Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Estimation sur petits domaines hiérarchique bayésienne sous un modèle spatial avec application à des données d'enquête sur la santé

par Yong You et Qian M. Zhou

Juin 2011

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2011
Vol. 37, N° 1, pp. 31-44
Statistique Canada, N° 12-001-X au catalogue

Estimation sur petits domaines hiérarchique bayésienne sous un modèle spatial avec application à des données d'enquête sur la santé

Yong You et Qian M. Zhou [1]

Résumé

Dans le présent article, nous étudions l'estimation sur petits domaines en nous servant de modèles au niveau du domaine. Nous considérons d'abord le modèle de Fay-Herriot (Fay et Herriot 1979) pour le cas d'une variance d'échantillonnage connue lissée et le modèle de You-Chapman (You et Chapman 2006) pour le cas de la modélisation de la variance d'échantillonnage. Ensuite, nous considérons des modèles spatiaux hiérarchiques bayésiens (HB) qui étendent les modèles de Fay-Herriot et de You-Chapman en tenant compte à la fois de l'hétérogénéité géographiquement non structurée et des effets de corrélation spatiale entre les domaines pour le lissage local. Les modèles proposés sont mis en œuvre en utilisant la méthode d'échantillonnage de Gibbs pour une inférence entièrement bayésienne. Nous appliquons les modèles proposés à l'analyse de données d'enquête sur la santé et comparons les estimations fondées sur le modèle HB aux estimations directes fondées sur le plan. Nos résultats montrent que les estimations fondées sur le modèle HB ont de meilleures propriétés que les estimations directes. En outre, les modèles spatiaux au niveau du domaine proposés produisent des CV plus petits que les modèles de Fay-Herriot et de You-Chapman, particulièrement pour les domaines ayant trois domaines voisins ou plus. Nous présentons aussi une comparaison des modèles bayésiens et une analyse de l'adéquation du modèle.

Mots clés : Modèle au niveau du domaine ; comparaison de modèles bayésiens ; taux de prévalence de la maladie ; échantillonnage de Gibbs ; modèle spatial hiérarchique ; vérification du modèle prédictif *a posteriori* ; variance d'échantillonnage.

1. Introduction

Les méthodes d'estimation sur petits domaines fondée sur un modèle ont été utilisées à grande échelle en pratique en raison de la demande croissante d'estimations précises pour des régions locales et divers petits domaines. En général, les enquêtes par sondage sont conçues pour fournir des estimations fiables pour de grandes régions ou des agrégats de petits domaines, tel que le pays dans son ensemble et les provinces. Des estimations par sondage directes fondées uniquement sur des données d'échantillon propres au domaine fournissent habituellement des estimations fiables du paramètre d'intérêt pour les grands domaines. Pour les petits domaines, particulièrement certaines petites régions géographiques ou des petits domaines particuliers, les estimations directes sont susceptibles de produire de grandes erreurs-types, à cause de la petite taille des échantillons dans ces petits domaines. Par conséquent, en inférence pour les petits domaines, il est nécessaire d'emprunter de l'information à des domaines connexes pour produire des estimations indirectes qui augmentent les tailles effectives d'échantillons et donc, la précision des estimations. Il est généralement reconnu aujourd'hui que ces estimations indirectes doivent être fondées sur des modèles explicites qui fournissent des liens avec les domaines connexes grâce à l'utilisation de données supplémentaires, telles que des chiffres de recensement ou des données de dossiers administratifs ; voir, par exemple, Rao (2003) et Jiang et Lahiri (2006) pour une discussion plus approfondie des méthodes d'estimation sur petits domaines fondée sur un modèle. Les estimations fondées sur un modèle sont obtenues en vue d'améliorer les estimations directes fondées sur le plan en ce qui a trait à la précision et à la fiabilité, c'est-à-dire réduire les coefficients de variation (CV). Ces modèles d'estimation sur petits domaines se répartissent en deux grandes catégories, à savoir les modèles au niveau du domaine et les modèles au niveau de l'unité. Les modèles au niveau du domaine sont fondés sur des estimations directes au niveau du domaine et les modèles au niveau de l'unité, sur des observations individuelles faites dans les petits domaines. Dans le présent article, nous nous concentrons sur les modèles au niveau du domaine qui empruntent de l'information à diverses régions pour améliorer les estimations directes sous le plan.

Parmi les modèles au niveau du domaine, celui de Fay-Herriot (Fay et Herriot 1979) est un modèle élémentaire très utilisé en pratique pour obtenir des estimations fondées sur un modèle fiables pour de petits domaines. Fondamentalement, le modèle de Fay-Herriot possède deux composantes, à savoir un modèle d'échantillonnage pour les estimations directes et un modèle de lien pour les paramètres d'intérêt. Le modèle d'échantillonnage comprend l'estimation par sondage direct et la variance d'échantillonnage correspondante. Le modèle de Fay-Herriot repose sur l'hypothèse que la variance d'échantillonnage est connue dans le modèle. Habituellement, on obtient un estimateur lissé de la variance

1. Yong You, Division de la recherche et de l'innovation en statistiques, Statistique Canada. Courriel : yongyou@statcan.gc.ca ; Qian M. Zhou, Département de biostatistique, Université Harvard.

d'échantillonnage que l'on traite ensuite comme étant connue dans le modèle. Wang et Fuller (2003), ainsi que You et Chapman (2006) ont étudié la situation où les variances d'échantillonnage sont inconnues et modélisées séparément par des estimateurs directs. Dans le présent article, nous examinerons les méthodes de lissage et de modélisation pour les variances d'échantillonnage dans le modèle d'échantillonnage.

Le modèle de lien relie le paramètre d'intérêt à un modèle de régression contenant des effets aléatoires propres au domaine. Dans le modèle de Fay-Herriot, on suppose habituellement que les effets aléatoires de domaine sont des variables aléatoires indépendantes et identiquement distribuées (*iid*) qui suivent une loi normale pour refléter les variations géographiquement non structurées entre les domaines. Cependant, dans certaines applications aux petits domaines, particulièrement dans les problèmes d'estimation de la santé publique, la variation géographique de la prévalence d'une maladie est un sujet d'intérêt, et l'estimation des profils spatiaux globaux de risque et l'emprunt d'information à diverses régions pour réduire les variances des estimations finales sont deux éléments importants. Donc, il pourrait être plus raisonnable de construire des modèles spatiaux sur les effets aléatoires propres aux domaines pour traduire l'interdépendance spatiale de ces effets. Les modèles spatiaux sont généralement utilisés pour l'estimation sur petits domaines ayant trait à la santé, et plusieurs de ces modèles ont été proposés (par exemple, Cressie 1990 ; Ghosh, Natarajan, Stroud et Carling 1998 ; Maiti 1998 ; Ghosh, Natarajan, Walter et Kim 1999 ; He et Sun 2000 ; Moura et Migon 2002 ; Singh, Shukla et Kundu 2005 ; Souza, Moura et Migon 2009). Best, Richardson et Thomson (2005) ont présenté une revue complète des modèles spatiaux pour la cartographie des maladies. Rao (2003) a également discuté de plusieurs modèles spatiaux pour petits domaines.

L'objectif du présent article est d'examiner des modèles à corrélation spatiale pour l'estimation sur petits domaines et d'illustrer leur utilité en les appliquant à des données d'enquête sur la santé. La présentation de l'article est la suivante. À la section 2, nous commençons par étudier les modèles au niveau du domaine, y compris le modèle de Fay-Herriot et les modèles de lien à corrélation spatiale. Ensuite, à la section 3, nous proposons des modèles d'estimation sur petits domaines hiérarchiques bayésiens (HB) avec corrélation spatiale et obtenons l'inférence HB pour les paramètres de petit domaine grâce à la méthode d'échantillonnage de Gibbs. À la section 4, nous appliquons les modèles proposés à l'analyse de données sur des petits domaines provenant de l'Enquête sur la santé dans les collectivités canadienne. Nous comparons la performance des estimations fondées sur un modèle aux estimations directes fondées sur le plan et, en outre, nous comparons les modèles proposés au modèle de Fay-Herriot et au modèle de You-Chapman (You et Chapman 2006) afin d'étudier les effets de la prise en compte de la structure spatiale sur les effets aléatoires de domaine. Nous présentons aussi une comparaison des modèles bayésiens et une analyse de l'adéquation du modèle. Enfin, à la section 5, nous tirons certaines conclusions.

2. Modèles et inférence pour petits domaines

2.1 Modèle de Fay-Herriot

Soit θ_i le paramètre d'intérêt pour le i^{e} domaine, où $i = 1, ..., m$, et m est le nombre total de domaines. Le modèle de Fay-Herriot repose sur l'hypothèse que les θ_i sont reliés à des données auxiliaires particulières au domaine $\mathbf{x}_i = (x_{i1}, ..., x_{ip})'$ au moyen du modèle de régression linéaire qui suit :

$$\theta_i = \mathbf{x}_i'\beta + v_i,\ i = 1, ..., m \qquad (1)$$

où $\beta = (\beta_1, ..., \beta_p)'$ est le vecteur de dimension $p \times 1$ des coefficients de régression, et les v_i sont les effets aléatoires propres au domaine que l'on suppose être *iid* avec $\mathrm{E}(v_i) = 0$ et $\mathrm{Var}(v_i) = \sigma_v^2$. L'hypothèse de normalité peut également être incluse. Ce modèle est appelé modèle de lien pour θ_i. Le modèle de Fay-Herriot repose aussi sur l'hypothèse qu'un estimateur direct fondé sur le plan y_i, qui est habituellement sans biais sous le plan pour le paramètre d'intérêt θ_i, existe quand la taille d'échantillon du domaine $n_i > 1$. On suppose habituellement que

$$y_i = \theta_i + e_i,\ i = 1, ..., m \qquad (2)$$

où les e_i sont les erreurs d'échantillonnage associées à l'estimateur direct y_i. Nous supposons en outre que les e_i sont des variables aléatoires normales indépendantes de moyenne $E(e_i \mid \theta_i) = 0$ et de variance d'échantillonnage $\mathrm{Var}(e_i \mid \theta_i) = \sigma_i^2$. Le modèle (2) est appelé modèle d'échantillonnage pour l'estimateur direct y_i. La combinaison des deux composantes (1) et (2) donne un modèle linéaire à effets mixtes, le modèle de Fay-Herriot, de la forme

$$y_i = \mathbf{x}_i'\beta + v_i + e_i,\ i = 1, ..., m. \qquad (3)$$

Dans le modèle de Fay-Herriot élémentaire (3), on suppose habituellement que les variances d'échantillonnage σ_i^2 sont connues, ce qui est une hypothèse très forte. En général, nous pouvons utiliser les estimations directes des variances d'échantillonnage d'après les données d'enquête, mais ces estimations sont instables si les tailles d'échantillon sont faibles. Par conséquent, en pratique, on introduit dans le modèle un estimateur lissé de σ_i^2 que l'on traite alors comme étant connue. La fonction de variance généralisée est habituellement appliquée en pratique pour obtenir un estimateur lissé de la variance d'échantillonnage (par exemple, Dick 1995). Ces dernières

années, une méthode de lissage des effets de plan a été élaborée et utilisée en pratique pour obtenir des estimateurs de variance lissés (par exemple, Singh, Folsom et Vaish 2005 ; You 2008a ; Liu, Lahiri et Kalton 2008). En particulier, You (2008a) a appliqué une approche de modélisation à effets de plan égaux pour obtenir des estimations lisses des variances d'échantillonnage. L'effet de plan pour le i^{e} domaine peut s'écrire approximativement sous la forme

$$\text{deff}_i = \frac{s_i^2}{s_{ri}^2}, \text{ pour } i = 1, ..., m,$$

où s_i^2 est l'estimation directe sans biais de la variance d'échantillonnage fondée sur le plan d'échantillonnage complexe, et s_{ri}^2 est l'estimation de la variance d'échantillonnage sous un plan d'échantillonnage aléatoire simple. Pour chaque domaine, en émettant l'hypothèse d'un effet de plan commun, un facteur deff lissé peut être obtenu en appliquant $\text{deff} = \sum_{i=1}^{m} \text{deff}_i/m$. Ensuite, une estimation lissée de la variance d'échantillonnage $\tilde{\sigma}_i^2$ peut être obtenue sous la forme $\tilde{\sigma}_i^2 = s_{ri}^2 \cdot \text{deff}$.

Au lieu d'introduire les estimations lissées des variances d'échantillonnage dans le modèle, nous pouvons modéliser directement ces variances. Dans Wang et Fuller (2003) et You et Chapman (2006), ces auteurs supposent que la variance d'échantillonnage σ_i^2 est inconnue et estiment σ_i^2 au moyen d'un estimateur direct sans biais s_i^2, qui est indépendant de l'estimateur direct sous le plan y_i. Ils supposent aussi que $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$, où $d_i = n_i - 1$ et n_i est la taille d'échantillon pour le i^{e} domaine. You et Chapman (2006) ont examiné l'approche HB complète avec la méthode d'échantillonnage de Gibbs qui tient compte automatiquement de l'incertitude supplémentaire associée à l'estimation de σ_i^2. Dans le présent article, nous considérons à la fois les approches de lissage et de modélisation pour les variances d'échantillonnage.

2.2 Modèles spatiaux

En vue d'intégrer les effets aléatoires spatialement corrélés dans le modèle de lien, un moyen simple et évident consiste à ajouter un effet aléatoire spatial u_i dans le modèle de lien indépendant (1) comme il suit :

$$\theta_i = \mathbf{x}_i' \beta + v_i + u_i, \qquad (4)$$

où les u_i suivent le modèle conditionnel autorégressif (CAR) intrinsèque bien connu donné par

$$u_i \mid u_{-i} \sim \text{N}\left(\frac{\sum_{j \neq i} w_{ij} u_j}{\sum_{j \neq i} w_{ij}}, \frac{\sigma_u^2}{\sum_{j \neq i} w_{ij}} \right), \qquad (5)$$

où u_{-i} désigne les valeurs des effets aléatoires spatiaux u_j dans tous les autres domaines avec $j \neq i$, les poids w_{ij} sont des constantes fixées et σ_u^2 est une composante de variance inconnue. En pratique, un choix fréquent pour w_{ij} consiste à poser que $w_{ij} = 0$ à moins que les domaines i et j soient voisins (c'est-à-dire aient une limite commune), auquel cas $w_{ij} = 1$. Le modèle (4) est proposé par Besag, York et Mollie (1991) pour séparer les effets spatiaux de l'hétérogénéité globale dans les domaines. Dans le modèle (4), les effets aléatoires indépendants v_i traduisent l'hétérogénéité géographiquement non structurée entre les domaines, et les effets aléatoires spatiaux u_i traduisent la dépendance spatiale entre les domaines. De cette façon, le degré de dépendance spatiale globale peut être exprimé en se basant sur la proportion de la variation totale dans les $v_i + u_i$ reflétées par chaque composante.

En pratique, le choix entre un modèle non structuré (par exemple, le modèle de lien élémentaire) donné par (1) et un modèle purement structuré spatialement (par exemple, le modèle autorégressif intrinsèque) donné par (5) n'est souvent pas clair. Pour le modèle (4), l'inférence *a posteriori* au sujet de la dépendance spatiale est fondée sur la proportion de la variation totale de la somme de $v_i + u_i$ reflétée par chaque composante. Cependant, bien que les lois conditionnelles univariées de la composante spatiale (5) soient bien définies, la loi conjointe correspondante est impropre (de moyenne non définie et de variance infinie). De surcroît, le modèle (4) pose un problème éventuel d'identifiabilité où seule la somme des effets aléatoires $v_i + u_i$ est bien définie par les données ; voir, par exemple, Best et coll. (2005), pour une discussion plus détaillée.

Nous pouvons également considérer une autre paramétrisation spatiale étudiée par Leroux, Lei, et Breslow (1999) et par MacNab (2003), qui permet d'éviter le problème d'identifiabilité qui se pose avec le modèle (4). Soit $\theta_i = \mathbf{x}_i' \beta + b_i$, et $\mathbf{b} = (b_1, ..., b_m)'$. À l'instar de Leroux et coll. (1999) et de MacNab (2003), nous appliquons le modèle conditionnel autorégressif (CAR) qui suit aux effets spatiaux de domaine $\mathbf{b} = (b_1, ..., b_m)'$:

$$\mathbf{b} \sim \text{MVN}(\mathbf{0}, \Sigma(\sigma_b^2, \lambda)) \qquad (6)$$

$$\Sigma(\sigma_b^2, \lambda) = \sigma_b^2 \mathbf{D}^{-1}, \ \mathbf{D} = \lambda \mathbf{R} + (1 - \lambda)\mathbf{I} \qquad (7)$$

où σ_b^2 est un paramètre de dispersion spatiale et λ est un paramètre d'autocorrélation spatiale, $0 \leq \lambda \leq 1$; $\mathbf{I}$ est une matrice identité de dimension m ; $\mathbf{R}$ est une matrice, habituellement appelée matrice de voisinage (*neighbourhood matrix*), dont le i^{e} élément diagonal est égal au nombre de voisins du domaine i, et dont les éléments hors diagonale dans chaque ligne sont égaux à -1 si les domaines correspondants sont voisins et à 0 autrement. Le modèle CAR donné par les expressions (6) et (7) correspond à la loi conditionnelle de b_i suivante :

$$b_i \mid b_{-i} \sim \text{N}\left(\frac{\lambda}{1 - \lambda + \lambda w_{i+}} \sum_{j \neq i} w_{ij} v_j, \frac{\sigma_b^2}{1 - \lambda + \lambda w_{i+}} \right),$$

où $w_{i+} = \sum_{j \neq i} w_{ij}$. Le modèle CAR (6) - (7) devient le modèle autorégressif intrinsèque (5) si $\lambda = 1$. Par ailleurs, si $\lambda = 0$, il se réduit au modèle de lien indépendant (1) dans lequel les effets aléatoires propres au domaine v_i sont supposés indépendants. Il convient de souligner que la moyenne et les variances conditionnelles de $b_i \mid b_{-i}$ sont les sommes pondérées des moments lissés globaux provenant du modèle de lien élémentaire (1) et des moments lissés locaux provenant du modèle autorégressif intrinsèque :

$$\mathrm{E}(b_i \mid b_{-i}) = \frac{1-\lambda}{1-\lambda+\lambda w_{i+}} \times 0 + \frac{\lambda w_{i+}}{1-\lambda+\lambda w_{i+}} \left(\sum_{j \neq} w_{ij} b_j / w_{i+} \right)$$

$$\mathrm{Var}(b_i \mid b_{-i}) = \frac{1-\lambda}{1-\lambda+\lambda w_{i+}} \times \sigma_b^2 + \frac{\lambda w_{i+}}{1-\lambda+\lambda w_{i+}} (\sigma_b^2 / w_{i+}).$$

Donc, le modèle (6) - (7) représente un équilibre entre le modèle de lien indépendant (1) et le modèle CAR intrinsèque (5). Le paramètre de corrélation spatiale λ mesure l'importance des effets spatiaux pour le lissage local des domaines voisins. La structure de modélisation (6) traduit à la fois l'hétérogénéité non structurée entre les domaines et les effets de corrélation spatiale du domaine voisin.

2.3 Modèles hiérarchiques bayésiens et inférence

Afin d'estimer θ_i, le paramètre d'intérêt, nous appliquons une approche hiérarchique bayésienne (HB) en utilisant la méthode d'échantillonnage de Gibbs. Comparativement à d'autres approches, telles que l'EBLUP et l'approche empirique bayésienne (EB), l'approche HB est simple et les inférences concernant θ_i sont exactes contrairement aux inférences EB ou EBLUP. En outre, l'approche HB donne le moyen de traiter des modèles pour petits domaines complexes en utilisant la méthode Monte Carlo par chaîne de Markov (MCMC), qui permet de surmonter en grande partie les difficultés que posent le calcul des intégrales multidimensionnelles des quantités *a posteriori*.

Soit $\mathbf{y} = (y_1, ..., y_m)'$, $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)'$, et $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_m)'$. Nous commençons par construire deux modèles HB sans et avec structure spatiale sous l'hypothèse que les variances d'échantillonnage sont connues et remplacées par l'estimation lissée σ_i^2.

Modèle 1 : Modèle de Fay-Herriot, désigné par MFH (Fay et Herriot 1979 ; Rao 2003).

- $y_i \mid \theta_i \sim \mathrm{N}(\theta_i, \sigma_i^2 = \tilde{\sigma}_i^2)$, pour $i = 1, ..., m$;
- $\theta_i \mid \beta, \sigma_v^2 \sim \mathrm{N}(\mathbf{x}_i'\beta, \sigma_v^2)$, pour $i = 1, ..., m$;
- Priors pour les paramètres $(\beta, \sigma_v^2) : \pi(\beta) \propto 1$; $\pi(\sigma_v^2) \sim \mathrm{GI}(a_0, b_0)$, où a_0, b_0 sont des constantes connues choisies très petites pour refléter les connaissances vagues au sujet de σ_v^2. N désigne la loi normale et GI, la loi gamma inverse.

Modèle 2 : Modèle CAR au niveau du domaine proposé, en tant qu'extension du modèle de Fay-Herriot, désigné par CAR-MFH.

- $\mathbf{y} \mid \boldsymbol{\theta} \sim \mathrm{MVN}(\boldsymbol{\theta}, \mathbf{E})$, où $\mathbf{E}$ est une matrice diagonale dont le i^{e} élément diagonal est $\sigma_i^2 = \tilde{\sigma}_i^2$;
- $\boldsymbol{\theta} \mid \beta, \sigma_v^2 \sim \mathrm{MVN}(\mathbf{X}\beta, \sigma_v^2 \mathbf{D}^{-1})$, où $\mathbf{D} = \lambda \mathbf{R} + (1-\lambda)\mathbf{I}$, avec $\mathbf{I}$ désignant une matrice identité de dimension m et $\mathbf{R}$, la matrice de voisinage ;
- Priors pour les paramètres $(\beta, \lambda, \sigma_v^2) : \pi(\beta) \propto 1$; $\pi(\lambda) \sim \mathrm{Uniform}(0,1)$, où $0 \leq \lambda \leq 1$; $\pi(\sigma_v^2) \sim \mathrm{GI}(a_0, b_0)$, où a_0, b_0 sont des constantes connues choisies très petites. MVN désigne la loi normale multivariée.

Il convient de souligner que le modèle CAR-MFH proposé se réduit au modèle MFH lorsque le paramètre d'autocorrélation spatial est $\lambda = 0$.

Nous considérons aussi deux modèles HB dont la variance d'échantillonnage σ_i^2 est inconnue et modélisée par l'estimateur sans biais direct s_i^2.

Modèle 3 : Modèle de You-Chapman désigné par MYC (You et Chapman 2006).

- $y_i \mid \theta_i, \sigma_i^2 \sim \mathrm{N}(\theta_i, \sigma_i^2)$, pour $i = 1, ..., m$;
- $d_i s_i^2 \mid \sigma_i^2 \overset{\mathrm{ind}}{\sim} \sigma_i^2 \chi_{d_i}^2$ où $d_i = n_i - 1$, pour $i = 1, ..., m$;
- $\theta_i \mid \beta, \sigma_v^2 \sim \mathrm{N}(\mathbf{x}_i'\beta, \sigma_v^2)$, pour $i = 1, ..., m$;
- Priors pour les paramètres inconnus $(\beta, \sigma_v^2, \sigma_i^2, i = 1, ..., m) : \pi(\beta) \propto 1$; $\pi(\sigma_v^2) \sim \mathrm{GI}(a_0, b_0)$, $\pi(\sigma_i^2) \sim \mathrm{GI}(a_i, b_i)$ pour $i = 1, ..., m$, où a_i, b_i $(0 \leq i \leq m)$ sont des constantes connues choisies très petites pour refléter les connaissances vagues au sujet de σ_i^2 et σ_v^2.

Modèle 4 : Modèle CAR au niveau du domaine proposé avec variances d'échantillonnage inconnues, en tant qu'extension du modèle de You-Chapman, désigné par CAR-MYC.

- $\mathbf{y} \mid \boldsymbol{\theta}, \sigma_1^2, ..., \sigma_m^2 \sim \mathrm{MVN}(\boldsymbol{\theta}, \mathbf{E})$, où la matrice $\mathbf{E}$ contient les éléments diagonaux σ_i^2 ;
- $d_i s_i^2 \mid \sigma_i^2 \overset{\mathrm{ind}}{\sim} \sigma_i^2 \chi_{d_i}^2$, où $d_i = n_i - 1$, pour $i = 1, ..., m$;
- $\boldsymbol{\theta} \mid \beta, \sigma_v^2 \sim \mathrm{MVN}(\mathbf{X}\beta, \sigma_v^2 \mathbf{D}^{-1})$, où $\mathbf{D} = \lambda \mathbf{R} + (1-\lambda)\mathbf{I}$;
- priors pour les paramètres $(\beta, \lambda, \sigma_v^2, \sigma_i^2, i = 1, ..., m)$: $\pi(\beta) \propto 1$; $\pi(\lambda) \sim \mathrm{Uniform}(0, 1)$, où $0 \leq \lambda \leq 1$; $\pi(\sigma_v^2) \sim \mathrm{GI}(a_0, b_0)$; $\pi(\sigma_i^2) \sim \mathrm{GI}(a_i, b_i)$ pour $i = 1, ..., m$, où a_i, b_i $(0 \leq i \leq m)$ sont des constantes connues choisies très petites.

De nouveau, soulignons que le modèle proposé CAR-MYC se réduit au modèle de You-Chapman quand $\lambda = 0$. Le modèle 3 ainsi que le modèle 4 comportent l'hypothèse implicite que la taille d'échantillon de domaine $n_i \geq 2$. Si des priors plats (*flat*) sont utilisés pour σ_i^2, il faudrait que $n_i \geq 4$ pour être certains que les lois *a posteriori* soient appropriées (You et Chapman 2006).

Nous appliquons la méthode d'échantillonnage de Gibbs pour estimer la moyenne *a posteriori* $\mathrm{E}(\theta_i \mid \mathbf{y})$ et la variance *a posteriori* correspondante $\mathrm{Var}(\theta_i \mid \mathbf{y})$. Les lois conditionnelles complètes des paramètres requises sous les divers modèles sont données à l'annexe A. Pour les modèles de Fay-Herriot et de You-Chapman, toutes les lois conditionnelles complètes possèdent des formes explicites et le tirage d'échantillons à partir de ces lois est simple. Pour les modèles spatiaux au niveau du domaine proposés CAR-MFH et CAR-MYC, la loi conditionnelle du paramètre de corrélation spatiale λ ne possède pas de forme explicite. Nous utilisons l'algorithme de Metropolis-Hastings avec l'échantillonneur de Gibbs (Chip et Greenberg 1995) pour mettre à jour λ. Sous le modèle CAR-MFH, la loi conditionnelle complète de λ dans l'échantillonneur de Gibbs peut s'écrire sous la forme

$$[\lambda \mid \boldsymbol{\theta}, \beta, \sigma_v^2] \propto \mathrm{h}(\lambda)\,\mathrm{f}(\lambda)$$

où $\mathrm{f}(\lambda)$ est une densité de probabilité de la loi uniforme Uniform (0, 1) donnée par

$$\mathrm{f}(\lambda) \propto 1, \text{ où } 0 \leq \lambda \leq 1$$

et $\mathrm{h}(\lambda)$ est une fonction donnée par

$$\mathrm{h}(\lambda) \propto \left|\left[\lambda \mathbf{R} + (1-\lambda)\mathbf{I}\right]^{-1}\right|^{-1/2} \times \exp\left\{-\frac{1}{2\sigma_v^2}(\boldsymbol{\theta} - \mathbf{X}\beta)'\left[\lambda \mathbf{R} + (1-\lambda)\mathbf{I}\right](\boldsymbol{\theta} - \mathbf{X}\beta)\right\}.$$

Nous utilisons $\mathrm{f}(\lambda)$ comme densité de probabilité génératrice « candidate » dans l'étape de mise à jour de l'algorithme de Metropolis-Hastings. Pour mettre λ à jour à partir des valeurs courantes de $(\boldsymbol{\theta}^{(k)}, \beta^{(k)}, \sigma_v^{2(k)})$, la procédure est la suivante :

1. tirer λ^* d'une loi uniforme ;
2. calculer la probabilité d'acceptation $\alpha(\lambda^*, \lambda^{(k)}) = \min\{h(\lambda^*)/h(\lambda^{(k)}), 1\}$;
3. générer u à partir d'une loi uniforme ; si $u < \alpha(\lambda^*, \lambda^{(k)})$, la valeur candidate λ^* est acceptée, c'est-à-dire $\lambda^{(k+1)} = \lambda^*$; sinon, λ^* est rejetée, et fixée à $\lambda^{(k+1)} = \lambda^{(k)}$.

Pour le modèle CAR-MYC, une procédure similaire peut être appliquée au moment du tirage des échantillons à partir de la loi conditionnelle de λ.

3. Analyse des données

3.1 Description des données et mise en œuvre

L'Enquête sur la santé dans les collectivités canadiennes (ESCC) est une enquête fédérale réalisée par Statistique Canada. L'objectif principal de l'ESCC est de fournir des estimations à jour et fiables des déterminants de la santé, de l'état de santé et de l'utilisation du système de santé au Canada. Il s'agit d'une enquête transversale dont le cycle de collecte est de deux ans. La première année de l'enquête, le cycle « x.1 » a pour cible les membres de 12 ans et plus de la population à domicile et est une enquête générale sur la santé de la population réalisée auprès d'un grand échantillon (130 000 personnes) conçu en vue de fournir des estimations fiables au niveau des régions sociosanitaires, des provinces et du Canada. La deuxième année de l'enquête, le cycle « x.2 » est réalisé auprès d'un échantillon plus petit (30 000 personnes), réparti en fonction des achats d'unités d'échantillonnage supplémentaires des provinces et conçu pour fournir des résultats au niveau provincial et national sur des thèmes particuliers liés à la santé. Bien que les estimations nationales et provinciales soient très importantes, la demande de données sur la santé à des niveaux plus fins d'agrégation géographique augmente dans un certain nombre de provinces, dont la Colombie-Britannique (C.-B.), l'Île-du-Prince-Édouard (Î.-P.-É.), le Québec et d'autres. Le cycle « x.1 » de l'ESCC permet de recueillir des données pour 136 régions sociosanitaires réparties dans les dix provinces et les trois territoires. Il est réalisé essentiellement à l'aide de deux bases de sondage. La première, qui est la base principale, est fondée sur la base de sondage aréolaire conçue pour l'Enquête sur la population active du Canada. Cette base aréolaire est utilisée pour échantillonner les logements conformément à un plan d'échantillonnage en grappes stratifiées à plusieurs degrés. La deuxième base de sondage consiste en une liste de numéros de téléphone. La méthode de composition aléatoire est utilisée dans certaines régions sociosanitaires pour des raisons de coût. Des renseignements plus détaillés sur le plan de sondage sont fournis dans Béland (2002). Dans le présent article, nous utilisons un petit ensemble de données provenant du cycle 1.1 pour illustrer l'analyse. Nous voulons estimer le taux de prévalence de la maladie dans les régions sociosanitaires à l'intérieur des provinces. En particulier, nous appliquons les quatre modèles décrits à la section 2 pour estimer le taux de prévalence de l'asthme dans 20 régions sociosanitaires dans la province de la Colombie-Britannique en utilisant les données provenant du cycle 1.1. La figure 1 montre la carte des 20 régions sociosanitaires de la Colombie-Britannique. Nous utilisons cette carte pour définir la matrice de corrélation de voisinage utilisée dans les modèles spatiaux. L'annexe B donne la liste des régions sociosanitaires et les structures spatiales connexes.

Soit θ_i le taux réel de prévalence de l'asthme dans la i^e région sociosanitaire en C.-B., $i = 1, ..., 20$. D'après les données du cycle 1.1 de l'enquête, nous obtenons l'estimation directe y_i de θ_i comme étant le ratio du nombre de personnes asthmatiques (estimation directe) divisé par la taille de population correspondante (constante connue). Nous avons également inclus six variables auxiliaires au niveau du domaine utilisées dans le modèle, à savoir la taille totale de population, le nombre de personnes dont l'asthme est l'un des symptômes de maladie chronique, le nombre de personnes dont l'asthme est le symptôme principal de maladie chronique, le nombre de personnes dont le diabète est l'un des symptômes de maladie chronique, le nombre de personnes dont le diabète est le symptôme principal de maladie chronique, et le nombre de visites à l'hôpital. Notons que, dans la littérature traitant de la cartographique des maladies (par exemple, Mollié 1996 ; Maiti 1998 ; MacNab 2003), la loi de Poisson ou la loi binomiale est habituellement celle qui est supposée dans le modèle d'échantillonnage pour l'estimation directe y_i. Cependant, dans l'estimation sur petits domaines, l'estimation directe y_i s'obtient en se fondant sur le plan d'échantillonnage complexe utilisé dans l'enquête. Donc, on émet habituellement l'hypothèse d'un modèle d'échantillonnage normal pour les estimations directes y_i ; voir, par exemple, Datta, Lahiri, Maiti et Lu (1999), Rao (2003), Mohadjer, Rao, Liu, Krenzke et Van de Kerckhove (2007), et You (2008a). Il convient de souligner que nous n'avons pris en considération qu'un seul type de données sur la prévalence de la maladie provenant d'une seule province dans notre étude, et utilisé cet exemple pour illustrer le modèle proposé et évaluer les effets de la modélisation spatiale dans les modèles pour petits domaines.

Pour appliquer l'échantillonnage de Gibbs, nous utilisons $L = 5$ exécutions parallèles ayant chacune une longueur de « rodage » de $B = 2\,000$ et une taille d'échantillonnage de Gibbs de $G = 5\,000$. Pour les modèles CAR-MFH et CAR-MYC proposés, afin de réduire l'autocorrélation qui résulte de l'algorithme d'acceptation–rejet dans l'exécution, nous prenons une itération sur cinq après la période de « rodage ». Par conséquent, pour les modèles MFH et MYC, nous avons $n = 5\,000$ échantillons pour chaque exécution, et pour les modèles CAR-MFH et CAR-MYC, nous avons $n = 1\,000$ échantillons pour chaque exécution. Nous surveillons la convergence de l'échantillonnage de Gibbs pour le paramètre de petits domaines θ_i et d'autres paramètres inconnus dans le modèle en utilisant le facteur de réduction potentiel d'échelle (*potential scale reduction factor*) (Gelman et Rubin 1992 ; Gelman, Carlin, Stern et Rubin 2004, pages 296-297). Nous avons calculé les facteurs de réduction pour tous les paramètres surveillés dans le modèle dans l'échantillonnage de Gibbs. Les valeurs du facteur sont toutes proches de 1 (inférieures à 1,05), ce qui donne à penser que la convergence souhaitée pour ces paramètres est réalisée par l'échantillonneur de Gibbs.

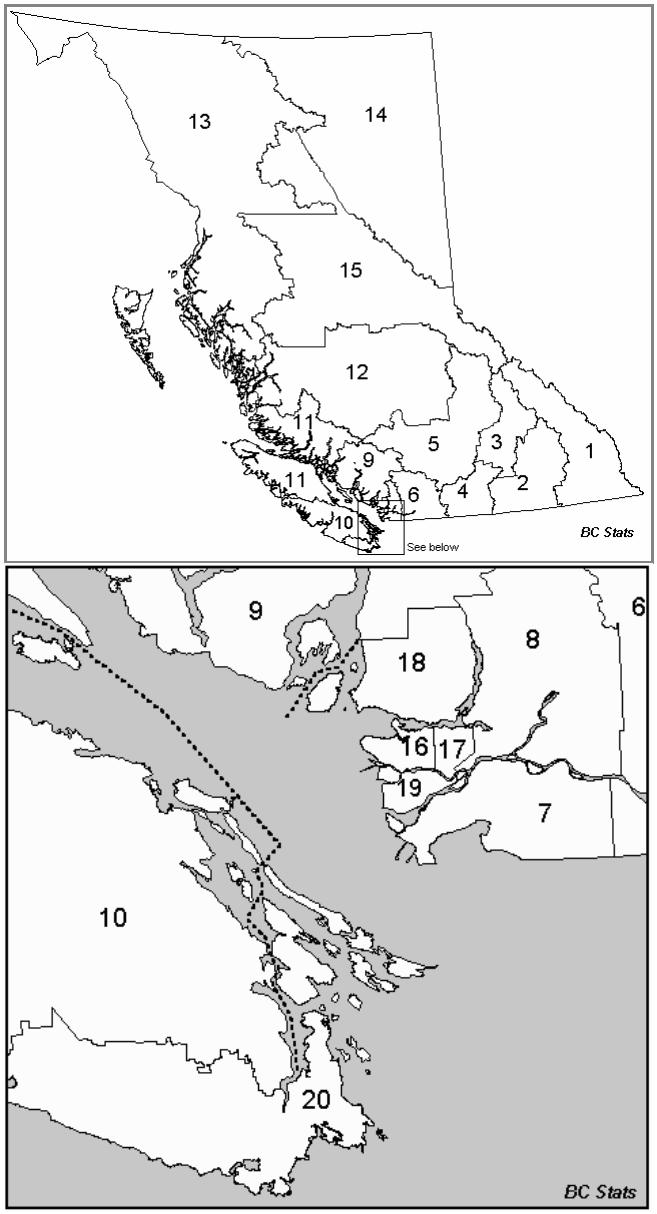

Figure 1 Carte des 20 régions sociosanitaires de la Colombie-Britannique

Nous avons utilisé des priors vagues pour les hyperparamètres du modèle comme il est d'usage en pratique pour l'estimation sur petits domaines de type HB. En particulier, il est fréquent d'utiliser le prior plat pour le paramètre de régression $\pi(\beta) \propto 1$ et les priors gamma inverses appropriés pour les composantes de la variance (par exemple, Arora et Lahiri 1997 ; Ghosh et coll. 1998 ; Datta et coll. 1999 ; You et Rao 2000 ; Rao 2003, page 237 ; Souza et coll. 2009). À l'exemple de MacNab (2003), nous avons

utilisé le prior uniforme $\pi(\lambda) \sim$ pour le paramètre d'autocorrélation. Les priors uniformes sont aussi utilisés fréquemment pour les paramètres d'autocorrélation dans les modèles spatiaux (par exemple, Maiti 1998 ; He et Sun 2000 ; Rao 2003, page 266). Nous avons également essayé plusieurs valeurs différentes pour les priors gamma inverses. Les estimations HB sont assez stables et non sensibles au choix des priors vagues appropriés. Une discussion plus détaillée de l'analyse de sensibilité peut être consultée, par exemple, dans You et Chapman (2006) pour des modèles similaires.

3.2 Comparaison des résultats

Pour commencer, nous présentons les estimations HB du taux de prévalence de l'asthme sous les modèles MFH et CAR-MFH dans lesquels les variances d'échantillonnage σ_i^2 sont supposées connues. Nous avons utilisé l'estimation lissée $\tilde{\sigma}_i^2$ obtenue par la technique de lissage exposée dans You (2008a) comme il est décrit à la section 2. La figure 2 donne les estimations directes et les estimations fondées sur le modèle HB sous les modèles MFH et CAR-MFH pour les 20 régions sociosanitaires de la Colombie-Britannique. Les régions sociosanitaires sont représentées sur l'axe des x, classées par ordre de taille d'échantillon, de la plus petite (Peace Liard) à gauche à la plus grande (South Fraser Valley) à droite. Le modèle 1 (MFH) et le modèle 2 (CAR-MFH) donnent des estimations ponctuelles similaires et les deux estimations fondées sur un modèle mènent à des estimations modérément lisses comparativement aux estimations directes. En outre, les estimations directes et les deux estimations HB du taux de prévalence de la maladie sont très proches pour certaines régions sociosanitaires dont la taille d'échantillon est grande, mais diffèrent dans une certaine mesure pour certaines régions dont la taille d'échantillon est plus petite. Des résultats similaires sont obtenus sous le modèle 3 (MCY) et le modèle 4 (CAR-MCY).

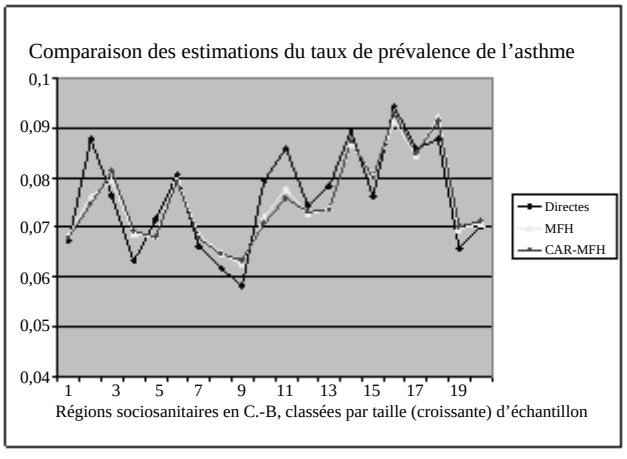

Figure 2 Estimations directes et fondées sur un modèle HB sous les modèles MFH et CAR-MFH

La figure 3 donne les CV des estimations directes et des deux estimations fondées sur un modèle HB pour les régions sociosanitaires classées selon la taille d'échantillon, de la plus petite à la plus grande, comme à la figure 2. Les CV des estimations HB sont obtenus en divisant la racine carrée de la variance *a posteriori* par la moyenne *a posteriori*. Comme prévu, les CV des estimations directes ont nettement tendance à diminuer à mesure que la taille d'échantillon augmente. Par contre, les deux estimations fondées sur un modèle HB donnent des CV plus lisses. En outre, ces deux estimations offrent une amélioration importante par rapport aux estimations directes fondées sur le plan en ce qui a trait à la précision et à la fiabilité, c'est-à-dire la réduction des CV. Comparativement aux estimations directes, la réduction moyenne du CV des estimations HB sous le modèle MFH est de l'ordre de 22,7 %, variant de 7,8 % à 40,5 %, et la réduction moyenne du CV des estimations HB sous le modèle CAR-MFH proposé est de 27,8 %, variant de 12,5 % à 52,1 %. Donc, il est clair que le modèle spatial proposé CAR-MFH est supérieur au modèle de Fay-Herriot. Nous avons également obtenu des résultats similaires pour les modèles MYC et CAR-MYC quand la variance d'échantillonnage est modélisée directement. La réduction moyenne du CV sous MYC est de 23,9 %, tandis qu'elle est de 29,0 % sous le modèle spatial proposé CAR-MYC. Des résultats détaillés, y compris les estimations ponctuelles et les CV correspondants, sont présentés dans un tableau à l'annexe C. Dans notre exemple, la taille de l'échantillon au niveau de la région sociosanitaire est relativement grande. Néanmoins, les estimations fondées sur un modèle révèlent une amélioration importante par rapport aux estimations directes fondées sur le plan de sondage. Nos résultats indiquent que les modèles pour petits domaines proposés peuvent être utilisés afin d'améliorer les estimations directes, même quand la taille d'échantillon est relativement grande. Il convient de souligner que les intervalles de crédibilité bayésiens pour les paramètres de petit domaine peuvent être construits facilement en utilisant la sortie MCMC de l'échantillonneur de Gibbs, si cela est nécessaire en pratique. Il s'agit de l'un des avantages de l'utilisation de l'inférence HB par la voie de l'échantillonnage MCMC. Cependant, ici, nous ne présentons que les estimations ponctuelles fondées sur un modèle et les CV correspondants, car notre objectif principal est de comparer les estimations fondées sur un modèle aux estimations directes et de montrer les gains d'efficacité des modèles. Le gain d'efficacité est clairement appréciable.

Afin d'étudier les effets de l'intégration de la structure spatiale dans le modèle, à la figure 4, nous présentons les CV des estimations directes et HB selon les régions sociosanitaires classées en fonction du nombre de régions voisines, en allant du plus petit (deux voisines) au plus grand

(sept voisines). La figure montre que les estimations HB d'après le modèle CAR-MFH proposé ont un CV plus petit que celles produites au moyen du modèle de Fay-Herriot. En outre, l'amélioration que représente le modèle CAR-MFH par rapport au modèle de Fay-Herriot est nettement plus marquée dans les régions ayant un grand nombre de voisines, alors que les deux modèles donnent des CV très proches dans les régions dont le nombre de régions adjacentes est plus petit. Nous obtenons des résultats très semblables pour le modèle CAR-MYC comparativement au modèle MYC. Le tableau 1 donne la réduction moyenne des CV pour les régions sociosanitaires ayant le même nombre de voisines. Les résultats du tableau 1 représentent la réduction du CV des modèles spatiaux proposés quand la variance d'échantillonnage est connue ou inconnue. Par exemple, pour σ_i^2 connue ($\tilde{\sigma}_i^2$ lissée), pour les régions ne comptant que deux voisines, la réduction moyenne du CV sous le modèle CAR-MFH par rapport au modèle de Fay-Herriot n'est que de 0,9 % environ, tandis que pour les régions comptant sept voisines, la réduction moyenne du CV sous le modèle CAR-MFH par rapport au modèle MFH peut aller jusqu'à environ 20 %. Pour le cas de σ_i^2 inconnue, nous obtenons des résultats similaires pour CAR-MYC par rapport MYC. Les résultats numériques du tableau 1 confirment la tendance nette à une plus grande réduction des CV sous le modèle spatial proposé que sous le modèle MFH ou MYC à mesure que le nombre de régions voisines augmente. Donc, un plus grand nombre de régions voisines peuvent fournir plus de renseignements sur la structure spatiale en vue d'améliorer la précision et la fiabilité des estimations HB.

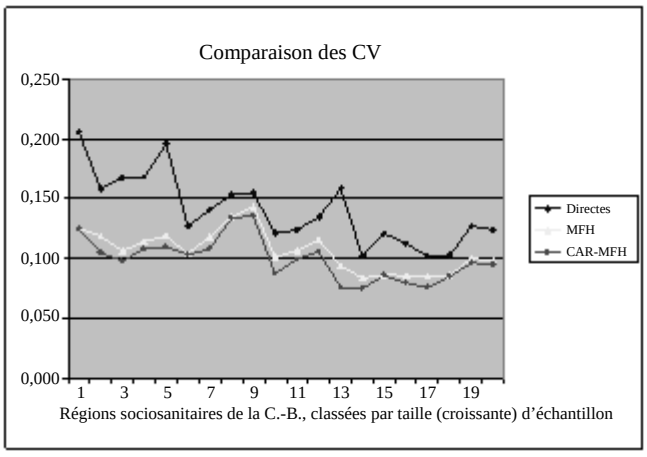

Figure 3 CV des estimations directes et HB sous les modèles MFH et CAR-MFH

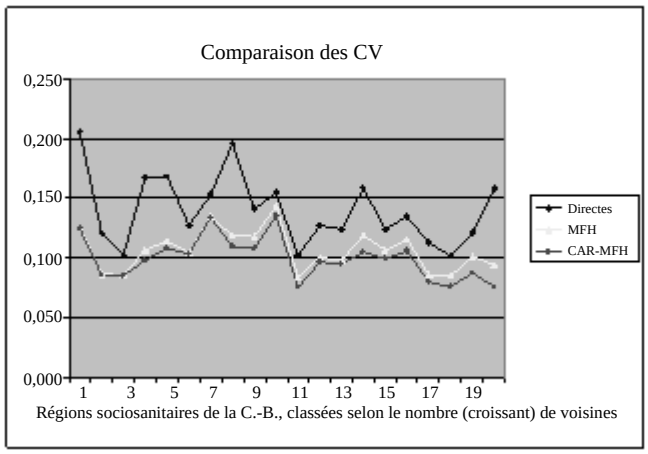

Figure 4 CV des estimations directes et HB sous les modèles MFH et CAR-MFH avec les régions sociosanitaires classées selon le nombre de régions voisines

Tableau 1
Comparaison de la réduction moyenne du CV

Nombre de régions voisines	Réduction moyenne du CV	
	CAR-MFH par rapport à MFH	CAR-MYC par rapport à MYC
2	0,9 %	1.8 %
3	3,7 %	3.5 %
4	6,3 %	6.0 %
5	8,9 %	8.7 %
6	13,7 %	11,0 %
7	19,2 %	20,7 %

3.3 Comparaison des modèles bayésiens

À la présente section, nous comparons les modèles proposés CAR-MFH et CAR-MYC aux modèles MFH et MYC, respectivement. Pour la comparaison de modèles hiérarchiques bayésiens, on peut se servir du critère d'information de déviance (DIC pour *Deviance Information Criterion*) proposé par Spiegelhalter, Best, Carlin et van der Linde (2002). Ce critère a été utilisé fréquemment ces dernières années pour comparer les modèles bayésiens à effets non emboîtés et mixtes. Le DIC est basé sur la déviance du modèle $D(\theta)$, qui est égale à moins deux fois la log-vraisemblance du modèle, et il est habituellement calculé sous la forme $\text{DIC} = D(\hat{\theta}) + 2p_D$, où $D(\hat{\theta})$ est la déviance du modèle, évaluée à la moyenne *a posteriori* des paramètres du modèle, qui résume la qualité de l'ajustement du modèle, et p_D est le nombre effectif de paramètres, qui traduit la complexité du modèle. p_D est défini comme étant $p_D = \bar{D}(\theta) - D(\hat{\theta})$, où $\bar{D}(\theta)$ est la moyenne *a posteriori* de la déviance du modèle. Donc, le DIC est défini comme la somme de l'adéquation du modèle et de la complexité du modèle. L'ajustement du modèle est d'autant meilleur que la valeur du DIC est faible. Le calcul du DIC est relativement

simple à condition qu'il existe une forme explicite pour la déviance, et p_D peut être calculé après l'exécution de l'échantillonnage de Gibbs en prenant la moyenne d'échantillon des valeurs simulées de $D(\theta)$ dont on soustrait l'estimation de la déviance $D(\hat{\theta})$ obtenue par insertion de valeurs (*plug-in*). Pour les quatre modèles présentés à la section 2, nous avons calculé les valeurs du DIC correspondantes, qui sont présentées au tableau 2. Il est clair que les modèles spatiaux proposés CAR-MFH et CAR-MYC ont tous deux un DIC plus petit que les modèles non spatiaux MFH et MYC, respectivement, ce qui signifie que les modèles spatiaux sont meilleurs que les modèles non spatiaux dans notre étude. Les deux modèles spatiaux CAR-MFH et CAR-MYC donnent de bons résultats dans cet exemple. Les résultats de la comparaison des modèles corroborent les résultats d'estimation présentés à la section 3.2.

Tableau 2
Comparaison des valeurs du DIC pour les quatre modèles hiérarchiques

Modèle	Valeur du DIC
MFH	27,1
CAR-MFH	24,6
MYC	26,8
CAR-MYC	24,5

3.4 Test d'adéquatoin du modèle

Afin de vérifier l'adéquation globale des modèles proposés CAR-MFH et CAR-MYC, nous avons utilisé la méthode de la loi prédictive *a posteriori*. Soit y_{rep} l'observation répétée sous le modèle. La loi prédictive *a posteriori* de y_{rep} sachant les données observée y_{obs} est définie comme étant $f(y_{\text{rep}}|y_{\text{obs}}) = \int f(y_{\text{rep}}|\theta)\, f(\theta|y_{\text{obs}})\, d\theta$. Dans cette approche, nous pouvons définir une statistique de test $T(y, \theta)$ qui dépend des données *y* et éventuellement du paramètre θ, et comparer la valeur observée $T(y_{\text{obs}}, \theta \mid y_{\text{obs}})$ à la loi prédictive *a posteriori* de $T(y_{\text{rep}}, \theta \mid y_{\text{obs}})$, tout écart significatif indiquant l'échec du modèle. Le manque d'ajustement aux données par rapport à la loi prédictive *a posteriori* peut être mesuré par la valeur *p* de la statistique de test (Meng 1994 ; Gelman, Meng et Stern 1996). La valeur *p* prédictive *a posteriori* est définie comme étant $p = P(T(y_{\text{rep}}, \theta) \geq T(y_{\text{obs}}, \theta) \mid y_{\text{obs}})$. Si le modèle donné est bien ajusté aux données observées, $T(y_{\text{obs}}, \theta \mid y_{\text{obs}})$ doit être proche de la partie centrale de l'histogramme des valeurs de $T(y_{\text{rep}}, \theta \mid y_{\text{obs}})$ si y_{rep} est générée de manière répétée à partir de la loi prédictive *a posteriori*. Par conséquent, la valeur *p* prédictive *a posteriori* devrait, en principe, être proche de 0,5 si le modèle est adéquatement ajusté aux données. Des valeurs *p* extrêmes (proche de 0 ou de 1) suggèrent un mauvais ajustement. La vérification des modèles au moyen de la valeur *p* prédictive *a posteriori* a été critiquée comme étant trop prudente, à cause de la double utilisation des données observées ; voir, par exemple, Bayarri et Berger (2000). Ces auteurs ont proposé des mesures de la valeur *p* de rechange pour la vérification des modèles, appelées valeur *p* prédictive *a posteriori* partielle et valeur *p* prédictive conditionnelle. Cependant, leurs méthodes sont plus difficiles à mettre en œuvre et à interpréter (Rao 2003 ; Sinharay et Stern 2003). Comme l'ont fait remarquer Sinharay et Stern (2003), la valeur *p* prédictive *a posteriori* est particulièrement utile si nous considérons le modèle courant comme un point final plausible auquel des modifications ne doivent être apportées que s'il s'avère que l'adéquation de l'ajustement laisse beaucoup à désirer.

Pour exécuter la vérification du modèle prédictif *a posteriori*, nous devons spécifier une statistique de test $T(y, \theta)$. You (2008b) a étudié par simulation plusieurs statistiques de test pour la vérification du modèle prédictif *a posteriori* dans le cas des modèles pour petits domaines et à proposé une statistique de test donnée par

$$T(y, \theta) = \left| \max(y_i) - \text{moyen.}(\theta_i) \right| - \left| \min(y_i) - \text{moyen.}(\theta_i) \right|.$$

You (2008b) montre que la statistique de test proposée $T(y, \theta)$ est sensible au choix de la distribution des effets aléatoires et de différentes fonctions de moyenne sous le modèle de Fay-Herriot. Une statistique de test similaire est également proposée dans Gelman et coll. (2004) pour la vérification du modèle prédictif *a posteriori*. Dans notre étude, sous le modèle proposé CAR-MFH, la valeur *p* estimée est de 0,472, et sous le modèle CAR-MYC, elle est de 0,453. Rien n'indique donc un manque d'ajustement du modèle et les deux modèles spatiaux proposés sont assez bien ajustés aux données.

Pour examiner l'ajustement du modèle au niveau d'observation individuel, nous avons également calculé les valeurs des probabilités prédictives individuelles p_i^* sous la forme $p_i^* = P(y_{i(\text{rep})} < y_{i(\text{obs})} \mid y_{\text{obs}})$; voir, par exemple, Gelfand (1996), ainsi que Daniels et Gatsonis (1999). Ces probabilités prédictives individuelles renseignent sur le degré de surestimation ou de sous-estimation systématique des données observées. Pour le modèle CAR-MFH, les p_i^* varient de 0,325 à 0,768 avec une moyenne de 0,517 et une médiane de 0,496 ; pour le modèle CAR-MYC, les p_i^* varient de 0,316 à 0,772 avec une moyenne de 0,511 et une médiane de 0,497. Les deux modèles donnent des résultats fort semblables, et les valeurs moyennes et médianes sont de l'ordre de 0,5. Il n'existe aucune indication d'une surestimation ou d'une sous-estimation systématique des modèles proposés. Les valeurs *p* globales et les probabilités prédictives individuelles montrent que les modèles spatiaux pour petits domaines proposés sont assez bien ajustés aux données.

3.5 Diagnostic du biais

Afin d'évaluer le biais éventuel des estimations fondées sur les modèles proposés par rapport aux estimations directes sous le plan de sondage, comme Brown, Chambers, Heady et Heasman (2001), nous appliquons une simple méthode d'analyse de régression aux estimations directes et aux estimations fondées sur un modèle HB. You (2008a) a également utilisé la méthode d'analyse de régression pour diagnostiquer le biais d'un modèle. Si les estimations fondées sur le modèle sont proches des valeurs réelles des taux de prévalence de la maladie dans les petits domaines, les estimations directes sous le plan de sondage, qui sont supposées fournir des estimations sans biais par rapport aux taux réel de prévalence de la maladie, devraient se comporter comme des variables aléatoires dont les valeurs prédites correspondent aux valeurs des estimations fondées sur un modèle. Autrement dit, les estimations fondées sur un modèle devraient être des prédicteurs sans biais des estimations directes. En ce qui concerne l'analyse de régression, nous ajustons essentiellement le modèle de régression $Y = \alpha + \beta X$ aux données et estimons les coefficients, et nous voyons dans quelle mesure la droite de régression s'approche de $Y = X$. Soit Y les estimations directes et X les estimations fondées sur le modèle. Sous le modèle proposé CAR-MFH, nous obtenons la droite de régression $Y = -0,0021(0,011) + 1,0365(0,1445)X$; sous le modèle proposé CAR-MYC, nous obtenons la droite de régression $Y = -0,0028(0,0108) + 1,0458(0,1427)X$. Donc, les deux droites de régression diffèrent fort peu de $Y = X$. Nous concluons par conséquent que les estimations fondées sur le modèle convergent vers les estimations directes sans biais supplémentaire éventuel induit par les modèles proposés. Les résultats donnent aussi une indication de l'absence de tout biais dû à une erreur éventuelle de spécification du modèle.

4. Conclusion

Dans le présent article, nous avons discuté de deux modèles au niveau du domaine, à savoir le modèle bien connu de Fay-Herriot dans lequel la variance d'échantillonnage est supposée être connue, et le modèle de You-Chapman dans lequel la variance d'échantillonnage est inconnue et modélisée séparément par son estimateur direct. Dans l'un et l'autre modèle, il est supposé que les effets aléatoires de domaine sont des variables aléatoires *iid* normales pour traduire les effets de l'hétérogénéité inexpliquée des domaines. Après avoir comparé diverses formes de modèles CAR gaussiens proposés dans la littérature (par exemple Best et coll. 2005) pour la cartographie des maladies en vue d'intégrer les effets spatialement corrélés, nous avons étendu le modèle avec effets de domaine indépendants à un modèle de corrélation spatiale, et l'avons combiné aux modèles pour petits domaines classiques. Les modèles à corrélation spatiale pour petits domaines CAR-MFH et CAR-MYC que nous proposons comprennent les modèles d'échantillonnage pour petits domaines et un modèle de lien à corrélation spatiale qui reflète l'hétérogénéité non structurée entre les domaines, ainsi que les effets de corrélation spatiale des domaines voisins. Le paramètre d'autocorrélation spatiale ne doit pas être spécifié dans le modèle et est estimé d'après les données.

Dans l'analyse des données, nous avons comparé les modèles spatiaux proposés aux modèles à effets non spatiaux pour estimer le taux de prévalence de l'asthme dans les 20 régions sociosanitaires de la Colombie-Britannique. Nous constatons que, comparativement aux estimations directes, les estimations fondées sur un modèle représentent une amélioration importante, qui se traduit par des estimations ponctuelles modérément lissées et des CV beaucoup plus petits. En particulier, les modèles proposés sont supérieurs au modèle de Fay-Herriot ou à celui de You-Chapman, que les variances d'échantillonnage soient supposées connues ou inconnues. En outre, la réduction des CV produits par les modèles spatiaux proposés comparativement au modèle de Fay-Herriot ou au modèle de You-Chapman est plus importante pour les domaines ayant un grand nombre de voisins. La comparaison des modèles bayésiens et l'analyse d'adéquation du modèle donnent aussi des résultats favorables aux modèles spatiaux pour petits domaines proposés.

Dans de futurs travaux, les modèles spatiaux pour petits domaines proposés pourront être étendus aux modèles d'échantillonnage et de lien non appariés (You et Rao 2002) sous variance d'échantillonnage connue ou inconnue. Nous prévoyons évaluer les effets de divers modèles spatiaux, ainsi que les effets des structures spatiales sur l'estimation. Pour l'analyse des données, nous produirons des estimations de l'état de santé fondées sur les modèles proposés pour les régions sociosanitaires des diverses régions du Canada et étudierons la possibilité d'étendre l'approche fondée sur un modèle à la production d'estimations à un niveau de détail plus fin, tel que les domaines âge-sexe à l'intérieur des régions sociosanitaires. Nous prévoyons également examiner la méthode de clonage des données (Lele, Dennis et Lutscher 2007 ; Lele, Nadeem et Schmuland 2010) pour les modèles spatiaux. Un avantage de cette méthode est que les résultats sont indépendants du choix des priors. Cependant, la demande de ressources informatiques pourrait être considérable.

Remerciements

Nous remercions le rédacteur associé et un examinateur de leurs commentaires et suggestions détaillés. Les travaux

de recherche de Yong You ont été financés par les ressources de financement global de la recherche de la Direction de la méthodologie de Statistique Canada. Les travaux de Qian M. Zhou ont été exécutés sous la supervision de Yong You durant un stage de recherche à Statistique Canada financé par le MITACS/PNSDC. Q.M. Zhou a présenté les modèles proposés et certains résultats décrits dans l'article à l'Assemblée annuelle de 2008 de la Société statistique du Canada (SSC) tenue à Ottawa et a été la lauréate 2008 du prix du meilleur article présenté par un étudiant décerné par la Section des méthodes d'enquête de la SSC.

Annexe A

Lois conditionnelles complètes

A.1. Lois conditionnelles complètes pour l'échantillonnage de Gibbs sous le modèle 1 : MFH.

- $[\theta_i \mid y_i, \beta, \sigma_v^2] \sim \mathrm{N}[\gamma_i y_i + (1-\gamma_i)\mathbf{x}_i'\beta, \tilde{\sigma}_i^2\gamma_i]$, où $\gamma_i = \sigma_v^2/(\sigma_v^2+\tilde{\sigma}_i^2)$, pour $i = 1, ..., m$;
- $[\beta|\boldsymbol{\theta}, \sigma_v^2] \sim \mathrm{N}\left[\left(\sum_{i=1}^m \mathbf{x}_i\mathbf{x}'\right)^{-1}\left(\sum_{i=1}^m \mathbf{x}_i\theta_i\right), \sigma_v^2\left(\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right]$;
- $[\sigma_v^2 \mid \boldsymbol{\theta}, \beta] \sim \mathrm{GI}\left[a_0 + \frac{1}{2}m, b_0 + \frac{1}{2}\sum_{i=1}^m (\theta_i - \mathbf{x}_i'\beta)^2\right]$.

A.2. Lois conditionnelles complètes pour l'échantillonnage de Gibbs sous le modèle 2 : CAR-MFH.

- $[\boldsymbol{\theta} \mid \mathbf{y}, \beta, \lambda, \sigma_v^2] \sim \mathrm{MVN}(\boldsymbol{\Lambda}\mathbf{y} + (\mathbf{I}-\boldsymbol{\Lambda})\mathbf{X}\beta, \boldsymbol{\Lambda}\mathbf{E})$, où $\boldsymbol{\Lambda} = (\mathbf{E}^{-1}+\mathbf{D}/\sigma_v^2)^{-1}\mathbf{E}^{-1}$ avec $\mathbf{E} = \mathrm{diag}\{\tilde{\sigma}_1^2, ..., \tilde{\sigma}_m^2\}$ et $\mathbf{D} = \lambda\mathbf{R} + (1-\lambda)\mathbf{I}$;
- $[\beta|\boldsymbol{\theta}, \lambda, \sigma_v^2] \sim \mathrm{MVN}[(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\boldsymbol{\theta}, \sigma_v^2(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}]$;
- $[\lambda|\boldsymbol{\theta}, \beta, \sigma_v^2] \propto \left|[\lambda\mathbf{R} + (1-\lambda)\mathbf{I}]^{-1}\right|^{-1/2} \times \exp\left\{-\frac{1}{2\sigma_v^2}(\boldsymbol{\theta}-X\beta)'[\lambda\mathbf{R}+(1-\lambda)\mathbf{I}](\boldsymbol{\theta}-X\beta)\right\}$;
- $[\sigma_v^2|\boldsymbol{\theta}, \beta, \lambda] \sim \mathrm{GI}\left[a_0 + \frac{m}{2}, b_0 + \frac{1}{2}(\boldsymbol{\theta}-X\beta)'\mathbf{D}(\boldsymbol{\theta}-X\beta)\right]$.

A.3. Lois conditionnelles complètes pour l'échantillonnage de Gibbs sous le modèle 3 : MYC.

- $[\theta_i \mid y_i, \beta, \sigma_i^2, \sigma_v^2] \sim \mathrm{N}[\gamma_i y_i + (1-\gamma_i)\mathbf{x}_i'\beta, \sigma_i^2\gamma_i]$, où $\gamma_i = \sigma_v^2/(\sigma_v^2+\sigma_i^2)$, pour $i = 1, ..., m$;
- $[\beta|\boldsymbol{\theta}, \sigma_v^2] \propto \mathrm{N}\left[\left(\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^m \mathbf{x}_i\theta_i\right), \sigma_v^2\left(\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right]$;
- $[\sigma_i^2 \mid y_i, \theta_i] \sim \mathrm{GI}\left(a_i + \frac{d_i+1}{2}, b_i + \frac{(y_i-\theta_i)^2 + d_i s_i^2}{2}\right)$, où $d_i = n_i - 1$, pour $i = 1, ..., m$;
- $[\sigma_v^2 \mid \boldsymbol{\theta}, \beta] \sim \mathrm{GI}\left[a_0 + \frac{1}{2}m, b_0 + \frac{1}{2}\sum_{i=1}^m (\theta_i - \mathbf{x}_i'\beta)^2\right]$.

A.4. Lois conditionnelles complètes pour l'échantillonnage de Gibbs sous le modèle 4 : CAR-MYC.

- $[\boldsymbol{\theta} \mid \mathbf{y}, \beta, \lambda, \sigma_v^2, \sigma_i^2] \sim \mathrm{MVN}(\boldsymbol{\Lambda}\mathbf{y} + (\mathbf{I}-\boldsymbol{\Lambda})\mathbf{X}\beta, \boldsymbol{\Lambda}\mathbf{E})$, où $\boldsymbol{\Lambda} = (\mathbf{E}^{-1}+\mathbf{D}/\sigma^2)^{-1}\mathbf{E}^{-1}$, et $\mathbf{E} = \mathrm{diag}\{\sigma_1^2, ..., \sigma_m^2\}$, $\mathbf{D} = \lambda\mathbf{R} + (1-\lambda)\mathbf{I}$;
- $[\beta \mid \boldsymbol{\theta}, \lambda, \sigma_v^2] \sim \mathrm{MVN}[(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\boldsymbol{\theta}, \sigma_v^2(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}]$;
- $[\lambda \mid \boldsymbol{\theta}, \beta, \sigma_v^2] \propto \left|[\lambda\mathbf{R}+(1-\lambda)\mathbf{I}]^{-1}\right|^{-1/2} \times \exp\left\{-\frac{1}{2\sigma_v^2}(\boldsymbol{\theta}-X\beta)'[\lambda\mathbf{R}+(1-\lambda)\mathbf{I}](\boldsymbol{\theta}-X\beta)\right\}$

  $[\sigma_i^2 \mid y_i, \theta_i] \sim \mathrm{GI}\left(a_i + \frac{d_i+1}{2}, b_i + \frac{(y_i-\theta_i)^2 + d_i s_i^2}{2}\right)$, où $d_i = n_i - 1$, pour $i = 1, ..., m$;
- $[\sigma_v^2 \mid \boldsymbol{\theta}, \beta, \lambda] \sim \mathrm{GI}\left[a_0 + \frac{m}{2}, b_0 + \frac{1}{2}(\boldsymbol{\theta}-X\beta)'\mathbf{D}(\boldsymbol{\theta}-X\beta)\right]$.

Annexe B

Liste des 20 régions sociosanitaires de la Colombie-Britannique avec les tailles d'échantillon et les structures spatiales correspondantes

Numéro d'ID	Nom de la région sociosanitaire	Taille de l'échantillon	Nombre de régions voisines	Régions voisines
1	East Kootenay	645	3	2, 3, 15
2	West Kootenay-Boundary	705	3	1, 3, 4
3	North Okanagan	890	5	1, 2, 4, 5, 15
4	South Okanagan Similameen	1 063	4	2, 3, 5, 6
5	Thompson	982	7	3, 4, 6, 9, 11, 12, 15
6	Fraser Valley	1 125	5	4, 5, 7, 8, 9
7	South Fraser Valley	1 437	4	6, 8, 17, 19
8	Simon Fraser	1,165	5	6, 7, 9, 17, 18
9	Coast Garibaldi	623	5	5, 6, 8, 11, 18
10	Central Vancouver Island	1 077	2	11, 20
11	Upper Island/Central Coast	746	4	5, 9, 10, 12
12	Cariboo	673	4	5, 11, 13, 15
13	North West	650	3	12, 14, 15
14	Peace Liard	611	2	13, 15
15	Northern Interior	859	6	1, 3, 5, 12, 13, 14
16	Vancouver	1 285	4	17, 18, 19, 20
17	Burnaby	871	5	7, 8, 16, 18, 19
18	North Shore	842	4	8, 9, 16, 17
19	Richmond	828	3	7, 16, 17
20	Capital	1 225	2	10, 16

Nota : Vancouver (n°16) et Capital (n°20) ne sont pas des régions adjacentes sur la carte, puisqu'elles sont séparées par l'océan. Cependant, étant donné le lien intensif et étroit entre ces deux régions, nous les définissons comme voisines dans notre étude à titre d'illustration seulement.

Annexe C

Estimations ponctuelles directes et fondées sur un modèle et CV

Comparaison des estimations ponctuelles					
ID du domaine	**Est. directe**	**MFH**	**CAR-MFH**	**MYC**	**CAR-MYC**
1	0,0765	0,0793	0,0812	0,0795	0,0812
2	0,0804	0,0795	0,0793	0,0797	0,0794
3	0,0745	0,0726	0,0731	0,0725	0,0729
4	0,0893	0,0868	0,0874	0,0867	0,0873
5	0,0782	0,0739	0,0736	0,0729	0,0731
6	0,0943	0,0914	0,0927	0,0918	0,0928
7	0,0702	0,0707	0,0712	0,0711	0,0717
8	0,0858	0,0845	0,0848	0,0844	0,0849
9	0,0877	0,0763	0,0745	0,0765	0,0747
10	0,0763	0,0805	0,0799	0,0805	0,0796
11	0,0661	0,0685	0,0678	0,0679	0,0676
12	0,0717	0,0681	0,0681	0,0678	0,0677
13	0,0631	0,0687	0,0692	0,0690	0,0693
14	0,0673	0,0685	0,0680	0,0685	0,0686
15	0,0793	0,0721	0,0707	0,0728	0,0713
16	0,0657	0,0696	0,0702	0,0697	0,0704
17	0,0859	0,0778	0,0759	0,0773	0,0759
18	0,0583	0,0626	0,0633	0,0618	0,0626
19	0,0619	0,0649	0,0647	0,0653	0,0647
20	0,0877	0,0923	0,0914	0,0917	0,0908
Comparaison des CV					
ID du domaine	**Est. directe**	**MFH**	**CAR-MFH**	**MYC**	**CAR-MYC**
1	0,168	0,107	0,099	0,107	0,100
2	0,127	0,105	0,104	0,097	0,093
3	0,135	0,116	0,106	0,110	0,097
4	0,102	0,084	0,076	0,079	0,072
5	0,158	0,094	0,076	0,105	0,083
6	0,113	0,086	0,080	0,086	0,081
7	0,124	0,099	0,096	0,106	0,101
8	0,102	0,085	0,076	0,081	0,073
9	0,158	0,119	0,105	0,117	0,105
10	0,121	0,087	0,086	0,086	0,084
11	0,141	0,118	0,108	0,109	0,105
12	0,196	0,119	0,109	0,130	0,116
13	0,168	0,115	0,108	0,111	0,108
14	0,206	0,126	0,125	0,136	0,133
15	0,121	0,101	0,087	0,094	0,083
16	0,127	0,101	0,097	0,103	0,097
17	0,124	0,107	0,100	0,105	0,096
18	0,155	0,143	0,136	0,134	0,130
19	0,154	0,135	0,134	0,128	0,128
20	0,103	0,086	0,085	0,083	0,082

Bibliographie

Arora, V., et Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

Bayarri, M.J., et Berger, J.O. (2000). P values for composite null models. *Journal of the American Statistical Association*, 95, 1127-1142.

Béland, Y. (2002). Enquête sur la santé dans les collectivités canadiennes – Aperçu de la méthodologie. Rapports sur la santé, Statistique Canada, n° 82-003 au catalogue, 13, 3, 9-15.

Besag, J., York, J. et Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (avec discussion). *Annals of the Institute of Statistical Mathematics*, 43, 1-59.

Best, N., Richardson, S. et Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14, 35-39.

Brown, G., Chambers, R., Heady, P. et Heasman, D. (2001). Évaluation des méthodes d'estimation régionale dans leur application aux estimations du chômage tirées de l'Enquête sur la population active au Royaume-Uni. Recueil : Symposium 2001, *La qualité des données d'un organisme statistique* : *une perspective méthodologique*, Statistique Canada, CD-ROM, 1-10.

Chip, S., et Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49, 327-335.

Cressie, N. (1990). Prédiction du sous-dénombrement pour les petites régions à l'aide du modèle linéaire général. Recueil : Symposium 1990, *Mesure et amélioration de la qualité des données*, Statistique Canada, 103-116.

Daniels, M.J., et Gatsonis, C. (1999). Hierarchical generalized linear models in the analysis of variations in health care utilization. *Journal of the American Statistical Association*, 94, 29-42.

Datta, G.S., Lahiri, P., Maiti, T. et Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.

Dick, P. (1995). Modélisation du sous-dénombrement net dans le recensement du Canada de 1991. *Techniques d'enquête*, 21, 51-61.

Fay, R.E., et Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 268-277.

Gelfand, A.E. (1996). Model determination using sampling-based methods. Dans *Markov Monte Carlo in Practice* (Éds., W.R. Gilks, S. Richardson et D.J. Spiegelhalter), Londres : Chapman & Hall, 145-161.

Gelman, A., Carlin, J.B., Stern, H.S. et Rubin, D.B. (2004). *Bayesian Data Analysis*, 2e Édition. Chapman & Hall/CRC.

Gelman, A., Meng, X.L. et Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (avec discussion). *Statistica Sinica*, 6, 733-807.

Gelman, A., et Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.

Ghosh, M., Natarajan, K., Stroud, T.W.F. et Carlin, B.P. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association*, 93, 273-282.

Ghosh, M., Natarajan, K., Walter, L.A. et Kim, D.H. (1999). Hierarchical Bayes GLMs for the analysis of spatial data: An application to disease mapping. *Journal of Statistical Planning and Inference*, 75, 305-318.

He, Z., et Sun, D. (2000). Hierarchical Bayes estimation of hunting success rates with spatial correlations. *Biometrics*, 56, 360-367.

Jiang, J., et Lahiri, P. (2006). Mixed model prediction and small area estimation (avec discussion). *Test*, 15, 1-96.

Lele, S.R., Dennis, B. et Lutscher, F. (2007). Data cloning: Easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters*, 10, 551-563.

Lele, S.R., Nadeem, K. et Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105, 1617-1625.

Leroux, B.G., Lei, X. et Breslow, N. (1999). Estimation of disease rates in small areas: A new mixed model for spatial dependence. Dans *Statistical Models in Epidemiology, the Environment and Clinical Trials*, (Éds., M.E. Halloran et D. Berry). New York : Springer Verlag, 135-178.

Liu, B., Lahiri, P. et Kalton, G. (2008). Hiearchical Bayes modeling of survey weighted small area proportions. Manuscript non-publié.

Maiti, T. (1998). Hierarchical Bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, 69, 339-348.

Mohadjer, L., Rao, J.N.K., Liu, B., Krenzke, T. et van de Kerckhove, W. (2007). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *Proceedings of the American Statistical Association, Section of Survey Method Research.*

Moura, F.A.S., et Migon, H.S. (2002). Bayesian spatial models for small area proportions. *Statistical Modelling*, 2, 3, 183-201.

MacNab, Y.C. (2003). Hierarchial Bayesian spatial modeling of small-area rates of non-rare disease. *Statistics in Medicine*, 22, 1761-1773.

Maiti, T. (1998). Hierarchical Bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, 69, 339-348.

Meng, X.L. (1994). Posterior predictive p value. *The Annals of Statistics*, 22, 1142-1160.

Mollié, A. (1996). Bayesian mapping of disease. Dans *Markov Chain Monte Carlo in Practice*. Londres : Chapman and Hall, 359-379.

Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.

Singh, A.C., Folsom, R.E., Jr. et Vaish, A.K. (2005). Small area modeling for survey data with smoothed error covariance structure via generalized design effects. Federal Committee on Statistical methods Conference proceedings, Washington, D.C., www.fcsm.gov.

Singh, B.B., Shukla, G.K. et Kundu, D. (2005). Modèles spatio-temporels pour l'estimation pour petits domaines. *Techniques d'enquête*, 31, 201-214.

Sinharay, S., et Stern, H.S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, 111, 209-221.

Spiegelhalter, D.J., Best, N., Carlin, B.P. and van de Linde, A. (2002). Bayesain measures of model complexity and fit. *Journal of Royal Statistical Society*, B, 64, 583-639.

Souza, D.F., Moura, F.A.S. et Migon, H.S. (2009). Prédiction de la population de petits domaines au moyen de modèles hiérarchiques. *Techniques d'enquête*, 35, 221-234.

Wang, J., et Fuller, W. A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y. (2008a). Une approche intégrée de modélisation de l'estimation du taux de chômage pour les régions infraprovinciales au Canada. *Techniques d'enquête*, 34, 21-31.

You, Y. (2008b). Small area estimation using area level models with model checking and applications. *Proceedings of Survey Methods Section, Statistical Society of Canada*.

You, Y., et Chapman, B. (2006). Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances d'échantillonnage. *Techniques d'enquête*, 32, 107-114.

You, Y., et Rao, J.N.K. (2000). Estimation bayésienne hiérarchique des moyennes pour petites régions à l'aide de modèles à plusieurs niveaux. *Techniques d'enquête*, 26, 197-206.

You, Y., et Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 20, 3-15.