Comparaison de deux modélisations pour la classification de données géostatistiques

Nathalie Peyrard[1] *Denis Allard*[1] *et Florence Forbes*[2]

[1] *Institut National de la Recherche Agronomique, Avignon, France*

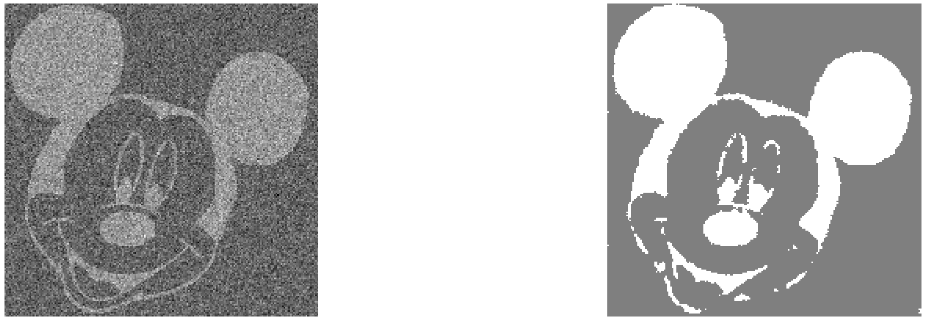
[2] *Institut National de la Recherche en Informatique et Automatique, Grenoble, France*

`{peyrard,allard}@avignon.inra.fr, florence.forbes@inrialpes.fr`

XXXVIIèmes Journées de Statistique, Pau, 6-10 juin 2005

Introduction

- Classification de données spatiales :
  - Des données observées en plusieurs sites sur un domaine
  - Domaine composé de K sous domaines correspondant à des processus différents de génération des données
  - Problème : à partir des données, retrouver les K sous domaines en prenant en compte l'information spatiale
- Exemple



- Cas des données géostatistiques
  - Sites d'observation irrégulièrement répartis dans l'espace
  - Dans chaque sous domaine, les observations ne sont pas indépendantes
  - Nombre d'observations faible à modéré ($\sim$ 100 – 500)
- Exemple

- Formalisation statistique
  - n sites d'observation $s_1, \ldots, s_n$
  - X_i, variable aléatoire, valeur de la donnée au site s_i

    $\mathbf{X} = \{X_1, \ldots, X_n\}$: **champ observé**
  - K groupes/sous domaines (K supposé connu ici)
  - Z_i, variable aléatoire, groupe auquel appartient le site i

$$\begin{aligned} Z_{ik} &= 1 \quad \text{si } s_i \text{ est dans le groupe } k, \\ Z_{ik} &= 0 \quad \text{sinon} \end{aligned}$$

$\mathbf{Z} = \{Z_1, \ldots, Z_n\}$: **champ caché**

  - Vraisemblance complète

$$p(\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = p(\mathbf{X} = \mathbf{x} \mid \mathbf{Z} = \mathbf{z}) p(\mathbf{Z} = \mathbf{z})$$

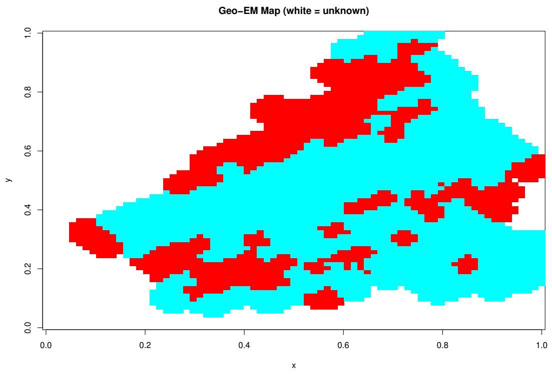
Choix de modélisation

- Où est-il pertinent d'intégrer l'information spatiale?
  - au niveau du champ caché $\mathbf{Z}$: champ de Markov caché (CMC)
  - au niveau du champ observé $\mathbf{X}$: champ gaussien (CG)
  - aux deux niveaux $\mathbf{Z}$ et $\mathbf{X}$: modèle complet (MC)
- Le modèle complet sera notre modèle de référence, supposé représenter le mieux la réalité

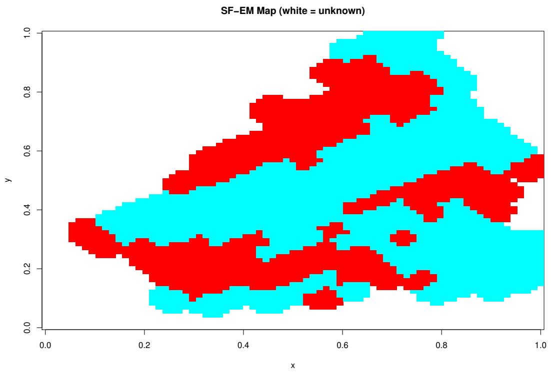
Illustration

- Données géologiques du Jura Suisse (Atteia, Dubois, Webster 1994)
  - 359 sites d'observation
  - en chaque site, mesure de la concentration en Co, Ni ...

Co, modèle CG



Co, modèle CMC



Modèle champ de Markov caché

- champ de Markov pour la loi de Z

$$p(\mathbf{Z} = \mathbf{z}) \propto \exp\{-\sum_{C} \Phi_C(z_c; \beta)\}$$

  où C sont les cliques associées au système de voisinage

- indépendance conditionnelle des observations $\mathbf{X}$

$$p(\mathbf{X} = \mathbf{x} \mid \mathbf{Z} = \mathbf{z}) = \prod_{i} p(X_i = x_i \mid Z_i = z_i, \theta)$$

Généralement utilisé pour des **grilles régulières** (analyse d'image)

Dans le cas de données échantillonnées irrégulièrement, les lois locales sont modifiées par le déplacement d'un point du voisinage.

- Estimation des paramètres : algorithme EM
  calcul de l'estimateur du maximum de vraisemblance dans le cas de données manquantes → algorithme itératif en deux étapes.
  Soit $\Psi = (\theta, \beta)$

  (E) calcul de

  $$Q(\Psi \mid \Psi^{(q-1)}) \quad = \quad E[\log p(\mathbf{x}, \mathbf{Z} \mid \Psi) \mid \mathbf{x}, \Psi^{(q-1)}]$$

  (M) $\Psi^{(q)} \; = \; \arg\max_{\Psi} Q(\Psi \mid \Psi^{(q-1)})$

Complexité calculatoire : pas de mise en œuvre directe possible

- algorithme Simul-Field EM : on se ramène à un modèle de mélange indépendant

  Itération q:

  (1) Génération d'une configuration pour le voisinage $\tilde{\mathbf{z}}^{(q)}$, from x and $\Psi^{(q-1)}$

  $$p(\mathbf{z} \mid \beta) \approx \prod_i p(z_i | \, \tilde{z}^{(q)}_{N(i)}, \beta)$$

  (2) EM pour le modèle défini par l'approximation $\longrightarrow \Psi^{(q)}$

  $$p(\mathbf{z}, \mathbf{x} \mid \Psi) \approx \prod_i p(x_i | \, z_i, \theta) p(z_i | \, \tilde{z}^{(q)}_{N(i)}, \beta)$$

Modèle champ gaussien

- Indépendance des classes $\mathbf{Z}$: $p(\mathbf{z}) = \prod_i p(z_i)$, $p(Z_i = k) = \pi_k$
- Dépendance des X_i intra classes mais indépendance inter classes:
  - Pour s_i et s_j dans la classe k,

    $$E[X(s_i)] = \mu_k, \quad \mathbf{\Sigma}_k[ij] = \text{Cov}(X(s_i), X(s_j)) = \sigma_k^2 \rho_k(s_i - s_j),$$
  - Pour s_i et s_j dans des classes différentes

    $$\text{Cov}(X(s_i), X(s_j)) = 0$$
  - produit de K champs gaussiens indépendants pour $(\mathbf{X} \mid \mathbf{Z})$

    $$p(\mathbf{x} \mid \mathbf{z}) \propto e^{\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu})^t \mathbf{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})}, \quad \mathbf{\Sigma} = \sum_k \sigma_k^2 R_k$$

    $$R_k[i,i] = z_{ik}, \quad R_k[i,j] = z_{ik}.z_{jk}.\rho_k(s_i - s_j) \text{ si } i \neq j$$
- Pas d'algorithme de classification satisfaisant

Modèle complet

- champ de Markov pour $\mathbf{Z}$
- produit de K champs gaussiens indépendants pour $(\mathbf{X} \mid \mathbf{Z})$

- modèle de référence pour la comparaison de CMC et CG

Algorithme d'estimation du modèle CG

- la mise en œuvre directe de EM est impossible
- approche adoptée : s'inspirer des formules de EM pour mélange indépendant ...

  1. mise à jour de l'estimation de la probabilité a posteriori que s_i soit dans la classe k (classification floue)

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

  2. mise à jour des paramètres

  $$\pi_k^{q+1} = \frac{\sum_i t_{ik}^{q+1}}{n}, \quad \mu_k^{q+1} = \frac{\sum_i t_{ik}^{q+1} x_i}{\sum_i t_{ik}^{q+1}}, \quad (\sigma_k^{q+1})^2 = \frac{\sum_i t_{ik}^{q+1} (x_i - \mu_k^{q+1})^2}{\sum_i t_{ik}^{q+1}}$$

- ... et rajouter du spatial en modifiant

$$g(x_i \mid \mu_k^q, \sigma_k^q)$$

Algorithme Geo-EM

1. mise à jour des t_{ik}:

$$t_{ik}^{q+1} = \frac{\pi_k^q g(x_i \mid x_{-i}, t_{ik} = 1, t_{-i}^q, \theta^q)}{\sum_l \pi_l^q g(x_i \mid x_{-i}, t_{il} = 1, t_{-i}^q, \theta^q)},$$

où $g(x_i \mid x_{-i}, t_{ik} = 1, t_{-i}^q, \theta^q)$ est la densité conditionnelle de x_i, sachant que s_i est dans la classe k et sachant tous les autres x_j et t_{jk} flous
$\rightarrow$ densité de loi gaussienne d'espérance m_{ik}^q et d'écart type σ_{ik}^q

$$m_{ik} = \mu_k^q + (\mathbf{\Sigma}_{ik}^q)^t (\mathbf{\Sigma}_{-i}^q)^{-1} (x_{-i} - \eta_{-i}^q)$$

$$(\sigma_{ik}^q)^2 = (\sigma_k^q)^2 - (\mathbf{\Sigma}_{ik}^q)^t (\mathbf{\Sigma}_{-i}^q)^{-1} \mathbf{\Sigma}_{ik}^q,$$

avec

$$\eta^q[j] = \sum_k z_{jk}^q \mu_k^q \quad \text{et} \quad \mathbf{\Sigma}_{ik}^q = i\text{ème colonne de } \mathbf{\Sigma}_k^q$$

Algorithme Geo-EM (suite)

2. mise à jour des paramètres:
   - comme EM pour mélange indépendant pour π_k^q, μ_k^q et σ_k^q
   - méthode des moindres carrés pondérés sur le variogramme pour l'estimation du paramètre de la fonction de corrélation ρ_k

Variante Geo-EM(A)

1. mise à jour des t_{ik}:
   méthode exacte : pour chaque itération, inversion de n matrices de taille $(n-1) \times (n-1)$
   on remplace la densité conditionnelle de x_i par

$$t_{ik}^{q+1} = \frac{\pi_k^q g(x_i \mid x_{-i}, t = t^q, \theta^q)}{\sum_l \pi_l^q g(x_i \mid x_{-i}, t = t^q, \theta^q)}$$

2. mise à jour des paramètres:
   - comme GeoEM pour π_k^q, μ_k^q et le paramètre de la fonction de corrélation
   - pour σ_k^q, surestimation de la variance avec la méthode exacte

$$(\sigma_k^q)^2 = \frac{\sum_i t_{ik}^q (x_i - \eta_i^q)^2}{\sum_i t_{ik}^q}$$

Résultats de classification sur données simulées

- modèle CMC : $\mathbf{Z}$, Potts à K couleurs avec champ externe

$$p(\mathbf{z}|\beta, \mathbf{v}) = W(\beta, \mathbf{v})^{-1} e^{\sum_{i \in S} z_i^t \mathbf{v} + \beta \sum_{j \sim i} z_i^t z_j}$$

  - $\beta = 0$: classes indépendantes
  - $\mathbf{v} = (v_1, \ldots, v_K)$, vecteurs des poids des classes
- modèle CG : fonction de corrélation $\rho_k(h) = \exp(-\frac{|h|}{b_k})$
- critère = proportion de sites mal classés
- résultats moyennés sur 100 simulations

Performances de Geo-EM et Geo-EM(A)

modèle simulé						*Err*	
β	v_1	v_2	b_1	b_2	μ_2	Geo-EM	Geo-EM(A)
0	0	0	0.1	0.1	2	28.9	15.8
0	0	0	0.1	0.1	3	14.6	7.3
0	0	0	0.2	0.2	3	11.5	6.7
0	1	3	0.1	0.1	2	45.9	22.3
0	1	3	0.1	0.1	3	39.6	6.7

$$K = 2,\ \mu_1 = 0,\ \sigma_k = 1$$

Geo-EM(A) et SF-EM pour la classification de données issues du modèle complet

modèle simulé						*Err*	
β	v_1	v_2	b_1	b_2	μ_2	Geo-EM(A)	SF-EM
0.15	0	0	0.1	0.1	2	16.4	26.2
0.15	0	0	0.1	0.1	3	7.4	10.4
0.15	0	0	0.2	0.2	2	17.1	23.9
0.15	0	0	0.2	0.2	3	6.8	11.8

$$K = 2,\ \mu_1 = 0,\ \sigma_k = 1$$

modèle simulé						*Err*	
β	v_1	v_2	b_1	b_2	μ_2	Geo-EM(A)	SF-EM
0.4	0	0	0.1	0.1	2	17.7	23.3
0.4	0	0	0.1	0.1	3	6.9	10.7
0.6	0	0	0.1	0.1	2	18.4	24.3
0.6	0	0	0.1	0.1	3	7.1	11.4
0.8	0	0	0.1	0.1	2	19.6	21.9
0.8	0	0	0.1	0.1	3	7.3	11.2

$$K = 2,\ \mu_1 = 0,\ \sigma_k = 1$$

modèle simulé				*Err*	
β	v_1	v_2	v_3	Geo-EM(A)	SF-EM (V estimé)
0.15	0	0	0	17.1	18.9
0.15	1	2	3	33.3	56.1
0.4	0	0	0	17.4	21.8
0.4	1	2	3	38.2	59.2
0.6	0	0	0	17.3	22.8
0.6	1	2	3	44.4	62.5
0.8	0	0	0	18.3	24.2
0.8	1	2	3	56.6	63.5

$$K = 3,\ b_1 = b_2 = b_3 = 0.1,\ \mu_1 = 0,\ \mu_2 = 2,\ \mu_3 = 5,\ \sigma = 1$$

Conclusion

- Nous avons comparé deux façon de représenter la structure spatiale dans un problème de classification de données irrégulièrement réparties:
  - Le modèle avec champ gaussien apparaît plus robuste face à des données “complètement” spatialisées
  - Le modèle de champ de Markov caché conduit à des classifications plus lisses
- En cours : un algorithme pour l'estimation/classification dans le cas du modèle complet