Recueil du Symposium 2001 de Statistique Canada
La qualité des données d'un organisme statistique : une perspective méthodologique

DE LA MESURE DE LA QUALITÉ DES ESTIMATIONS RÉGIONALES INDIRECTES

J.N.K Rao[1]

RÉSUMÉ

Généralement, la taille des échantillons régionaux est très petite, si bien que les estimateurs directs habituels, propres à une région, de la moyenne régionale produisent des estimations de qualité inacceptable si l'on en juge par l'EQM. Par conséquent, on a largement recours, à l'heure actuelle, à des estimateurs indirects dont l'efficacité est renforcée au moyen de données empruntées à des régions connexes grâce à des modèles de liaison fondés sur des données auxiliaires. Ces modèles de liaison peuvent être implicites (comme dans le cas des estimateurs synthétiques) ou explicites (comme dans le cas des estimateurs modélisés). En théorie fréquentiste, on mesure la qualité d'un estimateur indirect par estimation de son EQM, tandis qu'en théorie bayésienne, on s'appuie sur la variance a posteriori de la moyenne régionale. Le présent article a pour but de passer en revue certains travaux récents sur l'estimation de l'EQM et sur l'évaluation de la variance a posteriori.

MOTS CLÉS : Estimateurs composites; modèles de liaison; EQM; variance a posteriori.

1. INTRODUCTION

Les enquêtes par sondage sont généralement conçues pour produire des estimations directes sur échantillon caractérisées par un faible coefficient de variation (c.v.) pour les grandes régions (ou domaines). En fait, les statisticiens d'enquête insistent sur le fait que les erreurs non dues à l'échantillonnage, y compris les erreurs de mesure de couverture et la non-réponse, contribuent nettement plus que les erreurs d'échantillonnage à l'erreur quadratique moyenne (EQM) totale qui est souvent utilisée comme indicateur de la qualité des estimateurs. Par contre, dans le cas des estimations régionales, les erreurs d'échantillonnage jouent un rôle dominant, parce que, pour les petites régions géographiques, il est rare que la taille de l'échantillon soit suffisamment grande pour que l'on puisse produire des estimateurs directs, propres à la région, d'une qualité acceptable si l'on en juge par l'EQM (ou le c.v.). En fait, pour des nombreuses régions étudiées, la taille de l'échantillon est nulle. Par exemple, aux États-Unis, dans le cas de l'estimation du nombre d'enfants pauvres d'âge scolaire selon le comté ou le district scolaire d'après les données de la Current Population Survey (CPS), du Recensement et des dossiers administratifs, la taille de l'échantillon de la CPS est nulle pour nombre de comtés (National Research Council, 2000).

Il est donc nécessaire d'utiliser, pour calculer les estimations régionales, des estimateurs indirects contenant des informations empruntées à des régions adjacentes grâce à des modèles de liaison basés sur des données de recensement et des données administratives. Ces modèles de couplage peuvent être implicites ou explicites. Les estimateurs indirects classiques fondés sur des modèles implicites incluent les estimateurs synthétiques et composites. Malheureusement, l'estimation de l'EQM de ces estimateurs pose des difficultés (section 2). En revanche, ces dernières années, beaucoup d'attention a été accordée aux estimateurs indirects fondés sur des modèles explicites, parce qu'ils ont sur les estimateurs indirects classiques les avantages qui suivent : i) les méthodes fondées sur un modèle explicite tiennent compte spécifiquement de la variation locale grâce à l'inclusion de structures d'erreur complexes dans le modèle destiné à établir le lien entre les régions; ii) les modèles peuvent être validés d'après des données

[1] J.N.K. Rao, Carleton University, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6

d'échantillon; iii) les méthodes permettent de traiter certains cas complexes, comme les données transversales, les données chronologiques et les données multivariées; iv) les mesures de la variabilité des estimations au niveau de la région sont stables, contrairement aux mesures globales utilisées habituellement dans le cas des estimateurs classiques indirects.

Dans le présent article, nous donnons un aperçu des travaux ayant trait à l'estimation de l'EQM des estimateurs indirects en insistant sur les méthodes fondées sur des modèles explicites. Nous décrivons aussi brièvement l'évaluation de la variance a posteriori selon les méthodes hiérarchiques de Bayes (HB), nous utilisons la variance a posteriori comme mesure de la qualité des estimateurs dans le cas de la méthode bayésienne.

2. ESTIMATEURS INDIRECTS CLASSIQUES

Les estimateurs synthétiques, $\hat{Y}_i(S)$, des totaux régionaux, Y_i, sont d'un usage très répandu, en raison de leur simplicité et de leur efficacité. Supposons que l'on dispose à la fois d'un estimateur direct fiable, $\hat{\bar{Y}}.g$, de la moyenne de grandes strates a posteriori, $\bar{Y}.g$, et de chiffres de population de cellule $\{N_{ig}\}$. Un estimateur synthétique simple de Y_i prend alors la forme

$$\hat{Y}_i(S) = \sum_g N_{ig} \hat{\bar{Y}}.g \ . \tag{2.1}$$

La variance de cette estimateur est faible et son EQM est satisfaisante si le modèle implicite $\bar{Y}_{ig} \approx \bar{Y}.g$ s'applique pour tout *i*, où $\bar{Y}_{ig}$ représente la moyenne de la cellule (i,g).

Un estimateur non biaisé de l'EQM de $\hat{Y}_i(s)$ est donné par

$$\begin{aligned} eqm\left[\hat{Y}_i(S)\right] &= \left(\hat{Y}_i(S) - \hat{Y}_i\right)^2 - v\left(\hat{Y}_i(S) - \hat{Y}\right) + v\left(\hat{Y}_i(S)\right), \\ &\approx \left(\hat{Y}_i(S) - \hat{Y}\right)^2 - v\left(\hat{Y}_i\right), \end{aligned} \tag{2.2}$$

où $\hat{Y}_i$ représente un estimateur direct non biaisé de Y_i et $v(.)$ représente un estimateur de variance. Les estimateurs de la variance de $\hat{Y}_i(S) - \hat{Y}_i$ et $\hat{Y}_i(S)$ s'obtiennent facilement par la méthode du jackknife, particulièrement dans le cas des plans de sondage stratifiés à plusieurs degrés, tels que celui de la CPS. Bien qu'il ne soit pas biaisé, l'estimateur (2.2) est très instable et peu prendre des valeurs négatives. Une méthode adoptée fréquemment pour éviter l'instabilité consiste à calculer la moyenne de $eqm\left[\hat{\bar{Y}}_i(S)\right] = eqm\left[\hat{Y}_i(S)\right] / N_i^2$ sur les régions *i*, puis à utiliser

$$eqm_a\left(\hat{Y}_i(S)\right) = eqm_a\left(\hat{\bar{Y}}_i(S)\right) N_i^2 \tag{2.3}$$

comme estimateur de l'EQM, où $\hat{\bar{Y}}_i(s) = \hat{Y}_i(s) / N_i$ est l'estimateur synthétique de la moyenne $\bar{Y}_i$, N_i est la taille connue de la région et

$$eqm_a\left[\hat{\bar{Y}}_i(S)\right] = \tfrac{1}{m} \sum_i \ eqm\left[\hat{\bar{Y}}_i(S)\right] \cdot \tag{2.4}$$

L'estimateur (2.3) de l'EQM est stable, mais il n'est pas particulier à la région, sauf pour le multiplicateur N_i^2.

Marker (1995) a proposé un autre estimateur de l'EQM qui est plus spécifique à la région que (2.3) et est également stable. Il suppose que le carré du biais, $b^2\left[\hat{\bar{Y}}_i(S)\right]$, de $\hat{\bar{Y}}_i(S)$ est approximativement égal au carré moyen du biais :

$$b^2\left[\hat{\bar{Y}}_i(S)\right] \approx b_a^2\left[\hat{\bar{Y}}_i(S)\right] = eqm_a\left[\hat{\bar{Y}}_i(S)\right] - \tfrac{1}{m}\sum_i \tfrac{1}{N_i^2} v\left[\hat{Y}_i(S)\right]. \quad (2.5)$$

Dans ces conditions hypothétiques, l'estimateur de l'EQM de Marker est donné par

$$eqm_M\left[\hat{Y}_i(S)\right] = v\left[\hat{Y}_i(S)\right] + N_i^2 b_a^2\left[\hat{\bar{Y}}_i(S)\right]. \quad (2.6)$$

Pour l'estimateur synthétique simple (2.1), $eqm_M\left[\hat{Y}_i(S)\right]$ dépend des chiffres de population des cellules propres à la région $\{N_{ig}\}$ par la voie de $v\left[\hat{Y}_i(S)\right]$, tandis que $eqm_a\left[\hat{Y}_i(S)\right]$ dépend uniquement de la taille globale N_i de la $i^{\text{ième}}$ région. En ce sens, l'estimateur de l'EQM de Marker est plus spécifique à la région.

On utilise aussi fréquemment des estimateurs composites de la forme

$$\hat{Y}_i(C) = \hat{\phi}_i \hat{Y}_i + \left(1 - \hat{\phi}_i\right)\hat{Y}_i(S) \quad (2.7)$$

où $\hat{\phi}_i$ est un estimateur du coefficient de pondération optimal $\phi_i(opt)$ qui minimise $EQM\left(\hat{Y}_i(C)\right)$. Les coefficients de pondération de ce genre prennent la forme

$$\hat{\theta} = eqm\left[\hat{Y}_i(S)\right] / \left(\hat{Y}_i(S) - \hat{Y}_i\right)^2 \quad (2.8)$$

et sont très instables. Pour éviter l'instabilité de $\hat{\phi}_i$, on utilise généralement un coefficient de pondération moyen $\hat{\phi}.$ dans l'équation (2.7). Selon le même raisonnement, on peut obtenir un estimateur de $EQM\left[\hat{Y}_i(C)\right]$ pour $\hat{Y}_i(S)$ en remplaçant $\hat{Y}_i(S)$ par $\hat{Y}_i(C)$ dans les équations (2.2) à (2.6). Par conséquent, $\hat{Y}_i(C)$ présente les mêmes limitations que $\hat{Y}_i(S)$, en plus de la difficulté d'obtenir un coefficient de pondération $\hat{\phi}_i$ stable.

Les évaluations externes des estimateurs indirects sont souvent réalisées par comparaison des estimateurs aux valeurs réelles. Gonzalez et coll. (1996) ont étudié les estimateurs synthétiques de dénombrement

$$\hat{P}_i(S) = \left(\sum_g N_{ig}\hat{P}_{ig}\right) / \left(\sum_g N_{ig}\right) \quad (2.9)$$

pour le calcul des proportions P_i, où $\hat{P}_{.g}$ est l'estimateur direct de la proportion dans la strate a posteriori $P_{.g}$. Ils ont évalué l'efficacité de $\hat{P}_i(S)$ comparativement à l'estimateur direct $\hat{P}_i$, au moyen de données provenant de la U.S. National Natality Survey 1980. Pour l'évaluation externe, ils ont obtenu la valeur réelle de P_i pour trois caractéristiques (faible poids de naissance, soins prénataux et indice d'Apgar) et ont utilisé l'estimateur non biaisé de l'EQM.

$$eqm\left[\hat{P}_i(S)\right] = \left(\hat{P}_i(S) - P_i\right)^2. \quad (2.10)$$

Ils se sont servis, pour évaluer l'efficacité de $\hat{P}_i(S)$, des estimations des c.v. calculés d'après l'équation (2.10). Par ailleurs, ils ont utilisé la méthode des répétitions équilibrées répétées (BRR pour *balanced repeated replication*) pour estimer la variance de $\hat{P}_i$. Cependant, les résultats de ce genre de comparaison doivent être interprétés prudemment, car l'estimateur (2.10) est très instable. Dans ce cas, seules les comparaisons globales seraient fiables. Par exemple, on pourrait comparer l'erreur relative absolue (*ERA*), moyenne ou médiane, où $ERA_i = \left|est_i - réelle_i\right| / réelle_i$ pour la $i^{\text{ième}}$ région. Une comparaison plus fine consiste à comparer l'ERA moyenne ou médiane dans des catégories précisées des petites régions. Le National Research Council (2000) a communiqué les résultats de ce genre de comparaisons dans le contexte de l'estimation du nombre d'enfants pauvres d'âge scolaire.

3. MODÈLES RÉGIONAUX

Deux catégorie de modèles régionaux de base ont fait l'objet de publications. Dans le cas des modèles de la première catégorie, appelés modèles de base au niveau de la région, seules des données auxiliaires particulières à la région $z_i = (z_{1i},..., z_{pi})^T$, liées à la moyenne régionale $\overline{Y}_i (i = 1,..., m)$ ou à une fonction appropriée $\theta_i = g(\overline{Y}_i)$, sont utilisées pour élaborer les modèles de liaison de la forme $\theta_i \overset{ind}{\sim} N(z_i^T \beta, \sigma_v^2)$. Le modèle de liaison est combiné au modèle d'échantillonnage $\hat{\theta}_i | \theta_i \overset{ind}{\sim} N(\theta_i, \psi_i)$ où $\hat{\theta}_i$ est un estimateur direct de $\hat{\theta}_i$ dont la variance d'échantillonnage ψ_i est connue. Les estimateurs axés sur un modèle de θ_i et $\overline{Y}_i = g^{-1}(\theta_i)$ sont obtenus à partir du modèle combiné en utilisant la méthode empirique de Bayes (EB) ou la méthode hiérarchique de Bayes (HB). Des mesures de la variabilité associée aux estimateurs sont aussi obtenues à titre d'indicateurs de la qualité de ces estimateurs. Le modèle de base au niveau de la région avec $\theta_i = \log Y_i$ a été utilisé récemment pour produire, par la méthode EB, des estimations par comté du nombre d'enfants pauvres d'âge scolaire aux États-Unis (National Research Council, 2000). Chaque année, le US Department of Education se fonde sur ces estimations pour répartir plus de 7 milliards de dollars du budget fédéral entre les comtés. Diverses extensions du modèle de base au niveau de la région ont été proposées pour tenir compte de la corrélation des erreurs d'échantillonnage, de la dépendance spatiale des θ_i, des données chronologiques et transversales et d'autres facteurs (consulter Rao, 1999 pour une vue d'ensemble récente).

Dans le cas des modèles de la deuxième catégorie, appelés modèles de base au niveau de l'unité, les variables auxiliaires au niveau de l'unité d'échantillonnage $\mathbf{x_{ij}} = (x_{1ij},..., x_{pij})^T$ sont reliées aux valeurs unitaires, y_{ij}, de y au moyen d'un modèle de régression unidimensionnel à erreur emboîtée $y_{ij} = x_{ij}^T \beta + v_i + e_{ij}$, où les $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ sont indépendants des $e_{ij} \overset{iid}{\approx} N(0, \sigma_v^2)$. Les paramètres étudiés sont les moyennes régionales $\overline{Y}_i = \sum_j y_{ij} / N_i \, (j = 1,..., N_i, i = 1,..., m)$. Diverses extensions du modèle de base au niveau de l'unité de population ont été proposées dans le cas de réponses binaires, de réponses multivariées, de l'échantillonnage à deux degrés dans les petites régions et d'autres éléments (consulter Rao, 1999). Dans le présent article, nous nous concentrons sur le modèle élémentaire au niveau de la région par souci de simplicité et examinons certains travaux récents sur l'estimation de l'EQM et l'évaluation de la variance a posteriori.

4. MÉTHODE EMPIRIQUE DE BAYES (EB)

4.1 Estimateur de θ_i

Dans le cas du modèle de base au niveau de la région, le meilleur estimateur de θ_i, si l'on considère la réduction de l'EQM au minimum, est donné par l'espérance conditionnelle $E(\theta_i | \hat{\theta}_i)$:

$$E(\theta_i | \hat{\theta}_i) = \gamma_i \hat{\theta}_i + (1 - \gamma_i) \mathbf{z_i}^{\mathrm{T}} \boldsymbol{\beta}, \tag{4.1}$$

où $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \psi_i)$. Cet estimateur, qui est également appelé estimateur de Bayes, est représenté par $\hat{\theta}_i^B$. Il dépend des paramètres inconnus du modèle $(\boldsymbol{\beta}, \sigma_v^2)$. Si nous remplaçons $(\boldsymbol{\beta}, \sigma_v^2)$ par des estimateurs appropriés $(\hat{\boldsymbol{\beta}}, \hat{\sigma}_v^2)$, obtenus d'après la distribution marginale des $\hat{\theta}_i$, à savoir $\hat{\theta}_i \overset{ind}{\sim} N(\mathbf{z_i}^{\mathrm{T}} \boldsymbol{\beta}, \sigma_v^2 + \psi_i)$, nous obtenons l'estimateur empirique de Bayes (EB) ou meilleur estimateur empirique, $\hat{\theta}_i^{EB}$:

$$\hat{\theta}_i^{EB} = \hat{\gamma}_i \hat{\theta}_i + (1-\hat{\gamma}_i)\mathbf{z}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}\,, \tag{4.2}$$

où $\hat{\gamma}_i = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \psi_i)$. L'équation (4.2) montre que $\hat{\theta}_i^{EB}$ est une moyenne pondérée de l'estimateur direct, $\hat{\theta}_i$, et de l'estimateur synthétique par régression, $\mathbf{z}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$, pour laquelle les coefficients de pondération sont $\hat{\gamma}_i$ et $1-\hat{\gamma}_i$; $\hat{\gamma}_i$ est une mesure de la variabilité entre régions relativement à la variabilité totale associée à la région *i*. L'estimateur $\hat{\theta}_i^{EB}$ est dépourvu de biais lié au modèle au sens où $E(\hat{\theta}_i^{EB} - \theta_i) = 0$.

Pour un σ_v^2 donné, l'estimateur de $\boldsymbol{\beta}$ est l'estimateur par les moindres carrés pondérés $\tilde{\boldsymbol{\beta}}(\sigma_v^2)$ et ne nécessite aucune hypothèse de normalité. De même, si l'on n'émet pas l'hypothèse qu'il obéit à la loi normale, σ_v^2 peut être estimé par une simple méthode des moments (Prasad et Rao, 1990) ou par résolution itérative de l'équation de moment pour σ_v^2 (Fay et Herriot, 1979) :

$$h(\sigma_v^2) = \sum_i \left(\hat{\theta}_i - z_i^T \tilde{\beta}(\sigma_v^2)\right)^2 / (\sigma_v^2 + \psi_i) = m - p\,. \tag{4.3}$$

L'introduction des estimateurs $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\sigma}_v^2)$ et $\hat{\sigma}_v^2$ ainsi obtenus dans l'équation (4.1) donne l'estimateur empirique de la meilleure prédiction linéaire non biaisée (EMPLNB) qui ne dépend pas de l'hypothèse de normalité. Cet estimateur est identique à $\hat{\theta}_i^{EB}$ donné par l'équation (4.2). Dans les conditions de normalité, on pourrait utiliser les estimateurs du maximum de vraisemblance (MV) ou du maximum de vraisemblance restreint (MVR), $\hat{\boldsymbol{\beta}}$ et $\hat{\sigma}_v^2$, de $\boldsymbol{\beta}$ et σ_v^2 qui demeurent convergents même si les conditions de normalité ne sont pas respectées (Jiang, 1996). Nous considérons l'estimateur de $\bar{Y}_i$ comme étant $g^{-1}(\hat{\theta}_i^{EB})$. Des estimateurs plus complexes de $\bar{Y}_i$, à biais réduit, ont également été proposés. Par exemple, si $g(\bar{Y}_i) = \log \bar{Y}_i$ nous pourrions utiliser $\exp\left[\hat{\theta}_i^{EB} + \frac{1}{2} eqm(\hat{\theta}_i^{EB})\right]$ en supposant que $\hat{\theta}_i^{EB}$ obéit à la loi normale (ou que $\exp(\hat{\theta}_i^{EB})$ obéit à la loi lognormale), où $eqm(\hat{\theta}_i^{EB})$ est un estimateur de $EQM(\hat{\theta}_i^{EB}) = E(\hat{\theta}_i^{EB} - \theta_i)^2$; à cet égard, consulter National Research Council (2000).

4.2 Estimateurs de l'EQM

Si la distribution est normale, la distribution conditionnelle (ou a posteriori) de θ_i donnée par $\hat{\theta}_i$ est $N(\hat{\theta}_i^B, g_{1i}(\sigma_v^2))$, où

$$g_{1i}(\sigma_v^2) = \gamma_i \psi_i\,. \tag{4.4}$$

Une approche EB naïve consiste à utiliser la distribution a posteriori estimée $N(\hat{\theta}_i^{EB}, g_{1i}(\hat{\sigma}_v^2))$ pour faire une inférence au sujet de θ_i. Plus précisément, nous utilisons la moyenne de la distribution a posteriori estimée, $\hat{\theta}_i^{EB}$, comme estimateur de θ_i, et la variance de la distribution a posteriori estimée, $g_{1i}(\hat{\sigma}_v^2)$, comme mesure de la variabilité. Cette méthode peut donner lieu à une sous-estimation grave de l'EQM de $\hat{\theta}_i^{EB}$ parce que $g_{1i}(\hat{\sigma}_v^2)$ ne tient pas compte de la variabilité associée à $\hat{\boldsymbol{\beta}}$ et $\hat{\sigma}_v^2$. Dans le cas de la méthode EMPLNB, un estimateur naïf de l' $EQM(\hat{\theta}_i^{EB})$, sans hypothèse de normalité, est donné par

$$eqm_N(\hat{\theta}_i^{EB}) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2), \tag{4.5}$$

où

$$g_{2i}(\sigma_v^2) = (1-\gamma_i)^2 \mathbf{z}_i^{\mathrm{T}} \left[\sum_i \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}} / (\sigma_v^2 + \psi_i)\right]^{-1} \mathbf{z}_i\,. \tag{4.6}$$

Dans l'équation (4.5), le dernier terme $g_{2i}(\hat{\sigma}_v^2)$ représente de la variabilité de $\hat{\boldsymbol{\beta}}$ mais non celle de $\hat{\sigma}_v^2$. Notons que $eqm_N(\hat{\theta}_i^{EB})$ donne un meilleur résultat que la mesure EB naïve, $g_{1i}(\hat{\sigma}_v^2)$.

Les méthodes qui tiennent compte simultanément de la variabilité de $\hat{\beta}$ et $\hat{\sigma}_v^2$ ont fait l'objet de nombreuses études ces dernières années. Nous résumons certains travaux ici. Dans les conditions de normalité, une approximation correcte de $EQM\left(\hat{\theta}_i^{EB}\right)$ est donnée par

$$EQM\left(\hat{\theta}_i^{EB}\right) \approx g_{1i}\left(\sigma_v^2\right) + g_{2i}\left(\sigma_v^2\right) + g_{3i}\left(\sigma_v^2\right), \tag{4.7}$$

où

$$g_{3i}\left(\sigma_v^2\right) = \left[\psi_i^2 / \left(\sigma_v^2 + \psi_i\right)^4\right] E\left(\hat{\theta}_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 h\left(\sigma_v^2\right) \tag{4.8}$$

$$= \left[\psi_i^2 / \left(\sigma_v^2 + \psi_i\right)^3\right] h\left(\sigma_v^2\right) \tag{4.9}$$

et $h\left(\sigma_v^2\right)$ représente la variance asymptotique de $\hat{\sigma}_v^2$ pour une valeur élevée de m. Les termes que nous laissons tomber dans l'approximation (4.7) sont d'ordre inférieur à m^{-1}. L'approximation (4.7) est valide pour toutes les méthodes d'estimation de σ_v^2 considérées à la section 4.1.

Si nous utilisons l'estimateur du moment de Prasad-Rao (PR) pour calculer σ_v^2, nous avons

$$h_{PR}\left(\sigma_v^2\right) = 2m^{-2}\sum_i \left(\sigma_v^2 + \psi_i\right)^2 . \tag{4.10}$$

Si nous utilisons l'estimateur de Fay-Herriot (FH) pour calculer σ_v^2 la variance asymptotique est donnée par

$$h_{FH}\left(\sigma_v^2\right) = 2m\left[\sum_i \left(\sigma_v^2 + \psi_i\right)^{-1}\right]^{-2} \tag{4.11}$$

(Datta, Rao et Smith, 2001). Il découle de (4.10) et (4.11) que

$$h_{FH}\left(\sigma_v^2\right) \le h_{PR}\left(\sigma_v^2\right) \tag{4.12}$$

où l'égalité est vérifiée lorsque $\psi_i = \psi$ pour tous les *i*. Pour les estimateurs MV et MVR de σ_v^2, nous obtenons

$$h_{MV}\left(\sigma_v^2\right) = h_{MVR}\left(\sigma_v^2\right) = 2\left[\sum_i \left(\sigma_v^2 + \psi_i\right)^{-2}\right]^{-1} \le h_{FH}\left(\sigma_v^2\right). \tag{4.13}$$

Il découle de (4.12) et (4.13) que

$$h_{MV}\left(\sigma_v^2\right) = h_{MVR}\left(\sigma_v^2\right) \le h_{FH}\left(\sigma_v^2\right) \tag{4.14}$$

où l'égalité est vérifiée si $\psi_i = \psi$ pour tous les *i*. Donc, les estimateurs MV et MVR sont ceux qui produisent l'EQM la plus faible; vient ensuite l'estimateur FH.

Passons maintenant à l'estimation de l'EQM. Un estimateur correct jusqu'au même ordre d'approximation que (4.7) est donné par

$$eqm\left(\hat{\theta}_i^{EB}\right) \approx g_{1i}\left(\hat{\sigma}_v^2\right) + g_{2i}\left(\hat{\sigma}_v^2\right) + 2g_{3i}\left(\hat{\sigma}_v^2\right). \tag{4.15}$$

Cet estimateur est presque dépourvu de biais pour $EQM\left(\hat{\theta}_i^{EB}\right)$ en ce sens que son biais est d'ordre inférieur à m^{-1}. L'approximation (4.15) est valide pour les estimateurs MVR et PR de σ_v^2 mais non pour les estimateurs MV et FH.

Lahiri et Rao (1995) ont montré que, dans le cas de l'estimateur PR, l'estimateur (4.15) est robuste même si les θ_i n'obéissent pas à la loi normale, en ce sens que la quasi absence de biais demeure valide. Notons que l'hypothèse de la distribution normale des estimateurs directs θ_i est maintenue mais qu'elle est moins contraignante que l'hypothèse de normalité des θ_i à cause de l'effet du théorème central limite sur les $\hat{\theta}_i$. Nous ne savons pas si la robustesse persiste si l'on utilise l'estimateur MVR de σ_v^2.

Pour les estimateurs MV et FH, un terme supplémentaire $g_{10}\left(\hat{\sigma}_v^2\right)$ est ajouté dans (4.15). Pour l'estimateur MV, ce terme supplémentaire est positif (Datta et Lahiri, 2000). Par conséquent, si l'on ne tient pas compte de ce terme et que l'on utilise (4.15) avec l'estimateur MV de $\hat{\sigma}_v^2$, l'EQM sera sous-estimée. Par contre,

pour l'estimateur FH, le terme supplémentaire est négatif (Datta, Rao et Smith, 2001). Par conséquent, si l'on ne tient pas compte de ce terme et que l'on utilise (4.15) avec l'estimateur FH de $\hat{\sigma}_v^2$, l'EQM sera surestimée.

L'une des critiques de l'estimateur (4.15) de l'EQM et de sa modification pour les estimateurs MV ou FH tient au fait qu'il n'est pas spécifique à la région, en ce sens qu'il ne dépend pas de l'estimateur direct $\hat{\theta}_i$, même si $\mathbf{z_i}$ figure dans le terme g_{2i}. Toutefois, il est facile de faire d'autres choix, en utilisant la formule (4.8) pour $g_{3i}(\sigma_v^2)$. Par exemple, nous pouvons utiliser

$$eqm_1(\hat{\theta}_i^{EB}) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + g_{3i}(\hat{\sigma}_v^2) + [\psi_i^2/(\sigma_v^2+\psi_i)^4][\hat{\theta}_i - \mathbf{z_i^T}\hat{\boldsymbol{\beta}}]^2\, h/(\hat{\sigma}_v^2) \qquad (4.16)$$

(Rao, 2000). Le dernier terme de l'équation (4.16) est moins stable que $g_{3i}(\hat{\sigma}_v^2)$, mais il est d'ordre plus faible que le premier terme $g_{1i}(\hat{\sigma}_v^2)$. Par conséquent, la variabilité de $eqm_1(\hat{\theta}_i^{EB})$ devrait être comparable à la variabilité de $eqm(\hat{\theta}_i^{EB})$, du moins pour les valeurs moyennes à élevées de *m*.

Dans la discussion qui précède, nous avons utilisé l'EQM pour mesurer la variabilité de $\hat{\theta}_i^{EB}$. Une autre approche consiste à imiter l'approche hiérarchique de Bayes (HB) pour corriger la sous-estimation que cause l'estimation a posteriori de la distribution (Section 5). Il faut pour cela établir une distribution a priori des paramètres $\boldsymbol{\beta}$ et σ_v^2 du modèle pour obtenir la variance a posteriori $V(\theta_i|\hat{?}_i)$ qui est utilisée comme mesure de la variabilité dans le contexte de la méthode HB, où $\hat{?} = (\hat{\theta}_1,...,\hat{\theta}_m)^T$. Kass et Steffey (1989) utilisent une approximation asymptotique de première ordre de la variance a posteriori qui ne dépend pas de la forme de la distribution a priori de $\boldsymbol{\beta}$ et σ_v^2. Laird et Louis (1987) commencent par exprimer $V(\theta_i|\hat{?})$ sous la forme

$$V(\theta_i|\hat{?}) = E_{\boldsymbol{\beta},\sigma_v^2}[V(\theta_i|\hat{?}_i,\boldsymbol{\beta},\sigma_v^2)] + V_{\boldsymbol{\beta},\sigma_v^2}[E(\theta_i|\hat{?}_i,\boldsymbol{\beta},\sigma_v^2)], \qquad (4.17)$$

où $E_{\boldsymbol{\beta},\sigma_v^2}$ et $V_{\boldsymbol{\beta},\sigma_v^2}$ représentent, respectivement, l'espérance mathématique et la variance pour la distribution a posteriori $f(\beta,\sigma_v^2|\hat{?})$. Puis ils estiment séparément les deux derniers termes de (4.17) par la méthode bootstrap paramétrique (voir Ghosh et Rao, 1994). Notons que le dernier terme de (4.17) tient compte de la sous-estimation, tandis que le deuxième terme est à peu près égal à $g_{1i}(\hat{\sigma}_v^2)$, c'est-à-dire la variance de l'estimation a posteriori de la distribution de θ_i. Butar et Latiri (1997), quant à eux, montrent que le biais de l'estimateur de Laird-Louis (en tant qu'estimateur de l'EQM) est d'ordre m^{-1}, contrairement au biais des estimateurs (4.15) ou (4.16). Après correction de ce biais, ils obtiennent un estimateur de l'EQM identique à $eqm_1(\hat{\theta}_i^{EB})$ donné par (4.16). L'estimateur de premier ordre de Kass et Steffey (1989) est également biaisé. L'estimateur de deuxième ordre de Kass et Steffey dépend de la distribution a priori de $\boldsymbol{\beta}$ et σ_v^2.

4.3 Estimateur Jackknife de l'EQM

Jiang, Lahiri et Wang (1999) proposent une méthode jackknife d'estimation de l'EQM applicable aux modèles généraux dont la structure des corrélations est une matrice à diagonale par blocs, où les blocs correspondent aux petites régions. Nous illustrons la méthode du jackknife pour le modèle de base au niveau de la région. Elle se fonde sur la décomposition orthogonale de l'EQM qui suit :

$$EQM(\hat{\theta}_i^{EB}) = E(\hat{\theta}_i^{B} - \theta_i)^2 + E(\hat{\theta}_i^{EB} - \hat{\theta}_i^{B})^2$$

$$= g_{1i}(\sigma_v^2) + E\left(\hat{\theta}_i^{EB} - \hat{\theta}_i^{B}\right)^2. \tag{4.18}$$

Écrivons $\hat{\theta}_i^{EB}$ sous la forme $\hat{\theta}_i = k\left(\hat{\theta}_i, \hat{d}\right)$, où $d = \left(\boldsymbol{\beta}^T, \sigma_v^2\right)^T$. L'estimateur EB, $\hat{\theta}_i^{EB}$, peut être exprimé sous la forme $k\left(\hat{\theta}_i, \hat{\boldsymbol{\delta}}\right)$. Les étapes du jackknife sont les suivantes :

i) Calculer $\hat{d}(l) = \left(\hat{\boldsymbol{\beta}}(l), \hat{\sigma}_v^2(l)\right)$, l'estimateur de d lorsque l'on supprime les données de la $l^{\text{ième}}$ région $\left(\hat{\theta}_l, \mathbf{z}_\ell\right)$. Poser que $\hat{\theta}_i^{EB}(l) = k\left(\hat{\theta}_i, \hat{d}(l)\right)$. À noter que $\hat{\theta}_i$ ne varie pas dans $\hat{\theta}_i^{EB}(l)$.

ii) Calculer

$$M_{2i} = \tfrac{m-1}{m} \sum_{i=1}^{m} \left(\hat{\theta}_i^{EB}(l) - \hat{\theta}_i^{EB}\right)^2. \tag{4.19}$$

M_{2i} est un estimateur jackknife du dernier terme de (4.18).

iii) Ajuster le biais de $g_{1i}(\hat{\sigma}_v^2)$ (en tant qu'estimateur de $g_{1i}(\sigma_v^2)$) par la méthode jackknife de réduction du biais. L'estimateur ajusté de $g_{1i}(\sigma_v^2)$ est

$$\hat{M}_{1i} = g_{1i}(\hat{\sigma}_v^2) - \tfrac{m-1}{m} \sum_{i=1}^{m} \left[g_{1i}(\hat{\sigma}_v^2(l)) - g_{1i}(\hat{\sigma}_v^2)\right]. \tag{4.20}$$

iv) Calculer l'estimateur jackknife de l'EQM comme étant

$$eqm_J\left(\hat{\theta}_i^{EB}\right) = \hat{M}_{1i} + \hat{M}_{2i}. \tag{4.21}$$

L'estimateur jackknife de l'EQM (4.21) est approximativement non biaisé au sens où le biais est d'ordre inférieur à m^{-1}. Il est également propre à la région, comme l'estimateur (4.16). Cependant, la méthode peut être fastidieuse si elle comporte un calcul itératif, comme cela est le cas pour les estimateurs MV et MVR, parce que les estimations des paramètres $\boldsymbol{\delta}$ du modèle doivent être recalculées *m* fois en supprimant chaque région l'une après l'autre. Pour ces deux estimateurs, on peut simplifier considérablement les calculs si l'on utilise l'étape unique de l'algorithme de Newton-Raphson en prenant l'estimation de $\boldsymbol{\delta}$ d'après l'échantillon complet comme valeur de départ.

Pfeffermann et Tiller (2001) utilisent une version bootstrap de la méthode de Jiang-Lahiri dans le contexte des modèles de séries chronologiques à filtre de Kalman. Cette méthode devrait également être applicable aux modèles régionaux.

4.4 Autres estimateurs de l'EQM

Booth et Hobert (1998) soutiennent que l'EQM conditionnelle de l'estimateur EB, étant donné les données sur la $i^{\text{ième}}$ région, est une mesure de la variabilité plus pertinente que l'EQM inconditionnelle, parce que la première est particulière à la région. Fuller (1989) avait proposé antérieurement une mesure comparable dans le contexte des modèles linéaires mixtes. Cependant, l'estimateur de l'EQM (4.16) et l'estimateur jackknife (4.21) montrent qu'il est possible d'obtenir un estimateur particulier à la région de l'EQM inconditionnelle. En fait, pour le modèle de base au niveau de la région, l'estimateur (4.16) est étroitement lié à celui de EQM conditionnelle de Fuller.

Les statisticiens d'enquête préfèrent considérer l'estimation, d'après les données d'échantillon, de l'EQM de $\hat{\theta}_i^{EB}$, c.-à-d. $EQM_p\left(\hat{\theta}_i^{EB}\right) = E_p\left(\hat{\theta}_i^{EB} - \theta_i\right)^2$, où l'espérance mathématique, E_p, a trait à la distribution d'échantillon $f\left(\hat{\theta}_i \middle| \theta_i\right)$. Rivest et Belmonte (2000) calculent un estimateur non biaisé par rapport au plan de sondage, $eqm_p\left(\hat{\theta}_i^{EB}\right)$, en se servant de l'estimateur PR de σ_v^2. Dans le cas simple où l'on connaît $\boldsymbol{\beta}$ et σ_v^2, nous obtenons $\hat{\theta}_i^{EB} = \hat{\theta}_i^{B}$ et

$$eqm_p\left(\hat{\theta}_i^B\right) = \psi_i \sigma_i + \left(\frac{\psi_i}{\psi_i + \sigma_v^2}\right)^2 \left[\left(\hat{\theta}_i - z_i^T \beta\right)^2 - \psi_i - \sigma_v^2\right] \cdot \tag{4.22}$$

Par ailleurs, $eqm\left(\hat{\theta}_i^B\right) = g_{1i}\left(\sigma_v^2\right) = \psi_i \delta_i$. Rivest et Bellmonte (2000) étudient les propriétés relatives de $eqm_p\left(\hat{\theta}_i^B\right)$ et de $eqm\left(\hat{\theta}_i^B\right)$ lors de l'estimation de $EQM_p\left(\hat{\theta}_i^B\right)$ pour le cas spécial où $\psi_i = \psi$. Ils calculent le rapport de l'EQM moyenne de $eqm_p\left(\hat{\theta}_i^B\right)$ à l'EQM moyenne de $eqm_p\left(\hat{\theta}_i^B\right)$, où la moyenne est calculée sur l'ensemble des petites régions *i*. Ce rapport se réduit à $R = \left(\psi^2 + 2\psi\sigma_v^2\right)/\sigma_v^4$ qui est supérieur à 1 si $\sigma_v^2/\psi < 2.4$. Lorsque le rétrécissement est important, c.-à-d. si γ_i est petit, $eqm\left(\hat{\theta}_i^B\right)$ donne de meilleurs résultats que $eqm_p\left(\hat{\theta}_i^B\right)$. Par exemple, si $\gamma_i = \frac{1}{2}$ ou $\sigma_v^2/\psi = 1$, nous obtenons $R = 3$.

4.5 Comparaison des estimateurs de l'EQM

Datta, Rao et Smith (2001) ont réalisé une étude en simulation du biais relatif des estimateurs de l'EQM étudiés à la Section 4.2. Ils ont considéré un modèle simple : $\hat{\theta}_i \overset{ind}{\sim} N(\theta_i, \psi_i)$ et $\theta_i \overset{ind}{\sim} N(0,1), i = 1,..., m$ et $m = 15,30$, et deux profils pour ψ_i : a) variation moyenne sur i : de 0,7 à 0,3 ; b) variation importante sur i : de 4,0 à 0,1. Ils ont produit 10 000 échantillons pour chaque profil de ψ_i et m. Pour le profil a), les estimateurs de l'EQM fondé sur les estimateurs FH, MV, MVR et PR sont comparables en ce qui concerne le biais relatif, mais FH donne de bons résultats pour toutes les combinaisons de m et des profils de ψ_i. Par conséquent, il semble vraiment que l'estimateur de l'EQM fondé sur l'estimateur FH soit robuste pour les divers profils de ψ_i tandis que l'estimateur fondé sur FH, $\hat{\theta}_i^{EB}$, demeure efficace.

5. MÉTHODE HIÉRARCHIQUE DE BAYES (HB)

Un inconvénient de $\hat{\theta}_i^{EB}$ tient au fait que le coefficient de pondération $\hat{\gamma}_i$ peut prendre une valeur nulle, auquel cas $\hat{\theta}_i^{EB}$ se réduit à l'estimateur synthétique par régression $\mathbf{z}_i^T \hat{\boldsymbol{\beta}}$. Par conséquent, un coefficient de pondération nul est appliqué aux estimateurs directs, $\hat{\theta}_i$, même pour certaines régions où la taille d'échantillon n'est pas faible. Cette difficulté s'est posée lors de l'utilisation d'un modèle au niveau de l'État pour produire des estimations EB du nombre d'enfants pauvres d'âge scolaire selon l'État aux États-Unis (National Research Council, 2000). La méthode HB permet d'éviter cette difficulté, car elle produit des coefficients de pondération positifs dans tous les cas. De surcroît, la méthode HB est simple, les inférences sont « exactes » et les problèmes complexes peuvent être résolus par application des méthodes de Monte Carlo avec chaîne de Markov (MCCM) mises au point récemment. Nous précisons la distribution a priori des paramètres du modèle $\boldsymbol{\delta} = \left(\boldsymbol{\beta}^T, \sigma_v^2\right)^T$ et nous fondons les inférences sur la distribution a posteriori $f\left(\theta_i | \hat{\boldsymbol{\theta}}\right)$; plus précisément, nous estimons θ_i par la moyenne a posteriori $E\left(\theta_i | \hat{\boldsymbol{\theta}}\right)$ et nous déterminons sa précision par sa variance a posteriori $V\left(\theta_i | \hat{\boldsymbol{\theta}}\right)$ donnée par (4.17).

Nous utilisons les méthodes MCCM pour produire des échantillons simulés $\left\{\theta_1^{(j)},..., \theta_m^{(j)}; j = 1,..., J\right\}$ d'après la distribution conjointe a posteriori $f\left(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}\right)$. Pour le modèle de base au niveau de la région, nous pouvons produire facilement des échantillons simulés à partir des distributions conditionnelles $\boldsymbol{\beta} | \boldsymbol{\theta}, \sigma_v^2, \hat{\boldsymbol{\theta}}, \boldsymbol{\theta} | \boldsymbol{\beta}, \sigma_v^2, \hat{\boldsymbol{\theta}}$ et $\sigma_v^{-2} | \boldsymbol{\beta}, \boldsymbol{\theta}, \hat{\boldsymbol{\theta}}$ au moyen de l'échantillonnage de Gibbs. Nous obtenons

$$\hat{\theta}_i^{HB} = E\left(\theta_i \mid \hat{?}\right) \approx \frac{1}{J}\sum_{j=1}^{J} \theta_i^{(j)} = \theta_i^{(1)} \tag{5.1}$$

et

$$V\left(\theta_i \mid \hat{\boldsymbol{\theta}}\right) \approx \frac{1}{J}\sum_{j=1}^{J} \left(\theta_i^{(j)} - \theta_i^{(.)}\right)^2 . \tag{5.2}$$

Nous pouvons aussi obtenir des estimateurs plus efficaces (en ce qui concerne la réduction de l'erreur de simulation). Il est également facile d'obtenir l'estimation HB de $\overline{Y}_i$ et sa variance a posteriori à partir de (5.1) et (5.2) en remplaçant $\theta_i^{(j)}$ par $g^{-1}\left(\theta_i^{(j)}\right) = \overline{Y}_i^{(j)}$.

Bell (1999) applique la méthode HB au modèle au niveau de l'état susmentionné en se servant de la fonction de densité a priori impropre (diffuse) pour $\boldsymbol{\beta}$ et $\sigma_v^2 : f(\boldsymbol{\beta})$ = constante et $f\left(\sigma_v^2\right)$ = constante $\left(0 < \sigma_v^2 < \infty\right)$. Il obtient ainsi des estimations HB dont le coefficient de pondération est positif dans tous les cas.

Récemment, Youg et Rao (2002) ont appliqué la méthode HB pour traiter le cas de modèles d'échantillonnage et de liaison non concordants dans le contexte de l'estimation du sous-dénombrement au recensement au Canada. Dans cette application, c_i = dénombrement au recensement, u_i = nombre d'unités non dénombrées et y_i représente un estimateur de u_i fondé sur les données d'une enquête post-censitaire dont on connaît la variance ξ_i. Le modèle d'échantillonnage est donné par $y_i \mid u_i \overset{ind}{\sim} N\left(u_i, \xi_i\right)$ et le modèle de liaison, par $\theta_i = \log\left\{u_i / \left(u_i + c_i\right)\right\} \overset{ind}{\sim} N\left(\mathbf{z_i^T}\boldsymbol{\beta}, \sigma_v^2\right)$. Notons que le modèle de base au niveau de la région se fonde sur des modèles d'échantillonnage et de liaison concordants, mais que l'hypothèse que $E_p\left[\hat{\theta}_i\right] = \theta_i$ pourrait ne pas être vérifiée si la taille des échantillons régionaux est faible, où $\hat{\theta}_i = \log\left\{y_i / \left(y_i + c_i\right)\right\}$.

La méthode HB est intéressante, mais les méthodes MCCM doivent être utilisées avec prudence (voir, p. ex., Rao (1999)). Ainsi, l'échantillonneur de Gibbs produit parfois des inférences en apparence raisonnable au sujet d'une distribution a posteriori inexistante, lorsque cette dernière est incorrecte mais que toutes les distributions conditionnelles de Gibbs sont correctes (Hobert et Casella, 1996). Une autre difficulté que posent les méthodes MCCM tient au fait que les outils de diagnostic de la convergence ne permettent pas toujours de repérer les formes d'échec de convergence qu'ils sont censés déceler (Cowles et Carlin, 1996).

BIBLIOGRAPHIE

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

Booth, J.G., et Hobert, J.P. (1998), "Standard Errors of Predictors in Generalized Linear Mixed Models", *Journal of the American Statistical Association*, 6, pp. 363-372.

Butar, P.B., et Lahiri, P. (1997), "On the Measure of Uncertainty of Empirical Bayes Small Area Estimators", document non publié, Lincoln, Nebraska: University of Nebraska-Lincoln.

Cowles, M.K., et Carlin, B.P. (1996), "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review", *Journal of the American Statistical Association*, 91, pp. 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, pp. 613-627.

Datta, G.S, Rao, J.N.K., et Smith, D.D. (2001), "On Measures of Uncertainty of Small Area Estimates in the Fay-Herriot Model", document non publié, Athens, Georgia: University of Georgia.

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

Fuller, W.A. (1989), "Prediction of True Values for the Measurement Error Model", présenté a la Conference of Statistical Analysis of Measurement Error Models, Humboldt State University.

Ghosh, M., et Rao, J.N.K. (1994), "Small Area Estimation: an Appraisal", (avec discussion), *Statistical Science*, 9, pp. 55-93.

Gonzalez, Jr., J.F., Placek, P.J. et Scott, C. (1996), "Synthetic Estimation in Follow back Surveys at the National Center for Health Statistics", dans W.L. Schaible (ed.) *Indirect Estimators in U.S. Federal Programs*, New York: Springer, pp. 16-27.

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, pp. 1461-1479.

Jiang, J. (1996), "REML Estimation: Asymptotic Behavior et Related Topics", *Annals of Statistics*, 24, pp. 255-286.

Jiang, J., Lahiri, P., et Wan, S. (1999), "Jackknifing the Mean Squared Error of Empirical Best Predictor", document non publié, Lincoln, Nebraska: University of Nebraska-Lincoln.

Kass, R.E, et Steffey, D. (1989), "Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models)", *Journal of the American Statistical Association*, 84, pp. 717-726.

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, pp. 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, pp. 739-750.

Marker, D.A. (1995), *Small Area Estimation: A Bayesian Perspective*, thèse de doctorat non publié, Ann Arbor, Michigan: University of Michagan.

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

Pfefferman, D, et Tiller, R. (2001), "Bootstrap Approximation to Prediction MSE for state-space Models with Estimated Parameters", document non publié, Jerusalem, Israel: Hebrew University.

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

Rao, J.N.K. (1999), "Quelques progrès récents concernant l'estimation régionale fondée sur un modèle", *Techniques d'enquête*, 25, pp. 199-212.

Rao, J.N.K. (2000), "EB and EBLUP in Small Area Estimation" in S.E. Ahmed and N. Reid (eds.) *Empirical Bayes Likelihood Inference*, New York: Springer, pp. 33-43.

Rivest, L-P., et Belmonte, E. (2000), "Une secteur quadratique moyenne conditionelle des estimateurs regionaux", *Techniques d'enquête*, 26, pp. 79-90.

You, Y., et Rao, J.N.K. (2002), "Small Area Estimation Using Unmatched Sampling and Linking Models", *Canadian Journal of Statistics*, 29, in press.