PUBLICATIONS MATHÉMATIQUES ET INFORMATIQUES DE RENNES

RÉGIS GRAS

# L'analyse de données : une méthodologie de traitement de questions de didactique

*Publications de l'Institut de recherche mathématiques de Rennes*, 1991, fascicule S6
« Vième école d'été de didactique des mathématiques et de l'informatique », , p. 115-118

<http://www.numdam.org/item?id=PSMIR_1991___S6_115_0>

THEME 5

**Cours** : "*L'analyse de données : une méthodologie de traitement de questions de didactique*"

par Régis GRAS

I.R.E.S.T.E., Université de Nantes, "La Chantrerie"
44087 NANTES Cédex 03
I.R.M.A.R. Université de Rennes 35042 RENNES Cédex

## I- PROBLEMATIQUE DIDACTIQUE

La didactique, tant du côté de l'enseignant que du côté du chercheur et à l'exclusion de certains domaines, ne dispose pas actuellement de réponses tranchées relativement à la plupart des questions qui lui sont posées. Or, pour se fonder scientifiquement, en dépassant la simple opinion, elle doit pouvoir formuler et avancer des hypothèses correspondant à ces questions. Elle doit pouvoir mettre en place un dispositif de recueil et de traitement de données susceptibles de conforter ou infirmer les hypothèses, d'avancer des conclusions. Certes, cette stratégie ambitieuse mais rigoureuse ne peut pas s'enclencher dès les premières approches des phénomènes à observer et d'où surgissent les questions. Elle s'impose, cependant, ultérieurement si l'on souhaite que les décisions didactiques s'appuient sur une stabilité et une pertinence de réponses et gagnent ainsi précision, validité et prédictibilité.

Par exemple, à travers une analyse a priori d'une situation - problème, on conjecture l'apparition de certaines procédures de résolution et une hiérarchie de difficultés. L'observation fait apparaître un écart entre le modèle a priori et l'ensemble des procédures effectivement observées. Quelles conclusions peut-on tirer de la distorsion ?

Des difficultés surgissent à tout moment et le chercheur isolé se trouve démuni devant les choix présentés et les décisions à prendre. Par exemple, comment traiter les informations qualitatives ? Comment coder les données ? A partir de quel effectif d'élèves la crédibilité d'un résultat est-elle assurée ? Quelle méthode statistique peut-on adopter ? Comment interpréter les résultats ? Il s'agit de trouver un juste équilibre, dans la recherche d'une validation d'hypothèses, entre la péjoration des méthodes statistiques, le refus d'investissement dans ce domaine et la statisticomanie qui conduit à une pléthore de résultats inexploitables, accompagnée de l'illusion de la transparence.

## II- REPONSES DE LA STATISTIQUE CLASSIQUE

La statistique descriptive classique permet, entre autre, de décrire quelquefois de façon très suggestive la distribution d'une variable, ou mieux l'interaction de deux variables. Elle apparaît cependant limitée dans ses objectifs et mutilante dans le cas de données multidimensionnelles.

Les modes de réponse de la statistique inférentielle et décisionnelle sont compatibles avec les démarches dites scientifiques.

Elles peuvent être résumées ainsi : à partir d'un échantillonnage d'une population, l'échantillonnage étant supposé effectué par des épreuves indépendantes et de même distribution (dans le cas de la statistique dite paramétrique) :

. on estime en vue de construire des modèles probabilistes, des paramètres relatifs aux variables observées dans l'expérience ou, tout au moins, des intervalles de confiance censés contenir ces paramètres ;

. on ajuste inductivement, à une ou deux autres variables, une variable dépendante dans un but de prédiction ou d'explication ;

. on décide de valider ou d'invalider des jugements et des hypothèses de modèles par des tests dits d'hypothèse (ceci est encore le cas en statistique non paramétrique).

Mais, on le voit, ces conduites se légitiment à plusieurs conditions :

. avoir clairement identifié, séparé et probabilisé les variables : or en didactique, données et variables peuvent être très nombreuses et intriquées ;

. présupposer le plus souvent des hypothèses de normalité des variables, hypothèses restrictives et quelquefois difficilement vérifiables ;

. accepter des méthodes longues et fastidieuses de croisements 2 à 2 de ces variables, en supposant qu'elles soient identifiées ;

. savoir formuler des hypothèses dites "nulles" i.e. mises en question et réfutables à l'issue de l'expérience avec un risque d'erreur donné ;

. écraser l'individualité des sujets dans l'échantillonnage.

Or ces conditions sont encore difficilement satisfaites au stade de scientificité actuel de la didactique. Si celui-ci autorise l'émission de conjectures, il apparaît indispensable d'avoir recours à d'autres méthodes que les précédentes pour synthétiser et structurer les données et ainsi pouvoir identifier les variables, les facteurs en jeu, leurs liaisons, leur hiérarchie, etc.

## III- RUPTURE EPISTEMOLOGIQUE DE LA STATISTIQUE : L'ANALYSE DES DONNEES, SES POSSIBILITES ET SES MODES DE REPONSE

Une double conjoncture va permettre d'apporter des réponses plus satisfaisantes à notre problématique :

. d'une part, la formalisation de l'algèbre linéaire, de la géométrie, des probabilités va permettre d'élaborer de nouvelles méthodes de traitement de données ;

. d'autre part, l'ordinateur va permettre de les engranger, de pratiquer des calculs rapides sur des structures complexes sans mutiler la taille des tableaux à traiter et de fournir des représentations variées de l'information obtenue.

En effet, l'analyse des données, méthodologie de traitement des données en vue de visualiser mais aussi structurer, modéliser et expliquer des phénomènes, fournit à ce jour de multiples méthodes, dites analyses de données, qui permettront d'obtenir, contrairement à leur désignation, des synthèses des données, en vision holographique, des facteurs discriminants, des typologies, des hiérarchies, etc.

La rupture épistémologique concerne donc à la fois les objectifs visés et atteints, les moyens techniques pour y parvenir (informatique), les données traitées (nombre, nature, variété, ...) les sujets de l'analyse (variables ou individus), les modes de restitution de l'information, les démarches (aller des données vers les modèles et non l'inverse), les méthodes mathématiques employées, les concepts en jeu dans celles-ci, etc.

Mais les nouvelles perspectives offertes créent ou entretiennent la fiction que des données recueillies et traitées sans choix opportun de la méthode et sans hypothèses préalables, vont fournir des informations en clair et des résultats organisés. Trop de chercheurs qui ont d'ailleurs ensuite abandonné cette méthodologie pour cette raison, se sont retrouvés avec des amas de papier de données non exploitables. Gâchis économique et intellectuel ! Il me parait indispensable, près de vingt années après mes premières rencontres avec l'analyse de données, de procéder ainsi :

. formuler des hypothèses, sans entrer dans l'illusion qu'elles seront réfutables ou définitivement acquises, mais seulement mises en doute ou confortées ;

. choisir une méthode d'analyse adaptée parmi les deux grandes classes : analyse factorielle et classification automatique ; par exemple, si l'on cherche à mettre en évidence :

* les principaux facteurs discriminants dans une population à travers des variables : une analyse factorielle,
* une partition parmi des variables : les nuées dynamiques,
* une typologie ou une classification : une classification hiérarchique des similarités,
* une implication entre variables ou classes de variables : un arbre implicatif ou une hiérarchie implicative, etc.

. coder et élaborer un tableau de données exhaustif, pertinent, homogène, compatible avec la méthode choisie ;

. connaître succinctement les concepts mathématiques à la base des synthèses (espaces vectoriels euclidiens, algorithmes et critères de classification), connaissance qui contrôle et facilite l'interprétation ;

. interpréter les résultats numériques et graphiques de façon synthétique par un certaine distanciation et savoir étendre ou restreindre les données sur lesquelles un deuxième passage apparaît nécessaire pour confirmer ou critiquer les premières interprétations. Eventuellement, dans ce cas, pratiquer une méthode inférentielle.

Il sera nécessaire, pour le chercheur, de dépasser les évidences de certains résultats, de se servir de cet accord pour crédibiliser les interprétations plus cachées, plus surprenantes qui justifient à elles-seules l'emploi d'une méthode sophistiquée.

## IV- PRESENTATION BREVE DE DEUX METHODES

### 4.1- L'ANALYSE FACTORIELLE DES CORRESPONDANCES

Elle se propose de donner une représentation géométrique, dans un espace de grande dimension en général, d'une distribution conjointe de deux ensembles E (en général des sujets) et V (en général des variables ou des modalités de variables).

La méthode analytique consiste alors à extraire, à partir des espaces de représentation, des sous-espaces dont la dimension est réduite, mais tels que le nuage de points E ou V y soit représenté de façon optimale par ses différentes projections. Aux axes de ces sous-espaces correspondent des facteurs discriminants dans les ensembles E et V : le premier est le plus informatif à ce sujet, le second l'est moins, etc. Le didacticien doit alors, par un jeu d'opposition - ressemblance des projections de points, déterminer la signification de ces facteurs. Elle lui servira à analyser puis à interpréter les informations qui sont plus cachées et qui découlent de cette signification. Il s'intéressera aux contributions de certains points à ces facteurs et, entre autres, par exemple, aux positions relatives de sous-groupes de la population étudiée.

### 4.2- L'ANALYSE HIERARCHIQUE DES SIMILARITES SELON IC. LERMAN

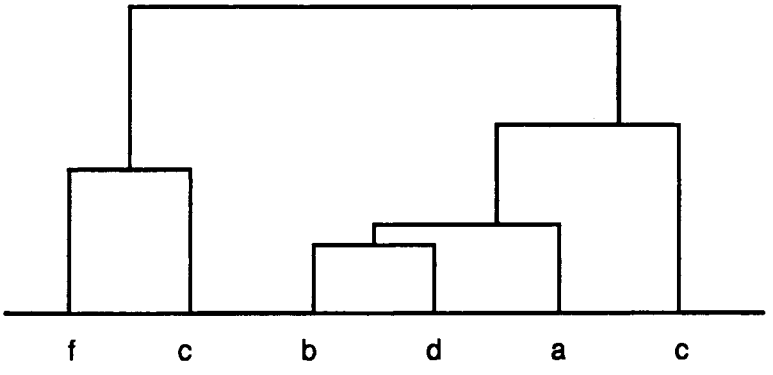
Comme dans toutes les méthodes de classification, on cherche à constituer sur l'ensemble V des variables des partitions de moins en moins fines, construites de façon ascendante en un arbre à l'aide d'un critère de similarité entre variables. Le didacticien s'intéresse à ce type d'analyse qui lui permet d'étudier puis d'interpréter en termes de typologie et de ressemblance (et de dissemblance) décroissante des noyaux de variables, constitués significativement à certains niveaux de l'arbre et s'opposant à d'autres à ces mêmes niveaux.

Le critère de similarité s'exprime de la façon suivante :

> 2 variables a et b se ressemblent d'autant plus que l'effectif des sujets les satisfaisant ( $A \cap B$) est important eu égard d'une part à ce qu'il aurait été dans le cas d'absence de lien a priori entre a et b et d'autre part, aux cardinaux, de A et B. On mesure cette ressemblance par la probabilité de son invraisemblance.

L'indice entre les variables qui lui correspond n'est donc pas biaisé par les effectifs. Il sert ensuite à définir un indice de similarité entre deux classes de variables.

Ainsi, pour construire un arbre de classification, on réunit en une classe au plus bas niveau, tout d'abord, les 2 variables qui se ressemblent le plus à travers cet indice, puis 2 autres variables ou une variable et la classe déjà formée, puis d'autres variables ou des classes de variables.

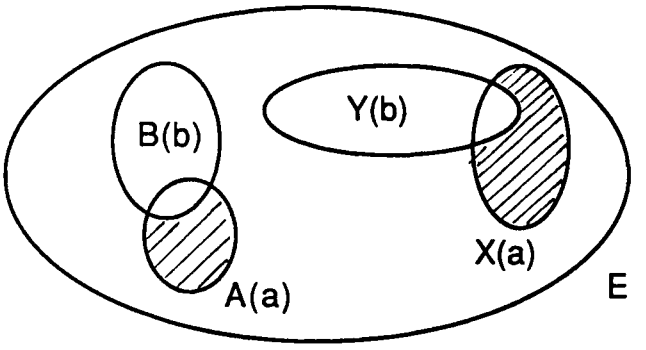



Dans ma propre thèse, ce type d'analyse met en évidence une typologie intéressante de tâches dans la résolution d'exercices portant sur la symétrie centrale, typologie où s'opposent en particulier les traitements dans le cadre géométrique et le cadre algébrique. La contribution de certaines catégories d'élèves à cette typologie présente un grand intérêt.

## V- L'ANALYSE IMPLICATIVE

### 5.1- IMPLICATION ENTRE VARIABLES

Contrairement aux méthodes citées précédemment, où distance et indice de similarité sont symétriques, la méthode implicative est non symétrique. L'élaboration de cette méthode prend son origine dans une question que je me posais en 1978 au sujet d'une hiérarchie de complexité organisée selon un ordre partiel. La problématique générale qui l'introduit est la suivante, dans le cas où les variables considérées sont binaires (un individu satisfait ou non une variable):




Si A et B sont les sous-populations des sujets ayant respectivement satisfait les variables a et b, dans quelle mesure peut-on dire : "si a alors b", l'implication ne devant pas être connotée a priori de causalité.

Si A C B, la proposition est vérifiée, mais généralement les cas courants présentent une intersection $A \cap \bar{B}$ non vide.

L'indice d'implication mesure, d'une façon comparable à la similarité, le degré d'"étonnement" de la petitesse de $A \cap \bar{B}$ eu égard à l'indépendance a priori et aux effectifs observés. Ainsi on dira, par exemple, que X et Y étant 2 parties aléatoires de E de mêmes cardinaux respectifs que A et B,

"a ==> b" est admissible au niveau de confiance ou avec l'intensité implicative 0,95 si et seulement si : Prob [card $(X \cap \bar{Y})$ < card $(A \cap \bar{B})$ ] < 0,05.

Cette notion est étendue, depuis la thèse d' A.Larher, à des variables modales et numériques, unifiées en variables fréquentielles, prenant leurs valeurs sur [0,1]. Un arbre implicatif rend compte de l'ordre partiel induit par cette intensité d'implication.

### 5.2- IMPLICATION ENTRE CLASSES DE VARIABLES

Insuffisamment synthétique, l'implication entre variables est conceptuellement prolongeable en une implication entre classes de variables selon le voeu qu'avait formulé G.Vergnaud après ma thèse. L'examen d'une telle relation entre deux classes n'ayant véritablement un sens que dans le cas d'une "bonne fermeture" des classes, nous définissons le concept de cohésion d'une classe comme antinomique à celui de "désordre implicatif" (au sens de l'entropie dans la théorie de l'information). De là, l'implication entre deux classes bien "cohésives", i.e. déjà ordonnées en leur sein, traduit la force implicative de l'une sur l'autre.

De façon générale, les deux méthodes basées sur la similarité et l'implication nous ont permis, dans des travaux divers, de définir, à partir de classes de comportements, des procédures voire des conceptions stables et consistantes permettant de les repérer, de façon anticipée, dès l'apparition de quelques signes. On verra, en Intelligence Artificielle, l'avantage de partir de conceptions réellement synthétisées en vue d'une modélisation de l'élève au lieu de partir d'un hypothétique écart au modèle ou à un sous-modèle de l'expert.

En conclusion, soulignons deux points qui nous paraissent importants dans l'emploi d'une telle méthodologie d'analyse didactique :

. les méthodes ayant des fondements mathématiques différents, comme nous l'avons vu, conduisent à des résultats qui, certes fréquemment se confortent, mais le plus souvent se complètent ; ceci doit nous encourager à doubler telle méthode par telle autre ;

. l'interprétation ne peut se faire qu'à partir de questions préalablement posées : il n'est ni superflu, ni péjorant de retrouver une information évidente, cette absence de contradiction avec le "connu" ou le "vraisemblable" doit au contraire crédibiliser l'information nouvelle et inattendue.