Modèles mixtes et méthodes bayésiennes

Sylvia Richardson

Department of Epidemiology and Public Health,
Imperial College, London

Plan

1. Introduction
2. Hiérarchie normale
3. Modèles linéaires généralisés mixtes
   - Estimation
   - Structure des effets aléatoires
4. Modèles mixtes avec structure spatiale
5. Discussion

1 – Introduction

De nombreux problèmes statistiques font intervenir de multiples paramètres

Pourquoi ?

Nécessaire pour rendre compte de structures complexes des observations: hétérogénéité, dépendance spatiale ou temporelle, erreurs de mesure ...

Par example :

- "effet sujet" dans les modèles longitudinaux avec données repétées
- "fragilité" dans les données de survie corrélées (e.g. familiale)
- "effet génotype " dans les études d'association gène -maladie
- "effet centre " dans les méta-analyses
- risques relatifs de maladies par zone/age/périodes temporelles

- Comment faire une inférence statistique sur des paramètres multiples, $\{\theta_1, \ldots \theta_n\}$, mesurés sur n unités (gènes, personnes, centres, zones, ...), paramètres *qui sont reliés par la structure du problème ?*

"When a model has many parameters, it may be the case that we can consider them as a sample from some distribution. In this way we model the parameters with another set of parameters and build a model with different levels of hierarchy."

(Schervish, 1995)

$\longrightarrow$ Structure hiérarchique

Les modèles mixtes sont des examples de modèles hiérarchiques. Les différents niveaux correspondent à des hypothèses successives sur la structure des 'effets aléatoires' θ_i.

- Comment modéliser la distribution des 'effets aléatoires' θ_i?
  Relation entre la notion de similarité des θ_i et la représentation des θ_i comme provenant d'une distribution
  (Théorème de De Finetti sur la 'Représentation')

→ idée **d'échangeabilité**

  – Les θ_is sont un échantillon indépendant d'une distribution a priori commune

  – Une des hypothèses courantes dans la formulation de modèles hiérarchiques

  – Dans de nombreuses situations, cette hypothèse est trop 'simpliste'

→ **Echangeabilité Partielle**

  – Les θ_is sont partionnés en sous-groupes, de telle sorte qu'ils soient échangeables par sous-groupe.

→ **Autres modèles?**

Modéles Bayésiens hiérarchiques

Observations y,

Paramètres $\theta = (\theta_1, \ldots, \theta_n)$

- vraisemblance $p(y|\theta)$ (1er niveau)
  modélise la structure des observations

- loi a priori sur les 'effets aléatoires': $p(\theta)$
  est décomposée en distributions conditionelles successives: $p(\theta|\eta_2)$ (2ème niveau), $p(\eta_2|\eta_3)$... en se basant sur des considérations structurelles (eg échangeabilité ...) et une distribution marginale $p(\eta_m)$ de telle sorte que

$$p(\theta) = \int p(\theta|\eta_2)p(\eta_2|\eta_3)\ldots p(\eta_{m-1}|\eta_m)p(\eta_m)d\eta_2 d\eta_3 \ldots d\eta_m$$

les η_l sont appelés les hyperparamètres du niveau l

Principe bayésien d'estimation

- Toutes les quantités non observables, *c.a.d. effets aléatoires et paramètres*, sont estimés à partir de leur distribution conditionelle aux observations
- Dans cette distribution conditionelle, toutes les autres quantités non observables ont été intégrées
- Ceci permet de propager de manière cohérente l'incertitude sur les quantités non observables dans l'inférence sur les quantités d'intêret
- De manière pratique, on va simuler le plus souvent *la distribution conjointe* de toutes les quantités non observables (par Méthodes de Monte Carlo ou de Monte Carlo Chaînes de Markov)
- L'intégration par rapport à n'importe quel sous-ensemble de quantités aléatoires se fera alors facilement

2 – Hiérarchie normale

Modèles linéaires mixtes avec effets fixes (a) et effets aléatoires (β_i).

Plaçons nous dans le cas simple d'observations Y_{ij}, où i est l'indice du groupe, $i = 1, \ldots, I$ et j celui des observations dans le groupe, $j = 1, \ldots, n_i$.

- 1er Niveau

$$p(Y_{ij}|a, \beta_i, \sigma^2) \sim N(X_{ij}a + Z_{ij}\beta_i, \sigma^2).$$

  X_{ij} et Z_{ij} sont des covariables pour les effets fixes et aléatoires respectivement.

- 2ème Niveau

$$p(a|a^*, V_A) \sim N(a^*, V_A)$$

$$p(\beta|\beta^*, V_B) \sim N(\beta^*, V_B)$$

  où $\beta = (\beta_i, 1 \leq i \leq I)$.

- 3ème Niveau

$$p(a^*, \beta^*, V_A, V_B, \sigma^2|\text{Hyperparamètres})$$

$\rightarrow$ effets fixes et aléatoires sont traités symétriquement

A partir de la formulation précédente, on obtient des expressions explicites pour: $R = V(Y|a, \beta), V = V(Y|a) = R + ZV_BZ^T, C = V(Y|\beta) = R + XV_AX^T, V(Y) = V + XV_AX^T = C + ZV_BZ^T$.

Lois a posteriori

Pour simplifier, supposons qu'il n'y a pas de 3ème niveau $\Rightarrow a^*, \beta^*, V_A, V_B, \sigma^2$ soient connus.

On trouve

$$p(\beta|Y) \sim N\left(E(\beta|Y), V(\beta|Y)\right)$$

avec

$$V(\beta|Y)^{-1} = [V_B^{-1} + Z^TC^{-1}Z]$$

$$E(\beta|Y) = V(\beta|Y)[Z^TC^{-1}(Y - Xa^*) + V_B^{-1}\beta^*]$$

et une expression symétrique pour $p(a|Y)$.

Supposons que l'on choisit

- une loi a priori vague pour a, c.a.d. $V_A^{-1} = 0$
- $\beta^* = 0$
  $\Rightarrow$ on obtient:

$$E(\beta|Y) = BLUP = V_B Z^T V^{-1}(Y - X\hat{a}).$$

De plus, pour les effets fixes, on obtient l'estimateur des moindres carrés généralisés:

$$E(a|Y) = BLUE = (X^T V^{-1} X)^{-1} X^T V^{-1} Y = \hat{a}.$$

De manière générale, les estimations obtenues à l'aide du modèle hiérarchique bayésien seront différentes des estimateurs classiques

$\longrightarrow$ l'incertitude sur les valeurs des variances est prise en compte grâce au 3ème niveau.

C'est un des avantages de l'approche bayésienne

3 – Modèles linéaires généralisés mixtes

Plaçons nous à nouveau dans le cas simple d'observations Y_{ij}, i est l'indice du groupe, et j celui des observations dans le groupe

- 1er Niveau

$$p(Y_{ij}|a, \beta_i, X_{ij}), \text{indépendamment } \ 1 \leq j \leq n_i$$

$$E(Y_{ij}|a, \beta_i, X_{ij}) = u_{ij}, \ h(u_{ij}) = X_{ij}a + \beta_i$$

  X_{ij} covariables pour les effets fixes a, h fonction de lien

- 2ème Niveau

$$p(\beta_i|\beta*, \gamma, Z_i)$$

  Z_i covariables au niveau du groupe, coefficients de regression γ

- 3ème Niveau

$$p(\beta^*, a, \gamma|\text{Hyperparamètres})$$

Example de Modèle mixte Poisson-Gamma

Estimation du risque d'évènements rares

y_i : nombre observé d'évènements rares dans une zone i (il n'y a pas de répétitions)
E_i : nombre attendu d'évènements dans la zone i, (après ajustement sur l'age et le sexe
– Paramètres θ_i : risque relatif sous-jacent pour la zone i

- 1er niveau – variation locale dans la zone:

$$y_i \sim \text{Poisson}(E_i \theta_i)$$

- 2ème niveau – échangeabilité

$$\theta_i \sim \text{Gamma}(\alpha, \beta)$$

- 3ème niveau – loi des hyperparamètres (e.g. loi exponentielle)
  $\alpha \sim p(\alpha), \quad \beta \sim p(\beta)$

Si l'on avait des covariables Z_i, on pourrait considérer de manière alternative au 2ème niveau:

$$\log \theta_i \sim N(\theta^* + \gamma Z_i, \sigma^2)$$

Tout d'abord, supposons que α et β soient connus, alors:

$$p(\theta_i|\alpha, \beta, Y, E) \sim \text{Gamma}(y_i + \alpha, E_i + \beta)$$

$$E(\theta_i|\alpha, \beta, Y, E) = \frac{y_i + \alpha}{E_i + \beta}$$

Paramètres θ_i sont stabilisés et rapprochés de l'espérance de la population: α/β

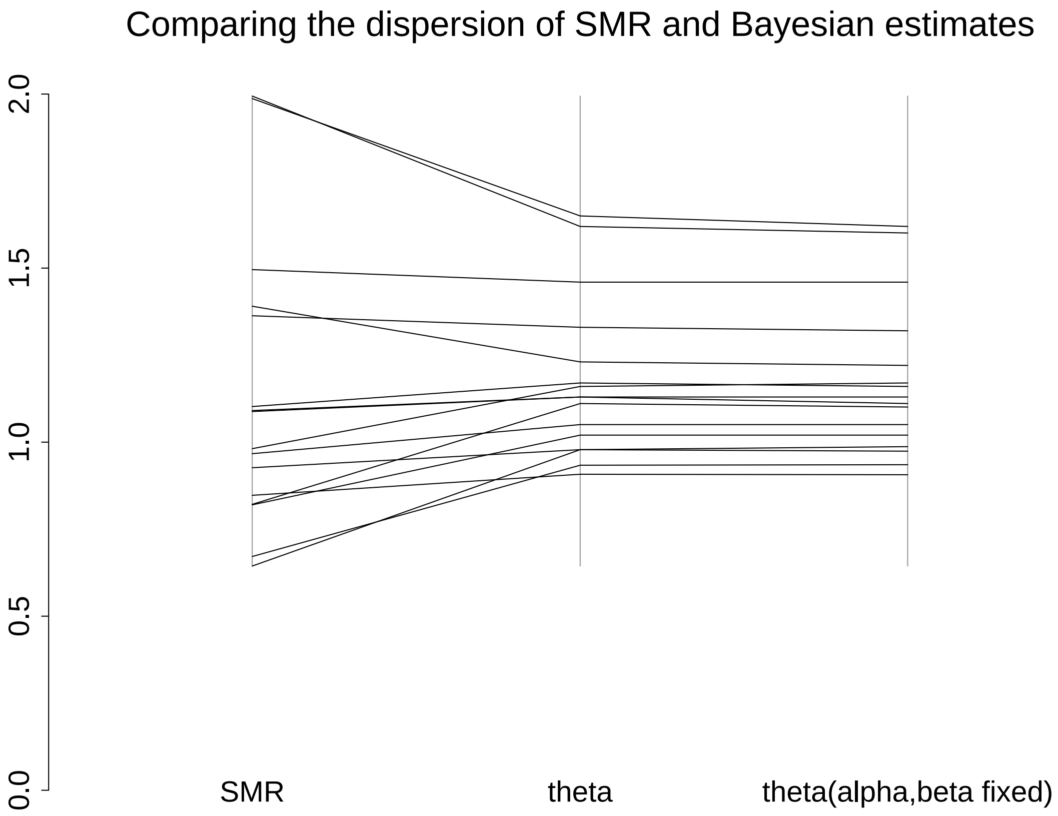
En général, on modélise l'incertitude sur les hyperparamètres à un 3ème niveau
$\Rightarrow$ quand on calcule $p(\theta_i|Y, E)$, l'incertitude sur α, β et les autres $\theta_j s$ est ainsi intégrée.

Example épidémiologique

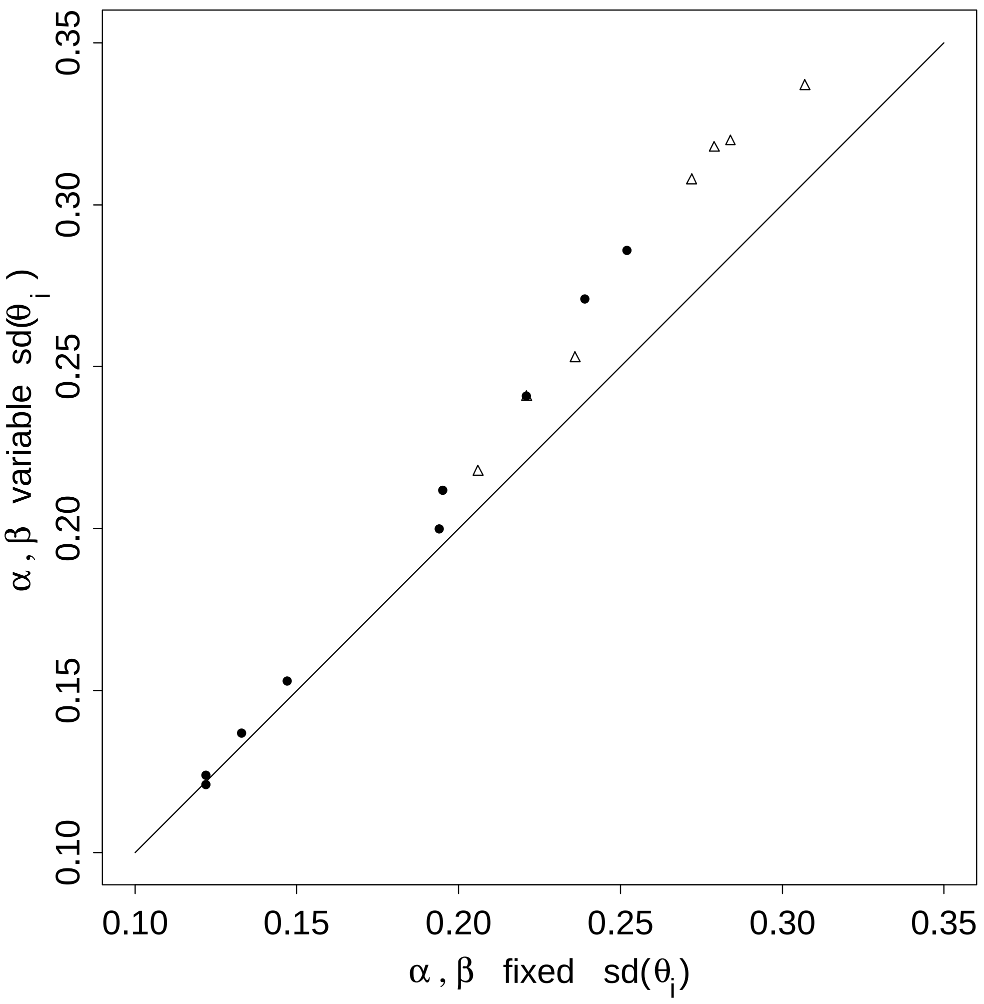
Question : le cancer de l'estomac est-il élevé dans les zones côtières où les coquillages sont contaminés par des toxines diarrhéiques ?

projet de S. Cordier (INSERM, Rennes) avec la collaboration de IFREMER

– Data : 16 zones choisies pour leur niveau d'exposition potentielle



Comparaison des écarts types a posteriori obtenus pour les risques relatifs θ_i : α and β fixés versus α and β suivant des lois a priori exponentielles



Les triangles indiquent les zones avec $y_i < 10$.

Estimation des effets aléatoires

Inférence classique: en dehors du cadre Gaussien, il est souvent fait appel à des évaluations numériques de vraisemblance ou à des approximations.

Etude de *Browne et Draper (2000)* dans le cadre du modèle mixte logistique avec un emboitement à trois niveaux (y_{ijk})

Ils comparent par simulation la performance de méthodes MQL: "marginal quasi-likelihood", PQL: "penalised quasi-likelihood" (Breslow et Clayton, 1993) et des méthodes bayésiennes estimées par des algorithmes MCMC

Ils concluent:

- Les méthodes de quasi vraisemblance n'ont pas une performance satisfaisante pour estimer les variances des effets aléatoires (biais de sous-estimation et couverture insuffisante)
- Les méthodes bayésiennes donnent des résultats satisfaisants, mais avec un temps de calcul bien supérieur

L'estimation de modèles linéaires généralisés mixtes est implémentée dans des logiciels tels BUGS (Bayesian inference Using Gibbs Sampling)

$\longrightarrow$ stratégie à deux étapes?
– exploratoire avec des méthodes approchées
– inférence bayésienne via MCMC pour affiner l'estimation des variances

Structure des effets aléatoires

- **Echangeabilité**
- **Echangeabilité partielle**
  – Suppose qu'on connait les sous-groupes

  $\longrightarrow$ Comment procéder quand les θ_is doivent capturer une hétérogénéité résiduelle mais que les sous-groupes sont inconnus?
- **Modèles de mélange**
  – Permettent d'approcher n'importe quelle forme paramétrique
- **Modèles de partition**

  $\longrightarrow$ Comment procéder quand les θ_is sont corrélés?

- **Modèle mixte spatial:** comparaison entre plusieurs types de structures

Modèles de mélange

Richardson et Green (1997)
Sont utiles pour modéliser l'hétérogénéité résiduelle (après prise en compte de covariables explicatives, ...)
→ Modèle de mélange au 2ème niveau

$$\theta_i \sim \sum_{j=1}^{k} w_j f(\cdot|\phi_j) \quad \text{independently for } i = 1, 2, \ldots, n$$

$f(\cdot|\phi)$ famille paramétrique donnée
$\{\phi_j\}, \{w_j\}, k$ inconnus

→ on suppose qu'il y a des groupes $j = 1, 2, \ldots, k$, de taille $\propto w_j$. A chaque observation i est associée une *variable d'allocation* z_i:

$$p(z_i = j) = w_j \quad \text{indépendamment pour} i = 1, 2, \ldots, n$$

$$\theta_i | z \sim f(\cdot|\phi_{z_i}) \quad \text{indépendamment pour } i = 1, 2, \ldots, n$$

Example sur le risque d'évènements rares (continué)

Au lieu de sélectionner arbitrairement des zones côtières 'exposées', on modélise la variation du rique relatif dans un ensemble de zones côtières où l'exposition est monitorée.

– Données :
(y_i, E_i) dans 59 zones côtières,
covariable x_i mesurant l'exposition à certaines toxines par le nombre de jours où le niveau de ces toxines excède une limite.

– On utilise un mélange de distributions de Poisson, avec un nombre inconnu de composantes, pour estimer le risque relatif pour chaque zone côtière.

Analyse des 59 zones côtières

Hétérogénéité manifeste, a posteriori pas de probabilité sur $k = 1$

$P(k = 2|y, E) = 0.25, P(k = 3|y, E) = 0.23$
$P(k = 4|y, E) = 0.16$

Valeurs des paramètres pour $k = 3$ (à titre illustratif)
$w_1 = 0.39, \theta_1 = 0.74$
$w_2 = 0.31, \theta_2 = 1.03$
$w_3 = 0.30, \theta_3 = 1.49$

Hétérogénéité résiduelle après inclusion de la covariable x_i:

$P(k = 2|y, E) = 0.38, P(k = 3|y, E) = 0.25$
$P(k = 4|y, E) = 0.16$

$w_1 = 0.36, \theta_1 = 0.67$
$w_2 = 0.32, \theta_2 = 0.96$
$w_3 = 0.32, \theta_3 = 1.37$

$\rightarrow$ Mais pas de relation entre les zones attribuées a posteriori à la 3ème composante et celles où x_i a des valeurs élevées.

Modèles Bayésiens de Partition (BPM)

Introduits en regression, classification et dans les analyses en plan factoriel
(Denison et al. 1998, Holmes et al. 2000, Consonni and Veronese 1995, Nobile and Green 2000)

Introduits dans les analyse spatiales de risque par *Knorr-Held and Rasser* (2000) et *Denison and Holmes* (2000)

- Une alternative flexible pour modéliser la structure de paramètres au 2ème niveau
- Suppose qu'il y a un nombre, a priori inconnu, de groupes ou de strates d'allocation
- Chaque groupe a en commun le paramètre θ_j
  $\rightarrow$ Il faut définir:
  – Le modèle a priori sur la structure de partition
  – Le modèle d'allocation des θ_is dans les groupes
  $\rightarrow$ En général, nécessité d'incorporer une information a priori dans la structure des partitions pour "borrowing strenght" de manière effective

Modèles mixtes avec structure spatiale

– Les cadres d'applications sont nombreux, **épidémiologie** (structure spatiale des maladies), traitement d'image (reconstruction), écologie (répartition des espèces).

– Nous allons nous restreindre au cadre épidémiologique d'analyse des variations spatiale du risque pour une maladie rare

– Une hypothèse raisonnable est de supposer que des zones voisines auront tendance à avoir des risques semblables (partage de facteurs de risque commun)

– Pour une maladie rare ou des petites zones:

$\rightarrow$ fluctuations de Poisson des cas observés y_i autour du leur espérance $E_i\theta_i$, où E_i sont les nombres attendus sur la base de taux de référence (après ajustement sur l'âge et le sexe).

θ_i est le risque relatif de la zone i et que l'on cherche à estimer.

Modèle hiérarchique spatial

- 1er niveau – variabilité locale:

$$y_i \sim \text{Poisson}(E_i\theta_i), i = 1, \ldots, n$$

- 2ème niveau – modèle pour les $\{\theta_i\}$ qui prend en compte hétérogénéité et dépendance spatiale

  Champs markoviens

  Modèles d'allocation ou de partition spatiale

Note: on pourrait facilement considérer un modèle binomial au 1er niveau

$\rightarrow$ comment introduire une dépendance spatiale?

On utilise un graphe avec des voisins définis par la contiguité: $j \sim i$ indique que i et j sont voisins

Champs markoviens (Markov random fields, MRF): Besag, York and Mollié (1991), Clayton and Bernardinelli (1992)

2ème niveau – modèle mixte log lineaire

$$\log \theta_i = \mu + u_i + v_i$$

Les effets aléatoires, u_i et v_i, avec $\sum_i u_i = 0, \sum_i v_i = 0$ ont des fonctions différentes:

u_i représente une composante d'hétérogénéité non structurée (échangeable):

$p(u_i) \sim N(0, \tau_u^{-1})$ indépendamment pour $i = 1, 2, \ldots, n$

v_i représente un effet de "clustering" et suit un modèle autorégressif :

$$p(v_i | v_j, j \neq i) \sim N(\overline{v}_i, (n_i \tau_v)^{-1})$$

où $\overline{v}_i = \sum_{j \sim i} v_j / n_i$, et n_i est le nombre de voisins de la zone i.

3ème niveau Lois a priori pour τ_u^{-1} et τ_v^{-1}: Lois Gamma ou du χ^2, dépendant d'hyperparamètres fixés

Remarques

Les paramètres τ_u^{-1} et τ_v^{-1} ne sont pas comparables

$\rightarrow \tau_u^{-1}$ représente une variabilité marginale
$\rightarrow \tau_v^{-1}$ représente une variance conditionelle

Il y a parfois des problémes d'identifiabilité entre les effets aléatoires u_i et v_i

La formulation markovienne conditionelle facilite l'implémentation d'algorithmes de simulation efficaces.

Ceci ne serait pas le cas si on avait spécifié directement une distribution multivariée pour les $\log \theta_i$ avec une matrice de variance-covariance dépendante de paramètres reflétant une structure spatiale

Modèles d'allocation ou de partition

⇒ Dans la formulation MRF, les paramètres caractérisant la dépendance spatiale, τ_u^{-1} et τ_v^{-1} ont un effet global

⇒ Risque de masquer des discontinuités potentielles par un 'lissage spatial' trop fort

⇒ Dévelopement de modèles qui remplacent le champs spatial continu des $\{\theta_i\}$ par des modèles d'allocation ou de partition, permettant des discontinuités.

Points communs:

- nombre d'éléments dans la partition ou le mélange est inconnu
  ⇒ utilisation d'algorithmes MCMC à saut réversible (Green, 1995)
- Estimation des θ_i obtenue en marginalisant la distribution des allocations
  ⇒ distribution a posteriori de forme continue

Modèles de mélange avec allocations correlées

(Green and Richardson, 2002)

La structure spatiale est introduite dans le modèle d'allocation

Forme du modèle:

$$\theta_i = \theta_{z_i}$$

où $\{\theta_j, j = 1, 2, \ldots, k\}$ caractérisent k différentes composantes, et $\{z_i, i = 1, 2, \ldots, n\}$ sont *les variables d'allocation* à valeurs dans $\{1, 2, \ldots, k\}$.

Lois a priori pour $\{\theta_j, j = 1, 2, \ldots, k\}$:

$$\theta_j \sim \Gamma(\alpha, \beta) \quad \text{indépendamment pour } j = 1, 2, \ldots, k.$$

Modèle pour les variables d'allocation $\{z_i\}$*:*

Le nombre de composantes k est inconnu avec loi a priori uniforme dans un interval

Les allocations $\{z_i | k\}$ sont corrélées spatialement suivant le modèle de Potts.

Le modéle de Potts

– Inspiré par le traitement d'image

Formulation du modéle de Potts:

$$p(z) = \frac{e^{\psi U(z)}}{c_k(\psi)}$$

where

$$U(z) = \sum_{i \sim i'} I[z_i = z_{i'}] \quad \text{et} \quad c_k(\psi) = \sum_{z \in \{1,2,\ldots,k\}^n} e^{\psi U(z)}$$

sont le *nombre de pairs de zones de mme 'couleurs'*, et une constante de normalisation

Pour le paramètre d'intéraction ψ, on choisit une loi a priori discréte uniforme sur une grille de valeurs

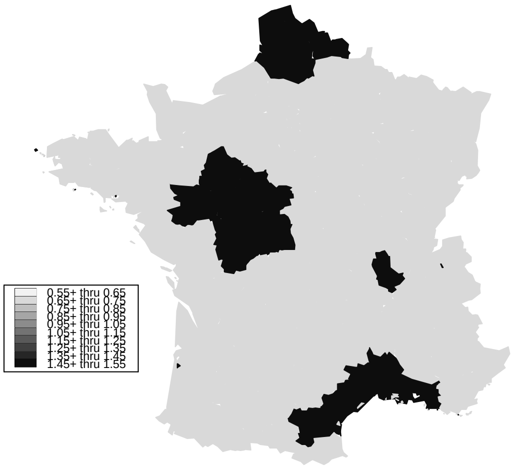
Performance et comparaison entre le modèle MRF de Besag, York and Mollié (BYM) et le modéle de mélange – Potts (MIX)

Investigation sur des données simulées dans les 94 départements

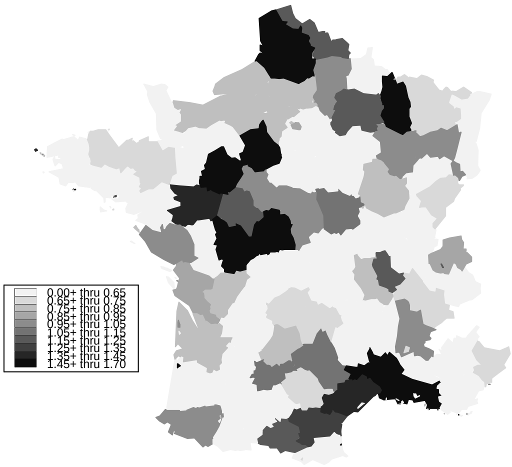
Pout tester différentes caractéristiques des modéles, on a généré 3 examples correspondant à différents types de de variation géographique du risque:

- "Block4" 4 groupes de zones ayant un risque élévé de 1.5 sur un 'arriére-plan' de 0.7
- "North-South" simule un simple contraste Nord/Sud avec des risques de 0.8 dans le Nord et 1.2 dans le Sud
- "GradientNS" correspond à des risques relatifs décroissants linéairement du Nord au Sud

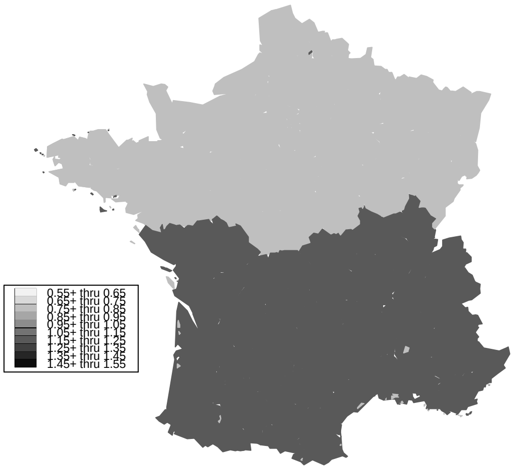
Block4, true RR



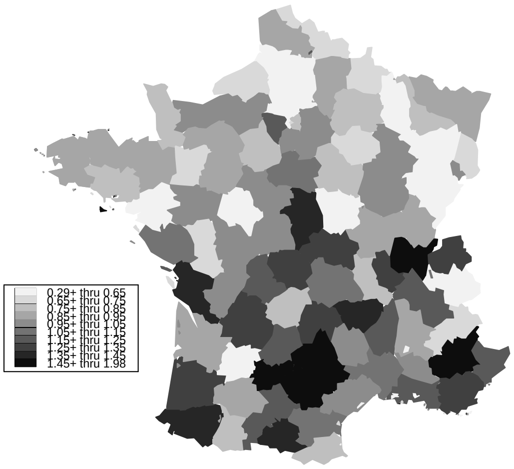
Block4, SMR



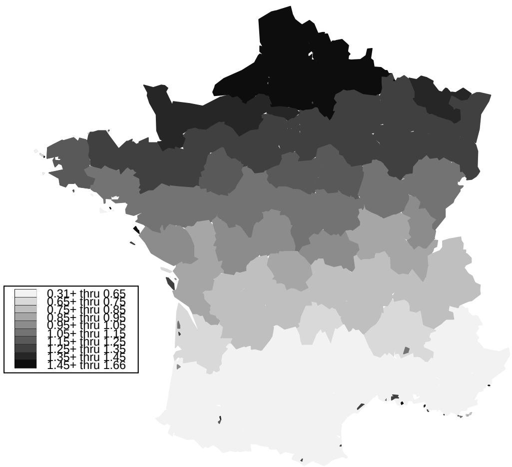
North-South, true RR



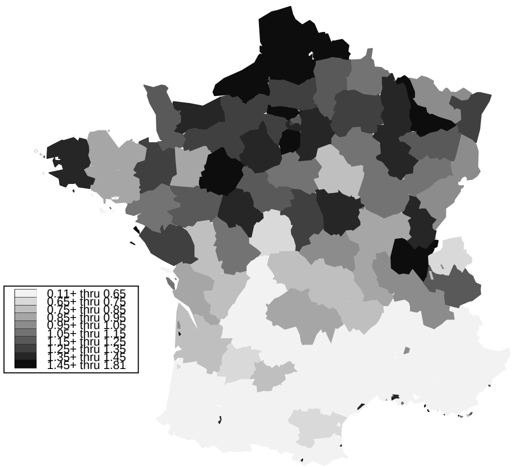
North-South, SMR



Gradient NS, true RR



Gradient NS, SMR

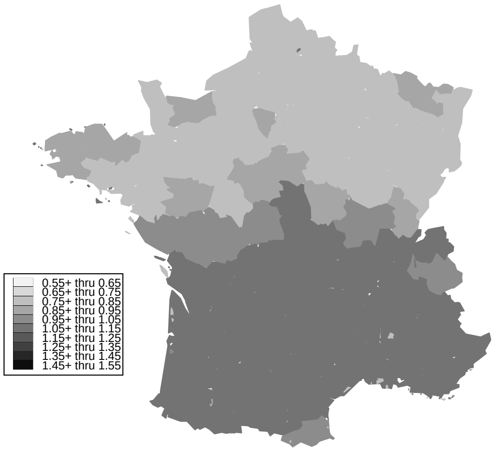


Estimation a posteriori des risques $\{\theta_{z_i}\}$

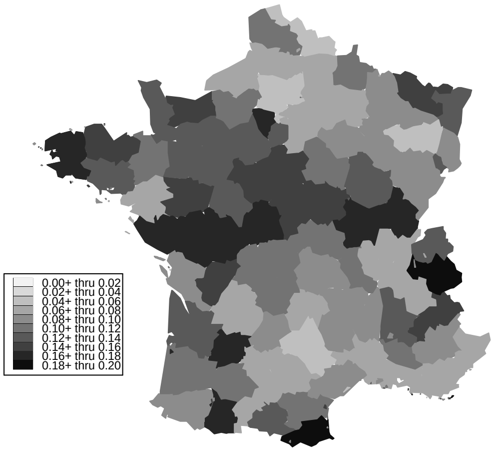
- Bonne reconstruction des risques sous-jacents pour MIX, image plus bruitée pour BYM
  $\rightarrow$ meilleure erreur quadratique moyenne pour MIX
- Variabilité augmentée pour les zones de contraste pour MIX
- Variabilité influencée par la taille des risques pour BYM
- Intérêt à représenter la distribution a posteriori de $\{\theta_{z_i}\}$ par zone

North-South

Posterior means of RR

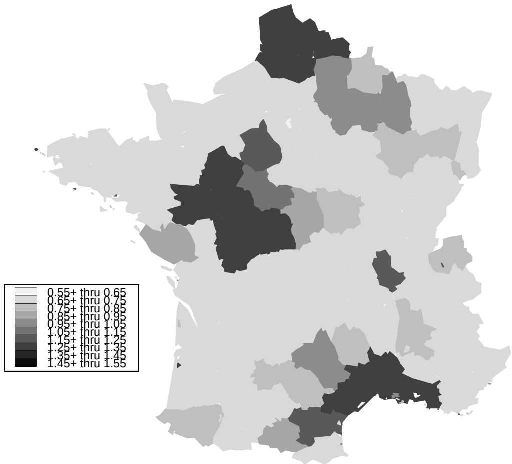


Posterior s.d. of RR



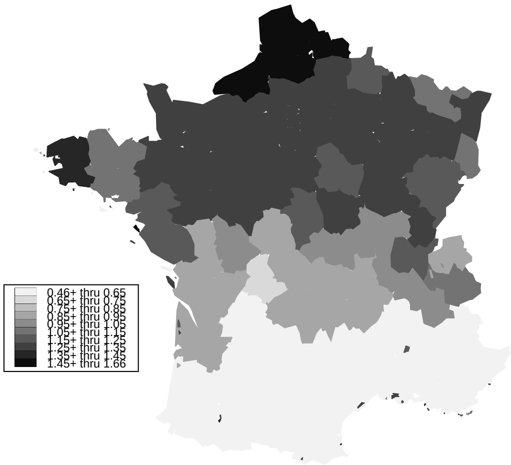
Block4

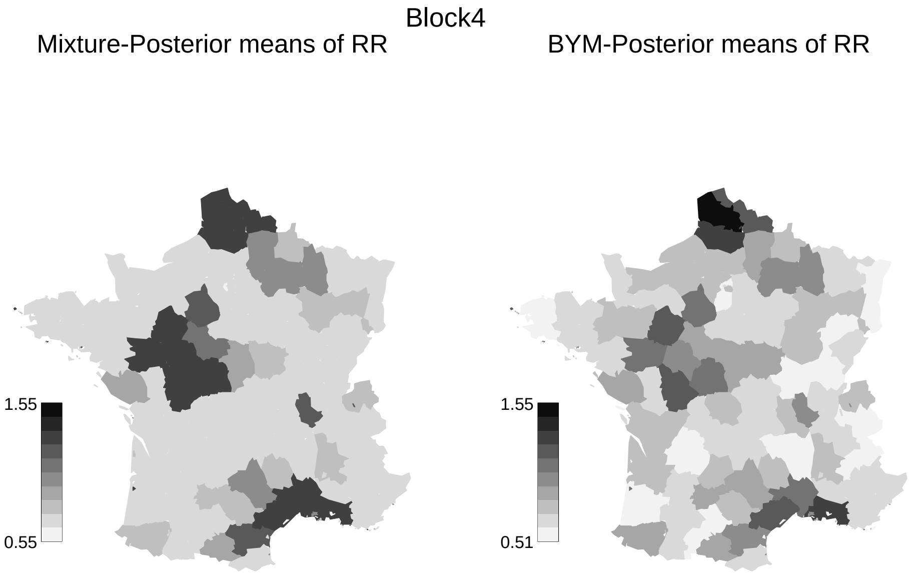
Posterior means of RR

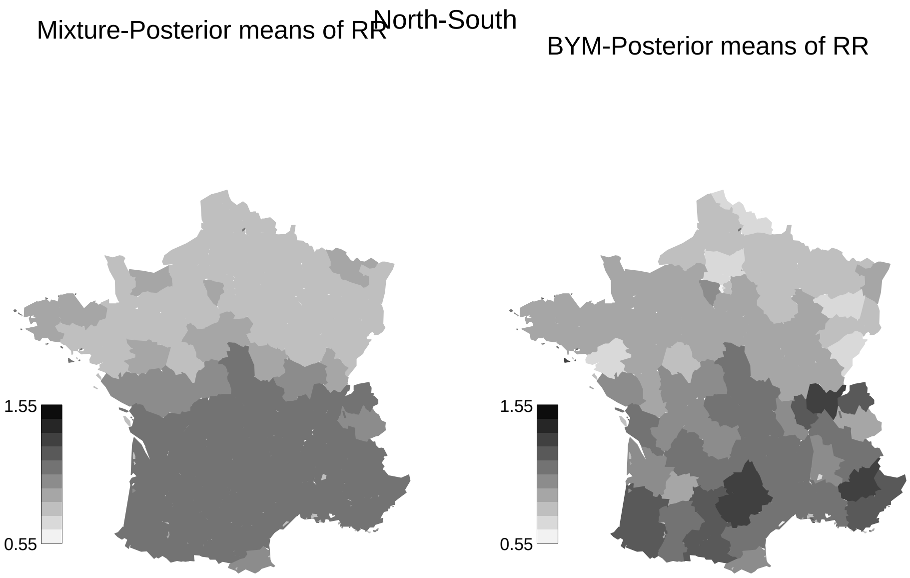


Gradient NS

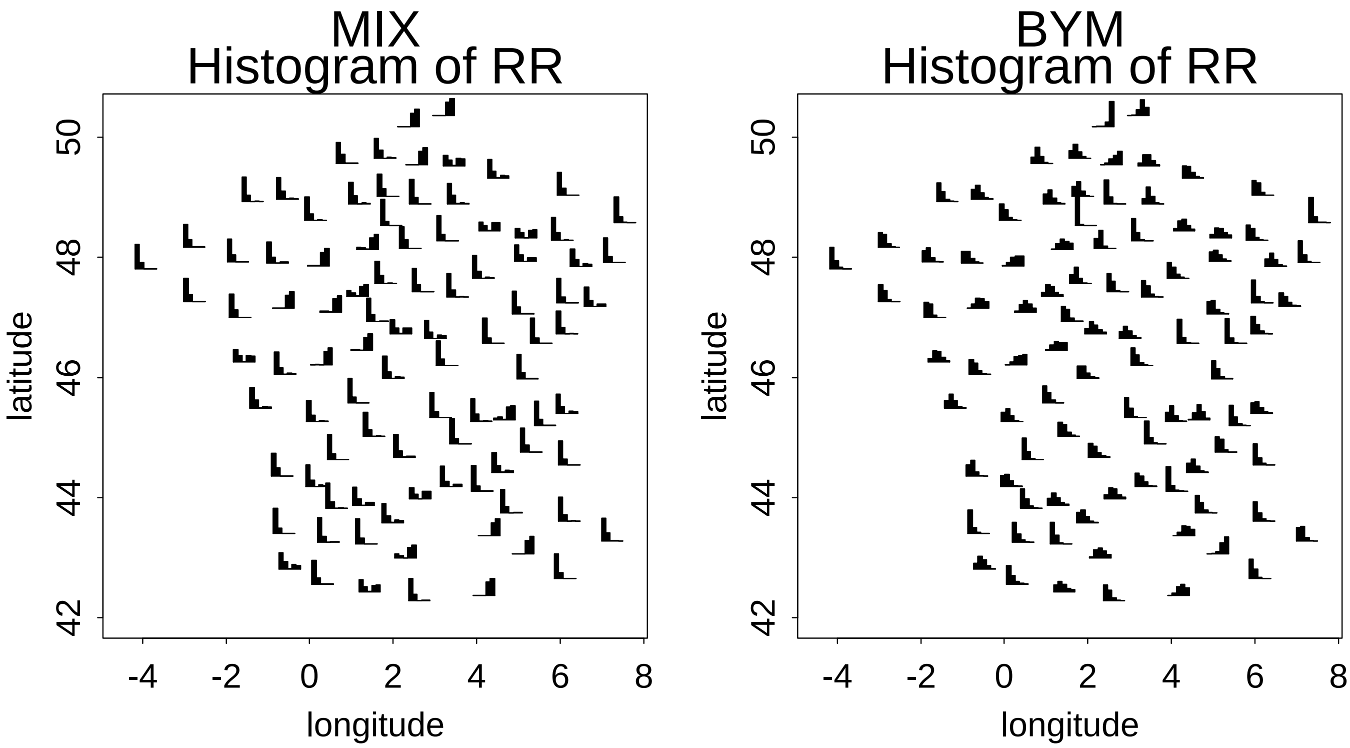
Posterior means of RR




Block4
Mixture-Posterior means of RR
BYM-Posterior means of RR
1.55
0.55
1.55
0.51


North-South
Mixture-Posterior means of RR
BYM-Posterior means of RR
1.55
0.55
1.55
0.55

Block4


MIX
Histogram of RR
BYM
Histogram of RR
latitude
longitude
50
48
46
44
42
-4
-2
0
2
4
6
8

Discussion

- Approche bayésienne des modèles mixtes permet une grande flexibilité

  Définition de la hiérarchie

  Prise en compte de différentes sources de dépendance et d'hétérogénéité

  Choix des distributions à tous les niveaux

  Ne requiert pas d'approximations

- Implémentation nécessite le plus souvent des algorithms de simulations, type MCMC

  Des logiciels tel BUGS permettent d'estimer la majorité des modèles linéaires généralisés mixtes

- Le choix de la structure et des formes de la loi des effets aléatoires a une influence

  modèles paramétriques $\rightarrow$ semi-paramétriques, modèles de partition, de mélange.