BUREAU DE RECHERCHES GÉOLOGIQUES ET MINIÈRES
74, rue de la Fédération, 75 Paris (15e) — Tél.: (1) 783.94.00
SERVICE GÉOLOGIQUE NATIONAL
B.P. 6009 — 45 Orléans (02) — Tél.: (38) 66.06.60

OPTIMISATION DES RÉSEAUX DE MESURES HYDROLOGIQUES

Sélection des stations à conserver dans un réseau surabondant

par

M. CANCEILL

Département d'hydrogéologie
Service hydrodynamique et
informatique hydrogéologique
45 — ORLÉANS-LA SOURCE

70 SGN 351 HYD

Décembre 1970

R E S U M E

Les techniques statistiques, ainsi que la puissance des moyens de calcul modernes, permettent de poser les problèmes d'optimisation (au sens large du terme) de réseaux de mesures, et, parfois, de les résoudre.

Ces méthodes sont encore trop récentes et trop empiriques pour, à l'heure actuelle, être considérées comme parfaitement adaptées ; employées judicieusement, elles peuvent cependant permettre de réduire certains réseaux surabondants.

Les techniques statistiques utilisées sont l'analyse factorielle et l'analyse spectrale ; leur but est avant tout descriptif : mettre en évidence quelques indices plus parlants qu'un amas de données trop nombreuses pour être interprétées ; selon la valeur numérique de ces indices, et avec prudence, on saura localiser les points où l'information semble redondante.

Il reste à intégrer ces techniques descriptives dans des schémas plus élaborés, en termes de théorie de la décision ; ceci nous paraît prématuré, mais il faut prendre conscience que c'est en cherchant dans cette voie qu'on se rapprochera le plus d'une "optimisation" au sens propre du terme.

INTRODUCTION

La vogue du terme "optimisation", issu de la recherche opérationnelle et de l'économie, semble atteindre les sciences de la Terre. Même si le mot semble ambitieux, son emploi montre bien qu'un problème de gestion des réseaux de mesures se pose actuellement. Cette courte note est une tentative de délimitation du problème ; on y suggère une certaine approche mathématique, qui reste à expérimenter. Des applications en cours au B.R.G.M. nous permettront, dans les mois qui viennent, de compléter une réflexion qui n'est que spéculative par des résultats numériques.

1. "Optimisation" ou "réduction" ?

Mathématiquement, un problème d'optimisation s'énonce ainsi :

- trouver un extremum d'une certaine fonction $f(x_1 , x_2 \ldots x_n)$ dite "fonction économique" sous un ensemble de contraintes. (Les contraintes étant des équations ou des inéquations portant sur les variables $(x_1 , x_2 \ldots x_n)$).

Ce problème est donc celui de la recherche de l'optimum d'une fonction de n variables dans un sous-ensemble de l'espace à n dimensions, ce sous-ensemble étant défini par les contraintes.

On conçoit qu'il n'existe pas de solution générale à un problème aussi vaste ; il y a deux cas où des techniques classiques peuvent s'appliquer :

- La fonction et le domaine défini par des contraintes présentent des propriétés de régularités suffisantes ; on sait, alors, trouver un extremum lié (multiplicateurs de LAGRANGE).

- La fonction et les contraintes sont linéaires ; le domaine défini par les contraintes est un hyper-polyèdre, et le problème relève de la programmation linéaire.

Dans la pratique, bien peu de problèmes acceptent d'être réduits au premier "cas d'école" ; quant au deuxième, s'il est d'application plus fréquente, il est loin d'être universel, et, en tous les cas, sa résolution suppose d'importants moyens de calcul dès que le nombre de variables n'est plus négligeable. En ce qui concerne le cas non linéaire, de nombreuses études spécialisées sont faites tous les ans sur des problèmes particuliers (certains programmes quadratiques, par exemple), mais il n'y a pas de solution générale.

Comment, dans ce cadre, poser un problème d'optimisation de réseau de mesures ?

On définira deux fonctions, la fonction de précision (faisant intervenir des paramètres comme la variance, ou l'entropie d'information) et la fonction de coût. On cherchera alors, selon les circonstances, soit un maximum de précision à coût constant, soit un minimum de coût à précision constante. On voit que chaque fonction peut servir de fonction économique, ou être égalée à une constante pour devenir une contrainte.

Cette belle théorie est malheureusement peu applicable à l'heure actuelle, pour deux raisons :

- il est souvent très laborieux de construire une fonction de coût
- il est encore plus laborieux, voire impossible dans certains cas, de construire une fonction de précision cohérente ; l'information étudiée, en effet, est souvent répartie dans l'espace et dans le temps (ensemble de séries chronologiques), et le problème relève de la théorie des fonctions aléatoires du second ordre et à n dimensions, théorie encore incomplète.

Si, donc, ce sujet nous paraît mériter un grand intérêt, il n'en existe pas encore, à notre connaissance, d'application, et il ne nous semble pas qu'il puisse en exister avant plusieurs années.

Des tentatives, plus limitées dans leurs ambitions, ont eu lieu, d'autres ont lieu actuellement, et c'est dans ce sens que nous travaillons.

Ces tentatives ont pour but la recherche d'une réduction du réseau de mesures qui ne fasse pas perdre " trop de représentativité". On peut dire que les fonctions de coût et de précision ne sont pas exprimées sous leur forme analytique, mais sous forme littéraire... Ces réductions, d'autre part, jouent dans l'espace, ou dans le temps, mais rarement dans les deux simultanément. C'est pourquoi nous pensons que le mot réduction qualifie mieux ce genre d'opération que le mot optimisation.

2. Réduction d'un réseau de mesures - problème général

L'information de base est celle fournie par N stations de mesures (pluviométriques, hydrométriques, piézométriques, etc...) observées régulièrement pendant la même période, soit T observations.

Nous n'examinerons ici ni le problème d'acquisition des données (calculs des débits à partir des hauteurs, lecture automatique de limnigrammes, etc...), ni le problème des observations manquantes ; remarquons cependant, qu'il est impossible de les ignorer lors des applications, et que leur étude peut être aussi lourde que celle de l'optimisation proprement dite...

A partir de ce réseau, en place, on cherche à en définir un nouveau, par :

- suppression de certains points de mesure
- abaissement de la fréquence des mesures.

L'importance relative de ces deux points étant déterminée par le contexte : la topographie du réseau jouera un rôle important (-sorte de paramètre de cette fonction de coût qu'on veut ignorer mais qui reparaît sans cesse...) dans le premier objectif ; le poids du second sera négligeable, ou, au contraire, grand, selon que le réseau sera ou ne sera pas équipé de systèmes d'enregistrement automatiques.

Nous avons avoué n'être pas en mesure pour le moment de résoudre le problème global ; il est nécessaire de n'aborder les difficultés que les unes après les autres, mais dans quel ordre ?

On peut considérer que, des deux problèmes, c'est celui de la réduction dans le temps qui est le plus délicat ; c'est pourquoi nous suggérons de procéder d'abord à une analyse spatiale, dans un double objectif :

- résoudre une partie du problème de la réduction dans l'espace
- dégrossir les données pour faciliter l'analyse temporelle

L'analyse temporelle, alors, permettra :

- d'achever la réduction spatiale
- de procéder à la réduction temporelle.

Ces considérations peuvent sembler obscures ; la justification de ce découpage en quatre phases, et de leur enchaînement, apparaîtra dans les paragraphes suivants au fur et à mesure que nous y décrirons les méthodes statistiques que nous suggérons d'employer.

3. Réduction dans l'espace : analyse factorielle

Les techniques d'analyse factorielle ont depuis longtemps été employées à des fins variées. On distingue, en gros, deux grandes classes d'applications de ces techniques :

- but explicatif
- but descriptif

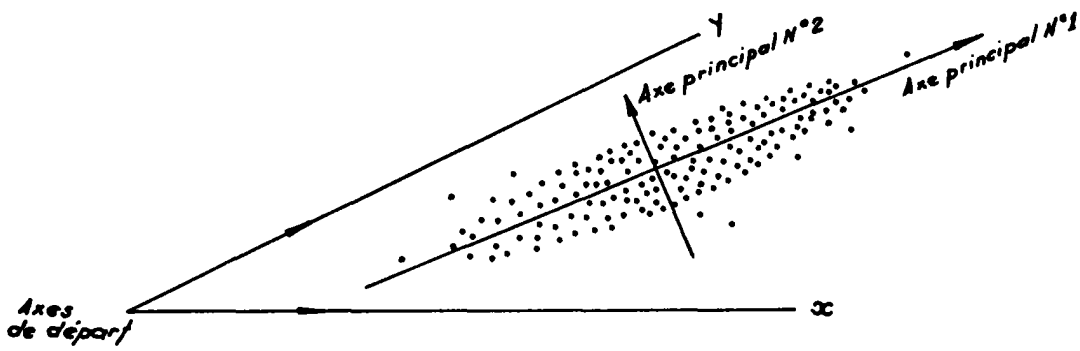
c'est le second point qui nous intéressera ici (remarquons que dans cette deuxième optique, au contraire de la première, aucune hypothèse de structure n'est nécessaire à la validité de la méthode).

L'analyse factorielle permet, à partir d'observations multidimensionnelles liées, de passer par un changement de variable linéaire, à des observations multidimensionnelles indépendantes,et, en général, situées dans un sous-espace de dimension plus faible.

Sans entrer dans le détail mathématique, on peut examiner les interprétations géométriques possibles :

On considère les T observations d'un vecteur aléatoire à N dimensions, dans l'espace euclidien réel $\mathbb{R}^N$. Chaque "axe" de l'espace correspond à une "variable" (une station de mesures), chaque point à une observation à l'instant t sur l'ensemble du réseau. Si les N variables s'avèrent indépendantes en probabilité, ce qui dans notre représentation correspond à l'orthogonalité des axes, aucune réduction n'est possible. Si, au contraire, certaines des variables ont le bon goût d'être corrélées, c'est-à-dire si les axes représentant ces variables sont fortement obliques, il sera possible de grouper ces axes dans un certain cône d'angle au sommet faible. L'axe de ce cône sera baptisé "axe principal", et on pourra trouver N axes principaux orthogonaux. Dans l'espace $\mathbb{R}^N$, le carré de la distance moyenne des points-observations à leur centre de gravité, correspond à la notion statistique de variance ; d'après le théorème de PYTHAGORE, ce carré moyen est égal à la somme des carrés de ses projections sur les axes principaux (car ceux-ci sont orthogonaux). On peut donc hiérarchiser les axes principaux par la proportion de distance moyenne qui se projette sur eux. L'oblicité des axes de départ donnant au "nuage" de points une forme aplatie, les axes principaux seront fortement hiérarchisés : un pourcentage élevé de la distance moyenne des points du nuage à leur centre se projettera sur les premiers axes principaux.

Exemple à deux dimensions :



Ce pourcentage baptisé " part de variance expliquée" définit l'unité de longueur sur l'axe principal concerné ; si le nuage de points est suffisamment aplati, on pourra le représenter en ne le déformant que très peu,par une projection dans l'espace défini par les premiers axes principaux. Dans la pratique, compte tenu de l'avantage de la représentation plane, on projette sur le plan des deux premiers axes, ce qui est parfois suffisant pour expliquer plus de 90 % de la variance totale. Nous appellerons cette projection "graphique des observations".

Il existe une autre possibilité de représentation, tout aussi importante : au lieu de considérer T points dans $\mathbb{R}^N$, on va considérer N points dans $\mathbb{R}^T$. Les axes correspondent aux instants de mesure, les points aux stations (point k = vecteur ayant pour composantes toutes les mesures à la station k). Des axes obliques correspondent à des instants de mesure non indépendants. On peut, comme précédemment, procéder à l'extraction des axes principaux, et à la représentation plane des points-variables sur le plan des deux premiers axes principaux ; cette seconde représentation sera baptisée "graphique des variables".

La tentation est grande à ce moment de considérer le problème comme résolu, en ne retenant que les stations proches des axes principaux sur le graphique des variables. Nous pensons que cette information, pour intéressante qu'elle soit, n'est pas suffisante pour conclure ; c'est pourquoi nous considérons que ce n'est qu'une première partie de la réduction spatiale: "phase $\propto$".

Les proximités décelées par l'analyse factorielle, en effet, sont des proximités "instantanées" : si, sur plusieurs centaines d'observations, la station X et la station Y présentent le même profil, les points les représentant sur le graphique des variables seront très proches. Mais si la station X et la station Y présentent des profils identiques avec décalage de trois mesures, par exemple, ceci n'apparaîtra pas. Il est donc nécessaire d'analyser des "corrélations retardées" par une autre méthode.

Cette analyse des "corrélations retardées" relève de l'étude des processus aléatoires du second ordre, et sera examinée au paragraphe suivant. Notons simplement que l'analyse factorielle, permettant de regrouper les stations proches, est un préliminaire très important à l'analyse des corrélations retardées (c'est ce qui fait, à notre connaissance, l'originalité de la méthode que nous proposons). Ce regroupement correspond à ce que nous avons baptisé "phase β".

L'analyse du graphique des observations, enfin, permettant d'étudier, en première approximation, la corrélation entre instants de mesure, est une préparation intéressante à la phase suivante.

Remarquons, pour conclure, que plusieurs techniques d'analyse factorielle se présentent à nous ; le modèle que nous avons décrit est celui de l'analyse en composantes principales (HOTELLING, 1930) ; il en existe depuis peu (BENZECRI, 1968) sous le nom d'analyse factorielle des correspondances, une variante extrêmement intéressante qui présente, entre autres, l'avantage de permettre la superposition du graphique des variables et du graphique des observations ; un ingénieux principe, baptisé par les auteurs "principe d'équivalence distributionnelle", permet d'interpréter en termes de corrélations des proximités entre variables et observations. .

4. Réduction dans le temps : processus aléatoires du second ordre

L'étude de fonctions aléatoires du second ordre (i.e. admettant des moments - moyenne et variance - aux deux premiers ordres) suppose, presque toujours, la stationnarité de ces fonctions. On entend, par stationnarité, la stabilité de la loi de probabilité dans une translation de l'axe des temps.

Si il y a stationnarité, on pourra considérer que la mesure à l'instant t et la mesure à l'instant $t + \tau$ sont stochastiquement liées de la même manière que la mesure à l'instant t' et la mesure à l'instant $t' + \tau$, c'est-à-dire que la corrélation entre deux mesures ne dépend que de l'intervalle de

temps entre ces deux mesures. On pourra alors définir, pour chaque station (assimilée, donc à une réalisation de fonction aléatoire du second ordre), une fonction d'auto-corrélation.

L'étude de cette fonction d'auto-corrélation permet de déceler le "pas de variation" d'une station de mesure, c'est-à-dire l'intervalle de temps séparant deux mesures indépendantes. Cet intervalle de temps sera baptisé, par abus de langage, "fréquence optimale". Ceci devrait donc permettre la réduction dans le temps ("phase δ").

On peut, par ailleurs, toujours à l'aide de l'hypothèse de stationnarité, définir une fonction d'inter-corrélation entre deux stations de mesures, mettant en évidence la "corrélation retardée" entre des stations de mesures distinctes, et donc achever la réduction spatiale ("phase γ").

Ce type d'analyse est extrêmement puissant, quand il est employé à bon escient. Il nous paraît nécessaire de faire à ce sujet les trois remarques suivantes :

- La validité statistique de la méthode suppose un grand nombre de mesures par station, quelques centaines au minimum.

- Une certaine tendance s'est manifestée en hydrologie, d'autre part, consistant à utiliser les techniques de l'analyse spectrale. Ces techniques, utilisant un appareil mathématique inquiétant pour le profane (transformation de FOURIER), n'apportent rien à la méthode elle-même ; leur seul avantage est de faciliter l'interprétation des résultats.

- En troisième lieu, et c'est là le point le plus délicat, il est toujours dangereux d'émettre l'hypothèse de stationnarité avec légèreté. On peut tenter de la vérifier grossièrement quand les mesures sont assez nombreuses (plusieurs milliers par station), ce qui est rarement le cas.

C'est là qu'apparaît l'intérêt du groupement effectué lors de la "phase β " : on considérera plusieurs stations d'un même groupe comme des réalisations distinctes d'une même fonction aléatoire du second ordre ; il sera alors possible, même avec un nombre d'observations limité à quelques centaines, de vérifier l'hypothèse de stationnarité, et, en cas de résultat négatif, de chercher un changement de variable permettant de s'y ramener. L'expérience a montré, en effet, que la stationnarité d'une série d'observations, en hydrométrie par exemple, dépend de beaucoup de conditions physiques d'une manière assez mal connue, et qu'il est toujours dangereux de la postuler sans vérification.

CONCLUSION

Nous espérons avoir montré l'intérêt du découpage de ce genre d'opération en quatre phases, la réduction dans l'espace précédant la réduction dans le temps, avec interaction étroite entre les deux démarches.

Malgré notre souci d'éviter toute formule mathématique, certains percevront ce programme de travail comme un schéma rigoureux et fixe, une sorte de "boîte noire" où l'on introduit des données et qui produit, en retour, des résultats.

Il ne saurait en être ainsi que en cas d'emploi maladroit : nous ne pensons pas, en effet, que l'on puisse réduire les mathématiques appliquées à l'emploi de simples recettes, si élaborées soient-elles ; nous avons tenté de formaliser des idées assez diverses, issues de cerveaux différents, chacune pour un problème différent. Il est bien évident qu'il ne peut y avoir de réduction ou d'optimisation que dans un but bien précis, et dans un contexte également précis. La manière de poser le problème dépendra de l'objectif aussi bien que du contexte : on ne traitera pas forcément de la même manière un réseau de piézomètres ou un réseau de stations hydrométriques ; l'échelle géographique du réseau aura une importance déterminante,

ainsi que le régime climatique ; on "n'optimisera" pas dans l'absolu, mais bien dans le but de tenir à jour un bilan hydrologique, ou de prévoir des crues ou des étiages, etc...

Il ne faut pas non plus s'illusionner sur la puissance de ces méthodes : elles ne valent que ce que valent des données à traiter ; elles ne sauraient détecter une structure là où il n'y en a aucune... La réduction peut porter, selon les cas, sur 10 % ou sur 80 % (le cas s'est présenté dans un réseau très structuré) des mesures.

Nous espérons seulement contribuer, par cette note qui n'a aucun caractère définitif, à une mise en ordre et à une formalisation de concepts en pleine évolution.