Deux algorithmes pour la classification non supervisée de données géostatistiques

Thomas Romary [1]

[1] *Mines ParisTech*
Centre de Géosciences/Géostatistique
35 rue Saint-Honoré
77305 Fontainebleau
France
thomas.romary@mines-paristech.fr

Résumé. Avec le développement des plateformes de télédétection, aéroportées ou satellites, et l'évolution des moyens d'échantillonnage des compagnies minières ou pétrolières, les jeux de données spatiales deviennent de plus en plus grands, renseignent un nombre croissant de variables et couvrent des étendues de plus en plus larges. De fait, il devient souvent nécessaire de séparer le domaine d'étude en différentes zones homogènes afin de simplifier l'étape de modélisation. La définition de ces zones peut se voir comme un problème de classification non supervisée où l'on cherche à découper le domaine d'étude en zones homogènes en fonction des variables renseignées. L'application des méthodes de classification pour des observations indépendantes ne permet généralement pas de conserver une cohérence spatiale dans les zones ainsi formées. Les algorithmes de segmentation d'image, basés sur des champs de Markov, ne sont quant à eux pas adaptées lorsque le plan d'échantillonnage n'est pas régulier (Ambroise et al., 1995 [2]). Les approches existantes (cf. *e.g.* Allard et Guillot, 2000 [1] et Guillot et al., 2006 [4]), basées sur une estimation de mélange de fonctions aléatoires gaussiennes par l'algorithme E-M, sont limitées à des tailles d'échantillon raisonnables pour un faible nombre de variables. Nous proposons dans ce travail deux algorithmes basés sur des adaptations d'algorithmes classiques, qui permettent de traiter un large volume de données. Le premier procède par classification ascendante hiérarchique tandis que le second est basé sur la méthode de classification spectrale. Les deux algorithmes sont appliqués à des jeux de données synthétiques et à un jeu de données minières.

Mots-clés. Données spatiales, classification, grands jeux de données, géostatistique

Abstract. With the increasing development of remote sensing platforms, and the evolution of sampling means in mining or oil industries spatial datasets are becoming increasingly large, inform a growing number of variables and cover areas increasingly wide. In fact, it is often necessary to split the field of study into different homogeneous areas to simplify the modeling step. The definition of these areas can be seen as a problem of unsupervised classification where we try to divide the study area into homogeneous

areas with respect to the variables at stake. The application of clustering methods for independent observations does not generally maintain a spatial coherence in the areas thus defined. Image segmentation algorithms based on Markov random fields are not adapted when the sampling is not regular (Ambroise et al., 1995 [2]). Existing approaches (see *e.g.* Allard and Guillot, 2000 [1] and Guillot et al., 2006, [4]), based on estimated mixtures of Gaussian random functions via the E-M algorithm, are limited to sample sizes reasonable for a small number of variables. We propose in this work two algorithms based on adaptations of classical algorithms, that can handle a large volume of data. The first one proceeds by agglomerative hierarchical clustering while the second is based on spectral clustering. Both algorithms are applied to toy examples and a mining dataset.

Keywords. Spatial data, clustering, large datasets, geostatistics

1 Introduction

La classification non supervisée en contexte spatial a été largement étudiée en analyse d'image ou de données de télédétection où la modélisation par champ de Markov donne de bons résultats, voir par exemple l'ouvrage de Xavier Guyon (1995) [5], et Celeux et al. 2003 [3].
La classification de données issues d'un échantillonnage spatial non régulier a été en revanche moins étudié. Il peut s'agir par exemple de données minières, où l'on dispose de sondages verticaux à l'échelle d'un gisement minier. Dans ce cadre, les approches par champs de Markov sont peu adaptées du fait que le graphe sur lequel un tel modèle pourrait être appliqué dépend fortement de l'échantillonnage. Si l'on souhaite appliquer une approche similaire, via l'algorithm EM, le modèle probabiliste le plus adapté à ces données est celui de mélanges de fonctions aléatoires (cf. Allard et Guillot, 2000) qui se prête peu au traitement de grands jeux de données multidimensionnelles : le calcul de la vraisemblance nécessite l'inversion d'une matrice de covariance de très grande taille, dense qui plus est. L'intégration de données catégorielles devient également délicate dans ce cadre.

2 Algorithmes

Les algorithmes de classification non supervisée que nous proposons sont tous deux basés sur la même idée. Elle consiste à structurer les données disponibles selon un graphe (Delaunay ou basé sur des voisinages) fait d'une unique structure connexe. L'algorithme de classification consiste alors à partitionner ce graphe en autant classe que désiré. La structure ainsi imposée aux données permet de s'assurer de la cohérence spatiale des classe obtenues : elles sont en effet connexes (au sens du graphe) par construction.
Dans ce cadre, deux algorithmes sont proposés : l'un procède de manière hiérarchique ascendante et le second résulte d'une adaptation de l'algorithme de classification spectrale.

Ils sont présentés ci-dessous. Tous deux dépendent également du choix d'une distance, adaptée au cas multivarié. Pour un échantillon $(x_1, \ldots, x_n) \in \mathbb{R}^(n \times p)$, avec p le nombre de variables, on définit la distance d :

$$d(x_i, x_j) = \sum_{k=1}^{p} w_k d^{(k)}(x_i^{(k)}, x_j^{(k)}).$$

d est ainsi une somme pondérée de distances individuelles qui sont choisies en fonction de la variable associée : si celle-ci est quantitative, on utilise généralement la distance euclidienne au carré, s'il sagit d'une variable catégorielle, une distance adaptée est employée, par exemple une distance qui prend la valeur 0 lorsque les deux observations prennent deux valeurs égales pour cette variable et 1 sinon. L'algorithme de classification hiérarchique géostatistique (CHG) prend la forme suivante :

1. *Former la matrice de distance* $D \in \mathbb{R}^{n \times n}$, *telle que* $D_{ij} = d(x_i, x_j),\ j < i$, *si* $i \leftrightarrow j$, *0 sinon,*

2. *Trouver* $k < l$ *tels que* $D_{kl} = \min_{\{i,j,i \leftrightarrow j\}} D_{ij}$

3. *Fusionner* k *et* l *en* $\{kl\}$, *et mettre* D *à jour selon*

$$D_{ki} = \max(D_{ki}, D_{li}) \text{ si } i \leftrightarrow \{kl\} \text{ et } i < k$$

$$D_{ik} = \max(D_{ki}, D_{li}) \text{ si } i \leftrightarrow \{kl\} \text{ et } k > i$$

   supprimer la ligne et la colonne l *de* D

4. *Répéter 2 et 3 jusqu'à ce que* D *ne contienne plus qu'un élément.*

L'algorithme de classification spectrale géostatistique (CSG) nécessite de donner a priori un nombre K de classes. Il prend la forme suivante :

1. *Former la matrice de similarité* $W \in \mathbb{R}^{n \times n}$, *telle que* $W_{ij} = \exp\left(-\frac{d(x_i,x_j)}{\sigma^2}\right)$, *si* $i \leftrightarrow j$, *0 sinon,*

2. *Former la matrice diagonale* D *telle que* $D_{ii} = \sum_{j=1}^{n} W_{ij}$

3. *Calculer la matrice Laplacienne du graphe*

$$L = D^{-1/2} W D^{-1/2}$$

4. *Calculer les* K *plus grandes valeurs propres de* L *et former la matrice* $V \in \mathbb{R}n \times K$ *dont les colonnes sont les* K *premiers vecteurs propres de* L

5. *Normaliser les lignes de* V

6. *Appliquer l'algorithme des K-means aux lignes de V*

7. *Assigner le point x_i à la classe à laquelle la ligne i de V l'a été.*

L'efficacité numérique de ces deux algorithmes repose sur le nombre de connexions du graphe utilisé. Dans le cas du CHG, seules les distances entre points connectés sont nécessaire au lancement de l'algorithme, les distances entre points non connectés nécessaires lors de l'étape 3 de l'algorithme peuvent être calculées à la volée. En ce qui concerne le CSG, l'utilisation du graphe induit une structure creuse dans toutes les matrices impliquées dans l'algorithme, ce qui permet d'employer les algorithmes d'algèbre linéaire adaptée.

3 Résultats

Les algorithmes ci-dessus ont été appliqués à un jeu de données minières. La première étape consiste à sélectionner les variables à considérer dans la classification :

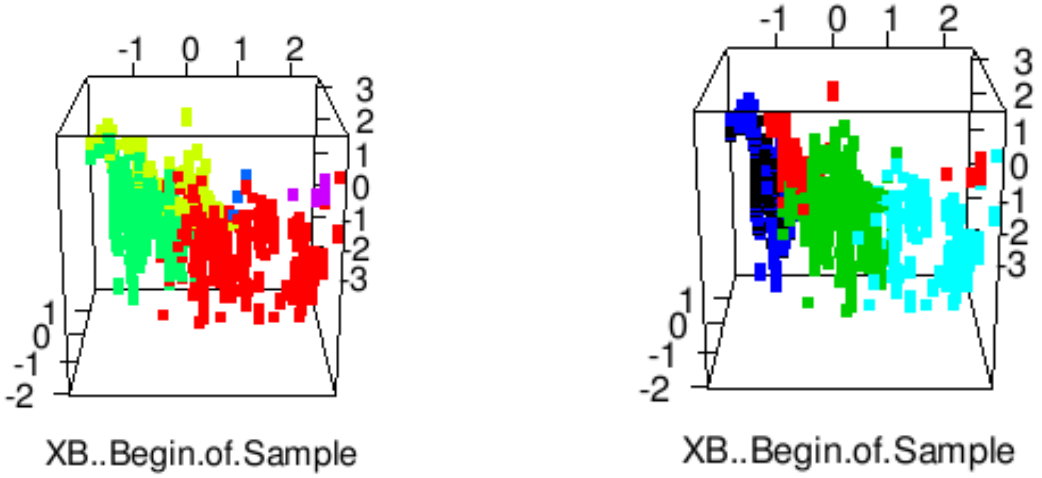
- les coordonnées,
- la teneur en uranium,
- un facteur géologique décrivant le socle,
- le degré d'hématisation (altération)

Le choix de ces variables a été guidé par une analyse exploratoire préliminaire et des discussions avec les géologues. Les variables ont été ensuite transformées :

- les coordonnées ont été normalisées
- le logarithme des teneurs a été considéré puis normalisé
- le degré d'hématisation a été considéré comme une variable continu (variable ordinale).

Les deux algorithmes ont ensuite été lancés sur ce jeu de données avec les mêmes pondérateurs pour le calcul de la distance. Les résultats obtenus pour 5 classes sont présentés en figure 1.

On peut voir certaines similarités entre les classifications obtenues par les deux algorithmes. En particulier, elles présentent toutes deux les propriétés de connexité désirées. Elles isolent également toutes deux un ensemble de points au sud est du gisement qui correspond à de fortes teneurs se trouvant dans le socle.



a. *b.*

Figure 1: Classifications obtenues: CHG *a.* et CSG *b.*

4 Discussion

Deux algorithmes de classification de données géostatistiques ont été présentés. Leur principal avantage est de générer des classes connexes tout en permettant, par construction, de traiter un grand volume de données spatiales multivariées.
De nombreux points restent à étudier, en particulier en ce qui concerne le choix des pondérateurs de la fonction de distance ou le paramètre d'échelle de l'algorithme de classification spectrale.

References

[1] Allard, D., and Guillot, G. Clustering geostatistical data. In *Proceedings of the sixth geostatistical conference* (2000).

[2] Ambroise, C., Dang, M., and Govaert, G. Clustering of spatial data by the EM algorithm. In *geoENV I - Geostatistics for Environmental Applications* (1995), A. S. et al., Ed., Kluwer Academic Publishers, pp. 493–504.

[3] Celeux, G., Forbes, F., and Peyrard, N. Em procedures using mean field-like approximations for markov model-based image segmentation. *Pattern recognition 36*, 1 (2003), 131–144.

[4] Guillot, G., Kan-King-Yu, D., Michelin, J., and Huet, P. Inference of a hidden spatial tessellation from multivariate data: application to the delineation of homogeneous regions in an agricultural field. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 55*, 3 (2006), 407–430.

[5] Guyon, X. *Random fields on a network*. Springer, 1995.