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

Article

Estimation des intervalles de confiance des paramètres de petit domaine avec rétrécissement des moyennes et des variances

par Sarat C. Dass, Tapabrata Maiti, Hao Ren et Samiran Sinha

Décembre 2012

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 nº 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 ».

Ce produit est aussi disponible en version imprimée standard au prix de 30 $CAN l'exemplaire et de 58 $CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 $CAN	12 $CAN
Autres pays	10 $CAN	20 $CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par les moyens suivants :

- Téléphone (Canada et États-Unis) 1-800-267-6677
- Télécopieur (Canada et États-Unis) 1-877-287-4369
- Courriel infostats@statcan.gc.ca
- Poste Statistique Canada
Finances
Immeuble R.-H.-Coats, 6e étage
150, promenade Tunney's Pasture
Ottawa (Ontario) K1A 0T6
- En personne auprès des agents et librairies autorisés.

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

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

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
0s	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, décembre 2012
Vol. 38, N° 2, pp. 187-203
Statistique Canada, N° 12-001-X au catalogue

Estimation des intervalles de confiance des paramètres de petit domaine avec rétrécissement des moyennes et des variances

Sarat C. Dass, Tapabrata Maiti, Hao Ren et Samiran Sinha [1]

Résumé

Nous proposons une nouvelle approche d'estimation sur petits domaines fondée sur la modélisation conjointe des moyennes et des variances. Le modèle et la méthodologie que nous proposons améliorent non seulement les estimateurs sur petits domaines, mais donnent aussi des estimateurs « lissés » des vraies variances d'échantillonnage. Le maximum de vraisemblance des paramètres du modèle est estimé au moyen de l'algorithme EM en raison de la forme non classique de la fonction de vraisemblance. Les intervalles de confiance des paramètres de petit domaine sont obtenus en adoptant une approche de la théorie de la décision plus générale que l'approche classique de minimisation de la perte quadratique. Les propriétés numériques de la méthode proposée sont étudiées au moyen d'études par simulation et comparées à celles de méthodes concurrentes proposées dans la littérature. Une justification théorique des propriétés effectives des estimateurs et intervalles de confiance résultants est également présentée.

Mots clés : Algorithme EM ; Bayes empirique ; modèles hiérarchiques ; échantillonnage réjectif ; variance d'échantillonnage ; estimation sur petits domaines.

1. Introduction

L'estimation sur petits domaines et les techniques statistiques qui s'y rapportent sont des sujets qui ont fait l'objet d'une attention croissante ces dernières années. De nombreux organismes, tant publics que privés, cherchent à obtenir des estimations sur petits domaines fiables pour prendre des décisions stratégiques utiles. La surveillance de la situation socioéconomique et de l'état de santé de divers groupes définis selon l'âge, le sexe et la race pour lesquels s'observent des tendances distinctes sur de petites régions géographiques est un exemple d'application pratique des techniques d'estimation sur petits domaines.

Il est aujourd'hui généralement reconnu que les estimations directes à partir de données d'enquête calculées pour les petits domaines ne sont d'ordinaire pas fiables parce que leurs erreurs-types et coefficients de variation sont très souvent grands. Il devient donc nécessaire d'obtenir de meilleures estimations, d'une plus grande précision. Des approches fondées explicitement ou implicitement sur un modèle sont élaborées pour relier des petits domaines et obtenir une plus grande précision par « emprunt d'information » à des domaines similaires. Cette technique d'estimation est également appelée estimation par rétrécissement, ou estimation à rétrécisseur, puisque les estimations directes sont « rétrécies » afin qu'elles se rapprochent de la moyenne globale. Les estimations directes d'après les données d'enquête et les variances d'échantillon sont les principaux ingrédients qui entrent dans la création des modèles d'estimation sur petits domaines de niveau agrégé. La stratégie de modélisation repose habituellement sur l'hypothèse que les variances d'échantillonnage sont connues, tandis qu'un modèle de régression linéaire approprié est utilisé pour les moyennes. Pour des renseignements détaillés sur ces développements, le lecteur est invité à consulter Ghosh et Rao (1994), Pfeffermann (2002) et Rao (2003). Les modèles habituels au niveau du domaine suscitent deux critiques importantes. Premièrement, en pratique, les variances d'échantillonnage sont des quantités estimées qui sont donc sujettes à d'importantes erreurs. Il en est ainsi parce qu'elles sont souvent fondées sur des tailles d'échantillon équivalentes à celles qui servent au calcul des estimations directes. Deuxièmement, en raison de l'hypothèse que les variances d'échantillonnage sont connues et fixes formulée dans les modèles d'estimation sur petits domaines classiques, l'incertitude que comporte l'estimation de la variance n'est pas prise en compte dans la stratégie d'inférence globale.

Des tentatives en vue de modéliser uniquement les variances d'échantillonnage ont été faites antérieurement ; voir, par exemple, Maples, Bell et Huang (2009), Gershunskaya et Lahiri (2005), Huff, Eltinge et Gershunskaya (2002), Cho, Eltinge, Gershunskaya et Huff (2002), Valliant (1987), et Otto et Bell (1995). Dans leurs articles, Wang et Fuller (2003) et Rivest et Vandal (2003) ont étendu l'estimation de l'erreur quadratique moyenne (EQM) asymptotique des estimateurs sur petits domaines au cas où l'on estime les variances d'échantillonnage au lieu de s'appuyer sur l'hypothèse classique que les variances sont connues. En outre, You et Chapman (2006) ont considéré la modélisation des variances d'échantillonnage avec inférence en appliquant des techniques d'estimation entièrement bayésiennes.

De nombreux praticiens ont jugé nécessaire de modéliser la variance. Les progrès les plus récents dans ce domaine

1. Sarat C. Dass et Tapabrata Maiti, Department of Statistics & Probability, Michigan State University. Courriel : maiti@stt.msu.edu ; Hao Ren, CTB/McGraw-Hill, 20 Ryan Ranch Rd, Monterey, CA 93940 ; Samiran Sinha, Department of Statistics, Texas A & M University.

sont résumés élégamment dans un article publié en 2008 par William Bell, du *United States Census Bureau*. Ce dernier a examiné minutieusement les conséquences de ces problèmes dans le contexte de l'estimation de l'EQM des estimateurs sur petits domaines fondés sur un modèle. Il a également donné des preuves numériques de l'estimation de l'EQM pour le modèle de Fay-Herriot (donné dans l'équation 1) quand il est supposé que les variances d'échantillonnage sont connues. Les progrès exposés jusqu'à présent dans la littérature traitant des petits domaines peuvent être considérés « grosso modo » comme étant i) le lissage des estimations directes des variances des erreurs d'échantillonnage pour obtenir des estimations des variances plus stables dont le biais est faible et ii) la prise en compte (partielle) de l'incertitude dans les variances d'échantillonnage en étendant le modèle de Fay-Herriot.

Manifestement, l'effort en vue de bien tenir compte des variances d'échantillonnage dans la modélisation de la moyenne a été faible, voire nul, comparativement au nombre d'études consacrées à la modélisation et à l'inférence des moyennes. Le développement systématique du « rétrécissement » des moyennes ainsi que des variances fait défaut dans la littérature traitant de l'estimation sur petits domaines. Autrement dit, nous aimerions exploiter la technique de l'« emprunt d'information » à d'autres petits domaines en vue d'« améliorer » les estimations de la variance, tout comme nous le faisons pour « améliorer » les estimations des moyennes de petits domaines. Nous proposons un modèle hiérarchique utilisant à la fois les estimations directes d'après les données d'enquête et les estimations des variances d'échantillonnage pour inférer les paramètres du modèle qui déterminent le système stochastique. Notre objectif méthodologique est d'élaborer l'estimation « par rétrécissement » double pour les moyennes ainsi que les variances de petit domaine, en exploitant la structure de la modélisation conjointe moyenne-variance afin que les estimateurs finaux soient plus précis. Des preuves numériques montrent l'efficacité du rétrécissement double appliqué aux estimations sur petits domaines de la moyenne si l'on prend pour critère l'EQM.

Une autre contribution importante du présent article est l'obtention d'intervalles de confiance pour les moyennes de petits domaines. La littérature relative à l'estimation sur petits domaines traite avant tout des estimations ponctuelles et de leurs erreurs-types ; pourtant, il est bien connu que la pratique classique consistant à utiliser [estimation ponctuelle $\pm\, q \times$ erreur-type], où q est la valeur seuil Z (normale standard) ou t, ne produit pas des probabilités de couverture exactes des intervalles ; voir Hall et Maiti (2006) et Chatterjee, Lahiri et Li (2008) pour plus de précisions. Les travaux antérieurs, qui sont fondés sur la technique du bootstrap, sont d'un usage limité en raison de l'estimation répétée des paramètres du modèle. Nous produisons des intervalles de confiance pour les moyennes dans une perspective de théorie de la décision. La construction des intervalles de confiance est facile à mettre en œuvre en pratique.

La présentation de la suite de l'article est la suivante. Le modèle hiérarchique proposé pour les moyennes et les variances d'échantillonnage est élaboré à la section 2. L'estimation des paramètres du modèle au moyen de l'algorithme EM est exposée à la section 3. La justification théorique des intervalles de confiance proposés et leurs propriétés de couverture sont présentées à la section 4. Une étude par simulation et un exemple fondé sur des données réelles sont présentés aux sections 5 et 6, respectivement. Enfin, une discussion et certaines conclusions sont présentées à la section 7. Une autre formulation du modèle pour les petits domaines, ainsi que des détails mathématiques sont donnés en annexe.

2. Modèle proposé

Supposons que l'on examine n petits domaines. Pour le i^{e} petit domaine, soit de (X_i, S_i^2) la paire comprenant l'estimation directe et la variance d'échantillonnage, pour $i = 1, 2, ..., n$. Soit $\boldsymbol{Z}_i = (Z_{i1}, ..., Z_{ip})^T$ le vecteur de p covariables disponibles à l'étape de l'estimation pour le i^{e} petit domaine. Nous proposons le modèle hiérarchique suivant :

$$\left.\begin{aligned} X_i \mid \theta_i, \sigma_i^2 &\sim \text{Normale}(\theta_i, \sigma_i^2) \\ \theta_i &\sim \text{Normale}(\boldsymbol{Z}_i^T \boldsymbol{\beta}, \tau^2) \end{aligned}\right\} \qquad (1)$$

$$\left.\begin{aligned} \frac{(n_i - 1)S_i^2}{\sigma_i^2} \,\Big|\, \sigma_i^2 &\sim \chi^2_{n_i - 1} \\ \sigma_i^{-2} &\sim \text{Gamma}(a, b), \end{aligned}\right\} \qquad (2)$$

indépendamment pour $i = 1, 2, ..., n$. Dans l'élaboration du modèle, n_i est la taille d'un échantillon aléatoire simple (EAS) tiré du i^{e} domaine, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ est le vecteur de dimension $p \times 1$ des coefficients de régression, et $\boldsymbol{B} \equiv (a, b, \boldsymbol{\beta}, \tau^2)^T$ est la série complète de paramètres inconnus dans le modèle. En outre, $\text{Gamma}(a, b)$ est la densité de probabilité Gamma dont les paramètres de forme et d'échelle a et b, respectivement, sont positifs, définis comme étant $f(x) = \{b^a \Gamma(a)\}^{-1} e^{-x/b} x^{a-1}$ pour $x > 0$, et 0 autrement. Le terme σ_i^2 inconnu est la variance réelle de X_i et est habituellement estimé par la variance d'échantillon S_i^2. On suppose généralement que les S_i^2 suivent une loi du khi-carré possédant $(n_i - 1)$ degrés de liberté (en raison de la normalité et de l'EAS), mais nous notons que sous des plans de sondage complexes, le nombre de degrés de liberté

doit être déterminé prudemment (par exemple, Maples et coll. 2009). Surtout, le rôle de taille d'échantillon dans l'estimation par rétrécissement de σ_i^2 est le suivant : l'estimation de σ_i^{-2} se rapproche davantage de la moyenne globale (ab) pour de faibles valeurs de n_i que pour des valeurs élevées. Donc, pour les variances, les tailles d'échantillon jouent le même rôle que la précision dans l'estimation par rétrécissement des moyennes de petit domaine. Nous notons que You et Chapman (2006) ont également envisagé un deuxième niveau de modélisation de la variance d'échantillonnage. Cependant, les hyperparamètres reliés à la loi a priori de σ_i^2 ne sont pas dictés par les données mais plutôt choisis de façon telle que la loi a priori soit vague. Donc, leur modèle peut être considéré comme la version bayésienne de modèle examiné dans Rivest et Vandal (2003) et dans Wang et Fuller (2003). Le deuxième niveau de modélisation de σ_i^{-2} dans (2) peut être étendu encore davantage à $\sigma_i^{-2} \sim \text{Gamma}(b, \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta}_2)/b)$ de sorte que $E(\sigma_i^{-2}) = \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta}_2)$ pour un autre jeu de p coefficients de régression $\boldsymbol{\beta}_2$ afin d'inclure l'information sur les covariables dans la modélisation de la variance.

Bien que notre modèle soit motivé par Hwang, Qiu et Zhao (2009), nous tenons à mentionner que Hwang et coll. (2009) ont considéré les moyennes et variances par rétrécissement dans le contexte de données micro vectorielles où ils ont préconisé une solution importante consistant à insérer un estimateur à rétrécisseur de la variance dans l'estimateur de la moyenne. L'estimateur par rétrécissement de la variance dans Hwang et coll. (2009) est une fonction de S_i^2 seulement et non de X_i ainsi que S_i^2 ; voir les remarques 2 et 3 à la section 2. Donc, l'inférence de la moyenne ne tient pas compte de toute l'incertitude dans l'estimation de la variance. En outre, leur modèle ne contient aucune information sur les covariables. L'étude par simulation décrite plus loin indique que notre méthode d'estimation donne de meilleurs résultats que celle de Hwang et coll. (2009).

Dans la formulation du modèle susmentionné, l'inférence pour le paramètre θ_i représentant la moyenne de petit domaine peut être faite en se basant sur la distribution conditionnelle de θ_i sachant toutes les données $\{(X_i, S_i^2, \boldsymbol{Z}_i), i = 1, ..., n\}$. Sous notre modèle, la distribution conditionnelle de θ_i est une distribution non standard qui ne possède pas de forme analytique et requiert donc des méthodes numériques, telles que la méthode de Monte Carlo et l'algorithme EM, pour l'inférence. Des renseignements détaillés sont fournis à la section suivante.

3. Méthodologie d'inférence

3.1 Estimation des paramètres inconnus au moyen de l'algorithme EM

En pratique, $\boldsymbol{B} \equiv (a, b, \boldsymbol{\beta}, \tau^2)^T$ est inconnu et doit être estimé d'après les données $\{(X_i, S_i^2, \boldsymbol{Z}_i), i = 1, 2, ..., n\}$. Nous proposons d'estimer $\boldsymbol{B}$ par la méthode du maximum de vraisemblance marginale : estimer $\boldsymbol{B}$ par $\hat{\boldsymbol{B}}$ où $\hat{\boldsymbol{B}}$ maximise la vraisemblance marginale $L_M(\boldsymbol{B}) = \prod_{i=1}^n L_{M,i}(\boldsymbol{B})$, où

$$L_{M,i} \propto \frac{\Gamma(n_i/2+a)}{\tau\Gamma(a)b^a} \int \exp\left\{-\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}\right\} \psi_i^{-(n_i/2+a)} d\theta_i, \quad (3)$$

et

$$\psi_i \equiv \left\{0,5(X_i - \theta_i)^2 + 0,5(n_i - 1)S_i^2 + \frac{1}{b}\right\}. \quad (4)$$

La vraisemblance marginale L_M contient des intégrales qui ne peuvent pas être évaluées en forme analytique, de sorte que l'on doit recourir à des méthodes numériques pour sa maximisation. L'un de ces algorithmes est la procédure itérative EM (espérance-maximisation) qui est utilisée quand on a affaire à ce genre d'intégrales. L'algorithme EM comprend l'augmentation de la vraisemblance observée $L_M(\boldsymbol{B})$ présentant des données manquantes ; dans notre cas, les variables de l'intégration, θ_i, $i = 1, 2, ..., n$, constituent cette information manquante. Sachant $\boldsymbol{\theta} \equiv \{\theta_1, \theta_2, ..., \theta_n\}$, la log-vraisemblance (ℓ_c) sous *données complètes* peut s'écrire

$$\ell_c(\boldsymbol{B}, \boldsymbol{\theta}) = \sum_{i=1}^n \Bigg[\log\{\Gamma(n_i/2 + a)\} - \log\{\Gamma(a)\} - a\log(b) - 0,5\log(\tau^2) - \frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2} - (n_i/2 + a)\log(\psi_i) \Bigg],$$

où l'expression de ψ_i est donnée par l'équation (4). Partant d'une valeur initiale de $\boldsymbol{B}$, disons $\boldsymbol{B}^{(0)}$, l'algorithme EM exécute itérativement une maximisation par rapport à $\boldsymbol{B}$. À la t^{e} étape, la fonction d'objectif maximisée est

$$Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)}) = E(\ell_c(\boldsymbol{B}, \boldsymbol{\theta})) = \sum_{i=1}^n \Bigg[\log\{\Gamma(n_i/2 + a)\} - \log\{\Gamma(a)\} - a\log(b) - 0,5\log(\tau^2) - \frac{E(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2} - (n_i/2 + a)E\{\log(\psi_i)\} \Bigg].$$

Dans $Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})$, l'espérance est prise par rapport à la distribution conditionnelle de chaque θ_i sachant les données, $\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}^{(t-1)})$, ce qui est

$$\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}) \propto \exp\{-0,5(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2/\tau^2\} \psi_i^{-(n_i/2+a)}. \quad (5)$$

L'une des difficultés ici est que les espérances ne sont pas disponibles sous une forme analytique. Donc, nous

recourons à une méthode de Monte Carlo pour évaluer l'expression. Supposons que R échantillons iid de θ_i soient disponibles, disons $\theta_{i,1}, \theta_{i,2}, ..., \theta_{i,R}$. Alors, chaque expression de la forme $E\{h(\theta_i)\}$ peut être approximée par la moyenne Monte Carlo

$$E\{h(\theta_i)\} \approx \frac{1}{R}\sum_{r=1}^{R} h(\theta_{i,k}). \quad (6)$$

Cependant, le tirage de nombres aléatoires de la distribution conditionnelle $\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}^{(t-1)})$ n'est pas simple non plus, puisqu'il ne s'agit pas d'une densité standard. Les échantillons sont sélectionnés par la procédure d'acceptation-rejet (Robert et Casella 2004) : pour tirer un échantillon de la densité cible f, tirer un échantillon x de la loi instrumentale g, et l'accepter comme étant un échantillon tiré de f avec la probabilité $f(x)/\{M^* g(x)\}$, où $M^* = \sup_x \{f(x)/g(x)\}$. Un avantage de la méthode d'acceptation-rejet est que la densité cible f ne doit être connue que jusqu'à une constante de proportionnalité, ce qui est le cas pour $\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}^{(t-1)})$ dans (5) ; étant donné la forme non standard de la densité, la constante de normalisation ne peut pas être obtenue sous une forme analytique. Pour l'algorithme d'acceptation-rejet, nous avons utilisé la densité normale $g(\theta_i) \propto \exp\{-0{,}5(\theta_i - \boldsymbol{Z}_i^T \boldsymbol{\beta})^2/\tau^2\}$ comme loi instrumentale. La probabilité d'acceptation se calcule comme étant $[\{1/b + 0{,}5(n_i - 1)S_i^2\}/\{1/b + 0{,}5(n_i - 1)S_i^2 + 0{,}5(\theta_i - X_i)^2\}]^{n_i/2+a}$. Si l'on veut augmenter la probabilité d'acceptation, on peut choisir une meilleure loi instrumentale ou un algorithme différent (tel que les algorithmes d'échantillonnage réjectif adaptatif ou d'enveloppe d'acceptation-rejet), mais la loi instrumentale que nous avons choisie a donné des résultats satisfaisants dans les études que nous avons effectuées.

Le maximiseur de $Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})$ à la t^{e} étape peut être décrit explicitement. Les solutions pour $\boldsymbol{\beta}$ et τ^2 sont disponibles sous les formes analytiques suivantes

$$\boldsymbol{\beta}^{(t)} = \left(\sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i^T\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{Z}_i E(\theta_i)\right)$$

et

$$(\tau^2)^{(t)} = \frac{1}{n}\sum_{i=1}^{n} E(\theta_i - \boldsymbol{Z}_i^T \boldsymbol{\beta})^2,$$

respectivement. En outre, $a^{(t)}$ et $b^{(t)}$ s'obtiennent en résolvant $S_a = \partial Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})/\partial a = 0$ et $S_b = \partial Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})/\partial b = 0$ par la méthode de Newton-Raphson où

$$S_a = \sum_{i=1}^{n} \frac{\partial}{\partial a} \log\{\Gamma(n_i/2 + a)\} - n\left\{\frac{\partial}{\partial a} \log\{\Gamma(a)\}\right\} - n\log(b) - \sum_{i=1}^{n} E\{\log(\psi_i)\}$$

et

$$S_b = -\frac{na}{b} + \sum_{i=1}^{n} \frac{(n_i/2 + a)}{b^2} E(\psi_i^{-1}).$$

Nous posons que $\boldsymbol{B}^{(t)} = (a^{(t)}, b^{(t)}, \boldsymbol{\beta}^{(t)}, (\tau^{(t)})^2)$ et procédons à la $(t+1)^{\text{e}}$ étape. Cette procédure de maximisation est répétée jusqu'à la convergence de l'estimation $\boldsymbol{B}^{(t)}$. L'EMV de $\boldsymbol{B}$ est $\hat{\boldsymbol{B}} = \boldsymbol{B}^{(\infty)}$ une fois que la convergence est établie.

3.2 Estimation ponctuelle et intervalle de confiance de θ_i

Selon la technique classique, nous posons que l'estimateur sur petits domaines de θ_i est

$$\hat{\theta}_i = E(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B})\Big|_{\boldsymbol{B} = \hat{\boldsymbol{B}}}, \quad (7)$$

l'espérance de θ_i par rapport à la densité conditionnelle $\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B})$ où l'estimation du maximum de vraisemblance $\hat{\boldsymbol{B}}$ est « inséré » pour remplacer $\boldsymbol{B}$. L'estimation $\hat{\theta}_i$ est calculée numériquement en utilisant la procédure Monte Carlo (6) décrite à la section précédente. Dans la suite, le paramètre $\boldsymbol{B}$ inconnu sera remplacé par $\hat{\boldsymbol{B}}$ dans toutes les quantités dans lesquelles il intervient, même si nous continuons d'utiliser la notation $\boldsymbol{B}$ pour simplifier.

En outre, nous élaborons un intervalle de confiance pour θ_i fondé sur une théorie de la décision. Comme l'ont fait Joshi (1969), Casella et Hwang (1991), Hwang et coll. (2009), considérons la fonction de perte associée à l'intervalle de confiance C donnée par $(k/\sigma)L(C) - I_C(\theta)$, où k est un paramètre de mise au point indépendant des paramètres du modèle, $L(C)$ est la longueur de C et $I_C(\theta)$ est la fonction indicatrice prenant la valeur 1 ou 0 selon que $\theta \in C$ ou non. Notons que cette fonction de perte tient compte à la fois de la probabilité de couverture et de la longueur de l'intervalle ; la quantité positive (k/σ) sert de poids relatif de la longueur comparativement à la probabilité de couverture de l'intervalle de confiance. Si $k = 0$, la longueur de l'intervalle n'est pas prise en considération, de sorte que la valeur optimale de C est $(-\infty, \infty)$ avec une probabilité de couverture de 1. Par ailleurs, pour $k = \infty$, la probabilité de couverture est égale à 0, de sorte que la valeur optimale de C est un ensemble de points. Pour obtenir l'intervalle de confiance de Bayes de θ_i il faut minimiser la fonction de risque (la perte prévue) $E\{[(k/\sigma)L(C) - I_C(\theta)] \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B})\}$. Le choix optimal de C est donné par

$$C_i(\boldsymbol{B}) = \{\theta_i : kE(\sigma_i^{-1} \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}) < \pi(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B})\}. \quad (8)$$

Puisque que $C_i(\boldsymbol{B})$ s'obtient en minimisant le risque a posteriori, on pourrait vouloir l'interpréter comme un

ensemble crédible bayésien. Cependant, à l'instar de Casella et Berger (1990, page 470), nous continuerons de donner à $C_i(\boldsymbol{B})$ le nom d'intervalle de confiance. Dans une perspective bayésienne empirique également, cette terminologie est plus appropriée. Nous montrerons à la section 3.3 comment le paramètre de mise au point k détermine le niveau de confiance de $C_i(\boldsymbol{B})$.

En supposant pour le moment que k est connu, nous suivons les étapes ci-après pour calculer $C_i(\boldsymbol{B})$. Les densités conditionnelles de σ_i^2 et θ_i sont données par

$$\pi(\sigma_i^2 \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}) \propto \frac{\exp\left[\dfrac{-0{,}5(X_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{(\sigma_i^2 + \tau^2) - \left\{0{,}5(n_i - 1)S_i^2 + \dfrac{1}{b}\right\}\left(\dfrac{1}{\sigma_i^2}\right)}\right]}{(\sigma_i^2)^{(n_i-1)/2+a+1}(\sigma_i^2 + \tau^2)^{1/2}} \quad (9)$$

et (5), respectivement, expressions qui, comme nous l'avons mentionné plus haut, n'ont pas de forme analytique. Donc, comme dans le cas de θ_i, nous calculons $E(\sigma_i^{-1} \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B})$ numériquement en utilisant la méthode Monte Carlo par approximation de la valeur prévue de la moyenne $1/N \sum_{k=1}^{N} 1/\sigma_{i,k}$, où $\sigma_{i,r}^2$, $r = 1, 2, ..., R$ sont les R échantillons tirés de la densité conditionnelle $\pi(\sigma_i^2 \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B})$. La procédure d'acceptation-rejet est utilisée pour tirer des nombres aléatoires de $\pi(\sigma_i^2 \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B})$ avec une loi instrumentale donnée par la loi Gamma inverse

$$\frac{\exp\left[-\left\{0{,}5(n_i - 1)S_i^2 + \dfrac{1}{b}\right\}\left(\dfrac{1}{\sigma_i^2}\right)\right]}{(\sigma_i^2)^{(n_i-1)/2+a+1}},$$

et la probabilité d'acceptation

$$\frac{\exp\left\{\dfrac{-0{,}5(X_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{(\sigma_i^2 + \tau^2)}\right\}}{(\sigma_i^2 + \tau^2)^{1/2}} \times \exp(0{,}5) \times |X_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}|.$$

L'étape suivante consiste à déterminer les valeurs des bornes de $C_i(\boldsymbol{B})$ en trouvant deux valeurs de θ_i qui satisfont l'équation $kE(\sigma_i^{-1} \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}) - \pi(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}) = 0$. Il faut pour cela que la constante de normalisation donnée en (5)

$$D_i = \int_{-\infty}^{\infty} \exp\{-0{,}5(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2/\tau^2\}\, \psi_i^{-(n_i/2+a)}\, d\theta_i$$

soit évaluée numériquement. Nous le faisons en procédant à l'intégration de Gauss-Hermite avec 20 nœuds.

3.3 Choix de k

Nous choisissons pour le paramètre de mise au point k dans (8) l'expression

$$k = k(\boldsymbol{B}) = u_{i,0}\,\phi\left(t_{\alpha/2}\sqrt{\frac{n_i + 2a + 2}{n_i - 1}}\right) \quad (10)$$

où ϕ est la distribution normale standard, $t_{\alpha/2}$ est le $(1 - \alpha/2)^{e}$ centile de la distribution t avec $(n_i - 1)$ degrés de liberté, et $u_{i,0} = \sqrt{1 + \sigma_i^2/\tau^2}$. Puisque $u_{i,0}$ fait intervenir σ_i^2 qui est inconnue, une version estimée $\hat{u}_{i,0}$ s'obtient en introduisant l'estimation du maximum a posteriori

$$\hat{\sigma}_i^2 = \hat{\sigma}_i^2(\hat{\boldsymbol{B}}) = \arg\max_{\sigma_i^2} \pi(\sigma_i^2 \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B})\Big|_{\boldsymbol{B}=\hat{\boldsymbol{B}}} \quad (11)$$

à la place de σ_i^2. En outre, nous remplaçons $\boldsymbol{B}$ par $\hat{\boldsymbol{B}}$ dans (11). Nous démontrons que la probabilité de couverture de $C_i(\hat{\boldsymbol{B}})$ avec ce choix de k s'approche de $1 - \alpha$. Les justifications théoriques sont présentées à la section 4.

3.4 Autres méthodes apparentées aux fins de comparaison

Nous donnons à notre méthode le nom de méthode I. Nous décrivons brièvement ci-dessous trois autres méthodes auxquelles nous la comparerons.

Méthode II : Wang et Fuller (2003) ont considéré le modèle d'estimation sur petits domaines de Fay-Herriot donné par (1). Leur principale contribution est la construction de la formule d'estimation de l'erreur quadratique moyenne pour les estimateurs sur petits domaines avec variances d'échantillonnage estimées. Ce faisant, ils ont construit deux formules désignées par $\widehat{\text{EQM}}_1$ et $\widehat{\text{EQM}}_2$. Pour nos comparaisons, nous utilisons $\widehat{\text{EQM}}_1$, qui a été dérivée en suivant l'approche de correction du biais de Prasad et Rao (1990). La différence fondamentale par rapport à notre approche est qu'ils n'ont pas lissé les variances d'échantillonnage, et n'ont tenu compte de l'incertitude que dans l'inférence au sujet des paramètres de petit domaine. La méthode d'estimation des paramètres, qui est fondée sur les moments pour tous les paramètres du modèle, diffère également de la nôtre.

Méthode III : Hwang et coll. (2009) ont considéré les modèles log-normal et Gamma inverse pour σ_i^{-2} dans (2) pour l'analyse des données micro vectorielles. Leur étude par simulation a montré que les propriétés des intervalles de confiance des estimateurs sur petits domaines étaient meilleures sous modèle log-normal que sous le modèle Gamma inverse. Nous avons donc modifié leur modèle log-normal afin d'ajouter des covariables et des tailles d'échantillon n_i inégales comme il suit :

$$\left.\begin{aligned} X_i|\theta_i,\sigma_i^2 &\sim \text{Normale}(\theta_i,\sigma_i^2) \\ \theta_i &\sim \text{Normale}(\boldsymbol{Z}_i^T\boldsymbol{\beta},\tau^2);\end{aligned}\right\} \quad (12)$$

$$\left.\begin{aligned} \log S_i^2 = \log(\sigma_i^2)+\delta_i\ ;\ \delta_i &\sim N(m_i,\sigma_{ch,i}^2) \\ \log(\sigma_i^{-2}) &\sim N(\mu_v,\tau_v^2),\end{aligned}\right\} \quad (13)$$

indépendamment pour $i = 1, 2, ..., n$. Notons que le modèle de la moyenne dans (12) est identique à celui figurant dans (1). Les quantités τ^2, m_i et $\sigma_{ch,i}^2$ sont supposées connues et sont données par $m_i = E[\log(\chi_{n_i-1}^2/(n_i-1))]$ et $\sigma_{ch,i}^2 = \text{Var}[\log(\chi_{n_i-1}^2/(n_i-1))]$. Donc, la taille d'échantillon n_i détermine la forme de la distribution χ^2 par la voie du paramètre de nombre de degrés de liberté mais surtout, comme nous l'avons mentionné plus haut, les tailles d'échantillon différentes expliquent différents degrés de rétrécissement du paramètre de variance réelle correspondant. Comme dans leur approche d'estimation, les paramètres μ_v et τ_v^2 inconnus du modèle sont estimés selon une méthode fondée sur le moment dans un cadre bayésien empirique donnant $\hat{\mu}_v$ et $\hat{\tau}_v^2$, respectivement. Notons que, dans Hwang et coll. (2009), des estimations sont obtenues en se basant sur le modèle hiérarchique pour σ_i^2 dans (13) *seulement*, sans se préoccuper de la modélisation (1) de la moyenne. Nous renvoyons le lecteur à la section 5 de leur article pour des renseignements détaillés sur l'estimation des hyperparamètres. Nous suivons la même procédure en utilisant uniquement (13) pour estimer μ_v et τ_v^2 dans le cas de tailles d'échantillon inégales.

La dérivation de l'estimation bayésienne de σ_i^2 est

$$\hat{\sigma}_{i,B}^2 = \exp\left[E\{\ln(\sigma_i^2)\mid \ln(S_i^2)\}\right]$$

$$= \left\{\frac{S_i^2}{\exp(m_i)}\right\}^{M_{v,i}} \exp\{\mu_v(1-M_{v,i})\}$$

où $M_{v,i} = \tau_v^2/(\tau_v^2+\sigma_{ch,i}^2)$ et avec insertion des estimations pour remplacer les quantités inconnues. La distribution conditionnelle de θ_i sachant (X_i, S_i^2), qui est donnée par

$$\pi(\theta_i\mid X_i,S_i^2) = \int_0^\infty \pi(\theta_i\mid X_i,S_i^2,\sigma_i^2)\,\pi(\sigma_i^2\mid X_i,S_i^2)\,d\sigma_i^2,$$

est approximée par $\pi(\theta_i\mid X_i,S_i^2) \approx \int_0^\infty \pi(\theta_i\mid X_i,S_i^2,\hat{\sigma}_{i,B}^2)\,\pi(\sigma_i^2\mid X_i,S_i^2)\,d\sigma_i^2 = \pi(\theta_i\mid X_i,S_i^2,\hat{\sigma}_{i,B}^2)$. Cela suggère l'estimateur bayésien approximatif des paramètres de petit domaine donné par

$$\hat{\theta}_i = E(\theta_i\mid X_i,\hat{\sigma}_{i,B}^2) = \hat{M}_i X_i + (1-\hat{M}_i)\boldsymbol{Z}_i^T\hat{\boldsymbol{\beta}}, \quad (14)$$

où $\hat{M}_i = \hat{\tau}_v^2/(\hat{\tau}_v^2+\hat{\sigma}_{i,B}^2)$. L'intervalle de confiance pour θ_i s'obtient sous la forme

$$C_i^H = \left\{\theta_i : \frac{|\theta_i-\hat{\theta}_i|}{\hat{M}_i\hat{\sigma}_{i,B}^2} < -2\ln\{k\sqrt{2\pi}\} - \ln(\hat{M}_i)\right\}. \quad (15)$$

À la section 3 de Hwang et coll. (2009), pages 269 à 271, l'intervalle C_i^H est apparié avec l'intervalle t à $100(1-\alpha)\,\%$ $[|\theta_i - X_i| < tS_i]$ pour obtenir l'expression de k comme $k \equiv k_i = \exp\{-t^2/2\}\exp\{m_i/2\}/(\sqrt{2\pi})$.

Méthode IV : Cette méthode comprend un cas particulier du modèle de Fay-Herriot donné en (1), mais avec l'estimation des paramètres du modèle empruntée à Qiu et Hwang (2007). Ces derniers ont considéré le modèle

$$\left.\begin{aligned} X_i\mid\theta_i,\sigma^2 &\sim \text{Normale}(\theta_i,\sigma^2) \\ \theta_i &\sim \text{Normale}(0,\tau^2),\end{aligned}\right\} \quad (16)$$

indépendamment pour $i = 1, 2, ..., n$, pour analyser des données micro vectorielles expérimentales. Dans le cas où les paramètres du modèle étaient connus, ils ont proposé l'estimateur ponctuel $\hat{\theta}_i = \hat{M}X_i, \hat{M} = (1-((n-2)\sigma^2/|X|^2))_+$ où a_+ désigne $\max(0,a)$ pour tout nombre a et $|X| = |(\sum_{i=1}^n X_i^2)^{1/2}$. L'intervalle de confiance pour θ_i est $\hat{\theta}_i \pm v_1(\hat{M})$, où $v_1^2(\hat{M}) = \sigma^2\hat{M}(q_1 - \ln(\hat{M}))$ avec q_1 désignant la valeur critique de la variable normale standard pour le niveau de confiance souhaité et $v_1(0) \equiv 0$. Ici, en vue de procéder à la comparaison avec notre méthode, nous modifions le premier niveau du modèle hiérarchique dans (16) comme il suit :

$$X_i = \boldsymbol{Z}_i^T\boldsymbol{\beta} + v_i + e_i$$

où $v_i \sim \text{Normale}(0,\tau^2)$ et $e_i \sim \text{Normale}(0,S_i^2)$ indépendamment pour $i = 1, 2, ..., n$, et S_i^2 est traité comme étant connu. Comme Qiu et Hwang (2007), nous estimons τ^2 par

$$\tilde{\tau}^2 = \frac{1}{n-p}\left[\sum_i \hat{u}_i^2 - \sum_i S_i^2\left\{1-\boldsymbol{Z}_i^T\left(\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T\right)^{-1}\boldsymbol{Z}_i^T\right\}\right]$$

et $\hat{\tau}^2 = \max(\tilde{\tau}^2, 1/n)$, où $\hat{u}_i = X_i - \boldsymbol{Z}_i^T\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{\beta}} = (\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T)^{-1}(\sum_{i=1}^n \boldsymbol{Z}_i X_i)$. Puis nous définissons $\hat{M}_{0i} = \hat{\tau}^2/(\hat{\tau}^2+S_i^2)$ et $\hat{M}_i = \max(\hat{M}_{0i}, M_1)$, où , dans la dernière expression, $\hat{M}_{0i}$ est tronqué par $M_{1i} = 1 - Q_\alpha/(n_i-2)$, et Q_α est le α^{e} quantile d'une distribution du khi-carré à n_i degrés de liberté. Cet $\hat{M}_i$ est utilisé dans la formule de l'intervalle de confiance de θ_i donnée plus haut. Quand nous avons appliqué cette méthode dans notre étude par simulation et notre analyse des données réelles, nous avons modifié le modèle afin de pouvoir utiliser les tailles d'échantillon inégales et l'information sur les covariables mentionnées plus haut.

Remarque 1. Hwang et coll. (2009) ont choisi k en prenant (15) égale à l'intervalle t fondé sur X_i seulement pour les paramètres de petit domaine θ_i. Notons que X_i est l'estimateur direct d'après les données d'enquête. Par conséquent, ce choix de k n'exerce aucun contrôle direct sur la probabilité de couverture de l'intervalle construit sous *estimation par rétrécissement*. Par ailleurs, notre choix proposé de k a été établi de manière à maintenir la couverture nominale sous, précisément, l'estimation par rétrécissement.

Remarque 2. Notons qu'en l'absence de toute hypothèse de modélisation hiérarchique, S_i et X_i sont indépendants car S_i^2 et X_i sont, respectivement, auxiliaire et la statistique exhaustive complète pour θ_i. Cependant, sous les modèles (1) et (2), la distribution conditionnelle de σ_i^2 et θ_i fait intervenir à la fois X_i et S_i^2, ce que l'on peut constater en examinant (5) et (9).

Remarque 3. Dans Hwang et coll. (2009), l'estimateur à rétrécisseur de σ_i^2 est fondé uniquement sur l'information au sujet de S_i^2, et non au sujet de X_i ainsi que S_i^2. L'estimateur bayésien de σ_i^2 est introduit par insertion dans l'expression de l'estimateur bayésien des paramètres de petit domaine. Donc, l'estimateur sur petits domaines de Hwang et coll. s'écrit sous la forme $E(\theta_i \mid X_i, \hat{\sigma}_{i,B}^2)$ dans (14) où $\hat{\sigma}_{i,B}^2$ est l'estimateur bayésien de σ_i^2. En raison de l'équation (9), l'estimateur à rétrécisseur de σ_i^2 dépend de $(X_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2$ en plus de S_i^2 contrairement à l'estimateur de Hwang et coll. (2009). Nous pensons que cela pourrait être l'explication de la meilleure performance de notre méthode comparativement à celle de Hwang et coll. (2009).

Remarque 4. Comme nous l'avons mentionné plus haut, le nombre de degrés de liberté associés à la distribution χ^2 pour la variance d'échantillonnage ne doit pas être simplement $n_i - 1$, n_i étant la taille de l'échantillon pour le i^e domaine. Il n'existe aucun résultat théorique fiable pour déterminer le nombre de degrés de liberté quand le plan de sondage est complexe. Wang et Fuller (2003) ont approximé la distribution χ^2 par une distribution normale fondée sur l'approximation de Wilson-Hilferty. Si l'on connaît le plan de sondage exact, les lignes directrices basées sur la simulation de Maples et coll. (2009) pourraient être utiles. Pour produire des estimations au niveau du comté en se servant des données de l'American Community Survey, Maples et coll. (2009) ont suggéré d'estimer le nombre de degrés de liberté par $0{,}36 \times \sqrt{n_i}$.

4. Justification théorique

À la présente section, nous donnons la justification théorique du choix de k suivant l'équation (10). Comme dans Hwang et coll. (2009), la distribution conditionnelle de θ_i sachant X_i et S_i^2 peut être approximée par $\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{B}) \approx \pi(\theta_i \mid X_i, S_i^2, \boldsymbol{B}, \hat{\sigma}_i^2)$, où $\hat{\sigma}_i^2$ est défini comme dans (11). De la même façon, nous approximons $E(\sigma_i^{-1} \mid X_i, S_i^2, \boldsymbol{B})$ par $E(\sigma_i^{-1} \mid X_i, S_i^2, \boldsymbol{B}) \approx \hat{\sigma}_i^{-1}$. Sur la base de ces approximations, nous avons $C_i(\boldsymbol{B}) \approx \tilde{C}_i(\boldsymbol{B})$ où $\tilde{C}_i(\boldsymbol{B})$ est l'intervalle de confiance de θ_i donné par $\tilde{C}_i(\boldsymbol{B}) = \{\theta_i : \pi(\theta_i \mid X_i, S_i^2, \boldsymbol{B}, \hat{\sigma}_i^2) \geq k\hat{\sigma}_i^{-1}\}$. De (1) il découle que la densité de probabilité conditionnelle $\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{B}, \sigma_i^2)$ est normale de moyenne μ_i et de variance ν_i, où μ_i et ν_i sont donnés par les expressions

$$\mu_i = w_i X_i + (1 - w_i)\boldsymbol{Z}_i^T\boldsymbol{\beta},$$
$$\nu_i = \left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)^{-1} = \sigma_i^2\left(1 + \frac{\sigma_i^2}{\tau^2}\right)^{-1}, \qquad (17)$$

et

$$w_i = \frac{1/\sigma_i^2}{(1/\sigma_i^2 + 1/\tau^2)}.$$

Maintenant, en choisissant

$$k = \hat{u}_0\phi\left(t_{\alpha/2}\sqrt{\frac{n_i + 2a + 2}{n_i - 1}}\right)$$

comme nous l'avons mentionné, l'intervalle de confiance $\tilde{C}_i(\boldsymbol{B})$ devient

$$\tilde{C}_i(\boldsymbol{B}) = \left\{\theta_i : \hat{u}_{0i}\frac{|\theta_i - \hat{\mu}_i|}{\hat{\sigma}_i} \leq t_{\alpha/2}\sqrt{\frac{n_i + 2a + 2}{n_i - 1}}\right\}, \quad (18)$$

où $\hat{\mu}_i$ est l'expression de μ_i dans (17) avec remplacement de σ_i^2 par $\hat{\sigma}_i^2$. Considérons maintenant le comportement de $\hat{\sigma}_i^2 \equiv \hat{\sigma}_i^2(\boldsymbol{B})$ quand τ^2 varie entre 0 et ∞. Quand $\tau^2 \to \infty$, $\hat{\sigma}_i^2$ converge vers

$$\hat{\sigma}_i^2(\infty) \equiv \hat{\sigma}_i^2(a, b, \boldsymbol{\beta}, \infty) = \frac{\frac{(n_i - 1)}{2}S_i^2 + \frac{1}{b}}{\frac{n_i - 1}{2} + a + 1} = \frac{(n_i - 1)S_i^2 + \frac{2}{b}}{n_i + 2a + 1}.$$

De même, quand $\tau^2 \to 0$, $\hat{\sigma}_i^2$ converge vers

$$\hat{\sigma}_i^2(0) \equiv \hat{\sigma}_i^2(a, b, \boldsymbol{\beta}, 0) = \frac{(X_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2 + (n_i - 1)S_i^2 + \frac{2}{b}}{n_i + 2a + 2}.$$

Pour toute valeur intermédiaire de τ^2, nous avons $\min\{\hat{\sigma}_i^2(0), \hat{\sigma}_i^2(\infty)\} \leq \hat{\sigma}_i^2 \leq \max\{\hat{\sigma}_i^2(0), \hat{\sigma}_i^2(\infty)\}$. Donc, il

est suffisant de considérer les deux cas suivants : i) $\hat{\sigma}_i^2 \geq \hat{\sigma}_i^2(\infty)$, où il s'ensuit que $(n_i + 2a + 2)\hat{\sigma}_i^2 = (n_i + 2a + 1)\hat{\sigma}_i^2 + \hat{\sigma}_i^2 \geq (n_i - 1)S_i^2 + 2/b + \hat{\sigma}_i^2 \geq (n_i - 1)S_i^2$, et ii) $\hat{\sigma}_i^2 \geq \hat{\sigma}^2(0)$, où il s'ensuit que $(n_i + 2a + 2)\,\hat{\sigma}_i^2 = (X_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2 + (n_i - 1)S_i^2 + 2/b \geq (n_i - 1)S_i^2$. Donc, dans les cas (i) ainsi que (ii),

$$(n_i + 2a + 2)\,\hat{\sigma}_i^2 \geq (n_i - 1)S_i^2. \tag{19}$$

Puisque $\theta_i - \mu_i \sim N(0, \sigma_i^2\tau^2/(\sigma_i^2 + \tau^2))$ et $(n_i - 1)S_i^2/\sigma_i^2 \sim \chi^2_{n_i - 1}$, l'intervalle de confiance

$$D_i = \left\{\theta_i : u_{0i}\frac{|\theta_i - \mu_i|}{S_i} \leq t_{\alpha/2}\right\} \tag{20}$$

a une probabilité de couverture de $1 - \alpha$. Donc, si u_0 et μ_i sont remplacés par $\hat{u}_0$ et $\hat{\mu}_i$, il faut s'attendre à ce que l'intervalle de confiance résultant $\tilde{D}_i$, disons, ait une probabilité de couverture d'environ $1 - \alpha$. De (19), nous obtenons

$$P\{\tilde{C}_i(\boldsymbol{B})\} \geq P(\tilde{D}_i) \approx 1 - \alpha, \tag{21}$$

ce qui établit une borne inférieure approximative de $1 - \alpha$ pour le seuil de confiance de $\tilde{C}_i(\boldsymbol{B})$.

Dans (21), $\boldsymbol{B}$ était supposé fixe et connu. Quand $\boldsymbol{B}$ est inconnu, nous le remplaçons par l'estimation de son maximum de vraisemblance marginale $\hat{\boldsymbol{B}}$. Puisque l'expression (21) est vérifiée quelle que soit la valeur réelle de $\boldsymbol{B}$, la substitution de $\hat{\boldsymbol{B}}$ à $\boldsymbol{B}$ dans (21) comportera une erreur d'ordre $O(1/\sqrt{N})$, où $N = \sum_{i=1}^{n} n_i$. Comparativement à chaque n_i pris individuellement, ce groupement des n_i devrait réduire l'erreur de manière significative, de manière que $\tilde{C}_i(\hat{\boldsymbol{B}})$ soit suffisamment proche de $\tilde{C}_i(\boldsymbol{B})$ pour satisfaire la borne inférieure de $1 - \alpha$ dans (21).

5. Une étude par simulation

5.1 Conditions de simulation

Nous considérons les conditions de simulation dans lesquelles nous utilisons un sous-ensemble de configurations des paramètres emprunté à Wang et Fuller (2003). Chaque échantillon employé dans l'étude par simulation a été obtenu en suivant les étapes que voici. Premièrement, générer des observations en utilisant le modèle

$$X_{ij} = \beta + u_i + e_{ij},$$

où $u_i \sim N(0, \tau^2)$ et $e_{ij} \sim N(0, n_i\sigma_i^2)$, indépendamment pour $j = 1, ..., n_i$ et $i = 1, ..., n$. Alors, le modèle à effets aléatoires pour la moyenne de petit domaine, X_i, est

$$X_i = \beta + u_i + e_i, \quad \text{indépendamment pour } i = 1, ..., n,$$

où $X_i \equiv \bar{X}_{i\cdot} \equiv n_i^{-1}\sum_{j=1}^{n_i} X_{ij}$ et $e_i \equiv \bar{e}_{i\cdot} \equiv n_i^{-1}\sum_{j=1}^{n_i} e_{ij}$. Donc, $X_i \sim N(\theta_i, \sigma_i^2)$, où $\theta_i = \beta + u_i$, $\theta_i \sim N(\beta, \tau^2)$ et $e_i \sim N(0, \sigma_i^2)$. Nous avons estimé σ_i^2 en nous servant de l'estimateur sans biais

$$S_i^2 = (n_i - 1)^{-1}n_i^{-1}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2,$$

et il s'ensuit que $(n_i - 1)S_i^2/\sigma_i^2 \sim \chi^2_{n_i - 1}$, indépendamment pour $i = 1, 2, ..., n$. Notons que le plan de simulation ne tenait pas compte de la modélisation des variances d'échantillonnage au deuxième niveau dans (2). Par conséquent, notre résultat indiquera une robustesse à l'erreur de spécification du modèle de variance.

Les étapes susmentionnées ont produit les données $(X_i, S_i^2), i = 1, ..., n$. Pour simplifier la simulation, nous ne choisissons aucune covariable $\boldsymbol{Z}_i$. À l'instar de Wang et Fuller (2003), nous donnons la valeur m à tous les n_i afin de faciliter la programmation. Cependant, nous choisissons quand même que les variances d'échantillonnage réelles soient inégales : la valeur d'un tiers des σ_i^2 est fixée à 1, celle d'un deuxième tiers est fixée à 4 et celle du dernier tiers est fixée à 16. Nous prenons $\beta = 10$ et trois valeurs différentes de $\tau^2 = 0{,}25$, 1 et 4. Nous avons choisi ces valeurs des paramètres en nous inspirant de Qiu et Hwang (2007). Pour chaque valeur de τ^2, nous avons généré 200 échantillons pour les deux combinaisons $(m, n) = (9, 36)$ et (18, 180).

Dans l'étude par simulation, nous comparons la méthode que nous proposons aux méthodes de Wang et Fuller (2003), Hwang et coll. (2009), et Qiu et Hwang (2007) que nous appelons méthodes I, II, III et IV, respectivement, en nous fondant sur le biais, l'erreur quadratique moyenne (EQM), la probabilité de couverture (PC) des intervalles de confiance et la longueur moyenne des intervalles de confiance (LMIC). Le tableau 1 donne les estimations des paramètres pour a, b, β et τ^2. Les résultats numériques indiquent que les estimations du maximum de vraisemblance des paramètres du modèle ont de bonnes propriétés ; les valeurs estimées de β et τ^2 sont proches des valeurs réelles, ce qui témoigne de bonnes propriétés de robustesse à l'erreur de spécification de la distribution au deuxième niveau de (2). L'obtention d'estimations statistiquement significatives pour a ainsi que b indique que les variances d'échantillonnage « rétrécies » sont intégrées dans la méthode proposée. Les tableaux 2, 3 et 4 donnent les moyennes des résultats numériques calculées sur les domaines qui, dans chaque groupe, ont les mêmes variances d'échantillonnage réelles. Les résultats des tableaux sont fondés sur 200 répliques.

Tableau 1
Résultats des simulations pour les paramètres du modèle, a (panneau supérieur gauche), b (panneau supérieur droit), β (panneau inférieur gauche) et τ^2 (panneau inférieur droit). Ici, É.-T. représente l'écart-type sur 200 répliques. Nous avons pris $\beta = 10$ et $\tau^2 = 0,25,\ 1$ et 4

	$n = 36,\ m = 9$		$n = 180,\ m = 18$			$n = 36,\ m = 9$		$n = 180,\ m = 18$	
τ^2	Moyenne	É.-T.	Moyenne	É.-T.	τ^2	Moyenne	É.-T.	Moyenne	É.-T.
		a					b		
0,25	1,0959	0,1540	1,0328	0,0442	0,25	0,3992	0,0983	0,4249	0,0323
1	1,0937	0,1555	1,0325	0,0445	1	0,4030	0,1012	0,4253	0,0326
4	1,0996	0,1577	1,0339	0,0450	4	0,3999	0,1017	0,4245	0,0328
		β					τ^2		
0,25	10,0071	0,3618	9,9951	0,1853	0,25	0,2558	0,0605	0,2575	0,0097
1	10,0142	0,3311	9,9970	0,1743	1	0,9418	0,3333	1,0426	0,1264
4	10,0282	0,4639	10,0048	0,2254	4	3,5592	1,3316	4,0817	0,5551

Tableau 2
Résultats des simulations pour la prédiction quand $\tau^2 = 0,25$. Ici, EQM, LMIC et PC représentent l'erreur quadratique moyenne, la longueur moyenne de l'intervalle de confiance et la probabilité de couverture de l'intervalle de confiance, respectivement

		$n = 36,\ m = 9$				$n = 180,\ m = 18$			
		Méthode				Méthode			
	σ_i^2	I	II	III	IV	I	II	III	IV
Biais	1	0,0048	0,0198	0,0272	0,0018	-0,0051	-0,0086	-0,0112	-0,0111
relatif	4	-0,0033	-0,0061	-0,0145	-0,0158	-0,0130	-0,0109	-0,0065	-0,0116
	16	0,0126	0,0370	0,0369	0,0096	-0,0046	-0,0045	-0,0080	-0,0061
EQM	1	0,3066	0,3890	0,6861	0,3805	0,2258	0,2680	0,4470	0,2922
	4	0,3281	0,5430	1,3778	0,7285	0,2595	0,3000	0,5805	0,3748
	16	0,3715	0,5240	1,6749	1,9316	0,2815	0,2850	0,4856	0,6383
LMIC	1	2,1393	2,5485	4,4906	3,0528	1,9220	1,6006	3,6466	2,4811
	4	2,2632	3,9574	6,8887	5,6842	2,0557	2,1524	5,2472	4,2160
	16	2,3221	4,5619	9,3335	11,1363	2,1046	2,3308	6,5273	7,8492
PC	1	0,9468	0,9770	0,9771	0,9708	0,9564	0,9710	0,9851	0,9631
	4	0,9468	0,9710	0,9829	0,9917	0,9555	0,9660	0,9967	0,9967
	16	0,9365	0,9660	0,9933	0,9975	0,9529	0,9610	0,9998	0,9999

Tableau 3
Résultats des simulations pour la prédiction quand $\tau^2 = 1$. Ici, EQM, LMIC et PC représentent l'erreur quadratique moyenne, la longueur moyenne de l'intervalle de confiance et la probabilité de couverture de l'intervalle de confiance, respectivement

		$n = 36,\ m = 9$				$n = 180,\ m = 18$			
		Méthode				Méthode			
	σ_i^2	I	II	III	IV	I	II	III	IV
Biais	1	-0,0152	0,0205	0,0255	0,0051	-0,0064	-0,0085	-0,0111	-0,0101
relatif	4	-0,0167	-0,0164	-0,0151	-0,0219	-0,0151	-0,0121	-0,0133	-0,0164
	16	-0,0323	0,0508	0,0515	0,0216	-0,0028	-0,0017	-0,0073	-0,0039
EQM	1	0,5645	0,6330	0,7238	0,6260	0,5288	0,5430	0,5673	0,6336
	4	0,8566	1,1100	1,5396	1,0992	0,8159	0,8770	0,9415	0,8948
	16	1,0482	1,3100	2,1059	2,3156	0,9786	1,0000	1,1024	1,1878
LMIC	1	3,4550	3,1822	4,4938	3,2117	3,1088	2,5094	3,6763	2,8676
	4	4,0321	5,8733	6,8984	5,7909	3,7844	4,2908	5,3323	4,5543
	16	4,4082	7,4286	9,3555	11,1555	4,1187	5,1590	6,6785	7,8937
PC	1	0,9704	0,9640	0,9762	0,9275	0,9660	0,9650	0,9786	0,8879
	4	0,9633	0,9560	0,9812	0,9808	0,9627	0,9680	0,9918	0,9740
	16	0,9533	0,9490	0,9912	0,9938	0,9613	0,9680	0,9974	0,9979

Tableau 4
Résultats des simulations pour la prédiction quand $\tau^2 = 4$. Ici, EQM, LMIC et PC représentent l'erreur quadratique moyenne, la longueur moyenne de l'intervalle de confiance et la probabilité de couverture de l'intervalle de confiance, respectivement

		$n = 36,\ m = 9$				$n = 180,\ m = 18$			
		Méthode				Méthode			
	σ_i^2	I	II	III	IV	I	II	III	IV
Biais	1	-0,0024	0,0248	0,0229	0,0180	-0,0084	-0,0098	-0,0122	-0,0106
relatif	4	-0,0343	-0,0310	-0,0210	-0,0340	-0,0110	-0,0092	-0,0174	-0,0132
	16	-0,0147	0,0702	0,0767	0,0467	0,0016	0,0024	-0,0059	0,0012
EQM	1	0,8822	0,8590	0,8579	1,0559	0,8359	0,8180	0,8541	0,8605
	4	2,0577	2,2900	2,1818	2,2422	2,0424	2,1000	2,0935	2,1130
	16	3,4516	3,7600	3,9267	3,8981	3,3153	3,3500	3,3939	3,3631
LMIC	1	4,6318	4,1936	4,5369	3,7677	4,0256	3,5346	3,9626	3,7499
	4	6,2015	10,9093	7,0376	6,4314	5,9000	9,0913	6,2217	6,1540
	16	7,7221	18,0039	9,6718	11,3341	7,4430	14,6665	8,3908	8,7537
PC	1	0,9791	0,9670	0,9733	0,9029	0,9674	0,9570	0,9600	0,9468
	4	0,9556	0,9670	0,9725	0,9496	0,9592	0,9610	0,9633	0,9573
	16	0,9510	0,9670	0,9796	0,9858	0,9573	0,9650	0,9718	0,9776

Comparaisons des biais : Dans la plupart des cas, les biais des quatre méthodes sont comparables. Il n'existe aucune preuve manifeste d'écarts significatifs entre elles pour ce qui est du biais. Une forte variance d'échantillonnage donne plus de poids à la moyenne de population par construction, ce qui rend l'estimateur plus proche de la moyenne au deuxième niveau. Par ailleurs, les méthodes I à III comprennent l'utilisation d'estimateurs à rétrécisseur des variances d'échantillonnage qui seraient donc inférieurs au maximum de l'ensemble des variances d'échantillonnage. Donc, les méthodes I à III ont tendance à présenter un biais un peu plus important. Cependant, en raison du rétrécissement des variances d'échantillonnage, on peut s'attendre à une amélioration de la variance des estimateurs qui, à son tour, réduit l'EQM. Parmi les méthodes I à III, la méthode I a donné de meilleurs résultats que les méthodes II et III, dont les propriétés étaient assez semblables. Le gain maximal en utilisant la méthode I au lieu de la méthode II est de 99 %.

Comparaison des EQM : En ce qui concerne l'EQM, la méthode I a donné systématiquement de meilleurs résultats que les trois autres dans tous les cas, sauf quand le ratio de σ_i^2 à τ^2 était le plus faible : $(\sigma_i^2 = 1) / (\tau^2 = 4) = 0{,}25$. Dans ce cas, la variance entre les petits domaines (variance du modèle) est beaucoup plus grande que la variance dans les domaines (variance d'échantillonnage). Lorsque notre méthode est utilisée pour estimer θ_i, l'information « empruntée » à d'autres domaines peut mal orienter l'estimation : la moyenne estimée de la loi Gamma pour σ_i^{-2} provenant du deuxième niveau de (2) est $\hat{a}\hat{b}$, qui est égale à 0,44 environ pour les deux combinaisons (m, n) correspondant à $(9, 36)$ et $(18, 180)$ (la valeur réelle est $ab = 0{,}4$). Donc, $E(\sigma_i^{-2} \mid X_i, S_i^2, \hat{B})$ est significativement plus petite que 1 en raison du rapprochement vers la moyenne pour le groupe pour lequel la valeur réelle est $\sigma_i^2 = 1$. En outre, puisque σ_i^2 est plus faible que τ^2, le poids de X_i devrait être beaucoup plus élevé comparativement à β, la moyenne globale. Cependant, étant donné la sous-estimation de σ_i^{-2} dans ce cas, l'estimateur résultant donne moins de poids à X_i, ce qui donne lieu à une EQM plus grande. Cependant, cette sous-estimation diminue pour les grandes tailles d'échantillon en raison de la cohérence des estimateurs de Bayes. Ce fait s'observe effectivement quand la taille d'échantillon passe de $n = 36$ à $n = 180$ pour $\sigma_i^2 = 1$ et $\tau^2 = 4$. Comparativement à la méthode II, la méthode I produit une amélioration dans la plupart des cas simulés ; le gain maximal est de 30 %, tandis que la seule perte observée est de 9 % pour la combinaison $\sigma_i^2 = 1$ et $\tau^2 = 4$ pour $n = 36$ et $m = 9$. De même, par rapport à la méthode III, le gain maximal donné par la méthode I est de 77 % et la seule perte est de 11 %, pour les mêmes spécifications de paramètres et de tailles d'échantillon.

Comparaisons des PC : Nous avons obtenu les intervalles de confiance au seuil de confiance de 95 %. Les méthodes I et III ne révèlent aucune sous-couverture, ce qui n'est pas étonnant étant donné la construction optimale de leurs intervalles de confiance. La méthode I produit le taux nominal de couverture plus fréquemment que n'importe quelle autre méthode. La méthode II présente une certaine sous-couverture, le taux pouvant être aussi faible que 82 %.

Comparaisons des LMIC : La méthode I produit en général des intervalles de confiance considérablement plus courts

que les autres méthodes. La méthode IV a produit des intervalles de longueur comparable à ceux des autres méthodes dans tous les cas sauf quand σ_i^2 était élevé, auquel cas les longueurs étaient considérablement plus grandes. L'intervalle de confiance proposée dans Qiu et Hwang (2007) n'a pas de bonnes propriétés en échantillon fini, particulièrement pour les petites valeurs de τ^2. Afin d'éviter un faible taux de couverture, ils ont proposé de tronquer $M_0 = \tau^2/(\tau^2 + \sigma_i^2)$ à l'aide d'un nombre positif $M_1 = 1 - Q_\alpha/(\nu - 2)$ pour σ_i^2 connu, où Q_α est le α^e quantile d'une distribution du khi-carré à ν degrés de liberté. Quand le ratio de la variance d'échantillonnage à la variance du modèle, σ_i^2/τ^2, est élevé, M_1 a tendance à être plus grand que M_0, ce qui donne le taux nominal de couverture, mais avec de plus grandes longueurs d'intervalle. Par exemple, dans le cas où $(\sigma_i^2, \tau^2) = (16, 0{,}25)$, la LMIC est de 11,13 pour la méthode IV, alors qu'elle est seulement de 2,78 et 4,56 pour les méthodes I et II, respectivement.

5.2 Étude de la robustesse

Afin d'étudier la robustesse de la méthode proposée aux écarts par rapport à l'hypothèse de normalité des erreurs, nous avons procédé à l'étude par simulation qui suit. Les données ont été générées comme précédemment, mais en tirant les e_{ij} d'une loi exponentielle double (loi de Laplace) et d'une loi uniforme. Les estimateurs des méthodes II et III ont eu peu d'effet. Cela pourrait tenir au fait que, dans ces méthodes, l'estimation des paramètres du modèle se fait par la méthode des moments. La méthode IV a produit de plus grandes valeurs du biais relatif, de l'EQM et de la LMIC, et une plus faible probabilité de couverture. L'EQM est systématiquement plus faible pour la méthode I que pour la méthode II. Quand $\tau^2 = 0{,}25$ et 1, la LMIC est plus petite pour la méthode I que pour la méthode II pour (n = 36, m = 9), mais le résultat inverse s'observe quand (n = 180, m = 18). Pour ce qui est de la PC, la méthode II produit une certaine sous-couverture (taux le plus faible égal à 80 %). Par contre, la méthode I ne produit aucune sous-couverture. Faute d'espace, nous présentons uniquement les résultats pour les paramètres a, b, β et τ^2 sous les erreurs laplaciennes (tableau 5).

6. Analyse de données réelles

Pour illustrer notre méthodologie, nous choisissons un exemple très souvent étudié. Le jeu de données, qui provient du U.S. Department of Agriculture, a été analysé pour la première fois par Battese (1988). Il s'agit de données sur les productions de maïs et de soja dans 12 comtés de l'Iowa. Les tailles d'échantillon pour ces domaines sont faibles, variant de 1 à 5. Faute d'espace, nous considérons uniquement le cas du maïs. Pour les modèles proposés, il faut nécessairement que l'on ait des tailles d'échantillon $n_i > 1$. Par conséquent, nous avons utilisé des données modifiées tirées de You et Chapman (2006) avec $n_i \geq 2$. Les nombres déclarés d'hectares consacrés à la culture du maïs (X_i), qui sont les estimations directes par sondage, sont présentés au tableau 6. Ce tableau donne aussi les variances d'échantillonnage qui sont calculées d'après les données originales sous l'hypothèse d'un échantillonnage aléatoire simple. L'écart-type d'échantillon varie fortement, de 5,704 à 53,999 (le coefficient de variation varie de 0,036 à 0,423). Deux covariables sont considérées dans le tableau 6 : Z_{i1}, le nombre moyen de pixels correspondant à du maïs et Z_{i2}, le nombre moyen de pixels correspondant à du soja, provenant des données de satellite LANDSAT.

Tableau 5
Résultats des simulations pour les paramètres du modèle, a (panneau supérieur gauche), b (panneau supérieur droit), β (panneau inférieur gauche) et τ^2 (panneau inférieur droit) quand les erreurs suivent une loi de Laplace. Ici, É.-T. représente l'écart-type sur 200 répliques. Nous avons pris $\beta = 10$ et $\tau^2 = 0{,}25$, 1 et 4

	$n = 36$, $m = 9$		$n = 180$, $m = 18$			$n = 36$, $m = 9$		$n = 180$, $m = 18$	
τ^2	Moyenne	É.-T.	Moyenne	É.-T.	τ^2	Moyenne	É.-T.	Moyenne	É.-T.
		a					b		
0,25	0,9624	0,1632	0,9471	0,0498	0,25	0,5793	0,1733	0,5279	0,0501
1	0,9628	0,1657	0,9476	0,0497	1	0,5816	0,1777	0,5275	0,0503
4	0,9689	0,1694	0,9487	0,0499	4	0,5758	0,1796	0,5263	0,0503
		β					τ^2		
0,25	9,9736	0,3775	9,9800	0,1773	0,25	0,2696	0,0882	0,2565	0,0074
1	9,9753	0,3709	9,9836	0,1662	1	1,0508	0,2501	1,0403	0,0668
4	9,9736	0,4835	9,9855	0,2161	4	3,9624	1,1719	4,1256	0,4201

Tableau 6
Données sur le maïs provenant de You et Chapman (2006)

Comté	n_i	X_i	Z_{1i}	Z_{2i}	$\sqrt{S_i^2}$
Franklin	3	158,623	318,21	188,06	5,704
Pocahontas	3	102,523	257,17	247,13	43,406
Winnebago	3	112,773	291,77	185,37	30,547
Wright	3	144,297	301,26	221,36	53,999
Webster	4	117,595	262,17	247,09	21,298
Hancock	5	109,382	314,28	198,66	15,661
Kossuth	5	110,252	298,65	204,61	12,112
Hardin	5	120,054	325,99	177,05	36,807

Les estimations de $\boldsymbol{B}$ sont les suivantes : $a = 1{,}707$, $b = 0{,}00135$, $\tau^2 = 90{,}58$ et $\boldsymbol{\beta} = (-186{,}0\ ;\ 0{,}7505\ ;\ 0{,}4100)$. La moyenne a priori estimée de $1/\sigma_i^2$ qui est la moyenne de la loi Gamma dont les paramètres sont a et b, est $ab = 0{,}002295$ dont la racine carrée est 0,048 (notons que $1/0{,}048 = 20{,}85$, valeur en harmonie avec l'intervalle de variation des écarts-types d'échantillon s'étendant de 5,704 à 53,999). Les estimations sur petits domaines et leurs intervalles de confiance sont résumés au tableau 7 et à la figure 1. Les estimations ponctuelles produites par les quatre méthodes sont comparables : les mesures sommaires comprenant la moyenne, la médiane et l'étendue des estimations des paramètres de petit domaine pour les méthodes I, II, III et IV sont (121,9 ; 124,1 ; 122,2 ; 122,6), (125,2 ; 120,4 ; 115,0 ; 114,5) et (23,1 ; 53,0 ; 58,4 ; 56,6), respectivement. Les distributions de $\hat{\theta}_i$ (représentées graphiquement en prenant en considération tous les i) sont résumées à la figure 2 qui révèle une différence significative de variabilité. La méthode I est celle dont la variabilité est la plus faible et qui est donc la meilleure en ce sens. En outre, le lissage des variances d'échantillonnage a de fortes répercussions sur la mesure de l'incertitude et donc de l'estimation de l'intervalle. La méthode proposée donne l'intervalle de confiance le plus court, en moyenne, comparativement à toutes les autres méthodes. Les méthodes II et III donnent des intervalles dont la borne inférieure est négative, ce qui paraît irréaliste, car la moyenne directe des superficies consacrées à la culture du maïs est positive et grande pour les 12 comtés [les intervalles de confiance bruts $(x_i \pm t_{0{,}025}\, S_i)$ ne contiennent non plus de valeur nulle pour aucun des domaines]. Il n'existe aucun soutien théorique pour les intervalles de confiance de la méthode II. Les méthodes II et III produisent des intervalles de confiance plus larges quand la variance d'échantillonnage est élevée. Par exemple, la taille d'échantillon pour les comtés de Franklin et de Pocahontas est de trois, mais les écarts-types d'échantillon sont de 5,704 et 43,406, respectivement. Alors que les intervalles de confiance sont comparables sous la méthode I, ils sont très différents sous les méthodes II et III. Il en est ainsi parce que, même si ces méthodes tiennent compte de l'incertitude dans les estimations de la variance d'échantillonnage, comme le lissage n'a pas été effectué en utilisant l'information provenant des estimations directes d'après l'enquête, les estimations de la variance d'échantillonnage sous-jacentes demeurent très variables (à cause de la petite taille d'échantillon). En fait, la variance de l'estimateur de variance (des estimations ponctuelles) est plus grande que celle obtenue lorsque l'on applique la méthode I. Cela est aussi confirmé par le fait que les écarts-types intuitifs des estimations sur petits domaines « lissées » (un quart de l'intervalle) sont plus faibles et moins variables sous la méthode I que sous les autres méthodes. Une autre caractéristique de notre méthode qui mérite d'être soulignée est que les largeurs des intervalles sont comparables pour les comtés pour lesquels la taille d'échantillon est la même. Cela pourrait être une indication que l'on obtient des estimateurs équi-efficaces pour des tailles d'échantillon équivalentes.

Choix du modèle : Afin de choisir le modèle le mieux ajusté, nous avons utilisé le critère d'information bayésien (BIC pour *Bayesian Information Criteria*) qui tient compte à la fois de la vraisemblance et de la complexité des modèles ajustés. Nous avons calculé le BIC pour les modèles utilisés dans les méthodes I et III (Hwang et coll. 2009). Ces deux modèles comprennent le même nombre de paramètres et ne diffèrent que par la façon dont ces paramètres sont estimés. Le BIC du modèle est égal à 210,025 pour la méthode I et à 227,372 pour la méthode III, ce qui témoigne de la supériorité de notre méthode. Nous n'avons pas pu calculer le BIC pour le modèle de Wang et Fuller (2003), car ils n'ont utilisé aucune fonction de vraisemblance explicite.

Tableau 7
Résultats de l'analyse des données sur le maïs. Ici, IC et LIC représentent l'intervalle de confiance et la longueur de l'intervalle de confiance, respectivement

Comté	$\hat{\theta}_i$	IC	LIC	$\hat{\theta}_i$	IC	LIC
		I : méthode proposée			II : Wang et Fuller (2003)	
Franklin	131,8106	104,085 ; 159,372	55,287	155,4338	124,151 ; 193,094	68,943
Pocahontas	108,7305	80,900 ; 136,436	55,536	102,3682	-38,973 ; 244,019	282,993
Winnebago	109,0559	81,430 ; 136,646	55,216	115,9093	-53,768 ; 279,314	333,083
Wright	131,6113	103,736 ; 159,564	55,828	131,0674	8,330 ; 280,263	271,932
Webster	113,1484	92,805 ; 133,348	40,543	109,4795	32,514 ; 202,675	170,161
Hancock	129,4279	111,781 ; 147,193	35,412	124,1028	56,750 ; 162,013	105,262
Kossuth	121,0071	103,451 ; 138,626	35,175	116,7147	68,049 ; 152,454	84,405
Hardin	130,2520	112,373 ; 148,114	35,741	137,7983	51,734 ; 188,373	136,638
		III : Hwang et coll. (2009)			IV : Qiu et Hwang (2007)	
Franklin	158,4677	128,564 ; 188,370	59,805	157,7383	146,999 ; 168,477	21,478
Pocahontas	100,1276	-44,039 ; 244,295	288,334	101,1661	19,444 ; 182,887	163,442
Winnebago	114,1473	0,065 ; 228,228	228,163	113,7746	56,263 ; 171,286	115,022
Wright	140,3717	-24,119 ; 304,862	328,982	143,2244	41,559 ; 244,889	203,330
Webster	115,7865	50,297 ; 181,275	130,978	115,2224	75,124 ; 155,320	80,196
Hancock	111,3087	66,213 ; 156,403	90,189	113,1766	83,691 ; 142,661	58,970
Kossuth	110,9585	74,366 ; 147,550	73,184	112,3239	89,520 ; 135,127	45,607
Hardin	126,6093	40,040 ; 213,178	173,137	123,9049	54,607 ; 193,202	138,594

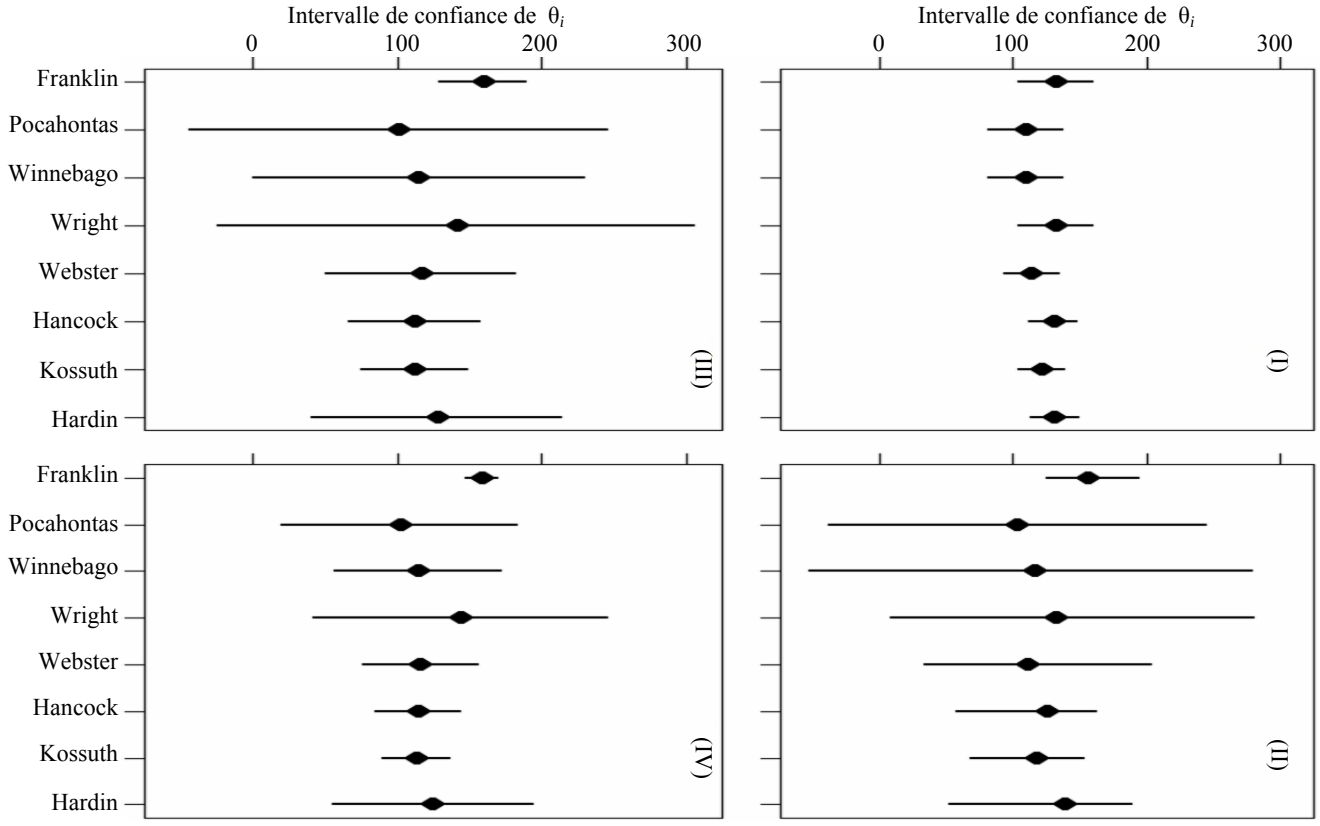

Figure 1 Estimation du nombre d'hectares consacrés au maïs. Pour chaque comté, la droite horizontale donne l'intervalle de confiance de $\hat{\theta}_i$, avec $\hat{\theta}_i$ marqué par le cercle, pour (I) la méthode proposée, (II) Wang et Fuller (2003), (III) Hwang et coll. (2009) et (IV) Qiu et Hwang (2007)

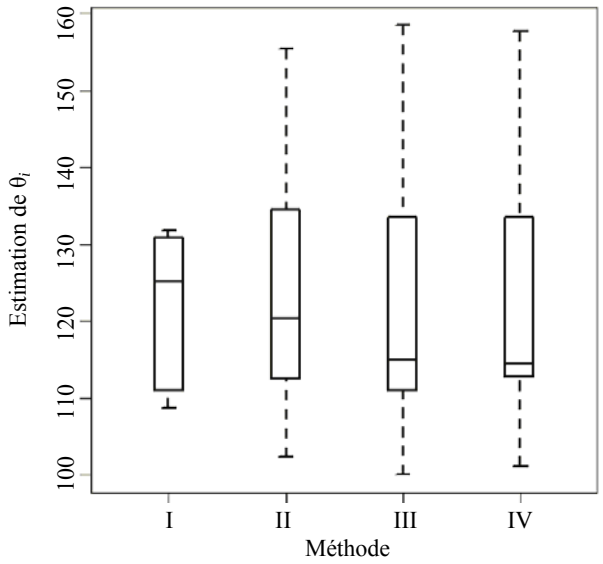

Figure 2 Boîtes à moustaches des estimations du nombre d'hectares consacrés au maïs p our chaque comté. (I) à (IV) sont les quatre méthodes correspondant à la figure I

7. Conclusion

Le présent article décrit la modélisation conjointe au niveau du domaine des moyennes et des variances pour l'estimation sur petits domaines. Il montre que les estimateurs sur petits domaines résultants sont plus efficaces que les estimateurs classiques obtenus en utilisant les modèles de Fay-Herriot qui ne rétrécissent que les moyennes. Bien que notre modèle soit le même que celui pris en considération dans Hwang et coll. (2009), notre méthode d'estimation diffère à deux égards, en ce qui concerne la détermination du paramètre de mise au point k et l'utilisation de $\pi(\sigma_i^2 \mid X_i, S_i^2, \boldsymbol{Z}_i)$ (qui dépend additionnellement de X_i), au lieu de $\pi(\sigma_i^2 \mid S_i^2, \boldsymbol{Z}_i)$ pour construire la distribution conditionnelle des paramètres θ_i de petit domaine. Nous avons démontré les propriétés de robustesse du modèle quand l'hypothèse que σ_i^2 est issue d'une loi Gamma inverse est violée. L'emprunt de l'information X_i pour estimer σ_i^2 ainsi que la robustesse à l'élicitation de la loi a priori démontre la supériorité de la méthode que nous proposons. Les valeurs des paramètres choisis dans l'étude par simulation diffèrent de celles utilisées dans l'analyse des données réelles. Cette dernière est présentée ici simplement en guise d'illustration. Notre objectif principal était d'élaborer la méthodologie de modélisation de la moyenne et de la variance, et de la comparer à certaines méthodes étroitement apparentée afin de montrer son efficacité. C'est pourquoi nous avons choisi de configurer les paramètres dans la simulation de la même façon que dans l'article traitant de l'estimation sur petits domaine bien connu de Wang et Fuller (2003).

L'obtention d'estimateurs améliorés de la variance d'échantillonnage est un produit secondaire de l'approche proposée. Nous avons fourni une technique d'estimation novatrice, qui est justifiée théoriquement et facile à utiliser. En ce qui concerne les calculs, la méthode est beaucoup plus simple que certaines méthodes concurrentes telles que les procédures MCMC bayésiennes ou les méthodes de ré-échantillonnage bootstrap. Notre méthode ne requiert qu'un seul échantillonnage à partir de la loi a posteriori durant l'estimation des paramètres du modèle, et les valeurs échantillonnées peuvent être utilisées par la suite à toute autre fin. Le logiciel peut être obtenu sur demande auprès des auteurs.

Remerciements

Les auteurs remercient deux examinateurs et le rédacteur associé de leurs commentaires constructifs qui leur ont permis d'améliorer considérablement l'article. L'étude a été financée en partie par les subventions SES 0961649, 0961618 et DMS 1106450 de la NSF.

Annexe

A. Obtention des distributions conditionnelles

Des équations (1) et (2) il découle que la distribution conjointe conditionnelle de $\{X_i, S_i^2, \theta_i, \sigma_i^2\}$, $\pi(X_i, S_i^2, \theta_i, \sigma_i^2 \mid a, b, \boldsymbol{\beta}, \tau^2)$ est

$$
\begin{aligned}
&\pi(X_i, S_i^2, \theta_i, \sigma_i^2 \mid \boldsymbol{Z}_i, \boldsymbol{B}) \\
&\quad = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{(X_i-\theta_i)^2}{2\sigma_i^2}\right\} \frac{1}{\Gamma\left(\frac{n_i-1}{2}\right) 2^{\frac{n_i-1}{2}}} \\
&\quad \times \left\{(n_i-1)\frac{S_i^2}{\sigma_i^2}\right\}^{\frac{n_i-1}{2}-1} \exp\left\{-\frac{(n_i-1)S_i^2}{2\sigma_i^2}\right\} \\
&\quad \times \left(\frac{n_i-1}{\sigma_i^2}\right) \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{-\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}\right\} \\
&\quad \times \frac{1}{\Gamma(a)b^a}\left(\frac{1}{\sigma_i^2}\right)^{a+1} \exp\left(-\frac{1}{b\sigma_i^2}\right) \\
&\quad \propto \exp\left[-\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2} - \left\{\frac{(X_i-\theta_i)^2}{2}\right.\right. \\
&\qquad \left.\left. + \frac{(n_i-1)S_i^2}{2} + \frac{1}{b}\right\}\frac{1}{\sigma_i^2}\right] \\
&\quad \times \left(\frac{1}{\sigma_i^2}\right)^{\frac{n_i}{2}+a+1} \left(\frac{1}{\tau^2}\right)^{\frac{1}{2}} \frac{1}{\Gamma(a)b^a}.
\end{aligned}
$$

Par conséquent, les distributions conditionnelles de σ_i^2 et θ_i sachant les données et $\boldsymbol{B}$ sont

$$\pi(\sigma_i^2 \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}) = \int \pi(X_i, S_i^2, \theta_i, \sigma_i^2 \mid \boldsymbol{Z}_i, \boldsymbol{B})\, d\theta_i \propto \frac{1}{(\sigma_i^2)^{(n_i-1)/2+a+1}(\sigma_i^2+\tau^2)^{1/2}} \exp\left[-\frac{(X_i - \boldsymbol{Z}_i^T, \boldsymbol{\beta})^2}{2(\sigma_i^2+\tau^2)} - \left\{\frac{1}{2}(n_i-1)S_i^2 + \frac{1}{b}\right\}\left(\frac{1}{\sigma_i^2}\right)\right],$$

$$\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{Z}_i, \boldsymbol{B}) = \int \pi(X_i, S_i^2, \theta_i, \sigma_i^2 \mid \boldsymbol{Z}_i, \boldsymbol{B})\, d\sigma_i^2 \propto \exp\left\{-\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}\right\}\psi_i^{-\left(\frac{n_i}{2}+a\right)}$$

où ψ_i est définie dans l'équation (4).

B. Détails de l'algorithme EM

La maximisation de $Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})$ est effectuée en posant que les dérivées partielles par rapport à $\boldsymbol{B}$ sont nulles, c'est-à-dire

$$\frac{\partial Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})}{\partial \boldsymbol{B}} = 0. \tag{B.1}$$

Partant de l'expression de $Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})$ dans le corps du texte, nous obtenons des expressions explicites pour les dérivées partielles par rapport à chaque composante de $\boldsymbol{B}$. La dérive partielle correspondant à β est

$$\frac{\partial Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})}{\partial \beta} = \sum_{i=1}^{n} \frac{\int \boldsymbol{Z}_i \left(\frac{\theta_i - Z_i^T\beta}{\tau^2}\right) \exp\left\{-\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}\right\}\psi_i^{-\left(\frac{n_i}{2}+a\right)} d\theta_i}{\int \exp\left\{-\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}\right\}\psi_i^{-\left(\frac{n_i}{2}+a\right)} d\theta_i} = \sum_{i=1}^{n} E\left\{\boldsymbol{Z}_i\left(\frac{\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}}{\tau^2}\right)\right\}$$

où l'espérance est calculée par rapport à la distribution conditionnelle de θ_i, $\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{B})$. L'expression de la dérivée partielle correspondant à τ^2 est :

$$\frac{\partial Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})}{\partial \tau^2} = -\frac{n}{2\tau^2} + \sum_{i=1}^{n} \frac{\int \frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2(\tau^2)^2} \exp\left\{-\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}\right\}\psi_i^{-\left(\frac{n_i}{2}+a\right)} d\theta_i}{\int \exp\left\{-\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}\right\}\psi_i^{-\left(\frac{n_i}{2}+a\right)} d\theta_i} = -\frac{n}{2\tau^2} + \sum_{i=1}^{n} E\left\{\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2(\tau^2)^2}\right\}.$$

De même, pour a et b, nous obtenons les solutions en posant que $S_a = 0$ et $S_b = 0$, où S_a et S_b sont, respectivement, les dérivées partielles de $Q(\boldsymbol{B} \mid \boldsymbol{B}^{(t-1)})$ par rapport à a et b en utilisant les expressions données dans le corps du texte. Ces équations sont résolues par la méthode de Newton-Raphson qui nécessite la matrice des dérivées secondes par rapport à a et b. Celles-ci sont données par les expressions suivantes :

$$S_{aa} = \sum_{i=1}^{n}\left[\log''\left\{\Gamma\left(\frac{n_i}{2}+a\right)\right\} - \log''\{\Gamma(a)\} + \mathrm{Var}\{\log(\psi_i)\}\right]$$

$$S_{ab} = \sum_{i=1}^{n}\left[-\frac{1}{b} + \frac{1}{b^2}E\left(\frac{1}{\psi_i}\right) - \left(\frac{n_i}{2}+a\right)\frac{1}{b^2} \mathrm{Cov}\left\{\frac{1}{\psi_i}, \log(\psi_i)\right\}\right], \tag{B.2}$$

et

$$S_{bb} = \sum_{i=1}^{n}\left\{\frac{a}{b^2} - (n_i+2a)\frac{1}{b^3}E\left(\frac{1}{\psi_i}\right) + \left(\frac{n_i}{2}+a\right)\frac{1}{b^4} E\left(\frac{1}{\psi_i^2}\right) + \left(\frac{n_i}{2}+a\right)^2\frac{1}{b^4}\mathrm{Var}\left(\frac{1}{\psi_i}\right)\right\}$$

avec $S_{ba} = S_{ab}$. À la u^{e} étape, les mises à jour de a et b sont données par

$$\begin{bmatrix} a^{(u)} \\ b^{(u)} \end{bmatrix} = \begin{bmatrix} a^{(u-1)} \\ b^{(u-1)} \end{bmatrix} - \begin{bmatrix} S_{aa}^{(u-1)} & S_{ab}^{(u-1)} \\ S_{ba}^{(u-1)} & S_{bb}^{(u-1)} \end{bmatrix}^{-1} \begin{bmatrix} S_a^{(u-1)} \\ S_b^{(u-1)} \end{bmatrix}, \tag{B.3}$$

où l'indice supérieur $(u-1)$ sur S_{aa}, S_{ab}, S_{ba}, S_{bb}, S_a et S_b désigne ces quantités évaluées aux valeurs qu'avaient a et b à la $(u-1)^{\mathrm{e}}$ itération. Lorsque la procédure de Newton-Raphson converge, les valeurs de a et b à la t^{e} étape de l'algorithme EM sont fixées à $a^{(t)} = a^{(\infty)}$ et $b^{(t)} = b^{(\infty)}$.

C. Une autre formulation du modèle d'estimation sur petits domaines

Il est possible de réduire la largeur de l'intervalle de confiance $\tilde{C}(\boldsymbol{B})$ en se fondant pour l'estimation sur petits domaines sur un autre modèle hiérarchique qui présente une certaine élégance mathématique. Dans (19), le terme constant $n_i + 2a + 2$ devient $n_i + 2a$ dans cette autre formulation du modèle. Le modèle est donné par

$$X_i \mid \theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2), \tag{C.1}$$

$$\theta_i \mid \sigma_i^2 \sim N(\boldsymbol{Z}_i\boldsymbol{\beta}, \lambda\sigma_i^2), \tag{C.2}$$

$$\frac{(n_i-1)S_i^2}{\sigma_i^2}\,\Big|\,\sigma_i^2 \sim \chi^2_{n_i-1}, \qquad \text{(C.3)}$$

$$\sigma_i^2 \sim \text{Inverse} - \text{Gamma}(a,b), \qquad \text{(C.4)}$$

indépendamment pour $i = 1, 2, ..., n$. Notons que, dans la formule susmentionnée, il est supposé que la variance conditionnelle de θ_i est proportionnelle à σ_i^2, tandis que la variance marginale est constante (en éliminant σ_i^2 par intégration en utilisant (C.4)). Dans (1) et (2), la variance de θ_i est une constante, τ^2, indépendante de σ_i^2, et il n'existe pour θ_i aucune structure conditionnelle dépendant de σ_i^2. L'ensemble de tous les paramètres inconnus dans le modèle hiérarchique courant est $\boldsymbol{B} = (a, b, \boldsymbol{\beta}, \lambda)$. La procédure d'inférence pour ce modèle est donnée ci-après. Le modèle repose essentiellement sur l'hypothèse que les effets réels de petit domaine ne sont pas identiquement distribués, même après avoir éliminé les variations connues.

C.1 Méthodologie d'inférence

En reparamétrisant la variance comme dans (C.2), on obtient certaines simplifications analytiques pour dériver les lois a posteriori de θ_i et σ_i sachant X_i, S_i^2 et $\boldsymbol{B}$. Nous avons

$$\pi(\sigma_i^2 \mid X_i, S_i^2, \boldsymbol{B}) = GI\left(\frac{n_i}{2}+a, \left[\frac{(n_i-1)S_i^2}{2}+\frac{(X_i-\boldsymbol{Z}_i\boldsymbol{\beta})^2}{2(1+\lambda)}+\frac{1}{b}\right]^{-1}\right)$$

où $GI(a,b)$ représente la loi Gamma inverse dont les paramètres de forme et d'échelle sont a et b, respectivement. Sachant $\boldsymbol{B}$ et σ_i^2, la distribution conditionnelle de θ_i est

$$\pi(\theta_i \mid X_i, \sigma_i^2, \boldsymbol{B}) = \text{Normale}\left(\boldsymbol{Z}_i^T\boldsymbol{\beta}, \frac{\lambda\sigma_i^2}{1+\lambda}\right).$$

En éliminant σ_i^2 par intégration, on obtient la distribution conditionnelle de θ_i sachant X_i, S_i^2 et $\boldsymbol{B}$,

$$\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{B}) = \int_0^\infty \pi(\theta_i \mid X_i, \sigma_i^2, B)\,\pi(\sigma_i^2 \mid X_i, S_i^2, \boldsymbol{B})\,d\sigma_i^2$$

$$\propto \left\{\frac{(1+\lambda)}{2\lambda}(\theta_i-\boldsymbol{Z}_i^T\boldsymbol{\beta})^2+\frac{\delta^2}{2}\right\}^{-(n_i+2a+1)/2}, \qquad \text{(C.5)}$$

où $\delta^2 = (n_i-1)\,S_i^2 + (X_i-\boldsymbol{Z}_i\boldsymbol{\beta})^2/(1+\lambda) + 2/b$. Nous pouvons réécrire (C.5) sous la forme

$$\pi(\theta_i \mid X_i, S_i^2, \boldsymbol{B}) = \frac{\Gamma((n_i+1)/2+a)\sqrt{1+\lambda}}{\delta^*\Gamma(n_i/2+a)\sqrt{(n_i+2a)\lambda\pi}} \left\{1+\frac{(\theta_i-\mu_i)^2}{(n_i+2a)\delta^{*2}\lambda/(1+\lambda)}\right\}^{-(n_i+2a+1)/2}$$

qui peut être considéré comme une distribution t à échelle possédant n_i+2a degrés de liberté et le paramètre d'échelle $\delta^*\sqrt{\lambda/(1+\lambda)}$ avec $\delta^{*2} = \delta^2/(n_i+2a)$. D'où,

$$E(\sigma_i^{-1} \mid X_i, S_i, \boldsymbol{B}) = \frac{\Gamma((n_i+1)/2+a)(\delta^2/2)^{-\{(n_i+1)/2+a\}}}{\Gamma(n_i/2+a)(\delta^2/2)^{-(n_i/2+a)}} = \frac{\Gamma((n_i+1)/2+a)}{\Gamma(n_i/2+a)}\,\frac{\sqrt{2}}{\delta^*\sqrt{n_i+2a}}.$$

Dans ce contexte, en choisissant

$$k = k(\boldsymbol{B}) = \left\{1+\frac{t_{\alpha/2}^2}{n_i-1}\right\}^{-(n_i+2a+1)/2}\sqrt{\frac{1+\lambda}{\lambda}}\,\frac{1}{\sqrt{2\pi}},$$

l'intervalle de confiance donné par (8) se simplifie en

$$C_i(\boldsymbol{B}) \equiv \left\{\theta_i : \frac{|\,\theta_i-\mu_i\,|}{\sqrt{\frac{\lambda}{1+\lambda}\,\frac{(n_i+2a)\delta^{*2}}{n_i-1}}} \le t_{\alpha/2}\right\}. \qquad \text{(C.6)}$$

En utilisant les mêmes arguments qu'auparavant et en notant que $(n_i+2a)\delta^{*2} \ge (n_i-1)S_i^2$, nous avons $P\{C_i(\boldsymbol{B})\} \ge P(D_i) = 1-\alpha$, où D_i est l'intervalle de confiance donné par (20). Quand $\boldsymbol{B}$ est inconnu, nous le remplaçons par l'estimation de son maximum de vraisemblance marginale $\hat{\boldsymbol{B}}$. Nous nous attendons à ce que la technique de groupement donne une erreur suffisamment petite pour que $P\{C_i(\hat{\boldsymbol{B}})\} \approx P\{C_i(\boldsymbol{B})\} \ge 1-\alpha$.

Bibliographie

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satelite data. *Journal of the American Statistical Association*, 95, 28-36.

Bell, W. (2008). Examining sensitivity of small area inferences to uncertainty about sampling error variances. Rapport technique du U.S. Census Bureau.

Casella, G., et Hwang, J. (1991). Evaluating confidence sets using loss functions. *Statistica Sinica*, 1, 159-173.

Chatterjee, S., Lahiri, P. et Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Annals of Statistics*, 36, 1221-1245.

Cho, M., Eltinge, J., Gershunskaya, J. et Huff, L. (2002). Evaluation of generalized variance function estimators for the U.S. current employment survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 534-539.

Fay, R., et Herriot, R. (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.

Gershunskaya, J., et Lahiri, P. (2005). Variance estimation for domains in the U.S. current employment statistics program. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3044-3051.

Ghosh, M., et Rao, J. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 54-76.

Hall, P., et Maiti, T. (2006). Nonparametric estimation of mean squared prediction error in nested-error regression models. *Annals of Statistics*, 34, 1733-1750.

Huff, L., Eltinge, J. et Gershunskaya, J. (2002). Exploratory analysis of generalized variance function models for the U.S. current employment survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1519-1524.

Hwang, J., Qiu, J. et Zhao, Z. (2009). Empirical Bayes confidence intervals shrinking both mean and variances. *Journal of the Royal Statistical Society*, B, 71, 265-285.

Joshi, V. (1969). Admissibility of the usual confidence sets for the mean of a univariate or bivariate normal population. *The Annals of Mathematical Statistics*, 40, 1042-1067.

Maples, J., Bell, W. et Huang, E. (2009). Small area variance modeling with application to county poverty estimates from the american community survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 5056-5067.

Otto, M., et Bell, W. (1995). Sampling error modelling of poverty and income statistics for states. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 160-165.

Pfeffermann, D. (2002). Small area estimation - New developments and directions. *Revue Internationale de Statistique*, 70, 125-143.

Prasad, N., et Rao, J. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Qiu, J., et Hwang, J. (2007). Sharp simultaneous intervals for the means of selected populations with application to microarray data analysis. *Biometrics*, 63, 767-776.

Rao, J. (2003). Some new developments in small area estimation. *Journal of the Iranian Statistical Society*, 2, 145-169.

Rivest, L.-P., et Vandal, N. (2003). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*.

Robert, C., et Casella, G. (2004). *Monte Carlo Statistical Methods* (Deuxième édition).

Valliant, R. (1987). Generalized variance functions in stratified two-stage sampling. *Journal of the American Statistical Association*, 82, 499-508.

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

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, 1, 107-114.