Benzécri, J. P. (1984) L'analyse des données.

T.2 L'analyse des correspondances, partie A, pp 3-17

Paris : Dunod

IIA n° 1
[Principes]

LES PRINCIPES DE L'ANALYSE DES DONNÉES

Avec l'analyse des données fondée sur l'usage de l'ordinateur, c'est une nouvelle méthodologie que la statistique apporte à la science et notamment aux sciences de l'homme. On en propose ici des principes :

1er Principe. *Statistique n'est pas probabilité. Sous le nom de statistique mathématique, des auteurs (qui, je vous le dis en français, n'écrivent guère dans notre langue...) ont édifié une pompeuse discipline, riche en hypothèses qui ne sont jamais satisfaites dans la pratique. Ce n'est pas de ces auteurs qu'il faut attendre la solution de nos problèmes typologiques.*

Quant aux applications, deux questions se posent. La conception probabiliste est-elle conforme aux phénomènes étudiés ? La connaissance que nous avons de ces phénomènes permet-elle de faire des calculs ?

Considérons dans la nature trois types de lois probabilistes.

a) Les lois de symétrie : des issues qui se correspondent par les opérations d'un groupe laissant invariant l'ensemble d'un système sont également probables. Exemples : le dé a une chance sur six de tomber sur chacune de ses faces ; la bille de la roulette s'arrête avec une égale probabilité en chaque position de la circonférence ; une aiguille qu'on laisse choir tombera sur un sol horizontal avec une direction uniformément répartie en probabilité sur la circonférence (invariance par rotation autour d'un axe vertical). Ce dernier exemple est particulièrement instructif. Si le point d'où on libère l'aiguille est lui-même distribué quasi uniformément sur un plan horizontal, il en sera de même du point de chute, d'où l'on peut déduire l'assertion de Buffon : la probabilité pour que l'aiguille se pose en coupant l'interstice entre deux lattes d'un parquet dont l'intervalle est égal à sa longueur, est $(2/\pi)$. Or l'on sait d'expérience que la chute d'une aiguille sur un parquet ne permet pas de calculer π avec une approximation satisfaisante : non que la théorie de Buffon soit en défaut, mais l'irrégularité du sol (notamment) en interdit l'application rigoureuse. Notons encore que le calcul de Buffon repose sur une hypothèse de symétrie approchée : la distribution horizontale du point de départ est uniforme à l'échelle de la grandeur qui caractérise le phénomène — la longueur de l'aiguille (invariance pour le groupe des translations horizontales). Cette

hypothèse concerne les conditions initiales et relève du principe général suivant : la distribution probabiliste de l'issue est la transformée de la distribution initiale par une transition probabiliste (éventuellement une fonction : la loi de transfert du phénomène étudié).

b) Les lois ergodiques : supposons que l'objet étudié comporte un grand nombre de paramètres, qui d'ailleurs ne sont pas immédiatement observés. On peut dire qu'on étudie un point qui se meut dans un espace de configuration de dimension très élevée ; sous certaines hypothèses, toute trajectoire tend au cours du temps à remplir uniformément tout l'espace (muni d'un élément de volume convenable). L'observation ne porte pas sur l'ensemble des détails microscopiques mais sur quelques grandeurs choisies (généralement macroscopiques relatives à l'ensemble ; mais parfois aussi locales, relatives à un seul individu) ; la propriété d'uniformité qui n'est vérifiée dans l'espace de configuration lui-même que sur un temps très long, suffit à impliquer que la loi de grandeur mesurée (ou la loi conjointe des grandeurs mesurées) est avec précision celle que l'on doit avoir à la limite. L'exemple type est la théorie cinétique des gaz : de l'interaction d'innombrables molécules résultent d'une part les lois (non probabilistes) qui lient les grandeurs macroscopiques (pression, volume, température, etc.), d'autre part les distributions probabilistes des diverses fluctuations : ainsi, selon Maxwell, la vitesse d'une molécule est une variable normale tridimensionnelle. Mathématiquement, le modèle le plus général de ces multiples interactions enchevêtrées, est sans doute la théorie ergodique ; mais le théorème central limite de Laplace fournit un cadre plus simple, dans lequel on a voulu inscrire d'innombrables phénomènes. On sait que sous des hypothèses convenables, la moyenne d'une suite indéfinie de variables aléatoires dont chacune suit une loi quelconque est une variable normale. D'où la grande vogue des hypothèses de normalité, là même où il n'est nullement prouvé que soient exclusivement en jeu des causes infimes jouant additivement.

c) La mécanique quantique : outre qu'elle fait grand cas des groupes de symétrie (*a*) et des interactions de grands ensembles (*b*), la mécanique quantique a un principe probabiliste qui lui est propre : elle traite de fonctions d'ondes (sorte de champs, régis par des équations aux dérivées partielles, dont le carré de module fournit la loi de fluctuations statistiques seules observables). L'objet premier de ce cours est l'étude statistique des faits humains, sociaux ou individuels ; de la zoologie ; de la botanique ; voire de la géologie, non des particules élémentaires. Pourtant comme le dépouillement des clichés de chambre à bulle, produit sans doute les plus gigantesques tableaux de données que des hommes aient rassemblés, nous devions évoquer ces recherches fameuses qui semblent attendre une méthode d'analyse statistique. Mais nous n'y reviendrons pas dans cette leçon.

Dans les données soumises au statisticien, les symétries géométriques simples ne jouent pas un rôle important ; nous avons dit que les phénomènes

quantiques ne sont pas l'objet de notre étude. Il reste donc la possibilité que les processus complexes étudiés soient régis par ce que nous avons appelé des lois ergodiques, exprimant sous forme probabiliste l'équilibre de multiples causes. Le domaine des lois ergodiques est sans doute vaste : il s'en faut toutefois de beaucoup qu'on puisse concevoir l'existence de probabilités objectives partout où l'on fait usage de modèles probabilistes. De quoi nous donnerons deux raisons.

D'une part, on postule souvent qu'ont une probabilité d'être vraie des assertions mal définies. Exemple : probabilité qu'il fasse beau demain ; est-ce la probabilité qu'il fasse beau un 14 juillet (j'écris le 13) en telle ville ? ou la probabilité qu'au lendemain d'un jour d'été ensoleillé tel qu'aujourd'hui il n'éclate point d'orage ? à la limite, si l'on précise parfaitement les conditions d'aujourd'hui, demain est lui aussi rigoureusement fixé et il n'y a plus doute mais certitude. On répète souvent en citant H. Poincarré que la probabilité est la mesure de notre ignorance, entendant par là que c'est l'imprécision de nos connaissances qui définit le domaine objectif sur lequel on ne connaîtra que des probabilités (e.g. on renonce à suivre chaque molécule : on se contente de connaître la loi de Maxwell qui donne la distribution en probabilité des vitesses (cf. *c*)). Mais le propre de l'ignorance est d'être indéfinie ; non seulement on n'a pas de résultats précis, mais aussi on ignore les bornes du connu et de l'inconnu. Dans la plupart des problèmes humains (e.g. probabilité que tel homme politique soit élu), probabilité ne signifie que degré de présomption ; le modèle probabiliste mathématique (cf. 1°) diffère certes du réel par des inexactitudes quantitatives (tel paramètre est erroné) et des simplifications (on ne fait aucun cas de telle cause mineure) ; il en diffère plus essentiellement par la conception qualitative de ce qu'est le possible. Il est d'ailleurs de règle qu'en mathématique l'inventaire des possibles soit vite fait : point n'est besoin de parcourir une droite pour savoir qu'elle est homogène. Et c'est pourquoi la théorie des ensembles s'accommode d'une sorte d'infini en acte. Dans la nature il en va autrement. Un domaine potentiel n'a d'existence que s'il est effectivement parcouru ; en ce sens la nature a horreur du vide ; certes le possible, le potentiel se distingue de l'actuel ; mais le potentiel et l'actuel sont tous deux réels. L'esprit scientifique formaliste contemporain pèche à ignorer cela.

D'autre part, lors même qu'un domaine naturel est défini sans ambiguïté (e.g. probabilité pour qu'un artisan cordonnier laisse l'échoppe à son fils), les conditions peuvent se modifier trop vite pour que l'équilibre de causes infimes, présupposé par les lois ergodiques, soit jamais atteint. Peut-on faire un modèle théorique de l'économie française alors que, disons, le tiers des variations de prix est acquis au cours de crises qui, telles celle de 1968, ne relèvent pas d'un type général ?

Nous venons de mettre en doute l'existence d'un ordre des choses probabiliste ; il faut encore examiner si, là même où cet ordre a pu objectivement s'instaurer, nous le connaissons assez bien pour calculer tout ce que nous

désirons. Les données probabilistes nous sont connues par des observations individuelles, en nombre fini, d'après lesquelles on calcule notamment des fréquences. Souvent la rareté des observations n'est que la conséquence nécessaire de la rareté des phénomènes : il n'y a pas d'équilibre objectif, donc, nous l'avons dit, pas d'ordre probabiliste. Mais parfois le manque d'information n'a pour cause que le coût des mesures : notre connaissance est beaucoup moins ordonnée que le phénomène ne l'est lui-même. Par exemple l'emploi des noms et des adjectifs de couleur par Honoré de Balzac, n'est certes pas régi jusqu'à la dixième décimale par des probabilités ; mais pour l'instant, à notre connaissance, toute statistique manque ; on pourrait donc, en s'y appliquant, découvrir des lois inconnues. En pareil cas, on a une évaluation acceptable des erreurs commises en supposant que la nature est rigoureusement conforme à un modèle probabiliste, mais que la connaissance qu'on en a est perturbée par les fluctuations d'échantillonnage (auxquelles s'ajoutent, pour les grandeurs continues, des erreurs de mesure). On se gardera toutefois de fonder les calculs d'erreurs sur des hypothèses restrictives elles-mêmes plus sujettes à caution que les faits de structure que l'analyse des données aura fait apparaître. Lors même que l'existence d'un ordre des choses probabiliste est avérée, on ne dispose généralement pas d'un modèle probabiliste complet (*) : tel est notre 2e principe.

Concluons donc : les bases de l'analyse statistique sont plutôt algébriques ou géométriques (mais quand sont en jeu de multiples dimensions (cf. 3°), les vues géométriques se fondent par les calculs de l'algèbre), que probabilistes ; il vaut mieux parler de moyenne puis d'axes principaux d'inertie, etc... déterminés sur un ensemble fini de données actuelles, que d'espérance mathématique, etc... relatives à un univers potentiel indéfini. Mais les conceptions probabilistes suggèrent des opérations algébriques et permettent parfois d'en évaluer la portée.

2e Principe. *Le modèle doit suivre les données, non l'inverse. Autre trait fâcheux des mathématiques appliquées aux sciences humaines : l'abondance de modèles, forgés* a priori *puis confrontés aux données par ce qu'on appelle des « tests ». Et tantôt le « test » sert à justifier un modèle où il y a plus de paramètres à ajuster que l'on n'a déterminé de données. Tantôt au contraire il sert à rejeter sévèrement comme invalides les plus judicieuses remarques de l'expérimentateur. Mais ce dont nous avons besoin c'est d'une méthode rigoureuse qui extraie des structures à partir des données.*

Quand elles s'efforcent de prendre expression mathématique, les sciences humaines, psychologie, linguistique, sociologie, histoire..., recourent généralement à des modèles. Un modèle est, en bref, un système de formules qui permet de calculer, en fonction de variables inobservables, les quantités observées : ainsi, en psychologie, on postulera l'existence d'une variable de

(*) Modèle qui est à la base de la statistique bayésienne, et aussi de la théorie de l'information (cf. [Inf. Tab.] T I B, n° 5, § 1.4).

prédisposition et d'une variable d'accoutumance, et on calculera en fonction de ces variables la diminution du taux des erreurs commises par un sujet dans un apprentissage.

Le terme de variable inobservable est, en un certain sens, relatif à l'état de la science. Plus d'une grandeur physique, d'abord conjecturée, telle l'énergie ou la charge électrique, est maintenant reconnue pour exister, et se mesure comme l'espace ou le temps. Mais les sciences humaines tâtonnent encore pour établir des lois rigoureuses ; et tandis qu'en astronomie quelques axiomes très simples régissent le mouvement des systèmes les plus complexes, un psychologue put rarement se vanter que l'étude exhaustive de phénomènes élémentaires lui permît de prédire avec précision l'évolution d'un cas complexe : la psychologie mathématique attend son Leverrier lâchant la plume pour pointer du doigt vers Neptune (et peut-être les faits psychologiques, n'étant pas circonscrits, ne se prêteront-ils jamais à des calculs, fussent-ils astronomiques ?). Aussi la portée d'un modèle dépasse-t-elle rarement le champ déterminé d'observations pour lequel il fut conçu.

Le modèle fait, deux éventualités peuvent donc se présenter. Ou bien il est impossible de donner aux variables des valeurs telles que les résultats des formules s'accordent, dans les limites définies par la précision des mesures et la statistique des erreurs, avec les observations recueillies : et alors, il est clair que le modèle est invalidé. Ou bien il n'y a pas de désaccord, mais cela ne signifie pas pour autant qu'on puisse recevoir le modèle comme théorie vraie. Souvent, en effet, le modèle comporte tant de variables inobservables, et les expériences sont si difficiles (ou les observations si rares), que l'on n'a aucune peine à rendre compte de celles-ci par celles-là. Il n'est pas surprenant qu'on puisse engendrer dix nombres observés, au moyen d'une formule à dix paramètres ! Certes les chercheurs compétents et honnêtes évitent de se placer dans un si mauvais cas... Mais, d'une part beaucoup de spécialistes des sciences humaines ont sur les mathématiques des vues si brumeuses qu'ils peuvent en toute probité s'égarer (tandis que des statisticiens proposent aux expérimentateurs d'élégants mais peu réels édifices probabilistes, il se trouve encore des praticiens qui définissent une moyenne d'après trois cas, ou affirment qu'un fait qui s'est produit 7 fois est significativement moins probable qu'un autre qui est apparu 9 fois ; la juste mesure restant encore à trouver). D'autre part, il est impossible d'éliminer d'un modèle les conceptions *a priori* qui ont permis au chercheur d'en imaginer les formules ; formules certes confrontées avec l'expérience, mais seulement pour en ajuster les paramètres aux mesures faites. Ainsi le modèle permet au mieux de prédire, non de comprendre.

Quant à la valeur de prédiction des modèles, on prendra garde qu'elle est souvent réduite à néant par une erreur très répandue : l'emploi, en vue de prévisions, de modèles dont les paramètres n'ont été calculés que sur un seul état instantané.

Prenons un exemple : on cherche comment avec l'évolution du pouvoir

d'achat va progresser la vente des meubles. Pour cela on recensera dans 100 agglomérations, d'une part la population répartie en une dizaine de catégories socioprofessionnelles de revenu à peu près homogène, d'autre part les ventes de quelque cinq types de meubles. On postule un modèle linéaire (ou autre...) :

$$V_i = \text{vente du meuble } i = \Sigma_j\, a_{ij}\, E_j$$

(où E_j désigne l'effectif de la catégorie socioprofessionnelle j). On choisit les a_{ij} pour que les équations du modèle soient approximativement satisfaites dans les 100 agglomérations (éventuellement, on peut perfectionner le modèle, en supposant que les a_{ij} sont fonction de la population totale de l'agglomération, etc.). On se croit alors autorisé à prévoir les variations de vente ΔV_i en fonction des variations ΔE_j (recensées ou prévues d'autre part) des effectifs des catégories socioprofessionnelles E_j ; on pose :

$$\Delta V_i = \Sigma_j\, a_{ij}\, \Delta E_j\,.$$

Il y a là, selon nous, une grave faute de méthode. A supposer que le modèle utilisé explique la situation de la vente des meubles à l'instant t, ce qu'on aura tant bien que mal vérifié sur les données, rien ne permet d'en déduire qu'il rende compte avec un égal bonheur des vitesses d'évolution. En termes mathématiques disons ceci : de ce que le vecteur des ventes $V(t)$ est approximativement à l'instant t une fonction Φ du vecteur social $E(t)$ on conclura peut-être légitimement que la même relation fonctionnelle est encore vérifiée à l'instant $t + \Delta t$. Mais il est très aventureux de postuler que l'on a :

$$\partial V_i(t)/\partial t \;\#\; \Sigma_j(\partial\Phi/\partial E_j)\,(\partial E_j/\partial t)\,;$$

car des erreurs, admissibles dans l'estimation des états $V(t)$ et $V(t + \Delta t)$, peuvent être du même ordre que la variation même $(V(t + \Delta t) - V(t))$; en sorte que $\Phi(E(t + \Delta t)) - \Phi(E(t))$ ne fournit pas *a priori* une approximation intéressante pour $V(t + \Delta t) - V(t)$.

Et en termes économiques nous ajouterons que les changements dans le marché sont d'ordinaire dus à des causes étroitement localisées mais très dynamiques : ces causes ne pèsent guère dans un état instantané, mais elles contrôlent la vitesse, voire l'accélération. La création d'un nouveau type d'habitation, d'une nouvelle sous-classe sociale ; l'ouverture à un type de consommation, d'une classe dont le revenu a atteint lentement le niveau critique, etc., ne se laissent pas voir sur une statistique synchronique d'ensemble. Il faut au contraire les chercher, et souvent avec plus d'intuition que d'esprit de système, dans des variations de grande vitesse mais de faible masse. En tout cas, pour prédire des changements, mieux vaut observer des changements ; plutôt qu'une image globale détaillée mais immobile, considérer quelques faits pertinents : des mouvements.

Quant à notre aspiration à comprendre (non seulement à prédire), certains philosophes nominalistes prétendront que, les causes profondes étant inaccessibles, ou peut-être n'existant même pas, on ne doit espérer faire mieux que d'enfermer quelques faits dans le réseau d'un système de noms et de chiffres ; que c'est s'adonner à la métaphysique que de prétendre à plus. Mais sans préjuger de la solution qu'il convient de donner au problème de l'être, on peut affirmer que la rigueur des sciences positives ne s'accommode pas d'un nominalisme de cette espèce. Que si la notion de cause nous paraît incertaine, on ne peut la rejeter ; et c'est justement à l'observation qu'il faut recourir pour la préciser et la fonder aussi honnêtement que possible. Encore faut-il que le champ d'observation soit assez ample : tel est l'objet du 3e principe.

3e Principe. *Il convient de traiter simultanément des informations concernant le plus grand nombre possible de dimensions. Du coup, le problème de la validité (du « test ») — troublant parfois, avouons-le — passe au second plan. Nul ne sait si l'inégalité* $0{,}5 \neq 0{,}7$ *peut dans tel cas pratique être considérée comme un résultat empirique sûr, ou si ce n'est qu'un jeu de hasard. Tandis que de trouver dans un espace de dimension 2 cinquante points approximativement rangés sur un cercle est sûrement une découverte (à moins que la méthode de calcul ne soit une duperie* !).

Il est commun de distinguer entre l'*observation* pure et simple des faits tels qu'ils se produisent d'eux-mêmes et l'*expérimentation*, ou observation de séries de faits provoquées par le savant pour décider de questions qu'il s'est posées, pour accepter ou rejeter des hypothèses. Pendant trois siècles, depuis Bacon jusqu'à nos jours, la méthode expérimentale a permis d'établir de façon sûre un grand nombre de lois de la nature exprimées par des relations simples, logiques ou mathématiques. Mais il semble que depuis cinquante ans la science ne soit plus si assurée de cette tactique, et cherche de nouvelles voies.

D'abord, en mécanique quantique, il est apparu impossible de déterminer par des mesures simultanées tout ce qu'on jugerait *a priori* désirable de connaître sur un même objet élémentaire. La perturbation apportée par l'observateur dans ce qu'il observe semble n'être plus réductible à l'infini par des précautions expérimentales judicieuses : elle est bornée inférieurement par les relations d'incertitude de Heisenberg. Nous ne croyons certes pas qu'il en faille conclure que la réalité extérieure soit inconnaissable ; mais d'une part cette réalité, quand elle atteint un certain degré de finesse, n'est pas indifférente à l'appareil qui nous met en relation avec elle ; d'autre part, en prétendant déterminer un certain nombre de coordonnées (au sens le plus général du terme) on postule *a priori* un ordre du réel qui est souvent plus hypothétique que les formules quelles qu'elles puissent être reliant ces coordonnées mêmes. Plus encore que la position ou le mouvement, c'est l'espace qui est en question ; l'espace, cadre que la pensée ne doit point postuler comme vide, mais dont la découverte est indissoluble de celle de ce qui l'emplit.

Cependant, en statistique, les méthodes de Sir R. Fisher, tout en respectant la notion même d'expérimentation, en ont changé le rôle. Montrant qu'il était plus économique de recueillir simultanément des informations relatives à plusieurs questions, le créateur du Plan d'expérience, a bousculé les règles

baconiennes. A une progression claire et distincte, de question en question, s'est substituée une tactique, à la fois rigoureuse et complexe, qui vise à avancer, en quelque sorte, simultanément sur un front.

Enfin l'avènement des ordinateurs (cf. 4°) permet aujourd'hui de confronter et de mettre en ordre des ensembles immenses de données, naguère vouées au feu ou à la poussière. L'on aborde ainsi des problèmes — milieux humains, écologie des plantes ou des animaux... — où, comme en mécanique quantique encore que pour d'autres raisons, expérimenter ce peut être instaurer un système nouveau, autre que celui qu'on se proposait pour objet. L'examen exhaustif de toutes les éventualités, condition nécessaire à la démonstration expérimentale d'une hypothèse, n'est d'ailleurs plus possible quand l'esprit a peine à embrasser l'ensemble des variables. Avant de prétendre retrouver une hiérarchie de causes, il convient de reconnaître suivant quels axes se rangent les masses. Avec l'analyse des données, c'est l'observation (préalable toujours nécessaire à l'expérimentation) qui apparaît présentement comme la méthode principale de nombreuses disciplines.

Pour voir ce que peut la statistique multidimensionnelle dans l'étude de phénomènes et de leurs causes considérons un exemple d'un paralogisme assez commun ; où l'argumentation statistique, fondée sur l'examen hâtif de deux ou trois fréquences, apparaissant comme une forme mathématique de l'imposture, offre au public l'occasion de vilipender les mathématiques. « La proportion des gens qui lisent est plus élevée parmi les spectateurs de la télévision que parmi les non-spectateurs : l'image entrevue donne le désir d'approfondir », conclut une revue acquise aux passe-temps nouveaux, mais qui se targue de culture ! Il serait assurément plus juste de dire, nous le sentons bien, que, puisque seuls quelques seigneurs de l'esprit, dont la plupart lisent beaucoup (faute sans doute de s'être élevés assez haut pour que la méditation suffise à les nourrir), refusent la lucarne qu'ils pourraient s'offrir, ce sont les mêmes hommes qui, nantis de quelque fortune, achètent à la fois les livres et les récepteurs. Le seul calcul de quatre fréquences ne permet pas de s'élever aux causes. En effet, prenons un exemple extrême : dire que tous les gens qui ont A ont aussi B (tandis que certains ont B sans avoir A), s'écrit certes dans le symbolisme de la logique formelle $A \Rightarrow B$; mais il faut bien se garder de voir dans cette formule une relation causale ; car elle peut s'interpréter aussi bien comme : *avoir* B est une condition nécessaire de *avoir* A (B entre dans la cause de A), ou comme : *avoir* A est une condition suffisante de *avoir* B (A cause B) ; et il est vraisemblable que, en général, sera en cause une troisième propriété C (ou plutôt un complexe de propriétés) qui, à la fois suffit à B et est nécessaire à A (C cause B, et entre dans la cause de A).

Comment donc fonder sûrement la notion de cause ? Dans certains cas la succession des faits est un indice : *post hoc, ergo propter hoc*... cependant, d'une part on rencontre ici la distinction, philosophique elle aussi et non acceptée par tous, entre cause efficiente (l'agent qui produit et précède l'action) et cause finale (le résultat qui est aussi le but et donc précède, dans l'intention

du sujet, l'acte auquel il succède) ; d'autre part dans beaucoup de domaines, les faits forment un complexe qui s'est noué depuis si longtemps que le chercheur n'en peut observer le développement temporel. Notre principale méthode devra donc être, même en histoire, l'étude synchronique, la confrontation d'un grand nombre d'observations prélevées à peu près simultanément. Et c'est bien ainsi que les sciences ont procédé depuis des siècles : mais tandis que jadis la synthèse ne se pouvait faire que dans la faculté cognitive d'un seul homme, possédant dans sa mémoire l'impression d'une multitude de faits et qu'elle était donc implicite, la synthèse, aujourd'hui, peut être faite par le calcul. Non que, avouons-le, le progrès des théories statistiques et mathématiques nous rende supérieurs à nos aînés, mais nous disposons d'outils de calcul qui effectuent en un éclair des opérations auxquelles il y a un quart de siècle, il n'eût pas été raisonnable de songer : c'est le 4e principe que nous énoncerons ci-dessous.

C'est pourquoi la comparaison systématique des variables deux à deux, faite ordinairement sur de petits tableaux obtenus par tris croisés (un tableau donnant, e.g., à l'intersection de la ligne 2 et de la colonne 3 le nombre de sujets ayant fourni à une question A la réponse numérotée 2 et à une question B la réponse numérotée 3) nous paraît devoir être abandonnée. Car d'une part ces comparaisons binaires sont peu démonstratives ; et d'autre part elles sont fastidieuses et bien moins suggestives que l'analyse simultanée de l'ensemble des données (principalement par analyse factorielle du grand tableau de correspondance obtenu en juxtaposant tous les petits tableaux binaires : cf. [Prat. Corr.] T II A, n° 2, §§ 1.4 et 1.5, *in fine*). Tout au plus recourra-t-on à la comparaison binaire pour confirmer un rapprochement apparu sur un axe de rang élevé (donc moins sûr que les premiers) issu de l'analyse factorielle (cf. e.g. [Peurs] T I C, n° 13, § 6.4, note infrapaginaire).

Si l'on voit aujourd'hui sortir des imprimantes de longues listes de ces petits tableaux, c'est que trop de chercheurs, notamment dans les sciences humaines, n'ont pas accordé leur méthode à la puissance des nouveaux outils de calcul (cf. *infra*, n° 5) : ils sont semblables à un ingénieur qui, pour en bâtir un pont, dessinerait des blocs d'acier ayant la forme de pierres de taille. Parmi ceux mêmes qui ont adopté l'analyse des données multidimensionnelles, beaucoup ne peuvent faire taire en eux la logorrhée d'hypothèse qui est tout ce qu'engendre d'ordinaire, avant une puissante synthèse, le choc de peu de faits. Le calcul des concepts incertains n'est-il pas au philosophe ce qu'est au mathématicien le verbiage des modèles (cf. 2e) : le jeu d'un marteau sans maître ? Il suffit de profiter de la clarté fugitive des hypothèses pour récolter des données qui parleront ensuite. Qui s'épuise à échafauder des mots sur d'autres axes que ceux des faits (cf. [Prat. Corr.] T II A, n° 2, § 3.3 ; et [Peurs] T I C, n° 12, § 4, sur les rapports du langage et de l'analyse des données) est semblable à un ingénieur qui, tout en donnant au pont une charpente d'acier, demanderait encore aux pierres de tenir par elles-mêmes.

Mais, revenons, pour conclure, à l'exemple évoqué plus haut : supposons

qu'au lieu de caractériser un individu par deux propriétés « lit ou ne lit pas de livres », et : « regarde ou ne regarde pas la télévision », on recueille une suite, un vecteur d'observations : e.g. loyer payé, profession, âge... ; ou mieux : dépenses annuelles en produits alimentaires, en chauffage et éclairage, en transports, en objets de luxe, etc... ; on pourra à l'aide de ces observations faire une typologie de l'ensemble des sujets étudiés. A l'intérieur de chaque classe (dans chaque région de l'espace typologique) on observera alors à nouveau l'indice lecture et l'indice télé ; on verra que l'ensemble de types s'organise d'abord suivant un axe où s'opposent, e.g. les riches aux pauvres ; puis apparaît un deuxième axe qui opposera les travailleurs de l'industrie aux artisans et aux savants, etc... Sur le premier axe, la télé et la lecture progresseront l'une avec l'autre au fur et à mesure que s'élèvera le niveau de fortune ; tandis que le deuxième axe les séparera.

Voilà du moins ce que suggère le sens commun — c'est-à-dire la synthèse rationnelle, intérieure à nous. Mais seule l'expérience, ici une expérience ample et difficile ! puis le calcul statistique permettraient de voir, sans s'asservir à des idées *a priori*, ce qu'il en est en réalité, quelle est la hiérarchie des faits positifs et donc quel en est le système de causes.

4e Principe. *Pour l'analyse des faits complexes et notamment de faits sociaux, l'ordinateur est indispensable. Principe évidemment vrai... mais qu'en eussent pensé nos pères les Gaulois il y a* 15 *ans* ?

Notre génération a vu paraître de nouveaux outils et de nouveaux noms ; ceux-ci sauf exceptions, superflus et grotesques ; ceux-là souvent puissants, mais ils sont redoutables. Ne regardons ici qu'un outil : l'ordinateur, et un art : l'informatique ; efforçons-nous de rouvrir sur eux l'œil toujours lumineux de la sagesse antique.

Considérons le travail d'un modeleur : à une matière, l'argile, il confère une forme (celle d'un pot ou celle de Socrate) par une énergie qui est celle de sa main. Traitement de la matière même de l'objet à produire, élaboration de la forme, effort mécanique (ici assez réduit) sont comme un seul acte d'un seul artisan.

Il s'en faut de beaucoup que ces trois éléments, matière, travail et forme soient toujours aussi intimement unis. Ainsi le médailleur contemporain élabore la forme (de Socrate, non d'un pot) sur un gros flanc de terre à modeler qui cède sous un travail minime. Cette forme est ensuite réduite et transportée par une sorte de pantographe et devient celle d'un coin qui a la taille réelle des médailles. Enfin une puissante machine frappe les disques de bronze.

Déplorons ce que l'art a perdu en abandonnant la taille directe : matière et forme ne s'unissent parfaitement que dans la main. Qui conduit une pelleteuse sentira-t-il jamais la terre ? La société moderne se divise en pièces ; la science s'y résout en techniques infimes. Tout l'homme et toute la nature terrestre tiennent dans un chant d'Homère : quel aveugle inspiré retrouvera l'épopée dans la gangue de nos institutions enchevêtrées ? qui répondra en

l'an 2000 au panorama que, voici 70 ans, David Hilbert offrait aux 200 redingotes de la mathématique mondiale assemblée à Paris.

Dissocier ce qui était uni ; et rendre interchangeable, avec mille autres chacun des maillons de la chaîne qu'on a rompue ; ainsi croissent les forces de production. Le Siècle passé a vu la puissance motrice se séparer de la forme et de la matière (le même moteur — ou presque — moud, éclaire, lime, véhicule...). Le Siècle présent voit la forme pure soumise en quelques secondes à des milliards d'opérations — certes élémentaires — au sein du milieu de culture universel qu'est l'ordinateur. Ce nouvel outil conçu d'abord pour le bureau de calcul, va aussi sûrement que l'a fait le moteur, prendre sa belle part de tous les métiers.

Prenons pour exemple l'imprimerie. On composait naguère à la main en lettres de plomb les lignes et les formules de nos livres. Depuis la fin du siècle dernier des machines existent qui, commandées par clavier, fondent à la demande, des suites de lettres (soudées en lignes, avec la linotype ; ou séparées, avec la monotype). Sont maintenant en usage des composeuses photographiques qui délivrent, sur film, des suites de signes photographiés d'après un réservoir de matrices (ou modèles) : avec ce procédé, le poids du métal n'intervient plus dans la composition. Toutefois, pour servir toutes les sciences et toutes les langues, l'Imprimerie Nationale de France doit réunir en une foule d'ateliers de pesants trésors. Qu'en sera-t-il demain ? Qu'en peut-il être dès aujourd'hui ? Par un clavier dactylographique, on appelle, de la mémoire d'un ordinateur, un petit programme qui commande la formation d'une lettre sur un écran fluorescent (semblable à celui que la télévision a vulgarisé, mais, on le verra, perfectionné). Il est facile de concevoir qu'un ordre issu du clavier choisisse toute taille de lettre, tout corps, tout alphabet. Dans la mémoire de l'ordinateur, un signe tient peu de place : point n'est besoin de graver un moule, de fondre ou de frapper du plomb ni même de conserver une matrice en noir sur blanc : un programme de quelques instructions suffit à créer un signe qui désormais sera toujours disponible. A l'imprimerie contemporaine, où les informations circulent et se transforment sur des supports de poids divers et sont traitées par des machines plus pesantes encore, se substitueront des programmes inscrits sur des supports universels (bandes, disques...) traités avec le secours de l'ordinateur et de l'écran cathodique, outils universels, eux aussi.

Autre exemple, la conception et l'usinage d'une pièce. Le dessin y joue aujourd'hui un rôle essentiel ; mais du calcul de résistance au dessin, de même que du dessin à la machine l'information ne peut passer que par l'homme. Calcul de résistance, dessin (c'est-à-dire calcul de cotes et éventuellement projection de vues sur un écran cathodique), et commande des machines (par signaux électromagnétiques) tendent à se faire ensemble dans l'ordinateur, qui exécutant le programme de l'ingénieur produira un enregistrement sur bande des ordres d'usinage. Comme celle du texte à imprimer, la forme de la pièce sera entièrement conçue dans ce lieu idéal sans matière ni énergie

qu'est l'ordinateur. Et point n'est besoin de multiplier les exemples pour affirmer que non seulement les nombres, objets des calculs mathématiques usuels, et le contenu des fichiers et archives (ce que l'on appelle information non numérique), mais tout ce qui (en quelque matière), est forme, se pétrira désormais par ce qu'avec un rare bonheur, on a nommé informatique :

Pétrir les formes, et écraser l'homme ? Assurément la part potentielle de l'informatique est partout la part du lion. Hors du traitement des formes qu'y a-t-il sinon des spécialités de portée limitée assurant l'application de la forme à la matière ?

L'avènement de l'informatique est certes implacable. Aux Etats-Unis d'Amérique, où des prémisses aux conclusions, les forces de production se propagent d'ordinaire avec la vitesse de la lumière, toutes les corporations sont touchées : mais, fait nouveau, au lieu d'exulter sur la voie du progrès, elles résistent... une lutte sournoise oppose le dessinateur, le bibliothécaire, le comptable... au programmeur qui l'interroge : « They want to sit by my sides and look at me and make a program out of my job ». On conçoit qu'ainsi saboté, le système de traitement de l'information ne soit pas la merveille qu'on nous vante parfois... En Europe occidentale, la société, sans trop d'aigreur, résiste de sa propre masse qui est grande. Notre France est dit-on le pays de Descartes. Elle est aussi le pays de la douceur de vivre ; et Descartes lui-même dut se cacher aux confins des Gaules pour discourir à loisir selon sa méthode. Quelques-uns s'enthousiasment. D'autres disent : ce n'est qu'une vogue ; et s'étonnant de voir instruire des centaines de programmeurs, ils prédisent que l'informatique n'offrira bientôt plus d'emplois. Ne voit-on pas bouleversée l'électronique, qui il y a un quart de siècle à peine fit une entrée triomphale ? L'argument vaut qu'on s'y arrête. L'électronique, telle qu'elle fleurit dans les années 50, est l'art de traiter par des circuits (dont les éléments les plus nobles étaient lampes et deviennent transistors) les informations propres à un système particulier : gouvernail, tour, laminoir, colonne à distillation. En ce qu'elle traite des informations traduites en impulsions électromagnétiques, l'informatique paraît comprise dans l'électronique. Mais l'informatique a dévoré l'électronique en ce que la production des ordinateurs universels se substitue à celle des circuits particuliers. Il y a certes les circuits si nombreux des ordinateurs, mais ces circuits (devenus quasi microscopiques) sont produits en séries. Et aujourd'hui un problème de traitement d'information ne se résout pas par un réseau de circuits spécialement conçus pour lui mais par un programme (tout au plus y a-t-il à cabler un lien entre l'ordinateur et ce qu'il règle). L'électronique, pour se répandre partout, devait se diversifier ; tandis que l'informatique, qui traite toute forme en tant que telle dans un support universel, trouve partout des sujets prêts à subir immédiatement son règne.

La profession de programmeur pourra se transformer : il est difficile de concevoir qu'elle ne s'étende à l'infini. Tout ce que l'on peut imaginer pour sa fin, c'est que tout homme étant devenu capable de programmer, comme

on l'est depuis longtemps d'écrire, on ne songe pas plus à se dire programmeur que scribe.

Or il est une faculté humaine, à laquelle la nature a donné de traiter toute forme en tant que telle dans un support universel : ce support est le cerveau, cette faculté l'intelligence. L'ordinateur ne peut comme l'esprit échafauder des images que les analogies sensibles animent ; et n'ayant point de fin propre il n'a point de jugement. Mais d'une part, du calcul posé au résultat, du programme et des données à l'exécution, il passe en un éclair. En sorte que le calculateur humain n'effectue plus d'opérations particulières ; il conçoit des algorithmes qui possèdent désormais à la fois, l'universalité de la formule théorique, et l'efficacité immédiate de l'acte pratique. D'autre part, à la lente synthèse de trésors d'expériences élaborés par la mémoire, vient s'adjoindre le traitement expéditif des tableaux de données : traitement que l'on appelle communément *analyse*, mais qui est bien plutôt *synthèse*, car il est confrontation de faits nombreux *mis ensemble*, non résolution d'un tout ordonné, en ses éléments. On peut songer que toutes les opérations de tous les esprits se retrouvent bientôt réunies, indestructibles, dans un milieu universel qui serait comme la matérialisation électronique de l'intellect commun tel que le concevait Averroès. Cette tour de Babel n'est toutefois pas plus près de s'élever que l'autre : montant à l'assaut du ciel les praticiens de l'informatique succombent comme les maçons à la confusion des langues. Ici comme ailleurs, à défaut d'un ordre bienfaisant c'est la tyrannie tempérée par l'anarchie.

Avec des armes chaque jour mieux trempées, soyons sûrs que corps et esprits livrent sans cesse les mêmes combats.

5e Principe. *Utiliser un ordinateur implique d'abandonner toutes techniques conçues avant l'avènement du calcul automatique. Je dis techniques, non science : les principes, géométriques ou algébriques, de nos programmes étaient connus de Laplace, il y a* 150 *ans. Mais Laplace est également l'auteur d'un traité de mécanique céleste qu'on vient de rééditer à l'usage des techniciens de l'espace... Cela n'a pas suffi à Napoléon pour conquérir la lune* !

Reportons-nous moins de vingt ans en arrière :

En juillet 1955, à l'initiative du professeur H. Laugier, assisté de M. Reuchlin, se tint à Paris un colloque international du C. N. R. S. consacré à *L'analyse factorielle et* (à) *ses applications*. Les Actes du colloque (Paris C. N. R. S., 1956) sont un vivant témoignage sur une étape de l'analyse des données dont 1955 marque le terme.

Les participants aux débats s'expriment avec une correction académique que nous nous garderons de railler ; il semble qu'une familiarité intime avec de nombreux faits expérimentaux soutienne nombre des exposés ; et, tout à la fois, les problèmes généraux ne sont jamais perdus de vue. Témoin ce dilemme, premier point du Rapport de synthèse de M. Reuchlin (Actes, p. 396), où on reconnaît d'abord le 2e principe posé ci-dessus.

« Pour les uns, l'analyse factorielle est essentiellement une méthode de *description*, d'*exploration*, pouvant révéler des groupements de faits, suggérer

des idées, et devant n'impliquer au départ que des hypothèses aussi générales que possible en nombre aussi restreint que possible.

Pour les autres, l'analyse factorielle est essentiellement une méthode de vérification d'une hypothèse, de confrontation d'un modèle avec des faits, et s'applique d'autant mieux que l'hypothèse, le modèle, sont spécifiés avec plus de précision. »

En 1955, L. Thurstone, patriarche de l'analyse des données, est de ce monde (il mourra pendant l'impression des Actes). De la communication que sans se rendre à Paris il avait adressée au colloque, extrayons ce point d'histoire (Actes, p. 32) :

« Il convient semble-t-il que nous ouvrions nos débats en nous référant à l'homme qui fut le promoteur de l'analyse factorielle. Je veux parler du Professeur Charles Spearman...

Le problème central auquel Spearman consacra principalement son attention en analyse factorielle fut celui-ci : les coefficients d'une table expérimentale donnée de corrélations sont-ils en accord avec l'hypothèse d'un facteur général unique ?... Il semble étonnant que le problème de Spearman n'ait jamais été formulé en fonction du rang de la matrice des corrélations, bien que Spearman ait eu la collaboration de mathématiciens compétents...

En 1930, nous avons posé la question différemment : combien de facteurs sont nécessaires pour rendre compte des corrélations ? Nous avons appelé ceci analyse multifactorielle. La question de savoir si, dans un cas particulier quelconque, un seul facteur suffirait pour rendre compte des intercorrélations, devenait alors une question de fait expérimental. »

Sans prendre parti dans les querelles de priorité, associons donc le nom de Thurstone et la date de 1930 avec la reconnaissance de ce principe fondamental, qu'en termes géométriques on énoncera ainsi : il ne suffit pas d'un seul axe, il faut un espace (cf. 3e, ci-dessus).

Voilà déjà présentés au colloque deux de nos principes. Quant à la part des probabilités (1er principe) remarquons d'abord que Thurstone tout en donnant la première place à la notion algébrique de rang, ne songe pas à écarter le terme probabiliste de corrélation ; mais il prête attention aux recherches de C. Eckart sur la matrice des notes et admet de calculer des corrélations entre personnes donc d'attribuer un rôle symétrique aux lignes et aux colonnes. Pour Thurstone (Actes, p. 33) : « Méthodologiquement, il est aussi plausible de calculer les corrélations entre personnes. Ceci est une entreprise plus laborieuse, car il y a plus de personnes que de tests. » Notons ensuite que, parmi une majorité de psychologues, des probabilistes mathématiciens sont présents au colloque. Ceux-là accordent à ceux-ci l'attention qui est alors de règle ; sans toutefois faire plus que de « contempler un instant le nombre vertigineux des modèles qui s'offriront au factoriste de demain... » (M. Reuchlin, p. 398).

Tant de courtoisie ne doit pas faire illusion : les savants d'antan savaient croiser le fer sans perdre leur contenance. Mais les statisticiens de 1955 ont un plastron plus ouaté encore que leur belle éducation ; ce plastron est leur

impuissance à calculer. Les ordinateurs existent déjà (1955 est l'année de la naissance du FØRTRAN) mais ils ne sont pas répandus : ils ne furent au colloque l'objet d'aucun débat. Et dès lors que les calculs sont limités, tous, quelle que puisse être par ailleurs leur ambition mathématique ou philosophique, en sont réduits à des compromis où l'intuition, acquise par une longue pratique, est reine.

En peu d'années, cependant, les plus vertigineuses perspectives ont été dévallées : des conjectures lointaines subissent la sanction des faits ; et l'expérience du calcul, comme toute expérience, rappelle à l'esprit beaucoup des vérités mêmes qu'on avait pu découvrir sans elle ; ayant rendu à nos Pères l'hommage qui leur est dû, nous sommes pressés de substituer aux algorithmes confus, issus des nécessités du calcul manuel et simultanément justifiés par des mosaïques de théorèmes (nous songeons aux rotations), des formules claires et exécutables qui soient les meilleures possibles. C'est ce qu'on a tenté de faire dans ce cours.