ECOLE POLYTECHNIQUE

TRAITEMENT STATISTIQUE DES DONNEES GEOLOGIQUES

COURS GLQ3402

(Version 2.5; 2000)

par
Denis Marcotte

1. INTRODUCTION

Au niveau du contenu théorique, ce cours se veut un complément au cours de statistique 327. Le cours 327 traitait surtout des aspects monovariables (mesures de tendance centrale et de dispersion, étude des distributions et tests d'hypothèses) et bivariables (corrélations et régression linéaire). Le cours 7.431 porte sur les techniques multivariables d'analyse des données (régression linéaire multiple, analyse en composantes principales, analyse discriminante et classifications). Les applications illustrant ces méthodes couvrent différents domaines du génie géologique: exploration géochimique, géologie appliquée au génie civil, données géophysiques et autres. Le point de vue géométrique et descriptif de ces méthodes est favorisé par rapport à une optique plus statistique et inférentielle.

1.1 Objectifs de l'analyse des données

L'ingénieur géologue est souvent confronté à des masses considérables de données. La plupart du temps plusieurs variables (attributs, caractères) ont été mesurées pour chaque observation. L'étude individuelle de ces variables, bien qu'essentielle, ne permet pas de retirer toute l'information désirée de ce tableau de données. Les méthodes bivariables, en particulier les diagrammes binaires (scatterplots en anglais), sont déjà beaucoup plus riches d'informations puisqu'elles permettent d'établir des relations entre les variables. Les méthodes multivariables vont un peu plus loin et cherchent les relations simultanées entre plusieurs variables. Ces méthodes ne sont pas le propre des ingénieurs géologues; elles peuvent servir à analyser les données de toute provenance. Elles ont été développées tout au long du 20e siècle par des statisticiens et des chercheurs provenant surtout des domaines suivants: sciences sociales, psychologie, sciences biologiques et agriculture.

L'analyse des données vise donc essentiellement à décrire les liens entre les variables et les observations de notre matrice de donnée. Plus précisément, voici une série de questions auxquelles ces méthodes tenteront de fournir des réponses:

i. Peut-on prédire le comportement d'une variable à partir d'une ou plusieurs autres variables (problème de régression)? Quelle est la meilleure équation de prédiction?
ii. Peut-on identifier, voir interpréter, des facteurs pouvant expliquer les variations observées dans les différentes variables (analyse factorielle)?
iii. Peut-on filtrer de nos données l'effet dû à des facteurs indésirables (analyse factorielle)?
iv. Peut-on identifier les différences existant entre divers groupes parmi nos données (analyse discriminante)? Comment utiliser ces différences pour prédire le groupe auquel appartient une nouvelle observation?
v. Quelles sont les observations (ou les variables) ayant des ressemblances au niveau de leur comportement (classification automatique)?

En géologie, contrairement à plusieurs autres domaines, ces variables sont habituellement mesurées en un point précis de l'espace. Certaines de ces variables peuvent être autocorrélées spatialement (rappelez-vous vos notions de géostatistique acquises en géologie minière) de sorte que l'échantillon ne peut être considéré comme étant constitué d'observations indépendantes d'une même population. Cet état de fait impose la nécessité de précautions supplémentaires lors du prélèvement de l'échantillon. En effet si l'on veut que les descriptions de notre échantillon aient quelque pertinence que ce soit en regard de la population, il conviendra d'obtenir un échantillon spatialement le plus homogène possible. On évitera en particulier les sur-représentations de zones géographiques, de types de roches, etc...Signalons au passage que d'autres méthodes (e.g. géostatistique) permettent l'étude des variables spatialement dépendantes. Ces méthodes sont encore du domaine de la recherche et dépassent le cadre de ce cours.

1.2 Place du cours dans la formation de l'ingénieur géologue; objectifs

Ce cours vise à fournir les principaux outils multivariables nécessaires à l'ingénieur géologue pour sa pratique professionnelle ou pour ses études graduées, la matière vue au cours n'étant reprise dans aucun autre cours gradué du département. Ce cours utilise les notions vues au cours 327 et au cours d'algèbre linéaire (105).

L'étudiant devrait pouvoir comprendre les articles scientifiques de son domaine utilisant ces méthodes. Il devrait maîtriser les notions théoriques suffisamment pour pouvoir utiliser de façon éclairée les logiciels d'analyse multivariable disponibles commercialement. Il devrait clairement percevoir les buts de chaque méthode. A partir d'un problème concret, il devrait pouvoir déterminer quelles sont les méthodes d'analyse les plus appropriées. Il devrait également comprendre les hypothèses et les limites propres à chaque méthode.

Le cours fournit également à l'étudiant l'occasion d'apprendre un logiciel de programmation extrêmement puissant: MATLAB. Ce logiciel, d'usage très général, lui sera d'une grande utilité tout au long de sa carrière professionnelle.

1.3 Principales statistiques d'un échantillon

Mesures de tendance centrale:

- Moyenne arithmétique:

$$\bar{x} = \frac{1}{n}\sum x_i$$

- Moyenne géométrique (G) :

$$Log(G) = \frac{1}{n}\sum \log(x_i)$$

- Moyenne harmonique (H) :

$$\frac{1}{H} = \frac{1}{n}\sum \frac{1}{x_i}$$

- Mode
  Valeur la plus fréquente (habituellement sur un histogramme pour des variables continues)
- Médiane
  Valeur centrale de l'échantillon

Mesures de dispersion

- Variance

$$s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 = \frac{1}{n-1}\sum x_i{}^2 - n\bar{x}^2$$

ou

$$\hat{\sigma}^2 = \frac{(n-1)}{n} s^2$$

- Écart-type

$$s = \sqrt{s^2}$$

- Écart inter- quartile
  Différence, dans la série ordonnée, entre l'observation correspondant au 75^{e} percentile (3^{e} quartile) et l'observation correspondant au 25^{e} percentile (1er quartile).

1.4 Principaux paramètres d'une population

Mesures de tendance centrale:

- Moyenne ou espérance mathématique:

$$\mu=E[X]$$

- Mode

  Valeur la plus probable, i.e. x_0 est le mode de f(x) la fonction de densité, si $f(x_0)=\max(f(x))$

- Médiane

  Valeur centrale de l'échantillon, i.e x_0 est la médiane si $F(x_0)=0.50$, où F(x) est la fonction de répartition.

Mesures de dispersion

- Variance

$$\sigma^2 = E[(X-\mu)^2] = E[X^2]-\mu^2$$

- Écart-type

$$\sigma = \sqrt{\sigma^2}$$

- Écart inter-quartile

$$F^{-1}(0.75)-F^{-1}(0.25)$$

  où $F^{-1}(p)$ est la fonction inverse de la fonction de répartition.