N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

QUESTIONS PRATIQUES D'ESTIMATION RÉGIONALE PAR MODÈLE

J.N.K. Rao[1]

RÉSUMÉ

Les estimateurs directs habituels par région pourraient ne pas être d'une précision acceptable dans le cas de petites régions, car les tailles d'échantillon sont rarement assez grandes, d'où la nécessité de «renforcer » une estimation indirecte interrégionale par des modèles d'appariement. Les modèles par région et par unité ont amplement été étudiés par les auteurs spécialisés en vue de l'obtention d'estimateurs optimaux de prévision linéaire empirique sans biais (estimation MPLSE) et de Bayes empiriques (estimation BE) ou hiérarchisés (estimation BH) pour de petites régions avec les mesures liées de variabilité. Dans ce document, nous aborderons plusieurs questions pratiques d'estimation basée sur le modèle pour de petites régions, et notamment les questions de détermination du modèle, de choix de valeurs antérieures pour les paramètres de modélisation dans l'estimation BH, d'étalonnage en fonction d'estimateurs directs sûrs de grandes régions et de pondération d'enquête dans une estimation par modèle.

MOTS CLÉS : Estimation composée, modèle par région, modèle par unité.

1. INTRODUCTION

À cause du facteur coût et pour d'autres considérations, les enquêtes par sondage sont habituellement conçues pour donner des estimateurs par région (ou directs) comportant un petit coefficient de variation (CV) échantillonnal pour de grandes régions (ou domaines). En fait, les praticiens des enquêtes soulignent fréquemment que les erreurs non dues à l'échantillonnage, en particulier les erreurs de mesure et d'observation, contribuent bien plus que les erreurs d'échantillonnage à l'erreur quadratique moyenne (EQM) totale, laquelle sert souvent de mesure de qualité des estimateurs. Il reste que les erreurs d'échantillonnage jouent un rôle de premier plan dans l'estimation par petite région, puisque les tailles d'échantillon de petites régions sont rarement assez grandes pour que les estimateurs directs soient d'une qualité acceptable pour l'EQM (ou le CV) en échantillonnage. En réalité, la taille d'échantillon peut être nulle dans un grand nombre de petites régions d'intérêt. Ainsi, on se reporte aux données de la Current Population Survey (CPS) pour l'estimation par comté (ou par district scolaire) du nombre d'enfants pauvres d'âge scolaire aux États-Unis, mais les tailles d'échantillon CPS sont nulles dans beaucoup de comtés visés (National Research Council, 2000).

En raison des difficultés liées aux estimateurs directs, il est nécessaire dans bien des cas de recourir à des estimations indirectes qui « empruntent » de l'information à des régions liées grâce à des modèles explicites (ou implicites) d'appariement et à l'aide de données censitaires et administratives relatives aux petites régions. Ces dernières années, on s'est beaucoup intéressé aux estimateurs indirects par modèles explicites d'appariement, ceux-ci offrant les avantages suivants par rapport aux estimateurs indirects habituels par modèles implicites : (i) les méthodes de modélisation explicite tiennent expressément compte de la variation locale par des structures complexes d'erreur dans le modèle d'appariement de petites régions; (ii) il est possible de valider les modèles au moyen des données de l'échantillon; (iii) les méthodes peuvent traiter des cas complexes, qu'il s'agisse de données transversales ou longitudinales, binaires ou de dénombrement, en corrélation spatiale ou à variables multiples; (iv) on peut obtenir des mesures par région de variabilité des estimations à l'opposé des mesures globales couramment employées pour les estimateurs indirects habituels.

[1] Université Carleton, École de mathématiques et de statistique, Ottawa, Canada (courrier électronique : jrao@math.carleton.ca).

Les auteurs spécialisés ont amplement étudié les modèles par région et par unité en vue de l'obtention, en estimations de totaux (ou de moyennes) et les mesures liées de variabilité, d'estimateurs optimaux de prévision linéaire empirique sans biais (estimation MPLSE) et de Bayes empiriques (estimation BE) ou hiérarchisés (estimation BH). La méthode MPLSE est applicable à des modèles mixtes linéaires par région et par unité. Les méthodes BE et BH sont d'une application plus générale et s'étendent aux modèles mixtes linéaires généralisés qui visent des données catégoriques (binaires, par exemple) et de dénombrement. L'EQM sert de mesure de variabilité dans les méthodes MPLSE et BE; dans la méthode BH, c'est la variance postérieure qui est la mesure de variabilité, une distribution antérieure étant posée pour les paramètres de modélisation. Nous renvoyons le lecteur à Rao (2003) pour un traitement détaillé des méthodes MPLSE, BE et BH d'estimation régionale. Dans le présent document, j'aborderai plusieurs questions pratiques d'estimation régionale par modèle : détermination du modèle, méthodes d'estimation de l'EQM, choix de valeurs antérieures pour les paramètres de modélisation dans une estimation BH, étalonnage en fonction d'estimateurs directs sûrs de grandes régions, emploi de la pondération d'enquête dans l'estimation par modèle, etc.

2. MODÈLES RELATIFS À DE PETITES RÉGIONS

Les auteurs spécialisés ont étudié deux types de modèles de base pour les petites régions. Dans la première catégorie, dite modélisation de base par région, on se reporte uniquement à des données auxiliaires par région $z_i = (z_{1i},..., z_{pi})^T$, d' une certaine fonction appropriée $\theta_i = g(Y_i)$ de totaux de petites régions $Y_i (i = 1,..., m)$, pour élaborer un modèle d'appariement sous la forme $\theta_i = z_i^T \beta + v_i$ avec $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$, où σ_v^2 est la variance du modèle. On combine le modèle d'appariement et le modèle d'échantillonnage correspondant $\hat{\theta}_i = \theta_i + e_i$, où $\hat{\theta}_i = g(\hat{Y}_i)$ est un estimateur direct de θ_i et $e_i \mid \theta_i \overset{ind}{\sim} N(0, \psi_i)$ qui fait intervenir une variance d'échantillonnage connue ψ_i. Le modèle composé $\hat{\theta}_i = z_i^T \beta + v_i + e_i$ est un cas d'espèce du modèle mixte linéaire.

La modélisation de base par région comporte au moins deux limites. D'abord, l'hypothèse des variances d'échantillonnage connues ψ_i est restrictive, bien qu'on ait proposé des méthodes fondées sur des fonctions généralisées de variance (FGV) pour l'obtention d'estimations lissées de ψ_i. En second lieu, l'hypothèse $E(e_i \mid \theta_i) = 0$ peut ne pas être applicable si la taille d'échantillon de petite région n_i est petite et que θ_i est une fonction non linéaire du total Y_i, même là où l'estimateur direct $\hat{Y}_i$ est sans biais de plan d'échantillonnage pour Y_i. Il est plus réaliste d'utiliser le modèle d'échantillonnage $\hat{Y}_i = Y_i + f_i$ avec $E(f_i \mid Y_i) = 0$, ce qui nous dit simplement que $\hat{Y}$ est sans biais de plan d'échantillonnage pour Y_i. Posons en outre que $V(f_i \mid Y_i) = \sigma_i^2$, où la variance d'échantillonnage peut dépendre de Y_i. Ainsi, $\sigma_i^2 = Y_i^2 c_i^2$, où c_i est le coefficient connu de variation de $\hat{Y}_i$ vérifié par ajustement de modèle FGV. Le modèle d'échantillonnage est maintenant sans correspondance avec le modèle d'appariement, en ce sens que les deux ne peuvent être directement combinés en un modèle mixte linéaire. On a proposé diverses extensions de la modélisation de base par région (modélisation Fay-Herriot) pour le traitement des erreurs d'échantillonnage corrélées, de la dépendance spatiale des erreurs de modélisation v_i et des données longitudinales et transversales (voir Rao, 2003, chapitre 8).

Dans la seconde catégorie, dite modélisation de base par unité, on relie des variables auxiliaires par unité $x_{ij} = (x_{1ij},..., x_{pij})^T$ aux valeurs y d'unités y_{ij} par un modèle hiérarchisé de régression linéaire d'erreur $y_{ij} = x_{ij}^T \beta + v_i + e_{ij}$, où $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ et est indépendant de $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$. On a proposé diverses extensions de la modélisation de base par unité pour le traitement des réponses binaires, de l'échantillonnage intrarégional à deux degrés, des réponses à plusieurs variables et d'autres éléments (voir Rao, 2003, chapitres 8, 9 et 10). Ainsi, pour des réponses binaires y_{ij}, on peut poser que $y_{ij} \overset{ind}{\sim}$ Bernoulli (p_{ij}) et que les p_{ij} sont raccordés par recours à un

modèle de régression logistique $\log\{p_{ij}/(1-p_{ij})\} = x_{ij}^T\beta + v_i$, où $v_i \overset{iid}{\sim} N(0,\sigma^2)$. C'est un cas d'espèce des modèles mixtes linéaires généralisés.

3. MODÉLISATION DE BASE BE PAR RÉGION

3.1 Estimation de θ_i

Dans une modélisation de base par région, le meilleur estimateur de θ_i en ce qui concerne l'EQM minimale nous est donné par $E(\theta_i \mid \hat{\theta}_i, \beta, \sigma_v^2)$, qui dépend des paramètres de modélisation β et σ_v^2. Si nous remplaçons (β, σ_v^2) par des estimateurs appropriés $(\hat{\beta}, \sigma_v^2)$ tirés de la distribution marginale de $\hat{\theta}_i s$, en l'occurrence $\hat{\theta}_i \overset{ind}{\sim} N(z_i^T\beta, \sigma_v^2 + \psi_i)$, nous obtenons l'estimateur de Bayes empirique ou le meilleur estimateur empirique BE

$$\hat{\theta}_i^{EB} = \hat{\gamma}_i\hat{\theta}_i + (1-\hat{\gamma}_i)z_i^T\hat{\beta}, \tag{3.1}$$

où $\hat{\gamma}_i = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \psi_i)$. La formule (3.1) indique que l'estimateur BE de θ_i est une moyenne pondérée de l'estimateur direct $\hat{\theta}_i$ et de l'estimateur synthétique de régression $z_i^T\hat{\beta}$ où la pondération est respectivement $\hat{\gamma}_i$ et $1-\gamma_i$. La pondération $\hat{\gamma}_i$ est une mesure de la variabilité interrégionale par rapport à la variabilité totale liée à la région *i*. L'estimateur $\hat{\theta}_i^{EB}$ est sans biais pour θ_i dans le modèle composé, c'est-à-dire $E(\hat{\theta}_i^{EB} - \theta_i) = 0$, et par rapport de plan d'échantillonnage à mesure que la variance d'échantillonnage ψ_i va vers zéro, à condition que l'estimateur direct soit lui aussi par rapport de plan d'échantillonnage. Toutefois, l'estimateur résultant $g^{-1}(\hat{\theta}_i^{EB})$ de Y_i est biaisé si $g(\cdot)$ est non linéaire. Il convient de noter que $g^{-1}(\hat{\theta}_i^{EB})$ n'est pas égal à l'estimateur BE $\hat{Y}_i^{EB}$ obtenu par l'évaluation de $E[g^{-1}(\theta_i) \mid \hat{\theta}_i, \beta, \sigma_v^2]$ à $\hat{\beta}$ et $\hat{\sigma}_v^2$.

Sous l'hypothèse de normalité, la méthode du maximum de vraisemblance (MV) ou du maximum de vraisemblance en valeur résiduelle (MVR) peut servir à l'estimation de β et de σ_v^2 à partir de la distribution marginale $\hat{\theta}_i \overset{iid}{\sim} N(z_i^T\beta, \sigma_v^2 + \psi_i)$. Autre possibilité : σ_v^2 peut s'estimer par une méthode simple des moments (Prasad et Rao, 1990) ou par une solution itérative de l'équation suivante des moments pour ce même élément (Fay et Herriot, 1979) :

$$a(\sigma_v^2) = \sum_{i=1}^{m} (\hat{\theta}_i - z_i^T\tilde{\beta}(\sigma_v^2))^2/(\sigma_v^2 + \psi_i) = m - p, \tag{3.2}$$

où $\tilde{\beta}(\sigma_v^2)$ est l'estimateur des moindres carrés pondérés de β pour un σ_v^2 donné. Les estimateurs résultants $\hat{\sigma}_v^2$ et $\hat{\beta} = \tilde{\beta}(\hat{\sigma}_v^2)$ mènent à l'estimateur MPLSE de θ_i en (3.1). L'estimateur en question ne dépend pas d'une hypothèse de normalité.

3.2 Estimation de l'EQM

Les auteurs spécialisés ont étudié à fond les méthodes d'estimation $EQM(\hat{\theta}_i^{EB})$ qui tiennent compte de la variabilité de $\hat{\beta}$ et $\hat{\sigma}_v^2$. Dans ce cas, $EQM(\hat{\theta}_i^{EB}) = E(\hat{\theta}_i^{EB} - \theta_i)^2$ et l'espérance est celle du modèle composé

(voir Rao, 2003, chapitre 7). Une approximation fidèle de $EQM\left(\hat{\theta}_i^{EB}\right)$ sous une hypothèse de normalité nous est donnée par

$$EQM\left(\hat{\theta}_i^{EB}\right) \approx g_{1i}\left(\sigma_v^2\right) + g_{2i}\left(\sigma_v^2\right) + g_{3i}\left(\sigma_v^2\right), \tag{3.3}$$

où le terme principal $g_{1i}\left(\sigma_v^2\right) = \gamma_i \psi_i$ avec $\gamma_i = \sigma_v^2 / \left(\sigma_v^2 + \psi_i\right)$ est la contribution à l'EQM si β et σ_v^2 sont censés être connus, où

$$g_{2i}\left(\sigma_v^2\right) = \left(1-\gamma_i\right)^2 z_i^T \left[\sum_{i=1}^{m} z_i z_i^T / \left(\sigma_v^2 + \psi_i\right)\right]^{-1} z_i \tag{3.4}$$

rend compte de la variabilité de $\hat{\beta}$ et où le terme

$$g_{3i}\left(\sigma_v^2\right) = \left[\psi_i^2 / \left(\sigma_v^2 + \psi_i\right)^4\right] E\left(\hat{\theta}_i - z_i^T \beta\right)^2 h\left(\sigma_v^2\right) \tag{3.5}$$

$$= \left[\psi_i^2 / \left(\sigma_v^2 + \psi_i\right)^3\right] h\left(\sigma_v^2\right) \tag{3.6}$$

rend compte de la variabilité de $\hat{\sigma}_v^2$, $h\left(\sigma_v^2\right)$ étant la variance asymptotique de $\hat{\sigma}_v^2$ pour des valeurs *m* élevées. Les termes négligés dans l'approximation (3.3) sont d'un ordre inférieur à m^{-1} et l'approximation en question vaut pour les méthodes PR (Prasad-Rao), FH (Fay-Herriot), MV et MVR d'estimation de σ_v^2.

Si nous comparons le terme principal $\gamma_i \psi_i$ de (3.3) à ψ_i, l'EQM de l'estimateur direct de $\hat{\theta}_i$, il ressort que l'estimateur BE $\hat{\theta}_i^{EB}$ fait faire un important gain d'efficience là où γ_i est petit, c'est-à-dire là où σ_v^2, la variabilité des erreurs de modélisation v_i, est petite par rapport à la variabilité totale $\sigma_v^2 + \psi_i$. À noter que ψ_i est aussi la variance de $\hat{\theta}_i$ en plan d'échantillonnage.

Dans l'estimation de l'EQM, un estimateur convenant à l'approximation du même ordre qu'en (3.3) est donné par

$$eqm\left(\hat{\theta}_i^{EB}\right) \approx g_{1i}\left(\hat{\sigma}_v^2\right) + g_{2i}\left(\hat{\sigma}_v^2\right) + 2\, g_{3i}\left(\hat{\sigma}_v^2\right). \tag{3.7}$$

L'estimateur (3.7) est approximativement sans biais pour $EQM\left(\hat{\theta}_i^{EB}\right)$, en ce sens que son biais est d'un ordre inférieur à m^{-1}, à condition que $\hat{\sigma}_v^2$ soit fondé sur la méthode MVR ou PR. Pour les méthodes MV et FH d'estimation de σ_v^2, nous ajoutons un terme $g_{0i}\left(\hat{\sigma}_v^2\right)$ à (3.8). Pour la méthode MV, ce terme supplémentaire est positif (Datta et Lahiri, 2000). Si nous négligeons ce terme et utilisons (3.8) avec l'estimateur MV de $\hat{\sigma}_v^2$, il y aura donc sous-estimation de l'EQM. En revanche, le terme supplémentaire est négatif dans la méthode FH (Datta, Rao et Smith, 2002) et, par conséquent, nous nous trouverons à surestimer l'EQM si nous oublions ce terme et utilisons (3.8) avec l'estimateur FH de $\hat{\sigma}_v^2$.

Lahiri et Rao (1995) ont démontré que, dans l'estimation de l'EQM en (3.8), l'estimateur PR de σ_v^2 est robuste par rapport à l'absence d'hypothèse de normalité des effets aléatoires e_i, en ce sens que l'hypothèse d'absence approximative de biais demeure valable à condition que l'hypothèse de normalité des erreurs d'échantillonnage se vérifie. Cette dernière hypothèse est moins restrictive que celle de la normalité des v_i en raison de l'incidence du

théorème central limite sur les estimateurs directs de $\hat{\theta}_i$. Nous ignorons si la propriété de robustesse vaut aussi pour les méthodes MVR, MV et FH.

On peut reprocher à l'estimateur de l'EQM (3.8) et à sa modification pour les méthodes MV et FH de ne pas être « par région », c'est-à-dire de ne pas dépendre expressément de $\hat{\theta}_i$, bien que les données auxiliaires par région z_i se retrouvent dans le terme $g_{2i}\left(\hat{\sigma}_v^2\right)$. Rao (2000) s'est reporté à l'expression (3.5) pour $g_{3i}\left(\sigma_v^2\right)$ en vue d'obtenir un autre estimateur par région de ce dernier élément :

$$\tilde{g}_{3i}\left(\hat{\sigma}_v^2, \hat{\theta}_i\right) = \left[\psi_i^2 / \left(\sigma_v^2 + \psi_i\right)^4\right]\left(\hat{\theta}_i - z_i^T \hat{\beta}\right)^2 h\left(\hat{\sigma}_v^2\right). \tag{3.8}$$

Par (3.8), nous avons deux estimateurs différents par région de l'EQM pour les méthodes MVR et PR :

$$eqm_1\left(\hat{\theta}_i^{EB}\right) \approx g_{1i}\left(\hat{\theta}_v^2\right) + g_{2i}\left(\hat{\sigma}_v^2\right) + g_{3i}\left(\hat{\sigma}_v^2\right) + \tilde{g}_{3i}\left(\hat{\sigma}_v^2, \hat{\theta}_i\right) \tag{3.9}$$

et

$$eqm_2\left(\hat{\theta}_i^{EB}\right) \approx g_{1i}\left(\hat{\theta}_v^2\right) + g_{2i}\left(\hat{\sigma}_v^2\right) + 2\tilde{g}_{3i}\left(\hat{\sigma}_v^2, \hat{\theta}_i\right). \tag{3.10}$$

Le terme $\tilde{g}_{3i}\left(\hat{\sigma}_v^2, \hat{\theta}_i\right)$ est moins stable que $g_{3i}\left(\hat{\sigma}_v^2\right)$, mais il est d'un ordre inférieur à celui du terme principal $g_{1i}\left(\hat{\sigma}_v^2\right)$ dans (3.9) et (3.10). Ainsi, le coefficient de variation (CV) de $eqm_1\left(\hat{\theta}_i^{EB}\right)$ devrait être comparable au CV de $eqm\left(\hat{\theta}_i^{EB}\right)$, du moins pour des valeurs m qui varient de moyennes à grandes. Fuller (1989) a estimé l'EQM conditionnelle de $\hat{\theta}_i^{EB}$ pour le i^e estimateur direct donné par région de $\hat{\theta}_i$. Son estimateur EQM par région est étroitement lié à l'estimateur EQM inconditionnel (3.9). Butar et Lahiri (1997) ont obtenu un estimateur EQM par région en corrigeant le biais de l'estimateur de variabilité de Laird et Louis (1987) par la méthode paramétrique bootstrap. Cet estimateur EQM en correction de biais est identique à (3.9), qui est tiré d'une manière simple de la formule (3.5) pour $g_{3i}\left(\sigma_v^2\right)$.

Il plaira davantage aux praticiens des enquêtes de considérer l'estimation EQM en échantillonnage de $\hat{\theta}_i^{EB}$, c'est-à-dire $EQM_p\left(\hat{\theta}_i^{EB}\right) = E_p\left(\hat{\theta}_i^{EB} - \theta_i\right)^2$, où l'espérance E_p vise le plan d'échantillonnage $p(\cdot)$, plus précisément la distribution des erreurs d'échantillonnage pour des $\hat{\theta}_i$ donnés. Rivest et Belmonte (2000) ont obtenu un estimateur sans biais de plan d'échantillonnage de $EQM\left(\hat{\theta}_i^{EB}\right)$ par l'estimateur PR de σ_v^2. Le terme principal de cet estimateur EQM est par région, c'est-à-dire dépend de $\hat{\theta}_i$ contrairement au terme principal $g_{1i}\left(\hat{\sigma}_v^2\right)$ de l'estimateur EQM de modélisation $eqm\left(\hat{\theta}_i^{EB}\right)$. Il demeure toutefois hautement instable par rapport à $eqm\left(\hat{\theta}_i^{EB}\right)$ sauf si l'estimateur direct de $\hat{\theta}_i$ reçoit un plus grand poids, $1 - \hat{\gamma}_i$ étant petit dans ce cas.

3.3 Variances d'échantillonnage inconnues ψ_i

Aux sections 4.1 et 4.2, nous avons posé que les variances d'échantillonnage ψ_i sont connues, mais c'est là une hypothèse restrictive. Wang et Fuller (2003) et Rivest et Vandal (2002) ont étudié l'effet de l'estimation de ψ_i sur l'EQM de l'estimateur BE en (4.1) là où $\hat{\gamma}_i$ est remplacé par $\hat{\sigma}_v^2 / \left(\hat{\sigma}_v^2 + \hat{\psi}_i\right)$, $\hat{\psi}_i$ étant un estimateur de ψ_i. Si nous pouvons poser, par exemple, un échantillon aléatoire $y_{ij} \overset{iid}{\sim} N\left(\theta_i, \sigma^2\right)$, $j = 1, \ldots n_i (\geq 2)$ pour la i^e région et $\hat{\theta}_i = \bar{y}_i$, moyenne de l'échantillon. Dans ce cas, $\hat{\psi}_i = s_i^2 / n_i$ est sans biais de plan d'échantillonnage pour ψ_i, s_i^2 étant la

variance de l'échantillon. De plus, $\bar{y}_i$ et $\hat{\psi}_i$ sont distribués indépendamment avec $\hat{\psi}_i \approx N\left[\psi_i, \delta_i = 2\psi_i^2/(n_i - 1)\right]$. Dans ce cadre, Rivest et Vandal (2002) ont obtenu un estimateur approprié de l'EQM en ajoutant le terme $2\hat{\delta}_i\hat{\sigma}_v^4/\left(\hat{\psi}_i + \hat{\sigma}_v^2\right)^3$ à (3.8) pour tenir compte de l'estimation de ψ_i, où $\hat{\delta}_i = 2\hat{\psi}_i^2/(n_i - 1)$. Si les tailles d'échantillon n_i sont petites, l'application de (3.8) peut donner lieu à une large sous-estimation de l'EQM contrairement à l'estimateur de Rivest-Vandal. Si $\hat{\psi}_i$ est un estimateur lissé de ψ_i par ajustement de modèle FGV, la contribution du terme supplémentaire est du même ordre, $O\left(m^{-1}\right)$, que celle du terme g_{3i}.

4. ESTIMATION JACKKNIFE DE L'EQM

Jiang, Lahiri et Wan (2002) ont proposé une méthode jackknife d'estimation de l'EQM d'estimateurs BE qui est applicable à des modèles mixtes linéaires généralisés à structures de covariance diagonales en blocs, les blocs en question correspondant à de petites régions. Cette méthode donne aussi des estimateurs approximativement sans biais de l'EQM des estimateurs BE. Ainsi, on peut considérer le cas de réponses binaires $y_{ij} \overset{iid}{\sim}$ Bernoulli (p_i), $j = 1,..., n_i$ et $\log\{p_i/(1-p_i)\} = z_i^T\beta + v_i, i = 1,..., m$, où z_i est le vecteur de covariables par région, $v_i \overset{iid}{\sim} N\left(0, \sigma_v^2\right)$ et p_i, la proportion de la i^{e} région. L'estimateur de l'EQM minimale (ou de Bayes) de p_i est donné par $\hat{p}_i^B = E\left(p_i \mid y_{i.}, \beta, \sigma_v^2\right) =: k\left(y_{i.}, \beta, \sigma_v^2\right)$, où $y_{i.} = \sum_j y_{ij}$. L'estimateur BE de p_i est $\hat{p}_i^{EB} = k\left(y_{i.}, \hat{\beta}, \hat{\sigma}_v^2\right)$, où $\hat{\beta}$ et $\hat{\sigma}_v^2$ sont des estimateurs appropriés de β et σ_v^2 tirés de la distribution marginale des y_i.

La méthode jackknife fait intervenir la décomposition orthogonale suivante de $EQM\left(\hat{p}_i^{EB}\right)$:

$$EQM\left(\hat{p}_i^{EB}\right) = E\left(\hat{p}_i^B - p_i\right)^2 + E\left(\hat{p}_i^{EB} - \hat{p}_i^B\right)^2. \tag{4.1}$$

En fonction de la décomposition en (4.1), Jiang et coll. (2002) ont proposé les étapes suivantes d'un traitement jackknife d'estimation $EQM\left(\hat{\theta}_i^{EB}\right)$:

(1) On calcule $\hat{\beta}(\ell)$ et $\hat{\sigma}^2(\ell)$ en supprimant les données de la $^{\text{e}}$ région (y_i, x_i); soit $\hat{p}_i^{EB}(\ell) = k\left\{y_{i.}, \hat{\beta}(\ell), \hat{\sigma}^2(\ell)\right\}$ l'estimateur BE de p_i en fonction de $\hat{\beta}(\ell)$ et de $\sigma^2(\ell)$ (à noter que y_i est inchangé).

(2) On calcule l'estimateur jackknife du dernier terme de (4.1) comme

$$\hat{M}_{2i} = \frac{m-1}{m}\sum_{\ell=1}^{m}\left[\hat{p}_i^{EB}(\ell) - \hat{p}_i^{EB}\right]^2. \tag{4.2}$$

(3) Le premier terme $E\left(\hat{p}_i^B - p_i\right)^2$ peut s'écrire comme $E\left[\tilde{g}_{1i}\left(y_{i.}, \beta, \sigma_v^2\right)\right] =: g_{1i}\left(\beta, \sigma_v^2\right)$, où $\tilde{g}_{1i}\left(y_{i.}, \beta, \sigma_v^2\right) = V\left(p_i \mid y_{i.}, \beta, \sigma_v^2\right)$ est la variance postérieure de p_i compte tenu de y_i et de $\left(\beta, \sigma_v^2\right)$. On corrige le biais de $g_{1i}\left(\hat{\beta}, \hat{\sigma}_v^2\right)$ (en tant qu'estimateur de $g_{1i}\left(\beta, \sigma_v^2\right)$) à l'aide de la méthode jackknife de réduction de biais. L'estimateur en correction de biais est donné par

$$\hat{M}_{1i} = g_{1i}\left(\hat{\beta}, \hat{\sigma}_v^2\right) - \frac{m-1}{m}\sum_{\ell=1}^{m}\left[g_{1i}\left(\hat{\beta}(\ell), \hat{\sigma}_v^2(\ell)\right) - g_{1i}\left(\hat{\beta}, \hat{\sigma}_v^2\right)\right]. \tag{4.3}$$

Il convient de noter que le terme principal $g_{1i}\left(\hat{\beta},\hat{\sigma}_v^2\right)$ n'est pas « par région », en ce sens qu'il ne dépend pas de $y_{i.}$.

(4) On calcule l'estimateur jackknife de l'EQM comme

$$eqm_J\left(\hat{p}_i^{EB}\right)=\hat{M}_{1i}+\hat{M}_{2i}. \tag{4.4}$$

Booth et Hobert (1998) ont fait valoir que, pour des réponses non normales, l'estimateur EQM devrait être par région, puisque la variance postérieure de p_i compte tenu de $\left(\beta,\sigma_v^2\right)$ dépend de $y_{i.}$, contrairement à ce qui se passe dans des modèles mixtes linéaires. Ils ont proposé une estimation EQM conditionnelle compte tenu des données de la i^{e} région $\left(y_i,z_i\right)$ comme étant la mesure appropriée de variabilité. Ils ont alors estimé l'EQM conditionnelle. Rao (2003, chapitre 9) réagit à cette critique en modifiant simplement l'estimateur $\hat{M}_{1i}$ en correction de biais. Au lieu d'évaluer l'espérance de $\tilde{g}_{1i}\left(y_{i.},\beta,\sigma_v^2\right)$ pour la distribution marginale de y_i (par intégration numérique), il a voulu corriger le biais de $\tilde{g}_{1i}\left(y_{i.},\hat{\beta},\hat{\sigma}_v^2\right)$ comme estimateur de $g_{1i}\left(y_{i.},\beta,\sigma_v^2\right)$. Cela donne

$$\tilde{M}_{1i}\left(y_{i.}\right)=\tilde{g}_{1i}\left(y_{i.},\hat{\beta},\hat{\sigma}_v^2\right)-\frac{m-1}{m}\sum_{\ell=1}^{m}\left[\tilde{g}_{1i}\left(y_{i.},\hat{\beta}(\ell),\hat{\sigma}_v^2(\ell)\right)-\tilde{g}_{1i}\left(y_{i.},\hat{\beta},\hat{\sigma}_v^2\right)\right], \tag{4.5}$$

qui est un traitement par région avec le terme principal $\tilde{g}_{1i}\left(y_{i.},\hat{\beta},\hat{\sigma}_v^2\right)$. L'estimateur jackknife modifié de l'EQM est à son tour donné par

$$eqm_J^*\left(\hat{p}_i^{EB}\right)=\tilde{M}_{1i}\left(y_{i.}\right)+\hat{M}_{2i}. \tag{4.6}$$

On notera que (4.6) est non seulement par région, mais aussi d'un calcul plus simple que (4.4), car on n'a pas à évaluer l'espérance de $\tilde{g}_{1i}\left(y_{i.},\beta,\sigma_v^2\right)$ pour la distribution marginale de y_i. Dans un modèle mixte linéaire, $\tilde{M}_{1i}\left(y_{i.}\right)=\hat{M}_{1i}$ et, par conséquent, (4.6) est identique à (4.4).

5. MODÉLISATION DE BASE PAR UNITÉ, MÉTHODE PSEUDO-BE

Passons maintenant à la modélisation de base par unité $y_{ij}=x_{ij}^T\beta+v_i+e_{ij}$ et posons que le modèle vaut pour l'échantillon, c'est-à-dire qu'il n'y a pas de biais de sélection d'échantillon à l'échelle intrarégionale. S'il y a aussi échantillonnage de régions, nous posons également l'absence de biais à l'échelle interrégionale. On peut approcher la moyenne de la i^{e} région $\overline{Y}_i$ par $\mu_i=\overline{X}_i^T\beta+v_i$ en posant que le nombre d'unités de population dans la i^{e} région, N_i, est grand, $\overline{X}_{i.}$ étant la moyenne de population de *x* pour cette i^{e} région. Si nous supposons que $v_i \overset{iid}{\sim} N\left(0,\sigma_v^2\right)$ et indépendant de $e_{ij} \overset{iid}{\sim} N\left(0,\sigma^2\right)$ et que nous estimons les paramètres de modélisation β et σ_v^2 à partir de la distribution marginale des y_{ij} échantillonnés, nous obtenons l'estimateur BE de μ_i sous la forme

$$\hat{\mu}_i^{EB}=\hat{\gamma}_i\left[\bar{y}_i+\left(\overline{X}_i-\bar{x}_i\right)^T\hat{\beta}\right]+\left(1-\hat{\gamma}_i\right)\overline{X}_i^T\hat{\beta}, \tag{5.1}$$

où $\hat{\gamma}_i=\hat{\sigma}_v^2/\left(\hat{\sigma}_v^2+\hat{\sigma}_e^2/n_i\right)$, $\left(\bar{y}_i,\bar{x}_i\right)$ sont les moyennes d'échantillon de la i^{e} région et $\left(\hat{\beta},\hat{\sigma}_v^2\right)$ sont les estimateurs de $\left(\beta,\sigma_v^2\right)$ (voir Battese, Harter et Fuller, 1988). Cet estimateur est aussi l'estimateur MPLSE sans l'hypothèse de normalité, à condition d'estimer β et σ_v^2 par une méthode des moments comme la méthode d'ajustement de

constantes. Comme $n_i \to \infty$, l'estimateur BE converge vers l'estimateur de « régression d'enquête » $\bar{y}_i + \left(\bar{X}_i - \bar{x}_i\right)^T \hat{\beta}$; il converge vers l'estimateur « synthétique de régression » $\bar{X}_i^T \hat{\beta}$ à mesure que n_i décroît. Nous renvoyons le lecteur à Rao (2003), chapitre 7, pour un traitement détaillé des estimations MPLSE et EQM. Un inconvénient avec (5.1) est que le traitement est purement par modèle et ne se trouve pas à tenir compte de la pondération d'enquête w_{ij}. Ainsi, il n'est pas par rapport de plan d'échantillonnage à mesure que s'élève n_i, à moins que le plan ne soit autopondéré à l'échelle intrarégionale, c'est-à-dire que $w_{ij} = w_i$. En revanche, l'estimateur MPLSE en modélisation par région est cohérent pour le plan d'échantillonnage. La cohérence en plan d'échantillonnage est aussi souhaitable dans la modélisation par unité, parce que n_i pourrait être modérément élevé pour certaines des régions visées. Il convient en outre de s'assurer que les estimateurs des totaux des régions donnent automatiquement par addition l'estimateur direct de « régression d'enquête » du total de grande région. You et Rao (2002a) ont conçu un estimateur pseudo-MPLSE de μ_i qui offre l'une et l'autre des propriétés désirées.

Nous avons supposé que les effets aléatoires de petite région v_i sont distribués normalement dans la modélisation de base par unité $y_{ij} = x_{ij}^T \beta + v_i + e_{ij}$. L'estimateur EQM de l'estimateur $\hat{\mu}_i^{EB}$ purement par modèle sous une hypothèse de normalité de v_i n'est pas robuste par rapport à l'absence de normalité, contrairement à ce qui se passe dans la modélisation de base par région à la section 3.2. Il serait bon d'étudier l'inférence BE dans des représentations semi-non-paramétriques (SNP) de la densité de v_i. Zhang et Davidian (2001) ont procédé à l'approximation de la densité de v_i par une représentation SNP qui fait intervenir la normalité comme cas d'espèce. Elle donne de la souplesse pour la prise en compte de l'absence d'hypothèse de normalité grâce à un paramètre d'ajustement choisi par l'utilisateur. Maiti (2001) a pris un mélange fini de distributions normales pour la distribution de v_i et établi des estimations de Bayes hiérarchisées (BH) des moyennes de petite région en posant une distribution antérieure pour les paramètres de modélisation. Une estimation BE des moyennes de petite région et une estimation EQM liée pour de grandes catégories de densité de v_i comme nous venons de l'évoquer seraient utiles dans la pratique.

6. MÉTHODE DE BAYES HIÉRARCHISÉE (BH)

Nous allons illustrer le traitement BH dans la modélisation de base par région à la section 2. Dans ce cas, nous spécifions une distribution antérieure $\delta = \left(\beta^T, \sigma_v^2\right)$ pour les paramètres de modélisation et les inférences sont alors fondées sur la distribution postérieure, $f\left(\theta_i \mid \hat{\theta}\right)$, de θ_i compte tenu des données $\hat{\theta} = \left(\hat{\theta}_1, ..., \hat{\theta}_m\right)^T$. Plus précisément, nous estimons θ_i par sa moyenne postérieure $E\left(\theta_i \mid \hat{\theta}\right)$, appelée estimateur de BH $\hat{\theta}_i^{HB}$. La variabilité de $\hat{\theta}_i^{HB}$ se mesure par la variance postérieure $V\left(\theta_i \mid \hat{\theta}\right)$. Le calcul de $\hat{\theta}_i^{HB}$ et $V\left(\theta_i \mid \hat{\theta}\right)$ comporte des intégrations en fonction de la distribution postérieure de β, σ_v^2, $f\left(\beta, \sigma_v^2 \mid \hat{\theta}\right)$. On peut toutefois se servir de méthodes de Monte Carlo à chaîne de Markov (CMMC) pour tirer directement *J* échantillons en simulation $\left\{\theta_1^{(j)}, ..., \theta_m^{(j)}; j = 1, ..., J\right\}$ de la codistribution postérieure $f\left(\theta \mid \hat{\theta}\right)$, où $\theta = (\theta_1, ..., \theta_m)^T$ et où *J* est suffisamment grand. Par ces échantillons simulés, nous approximons $\hat{\theta}_i^{HB}$ par la moyenne $\theta_i^{(.)} = J^{-1} \sum_j \theta_i^{(j)}$ et $V\left(\theta_i \mid \hat{\theta}\right)$ par la variance $J^{-1} \sum_j \left(\theta_i^{(j)} - \theta_i^{(.)}\right)^2$ des échantillons simulés $\theta_i^{(j)}$. Nous obtenons l'approximation de l'estimateur BH du total Y_i par la moyenne $Y_i^{(.)} = J^{-1} \sum_j Y_i^{(j)}$ et la variance postérieure de Y_i, $V\left(Y_i \mid \hat{\theta}\right)$, par la variance $J^{-1} \sum_j \left(Y_i^{(j)} - Y_i^{(.)}\right)^2$, où $Y_i^{(j)} = g^{-1}\left(\theta_i^{(j)}\right)$.

Un avantage de la méthode BH est qu'elle est d'une application simple, que les inférences sont « exactes » à la différence de la méthode BE (ou MPLSE) et qu'il est possible de traiter des modèles complexes de petite région par les méthodes CMMC, mais avec l'exigence de la spécification d'une distribution antérieure $f\left(\beta,\sigma_v^2\right)$ pour les paramètres de modélisation. Il serait souhaitable de choisir une distribution antérieure d'appariement $f\left(\beta,\sigma_v^2\right)\propto f\left(\sigma_v^2\right)$ qui mène à des inférences bien étalonnées. Il faudrait en particulier que la variance postérieure soit approximativement sans biais pour $EQM\left(\hat{\theta}_i^{HB}\right)$, c'est-à-dire que $E\left[V\left(\theta_i \mid \hat{\theta}\right)\right]-EQM\left(\theta_i^{HB}\right)=o\left(m^{-1}\right)$; asymptotiquement, $\hat{\theta}_i^{HB}\approx\hat{\theta}_i^{EB}$. On aurait alors une justification fréquentiste de la variance postérieure comme mesure de variabilité. Datta, Rao et Smith (2002) ont démontré qu'une telle distribution antérieure d'appariement est donnée par

$$f_i\left(\sigma_v^2\right)\propto\left(\sigma_v^2+\psi_i\right)^2\sum_{\ell=1}^{m}\left(\sigma_v^2+\psi_\ell\right)^{-2}. \tag{6.1}$$

Elle dépend collectivement des variances d'échantillonnage ψ_ℓ pour l'ensemble des régions ℓ, ainsi que de la variance d'échantillonnage par région ψ_i. Pour le cas d'équilibre $\psi_i=\psi$, la distribution antérieure d'appariement se réduit à la distribution antérieure « plate » $f\left(\sigma_v^2\right)\propto 1$. À noter que la distribution antérieure (6.1) pour le paramètre commun σ_v^2 est conçue pour une inférence sur la i^{e} région, si bien que sa dépendance à l'égard de ψ_i peut ne pas faire problème.

Un inconvénient avec l'estimateur BE $\hat{\theta}_i^{EB}$ est que la valeur de pondération $\hat{\gamma}_i$ qui s'attache à l'estimateur direct devient nulle lorsque $\hat{\sigma}_v^2=0$, auquel cas on est ramené aux estimateurs synthétiques de régression $z_i^T\hat{\beta}$. Ainsi, l'estimateur direct de $\hat{\theta}_i$ reçoit une valeur nulle de pondération dans tous les cas même si les tailles d'échantillon ne sont pas petites pour un certain nombre de régions. C'est cette difficulté qui s'est présentée lorsqu'on a utilisé un modèle d'états pour produire des estimations BE de la population enfantine pauvre d'âge scolaire aux États-Unis (National Research Council, 2000). Avec la méthode BH, on évite la difficulté en produisant des valeurs positives de pondération dans tous les cas. Bell (1999) a appliqué cette méthode au modèle d'états avec la spécification antérieure $f\left(\beta,\sigma_v^2\right)=f(\beta)f\left(\sigma_v^2\right)$ où $f(\beta)\propto 1$ et $f\left(\sigma_v^2\right)\propto 1$ pour ainsi obtenir des valeurs positives partout. On ne sait au juste cependant si cette méthode permet des inférences bien étalonnées, puisque la distribution antérieure d'appariement (6.1) est différente de la distribution antérieure plate, surtout lorsque les valeurs ψ_i varient significativement comme dans le cas de Bell (1999) avec un $\max\left(\psi_i\right)/\min\left(\psi_i\right)$ aussi grand que 20.

You et Rao (2002b) ont pris la méthode BH pour traiter les cas de non-correspondance des modèles d'échantillonnage et d'appariement (section 2) et l'ont appliquée aux estimations de sous-dénombrement du recensement canadien. Dans cette application, $C_i=$ dénombrement censitaire, $Y_i=$ nombre d'unités manquantes; $\hat{Y}_i$ est un estimateur d'enquête postcensitaire de Y_i avec une variance d'échantillonnage connue σ_i^2 pour la i^{e} province canadienne $\left(i=1,...,m=10\right)$. You et Rao ont estimé les σ_i^2 par ajustement de modèle FGV de la forme $V\left(\hat{Y}_i\right)\propto C_i^\gamma$ et en traitant ensuite le tout comme si cette spécification était connue dans le modèle d'échantillonnage $\hat{Y}_i \mid Y_i \overset{ind}{\sim} N\left(Y_i,\sigma_i^2\right)$. Le modèle d'appariement est donné par $\theta_i=\log\left\{Y_i/\left(Y_i+C_i\right)\right\}=\beta_0+\beta_1\log C_i+v_i$ avec $v_i \overset{iid}{\sim} N\left(0,\sigma_v^2\right)$. Ces auteurs ont établi par les méthodes CMMC des estimations BH des sous-dénombrement Y_i, des taux de sous-dénombrement $U_i=Y_i/\left(Y_i+C_i\right)$ et des coefficients de variation associés (en fonction de la variance postérieure).

Singh, Folsom et Vaish (2003) ont étudié la modélisation de base par unité $y_{ij} = x_{ij}^T \beta + v_i + e_{ij}$ avec $v_i \overset{iid}{\sim} N\left(0, \sigma_v^2\right)$ indépendant de $e_{ij} \overset{iid}{\sim} N\left(0, \sigma_e^2\right)$ pour la population et ils ont tenu compte d'un biais de sélection d'échantillon à l'intérieur des petites régions. Ils ont employé à cette fin des méthodes reposant sur des fonctions d'estimation (FE) en pondération d'enquête. Ils ont aussi étendu l'application de telles méthodes à des modèles mixtes linéaires généralisés comme le modèle mixte logistique $y_{ij} \mid p_{ij} \overset{ind}{\sim}$ Bernoulli $\left(p_{ij}\right)$ et logit $\left(p_{ij}\right) = \log\left\{p_{ij} / \left(1 - p_{ij}\right)\right\} = x_{ij}^T \beta + v_i$. On trouvera dans Rao (2003, p. 253-254) une brève description des méthodes FE. En cas de sélection d'un échantillon de régions, on pose que les effets aléatoires v_i sont exempts de biais de sélection.

La méthode BH est puissante et attrayante, mais on doit user de prudence lorsqu'on recourt aux méthodes CMMC. Ainsi, des algorithmes CMMC pourraient faire faire des inférences en apparence acceptables au sujet d'une distribution postérieure qui n'existe pas, ce qui se produit lorsque celle-ci ne convient pas et, pourtant, que toutes les distributions conditionnelles de Gibbs à la base de la production d'échantillons en simulation CMMC conviennent, elles (Hobert et Casella, 1996). Une autre difficulté avec les méthodes CMMC est que les outils de diagnostic de convergence peuvent ne pas détecter le genre de défauts de convergence qu'ils sont là pour reconnaître (Cowles et Carlin, 1996). Dans Rao (2003), section 10.2.4, le lecteur pourra trouver un examen des questions pratiques d'application de ces méthodes.

Celles-ci servent largement à une détermination du modèle qui joue un rôle primordial dans l'établissement d'estimations de modélisation pour de petites régions. Notons en particulier que les méthodes fondées sur les facteurs Bayésien, densités prédictives de distribution postérieure et densités prédictives d'intervalidation sont mises au service de cette détermination (voir Rao (2003), section 10.2.6). On se reporte souvent au critère des probabilités prédictives postérieures pour vérifier l'ajustement général d'un modèle proposé. Sinharay et Stern (2003) ont réalisé une étude de simulation sur l'efficacité de ce critère comme moyen de vérification de modèles, en l'occurrence un modèle de base par région sans covariables. Cette étude indique la difficulté que l'on a à reconnaître par ce critère la non-normalité des effets aléatoires v_i sauf là où le degré de dérogation à cette hypothèse est très élevé. C'est donc avec prudence que l'on doit appliquer des critères BH à la détermination du modèle.

7. CONSIDÉRATIONS PRATIQUES

Dans cette section, nous ferons de brèves observations sur des questions pratiques d'estimation régionale.

(i) Questions de plan d'échantillonnage

Il importe de s'attacher aux questions de plan d'échantillonnage qui influent sur les estimations relatives à de petites régions. En y trouvant une bonne solution, on pourrait rendre plus sûres les estimations tant directes qu'indirectes pour des domaines (régions) planifiés ou non. Au stade de la planification d'échantillonnage, les mesures suivantes pourraient avoir pour effet d'atténuer le besoin de recourir à des estimateurs indirects, du moins dans certains domaines planifiés : on pourrait utiliser une base de sondage par liste pour remplacer des grappes dans la mesure du possible, exploiter un grand nombre de petites strates pour l'échantillonnage, altérer la répartition d'échantillon pour rendre l'estimation plus sûre tant pour les petites que pour les grandes régions, intégrer des enquêtes en harmonisant les questions d'enquêtes différentes auprès de la même population, mener des enquêtes à base de sondage multiple ou procéder à un échantillonnage par renouvellement comme méthode de traitement cumulé des données dans le temps. Le lecteur trouvera plus de détails dans Rao (2003), chapitre 2, Singh et coll. (1994) et Marker (2001).

Malgré ces mesures de prévention au stade de la conception, il faudra des estimations indirectes dans la pratique, puisqu'il est impossible de prévoir et de planifier tous les domaines (régions) et les usages possibles de données d'enquête et que « le client exigera toujours plus que ce qui a été spécifié au stade de la conception [TRADUCTION] » (Fuller, 1999, p. 344).

(ii) Sélection et validation de modèles

L'évolution méthodologique et les applications de l'estimation par modèle sont impressionnantes, mais les hypothèses de modélisation nous incitent à la prudence. Le fait de disposer d'une bonne information auxiliaire pour les variables d'intérêt joue un rôle primordial dans l'établissement de modèles d'appariement appropriés. On doit donc prêter une plus grande attention à l'obtention de variables auxiliaires qui constituent de bons prédicteurs des variables étudiées. Les spécialistes ou les utilisateurs finaux d'un domaine statistique devraient avoir voix au chapitre dans le choix des modèles en général et des variables auxiliaires en particulier. On devrait cependant se servir d'outils de diagnostic de modélisation pour trouver des modèles appropriés qui s'ajustent bien aux données, qu'il s'agisse d'une analyse de résidus permettant de relever les écarts par rapport aux modèles posés ou de diagnostics de sélection de variables auxiliaires et de suppression de cas pour la détection des observations dominantes. Nous renvoyons le lecteur à Rao (2003), chapitre 6, pour une description de certaines méthodes de validation de modèles dans un cadre fréquentiste.

La méthode de Bayes hiérarchisée (BH) a fait beaucoup d'adeptes ces dernières années, parce qu'elle était capable de traiter des modèles complexes par les méthodes CMMC, mais on doit se garder de choisir des distributions antérieures impropres pour les paramètres de modélisation, comme nous l'avons fait remarquer à la section 6. Il faut aussi connaître les limites de ces méthodes (lacunes des outils disponibles de diagnostic de convergence, par exemple). Carlin et Louis (2000) ont fait des commentaires importants à propos des dangers d'une application mécanique des méthodes CMMC en disant : « Pis encore, la puissance même des méthodes CMMC a fait que les gens ont été tentés d'ajuster les modèles en plus grand que ne le permettaient d'emblée les données, donc sans une structure antérieure très informative, ce qui a maintenant tout d'une rareté dans des travaux appliqués du type bayésien [TRADUCTION]. »

Les méthodes BH de validation de modèles dans un cadre CMMC sont très développées, mais on peut s'interroger sur l'efficacité avec laquelle s'appliquent certains critères de vérification de la modélisation, comme nous l'avons indiqué à la section 6. Il faudra pousser la recherche sur l'efficacité méthodologique de la vérification et de la modélisation.

(iii) Modèles par région et par unité

Les modèles par région sont d'une plus grande portée que les modèles par unité, l'information auxiliaire étant plus immédiatement disponible pour les premiers que pour les seconds. Ajoutons qu'on intègre la pondération d'échantillonnage en modélisant des estimateurs directs en pondération d'échantillonnage et que les estimateurs BE et BH ainsi obtenus sont cohérents en plan d'échantillonnage. Il reste que l'hypothèse des variances d'échantillonnage connues ψ_i est des plus restrictives. Les estimations lissées de ψ_i par ajustement de modèle FGV peuvent aussi être source de difficultés dans l'estimation de l'EQM (voir la section 3.3). Nous devons viser à obtenir de bonnes approximations des variances d'échantillonnage, ainsi que des méthodes qui tiennent compte de la variabilité liée aux variances estimées d'échantillonnage dans l'estimation EQM. La tâche se complique avec des modèles par région à plusieurs variables ou par série chronologique, car il faut également les covariances d'échantillonnage.

On jugera prometteurs les récents travaux d'intégration de la pondération d'enquête à l'estimation basée sur le modèle pseudo-BE ou pseudo-BH et, plus particulièrement, la propriété d'autoétalonnage mentionnée à la section 5. Mais l'hypothèse d'absence de biais de sélection d'échantillon peut ne pas se vérifier dans certaines applications. Les fonctions d'estimation (FE) de Singh, Folsom et Vaish (2003), dont nous avons fait mention à la section 6, tiennent compte d'un biais de sélection à l'échelle intrarégionale, mais on y suppose aussi que les effets aléatoires v_i sont exempts d'un tel biais à l'échelle interrégionale. Nous avons besoin de méthodes pour traiter ce dernier cas. De plus, Singh, Folsom et Vaish (2003) supposent avec leur méthode que les variances d'échantillonnage sont connues comme dans la modélisation de base par région; c'est une hypothèse qui est peut-être restrictive (voir Rao, 2003, section 10.5.4).

(iv) Estimation à trois usages

Nous avons surtout parlé d'estimations de modélisation de totaux ou de moyennes de petites régions, mais de telles estimations pourraient ne pas convenir si le but principal est de produire un ensemble d'estimations paramétriques dont la distribution est en un certain sens assez proche de la distribution de paramètres par région. Notre but peut être, par exemple, d'ordonner des régions ou de reconnaître celles qui se situent au-dessous ou au-dessus d'un certain niveau établi d'avance. Shen et Louis (1998) ont proposé des estimateurs à trois usages qui permettent d'établir de bons rangs, de réaliser un histogramme convenable et de former de bons estimateurs par région par une modélisation simple d'appariement. Il serait utile d'étendre l'application de leurs méthodes à une modélisation plus complexe se prêtant aux estimations relatives à de petites régions.

(v) Erreurs non dues à l'échantillonnage

Nous avons supposé que les réponses et/ou les covariables étaient exemptes d'erreurs de mesure et que la non-réponse ne contribuait pas à l'erreur. Il reste que les erreurs non dues à l'échantillonnage peuvent largement influer sur les estimations relatives à de petites régions et qu'il conviendrait de concevoir des plans d'échantillonnage et des méthodes d'estimation permettant de bien tenir compte de ces erreurs. Nandram et Choi (2002) se sont servis de modèles BH de non-réponse pour des données binaires et ont appliqué la théorie aux données de la National Crime Survey américaine afin d'estimer les proportions régionales. Même sous une hypothèse d'additivité, les erreurs de mesure des réponses peuvent mener à des biais d'estimation des quantiles et des histogrammes. Dans le contexte de l'estimation directe, Fuller (1995) a proposé, pour le stade de la conception, des méthodes qui peuvent donner des estimateurs des quantiles et des histogrammes en corrigées pour le biais.

RÉFÉRENCES

Battese, G.E., Harter, R.M., Fuller, W.A. (1988). An Error-Components Model for Prediction of Crop Areas Using Survey and Satellite Data, *Journal of the American Statistical Association*, **83**, 28-36.

Bell, W.R. (1999). Accounting for Uncertainty About Variances in Small Area Estimation. *Bulletin of the International Statistical Institute*.

Booth, J.G. et Hobert, J.P. (1998). Standard Errors of Predictors in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, **93**, 262-272.

Butar, F.B. et Lahiri, P. (2001). On Measures of Uncertainty of Empirical Bayes Small-Area Estimators. Technical Report, Department of Statistics, University of Nebraska-Lincoln.

Carlin, B.P. et Louis, T.A. (2000). Empirical Bayes: Past, Present and Future. *Journal of the American Statistical Association*, **95**, 1286-1289.

Cowles, M.R. et Carlin, B.P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, **91**}, 883-904.

Datta, G.S. et Lahiri, P. (2000). A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems. *Statistica Sinica*, **10**, 613-627.

Datta, G.S., Rao, J.N.K. et Smith, D.D. (2002). On Measures of Uncertainty of Small Area Estimators in the Fay-Herriot Model, Technical Report, University of Georgia, Athens.

Fay, R.E. et Herriot, R.A. (1979). Estimation of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, **74**, 269-277.

Fuller, W.A. (1989). Prediction of True Values for the Measurement Error Model, in *Conference on Statistical Analysis of Measurement Error Models and Applications*, Humboldt State University.

Fuller, W.A. (1995). Estimation in the Presence of Measurement Error. *International Statistical Review*, **63**, 121-147.

Fuller, W.A. (1999). Environmental Surveys Over Time. *Journal of Agriculatural, Biological and Environmental Statistics*, **4**, 331-345.

Hobert, J.P. et Casella, G. (1996). The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models. *Journal of the American Statistical Association*, **91**, 1461-1473.

Jiang, J., Lahiri, P. et Wan, S.-M. (2002). A Unified Jackknife Theory for Empirical Best Prediction with M-Estimation. *Annals of Statistics*, **30**, 1782-1810.

Lahiri, P. et Rao, J.N.K. (1995). Robust Estimation of Mean Squared Error of Small Area Estimators. *Journal of the American Statistical Association*, **82**, 758-766.

Laird, N.M. et Louis, T.A. (1987). Empirical Bayes Confidence Intervals Based on Bootstrap Samples. *Journal of the American Statistical Association*, **82**, 739-750.

Maiti, T. (2001). Robust Generalized Linear Mixed Models for Small Area Estimation. *Journal of Statistical Planning and Inference*, **98**, 225-238.

Marker, D.A. (2001). Producing Small Area Estimates from National Surveys: Methods for Minimizing Use of Indirect Estimators. *Survey Methodology*, **27**, 183-188.

Nandram, B. et Choi, J.W. (2002). Hierarchical Bayesian Nonresponse Models for Binary Data from Small Areas with Uncertainty About Ignorability. *Journal of the American Statistical Association*, **97**, 381-388.

National Research Council (2000). *Small-Area Estimates of School-Age Children in Poverty: Evaluation of Current Methodology*, C.F. Citro and G. Kalton (Eds.), Committee on National Statistics, Washington, D.C.: National Academy Press.

Prasad, N.G.N. et Rao, J.N.K. (1990). The Estimation of the Mean Squared Error of Small-Area Estimators, *Journal of the American Statistical Association*, **85**, 163-171.

Rao, J.N.K. (2001). BE and MPLSE in Small Area Estimation, in S.E. Ahmed and N. Reid (Eds.); *Empirical Bayes and Likelihood Inference*. Lecture Notes in Statistics 148, New York: Springer, pp. 33-43.

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

Rivest, L-P. et Belmonte, E. (2000). A Conditional Mean Squared Error of Small Area Estimators. *Survey Methodology*, **26**, 67-78.

Rivest, L-P. et Vandal, N. (2003). Mean Squared Error Estimation for Small Areas When the Small Area Variances are Estimated, in *Proceedings of the International Conference on Recent Advances in Survey Sampling*, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa, Canada (in press).

Singh, M.P., Gambino, J., Mantel, H.J. (1994). Issues and Strategies for Small Area Data, *Survey Methodology*, **20**, 3-22.

Singh, A.C., Folsom, Jr., R.E. et Vaish, A.K. (2003). Estimating Function Based Approach to Hierarchical Bayes Small Area Estimation for Survey Data, in *Proceedings of the International Conference on Recent Advances in Survey Sampling*, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa, Canada (in press).

Shen, W. et Louis, T.A. (1998). Triple-Goal Estimates in Two-Stage Hierarchical Models. *Journal of the Royal Statistical Society, Series B*, **60**, 455-471.

Sinharay, S. et Stern, H.S. (2003). Posterior Predictive Model Checking in Hierarchical Models. *Journal of Statistical Planning and Inference*, **111**, 209-221.

You, Y. et Rao, J.N.K. (2002a). A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights. *Canadian Journal of Statistics*, **30**, 431-439.

You, Y. et Rao, J.N.K. (2002b). Small Area Estimation Using Unmatched Sampling and Linking Models. *Canadian Journal of Statistics*, **30**, 3-15.

Wang, J. et Fuller, W.A. (2003). The Mean Squared Error of Small Area Predictors Constructed With Estimated Area Variances. *Journal of the American Statistical Association*, **98**, 716-723.

Zhang, D. et Davidian, M. (2001). Linear Mixed Models With Flexible Distributions of Random Effects for Longitudinal Data. *Biometrics*, **57**, 795-802.