, l'un de toutes les combinaisons des caractères ($\vec{x}^1, \ldots \vec{x}^p$) et l'autre de celles des caractères ($\vec{y}^1, \ldots \vec{y}^q$). Ces ensembles sont ici représentés par des plans. Sur la figure *a*, les caractères communs sont $\vec{\xi}^1$ et $\vec{\eta}^1$. On détermine ensuite les caractères (ou les vecteurs) $\vec{\xi}^2$ et $\vec{\eta}^2$ appartenant chacun à un ensemble et qui sont les plus proches l'un de l'autre (*voir la figure b*). Le cosinus θ de l'angle entre deux caractères mesure la corrélation entre ces deux caractères. Ainsi, le vecteur $\vec{\eta}$ de l'ensemble W_2 est celui qui se rapproche le plus de W_1 et celui qui prédit le mieux le caractère associé au vecteur $\vec{\xi}$. Il arrive que l'ensemble W_2 soit réduit à un seul élément $\vec{y}$ (*voir figure c*). Dans ce cas, le vecteur $\vec{y}^*$ est le vecteur de W_1 le mieux corrélé avec $\vec{y}$; connaissant $\vec{y}^*$, on peut prévoir, au mieux, $\vec{y}$.



9. CE CERCLE DE CORRÉLATION relie un premier ensemble de résultats dans diverses matières d'enseignement à un second ensemble constitué de tests psychotechniques T_1, T_2... T_8. Après avoir établi les deux caractères canoniques $\vec{\xi}^1$ et $\vec{\xi}^2$, on place les différents caractères en fonction de leur corrélation avec ces deux variables. Ainsi r_1 est le coefficient de corrélation entre la note en éducation physique et le caractère canonique $\vec{\xi}^1$. Les tests psychotechniques T_2, T_4, T_5 semblent bien déterminer les qualités scientifiques des individus interrogés. La proximité entre T_8 et la musique peut être illusoire car ces deux points projetés de corrélations faibles avec $\vec{\xi}^1$ et $\vec{\xi}^2$ sont peut-être en réalité loin de ce plan. Pour déterminer la valeur du test T_8, il faudrait utiliser une troisième variable. Cet exemple est fictif.

qu'ils donnent. Ainsi, sur un ensemble d'élèves, on peut chercher à comparer les résultats dans les matières scolaires (mathématiques, physique, français, dessin, musique, éducation physique) à des tests psychologiques; on désigne par $\vec{x}^1$; $\vec{x}^2$; .... $\vec{x}^p$, les caractères du premier groupe et $\vec{y}^1$; $\vec{y}^2$; ... $\vec{y}^p$ les caractères du deuxième groupe. On appelle W_1 l'ensemble des caractères que l'on peut obtenir en combinant linéairement les caractères $\vec{x}$. Ainsi $\vec{\xi}$ est un élément de W_1 si $\vec{\xi} = a_1\vec{x}^1 + a_2\vec{x}^2 + ... + a_p\vec{x}^p$, où a_1, a_2 .... a_p sont des coefficients. De même W_2 sera constitué de caractères que l'on peut obtenir en combinant les caractères $\vec{y}$. Les ensembles W_1 et W_2 sont les « potentiels de prévision » associés aux deux groupes de caractères. Ils contiennent ce que peuvent prévoir exactement les différents caractères des deux ensembles (on se limite ici à des prévisions de type linéaire). On dit, mathématiquement, que W_1 et W_2 sont deux sous-espaces de dimensions p et q de l'espace E des caractères (l'espace E a n dimensions).

Les caractères que l'on peut prévoir (c'est-à-dire obtenir) aussi bien par les $\vec{x}$ que par les $\vec{y}$ sont les éléments de l'intersection des deux ensembles, les éléments de $W_1 \cap W_2$. Plus cet espace est gros, plus les deux groupes de caractères se ressemblent car ils permettent de prévoir les mêmes phénomènes. Il est cependant fréquent que l'intersection $W_1 \cap W_2$ soit vide, c'est-à-dire ne renferme aucun caractère. Dans ce cas, pour comparer les ensembles W_1 et W_2, on cherche simultanément le caractère $\vec{\xi}^1$, combinaison linéaire des X et le caractère $\vec{\eta}^1$, combinaison linéaire des Y qui se « ressemblent » le plus, c'est-à-dire dont le carré du coefficient de corrélation linéaire est le plus grand. Ces nouveaux caractères $\vec{\xi}^1$ et $\vec{\eta}^1$ sont appelés caractères canoniques et leur coefficient de corrélation, le coefficient de corrélation canonique.

On continue ensuite le processus en déterminant un nouveau couple $\vec{\xi}^2$, $\vec{\eta}^2$ non corrélés avec les précédents (c'est-à-dire respectivement perpendiculaires à $\vec{\xi}^1$, $\vec{\eta}^1$) et au coefficient de corrélation maximum.

Pour poursuivre les analogies géométriques, fructueuses dans ce cas car on sait que deux caractères non corrélés sont perpendiculaires, on dessine les vecteurs correspondant aux caractères dans l'espace des caractères (voir figure 8). On se rappelle que dans cet espace, les cosinus des angles entre deux vecteurs mesurent les corrélations entre deux caractères. On détermine les caractères canoniques $\vec{\xi}$ et $\vec{\eta}$ par des considérations de projections orthogonales. Comme $\vec{\xi}$ est l'élément de l'ensemble W_1 faisant un angle minimum avec l'ensemble W_2, la projection de $\vec{\xi}$ sur W_2, $P_2\vec{\xi}$ est colinéaire avec le vecteur $\vec{\eta}$. Si on reprojette $P_2\vec{\xi}$ sur l'ensemble W_1, le vecteur $P_1P_2\vec{\xi}$ est colinéaire au vecteur $\vec{\xi}$. On a l'égalité $P_1P_2\vec{\xi} = \cos^2\theta\,\vec{\xi}$. Les opérateurs de projections P_1 et P_2 sont des caractéristiques des ensembles W_1 et W_2. Appliqué à n'importe quel vecteur, un projecteur le projette sur l'ensemble auquel ce projecteur est associé; le caractère canonique $\vec{\xi}$ est un vecteur propre du produit P_1P_2 (les projecteurs P sont des opérateurs linéaires, donc des matrices). De même, le caractère $\vec{\eta}$ est un vecteur propre de P_2P_1. Ainsi, les caractères $\vec{\xi}$ sont les combinaisons linéaires des x qui sont les plus proches des $\vec{y}$. Ils servent souvent de base pour représenter l'ensemble des caractères ($\vec{x}_1$, $\vec{x}_2$ ..... $\vec{x}^p$) et ($\vec{y}_1$, $\vec{y}_2$ .... $\vec{y}^q$) au moyen du « cercle des corrélations ».

Le « cercle des corrélations » est un cercle tracé dans un plan défini par deux caractères canoniques $\vec{\xi}^1$ et $\vec{\xi}^2$. Chaque caractère x ou y est représenté par un point dont l'abscisse est le coefficient de corrélation avec $\vec{\xi}^1$ et l'ordonnée le coefficient de corrélation avec $\vec{\xi}^2$. Tous les points sont à l'intérieur d'un cercle de rayon unité car la somme des carrés des corrélations d'un caractère relativement à $\vec{\xi}^1$ et $\vec{\xi}^2$ est inférieure à 1.

Si le deuxième groupe de caractères ne comprend qu'un seul élément $\vec{y}$, l'analyse canonique revient alors à chercher l'élément de l'ensemble W_1, c'est-à-dire la combinaison linéaire $a_1\vec{x}^1 + a_1\vec{x}^2 + ... + a_p\vec{x}^p = \vec{y}^*$, la plus proche, (c'est-à-dire la plus corrélée) de $\vec{y}$. Supposons que $\vec{y}$ soit le produit national brut et que parmi les $\vec{x}$ figurent les différents critères utilisés pour classifier les pays de l'OCDE. Parmi les combinaisons linéaires $\vec{y}^*$, on fera intervenir des

caractères $\vec{x}^1$ tels que le nombre de télévisions par habitant, le nombre de téléphones, le nombre de kilos de viande mangés par année, etc. qui traduisent le mieux l'état de développement d'un pays c'est-à-dire son produit national brut qu'on pourra prévoir au moyen de ces données.

On reconnaît alors que, dans ce cas particulier, l'élément $\vec{y}$ étant unique, l'analyse canonique se réduit à la régression linéaire multiple de $\vec{y}$ sur les $\vec{x}$. Cette méthode statistique est classique et se présente ici comme un cas particulier de l'analyse canonique.

Toutefois, l'analyse canonique dans le cas général où il y a plusieurs caractères $\vec{y}^1$, ..., $\vec{y}^q$, est d'emploi difficile car l'interprétation des caractères $\vec{\xi}$ est délicate. Nous allons examiner dans ce qui suit l'analyse des correspondances et l'analyse discriminante qui en sont des applications plus pratiques.

L'analyse des correspondances.

C'est une méthode d'exploration des dépendances entre caractères nominaux, c'est-à-dire des caractères prenant des modalités non numériques. Cette méthode, proposée en France par le Professeur J. P. Benzecri vers 1965 dans le cas de deux variables, a été depuis généralisée à de nombreux autres cas.

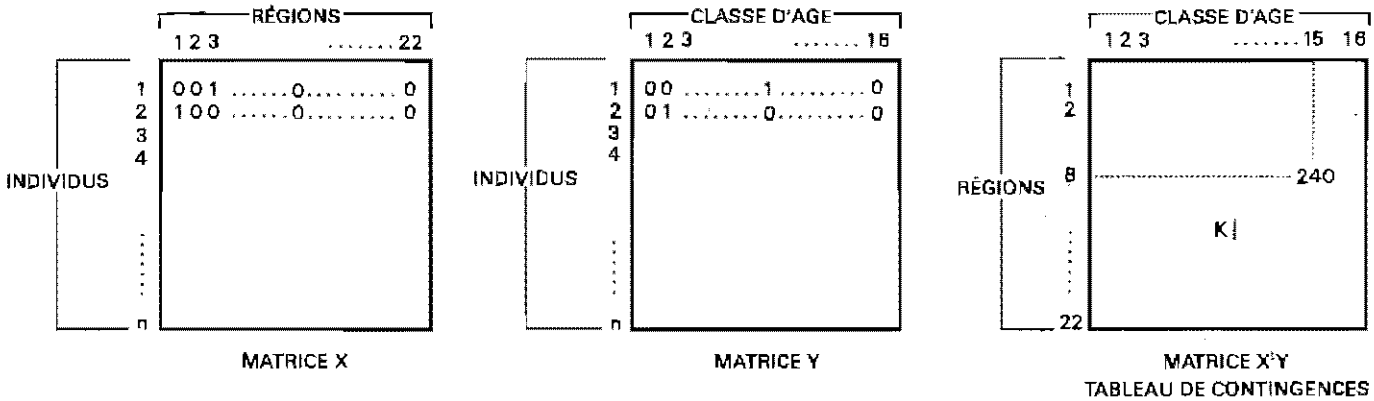
Nous exposerons la méthode dans le cas de deux caractères. Pour expliquer l'analyse des correspondances, il faut recourir à une nouvelle présentation des données : la forme disjonctive qui consiste essentiellement à donner des valeurs numériques 0 ou 1 à des caractères nominaux. Reprenons un instant l'exemple de la figure 3 avec le tableau de contingence croisant les caractères nominaux, classe d'âge et région d'habitation. Ce tableau résume toute l'information nécessaire concernant les liaisons entre les deux caractères. Une autre présentation est cependant imaginable, présentation qui fait réapparaître chacun des individus (ici les Français) : on « éclate » chaque caractère qualitatif en autant de nouveaux caractères qu'il y a de modalités. Aussi, au caractère région d'habitation, on associe 22 caractères correspondant aux 22 régions de France. Au caractère qualitatif nominal « région » nous avons associé un ensemble de caractères numériques 0 ou 1. Au caractère « classe d'âge », on associerait de la même manière un autre ensemble de caractères numériques formant un tableau de données Y de la même forme. Les caractères qualitatifs ont été ainsi mis sous forme *disjonctive*. On constate que par ce moyen, l'étude des relations entre deux caractères qualitatifs, dans ce cas la région et la classe d'âge, revient à étudier les relations entre deux ensembles de caractères numériques (les indicatrices des modalités); on peut alors appliquer l'analyse canonique à ce type de problème puisqu'elle permet de comparer deux ensembles de caractères numériques.

L'analyse des correspondances est ainsi une application de l'analyse canonique lorsque les deux ensembles de caractères représentent les modalités de deux caractères qualitatifs.

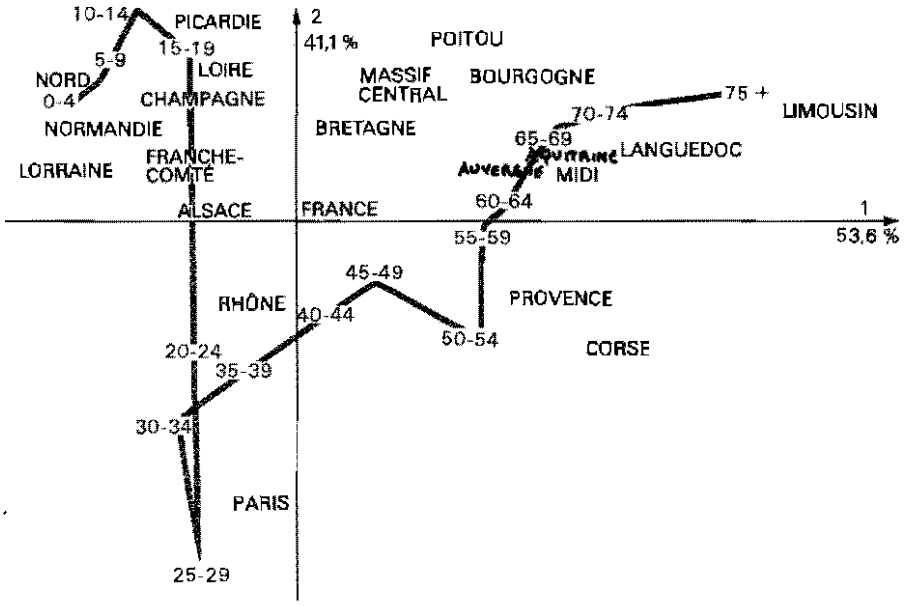
Dans la pratique, c'est-à-dire pour le programme de calcul, il n'est pas nécessaire de connaître les tableaux X et Y mais seulement le tableau de contingence contenant les k_i^j nombres d'individus de la classe d'âge j habitant la région i. La matrice des k_i^j est une réduction des deux matrices précédentes et s'obtient en effectuant à partir de X et de Y le produit matriciel X^tY où X^t est le tableau transposé de X obtenu en mettant les lignes à la place des colonnes. Nous avons emprunté cet exemple à M. Volle de l'INSEE qui en a fait l'analyse. La population étudiée est celle de la France au recensement de 1968; des caractères nominaux, la région (22 modalités) et l'âge, transformés en caractères qualitatifs par découpage en 16 classes d'âge engendrent les tableaux X et Y.

Ces caractères seront ce qu'on appelle les « indicatrices » qui ne peuvent prendre que les valeurs 0 ou 1 (1 si l'individu concerné habite la région considérée, 0 sinon). Ainsi, un individu qui habite la région n° 3 sera représenté par le vecteur (0010 .... 0), c'est-à-dire un 1 en troisième position et 21 zéros ailleurs. Le total des informations concernant la région où habite chacun des n individus recensés se présente sous la forme d'un tableau (ou matrice) X (voir figure 10).

La figure 11 donne la répartition des points représentant les différentes modalités sur le plan défini par les deux premiers caractères canoniques $\vec{\xi}^1$ et $\vec{\xi}^2$. Pour obtenir ces plans, on a comme d'habitude calculé les $\vec{\xi}^1$ et les $\vec{\xi}^2$ et projeté (selon le principe d'obtention du cercle de corrélation), les différentes variables sur le plan défini par ces deux vecteurs. Sur ce plan, chaque point représente une modalité, c'est-à-dire



10. L'ANALYSE DES CORRESPONDANCES est un exemple d'application de l'analyse canonique; elle sert à comparer deux ensembles d'individus présentant des modalités de caractères qualitatifs. Pour cette comparaison, on utilise les formes *disjonctives* pour établir facilement le tableau de contingence qui réunit la totalité des informations. Dans cet exemple, on établit une première matrice X indiquant la région (numérotée de 1 à 22) où habite chaque individu (numéroté de 1 à n). Le deuxième individu habite la première région, ce qui se traduit par 1000....00 dans la matrice X, le 1 désignant la région où habite cette personne. On a ainsi associé un caractère numérique à une modalité. De même, à chaque habitant, on a fait correspondre la classe d'âge correspondante dans la matrice Y. Pour obtenir le tableau de contingences, c'est-à-dire les répartitions des classes d'âges dans les différentes régions, on fait le produit X^tY où X^t est la matrice transposée de X; cette notation a l'avantage d'être automatique, aisément généralisable à plus de deux modalités et traitée facilement par ordinateur.



11. LES 22 RÉGIONS DE FRANCE sont ici représentées sur un graphique en même temps que les classes d'âge de la population française. Le premier axe, qui représente 53,6 % de l'information initiale contenue dans le tableau de contingence de la figure 10, oppose les régions jeunes aux régions vieilles, depuis la Lorraine jusqu'au Limousin. Le deuxième axe reprenant 41,1 % de l'information est d'interprétation plus difficile. Il montre le rôle particulier de la région parisienne dont la composition en individus proches de la trentaine est nettement plus grande que la moyenne. Un point région est d'autant plus proche d'une classe d'âge sur ce graphique que cette classe est plus représentée dans cette région. On voit apparaître des régions de structure similaire, par exemple le Poitou, la Bretagne, le Massif Central, la Bourgogne,

une région ou une classe d'âge. Le centre de cette figure 11 représente la France entière et les régions sont d'autant plus éloignées du centre que leur structure par âge diffère de la moyenne française. Pour faciliter la lecture, ces classes d'âges ont été reliées par un trait allant des plus jeunes aux plus vieilles. Le premier caractère canonique $\tilde{\xi}^1$ représenté par le premier axe oppose les classes jeunes, à gauche, aux classes vieilles à droite. Il est, dans ce cas, d'interprétation très facile : il classe les régions selon leur structure par âge. La région la plus jeune est la Lorraine et la plus vieille le Limousin, ce que l'on vérifie facilement en se reportant aux données brutes. Le deuxième caractère canonique perpendiculaire au premier oppose principalement les classes 25-29 et 30-34 ans à toutes les autres, et spécialement à la classe 10-14. Du point de vue des régions, cet axe distingue surtout la région parisienne dont la position est très particulière. Elle se distingue en effet par une forte proportion de population d'âge compris entre 25 à 34 ans. Inversement, les régions comme la Bretagne et le Centre se caractérisent par une proportion particulièrement faible de leur population dans ces classes.

Ces informations que l'on pourrait tirer d'une étude longue et approfondie de l'ensemble des données apparaissent ici immédiatement et sous une forme imagée, ce qui est un des principaux attraits de la méthode.

Ce genre de représentation graphique peut aussi mettre en évidence des phénomènes moins attendus qui eux-mêmes peuvent orienter des études ultérieures. On peut ainsi noter la position particulière de la classe 50-54 ans. Cette classe était, en 1968, une classe creuse due à la perte des naissances consécutive à la guerre de 1914-18. Le décrochage de cette classe montre que sa répartition géographique n'a pas obéi aux mêmes lois que les classes voisines et que cette classe est surreprésentée en Corse et en Provence. Diverses hypothèses peuvent alors être envisagées par le démographe mais elles sortent du domaine de l'analyse des données qui s'est contentée de faire ressortir les traits marquants du tableau de données.

On notera au passage que le fait d'avoir transformé le caractère « âge » en caractère qualitatif grâce au découpage en 16 classes apporte une information plus riche que si on avait gardé sa nature numérique : en effet, dans ce dernier cas, on se serait contenté de calculer quelques indicateurs simples mais assez pauvres comme l'âge moyen par région au lieu de garder toute la structure des données.

L'analyse des tableaux de contingence k^1 est très féconde et utile dans de nombreux domaines économiques ou sociologiques mais il arrive bien souvent, dans des enquêtes par exemple, qu'on se trouve en présence de beaucoup plus de deux caractères qualitatifs. L'analyse des correspondances se généralise aisément pour qu'on puisse traiter ces situations et constitue donc une méthode de choix pour le dépouillement d'enquêtes.

Grâce à sa souplesse d'utilisation et à la richesse de ses interprétations, l'analyse des correspondances est une des méthodes les plus utilisées actuellement; elle est au carrefour des techniques d'analyses des données et si nous avons préféré, par souci de cohérence, la présenter dans cet article à partir de l'analyse canonique, on peut également la considérer comme une analyse en composantes principales particulières où les individus seraient les lignes du tableau de contingence et les caractères, les colonnes ou vice versa. Ceci explique que l'on puisse parler, ici aussi, de pourcentage d'inertie expliquée. Dans l'exemple précédent, le pourcentage d'inertie expliquée par les deux axes du graphique était de 94,7 %, ce qui est tout à fait remarquable et indique que la structure par âge des régions françaises est presque parfaitement résumée par le graphique de la figure 11.

L'analyse discriminante.

Cette analyse est un autre exemple d'analyse canonique; elle permet d'expliquer un caractère qualitatif à l'aide d'un ensemble de caractères quantitatifs. Précisons cette méthode à l'aide d'un exemple traité par M. de Bonis (1970).

On s'intéresse à trois groupes de cinquante individus. Ces trois groupes sont constitués respectivement de sujets schizophrènes, névrosés ou normaux. Le caractère à expliquer prend donc trois modalités en fonction du groupe. Chaque sujet est soumis à un questionnaire d'anxiété. Quand il est soumis à ce questionnaire, il doit répondre à trente questions concernant des symptômes d'anxiété tels que la fatigue, les cauchemars, les crampes, etc. Les caractères ainsi mesurés sont quantitatifs, ils peuvent prendre quatre valeurs en fonction de réponses à des questions du genre : avez-vous des cauchemars? Toujours, souvent, rarement, jamais. On se pose alors les questions suivantes :

a) les questions sont-elles bonnes, c'est-à-dire permettent-elles de différencier les trois groupes?

b) sachant qu'on a affaire à un individu connu, par exemple un schizophrène, est-ce qu'on le place dans le bon groupe à l'aide des réponses aux questions?

c) quand un individu a répondu à un

certain nombre de questions, peut-on l'affecter à un de ces trois groupes, ceci avec une probabilité d'erreur minimale?

d) peut-on réduire le nombre de caractères quantitatifs, c'est-à-dire le nombre de questions de manière à ne garder que celles qui différencient le mieux les différents groupes?

Ces deux dernières opérations serviraient à éliminer les questions stupides. Par exemple : « aimez-vous le chocolat », quand il s'agit de distinguer des schizophrènes, des névrosés ou des individus normaux.

L'analyse discriminante à but descriptif, ou analyse factorielle discriminante, a pour but de répondre à la question *a;* l'analyse discriminante à but déeisionnel a pour but de répondre aux questions *b* et *c*. Les procédures pas à pas à but décisionnel et/ou descriptif permettent de répondre aux questions *a*, *b* et *d*.

Montrons quelle est l'application de l'analyse canonique à l'analyse discriminante. On appelle $\vec{x}^1$, $\vec{x}^2$, ... $\vec{x}^p$ les caractères quantitatifs. On suppose que le caractère qualitatif à expliquer détermine q groupes disjoints sur la population. Comme en analyse des correspondances, on construit q indicatrices notées $\vec{y}^1$, $\vec{y}^2$, ..., $\vec{y}^q$. Dans cet exemple $q = 3$ et le caractère $\vec{y}^2$ prend la valeur 1 sur les individus appartenant au deuxième groupe et zéro sur les autres. La question est : peut-on étudier les relations entre les q indicatrices et les p caractères quantitatifs? C'est le même problème que celui de l'analyse canonique; l'analyse factorielle discriminante revient donc à effectuer une analyse canonique entre les deux ensembles de caractères définis ci-dessus. Nous ne nous attarderons donc pas davantage sur cette méthode.

Le principe de l'analyse discriminante décisionnelle est également très simple. Il s'agit, au vu des caractéristiques quantitatives d'un individu de décider à quel groupe il appartient. Dans un premier temps, on utilise l'échantillon de départ afin d'élaborer une règle de décision, pour affecter quelqu'un à un groupe en fonction des réponses aux questions, par exemple. Concrètement, on peut calculer la distance d'un individu au centre de gravité de chaque groupe et décider de l'affecter au groupe le plus proche. Dans certains cas, on peut montrer que la distance (ou le score) d'un individu, noté x, à un groupe k peut s'écrire, à une constante près ne dépendant pas de k :

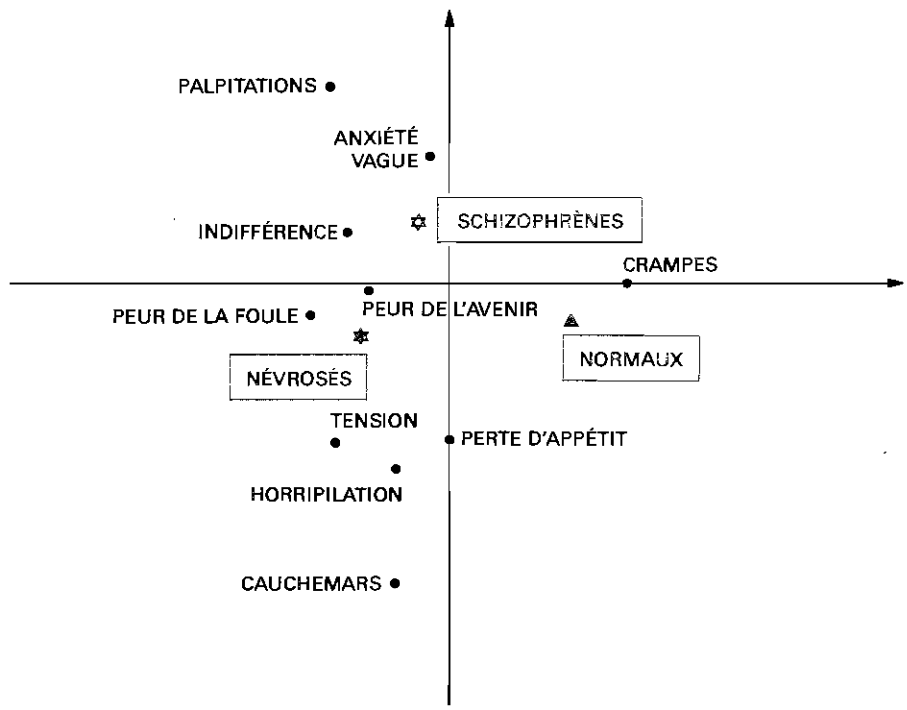
$f_k(\vec{x}) = \vec{c}_k + b_{ik}\,\vec{x}_i + \ldots + b_{pk}\,\vec{x}_p$

Les x_i sont les réponses aux différentes questions et la distance f_k les fait intervenir avec une certaine pondération. On construit k « fonctions discriminantes » de ce type. Ces mêmes fonctions servent, dans un deuxième temps, à prédire le groupe d'appartenance d'un nouvel individu. Cette règle de décision peut être validée en calculant les pourcentages d'individus bien et mal classés sur un échantillon d'origine et sur un échantillon de contrôle. D'autres règles sont suseeptibles d'être élaborées en faisant des hypothèses supplémentaires sur les lois de distribution des caractères quantitatifs sur les probabilités *a priori* d'appartenir à tel ou tel groupe et sur les coûts d'erreurs, c'est-à-dire sur les risques entraînés par l'erreur. Ces modèles sont fort utilisés dans la pratique.

Il existe enfin des procédures « pas à pas » permettant de réduire le nombre des caractères quantitatifs : on ne retient que les caractères peu corrélés entre eux et qui, globalement, différencient au mieux les divers groupes. Une telle procédure a été utilisée dans le cas des schizophrènes, des névrosés et des normaux. On a sélectionné dix questions sur les trente. Dans le plan des deux premiers caractères canoniques associés aux x^1, ... x^9, x^{10}, on obtient la représentation indiquée sur la figure 12 pour les dix questions et les trois indicatrices. Le premier axe sur ce graphique différencie les sujets normaux des autres tandis que le deuxième axe isole les schizophrènes. Cet axe oppose également les caractères somatiques (du côté des névrosés et des normaux) aux caractères non somatiques. L'analyse discriminante décisionnelle permet à partir de réponses aux dix questions sélectionnées d'affecter un nouveau sujet à l'un des trois groupes; un individu normal a 73 % d'être reconnu comme tel, un névrosé 2 % d'être eonsidéré comme normal!

Les développements récents.

La plupart des méthodes présentées ci-dessus ont été mises au point depuis fort longtemps sur le plan théorique. Avec le développement des moyens de calcul, elles ont connu un grand nombre d'utilisations. Elles sont considérées actuellement comme classiques mais leur formulation a beaucoup évolué. L'innovation la plus spectaculaire a été l'abandon des modèles probabilistes au profit des modèles géométriques. Il y a quelques années encore, la présentation de ces méthodes était fondée sur la recherche des liaisons entre variables aléatoires, ce qui supposait le choix *a priori* d'un modèle probabiliste de distribution multidimensionnelle. Ce choix était très



12. L'ANALYSE FACTORIELLE DISCRIMINANTE permet de distinguer trois groupes d'individus, schizophrèues, névrosés, et uormaux en fonction d'une batterie de dix questions. Les coordonnées des points questions sont égales à leur corrélation avec les deux caractères canoniques. La proximité de chaque groupe d'individus avec un point question est une mesure de l'intensité des réponses : les individus normaux ne sont sujets qu'aux crampes et les névrosés ont plus peur de la foule que les schizophrènes. Sur le plan somatique, les névrosés et les normaux apparaissent plus sujets aux cauchemars, aux troubles de l'appétit que ne le sont les schizophrènes qui semblent présenter des palpitations. Cet intéressant exemple d'application de l'analyse des données à la psychiatrie est due à M. De Bonis et ses collègues.

souvent arbitraire et peu réaliste. Une présentation purement géométriquc en terme de distance, de projection, de combinaison linéaire de vecteurs, élimine tout *a priori* sur les lois des phénomènes analysés. Elle permet de plus une formalisation claire et intuitive des calculs. Bien entendu, il est toujours possible de revenir à des hypothèses probabilistes, mais ce choix est fait sciemment lorsqu'une analyse purement descriptive des données nous y invite. Ce sont surtout les statisticiens français qui ont adopté cette démarche descriptive et géométrique tandis que les anglo-saxons restent plus fidèles à l'approche probabiliste et inductive. Les principales recherches actuelles consistent à appliquer ces méthodes quantitatives à des caractères qualitatifs grâce au codage optimal. De quoi s'agit-il? Un caractère qualitatif, nominal ou ordinal, peut être quantifié arbitrairement : pour cela on fait correspondre une valeur numérique à chaque modalité. Partant de cette remarque, certaines méthodes telles que l'analyse en composantes principales, l'analyse canonique, la régression multiple, peuvent être utilisées sur des caractères qualitatifs. On choisit alors une quantification optimale au sens d'un critère lié à la méthode : pourcentage de variance expliquée ou corrélation multiple.

Le développement de méthodes d'analyse de données correspond à un besoin très pratique des utilisateurs. Il s'agit de rendre disponibles des ensembles de programmes faciles à utiliser et compatibles entre eux. Des bibliothèques de programmes ont étc mises au point et sont largement diffusées. En France, de nombreuses équipes travaillent activement au développement et à l'application de l'analyse des données. Autour de J. P. Benzecri et L. Lebart, de nombreux chercheurs contribuent à diffuser et systématiser l'application de l'analyse des correspondances à tous les domaines de la science. Pour notre part, nous essayons avec M. Tenenhaus, de mettre au point des méthodes nouvelles d'analyse des données qualitatives. Sans pouvoir être exhaustif, on citera encore les travaux de E. Diday, Y. Escoufier, de J. P. Pagès et du laboratoire de H. Caussinus à l'Université de Toulouse.

Dans ce type de recherches, l'association d'un praticien (J. M. Bouroche) et d'un théoricien (G. Saporta) est féconde, le praticien apportant ses problèmes intéressants. Notre collaboration est née d'un problème relativement complexe : quels sont les bons critères permettant de découvrir les futurs bons payeurs à qui les banquiers peuvent accorder un crédit? ■

Belin

8, rue Férou - 75278 PARIS CEDEX 06

DYSLEXIE

DYSORTHOGRAPHIE

méthode pratique de rééducation de la lecture et de l'orthographe

par A. DE MEUR, P. H. NAVET

un volume broché de 317 pages....45,00 F

De nombreux enfants se trouvent handicapés dans leur progression scolaire par des perturbations qui atteignent leur aptitude à lire et à écrire correctement.
Cette méthode, fruit d'observations et d'études scientifiques sérieuses, répond à la nécessité d'aider ces enfants par des exercices nombreux, variés et progressifs selon les âges dans lesquels les troubles de la lecture et de l'orthographe sont envisagés et traités :

Orthopédie mentale - Structuration spatiale
Rééducation de la lecture
Rééducation de l'orthographe d'usage
Rééducation de l'orthographe grammaticale
Conjugaison
Confusions homonymiques.

Ce livre s'adresse particulièrement :

- **aux spécialistes de l'orthophonie,**
- **aux enseignants désireux d'aider certains élèves à surmonter un retard scolaire,**
- **aux parents qui voudront bien s'associer à ce travail de rééducation.**