BayesX: modèles additifs structurés de régression Exemple d'application sur des données géographiques médicales

Erik-A. Sauleau[1,2]

1 Pôle Santé publique - Santé au travail, Hôpitaux Universitaires de Strasbourg
2 Laboratoire de Biostatistiques, Faculté de médecine, Strasbourg

erik-andre.sauleau@medecine.u-strasbg.fr

AppliBUGS 27/11/2008

Plan

Plan

1. Rappel sur les modèles StAR
   - Le modèle des données
   - Les priors
     - Généralités
     - Données ponctuelles
     - Données groupées
   - StAR et GLMM
2. Présentation du logiciel BayesX
3. Comparaison entre BayesX et GeoBUGS
4. Application à des données géographiques médicales

Le modèle des données

Du GLM au modèle StAR

- GLM classique ...
  1. Réponse y, covariables $\boldsymbol{x}$ et paramètres γ
  2. La distribution de y appartient à la famille exponentielle
  3. $E(y|\boldsymbol{x},\gamma) = h(\mu) = h(\boldsymbol{x}'\gamma)$
- ...auquel, pour la flexibilité, on rajoute des fonctions de covariables
  - Prédicteur linéaire $\mu = \boldsymbol{u}'\gamma + \sum_j f_j(\Psi_j)$
  - Effets non linéaires, coefficients variables (VCM)
  - Tendances temporelles, saisonnalité
  - Surfaces multidimensionnelles

Le modèle des données

Le modèle StAR $\mu = \boldsymbol{u}'\gamma + \sum_j f_j(\Psi_j)$

De nombreux cas particuliers connus

1. Modèle additif généralisé (GAM), modèle mixte additif généralisé (GAMM)
2. Modèles géoadditifs à la Kammann et Wand $\mu = \boldsymbol{u}'\gamma + \sum_j f_j(\Psi_j) + f_{spat}(s)$
3. Modèles à coefficients variables, régression géographique pondérée $\mu = \boldsymbol{u}'\gamma + \sum_j f_j(\Psi_j) z_j$
4. Modèle d'interaction de type ANOVA

Modèles StAR $\mu = \boldsymbol{u}'\gamma + \sum_j f_j(\Psi_j) + f_{spat}(s)$

La forme générale des priors

- Avec $f_j(\Psi_j) = X_j\beta_j$, $\mu = X_1\beta_1 + \cdots + X_p\beta_p + \boldsymbol{u}'\gamma$
- $p\left(\beta_j|\tau_j^2\right) \propto \frac{1}{\left(\tau_j^2\right)^{\frac{\mathrm{rang}(K_j)}{2}}} \exp\left(-\frac{1}{2\tau_j^2}\beta_j' K_j \beta_j\right)$
- K_j est une matrice de pénalité

Modèles StAR $\mu = \boldsymbol{u}'\gamma + \sum_j f_j(\Psi_j) + f_{spat}(s)$

Les priors non spatiaux

- Effet fixe $p(\gamma) \propto cste$ ou aléatoire $\beta \sim N(0, \sigma^2)$
- Coefficients variables $f(x)z$, x est continue, spatiale ou catégorielle et z est le plus souvent catégorielle (pentes aléatoires)
- Random walks $\beta_t = \beta_{t-1} + \epsilon_m$ ou $\beta_t = 2\beta_{t-1} - \beta_{t-2} + \epsilon_m$ avec $\epsilon_m \sim N(0, \sigma^2)$
- P-splines $f(x) = \sum_j \beta_j B_j(x)$ où B est une B-spline, avec une pénalité en $\lambda \sum_{k+1}^{M} (\Delta^k \beta_t)^2$

Modèles StAR $\mu = \boldsymbol{u}'\gamma + \sum_j f_j(\Psi_j) + f_{spat}(s)$

Les priors spatiaux

Données ponctuelles

- *P-splines à 2 dimensions*
- GRF = Champ aléatoire Gaussien en spécifiant la fonction de corrélation

Données groupées

- GMRF = Champ aléatoire Gaussien de Markov

Priors pour les données ponctuelles

GRF = Gaussian Random Field

- $p\left(\beta_j | \tau_j^2\right) \sim \frac{1}{(\tau_j^2)^{\frac{rank(K_j)}{2}}} \exp\left(-\frac{1}{2\tau_i^j}\beta_j' K_j \beta_j\right)$
- La matrice de pénalité est $K = C^{-1}$ avec $C(i,j) = C(||s_i - s_j||)$
- $C(.)$ est une fonction isotropique de corrélation, pouvant prendre différentes formes, par exemple de la classe Matérn

Priors pour les données groupées

Processus conditionnel autorégressif

- GMRF (Gaussian Markov Random Field) : une multinormale avec certaines conditions d'indépendance conditionnelle
- CAR (Conditional AutoRegressive) : le GMRF présenté sous forme full conditional $E(x_i|x_{-i})$

$$\begin{cases} E\left(\phi_i|\phi_{-i}\right) & = & \mu_i + \sum_{j \neq i} \left(\beta \times w_{ij}\left(\phi_j - \mu_j\right)\right) \\ Var\left(\phi_i|\phi_{-i}\right) & = & sk_i \end{cases}$$

Priors pour les données groupées

Processus conditionnel autorégressif

- $E\left(\phi_i|\phi_{-i}\right) = \mu_i + \sum_{j \neq i}\left(\beta \times w_{ij}\left(\phi_j - \mu_j\right)\right)$
    - μ est une constante
    - β mesure la force de l'autocorrélation entre UG
    - $w_{ij} = \frac{q_{ij}}{\sum_j q_{ij}}$ mesure la dépendance entre UG
        - $q_{ij} = 1$ si i et j partagent une frontière, $q_{ij} = 0$ si non
        - $q_{ij} = exp(-\kappa d_{ij})$
- $Var\left(\phi_i|\phi_{-i}\right) = sk_i$
    - s mesure la force de la similarité entre UG

Priors pour les données groupées

Processus conditionnel autorégressif

- Pour plus de flexibilité une loi normale sans structure spatiale est souvent ajoutée (hétérogénéité)
- Un modèle particulier est le convolution prior (BYM)

Priors pour les données groupées

Le modèle Besag, York et Mollié

Autocorrélation	+	Hétérogénéité
CAR intrinsèque		Normale

$\beta = 1$

$\mu = 0$

$q_{ij} = 1$ si i et j partagent une frontière, $q_{ij} = 0$ si non

n_i : nombre de voisins de i

$\phi_i|\phi_{-i} \sim N\left(\frac{\sum_{j\in\partial}\phi_j}{n_i}, \frac{1}{\lambda n_i}\right)$ $\qquad$ $\theta_i \sim N\left(0, \frac{1}{\tau}\right)$

Precisions λ et τ

Présentation des modèles StAR en GLMM

- $\mu = \boldsymbol{u}'\gamma + \sum_j f_j(\Psi_j) = X_1\beta_1 + \cdots + X_p\beta_p + \boldsymbol{u}'\gamma$
- On définit pour chaque j : $\beta = X^u\beta^u + X^p\beta^p$
  1. $X^p = L(L'L)^{-1}$ avec $K = LL'$ (décomposition spectrale $K = \Gamma\Omega\Gamma'$ et $L = \Gamma\Omega^{\frac{1}{2}}$)
  2. $X^u = \mathbf{1}$ pour MRF et P-spline avec pas aléatoire d'ordre 1 et $X^u = (\mathbf{1}, \boldsymbol{\kappa})$ pour P-spline avec pas aléatoire d'ordre 2
- Si $\tilde{U}_j = X_j X_j^u$ et $\tilde{X}_j = X_j X_j^p$, alors $\mu = \tilde{\boldsymbol{U}}\boldsymbol{\beta}^u + \tilde{\boldsymbol{X}}\boldsymbol{\beta}^p$
  1. $\tilde{\boldsymbol{U}} = (\tilde{U}_1, \cdots, \tilde{U}_p, \boldsymbol{u})$ et $\boldsymbol{\beta}^u = ((\beta_1^u)', \cdots, (\beta_p^u)', \gamma')$
  2. $\tilde{\boldsymbol{X}} = (\tilde{X}_1, \cdots, \tilde{X}_p)$ et $\boldsymbol{\beta}^p = ((\beta_1^p)', \cdots, (\beta_p^p)')$
- On a $\frac{1}{\tau^2}\beta' K\beta = \frac{1}{\tau^2}(\beta^p)'\beta^p$
- Les priors : $p(\beta^u) \propto \text{constante}$ et $\beta^p \sim N(0, \tau^2)$
- GLMM avec effets fixes : $\boldsymbol{\beta}^u$ et effets aléatoires : $\boldsymbol{\beta}^p \sim N(0, T)$ où T est un réarrangement des τ^2

StAR et McMC

- Brezger A, Lang S. Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis* 2006(**50**) :967–91.
- Fahrmeir L, Lang S. Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society C (Applied Statistics)* 2001(**50**) :201–20.
- Fahrmeir L, Osuna, L. Structured additive regression for overdispersed and zero-inflated count data. *Applied Stochastic models in Business and Industry* 2006(**22**) :351–69.
- Hennerfeind A, Brezger A, Fahrmeir L. Geoadditive survival models. *Journal of the American Statistical Society* 2006(**101**) :1065–75.
- Lang S, Brezger A. Bayesian P-splines. *Journal of Computational and Graphical Statistics* 2004(**13**) :183–212.
- Kneib T, Hennerfeind A. Bayesian semiparametric multistate models. SFB 386, *discussion paper* 502, 2006.

StAR et modèle mixte

Fahrmeir L, Kneib T, Lang S. Penalized structured additive regression for space-time data : a bayesian perspective. *Statistica Sinica* 2004(**14**) :715–45.

Kneib T. Geoadditive hazard regression for interval censored survival times. *Computational Statistics and Data Analysis* 2006(**51**) :777–92.

Kneib T, Fahrmeir L. A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics* 2007(**34**) :207–28.

Kneib T, Fahrmeir L. Structured additive regression for categorical space-time data : a mixed model approach. *Biometrics* 2006(**62**) :109–18.

Et encore

Kammann EE, Wand MP. Geoadditive models. *Journal of the Royal Statistical Society C (Applied Statistics)* 2003(**52**) :1–18.

Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 1991(**43**) :1–59.

Plan

Interface

- Interface Java
- Routine statistiques en C++
- Uniquement plateforme Windows
- Quatre fenêtres
  1. Commande
  2. Résultats
  3. Historique
  4. Objets
- http ://www.stat.uni-muenchen.de/~bayesx

Structure générale

- Logiciel orienté objet
  1. Créer les objets

     typeobjet nomobjet

  2. Appliquer des méthodes aux objets

nomobjet.nommethode [*model*] [`weight` *nomvar*] [`if` *expr*]

[, *options*] [`using` *texte*]

- Utilisation de commandes en batch

`usefile` *nomfichier*

Inférence entièrement bayésienne par McMC

- Prior inverse-gamma sur les variances $p(\tau^2) \sim IG(a,b)$
- Hypothèses d'indépendance conditionnelle
  $p(\beta_1,\cdots,\beta_p,\tau_1^2,\cdots,\tau_p^2,\gamma|y) \propto$
  $L(y,\beta_1,\cdots,\beta_p,\gamma)\prod_1^p p(\beta_j|\tau_j^2)p(\tau_j^2)$
- Estimation par intégration de Monte Carlo et chaînes de Markov (McMC)

Inférence entièrement bayésienne par McMC

Réponse de la famille exponentielle

- IWLS : moindres carrés pondérés itératifs
  - Associe score de Fisher et algorithme de Metropolis-Hastings
  1. Utilisant les modes *a posteriori* (`proposal=iwlsmode`)
  2. Utilisant les modes courants (`proposal=iwls`)
3. Conditional prior proposal (`proposal=cp`)

Inférence bayésienne empirique (présentation GLMM)

- Les τ^2 sont des constantes inconnues à estimer de leur vraisemblance marginale
- L'ensemble des paramètres peut être estimé par IWLS et vraisemblance marginale (approchée) ou maximum de vraisemblance restreinte (REML)
- Répéter deux étapes
  1. Estimer les β^u et les β^p, solution de
  $$\begin{pmatrix} \tilde{\boldsymbol{U}}'\boldsymbol{W}\tilde{\boldsymbol{U}} & \tilde{\boldsymbol{U}}'\boldsymbol{W}\tilde{\boldsymbol{X}} \\ \tilde{\boldsymbol{X}}'\boldsymbol{W}\tilde{\boldsymbol{U}} & \tilde{\boldsymbol{X}}'\boldsymbol{W}\tilde{\boldsymbol{X}} + \tilde{\boldsymbol{\Lambda}}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^u \\ \boldsymbol{\beta}^p \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{U}}'\boldsymbol{W}\tilde{\boldsymbol{y}} \\ \tilde{\boldsymbol{X}}'\boldsymbol{W}\tilde{\boldsymbol{y}} \end{pmatrix}$$
  2. Estimer les τ^2 par maximisation de
  $-\frac{1}{2}(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{U}}\hat{\beta}^{\boldsymbol{u}})'\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{U}}\hat{\beta}^{\boldsymbol{u}})$ où $\boldsymbol{\Sigma} = \tilde{\boldsymbol{W}}^{-1} + \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{X}}'$

Les sept objets

`dataset`	Stockage et manipulation des données
`map`	Stockage et manipulation (très limitée) de cartes
`bayesreg`	Estimation par McMC de modèles StAR (famille exponentielle, survie, modèles multiétats)
`remlreg`	Estimation bayésienne empirique
`stepwisereg`	*Non documenté dans BayesX*
`graph`	Visualisation des données ou des estimations
`dag`	Estimation de modèles DAG par McMC (sauts réversibles)

Objet dataset

Méthode	Commentaire	Bugs
	Statistiques sur les variables	
descriptive	Statistiques résumées	N
tabulate	Tableaux de fréquence	N
pctile	1, 5, 25, 50, 75, 95 et 99%	N
	Manipulation de variables	
drop	Suppression (variables ou observations)	N
generate	Opérateurs, fonctions ou constantes	N
replace	Opérateurs, fonctions ou constantes	N
rename	Changement de nom	
	Manipulation de données	
sort	Tri sur plusieurs variables	N
infile	Import (ASCII)	O
outfile	Export (ASCII)	N

Objet map

Méthode	Commentaire	Bugs
`infile`	Import boundary (shp2bnd.r)	O
`outfile`	Export boundary	O
`reorder`	Diminution de l'enveloppe de la matrice d'adjacence (accélère McMC de `bayesreg`)	N

Objet graph

Méthode	Commentaire	Bugs
drawmap	Cartes, nombreuses options (couleurs, classes, . . .)	O
plot	Deux variables, nombreuses options (nuage, courbe, . . .)	O
plotautocor	Après autocor, tous les paramètres	O
plotsample	Après getsample, tous les paramètres	O

Objet dag

Méthode	Commentaire	Bugs
`estimate`	*Mea culpa*	N

Objet bayesreg

Méthode	Commentaire	Bugs
`regress`	*nomobjet*.`regress` *model* [`weight` *nomvar*] [`if` *expr*] [, *options*] `using` *dataset*	O
`autocor`	Tous les paramètres, itérations sauf burn-in et thinning	O
`getsample`	Tous les paramètres, itérations sauf burn-in et thinning	O

Objet bayesreg

model

- De la forme $depvar = term_1 + \cdots + term_r$
- *depvar* univariée uniquement
- Nombreux effets : offset, fixe, pas aléatoires (ordres 1 et 2), P-spline, saisonnalité, MRF, géospline, aléatoire, baseline
- Interactions : VCM, splines bidimensionnelles
- Nombreuses distributions de la réponse : gaussienne, gamma, binomiale (lien logit ou probit), multinomiale (lien logit ou probit), Poisson, négative-binomiale, ZIP, probit cumulés, survie, modèle multiétats
- Options propres à chaque effet et réponse

Objet bayesreg

P-spline bidimensionnelle

- Pour interaction entre deux variables continues
- Pour surface de lissage spatiale
- Produit tensoriel de deux B-splines avec mêmes noeuds
  1. $f(x,y) = \sum_{i=1}^{M} \sum_{j=1}^{M} \beta_{ij} B_i(x) B_j(w)$
  2. $\beta_{ij} | . \sim N\left(\frac{1}{4}\left(\beta_{i-1,j} + \beta_{i+1,j} + \beta_{i,j-1} + \beta_{i,j+1}\right), \frac{\tau^2}{4}\right)$

Objet bayesreg

option

- `burnin` = X, `step` = X, `iterations` = X
- `family` : type de réponse
- `level1` = 95, `level2` = 80 : intervalles de confiance
- `predict` : déviance, DIC, fichier des prédictions
- ...

Objet remlreg

Méthode	Commentaire	Bugs
`regress`	Idem `bayesreg`	N
	nomobjet.`regress` *model* [`weight` *nomvar*] [`if` *expr*] [, *options*] `using` *dataset*	

Objet `remlreg`

model

- De la forme $depvar = term_1 + \cdots + term_r$
- *depvar* univariée uniquement
- Effets
  - Identiques à `bayesreg`
  - Données catégorielles ordonnées, cumulées
  - `geokriging` : GRF stationnaire sur les centroïdes
- Interactions $\simeq$ identiques à `bayesreg`
- Réponses identiques à `bayesreg` mais
  - $+$ modèles pour données catégorielles ordonnées, cumulées
  - $-$ ZIP, négative-binomiale
- Options propres à chaque effet et réponse

Algorithmes

Gamerman D. Efficient sampling from the posterior distribution in generalized mixed models. *Statistics and Computing* 1997(**7**) :57–68.

Knorr-Held L. Conditional prior proposals in dynamic models. *Scandinavian Journal of Statistics* 1999(**26**) :129–44.

Plan

Philosophie générale

BayesX

Interface minimale mais fichiers en sortie

- Graphiques postscripts des effets, autocorrélations, historiques
- Fichier .tex et .dvi résumé du modèle estimé

GeoBUGS

Logiciel intégré avec de nombreux outils

- Diagnostics de convergence et stationnarité
- Statistiques sur les estimations
- Cartographie

BayesX est dévolu aux modèles StAR

BayesX

- Lissage en routine
- Nombre d'algorithmes McMC limité
- Structure hiérarchique imposée, pas de choix de priors

GeoBUGS

- Lissage difficile : matrices design issues de R
- Choix possible de nombreux algorithmes
- Langage de programmation souple et puissant (modèles)

Priors spatiaux

BayesX

- GMRF (`spatial`)
- MRF avec corrélation Matérn (`geokriging`)
- P-spline bidimensionnelle (`geospline`)

GeoBUGS

- CAR intrinsèque (`car.normal`), robuste (`car.L1`), propre (`car.proper`) et multivarié (`mv.car`)
- Kriging gaussien (`spatial.exp` et `spatial.disc`)
- Prédiction (`spatial.pred` et `spatial.unipred`)
- Moyenne mobile Poisson-Gamma (`poisson.cov`)

Format de cartes

BayesX

- Boundary
- Graph (équivalent à une matrice d'adjacence)

GeoBUGS

- Polygones
- S+©
- ArcInfo©, EpiMap© et ArcView©

D'autres différences

BayesX

- Univarié
- Pas de choix de départ des chaînes, une seule chaîne
- Manipulation de données
- Avantage : rapidité

GeoBUGS

- Univarié et multivarié
- WinBUGS, OpenBUGS
- Nombre d'utilisateurs importants, nombreuses contributions
- Pas de manipulation de données
- Avantage : lenteur

L'exemple

Cancer des lèvres en Ecosse

- Un exemple historique de GeoBUGS
- 56 comtés dont 3 îles
- Données
  1. Observés par comté : O_i
  2. Attendus (ajustés sur âge et sexe) : E_i
  3. Pourcentage de la population dans l'agriculture, la pêche ou les travaux forestiers : X_i
  4. Liste des adjacences
- Modèle des données $O_i \sim P(\mu_i)$ et $\log(\mu_i) = \log(E_i) + \alpha_0 + \alpha_1 \frac{x_i}{10} + b_i$
- Prior ICAR sur b_i

Préalable

Le problème des îles

- Pas d'adjacence $\Rightarrow$ supprimer les îles
- Dans GeoBUGS : supprimer les lignes des comtés 6, 8 et 11 dans la matrice d'adjacence
- Dans BayesX : supprimer les polygones des comtés 6, 8 et 11 dans le fichier boundary

Le code GeoBUGS

```
# Likelihood
 for (i in 1 : N) {
  O[i] ~ dpois(mu[i])
  log(mu[i]) <- log(E[i]) + alpha0 + alpha1*X[i]/10 + b[i]
 }
# CAR prior distribution for random effects:
 b[1:N] ~ car.normal(adj[],weights[],num[],tau)
 for(k in 1:sumNumNeigh) { weights[k] <- 1 }
# Other priors:
 alpha0  ~ dflat()
 alpha1 ~ dnorm(0.0, 1.0E-5)
 tau  ~ dgamma(0.5, 0.0005)
```

Le code BayesX

```
dataset d
d.infile using scotland.txt
map mappa
mappa.infile using scot.bnd
d.generate logE=log(Expected)
d.generate X10=Agri/10
bayesreg b
b.outfile=M0
b.regress Observed = District(spatial,map=mappa) + X10
          + logE(offset), predict family=poisson
          burnin=1000 iterations=11000 step=10 using d
b.autocor
b.plotautocor, outfile=test.ps
b.getsample
```

Les résultats

- GeoBUGS

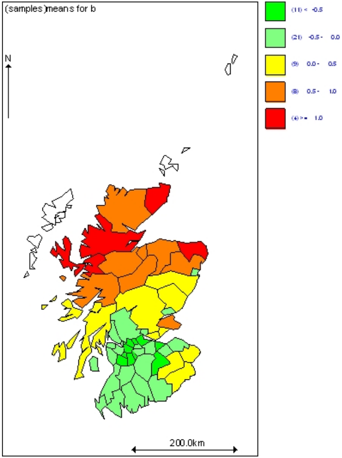
	mean	sd	MC_error	val2.5pc	median	val97.5pc
alpha0	-0.2934	0.1122	0.003409	-0.5137	-0.2924	-0.06169
alpha1	0.4509	0.1162	0.003403	0.2201	0.4518	0.6745
b[1]	1.136	0.3062	0.009266	0.5367	1.14	1.736
...						

- BayesX

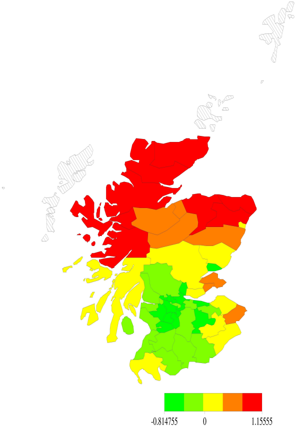
Variable	Mean	STD	2.5%-Qu.	Median	97.5%-Qu.
const	-0.252625	0.131151	-0.490781	-0.253055	0.000492825
X10	0.344362	0.139438	0.066433	0.347261	0.609835

Effet spatial ICAR

(a) GeoBUGS



(b) BayesX



Crainiceanu CM, Ruppert D, Wand MP. Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software* 2005(**14**) :14.

Clayton D, Kaldor J. Empirical bayes estimates of age-standardized relative risks for use of disease mapping. *Biometrics* 1987(**43**) :671–81.

Plan

Introduction

Objectifs

- Représentation de variations spatiales ("disease mapping")
- Etudes des corrélations géographiques : parallèle entre facteurs de risque et variations géographiques = études d'observation
- Recherche d'agrégation (clustering) et d'agrégats (cluster)
- Recherche d'agrégats autour d'un point source

épidémiologie ⇒ peu de cas observés par zone

cf. "Un modèle hiérarchique à composante spatiale partagée pour l'analyse du risque de tremblante du mouton", AppliBUGS 27/11/2008, Sophie Ancelot.

Niveau géographique de recueil

Données ponctuelles

- Localisation "exacte"
- Problèmes des données démographiques
- Mesures d'exposition individuelles ou caractéristiques d'un groupe

Données groupées

- Le plus courant
- Agrégation administrative = unités géographiques
- Grande variation des populations à risque entre UG
- Lien individuel entre exposition et effet sur la santé perdu : biais écologique
- Objet de l'inférence : risque relatif par UG

La spécification des effets pour données groupées

Méthodes classiques

- $o_i|\theta_i \sim P(e_i\theta_i)$
- La variable d'intérêt est le risque relatif par UG
- Le MLE de θ_i est $\frac{o_i}{e_i}$ (SMR ou SIR) mais $\widehat{var}(\hat{\theta}_i) = \frac{o_i}{e_i^2}$
- On peut construire un intervalle de confiance autour de θ_i avec $\theta_i \exp\left(\pm\frac{1{,}96}{\sqrt{o_i}}\right)$
- Il n'est pas sûr que l'estimation individuelle de chaque θ_i conduise à la meilleure représentation de l'ensemble des θ_i

Intérêt des méthodes bayésiennes

- Réduire la variabilité d'ensemble
- Partager l'information des différentes unités géographiques

Régression écologique

$$\ln(o_i) = \ln(e_i) + \mu + \boldsymbol{\beta} \boldsymbol{W}_i + \phi_i$$

- Spécification de ϕ type BYM
- Risque relatif ajusté $\exp(\beta_j)$, supposé constant sur les UG
- Hétérogénéité spatiale $\boldsymbol{\beta} \boldsymbol{W}_i$ avec prior spatial type BYM sur les β

Cancer de la prostate

- Le plus fréquent des cancers chez l'homme et la seconde de décès par cancer
- Rôle des métastases (os) sur la survie
- Données du Registre des Cancers du Haut-Rhin
  - Diagnostics entre 1988 et 1999
  - 2.494 cas (moyenne d'âge de 72 ans)

Survie brute

	#	1 an	2 ans	5 ans
Total	2.494	0,86 [84-87]	0,76 [74-77]	0,54 [52-56]
Métastases au diagnostic				
Sans	2.292	0,88 [87-90]	0,79 [77-81]	0,57 [55-59]
Os	76	0,49 [38-60]	0,24 [15-34]	0,08 [03-14]
Multiple	34	0,35 [20-51]	0,24 [11-39]	0,06 [01-17]
Autre	92	0,73 [63-81]	0,60 [49-69]	0,44 [34-54]

Modèle des données

1. $h(t_j|z_j) = h_0(t_j) \times \exp(z_j'\beta)$

1. Modèle de Cox

Modèle des données

$$
\begin{array}{lllll}
1. & h(t_j|z_j) & = & h_0(t_j) & \times & \exp(z_j'\beta) \\
2. & h(t_j|x_j, j \in i) & = & h_0(t_j) & \times & \exp(x_j'\beta + u_{j\in i})
\end{array}
$$

1. Modèle de Cox
2. Modèle de fragilité

Modèle des données

$$
\begin{array}{llcllcl}
1. & h(t_j|\boldsymbol{z}_j) & = & h_0(t_j) & \times & \exp(\boldsymbol{z}_j'\boldsymbol{\beta}) \\
2. & h(t_j|\boldsymbol{x}_j, j \in i) & = & h_0(t_j) & \times & \exp(\boldsymbol{x}_j'\boldsymbol{\beta} + u_{j\in i}) \\
3. & \log[h(t_j|\boldsymbol{x}_j, j \in i)] & = & \log[h_0(t_j)] & + & \boldsymbol{x}_j'\boldsymbol{\beta} + u_{j\in i}
\end{array}
$$

1. Modèle de Cox
2. Modèle de fragilité
3. Logtransformation

Modèle des données

$$
\begin{array}{llcll}
1. & h(t_j|\boldsymbol{z}_j) & = & h_0(t_j) & \times \quad \exp(\boldsymbol{z}_j'\boldsymbol{\beta}) \\
2. & h(t_j|\boldsymbol{x}_j, j \in i) & = & h_0(t_j) & \times \quad \exp(\boldsymbol{x}_j'\boldsymbol{\beta} + u_{j\in i}) \\
3. & \log[h(t_j|\boldsymbol{x}_j, j \in i)] & = & \log[h_0(t_j)] & + \quad \boldsymbol{x}_j'\boldsymbol{\beta} + u_{j\in i} \\
 & & & & \\
4. & \log[h(t_j|\boldsymbol{z}_j)] & = & \log[h_0(t_j)] & + \quad f_a(a_j) + f_p(p_j) + f_c(c_j) \\
 & & & & \quad\quad +f_{a,p}(a_j, p_j) + f_{a,c}(a_j, c_j) \\
 & & & & \quad\quad +f_{\text{spat}}(j \in i) \\
 & & & & \quad\quad +\gamma_1 \mathtt{m}_{1j} + \gamma_2 \mathtt{m}_{2j} + \gamma_3 \mathtt{m}_{3j}
\end{array}
$$

1. Modèle de Cox
2. Modèle de fragilité
3. Logtransformation
4. Modèle mixte géoadditif

Lois *a priori*

$$\log[h_0(t_j)] + f_a(a_j) + f_p(p_j) + f_c(c_j)$$

- Logbaseline et effet age-période-cohorte
  - P-splines bayésiennes : $f(x) = \sum_{k=1}^{m} \alpha_k B_k(x)$
  - B-splines cubiques avec 20 noeuds
  - Pas aléatoires d'ordre 2 sur les αs avec des erreurs $N(0, \frac{1}{\tau_\alpha})$
  - Prior gamma diffus sur τ_α

Lois *a priori*

$$\log[h_0(t_j)] + f_a(a_j) + f_p(p_j) + f_c(c_j) + f_{a,X}(a_j, X_j)$$

- Logbaseline et effet age-période-cohorte : P-splines bayésiennes
- Effect conjoint âge-période ou âge-cohorte
  - P-splines bayésiennes bidimensionnelles : $f(x,y) = \sum_{k=1}^{m} \sum_{l=1}^{m} \pi_{kl} B_k(x) B_l(y)$
  - B-splines cubiques avec 20 noeuds
  - Pas aléatoires d'ordre 1 sur les πs avec des erreurs $N(0, \frac{1}{\tau_\pi})$
  - Prior gamma diffus sur τ_π

Lois *a priori*

$$\log[h_0(t_j)] + f_a(a_j) + f_p(p_j) + f_c(c_j) + f_{a,X}(a_j, X_j) + f_s(j \in i)$$

- Logbaseline et effet age-période-cohorte : P-splines bayésiennes
- Effect conjoint âge-période ou âge-cohorte : P-splines 2D
- Effet spatial
  - ICAR de précision τ_ϕ
  - Avec ou sans une normale d'hétérogénéité $\theta_i \sim N(0, \frac{1}{\tau_\theta})$
  - Prior gamma diffus sur τ_ϕ and τ_θ

Lois *a priori*

$$\log[h_0(t_j)]+f_a(a_j)+f_p(p_j)+f_c(c_j)+f_{a,X}(a_j,X_j)+f_s(j \in i)+\sum_1^3 \gamma_k \mathtt{m}_{kj}$$

- Logbaseline et effet age-période-cohorte : P-splines bayésiennes
- Effect conjoint âge-période ou âge-cohorte : P-splines 2D
- Effet spatial : ICAR $\pm$ normal
- Effet des métastases
    - Variables indicatrices $\mathtt{m}_1$ (multiple metastases), $\mathtt{m}_2$ (bone metastasis), $\mathtt{m}_3$ (other metastases)
    - Prior normal vague sur les γs

Lois *a priori*

$$\log[h_0(t_j)]+f_a(a_j)+f_p(p_j)+f_c(c_j)+f_{a,X}(a_j,X_j)+f_s(j\in i)+\sum_1^3\gamma_k \mathtt{m}_{kj}$$

- Logbaseline et effet age-période-cohorte : P-splines bayésiennes
- Effect conjoint âge-période ou âge-cohorte : P-splines 2D
- Effet spatial : ICAR $\pm$ normal
- Effet des métastases : effets fixes

Implémentation - Choix d'un modèle

- Inférence par McMC : Metropolis block-update et Gibbs sampling
- Software BayesX
- Burn-in et thinning basés sur les autocorrelations et les diagnostics de Geweke
- Choix d'un meilleur modèle base sur le $DIC = \bar{D} + p_D$
  1. Adéquation au données
  2. Nombre effectif de paramètres
- Intervalles de crédibilité à 5%
- Rapport de risques ajustés

Le meilleur des modèles testés

Modèle	DIC (ajouter 26,000)	p_D
Effets de base		
Baseline seule	769	4.5
$+f_a(a_j)$	394	9.4
$+\sum^3 \gamma_k \mathtt{m}_{kj}$	598	7.5
$+\sum^3 \gamma_k \mathtt{m}_{kj} + f_a(a_j)$	216	11.4

Le meilleur des modèles testés

Modèle	DIC (ajouter 26,000)	p_D
Effets de base		
$log[h_0] + \sum^3 \gamma_k \mathtt{m}_{kj} + f_a(a_j)$	216	11.4
Ajouter les effets période et cohorte		
$+f_c(c_j)$	203	14.5
$+f_p(p_j)$	193	16.0
$+f_c(c_j) + f_{a,c}(a_j, c_j)$	201	14.6
$+f_p(p_j) + f_{a,p}(a_j, p_j)$	191	16.4

Le meilleur des modèles testés

Modèle	DIC (ajouter 26,000)	p_D
Effets de base		
$log[h_0] + \sum^3 \gamma_k \mathtt{m}_{kj} + f_a(a_j)$	216	11.4
Ajouter les effets période et cohorte		
$+f_p(p_j) + f_{a,p}(a_j, p_j)$	191	16.4
Ajouter les effets spatiaux		
$+ICAR(\tau_\phi)$	189	27.0
$+ICAR(\tau_\phi) + N\left(0, \frac{1}{\tau_\theta}\right)$	188	34.3

Effets fixes

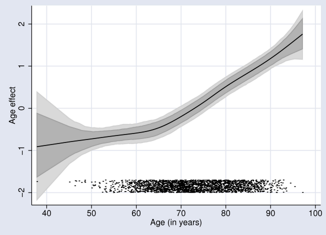
- Effet des métastases

	Posterior			
Métastases	Moyenne	Ecart-type	aHR	Cox HR
Multiple	1,81	0,18	6,11	6,61
Os	1,52	0,13	4,57	4,86
Autre	0,43	0,13	1,54	1,28

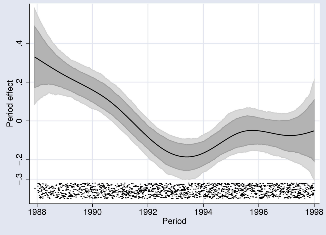
- Rapport de risques ajustés
  - Référence : absence de métastase
  - A comparer avec un modèle Cox (effet des métastases seul)

P-splines : baseline et APC

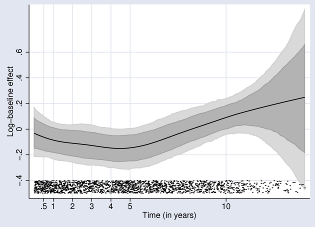
(c) Age



(d) Période

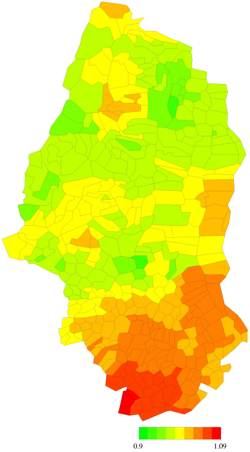


(e) Baseline

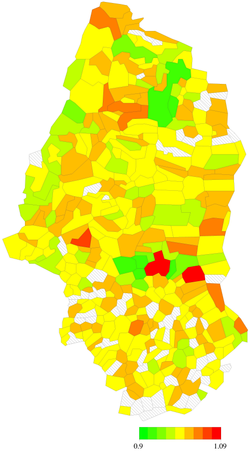


Effets spatiaux

(f) ICAR



(g) Normal



Cressie NAC. *Statistics for spatial data*. Wiley series in probability and mathematical statistics, John Wiley and Sons, New York, 1993.

Banerjee S, Carlin BP, Gelfand AE. *Hierarchical modeling and analysis for spatial data*. Monographs on statistics and applied probability 101, Chapman and Hall / CRC, Boca Raton, 2004.

Best N, Richardson S, Thomson A. A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research* 2005(**14**) :35–59.

Spiegelhalter D, Best N, Carlin B, Van der Linde A. Bayesian measures of model complexity and fit (with discussion) *Journal of the Royal Statistical Scoiety Series B* 2002(**64**) :583–639.

Hennerfeind A, Brezger A, Fahrmeir L. Geoadditive Survival Models. *Journal of the American Statistical Association* 2006(**101**) :1065–75.

Sauleau EA, Hennerfeind A, Buemi A, Held L. Age, period and cohort effects in Bayesian smoothing of spatial cancer survival with geoadditive models. *Statistics in Medicine* 2007(**26**) :212-29.

Merci de votre attention