MINISTÈRE DE L'INDUSTRIE

BUREAU DE RECHERCHES GEOLOGIQUES ET MINIERES

SERVICE GÉOLOGIQUE NATIONAL

B.P. 6009 - 45018 Orléans Cédex - Tél. (38) 63.80.01

ANALYSE STATISTIQUE MULTIDIMENSIONNELLE :

Propositions méthodologiques pour l'analyse de données concernant la chimie des eaux souterraines

par

M. CANCEILL et A. LANDREAU

Département hydrogéologie

B.P. 6009 - 45018 Orléans Cédex - Tél. (38) 63.80.01

78 SGN 001 HYD

Décembre 1978

RÉSUMÉ

Les techniques numériques de l'"analyse des données" ont été appliquées à l'hydrologie assez fréquemnent ces dernières années. Ces applications ont été souvent controversées ; c'est pourquoi il est apparu nécessaire d'en donner une vue d'ensemble, de les critiquer, de dégager quelques conclusions et d'émettre des propositions méthodologiques précises, en particulier pour les applications hydrochimiques. C'est l'objet de ce rapport, qui rend compte d'un travail de réflexion et d'expérimentation mené depuis plusieurs années au titre des études méthodologiques du département Hydrogéologie.

S O M M A I R E

1. INTRODUCTZON : BREF HISTORIQUE DE L'ANALYSE DES DONNEES

Ce qu'on appelle aujourd'hui en France "analyse des données" résulte de la convergence de deux courants :

a) Le développement des ordinateurs et du calcul scientifique a rendu accessibles à beaucoup les techniques de l'analyse statistique multidimensionnelle (*) du genre régression linéaire multiple, analyse factorielle, etc. ; ces techniques n'en sont pas devenues moins délicates pour autant : leur emploi correct est soumis à la vérification d'hypothèses souvent mal connues d'utilisateurs de plus en plus nombreux, et leur interprétation statistique nécessite un soin particulier. Ceci ne pouvait conduire qu'à des abus ; le premier courant dont on veut parler est une réaction à ces abus : "puisque la statistique mathématique pose des problèmes, oublions-la et donnons une interprétation purement algébrique et géométrique à ces techniques numériques qu'il est si facile de mettre en oeuvre...".

b) Le deuxième courant, beaucoup plus constructif, résulte de recherches numériques sur les problèmes de taxinomie et de classification [appliquée en particulier à la biométrie et aux sciences humaines) ; la difficulté qu'il y a à formuler ces problèmes en termes de variables aléatoires, l'importance de variables nominales ou ordinales, etc., ont conduit à des travaux de nature plutôt algébrique. On peut rattacher à ce courant les travaux du Pr. J.P. BENZECRI et de son équipe, en particulier ceux qui concernent l'analyse factorielle des correspondances, technique dont l'expansion a dépassé ce qu'on pouvait imaginer et à propos de laquelle on peut vraiment parler de mode avec les excès que cela suppose.

o°o

On vient de décrire schématiquement les deux courants dont la jonction a donné lieu à ce qu'on peut appeler l'école française d'analyse des données ; si le haut-lieu en reste le laboratoire de statistique mathématique de J.P. BENZECRI, à l'Université de Paris VI, de nombreuses autres équipes actives et compétentes ont essaimé dans l'université et dans l'industrie, Elles communiquent régulièrement dans divers séminaires [en particulier ceux du BURO, de 1'AFCET et de 1'IRIA (**) ; on voit apparaitre de nombreuses applications sur ce thème, qu'il s'agisse de recherches théoriques ou d'applications. Une revue spécialisée est même en train de naître.

(*) On préfèrera ici le terme au néologisme d'origine anglo-saxonne "analyse multivariate". Rappelons, à ce sujet, que le terme "variate" fut créé par les Anglo-saxons comme abréviation de "random variable" (variable aléatoire).

(55) BURO : Bureau Universitaire de Recherche Opérationnelle à l'Institut de Statistique des Universités de Paris (ISUP).

AFCET : Association Française pour la Cybernétique Economique et Technique.

IRIA : Institut de Recherches en Informatique et Automatique.

En conclusion, même si l'analyse des données est un domaine hybride [peut-on parler de discipline ?), si elle est parfois contestée et si elle a donné lieu à des excès, on doit reconnaftre que c'est aujourd'hui une activité vivante et féconde.

o°o

2. *LES TECHNIQUES VE L'ANALYSE DES VONNEES*

2.7. *Généralités*

Le bref aperçu historique du paragraphe précédent montre que l'analyse de données résulte, en fait, de la conjonction d'un certain nombre de techniques statistiques d'origines diverses, dont les points communs sont apparus quand on a commencé à les mettre en oeuvre à l'aide de moyens de calcul automatique,

Ces points communs sont :

- la possibilité théorique de traiter un nombre quelconque de variables
- la formalisation géométrique dans un espace euclidien à n dimensions. Qui dit géométrie euclidienne à n dimensions dit algèbre linéaire et calcul matriciel i on ne s'étonnera donc pas que l'outil mathématique et numérique universellement employé soit l'algèbre des matrices.

On n'a pas ici la place de décrire toutes ces techniques en détail ; donnons-en simplement une liste qui, sans être exhaustive, recouvre la plupart des applications :

- Régression linéaire multiple
- Analyse en composantes principales
- Analyse discriminante
- Analyse canonique
- Analyse factorielle des correspondances
- Classification automatique.

On trouvera une description des plus courantes de ces techniques, à un niveau relativement élémentaire, dans l'ouvrage de L. LEBART et J.P. FENELON (19711 ; un exposé plus complet et plus rigoureux dans un autre texte français récent, mais à un niveau plus élevé (F. CAILLIEZ et J.P. PAGES, 19761.

Comme ouvrage de référence de haut niveau, on doit citer l'américain T.W. ANDERSON (19581 et le britannique M.G. KENDALL(19661.

On ne s'intéressera ici qu'à quelques points, à savoir l'analyse en composantes principales, l'analyse factorielle des correspondances et l'analyse canonique.

CONCLUSIONS

L'objet de cette étude était de préciser les conditions les plus goureuses d'emploi et les résultats qu'on peut attendre de l'analyse factorielle s correspondances en hydrochimie.

En ce qui concerne les conditions d'emploi, nous pensons qu'il est nécessaire de traiter des tableaux de nombres entiers, ou de fréquence résultant d'un codage des données.

Nous avons indiqué une manière possible d'engendrer un tel tableau en introduisant en sus des paramètres chimiques descriptifs de la qualité de l'eau, un certain nombre de paramètres explicatifs. En outre, l'introduction de ces paramètres permet de pousser plus loin l'interprétation des données en essayant de mettre statistiquement en évidence leur influence sur la qualité des eaux. Les résultats à attendre de cette méthode sont évidemment fonction des paramètres explicatifs choisis.

Un essai en employant la méthode préconisée a été réalisé sur la chimie de la nappe des calcaires de Beauce.

Les relations qui en ressortent sont d'intérêt inégal, certaines comme la relation type de cultures-teneurs en nitrates ne pouvant être considérées comme un apport original.

Cet essai sur la nappe des calcaires de Beauce a également mis en évidence l'importance de la qualité des données de base et surtout l'attention que l'on se doit de porter à la représentativité de la "population" sur laquelle va porter l'analyse.

°°°