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

Article

Mod®lisation hi®rarchique bay®sienne de proportions dans de petits domaines pond®r®es par les poids de sondage

par Benmei Liu, Partha Lahiri et Graham Kalton

Juin 2014

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

• Service de renseignements statistiques	1-800-263-1136
• Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
• Télécopieur	1-877-287-4369

Programme des services de dépôt

Service de renseignements	1-800-635-7943
Télécopieur	1-800-565-7757

Comment accéder à ce produit

Le produit n° 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2014

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada (http://www.statcan.gc.ca/reference/licence-fra.html).

This publication is also available in English.

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, ses entreprises, ses administrations et les autres établissements. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^{s}	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Techniques d'enquête, juin 2014
Vol. 40, N° 1, pp. 1-15
Statistique Canada, n° 12-001-X au catalogue

Modélisation hiérarchique bayésienne de proportions dans de petits domaines pondérées par les poids de sondage

Benmei Liu, Partha Lahiri et Graham Kalton [1]

Résumé

L'article décrit les résultats d'une étude par simulation Monte Carlo réalisée en vue de comparer l'efficacité de quatre modèles hiérarchiques bayésiens d'estimation sur petits domaines pour estimer des proportions au niveau de l'État au moyen de données provenant d'échantillons aléatoires simples stratifiés tirés d'une population finie fixe. Deux des modèles reposent sur les hypothèses fréquentes selon lesquelles, pour chaque petit domaine échantillonné, la proportion pondérée par les poids de sondage estimée suit une loi normale et sa variance d'échantillonnage est connue. L'un de ces modèles comprend un modèle de lien linéaire et l'autre, un modèle de lien logistique. Les deux autres modèles utilisent tous deux un modèle de lien logistique et reposent sur l'hypothèse que la variance d'échantillonnage est inconnue. L'un de ces deux modèles suppose que le modèle d'échantillonnage obéit à une loi normale et l'autre, qu'il obéit à une loi bêta. L'étude montre que, pour chacun des quatre modèles, la couverture sous le plan de sondage de l'intervalle de crédibilité des proportions au niveau de l'État en population finie s'écarte considérablement du niveau nominal de 95 % utilisé pour construire les intervalles.

Mots-clés : Proportions pondérées; modélisation hiérarchique bayésienne; loi bêta; intervalle de crédibilité.

1 Introduction

Les méthodes d'estimation sur petits domaines sont souvent utilisées pour estimer les proportions d'unités ayant une caractéristique donnée dans de petits domaines. Ainsi, ces méthodes sont appliquées aux données du programme des Small Area Income and Poverty Estimates (SAIPE) du Census Bureau pour estimer les taux de pauvreté au niveau de l'État, du comté et du district scolaire (Citro et Kalton, 2000; Maples et Bell 2005), aux données de la National Survey on Drug Use and Health (NSDUH) pour estimer les taux de consommation de drogues au niveau de l'État (Wright, Sathe et Spagnola 2007), et aux données de la National Assessment of Adult Literacy (NAAL) pour estimer les proportions de personnes ayant le plus faible niveau d'alphabétisation au niveau de l'État et du comté (Mohadjer, Rao, Liu, Krenzke et Van De Kerckhove 2012). Dans chaque cas, les tailles d'échantillon dans les petits domaines sont trop faibles pour produire des estimations directes d'une précision suffisante. Une grande variété de méthodes ont été élaborées pour résoudre ces problèmes d'estimation sur petits domaines. Voir Rao (2003) et Jiang et Lahiri (2006a) pour une revue de ces méthodes, et Chattopadhyay, Lahiri, Larsen et Reimnitz (1999), Farrell, MacGibbon and Tomberlin (1997) ainsi que Malec, Sedransk, Moriarity et LeClere (1997) et Malec, Davis et Cao (1999) pour des méthodes conçues spécialement pour estimer les proportions dans les petits domaines. La gamme de méthodes englobe l'approche hiérarchique bayésienne (HB) et celle du meilleur prédicteur empirique (MPE), ainsi que les modèles élaborés au niveau du domaine et au niveau de l'unité. Nous nous concentrons ici sur les modèles HB au niveau du domaine.

Lorsqu'on fait appel à un modèle HB au niveau du domaine pour produire des estimations des proportions d'unités possédant une caractéristique donnée dans de petits domaines, on suppose

1. Benmei Liu, Division of Cancer Control and Population Sciences, National Cancer Institute, 9609 Medical Center Drive Room 4E524, Bethesda, Maryland 20892; Courriel : liub2@mail.nih.gov; Partha Lahiri, JPSM, University of Maryland, 1218 Lefrak Hall, College Park, Maryland 20742; Graham Kalton, Westar, 1600 Research Boulevard, Rockville, Maryland 20850. Une grande partie de ce travail de recherche a été effectué alors que la première auteure était étudiante au cycle supérieur du *Joint Program in Survey Methodology* de l'Université du Maryland.

habituellement que, pour chaque domaine échantillonné, la proportion pondérée par les poids de sondage suit une loi d'échantillonnage normale et que sa variance d'échantillonnage est connue. Cependant, ces hypothèses posent problème lorsque la taille de l'échantillon de petit domaine est petite ou que la proportion réelle est proche de 0 ou de 1. Afin de pouvoir appliquer le théorème central limite pour approximer la normalité de la distribution d'échantillonnage d'une proportion, les échantillons doivent être raisonnablement grands, particulièrement si la proportion dans la population est très petite ou très grande (p. ex., inférieure à 0,1 ou supérieure à 0,9). En outre, quand les proportions sont très petites ou très grandes, la variance d'échantillonnage d'une proportion dans l'échantillon est extrêmement sensible à la valeur réelle de la proportion, de sorte qu'il est difficile d'établir une valeur appropriée pour la variance d'échantillonnage. Pour tenter de contourner ces problèmes, nous proposons deux modèles de rechange pour estimer les proportions dans les petits domaines et nous les comparons à deux modèles utilisés fréquemment. Les modèles sont décrits à la section 3. Les quatre modèles sont comparés au moyen d'une étude par simulation Monte Carlo dans laquelle des échantillons aléatoires simples stratifiés sont générés à partir d'une population finie fixe. L'étude par simulation est décrite à la section 4 et les résultats sont présentés à la section 5. Certaines conclusions sont formulées à la section 6. Cependant, pour commencer, nous présentons la notation pour un plan de sondage aléatoire simple stratifié à la section 2.

2 Notation

Soit N_{ih} la taille de la population dans la strate h dans le domaine i d'une population finie ($i=1,...,m;\ h=1,...,H_i$). Soit y_{ihk} la réponse binaire pour la caractéristique d'intérêt pour l'unité k dans la strate h dans le domaine i $(k=1,...,N_{ih})$. Les paramètres à estimer sont les proportions dans les petits domaines $P_i=\sum_h\sum_k y_{ihk}/N_{ih}$.

Sous le plan de sondage aléatoire simple stratifié choisi pour l'étude, n_{ih} unités sont sélectionnées parmi les N_{ih} unités dans la strate (ih). L'estimateur direct classique pour P_i est :

$$p_{iw}=\frac{\sum_h^{H_i}\sum_k^{n_{ih}} w_{ih}y_{ihk}}{\sum_h^{H_i}\sum_k^{n_{ih}} w_{ih}},\ i=1,...,m, \tag{2.1}$$

où w_{ih} désigne le poids de sondage donné par $w_{ih}=N_{ih}/n_{ih}$.

La variance de p_{iw} peut s'exprimer sous la forme

$$VAR_{st}(p_{iw})=\frac{P_i(1-P_i)}{n_i}DEFF_i, \tag{2.2}$$

où $DEFF_i$ est l'effet de plan qui traduit l'effet du plan de sondage complexe (Kish 1965). Pour un échantillon aléatoire simple stratifié avec fractions d'échantillonnage négligeables dans toutes les strates, l'effet de plan est donné approximativement par :

$$DEFF_i=\frac{\sum_h W_{ih}^2P_{ih}(1-P_{ih})/n_{ih}}{P_i(1-P_i)/n_i}, \tag{2.3}$$

où $W_{ih}=N_{ih}/N_i$, $N_i=\sum_h N_{ih}$, $n_i=\sum_h n_{ih}$, et P_{ih} est la proportion dans la population de la strate h dans le domaine i.

L'effet de plan $DEFF_i$ est une fonction des P_{ih}, qui sont inconnus. Si $P_{ih}(1-P_{ih}) \approx P_i(1-P_i)$, une approximation de $DEFF_i$ peut être donnée par $deff_{iw} = n_i \sum_h W_{ih}^2 / n_{ih}$. La valeur de $deff_{iw}$ se calcule facilement, puisqu'elle ne dépend d'aucun des paramètres inconnus.

Les méthodes d'estimation sur petits domaines peuvent être utilisées pour résoudre le problème dû au fait que p_{iw} est très imprécis quand la taille de l'échantillon n_i est petite. La section 3 donne la description des modèles HB au niveau du domaine sur lesquels porte la présente étude.

3 Modèles étudiés

Un modèle général d'estimation sur petits domaines au niveau du domaine comprend deux composantes. L'une — le modèle d'échantillonnage — est un modèle pour l'erreur d'échantillonnage des estimations directes d'après l'enquête. L'autre — le modèle de lien — relie la valeur de population pour un domaine aux variables auxiliaires propres au domaine $x_i = (x_{i1},...,x_{ip})'$.

La section 3.1 décrit deux modèles au niveau du domaine souvent utilisés pour estimer les proportions dans les petits domaines et la section 3.2 décrit certains problèmes associés à ces modèles. La section 3.3 décrit deux autres modèles dont on peut se servir pour résoudre ces problèmes.

3.1 Deux modèles souvent utilisés

Nous étudions deux modèles souvent utilisés en vue de les comparer aux deux nouveaux modèles décrits à la section 3.4. Le premier est le modèle de Fay-Herriot (Fay et Herriot 1979), qui suppose que les variances d'échantillonnage sont connues et que les variables suivent des lois normales dans le modèle d'échantillonnage ainsi que dans le modèle de lien. Le second est le modèle normal logistique, qui ne diffère du modèle de Fay-Herriot que par le remplacement de la loi normale par une loi logit-normale dans le modèle de lien.

Modèle 1 : (modèle normal-normal de Fay-Herriot)

Modèle d'échantillonnage :

$$p_{iw} \mid P_i \overset{ind}{\sim} N(P_i,\ \psi_i) \tag{3.1}$$

Modèle de lien :

$$P_i \mid \beta, \sigma_v^2 \overset{ind}{\sim} N(x_i^{'}\beta, \sigma_v^2) \tag{3.2}$$

Modèle 2 : (modèle normal-logistique)

Modèle d'échantillonnage :

$$p_{iw} \mid P_i \overset{ind}{\sim} N(P_i,\ \psi_i) \tag{3.3}$$

Modèle de lien :

$$g(P_i) \mid \beta, \sigma_v^2 \overset{ind}{\sim} N(x_i'\beta, \sigma_v^2) \tag{3.4}$$

Les deux modèles reposent sur l'hypothèse que la variance d'échantillonnage ψ_i est connue. Le modèle 1 est considéré comme un modèle apparié parce que les modèles d'échantillonnage et de lien peuvent être combinés pour produire un modèle mixte linéaire relativement simple. Cependant, on préfère souvent utiliser un modèle de lien non linéaire pour modéliser les proportions, ce qui donne des modèles d'échantillonnage et de lien non appariés, comme dans le modèle 2 (voir, par exemple, You et Rao 2002). La fonction lien $g(\cdot)$ peut être déterminée empiriquement en vérifiant l'adéquation du modèle. Les fonctions lien *log* et *logit* ont été utilisées. Le modèle de lien $\text{logit}(P_i)$ est choisi ici afin d'être certain que l'estimation de P_i soit toujours comprise dans l'intervalle de valeurs permises (0, 1).

3.2 Problèmes posés par les modèles 1 et 2

Deux problèmes importants sont associés aux modèles 1 et 2. Le premier tient au fait que ces modèles reposent tous deux sur l'hypothèse que les variances d'échantillonnage ψ_i sont connues, alors qu'en pratique elles doivent être estimées. Une approche simple consiste à utiliser l'estimation directe de la variance, mais celle-ci est très imprécise quand la valeur de P_i est très petite ou très grande et que la taille de l'échantillon n_i est petite. Une autre approche, plus complexe, consiste à élaborer une estimation approximative de P_i, disons p_{isyn}, en partant d'un modèle simple, tel qu'un modèle logistique pour p_{iw} en fonction des variables auxiliaires, puis à utiliser cette estimation dans l'estimateur de variance synthétique suivant :

$$\text{var}_{stsyn} = \frac{p_{isyn}(1 - p_{isyn})}{n_i} deff_{iw}. \tag{3.5}$$

Lorsqu'on ne dispose pas de variables auxiliaires, la proportion dans l'échantillon global peut être utilisée pour p_{isyn} dans le calcul de l'estimateur de variance synthétique.

Le deuxième problème a trait à l'hypothèse de normalité dans le modèle d'échantillonnage, qui est fondée sur une approximation en grand échantillon. Comme il est mentionné à la section 1, lorsque la taille de l'échantillon n_i est petite et que P_i est proche de 0 ou de 1, ce qui est fréquent dans le cas de l'estimation sur petits domaines, cette hypothèse pose problème.

3.3 Deux modèles de rechange

Sous les modèles 1 et 2, les variances d'échantillonnage inconnues ψ_i sont estimées, puis les estimations résultantes sont traitées comme s'il s'agissait des valeurs vraies connues. Une autre approche possible consiste à considérer que les ψ_i sont des paramètres inconnus dans le modèle HB, comme cela a été fait dans un certain nombre d'études. Par exemple, Arora et Lahiri (1997) ont appliqué un modèle HB pour modéliser les variances sous le plan de sondage pour les estimations sur échantillon. Singh, Folsom et Vaish (2005) ont proposé d'utiliser un modèle à effet de plan généralisé pour lisser la matrice de

covariance d'échantillonnage dans la modélisation pour petits domaines au moyen de données d'enquête. Récemment, You (2008) a proposé d'utiliser des effets de plan égaux au cours du temps pour modéliser les variances d'échantillonnage dans l'estimation des taux de chômage dans les petits domaines en utilisant un modèle log-linéaire transversal et chronologique. Dans le modèle 3, en tant que variante du modèle 2, les variances d'échantillonnage ψ_i sont traitées comme étant inconnues. Afin de tenir compte de la non-normalité des distributions d'échantillonnage des proportions de petit domaine pondérées par les poids de sondage, l'hypothèse d'une loi normale peut-être remplacée par celle d'une autre loi. Cette approche est appliquée dans le modèle 4 en faisant l'hypothèse d'une distribution d'échantillonnage bêta, c'est-à-dire une distribution qui possède la propriété désirable d'avoir l'intervalle de valeurs (0,1). Pour le reste, le modèle 4 est semblable au modèle 3, y compris le traitement des ψ_i, $i = 1,...,m$ comme des paramètres inconnus. Le modèle 4 a été pris en considération antérieurement par Jiang et Lahiri (2006b) à titre d'exemple de l'estimation des moyennes de domaines de population finie en utilisant l'approche du meilleur prédicteur empirique.

Modèle 3 (modèle normal-logistique avec variance d'échantillonnage inconnue) :

Modèle d'échantillonnage :

$$p_{iw} \mid P_i \overset{ind}{\sim} N(P_i,\ \psi_i) \tag{3.6}$$

Modèle de lien :

$$logit(P_i) \mid \beta, \sigma_v^2 \overset{ind}{\sim} N(x'\beta, \sigma_v^2) \tag{3.7}$$

Modèle 4 : (modèle bêta-logistique avec variance d'échantillonnage inconnue)

Modèle d'échantillonnage :

$$p_{iw} \mid P_i \overset{ind}{\sim} beta(a_i, b_i) \tag{3.8}$$

Modèle de lien :

$$logit(P_i) \mid \beta, \sigma_v^2 \overset{ind}{\sim} N(x_i'\beta, \sigma_v^2) \tag{3.9}$$

La fonction de variance approximative utilisée dans les modèles 3 et 4 est $\psi_i = [P_i(1 - P_i)/n_i]deff_{iw}$. Dans le modèle 4, les paramètres a_i et b_i sont donnés par :

$$a_i = P_i\left(\frac{n_i}{deff_{iw}} - 1\right), \text{ et } b_i = (1 - P_i)\left(\frac{n_i}{deff_{iw}} - 1\right).$$

Les estimations HB sur petits domaines peuvent être calculées au moyen des quatre modèles en utilisant l'algorithme de Metropolis-Hastings dans l'échantillonneur de Gibbs. Une description détaillée de l'algorithme, qui tire des échantillons aléatoires en se basant sur les lois conditionnelles complètes des paramètres inconnus à partir d'un ou de plusieurs jeux de valeurs initiales, figure dans Robert et Casella

(1999) et dans Chen, Shao et Ibraham (2000). You et Rao (2002) décrivent également en détail comment l'algorithme de Metropolis-Hastings fonctionne dans l'échantillonneur de Gibbs pour des modèles similaires aux modèles 1 et 2. Pour les modèles 3 et 4, l'algorithme fonctionne de la même façon que pour le modèle 2. Les distributions conditionnelles complètes sous chaque modèle sont données à l'annexe A.

4 Étude par simulation

4.1 Population étudiée et plan de sondage

La présente section décrit l'étude par simulation réalisée pour comparer l'efficacité des estimations sur petits domaines fournies par les quatre modèles HB. L'étude avait pour base le fichier de données à grande diffusion sur la natalité de 2002 qui englobait toutes les naissances survenues aux États-Unis cette année-là. Le fichier contenait des données extraites des certificats produits pour les naissances enregistrées dans chaque État et territoire (pour des renseignements détaillés, voir U.S. National Center for Health Statistics, 2009).

La population finie étudiée a été limitée aux 4 024 378 enregistrements de naissances vivantes ayant eu lieu en 2002 dans les 50 États des États-Unis et dans le district de Columbia (DC) pour lesquelles le poids à la naissance était consigné. Le paramètre d'intérêt était le taux de faible poids à la naissance au niveau de l'État P_i, $i = 1,...,51$, où le faible poids de naissance est défini comme étant moins de 2 500 grammes. La valeur de P_i variait de 5 % à 11 % selon l'État.

Dans chaque État, un plan EAS stratifié a été utilisé pour tirer des échantillons d'enregistrements de naissance. La race de la mère (blanche, noire et autre) a servi de variable de stratification. La taille de l'échantillon national a été fixée à environ 1 500 enregistrements de naissance pour chaque groupe racial. Une fraction d'échantillonnage uniforme a été appliquée à tous les États pour chaque groupe racial, sous la contrainte que deux enregistrements de naissance soient échantillonnés pour chaque groupe racial dans chaque État. La taille de l'échantillon national résultant était de $n = 4\,526$ enregistrements de naissance. Les tailles d'échantillon au niveau de l'État n_i variaient de 7 (pour les petits États tel que le Vermont) à 690 (pour la Californie), avec une taille d'échantillon médiane de 61. Cette procédure d'échantillonnage a été répétée $R = 1\,000$ fois, pour créer 1 000 jeux de données d'échantillon indépendants. Les poids d'échantillonnage sont restés les mêmes pour les diverses exécutions de la simulation.

4.2 Calcul des estimations HB

Pour simplifier, les hypothèses qui suivent ont été faites pour les modèles HB :

1. Aucune variable auxiliaire n'a été utilisée, de sorte que $x_i'\beta = \mu$.

2. Pour les modèles 1 et 2, on a posé que les variances d'échantillonnage étaient données par $\psi_i = [p_w(1 - p_w)/n_i]deff_{iw}$, où $p_w = \sum\sum\sum w_{ih}y_{ihk} / \sum\sum n_i w_{ih}$ est l'estimation nationale de la proportion de naissances vivantes avec faible poids de naissance. (Une vérification de l'utilisation de $deff_{iw}$ comme approximation de $DEFF_i$ a montré que l'approximation était raisonnable : les

deux quantités étaient proches, avec un coefficient de corrélation de Pearson de 0,96 et un ratio moyen de 1,08 entre $deff_{iw}$ et $DEFF_i$.)

3. Loi a priori uniforme pour μ, c.-à-d. $f(\mu) \propto 1$, et gamma inverse pour σ_v^2, c.-à-d. $\sigma_v^2 \sim IG(0{,}001;\ 0{,}001)$.

Pour chaque jeu de données d'échantillon, la première étape consistait à calculer les estimations directes sur échantillon au niveau de l'État. Les estimations pour chaque jeu de données d'échantillon ont ensuite été utilisées tour à tour comme entrée dans le logiciel WinBUGS (Lunn, Thomas, Best et Spiegelhalter 2000) employé pour produire les estimations HB pour les quatre modèles.

Dans un nombre important d'États où n_i était petit, les estimations directes étaient nulles dans certains jeux de données d'échantillon. Puisque WinBUGS ne peut traiter les estimations directes nulles que pour le modèle 1, pour les autres modèles, les estimations directes nulles ont été perturbées afin de les transformer en nombres positifs très petits.

Pour chaque exécution de WinBUGS, on s'est servi de trois chaînes indépendantes. Pour chaque chaîne, on a effectué un rodage de 10 000 échantillons, qui a été suivi de 10 000 autres échantillons. Le nombre d'échantillons après rodage a été réduit d'un facteur deux afin de diminuer l'autocorrélation des échantillons MCMC. Les 15 000 échantillons MCMC résultant des trois chaînes après rodage ont ensuite servi à calculer la moyenne et les centiles a posteriori pour chaque modèle HB à partir de chaque jeu de données d'échantillon. Le facteur de réduction d'échelle possible (*potential scale reduction factor*) $\hat{R}$ a été utilisé comme mesure principale de la convergence (voir Gelman et Rubin 1992). Le code WinBUGS est donné à l'annexe B.

5. Résultats des simulations

À la section 5.1, nous présentons les principaux résultats relatifs aux intervalles de crédibilité obtenus pour les proportions de naissances vivantes avec faible poids de naissance au niveau de l'État en appliquant chacun des quatre modèles. Ensuite, à la section 5.2, nous examinons les biais et les racines carrées des erreurs quadratiques moyennes de ces estimations.

5.1 Estimations des modèles et intervalles de crédibilité

Soit P_i^{HB} un estimateur HB de P_i, le pourcentage de naissances vivantes avec faible poids de naissance dans l'État i, et soit $P_{i,q}^{HB}$ le q^e centile de la distribution a posteriori de P_i. Fondé sur les résultats des 1 000 jeux de données de simulation, le tableau 5.1 donne les résultats qui suivent pour chaque modèle : la probabilité de non-couverture des intervalles de crédibilité à 95 % de P_i, c.-à-d. la probabilité que l'intervalle allant de $P_{i;0,025}^{HB}$ à $P_{i;0,975}^{HB}$ ne contienne pas P_i, et la largeur moyenne des intervalles de crédibilité $P_{i;0,975}^{HB} - P_{i;0,025}^{HB}$. Les erreurs-types de simulation Monte Carlo correspondantes sont également présentées entre parenthèses dans le tableau.

Pour examiner l'effet de la taille de l'échantillon de l'État sur les résultats des simulations, les 50 États et le district de Columbia sont répartis en trois groupes en fonction de la taille de leur échantillon, à savoir

les 15 États dont l'échantillon est de petite taille $(n_i \leq 30)$, les 24 États dont l'échantillon est de taille moyenne $(30 < n_i \leq 100)$, et les 12 États dont l'échantillon est de grande taille $(n_i > 100)$. Les résultats présentés au tableau 5.1 sont les moyennes globales sur l'ensemble des États et les moyennes pour les trois groupes distincts.

L'examen de la moitié supérieure du tableau 5.1 montre que les intervalles de crédibilité pour le modèle de Fay-Herriot (M1) sont très prudents, ne donnant quasiment aucune non-couverture. La moitié inférieure du tableau révèle que ce résultat est obtenu au prix de la plus grande largeur moyenne de l'intervalle de crédibilité parmi les quatre modèles. La largeur des intervalles de crédibilité de M1 est très stable. Une faible proportion de ces intervalles possède une borne inférieure négative.

Le faible niveau de non-couverture observé pour M1 pourrait tenir au fait que les variances d'échantillonnage ont été surestimées, peut-être parce que $deff_{iw}$ a été utilisé au lieu de $DEFF_i$. Pour examiner cette possibilité, nous avons utilisé $DEFF_i$ pour calculer la variance d'échantillonnage et nous n'avons constaté presque aucune différence de taux de non-couverture. Nous avons également exécuté le modèle en nous servant de la vraie variance définie dans (2.2) et n'avons de nouveau observé aucune différence appréciable entre les taux de non-couverture. La non-normalité de la distribution d'échantillonnage de p_{iw} pourrait également être à l'origine de ce problème.

Tableau 5.1
Pourcentage de fois que les intervalles de crédibilité à 95 % ne contiennent pas P_i, largeur moyenne des intervalles de crédibilité à 95 %, avec les erreurs-types de simulation Monte Carlo fondées sur 1 000 simulations (en pourcentage)

Taille de l'échantillon de l'État n_i	M1*	M2	M3	M4
	Pourcentage de non-couverture (erreur-type de simulation Monte Carlo)			
Échantillon global	0,40 (0,028)	8,24 (0,109)	6,52 (0,101)	4,36 (0,088)
$n_i \leq 30$ (15 États)	0,05 (0,019)	11,39 (0,239)	8,45 (0,216)	6,21 (0,190)
$30 < n_i \leq 100$ (24 États)	0,46 (0,043)	9,44 (0,167)	7,61 (0,156)	4,52 (0,132)
$n_i > 100$ (12 États)	0,70 (0,076)	1,91 (0,122)	1,94 (0,124)	1,74 (0,119)
	Largeur moyenne de l'intervalle de crédibilité à 95 % (erreur-type de simulation Monte Carlo)			
Échantillon global	9,05 (0,004)	5,52 (0,009)	6,20 (0,009)	8,45 (0,014)
$n_i \leq 30$ (15 États)	10,27 (0,009)	5,94 (0,020)	6,78 (0,021)	9,30 (0,034)
$30 < n_i \leq 100$ (24 États)	9,16 (0,005)	5,60 (0,013)	6,28 (0,013)	8,71 (0,021)
$n_i > 100$ (12 États)	7,29 (0,004)	4,84 (0,012)	5,30 (0,013)	6,88 (0,017)

*Nota : Pour le modèle 1, une faible proportion d'intervalles de crédibilité possède une borne inférieure négative.

À 8,2 %, le taux global de non-couverture des intervalles de crédibilité pour le modèle normal-logistique (M2) est sensiblement supérieur au taux nominal de 5 %. Ce modèle donne la plus petite largeur moyenne de l'intervalle. Le taux de non-couverture pour le modèle normal-logistique avec variance

inconnue (M3) s'approche davantage du taux nominal, avec une largeur globale de l'intervalle un peu plus grande que dans le cas de M2.

Le taux global de non-couverture de 4,4 % observé pour le modèle bêta-logistique (M4) est celui qui est le plus proche du taux de non-couverture nominal. Cependant, la largeur moyenne des intervalles de crédibilité est plus grande que celle obtenue pour M2 et M3, et l'erreur-type Monte Carlo de la largeur de l'intervalle est plus grande que celle observée pour les trois autres modèles. Cette instabilité pourrait découler de la complexité de la distribution conditionnelle complète pour le modèle bêta. La forte proportion des 1 000 estimations directes qui étaient nulles pour certains États dont la taille d'échantillon était petite pourrait aussi avoir causé d'importants problèmes d'ajustement de la loi bêta.

Comme prévu, pour les quatre modèles, la largeur moyenne des intervalles de crédibilité diminue lorsque la taille de l'échantillon de l'État augmente, et la variation de la largeur diminue également lorsque la taille de l'échantillon augmente. Toutefois, malgré ces diminutions, les taux de non-couverture diminuent aussi lorsque la taille de l'échantillon augmente pour les modèles 2, 3 et 4. En fait, les taux de non-couverture sont très petits pour les États dont la valeur de n_i est grande, ce qui fait penser que les intervalles de crédibilité ne reflètent pas adéquatement l'effet de la plus grande précision des estimations directes dans les États où la taille d'échantillon est grande.

5.2 Biais et REQM des estimations fondées sur les modèles

Afin d'étudier ces résultats plus en détail, nous avons examiné le biais et la racine carrée de l'erreur quadratique moyenne (REQM) des estimations P_i^{HB} pour chaque modèle. Les résultats sont présentés au tableau 5.2 dans le même format qu'au tableau 5.1. Les biais des estimations sous les modèles M1, M2 et M3 présentent une tendance similaire : les biais observés pour les petits États sont grands et positifs, et sont compensés dans une certaine mesure par les biais négatifs relativement faibles observés pour les États de moyenne et de grande taille. Dans le cas du modèle M4, les biais des estimations présentent une tendance très différente : ils sont presque nuls pour les petits États et leur valeur est grande et négative pour les États de moyenne et de grande taille. Cela indique que M4 donnerait de meilleurs résultats que les trois autres modèles en ce qui concerne le biais lorsque les échantillons des petits domaines sont de petite taille.

Tableau 5.2
Biais et racine carrée de l'erreur quadratique moyenne des estimations de P_i fondées sur les quatre modèles (en pourcentage)

Taille de l'échantillon de l'État n_i	M1		M2		M3		M4	
	Biais	REQM	Biais	REQM	Biais	REQM	Biais	REQM
Échantillon global	0,165	1,518	0,071	1,346	-0,009	1,411	-0,214	1,712
$n_i \leq 30$ (15 États)	0,621	1,651	0,572	1,630	0,466	1,652	0,009	1,922
$30 < n_i \leq 100$ (24 États)	-0,006	1,547	-0,123	1,386	-0,201	1,452	-0,319	1,775
$n_i > 100$ (12 États)	-0,063	1,294	-0,167	0,911	-0,219	1,026	-0,283	1,323

6. Discussion

Le présent article décrit les résultats d'une étude par simulation réalisée à partir d'une population finie réelle pour évaluer les intervalles de crédibilité produits par quatre modèles hiérarchiques, en se basant sur leur longueur et sur leurs propriétés de couverture sous le plan de sondage. Autant que nous sachions, ce genre d'évaluation des intervalles de crédibilité (ou de confiance) fondée sur le plan de sondage n'a encore jamais été effectuée dans le contexte de l'estimation sur petits domaines.

Dans l'étude par simulation, nous avons comparé les propriétés de couverture sous le plan de sondage des intervalles de crédibilité résultant de divers modèles hiérarchiques bayésiens élaborés pour estimer les proportions dans de petits domaines sous un plan de sondage aléatoire simple stratifié. Globalement, aucun modèle ne l'emporte clairement sur les autres, si bien que nous ne pouvons pas en recommander un plutôt que les autres.

La version hiérarchique bayésienne du modèle bien connu de Fay-Herriot semble produire des intervalles de crédibilité exagérément prudents. La non-normalité du modèle d'échantillonnage ainsi que du modèle de lien pourrait être la source de ce problème. Pour le modèle hiérarchique bêta-logistique, les intervalles de crédibilité donnent presque la couverture nominale pour les proportions de population finie et, des quatre modèles étudiés, ce modèle est celui qui possède les meilleures propriétés de biais quand la taille d'échantillon est petite. Cependant, comme l'une des distributions conditionnelles complètes pour le modèle bêta-logistique fait intervenir les proportions pondérées par les poids de sondage, la méthode MCMC pose problème chaque fois qu'une de ces proportions est nulle. Les intervalles de crédibilité pour ce modèle sont également plus larges que ceux observés pour les deux autres modèles comportant un modèle de lien logistique. La largeur de l'intervalle de crédibilité sous le modèle bêta-logistique pourrait peut-être être réduite si le modèle est modifié, par exemple en utilisant un modèle de mélange de lois à effet aléatoire en deux parties approprié qui permettrait d'éviter le problème des proportions pondérées par les poids de sondage nulles. Une étude plus approfondie à ce sujet est nécessaire. Il pourrait aussi être utile d'envisager d'autres modèles, peut-être un modèle probabiliste discret pour le niveau 1, en vue d'améliorer l'estimation des intervalles des petites proportions pour les petits domaines.

L'étude par simulation a montré que la couverture des intervalles de crédibilité bayésiens des proportions dans la population finie s'écartait considérablement du niveau nominal de 95 % pour les quatre modèles, et une constatation semblable a été faite pour la couverture sous le plan de sondage du modèle de Fay-Herriot dont l'usage est très répandu. Compte tenu de ces résultats, nous avons effectué un certain nombre d'analyses supplémentaires en vue de trouver une explication. Ces analyses comprenaient l'ajout de variables prédictives aux modèles, l'utilisation d'une loi a priori uniforme pour σ_v^2 (fondée sur des arguments formulés par Gelman 2006), l'utilisation de l'approche de la meilleure prédiction empirique pour le modèle M1, l'accroissement de la taille d'échantillon dans les États ne comptant que quelques naissances en fixant le nombre minimum à 50, et l'application de ces méthodes pour estimer la proportion de naissances dont le poids à la naissance était inférieur à la médiane nationale dans chaque État. Même si les propriétés de couverture des intervalles des proportions dans la population finie au niveau de l'État présentaient certaines différences, aucune de ces analyses n'a produit des taux de couverture proches des taux nominaux. Le seul cas où ces derniers coïncidaient avec les taux de couverture réels était celui d'un jeu de données simulé construit sous le modèle M1 pour les proportions au niveau de l'État de naissances pour lesquelles le poids de naissance était inférieur à la médiane nationale; les taux de couverture moyens

étaient de 5,1 % et de 5,2 % pour les approches du meilleur prédicteur empirique (MPE) et HB, respectivement.

L'étude par simulation a été limitée à un plan de sondage à un seul degré. En outre, pour simplifier, aucune variable auxiliaire n'a été incluse dans les modèles de lien dans les analyses principales, alors qu'en pratique, l'ajout de ce genre de variables est habituel et presque essentiel. D'autres études par simulation doivent être réalisées en vue d'examiner différents plans de sondage et différentes tailles d'échantillon, et d'intégrer certaines variables auxiliaires dans les modèles de lien. Nous espérons que notre étude encouragera d'autres chercheurs à exécuter des simulations fondées sur le plan de sondage similaires pour évaluer les méthodes d'estimation sur petits domaines. À la lumière de nos résultats limités, il convient de mettre en garde les utilisateurs d'estimations sur petits domaines quant à l'interprétation des intervalles de crédibilité des estimations.

Remerciements

Les auteurs tiennent à remercier les rédacteurs associés ainsi que deux réviseurs pour leurs suggestions constructives ayant permis d'améliorer significativement l'article original. Les travaux de recherche du deuxième auteur ont été financés par la *National Science Foundation* SES-085100.

Annexe A

A1. Distributions conditionnelles complètes pour les paramètres de chaque modèle

Soit $\vec{p} = (p_{1w},...,p_{mw})^t$ et $r_i = \dfrac{\psi_i}{\psi_i + \sigma_v^2}$.

Les distributions conditionnelles complètes pour le modèle de Fay-Herriot (M1) sont les suivantes :

i) $\theta_i \mid \mu, \sigma_v^2, \vec{p} \sim N((1-r_i)p_{iw} + r_i\mu, \quad \psi_i(1-r_i))$;

ii) $\mu \mid \theta_i, \sigma_v^2, p \sim N\left(\dfrac{1}{m}\sum_{i=1}^{m}\theta_i, \dfrac{\sigma_v^2}{m}\right)$;

iii) $\sigma_v^2 \mid \mu, \theta_i, \vec{p} \sim ING\left(a + \dfrac{1}{2}m, b + \dfrac{1}{2}\sum_{i=1}^{m}(\theta_i - \mu)^2\right)$.

Les distributions conditionnelles complètes pour le modèle normal-logistique (M2) sont les suivantes :

i) $\theta_i \mid \mu, \sigma_v^2, \vec{p} \propto \dfrac{1}{\theta_i(1-\theta_i)\sigma_v\sqrt{\psi_i}} \exp\left(-\dfrac{(p_{iw}-\theta_i)^2}{2\psi_i} - \dfrac{(\text{logit}(\theta_i)-\mu)^2}{2\sigma_v^2}\right)$;

ii) $\mu \mid \theta_i, \sigma_v^2, p \sim N\left(\dfrac{1}{m}\sum_{i=1}^{m}\text{logit}(\theta_i), \ \dfrac{\sigma_v^2}{m}\right)$;

iii) $\sigma_v^2 \mid \mu, \theta_i, \vec{p} \sim ING\left(a + \frac{1}{2}m, b + \frac{1}{2}\sum_{i=1}^{m}(\text{logit}(\theta_i) - \mu)^2 \right)$.

Les distributions conditionnelles complètes pour le modèle normal-logistique avec variance inconnue (M3) sont les mêmes que pour le modèle M2 excepté que ψ_i est remplacé par $\theta_i(1-\theta_i)deff_{iw} / n_i$ pour la distribution de θ_i sachant les autres paramètres.

Soit $\delta_{iw} = \frac{n_i}{deff_{iw}} - 1$. Les distributions conditionnelles complètes pour le modèle bêta-logistique (M4) sont les suivantes :

i) $\theta_i \mid \mu, \sigma_v^2, \vec{p} \propto \frac{1}{\theta_i(1-\theta_i)\sigma_v} \frac{p_{iw}^{\theta_i \delta_{iw}-1}(1-p_{iw})^{(1-\theta_i)\delta_{iw}-1}}{\Gamma(\theta_i \delta_{iw})\Gamma((1-\theta_i)\delta_{iw})} \exp\left(-\frac{(\text{logit}(\theta_i) - \mu)^2}{2\sigma_v^2} \right)$;

ii) $\mu \mid \theta_i, \sigma_v^2, p \sim N\left(\frac{1}{m}\sum_{i=1}^{m}\text{logit}(\theta_i),\ \frac{\sigma_v^2}{m} \right)$;

iii) $\sigma_v^2 \mid \mu, \theta_i, \vec{p} \sim ING\left(a + \frac{1}{2}m, b + \frac{1}{2}\sum_{i=1}^{m}(\text{logit}(\theta_i) - \mu)^2 \right)$.

Annexe B

Code WinBUGS pour le modèle 1 :

```
Model {
      for ( i in 1 :N)  {
         pobs[i] ~ dnorm(theta[i], D[i])
         D[i] <- 1/varhat[i]
         theta[i]<-u+v[i]
         v[i]~dnorm(0, tau)
                          }
        u~dflat()
       tau~dgamma(0.001, 0.001)
       sigma_v2<-1/tau
          }
```

Code WinBUGS pour le modèle 2 :

```
Model {
      for ( i in 1 :N)  {
         pobs[i] ~ dnorm(theta[i], D[i])
         D[i] <- 1/varhat[i]
         logit(theta[i])<-u+v[i]
         v[i]~dnorm(0, tau)
                          }
```

```
      u~dflat()
      tau~dgamma(0.001, 0.001)
      sigma_v2<-1/tau
       }
```

Code WinBUGS pour le modèle 3 :

```
Model {
      for ( i in 1 :N)  {
         pobs[i] ~ dnorm(theta[i], E[i])
         E[i] <- SAMPn[i]/(theta[i]*(1-theta[i])*DEFF_kish[i])
         logit(theta[i])<-u+v[i]
         v[i]~dnorm(0, tau)
         D[i]<-1/E[i]
                          }
      u~dflat()
      tau~dgamma(0.001, 0.001)
      sigma_v2<-1/tau
         }
```

Code WinBUGS pour le modèle 4 :

```
Model {
      for ( i in 1 :N) {
         pobs[i] ~ dbeta(a[i], b[i])
         a[i] <- theta[i]*(theta[i]*(1-theta[i])/D[i]-1)
         b[i] <- (1-theta[i])*(theta[i]*(1-theta[i])/D[i]-1)
         logit(theta[i])<-u+v[i]
         v[i]~dnorm(0, tau)
         D[i]<-theta[i]*(1-theta[i])*DEFF_kish[i]/SAMPn[i]
                          }
      u~dflat()
      tau~dgamma(0.001, 0.001)
      sigma_v2<-1/tau
         }
```

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.

Chattopadhyay, M., Lahiri, P., Larsen, M., et Reimnitz, J. (1999). Estimation composite de la prevalence des drogues pour des zones infraétats. *Techniques d'enquetes*, 25, 91-97.

Chen, M., Shao, Q., et Ibraham, J.G. (2000). *Monte Carlo Methods in Bayesian Computation.* New York: Springer-Verlag.

Citro, C., et Kalton, G. (Eds.) (2000). *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond*. Washington, DC: National Academy Press.

Farrell, P.J., MacGibbon, B., et Tomberlin, T.J. (1997). Empirical Bayes estimators of small area proportions in multistage designs. *Statistical Sinica*, 7, 1065-1083.

Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

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

Jiang, J., et Lahiri, P. (2006a). Mixed model prediction and small area estimation. *Test*, 15, 111-999.

Jiang, J., et Lahiri, P. (2006b). Estimation of finite population domain means: A model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.

Kish, L. (1965). *Survey sampling*. New York: John Wiley.

Lunn, D.J., Thomas, A., Best, N., et Spiegelhalter, D. (2000). WinBUGS – A Bayesian modeling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325-337.

Malec, D., Davis, W., et Cao, X. (1999). Small area estimates of overweight prevalence using sample selection adjustment. *Statistics in Medicine*, 18, 3189-3200.

Malec, D., Sedransk, J., Moriarity, C.L., et LeClere, F.B. (1997). Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association,* 92, 815-826.

Maples, J., et Bell, W.R. (2005). Evaluation of school district poverty estimates: Predictive models using IRS income tax data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1322-1329.

Mohadjer, L., Rao, J.N.K., Liu, B., Krenzke, T., et Van De Kerckhove, W. (2012). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *Journal of the Indian Society of Agricultural Statistics,* 66 (1), 55-63.

Rao, J.N.K. (2003). *Small area estimation*. New York: John Wiley and Sons.

Robert, C.P., et Casella, G. (1999). *Monte Carlo Statistical Methods*. New York: Springer-Verlag.

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 Statistical Working Paper* No. 39. (http://www.fcsm.gov/05papers/Singh_etal_IIIC.pdf)

U.S. National Center for Health Statistics (2009). *National Vital Statistics System*. *Birth Data*. (http://www.cdc.gov/nchs/births.htm)

Wright, D., Sathe, N., et Spagnola, K. (2007). *State Estimates of Substance Use from the 2004-2005 National Surveys on Drug Use and Health.* (DHHS Publication No. SMA 07-4235, NSDUH Series H-31). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

You, Y. (2008). Une approche intégrée de modélisation de l'estimation du taux de chômage pour les régions infraprovinciales au Canada. *Techniques d'enquêtes*, 34, 21-30.

You, Y., et Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *Canadian Journal of Statistics*, 30, 3-15.