Comparaison de deux modélisations pour la classification de données géostatistiques

Nathalie Peyrard[1], Denis Allard[1] & Florence Forbes[2]

[1] *Unité Biométrie, INRA*
Domaine Saint Paul, Site Agroparc, 84914 AVIGNON cedex 9
[2] *INRIA Rhone-Alpes*
ZIRST, 655, avenue de l'Europe Montbonnot 38334 Saint Ismier Cedex

Résumé : L'objectif de la classification de données spatiales est de regrouper des mesures observées en différents points du plan en tenant compte de l'information spatiale. Nous comparons deux modélisations pour résoudre ce problème : le modèle de champ de Markov caché (CMC) et un modèle géostatistique basé sur un champ gaussien. L'estimation des paramètres des modèles est obtenue avec un algorithme de type EM. Si des approximations de EM existent pour le modèle CMC, ce n'est pas le cas pour le modèle géostatistique et nous proposons une approche mimant la solution existante dans le cas indépendant. Les performances des deux modèles sont comparées sur des données simulées et réelles : elles montrent la supériorité du modèle géostatistique en terme d'erreur de classification mais les classifications issues du modèle CMC ont l'avantage d'être plus lisses.
Mots-Clés : champ de Markov caché, champ aléatoire gaussien, algorithme EM.

Abstract : Clustering spatial data consists in grouping observations measured at differents points of the plane, taking into account the spatial information. We compare two models to solve this problem: hidden Markov random fields (HMRF) and a geostatistical model which uses Gaussian random fields. Parameters estimation is performed with an EM-like algorithm. If approximations of this algorithm exist for HMRF, this is not the case for the geostatistical model. We propose an approach mimicing the solution in the independent case. Models performances are compared on simulated and real data: they show the superiority of the geostatistical model in terms of classification error, but the classifications obtained with the CMC model are smoother.
Key Words: hidden Markov random field, Gaussian random field, EM algorithm.

1 Introduction

Considérons une région A du plan, partitionnée en K sous-régions inconnues ($A = \cup_k A_k$ et $A_k \cap A'_k = 0$ si $k \neq k'$). Des données univariées $\mathbf{y} = \{y_1, \ldots, y_n\}$ sont mesurées sur un ensemble $S = \{s_1, \ldots, s_n\} \in A^n$ de sites. L'objectif de la classification de données spatiales est d'estimer le vecteur $\mathbf{z} = \{z_1, \ldots, z_n\}$ codant les régions d'appartenance de chacun des sites, au vu des observations $\mathbf{y}$ et en tenant compte de l'information spatiale. Nous comparons ici deux modélisations qui permettent d'intégrer la composante spatiale du problème à deux niveaux différents. Dans un modèle de champ de Markov caché,

l'information spatiale est prise en compte au niveau des classes des observations. Cette hypothèse est classique en analyse d'image, domaine où ce modèle est largement utilisé. Par contre, ce cadre n'est pas celui adopté généralement dans le cas de données géostatistiques (i.e. irrégulièrement réparties) car il dépend fortement du système de voisinage choisi. Si un point du schéma d'échantillonnage est légèrement déplacé, cela peut entrainer la modification de son voisinage, et donc des lois locales, sans que cela reflète une réalité physique. Un modèle gaussien avec une structure de covariance est en général préféré (Allard et Guillot 2000). Avec ce second modèle, la dépendance spatiale est prise en compte uniquement au niveau des observations, alors que les classes sont indépendantes.
En classification, les paramètres du modèle sont généralement inconnus et une étape d'estimation est nécessaire. L'algorithme EM, adapté à l'estimation du maximum de vraisemblance pour des problèmes avec données manquantes, est largement utilisé dans ce contexte. Lorsque les modèles sont spatiaux, EM n'est pas applicable directement, et des approximations sont nécessaires. Si des solutions existent pour les modèles CMC (Chalmond 1989, Celeux et al. 2003), ce n'est pas le cas pour le modèle géostatistique. Nous proposons un algorithme EM approché, Geo-EM, basé sur les formules de EM pour les modèles de mélange indépendants. Les performances du modèle géostatistique associé à cet algorithme sont comparées à celles du modèle CMC sur des données simulées et réelles.

2 Modélisation en classification de données spatiales

2.1 Champ de Markov caché

Un champ de Markov discret est une collection de variables aléatoires discrètes $\mathbf{Z} = \{Z_i,\ i \in S\}$, chaque Z_i étant à valeur dans $\{1, \dots, K\}$, dont la distribution jointe est une distribution de Gibbs, $P(\mathbf{z}) = W^{-1} \exp(-\sum_c V_c(\mathbf{z}_c))$. La somme est sur l'ensemble des cliques définies par le système de voisinage choisi et les V_c sont des potentiels permettant de favoriser certains types de configurations. La constante de normalisation du modèle est égale à $W = \sum_{\mathbf{Z}} \exp(-\sum_c V_c(\mathbf{z}_c))$. Dans le modèle CMC, les données observées $\mathbf{Y} = \{Y_i,\ i \in S\}$ sont supposées indépendantes conditionnellement à la connaissance de $\mathbf{Z}$, $f(\mathbf{y} \mid \mathbf{z}) = \prod_{i \in S} f(y_i \mid z_i)$. Nous considérerons ici le cas où les densités conditionnelles $f(y_i \mid z_i)$ sont gaussiennes, de sorte que si $z_i = k$ alors $f(y_i \mid k)$ est la densité $g(y_i \mid \mu_k, \sigma_k)$ d'une loi gaussienne d'espérance μ_k et d'écart type σ_k.

2.2 Champ aléatoire gaussien

En géostatistique, un modèle avec une structure de covariance est souvent utilisé. Les observations $\mathbf{Y}$ sachant les classes sont modélisées par une fonction aléatoire $Y(s)$, mosaïque de K fonctions aléatoires stationnaires gaussiennes indépendantes sur chaque région A_k. Avec un tel modèle, nous avons, pour toute paire de sites s_i et s_j dans la classe k,

$E[Y_i] = E[Y_j] = \mu_k$, $\mathrm{Cov}(Y_i, Y_j) = \sigma_k^2 \rho_k(s_i - s_j)$. Nous utiliserons la fonction de corrélation $\rho_k(h) = \exp(-\frac{|h|}{b_k})$. Pour deux sites qui n'appartiennent pas à la même classe, $\mathrm{Cov}(Y_i, Y_j) = 0$. La densité jointe conditionnelle s'écrit alors $f(\mathbf{y} \mid \mathbf{z}) \propto e^{\frac{-1}{2}(\mathbf{y}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}$ avec μ le vecteur des espérances de chaque point, et $\boldsymbol{\Sigma} = \sum_k \boldsymbol{\Sigma}_k = \sum_k \sigma_k^2 R_k$. La matrice R_k exprime les coefficients de la matrice de corrélation en fonction des classes des sites considérés :

$$R_k[i,i] = t_{ik}, \quad R_k[i,j] = t_{ik}.t_{jk}.\rho_k(s_i - s_j) \text{ si } i \neq j. \tag{1}$$

Le vecteur $(t_{ik}, \forall 1 < k < K)$ est une autre façon de représenter l'information contenue dans z_i : $t_{ik} = 1$ si le site s_i est dans la classe k et 0 sinon. Enfin, les variables représentant les labels sont supposées i.i.d. : $P(\mathbf{z}) = \prod_i P(z_i)$, et l'on notera $\pi_k = P(Z_i = k)$.

3 Algorithme EM pour modèle géostatistique

3.1 Algorithme Geo-EM

Une mise en œuvre directe de EM n'est pas possible pour le modèle géostatistique : cela nécessiterait de nombreuses inversions de matrice et sommes sur un espace d'état trop large. Nous nous inspirons ici de la solution existante dans le cas des modèles de mélange indépendants. Pour un tel modèle, de poids des classes π_k et de lois conditionnelles $f(y_i \mid k)$ gaussiennes d'espérance μ_k et de variance σ_k^2, les formules de EM de mise à jour des paramètres à l'itération q sont les suivantes :

$$\pi_k^q = \frac{\sum_i t_{ik}^q}{n}, \quad \mu_k^q = \frac{\sum_i t_{ik}^q y_i}{\sum_i t_{ik}^q}, \quad (\sigma_k^q)^2 = \frac{\sum_i t_{ik}^q (y_i - \mu_k^q)^2}{\sum_i t_{ik}^q} \tag{2}$$

$$t_{ik}^{q+1} = \frac{\pi_k^q g(y_i \mid \mu_k^q, \sigma_k^q)}{\sum_{l=1}^K \pi_l^q g(y_i \mid \mu_l^q, \sigma_l^q)} \tag{3}$$

La quantité t_{ik}^q est l'estimation courante de la probabilité a posteriori que le site s_i soit dans la classe k sachant l'observation y_i. L'idée conduisant à Geo-EM est de prendre en compte les dépendances entre les y_i d'une manière similaire à la pseudo-vraisemblance, en remplaçant dans la formule (3) les densités gaussiennes par $g(y_i | \mathbf{y}_{-i}, t_{ik} = 1, \mathbf{t}_{-i} = \mathbf{t}_{-i}^q, \theta^q)$, avec $\mathbf{t}_{-i} = \{t_{jk}, \forall k, \forall j \neq i\}$, $\mathbf{t}_{-i}^q = \{t_{jk}^q, \forall k, \forall j \neq i\}$ et $\mathbf{y}_{-i} = \mathbf{y} \setminus \{y_i\}$. La densité $g(y_i | \mathbf{y}_{-i}, t_{ik} = 1, \mathbf{t}_{-i} = \mathbf{t}_{-i}^q, \theta^q)$ est la densité d'une loi gaussienne d'espérance m_{ik}^q et d'écart type σ_{ik}^q donnés par les formules suivantes :

$$m_{ik}^q = \mu_k^q + (\Sigma_{ik}^q)^t (\boldsymbol{\Sigma}_{-i}^q)^{-1} (\mathbf{y}_{-i} - \eta_{-i}^q), \; (\sigma_{ik}^q)^2 = (\sigma_k^q)^2 - (\Sigma_{ik}^q)^t (\boldsymbol{\Sigma}_{-i}^q)^{-1} (\Sigma_{ik}^q)$$

où $\eta_{-i}^q = \{\eta_j^q = \sum_{k=1}^K t_{jk}^q \mu_k^q, \forall j \neq i\}$, $\Sigma_{ik}^q = \{t_{jk}^q.(\sigma_k^q)^2.\rho_k^q(s_i - s_j), \; \forall j \neq i\}$ et $\boldsymbol{\Sigma}_{-i}^q$ est la matrice $\boldsymbol{\Sigma}^q = \sum_k (\sigma_k^q)^2 R_k^q$ privée de la ligne et de la colonne i (cf equation (1), les t_{ik} étant remplacés par les t_{ik}^q). Une fois les t_{ik}^q ainsi mis à jour, les poids, espérances et écarts type sont mis à jour selon (2). Enfin, pour la mise à jour du paramètre b_k de la fonction de corrélation, nous utilisons la méthode classique des moindres carrés pondérés (WLS, Cressie 1993) pour l'estimation du variogramme.

3.2 Variante

L'étape de mise à jour des t_{ik}^q est très demandeuse en temps de calcul. En effet, les évaluations de m_{ik}^q et de σ_{ik}^q nécessitent l'inversion de n matrices de taille $(n-1)\times(n-1)$ à chaque itération. Nous proposons une variante pour cette étape, qui permet de se ramener à une seule inversion d'une matrice $n \times n$ à chaque itération. Pour cela, la mise à jour des t_{ik}^q est remplacée par $t_{ik}^{q+1} = \pi_k^q g(y_i|\mathbf{y}_{-i}, \mathbf{t} = \mathbf{t}^q, \theta^q)/\sum_l \pi_l^q g(y_i|\mathbf{y}_{-i}, \mathbf{t} = \mathbf{t}^q, \theta^q)$, où $\mathbf{t}^q = \{t_{jl}^q, \forall i, \forall l\}$. La différence est que l'on conditionne par rapport à $t_{ik} = t_{ik}^q$, qui est flou, au lieu de $t_{ik} = 1$. $g(y_i|\mathbf{y}_{-i}, \mathbf{t} = \mathbf{t}^q, \theta^q)$ est toujours la densité d'une loi gaussienne, d'espérance $\tilde{m}_{ik}^q$ et d'écart type $\tilde{\sigma}_{ik}^q$. Ces paramètres sont donnés par les formules suivantes

$$\tilde{m}_{ik}^q = \mu_k^q + \sum_{j\neq i} \lambda_{ij}(y_j - \eta_j^q), \quad (\tilde{\sigma}_{ik}^q)^2 = \lambda_{ii} \ \forall i, \ \forall k$$

avec $\lambda_{ii} = \frac{1}{(\mathbf{\Sigma}^q)^{-1}[i,i]}$ et $\lambda_{ij} = \frac{-(\mathbf{\Sigma}^q)^{-1}[i,j]}{(\mathbf{\Sigma}^q)^{-1}[i,i]}$ si $i \neq j$. D'autre part des résultats sur données simulées ont montré que l'algorithme Geo-EM tend à surestimer la variance $(\sigma_k)^2$. Aussi, la mise à jour de l'estimateur de la variance est remplacée par $(\sigma_k^q)^2 = \frac{\sum_i t_{ik}^q (y_i - \eta_i^q)^2}{\sum_i t_{ik}^q}$. Les points tels que η_i^q n'est pas proche de μ_k^q voient ainsi leur influence diminuer. L'algorithme correspondant sera désigné par Geo-EM(A).

4 Résultats

Les performances des deux modélisations sont comparées sur différents types de données. Pour la classification par modèle CMC, le champ de Markov $\mathbf{Z}$ est un modèle de Potts à K couleurs avec champ externe défini par $P(\mathbf{z}|\beta, \mathbf{v}) = W(\beta, \mathbf{v})^{-1} e^{\sum_{i\in S} z_i^t \mathbf{V} + \beta \sum_{j\sim i} z_i^t z_j}$ (cette écriture suppose que Z_i est représenté par un vecteur à K composantes, toutes nulles sauf celle de la classe du site s_i). Le voisinage d'un site s_i est défini comme l'ensemble des sites de mesure contenus dans un disque de centre s_i et de rayon prédéfini r. Pour l'estimation des paramètres nous avons utilisé l'algorithme *Simulated Field* EM (SF-EM) présenté dans Celeux et al. (2003). Cet algorithme est dérivé du principe d'approximation en champ moyen d'une distribution de Gibbs.

4.1 Données simulées

Deux types de données ont été simulées : des réalisations du modèle géostatistique et des réalisations d'un modèle dit *complet*. Dans le modèle *complet* le champ des classes, $\mathbf{Z}$, est

modèle simulé						*Err*		
β	v_1	v_2	b_1	b_2	μ_2	Geo-EM	Geo-EM(A)	SF-EM
0	0	0	0.1	0.1	2	28.9	15.8	23.1
0	0	0	0.1	0.1	3	14.6	7.3	8.6
0	0	0	0.2	0.2	3	11.5	6.7	10
0.15	0	0	0.1	0.1	2	33.6	16.4	26.2
0.15	0	0	0.1	0.1	3	16.5	7.4	10.4
0.15	0	0	0.2	0.2	3	14.1	6.8	11.8
0	1	3	0.1	0.1	2	-	22.3	29.8
0	1	3	0.1	0.1	3	-	6.7	10.6

(a)

modèle simulé			*Err*	
v_1	v_2	v_3	Geo-EM(A)	SF-EM
1	1	1	16	21.7
1	2	3	26.5	53

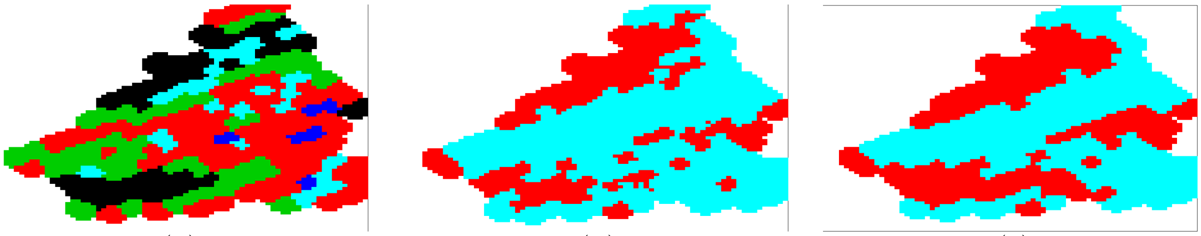
(b)

Table 1: Erreurs de classification pour différentes simulations : (a) $K = 2$, $\mu_1 = 0, \sigma = 1$, (b) $K = 3$, $\beta = 0, b_1 = b_2 = b_3 = 0.1, \mu_1 = 0, \mu_2 = 2, \mu_3 = 5, \sigma = 1$.

modélisé par un champ de Markov discret (ici le modèle de Potts avec champ externe) et le champ des observations, $\mathbf{Y} \mid \mathbf{Z}$, est modélisé par un champ aléatoire gaussien. Des simulations de ce modèle, plus riche que le CMC ou le modèle géostatistique, devraient mieux refléter la complexité présente dans des données réelles. Elles permettent donc de tester les deux méthodes de classification dans des situations a priori moins favorables. Les résultats en terme de taux d'erreur moyen (*Err*, pourcentage de pixels mals classés en moyenne sur 100 simulations) sont reportés dans les Tables 1 (a) et (b). Dans tous les tests $\sigma_k = \sigma$, identique pour toutes les classes. Lorsque le paramètre β est non nul, le modèle simulé correspond au modèle *complet* et lorsque $\beta = 0$ il s'agit du modèle géostatistique. Le vecteur de paramètres $\mathbf{v}$ joue sur la proportion des points dans chaque classe : plus ses coefficients sont distincts, plus les proportions sont tranchées. Ces résultats montrent la supériorité du modèle géostatistique (associé à Geo-EM(A)) quelle que soit la complexité des observations simulées.

4.2 Données réelles

Nous avons appliqué les deux modèles à la classification d'un jeu de données géologiques provenant du Jura suisse (Atteia et al. 1994). Un total de 359 positions de mesure ont été échantillonées et en chacun de ces points la concentration en six métaux, Co, Cr, Cu, Ni, Pb et Zn, a été mesurée. Par ailleurs une classification de la zone en terme de formations géologiques est disponible, notamment suivant les deux classes argovien/non argovien. La question se pose de l'existence ou non d'une corrélation entre la formation géologique et les concentrations en différents métaux. Les classifications obtenues avec les deux modèles (voir Figure 1 pour le cas du Co) paraissent aller dans le sens d'une telle corrélation. On constate cependant que quelquesoit le métal utilisé la classification obtenue avec le modèle CMC, du fait d'un modèle spatialisé sur les classes, a toujours un rendu visuel plus lisse que la classification obtenue avec le modèle géostatistique.



(a) (b) (c)

Figure 1: Classification du jeu de données Jura : (a) classification géologique (argovien en noir), (b) classification par Geo-EM(A), (c) classification par SF-EM.

5 Discussion

Nous avons présenté deux approches différentes de modélisation spatiale pour la classification de données géostatistiques : les champs de Markov cachés et le modèle géostatistique basé sur un champ gaussien. Au vu des résultats obtenus, le premier assure des classifciations lisses et le second est plus robuste lorsque la structure des données se complexifie. Ainsi les modèles CMC et géostatistiques apparaissent comme complémentaires : la complexité des mécanismes mis en jeu fait qu'en général l'hypothèse d'indépendance des observations sachant les classes n'est pas très réaliste. De la même manière, les régions A_k recherchées vont souvent être structurées spatialement et l'hypothèse d'indépendance des classes peut être restrictive. Le modèle *complet*, avec dépendence spatiale des classes et des observations est clairement plus réaliste et devrait permettre d'intégrer les bonnes propriétés observées chez les deux modèles. Ni l'algorithme SF-EM, ni l'algorithme Geo-EM ne sont adaptés au modèle complet et notre objectif est maintenant le développement d'une version de EM pour ce modèle.

Bibliographie

[1] Allard, D. and G. Guillot (2000, April). Clustering geostatistical data. In *Proceedings of the Sixth Geostatistical Conference*, Cape-Town South Africa.

[2] Atteia, O., J.-P. Dubois, and R. Webster (1994). Geostatistical analysis of soil contamination in the Swiss Jura. *Environmental Pollution 86*, 315–327.

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

[4] Chalmond, B. (1989). An iterative Gibbsian technique for reconstruction of m-ary images. *Pattern Recognition 22*(6), 747–761.

[5] Cressie, N. (1993). *Statistics for Spatial Data, revised edition.* New York: Wiley.