L'Analyse des Données

Beaucoup de chercheurs et d'ingénieurs regrettent une évolution qui les a conduit à une trop grande spécialisation. Chacun se rend compte que pour résoudre un problème réel complexe, il est nécessaire de mettre en oeuvre différentes techniques en mathématiques appliquées (analyse numérique, statistique, graphes, ...).

Nous avons pensé que les lecteurs de MATAPLI apprécieraient de trouver dans ce bulletin de brefs exposés d'introduction sur certains grands thèmes des mathématiques appliquées qu'ils connaissent peu. Nous commençons aujourd'hui par un article sur l'ANALYSE DES DONNEES, écrit par Gilbert SAPORTA, Directeur du Département d'Informatique de l'Institut Universitaire de Technologie de l'Université René Descartes, à Paris.

Nous serions heureux de recueillir vos suggestions sur les sujets de ces brefs exposés, ainsi que vos propositions pour écrire des articles de ce type dans votre spécialité.

L'Analyse des Données est une des branches les plus actives de la statistique et connait un succès grandissant auprès des utilisateurs: en dix ans, près de trente ouvrages ont été publiés par les éditeurs français. Le développement de ce domaine en France, influencé par la personnalité de Jean-Paul BENZECRI, a cependant suscité d'âpres contreverses entre tenants de la statistique mathématique et partisans de nouvelles méthodes. Le phénomène n'est pourtant pas limité à l'hexagone : si les praticiens français ont une prédilection marquée pour certaines techniques spécifiques comme l'analyse des correspondances, on retrouve sous d'autres formes la même démarche (retour aux données, méfiance vis-à-vis des modèles, accent sur les représentations graphiques) dans divers pays, notamment au Japon, en Italie, aux Pays-Bas, aux Etats-Unis (en particulier, aux Bell Laboratories) qui comptent des équipes très actives.

On a pu dire que l'Analyse des Données est née d'une démarche et d'une possibilité créées par le développement de l'informatique : demande de traitements statistiques adaptés aux fichiers de plus en plus vastes maintenant disponibles ; possibilité de réaliser de tels traitements grâce aux progrès du matériel et du logiciel.

Mais, en quoi consiste l'Analyse des Données ?

DONNEES ET METHODES

L'Analyse des Données se propose de décrire la structure d'un ensemble d'unités statistiques ou individus, repérés par un grand nombre de variables : un cas classique est celui du dépouillement d'un questionnaire administré à un vaste échantillon et comportant plusieurs dizaines, voire quelques centaines de questions.
Les méthodes diffèrent selon la nature des variables: numériques, comme l'âge, le salaire, la masse ; qualitatives comme la profession, le diplôme et leur regroupement en un seul ensemble ou plusieurs sous-ensembles. Le trait commun de toutes ces méthodes est d'être multidimensionnelles, en ce sens qu'elles exploitent les liaisons entre variables (ou corrélations), ce qui les distingue de la statistique descriptive traditionnelle qui procède variable par variable.

- Les variables ne forment qu'un groupe

On utilise alors deux grands types de méthodes :

- **la classification** qui vise à produire des groupes homogènes d'individus selon un certain critère lié à la dispersion des mesures. On peut chercher directement une partition en un nombre fixe de classes ou des chaines de partitions emboitées, obtenues en regroupant au fur et à

mesure les classes déjà obtenues (méthodes hiérarchiques) ;

- les analyses factorielles : l'analyse en composantes principales (ACP) est utilisée lorsque toutes les variables sont numériques. L'ensemble des individus forme un nuage dans un espace en dimension p où chaque axe représente une variable. Le principe de l'ACP consiste à créer de nouvelles variables (appelées composantes principales) par combinaison linéaire des variables initiales, qui soient non corrélées entre elles et de variance maximale afin de séparer au mieux les individus entre eux. On montre que la solution est fournie par les vecteurs propres de la matrice de variance-covariance des variables initiales. On peut alors visualiser les proximités entre individus en les projetant sur les axes-espace engendrés par les premiers axes principaux, ainsi que les relations entre variables à l'aide des corrélations avec les composantes principales. L'analyse factorielle des correspondances (AFC) peut être présentée comme une extension de l'ALP au cas des variables qualitatives : ici, chaque variable - la profession, par exemple - définit des sous-groupes d'individus appartenant à une même catégorie ; on représente alors chaque catégorie par le centre de gravité des individus qui la composent et chaque individu comme le barycentre des catégories auxquelles il appartient. Un des attraits de cette méthode est de pouvoir représenter simultanément individus et variables.

- Les variables forment deux groupes

C'est le cas, par exemple, où l'on veut comparer les résultats obtenus par deux laboratoires sur le même échantillon. Pour les variables numériques, l'analyse canonique permet de trouver les couples de combinaisons linéaires des variables de chaque groupe qui soient les plus corrélées. Sous son aspect empirique, la régression linéaire en est un cas particulier, celui où l'un des groupes est réduit à une seule variable.

Une mention spéciale doit être faite pour l'analyse factorielle discriminante : dans cette technique, on cherche une combinaison linéaires de p variables numériques permettant de séparer au mieux plusieurs groupes d'individus. Cette méthode est très prisée dans de nombreux problèmes de décision (aide au diagnostic médical, sélection de demandeurs de crédit).

Dans tous les cas, l'analyse des données cherche à réduire la dimensionalité du phénomène observé en créant de nouvelles variables synthétiques plus porteuses d'informations (variables numériques pour les analyses factorielles, variables qualitatives pour la classification). L'interprétation des résultats est alors aidée par une batterie d'indicateurs et de pratiques bien rôdées.

Il est remarquable que toutes ces méthodes n'utilisent que des mathématiques assez simples (algèbre linéaire et géométrie euclidienne), ce qui explique sans doute en partie leur succès. Bien plus, les principes en sont connus depuis longtemps : on peut ainsi faire remonter l'ACP à 1901 avec K. Pearson, l'AFC aux travaux de Guttman et Hoyoshi, dans les années 1940, etc.. Mais, c'est l'informatique qui les a rendu utilisables et l'usage qui en est fait pour l'analyse des données en a changé radicalement la nature.

L'ANALYSE DES DONNEES ET LA STATISTIQUE CLASSIQUE

- Les oppositions

En fait, ce n'est pas tant son caractère multidimensionnel que l'absence de recours à un modèle probabiliste qui distingue l'analyse des données de la statistique classique. En analyse des données, on ne cherche pas à estimer des paramètres, ni a ajuster un modèle : on explore simplement la structure des données sans supposer que les observations ont été tirées d'une population plus vaste, ni qu'elles présentent un caractère homogène.

En statistique classique, on suppose généralement que les individus ont été tirés au hasard dans une population unimodale : ils ne présentent pas d'intérêt en eux-mêmes et sont donc interchangeables, leur seul rôle étant de fournir des réalisations de variables aléatoires.

Au contraire, l'analyse des données redonne un statut aux individus qui peuvent être identifiés : on ne suppose plus qu'ils forment nécessairement un corpus homogène, et les techniques de l'analyse des données ont précisément pour but de mettre en évidence les hétérogénéités (existence de sous-groupes ou plurimodalité).

Comme l'a rappelé J.R. BARRA, l'opposition entre analyse des données et statistique classique n'est en réalité qu'un aspect de la dualité entre statistique inductive et statistique décisionnelle, la première visant à découvrir des faits, à proposer des hypothèses, la deuxième cherchant la confirmation des modèles en calculant des risques d'erreur.

- L'analyse des données dans la pratique statistique

Ces deux rôles précédents de la statistique ne sont pas antinomiques et sont associés successivement dans la démarche scientifique. Ces méthodes d'analyse des données trouvent donc leur place naturelle dans la phase exploratoire de toute étude statistique ; ses résultats permettront éventuellement de proposer des hypothèses ou d'ajuster à bon escient des modèles différents sur certains sous-groupes d'observations, alors qu'un modèle global aurait été fallacieux.

Est-ce à dire que l'analyse des données n'est que de la statistique descriptive améliorée et qu'elle exclut toute tentative inférentielle ? Certes non, et la pratique le contredit, en particulier pour les méthodes dites explicatives comme l'analyse discriminante qui est à la frontière entre décision et description. De plus, les travaux récents sur l'usage de procédure de simulation et de rééchantillonnage (bootstrap, jackknife) permettent d'éprouver la robustesse des méthodes et débouchent naturellement sur des perspectives inférentielles (intervalles de confiance, par exemple). Nous pensons d'ailleurs que la rencontre avec les techniques robustes et non paramétriques qui ne font pas non plus référence à des modèles trop contraignants et souvent irréalistes (comme la loi normale multidimensionnelle) promet d'être très féconde.

EN GUISE DE CONCLUSION

Le mouvement qui s'est constitué autour de l'analyse des données a permis de secouer quelques certitudes et de remettre l'accent sur l'observation avant la modélisation ; il a été, également, l'occasion d'un renouveau de la pensée statistique en France. Etre pour ou contre l'analyse des données nous semble désormais dénué de sens scientifique (sinon d'arrière-pensées corporatistes...). A côté des méthodes plus anciennes (régression, analyse de variance), les techniques d'analyse factorielle et de classification font désormais partie des outils de statistique appliquée dont la fonction essentielle est finalement l'analyse des données...

POUR EN SAVOIR PLUS

J.P. BENZECRI, L'Analyse des Données, Dunod
J.M. BOUROCHE, G. SAPORTA, L'Analyse des Données, Que-Sais-Je, PUF
F. CAILLIEZ, J.P. PAGES, Introduction à l'Analyse des Données, Smash
E. DIDAY, J.POUGET, J. LEMAIRE, F. TESTU, Eléments d'Analyse des Données, Dunod
L. LEBART, A. MORINEAU, N. TABARD, Techniques de la Description Statistique, Dunod
M. JAMBU, M. LEBEAUX, Classification Automatique pour l'Analyse des Données, Dunod
BARNETT, Interpreting Multivariate Data, Wiley
R. GNANADESIKAN, Methods for Statistical Data Analysis of Multivariate Observations, Wiley

Gilbert SAPORTA
Dpt Informatique
IUT-Univ. R.Descartes