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

Article

# Estimation régionale robuste au moyen d'un modèle simple à effets aléatoires


par N.G.N. Prasad et J.N.K. Rao

Juin 1999

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 1999
Vol. 25, N° 1, pp. 73-79
Statistique Canada, N° 12-001-X au catalogue



# Estimation régionale robuste au moyen d'un modèle simple à effets aléatoires

**N.G.N. Prasad et J.N.K. Rao** [1]

Résumé

Les auteurs étudient l'estimation régionale robuste en fonction d'un modèle simple à effets aléatoires constitué d'un modèle de base (ou à effets fixes) et d'un modèle de liaison qui traite les effets fixes comme des réalisations d'une variable aléatoire. À l'aide de ce modèle, on obtient un estimateur modélisé d'une moyenne régionale. Cet estimateur, qui dépend des poids d'enquête, demeure conforme au plan. On obtient également un estimateur à base de modèle de son erreur quadratique moyenne (EQM). Les résultats d'une simulation indiquent que l'estimateur proposé et l'estimateur modélisé de Kott (1989) sont également efficaces, et que l'estimateur de l'EQM proposé est souvent beaucoup plus stable que l'estimateur de l'EQM de Kott, même pour des écarts modérés du modèle de liaison. La méthode s'étend également à des modèles de régression à erreur emboîtée.

Mots clés : Conforme au plan; modèle de liaison; erreur quadratique moyenne; poids d'enquête.

## 1. Introduction

On utilise souvent, pour l'estimation régionale, des modèles à effets aléatoires au niveau de l'unité pour obtenir des estimateurs efficaces à base de modèle pour les moyennes régionales. Typiquement, ce genre d'estimateur ne fait pas appel aux poids d'enquête (par exemple Ghosh et Meeden 1986; Battese, Harter et Fuller 1988; Prasad et Rao 1990). Par conséquent, les estimateurs ne sont pas conformes au plan à moins que le plan d'échantillonnage ne soit autopondéré au sein des régions. On trouvera dans Ghosh et Rao (1994) une évaluation des méthodes d'estimation régionale.

Kott (1989) a recommandé l'utilisation d'estimateurs à base de modèles conformes au plan (par exemple des estimateurs modélisés) parce que de tels estimateurs offrent une protection contre la défaillance du modèle à mesure que la taille de l'échantillon des petites régions augmente. Il a calculé un estimateur conforme au plan pour une moyenne régionale à l'aide d'un modèle simple à effets aléatoires. Ce modèle englobe deux composantes : le modèle de base (ou à effets fixes) et le modèle de liaison. Le modèle de base est donné d'après

$$y_{ij} = \theta_i + e_{ij},\ j = 1, 2, ..., N_i;\ i = 1, 2, ..., m \qquad (1)$$

où les $y_{ij}$ sont les observations de l'échantillon et les $e_{ij}$ sont des erreurs aléatoires non corrélées de moyenne zéro et de variance $\sigma_i^2$ pour chaque petite région $i(=1, 2, ..., m)$. Par souci de simplicité, nous notons $\theta_i$ la moyenne régionale $\bar{Y}_i = \sum_j y_{ij}/N_i$, où $N_i$ est le nombre d'unités de la population de la $i^{\text{ième}}$ région. À noter que $\bar{Y}_i = \theta_i + \bar{E}_i$ et $\bar{E}_i = \sum_j e_{ij} / N_i \approx 0$ si $N_i$ est grand.

Le modèle de liaison suppose que $\theta_i$ est une réalisation d'une variable aléatoire qui satisfait au modèle

$$\theta_i = \mu + v_i \qquad (2)$$

où les $v_i$ sont des variables aléatoires non corrélées de moyenne zéro et de variance $\sigma_v^2$. De plus, $\{v_i\}$ et $\{e_{ij}\}$ sont supposés non corrélés.

En supposant que le modèle (1) reste valable pour l'échantillon $\{y_{ij},\ j = 1, 2, ..., n_i;\ i = 1, 2, ..., m\}$ et en combinant le modèle échantillon et le modèle de liaison, Kott (1989) a obtenu le modèle familier à effets aléatoires au niveau de l'unité

$$y_{ij} = \mu + v_i + e_{ij},\ j = 1, 2, ..., n_i;\ i = 1, 2, ..., m, \qquad (3)$$

appelé également modèle des composantes de la variance. On suppose habituellement des variances égales $\sigma_i^2 = \sigma^2$, même si le cas des variances d'erreur aléatoires a également été étudié (Kleffe et Rao 1992; Arora et Lahiri 1997).

En supposant $\sigma_i^2 = \sigma^2$, Kott (1989) a calculé un estimateur efficace $\hat{\theta}_{iK}$ de $\theta_i$ qui est à la fois non biaisé pour le modèle d'après (3) et conforme au plan. Il a également proposé un estimateur de son erreur quadratique moyenne (EQM) qui est non biaisé pour le modèle de base (1), tout en étant conforme au plan. Toutefois, cet estimateur de l'EQM peut être assez instable et peut même avoir des valeurs négatives, comme l'a remarqué Kott (1989) dans son exemple empirique. Kott (1989) a utilisé ses estimateurs de l'EQM surtout pour comparer la réduction globale de l'EQM lorsqu'on utilise $\hat{\theta}_{iK}$ au lieu d'un estimateur direct à base de plan $\bar{y}_{iw}$ donné d'après (4) ci-dessous. Il a fait remarquer qu'il faut des estimateurs de l'EQM plus stables.

Le présent exposé est axé surtout sur l'obtention d'un estimateur de meilleure prédiction linéaire non biaisée (MPLNB) pseudo-empirique de $\theta_i$ qui dépend des poids

1. N.G.N. Prasad, Department of Mathematical Sciences, University of Alberta, Edmonton, (Alberta), T6G 2G1; J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa (Ontario), K1S 5B6.

d'enquête et qui est conforme au plan (section 2). Un estimateur de l'EQM stable à base de modèle est également obtenu (section 3). Les résultats d'une étude de simulation à la section 4 montrent que l'estimateur de l'EQM proposé est souvent beaucoup plus stable que l'estimateur de l'EQM de Kott, mesuré selon le coefficient de variation, même pour des écarts modérés du modèle de liaison (2). Les résultats pour le modèle simple (3) s'étendent également à un modèle de régression à erreur emboîtée (section 5).

## 2. Estimateur de MPLNB pseudo-empirique

Soit $\tilde{w}_{ij}$ le poids du plan de base attribué à la $j^{\text{ième}}$ unité d'échantillonnage $(j = 1, 2, ..., n_i)$ de la $i^{\text{ième}}$ région $(i = 1, 2, ..., m)$. Un estimateur direct à base de plan de $\theta_i$ est alors donné d'après l'estimateur par quotient

$$\bar{y}_{iw} = \sum_j \tilde{w}_{ij} y_{ij} \Big/ \sum_j \tilde{w}_{ij} = \sum_j w_{ij} y_{ij} \qquad (4)$$

où $w_{ij} = \tilde{w}_{ij} / \sum_j \tilde{w}_{ij}$. L'estimateur direct $\bar{y}_{iw}$ est conforme au plan, mais il n'est pas renforcé par les autres régions.

Afin d'obtenir un estimateur plus efficace, nous examinons le modèle réduit ci-dessous obtenu du modèle combiné (3) avec $\sigma_i^2 = \sigma^2$ :

$$\begin{aligned} \bar{y}_{iw} &= \sum_j w_{ij} (\mu + v_i + e_{ij}) \\ &= \mu + v_i + \bar{e}_{iw}, \end{aligned} \qquad (5)$$

où les $\bar{e}_{iw}$ sont des variables aléatoires non corrélées de moyenne zéro et de variance $\delta_i = \sigma^2 \sum_j w_{ij}^2$. Le modèle réduit (5) est un modèle au niveau de la région qui est semblable au modèle bien connu de Fay-Herriot (Fay et Herriot 1979). Il s'ensuit d'après la théorie standard de la meilleure prédiction linéaire non biaisée (MPLNB) (voir Prasad et Rao 1990) que l'estimateur de MPLNB de $\theta_i = \mu + v_i$ pour le modèle réduit (5) est donné d'après

$$\tilde{\theta}_i = \tilde{\mu}_w + \tilde{v}_i, \qquad (6)$$

où

$$\tilde{v}_i = \gamma_{iw} (\bar{y}_{iw} - \tilde{\mu}_w)$$

avec $\tilde{\mu}_w = \sum_i \gamma_{iw} \bar{y}_{iw} / \sum_i \gamma_{iw}$ et $\gamma_{iw} = \sigma_v^2 / (\sigma_v^2 + \delta_i)$. À noter que $\tilde{\theta}_i$ est différent de l'estimateur de MPLNB pour le modèle intégral (3). Soit donc $\tilde{\theta}_i$ un estimateur pseudo-MPLNB. L'estimateur (6) s'écrit également sous forme de combinaison convexe de l'estimateur direct $\bar{y}_{iw}$ et $\tilde{\mu}_w$ :

$$\tilde{\theta}_i = \gamma_{iw} \bar{y}_{iw} + (1 - \gamma_{iw})\, \tilde{\mu}_w. \qquad (7)$$

L'estimateur $\tilde{\theta}_i$ dépend des paramètres $\sigma_v^2$ et $\sigma^2$ qui, concrètement, ne sont généralement pas connus. Par conséquent, nous remplaçons $\sigma_v^2$ et $\sigma^2$ dans l'équation (7) par des estimateurs conformes aux modèles $\hat{\sigma}_v^2$ et $\hat{\sigma}^2$ dans le cadre du modèle original au niveau de l'unité (3) de façon à obtenir l'estimateur

$$\hat{\theta}_i = \hat{\gamma}_{iw} \bar{y}_{iw} + (1 - \hat{\gamma}_{iw})\, \hat{\mu}_w, \qquad (8)$$

où

$$\hat{\gamma}_{iw} = \hat{\sigma}_v^2 \Big/ \left( \hat{\sigma}_v^2 + \hat{\sigma}^2 \sum_j w_{ij}^2 \right)$$

et

$$\hat{\mu}_w = \sum_i \hat{\gamma}_{iw} \bar{y}_{iw} \Big/ \sum_i \hat{\gamma}_{iw}.$$

Nous désignons l'estimateur $\hat{\theta}_i$ l'estimateur de MPLNB pseudo-empirique. Nous utilisons des estimateurs standard de $\sigma_v^2$ et de $\sigma^2$ en fonction des sommes des carrés intrarégion

$$Q_w = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$$

et des sommes des carrés interrégion

$$Q_b = \sum_i n_i (\bar{y}_i - \bar{y})^2,$$

où $\bar{y} = \sum_i n_i \bar{y}_i / \sum_i n_i$ est la moyenne globale de l'échantillon. Nous avons

$$\hat{\sigma}^2 = Q_w \Big/ \left( \sum_i n_i - m \right)$$

et $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$ où

$$\tilde{\sigma}_v^2 = [Q_b - (m-1)\hat{\sigma}^2] / n^*$$

avec

$$n^* = \sum_i n_i - \sum_i n_i^2 \Big/ \sum_i n_i.$$

Il est à noter que $\sigma_v^2$ et $\sigma^2$ ne peuvent pas être estimés ou sont estimés de façon imparfaite à partir du modèle réduit (5) à cause de problèmes d'identification. Suivant Kackar et Harville (1984), on peut montrer que l'estimateur de MPLNB pseudo-empirique $\theta_i$ n'est pas biaisé selon le modèle pour $\theta_i$ dans le cadre du modèle original (3) pour des erreurs $\{v_i\}$ et $\{e_{ij}\}$, réparties symétriquement, pas nécessairement normales. Il est aussi conforme au plan, à supposer que $n_i \sum_j w_{ij}^2$ soit limité à mesure que $n_i$ augmente, parce que $\hat{\gamma}_{iw}$ converge en probabilité vers 1 selon que $n_i \to \infty$ peu importe la validité du modèle (3), à supposer que $\hat{\sigma}_v^2$ et $\hat{\sigma}^2$ convergent en probabilité, par exemple vers $\hat{\sigma}_v^{*2}$ et $\hat{\sigma}^{*2}$.

On obtient l'estimateur à base de modèle de Kott (1989) de $\theta_i$ grâce à une combinaison pondérée de $\bar{y}_{iw}$ et de $\sum_{l \neq i} c_l^{(i)} \bar{y}_l$, c'est-à-dire,

$$f_i(\alpha_i, c^{(i)}) = (1 - \alpha_i)\, \bar{y}_{iw} + \alpha_i \sum_{l \neq i} c_l^{(i)} \bar{y}_l,$$

et en minimisant ensuite l'erreur quadratique moyenne (EQM) du modèle de $f_i(\alpha_i, c^{(i)})$ relativement à $\alpha_i$ et à $c_l^{(i)}$ sous réserve de l'absence de biais pour le modèle : $\sum_{l\neq i} c_l^{(i)} = 1$. Cela donne

$$\hat{\theta}_{iK} = f_i(\hat{\alpha}_i, \hat{c}^{(i)}) \tag{9}$$

avec

$$\hat{\alpha}_i = \sum_j \left[ w_{ij}^2 \Big/ \left\{ \sum_j w_{ij}^2 + \sum_{l\neq i} \hat{c}_l^{(i)^2} \Big/ n_i + \left( 1 + \sum_{l\neq i} \hat{c}_l^{(i)^2} \right) (\hat{\sigma}_v^2 / \hat{\sigma}^2) \right\} \right]$$

et

$$\hat{c}_l^{(i)} = \left[ (\hat{\sigma}_v^2 / \hat{\sigma}^2) + n_l^{-1} \right] \Big/ \sum_{h\neq i} \left[ (\hat{\sigma}_v^2 / \hat{\sigma}^2) + n_h^{-1} \right].$$

L'estimateur $\hat{\theta}_{iK}$ est également non biaisé pour le modèle et conforme au plan. Dans une version antérieure du présent exposé, nous avons proposé un estimateur semblable à (9). Il fait appel aux meilleurs estimateurs de $\mu$ dans le cadre du modèle au niveau de l'unité, d'après les moyennes non pondérées $\bar{y}_i$, plutôt que $\hat{\mu}_w$, le meilleur estimateur de $\mu$ dans le cadre du modèle réduit (4), d'après les moyennes $\bar{y}_{iw}$ pondérées en fonction de l'enquête.

## 3. Estimateurs de l'EQM

Le calcul de l'EQM de l'estimateur pseudo-MPLNB $\tilde{\theta}_i$ est simple dans le cadre du modèle au niveau de l'unité (3). Nous avons

$$\text{EQM}(\tilde{\theta}_i) = E(\tilde{\theta}_i - \theta_i)^2 = g_{1i}(\sigma_v^2, \sigma^2) + g_{2i}(\sigma_v^2, \sigma^2) \tag{10}$$

avec

$$g_{1i}(\sigma_v^2, \sigma^2) = (1 - \gamma_{iw})\, \sigma_v^2$$

et

$$g_{2i}(\sigma_v^2, \sigma^2) = \sigma_v^2 (1 - \gamma_{iw})^2 \Big/ \sum_i \gamma_{iw}.$$

Le terme dominant $g_{1i}(\sigma_v^2, \sigma^2)$ est d'ordre $O(1)$, tandis que le deuxième terme $g_{2i}(\sigma_v^2, \sigma^2)$, à cause de l'estimation de $\mu$, est d'ordre $O(m^{-1})$ pour un grand $m$.

On obtient un estimateur de l'EQM naïf de l'estimateur de MPLNB pseudo-empirique $\hat{\theta}_i$ en estimant l'EQM$(\hat{\theta}_i)$ donnée d'après (10) :

$$\text{eqm}_{\text{N}}(\hat{\theta}_i) = g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}^2). \tag{11}$$

Toutefois, (11) pourrait entraîner une nette sous-estimation de l'EQM$(\hat{\theta}_i)$ en laissant de côté l'incertitude associée à $\hat{\sigma}_v^2$ et à $\hat{\sigma}^2$. À noter que

$$\text{EQM}(\hat{\theta}_i) = \text{EQM}(\tilde{\theta}_i) + E(\hat{\theta}_i - \tilde{\theta}_i)^2 \tag{12}$$

en cas de normalité des erreurs $\{v_j\}$ et $\{e_{ij}\}$, de sorte que l'EQM$(\tilde{\theta}_i)$ est toujours plus petite que l'EQM$(\hat{\theta}_i)$; (voir Kackar et Harville 1984).

Afin d'obtenir un estimateur « correct » de l'EQM$(\hat{\theta}_i)$, nous établissons une approximation du terme de second ordre $E(\hat{\theta}_i - \tilde{\theta}_i)^2$ dans (12) pour un grand $m$, en supposant que $\{v_i\}$ et $\{e_{ij}\}$, sont distribués normalement. Suivant Prasad et Rao (1990), nous avons

$$E(\hat{\theta}_i - \tilde{\theta}_i)^2 = g_{3i}(\sigma_v^2, \sigma^2) \tag{13}$$

où les termes laissés de côté sont d'ordre inférieur à $m^{-1}$, et

$$g_{3i}(\sigma_v^2, \sigma^2) = \gamma_{iw}(1 - \gamma_{iw})^2 \sigma_v^{-2} \{V(\tilde{\sigma}_v^2) - 2(\sigma_v^2/\sigma^2)\text{Cov}(\tilde{\sigma}_v^2, \hat{\sigma}^2) + (\sigma_v^2/\sigma^2)^2 \text{Var}(\hat{\sigma}^2)\}; \tag{14}$$

(voir l'annexe 1). On trouvera aussi à l'annexe 1 les variances et covariances de $\hat{\sigma}_v^2$ et $\hat{\sigma}^2$. Il est possible de montrer que $g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}^2)$ est à peu près non biaisé pour $g_{1i}(\sigma_v^2, \sigma^2)$ du fait que son biais est d'ordre inférieur à $m^{-1}$ (voir l'annexe 2). De même, $g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}^2)$ et $g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}^2)$ sont à peu près non biaisés pour $g_{2i}(\sigma_v^2, \sigma^2)$ et $g_{3i}(\sigma_v^2, \sigma^2)$, respectivement. Il s'ensuit qu'un estimateur à peu près non biaisé pour le modèle de l'EQM$(\hat{\theta}_i)$ est donné d'après

$$\text{eqm}(\hat{\theta}_i) = g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + 2g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}^2). \tag{15}$$

Pour l'estimateur $\hat{\theta}_{iK}$ donné d'après (9), Kott (1989) a proposé comme estimateur de l'EQM

$$\text{eqm}(\hat{\theta}_{iK}) = (1 - 2\hat{\alpha}_i)\, v^*(\bar{y}_{iw}) + \hat{\alpha}_i^2 \left( \bar{y}_{iw} - \sum_{l\neq i} c_l^{(i)} \bar{y}_l \right)^2, \tag{16}$$

où $v^*(\bar{y}_{iw})$ est à la fois un estimateur conforme au plan de l'EQM pour le plan de $\bar{y}_{iw}$ et un estimateur non biaisé pour le modèle de la variance-modèle de $\bar{y}_{iw}$ dans le cadre du modèle de base (1). Puisque $\hat{\alpha}_i$ converge en probabilité vers zéro selon que $n_i \to \infty$, il s'ensuit d'après (16) que l'eqm$(\hat{\theta}_{iK})$ est aussi à la fois conforme au plan et non biaisée pour le modèle si l'on considère uniquement le modèle de base (1). Toutefois, l'eqm$(\hat{\theta}_{iK})$ est instable et peut même avoir une valeur négative lorsque $\hat{\alpha}_i$ dépasse 0,5, comme l'a fait remarquer Kott (1989).

À noter que notre estimateur de l'EQM$(\hat{\theta}_i)$, se fonde sur le modèle intégral (3) obtenu en combinant le modèle de base (1) et le modèle de liaison (2). Toutefois, les résultats de notre simulation de la section 4 indiquent qu'il peut fonctionner convenablement même en présence d'écarts modérés relativement au modèle de liaison.

## 4. Étude de simulation

Nous avons mené une étude de simulation limitée afin d'évaluer le fonctionnement de l'estimateur proposé $\hat{\theta}_i$, donné d'après (8), et de son estimateur de l'EQM, donné d'après (15), relativement à l'estimateur de Kott $\hat{\theta}_{iK}$, donné d'après (9), et à son estimateur de l'EQM, donné d'après (16). Nous avons mené l'étude en fonction de deux stratégies : i) pour chaque simulation, on établit une population finie de $m = 30$ petites régions avec $N_i = 200$ unités de population dans chaque région, en fonction du modèle proposé au niveau de l'unité, puis on tire indépendamment un échantillon avec PPT (probabilité proportionnelle à la taille) au sein de chaque petite région en fixant de $n_i = 20$; ii) on établit d'abord une population finie fixe en fonction du modèle proposé au niveau de l'unité, puis pour chaque simulation on tire indépendamment un échantillon avec PPT au sein de chaque petite région à l'aide de la population finie fixe. La première stratégie (i) se rapporte au plan aussi bien qu'au modèle de liaison, tandis que la deuxième stratégie (ii) se fonde sur le plan en ce sens qu'elle se rapporte uniquement au plan. Nous supposons que les erreurs $\{v_i\}$ et $\{e_{ij}\}$ sont distribuées normalement au moment de l'établissement des populations finies $\{y_{ij},\ i = 1, 2, ..., 30;\ j = 1, 2, ..., 200\}$. Nous avons considéré deux cas : 1) le modèle de liaison (2) est vrai pour $\mu = 50$; 2) on enfreint le modèle de liaison en laissant $\mu$ varier d'une région à l'autre : $\mu_i = 50,\ i = 1, 2, ..., 10;\ \mu_i = 55,\ i = 11, 12, ..., 20;\ \mu_i = 60,\ i = 21, 22, ..., 30$. Pour la mise en oeuvre de l'échantillonnage avec PPT dans chaque région, des mesures de la taille $z_{ij} (i = 1, 2, ..., 30;\ j = 1, 2, ..., 200)$ ont été établies à partir d'une distribution exponentielle de moyenne 200. À l'aide de ces valeurs $z$, nous avons calculé les probabilités de sélection $p_{ij} = z_{ij} / \sum_j z_{ij}$ pour chaque région $i$, puis nous les avons utilisées pour la sélection des échantillons de la taille $n_i = n$, avec PPT et avec remise, en prenant $n = 20$, et nous avons observé les valeurs d'échantillon associées $\{y_{ij}\}$.

Les poids de base du plan sont donnés d'après $\tilde{w}_{ij} = n^{-1} p_{ij}^{-1}$ de sorte que $w_{ij} = p_{ij}^{-1} / \sum_j p_{ij}^{-1}$. À l'aide de ces poids et des valeurs d'échantillon associées $y_{ij}$ nous avons calculé les estimations $\hat{\theta}_i$ et $\hat{\theta}_{iK}$ et les estimations associées de l'EQM, de même que l'estimation par quotient $\bar{y}_{iw}$ pour chaque simulation; on trouvera à l'annexe 3 la formule pour $v^*(\bar{y}_{iw})$ dans le cadre de l'échantillonnage avec PPT. Ce processus a été répété $R = 10\,000$ fois de façon à obtenir pour chaque exécution les estimations $r (= 1, 2, ..., R)$ $\hat{\theta}_i(r)$ et $\hat{\theta}_{iK}(r)$ et les estimations associées de l'$\text{eqm}_i(\hat{\theta}_i(r))$ et l'$\text{eqm}_i(\hat{\theta}_{iK}(r))$ de même que l'estimation directe $\bar{y}_{iw}(r)$. À l'aide de ces valeurs, nous avons calculé l'efficacité relative (ER) empirique de $\hat{\theta}_i$ et de $\hat{\theta}_{iK}$ pour $\bar{y}_{iw}$ sous la forme

$$\text{ER}(\hat{\theta}_i) = \text{EQM}_*(\bar{y}_{iw}) / \text{EQM}_*(\hat{\theta}_i)$$

et

$$\text{ER}(\hat{\theta}_{iK}) = \text{EQM}_*(\bar{y}_{iw}) / \text{EQM}_*(\hat{\theta}_{iK}),$$

où $\text{EQM}_*$ désigne l'EQM pour $R = 10\,000$ exécutions. Ainsi, $\text{EQM}_*(\hat{\theta}_i) = \sum_r [\hat{\theta}_i(r) - \bar{Y}_i(r)]^2 / R$, où $\bar{Y}_i(r)$ est la $i^{\text{ième}}$ moyenne de population aréolaire pour la $r^{\text{ième}}$ exécution. À noter que $\bar{Y}_i(r)$ demeure identique pour les exécutions $r$ dans le cadre de la stratégie fondée sur le plan puisque la population finie reste fixe pour les simulations.

De même, nous avons calculé le biais relatif des estimateurs de l'EQM sous la forme

$$\text{BR}[\text{eqm}(\hat{\theta}_i)] = [\text{EQM}_*(\hat{\theta}_i) - E_*\text{eqm}(\hat{\theta}_i)] / \text{EQM}_*(\hat{\theta}_i)$$

et

$$\text{BR}[\text{eqm}(\hat{\theta}_{iK})] = [\text{EQM}_*(\hat{\theta}_{iK}) - E_*\text{eqm}(\hat{\theta}_{iK})] / \text{EQM}_*(\hat{\theta}_{iK}),$$

où $E_*$ désigne l'espérance pour $R = 10\,000$ exécutions. Ainsi, $E_*\text{mse}(\hat{\theta}_i) = \sum_r \text{mse}(\hat{\theta}_i(r)) / R$. Enfin, nous avons calculé le coefficient de variation (cv) empirique des estimateurs de l'EQM sous la forme

$$\text{CV}[\text{eqm}(\hat{\theta}_i)] = [\text{EQM}_*\{\text{eqm}(\hat{\theta}_i)\}]^{1/2} / \text{EQM}_*(\hat{\theta}_i)$$

et

$$\text{CV}[\text{eqm}(\hat{\theta}_{iK})] = [\text{EQM}_*\{\text{eqm}(\hat{\theta}_{iK})\}]^{1/2} / \text{EQM}_*(\hat{\theta}_{iK}).$$

À noter que $\text{EQM}_*[\text{eqm}(\hat{\theta}_i)] = \sum_r [\text{eqm}(\hat{\theta}_i(r)) - \text{EQM}_*(\hat{\theta}_i)]^2 / R$ avec une expression semblable pour $\text{EQM}_*[\text{eqm}(\hat{\theta}_{iK})]$.

Le tableau 1 présente une mesure sommaire des valeurs ER en pourcentage, $|\text{BR}|$ et cv pour les cas (1) et (2) d'après la première stratégie. On trouvera au tableau 2 des mesures sommaires pour la deuxième stratégie. Les mesures sommaires considérées sont la moyenne et la médiane (med) pour les petites régions $i = 1, 2, ..., 30$.

Il est clair, d'après les tableaux 1 et 2, que $\hat{\theta}_{iK}$ et $\hat{\theta}_i$ fonctionnent de façon semblable relativement à une ER qui diminue à mesure que $\sigma_v / \sigma$ augmente. Dans le cadre de la deuxième stratégie, l'ER est grande à la fois pour le premier et le deuxième cas lorsque $\sigma_v / \sigma \leq 0,4$, tandis qu'elle diminue nettement dans le cadre de la première stratégie si le modèle de liaison est enfreint (deuxième cas); l'estimateur direct $\bar{y}_{iw}$ est plutôt instable dans le cadre de la deuxième stratégie.

Pour ce qui est du rendement des estimateurs de l'EQM dans le cadre de la première stratégie, le tableau 1 indique que le $|\text{BR}|$ de l'$\text{EQM}(\hat{\theta}_i)$ est négligeable ($< 4\,\%$) lorsque le modèle de liaison reste vrai (premier cas) et qu'il est petit ($< 10\,\%$) bien qu'il augmente même lorsque le modèle de liaison est enfreint. L'estimateur de l'$\text{eqm}(\hat{\theta}_{iK})$ comporte un $|\text{BR}|$ plus grand mais inférieur à 15 %. Le cv de l'$\text{eqm}(\hat{\theta}_i)$ est beaucoup plus petit que le cv de l'$\text{eqm}(\hat{\theta}_{iK})$ tant pour le premier que pour le deuxième cas. Ainsi, lorsque le modèle est vrai (premier cas), le cv médian est de 25 % pour l'$\text{eqm}(\hat{\theta}_i)$ comparativement à 148 % pour l'$\text{eqm}\,\hat{\theta}_{iK}$ lorsque $\sigma_v = 1$; le cv médian tombe à 8 % pour l'$\text{eqm}(\hat{\theta}_i)$ comparativement à 48 % pour l'$\text{eqm}(\hat{\theta}_{iK})$ lorsque $\sigma_v = 2$. Ce régime se maintient lorsque le modèle est

enfreint (deuxième cas). À noter que la probabilité d'une valeur négative de l'eqm $(\hat{\theta}_{iK})$ est assez grande $(> 0,3)$ lorsque $\sigma_v / \sigma \leq 0,4$.

**Tableau 1**
Efficacité relative (ER) des estimateurs, du biais relatif (|BR|) absolu et du coefficient de variation (CV) des estimateurs de l'EQM $(\sigma = 5,0,\ n = 20)$ : première stratégie

| $\sigma_v$ | | ER % | | \|BR\| % | | CV % | |
|---|---|---|---|---|---|---|---|
| | | $\hat{\theta}_{iK}$ | $\hat{\theta}_i$ | eqm$(\hat{\theta}_{iK})$ | eqm$(\hat{\theta}_i)$ | eqm$(\hat{\theta}_{iK})$ | eqm$(\hat{\theta}_i)$ |
| | | | | Premier cas | | | |
| 1 | Moyenne | 190 | 177 | 15,3 | 3,5 | 148 | 25 |
| | Médiane | 190 | 182 | 14,8 | 2,6 | 148 | 25 |
| 2 | Moyenne | 126 | 123 | 5,1 | 3,2 | 48 | 8 |
| | Médiane | 127 | 124 | 5,6 | 2,9 | 48 | 8 |
| 3 | Moyenne | 113 | 111 | 3,5 | 2,7 | 35 | 6 |
| | Médiane | 112 | 111 | 3,2 | 3,0 | 35 | 6 |
| | | | | Deuxième cas | | | |
| 1 | Moyenne | 108 | 103 | 10,4 | 7,9 | 39 | 6 |
| | Médiane | 108 | 104 | 11,1 | 7,7 | 38 | 5 |
| 2 | Moyenne | 108 | 104 | 13,3 | 8,9 | 39 | 6 |
| | Médiane | 108 | 104 | 13,6 | 7,9 | 37 | 6 |
| 3 | Moyenne | 104 | 103 | 11,5 | 7,2 | 37 | 5 |
| | Médiane | 105 | 105 | 13,1 | 8,0 | 36 | 6 |

Cas 1 : $\mu_i = 50,\ i = 1, 2, \ldots, 30$; Cas 2 : $\mu_i = 50,\ i = 1, 2, \ldots, 10$; $\mu_i = 55,\ i = 11, 12, \ldots, 20$; $\mu_i = 60,\ i = 21, 22, \ldots, 30$.

**Tableau 2**
Efficacité relative (ER) des estimateurs, du biais relatif (|BR|) absolu et du coefficient de variation (CV) des estimateurs de l'EQM $(\sigma = 5,0,\ n = 20)$ : deuxième stratégie

| $\sigma_v$ | | ER % | | \|BR\| % | | CV % | |
|---|---|---|---|---|---|---|---|
| | | $\hat{\theta}_{iK}$ | $\hat{\theta}_i$ | eqm$(\hat{\theta}_{iK})$ | eqm$(\hat{\theta}_i)$ | eqm$(\hat{\theta}_{iK})$ | eqm$(\hat{\theta}_i)$ |
| | | | | Premier cas | | | |
| 1 | Moyenne | 283 | 281 | 14,2 | 25,4 | 289 | 39 |
| | Médiane | 275 | 279 | 15,0 | 24,7 | 295 | 38 |
| 2 | Moyenne | 180 | 182 | 7,3 | 19,2 | 115 | 24 |
| | Médiane | 177 | 181 | 6,9 | 18,7 | 122 | 23 |
| 3 | Moyenne | 129 | 129 | 4,8 | 14,8 | 68 | 24 |
| | Médiane | 129 | 128 | 4,2 | 13,9 | 65 | 24 |
| | | | | Deuxième cas | | | |
| 1 | Moyenne | 278 | 276 | 15,7 | 26,8 | 291 | 41 |
| | Médiane | 271 | 275 | 16,6 | 26,2 | 297 | 40 |
| 2 | Moyenne | 175 | 177 | 8,8 | 20,7 | 117 | 26 |
| | Médiane | 173 | 177 | 8,5 | 20,3 | 124 | 25 |
| 3 | Moyenne | 124 | 124 | 6,3 | 16,2 | 70 | 25 |
| | Médiane | 125 | 124 | 6,8 | 15,5 | 67 | 26 |

Cas 1 : $\mu_i = 50,\ i = 1, 2, \ldots, 30$; Cas 2 : $\mu_i = 50,\ i = 1, 2, \ldots, 10$; $\mu_i = 55,\ i = 11, 12, \ldots, 20$; $\mu_i = 60,\ i = 21, 22, \ldots, 30$.

Pour ce qui est de la deuxième stratégie, le tableau 2 indique que le $|\mathrm{BR}|$ de l'eqm $(\hat{\theta}_i)$ est plus grand que la valeur relevant de la première stratégie et qu'il varie entre 15 % et 25 %. Par contre, le $|\mathrm{BR}|$ de l'eqm $(\hat{\theta}_{iK})$ est plus petit et varie entre 4 % et 15 %. Le cv de l'eqm $(\hat{\theta}_{iK})$, par contre, est beaucoup plus grand que pour la première stratégie. Ainsi, le cv médian pour le premier cas est de 295 % comparativement à 38 % pour l'eqm $(\hat{\theta}_i)$ lorsque $\sigma_v = 1$ avec diminution à 122 % comparativement à 23 % lorsque $\sigma_v = 2$. Un régime semblable est observé pour le deuxième cas, la population finie fixe étant établie en fonction du modèle avec des moyennes qui varient.

Afin de réduire le $|\mathrm{BR}|$ de l'eqm $(\hat{\theta}_i)$ dans le cadre de la deuxième stratégie, on pourrait le combiner à l'eqm $(\hat{\theta}_{iK})$ en établissant une moyenne pondérée, mais le choix des poids appropriés semble difficile. La moyenne pondérée sera plus stable que l'eqm $(\hat{\theta}_{iK})$.

## 5. Modèle de régression à erreur emboîtée

On peut étendre les résultats des sections 2 et 3 aux modèles de régression à erreur emboîtée

$$y_{ij} = x'_{ij}\beta + v_i + e_{ij},\ \ j = 1, 2, \ldots, n_i;\ \ i = 1, 2, \ldots, m \quad (17)$$

en utilisant les résultats de Prasad et Rao (1990), où $x_{ij}$ est un vecteur $p$ des variables auxiliaires de moyenne de population connue $\bar{X}_i$ et liées à $y_{ij}$, et $\beta$ est le vecteur $p$ des coefficients de régression. Le modèle réduit est donné d'après

$$\bar{y}_{iw} = \bar{x}'_{iw}\beta + v_i + \bar{e}_{iw} \quad (18)$$

avec $\bar{x}'_{iw} = \sum_j w_{ij}\, x_{ij}$. On obtient des estimations conformes au modèle $\hat{\sigma}_v^2$ et $\hat{\sigma}^2$ du modèle au niveau de l'unité (17), à l'aide de la méthode de l'ajustement des constantes (Prasad et Rao 1990) ou encore de l'estimation PMR (probabilité maximale restreinte) (Datta et Lahiri 1997).

La MPLNB pseudo-empirique de $\theta_i = \bar{X}'_i\beta + v_i$ est donnée d'après

$$\hat{\theta}_i = \hat{\gamma}_{iw}\bar{y}_{iw} + (1 + \hat{\gamma}_{iw})\bar{X}'_i\hat{\beta}_w, \quad (19)$$

où

$$\hat{\beta}_w = \left(\sum_i \hat{\gamma}_{iw}\bar{x}_{iw}\bar{x}'_{iw}\right)^{-1}\left(\sum_i \hat{\gamma}_{iw}\bar{x}_{iw}\bar{x}_{iw}\right).$$

Un estimateur approximatif non biaisé pour le modèle de l'EQM $(\hat{\theta}_i)$ est donné d'après (15) avec

$$g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) = (1 - \hat{\gamma}_{iw})\hat{\sigma}_v^2$$

comme auparavant,

$$g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}^2) =$$

$$\hat{\sigma}_v^2(\bar{X}_i - \hat{\gamma}_{iw}\bar{x}_{iw})'\left(\sum_i \hat{\gamma}_{iw}\,\bar{x}_{iw}\bar{x}'_{iw}\right)^{-1}(\bar{X}_i - \hat{\gamma}_{iw}\,\bar{x}_{iw}\,\bar{x}_{iw})$$

et $g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}^2)$, que l'on obtient de (14), comportant les variances et covariances estimatives de $\tilde{\sigma}_v^2$ et de $\hat{\sigma}^2$. On peut obtenir ce dernier élément de Prasad et Rao (1990) pour la méthode de l'ajustement des constantes et de Datta et Lahiri (1997) pour la probabilité maximale restreinte.

## 6. Conclusion

Nous avons proposé un estimateur modélisé d'une moyenne régionale dans le cadre d'un modèle simple à effets aléatoires au niveau de l'unité. Cet estimateur dépend des poids d'enquête et il est conforme au plan. Nous avons également obtenu un estimateur à base de modèle de l'EQM. Les résultats de notre étude de simulation ont indiqué que l'estimateur de l'EQM proposé fonctionne bien, même pour des écarts modérés du modèle de liaison. La stratégie proposée s'étend aussi à un modèle de régression à erreur emboîtée.

## Remerciements

Les auteurs ont reçu des subventions du Conseil de recherches en sciences naturelles et en génie du Canada. Nous remercions le rédacteur adjoint et l'examinateur de leurs remarques et suggestions constructives.

## Annexe 1

**Preuve de (13) :**

D'après les résultats généraux (Prasad et Rao 1990), nous avons

$$E(\hat{\theta}_i - \tilde{\theta}_i)^2 \approx tr[A_i(\sigma_v^2, \sigma^2)B_i(\sigma_v^2, \sigma^2)],$$

où $B_i(\sigma_v^2, \sigma^2)$ est la matrice de covariance 2 × 2 de $\tilde{\sigma}_v^2$ et de $\hat{\sigma}^2$, et $A_i(\sigma_v^2, \sigma^2)$ est la matrice de covariance 2 × 2 de

$$\left(\frac{\partial\theta_i^*}{\partial\sigma_v^2}, \frac{\partial\theta_i^*}{\partial\sigma^2}\right).$$

Or en notant que

$$\frac{\partial\theta_i^*}{\partial\sigma_v^2} = \frac{\partial\gamma_{iw}}{\partial\sigma_v^2}\bar{y}_{iw} = \left[\frac{\gamma_{iw}(1-\gamma_{iw})}{\sigma_v^2}\right]\bar{y}_{iw},$$

$$\frac{\partial\theta_i^*}{\partial\sigma^2} = \frac{\partial\gamma_{iw}}{\partial\sigma^2}\bar{y}_{iw} = -\left[\frac{\gamma_{iw}(1-\gamma_{iw})}{\sigma^2}\right]\bar{y}_{iw},$$

et $V(\bar{y}_{iw}) = \sigma_v^2 + \delta_i = \sigma_v^2/\gamma_{iw}$, nous avons

$$A_i(\sigma_v^2, \sigma^2) = [\gamma_{iw}(1-\gamma_{iw})^2\sigma_v^{-2}]\begin{bmatrix} 1 & -\sigma_v^2/\sigma^2 \\ -\sigma_v^2/\sigma^2 & (\sigma_v^2/\sigma)^2 \end{bmatrix},$$

et par conséquent le résultat (14).

**Matrice de covariance de $\tilde{\sigma}_v^2$ et de $\hat{\sigma}^2$ :**

En cas de normalité, nous avons

$V(\hat{\sigma}^2) = 2\sigma^4 / \left(\sum_i n_i - m\right)$,

$V(\tilde{\sigma}_v^2) = 2n_*^{-2}$

$$\left[\sigma^4(m-1)\left(\sum n_i - 1\right)\left(\sum n_i - m\right)^{-1} + 2n_*\sigma^2\sigma_v^2 + n_{**}\sigma_v^4\right]$$

et

$$\mathrm{Cov}(\hat{\sigma}^2, \tilde{\sigma}_v^2) = -(m-1)n_*^{-1}V(\hat{\sigma}^2),$$

où

$$n_{**} = \sum n_i^2 - 2\sum n_i^3 / \sum n_i + \left(\sum n_i^2\right)^2 / \left(\sum n_i\right)^2;$$

(voir Searle, Cassella et McCulloch 1992, page 428).

## Annexe 2

**Preuve de $E[g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + g_{zi}(\hat{\sigma}_v^2, \hat{\sigma}^2)] \approx g_{1i}(\sigma_v^2, \sigma^2)$ :**

Grâce à un développement de Taylor de $g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2)$ autour de $(\sigma_v^2, \sigma^2)$ au second ordre et en notant que $E(\hat{\sigma}^2 - \sigma^2) = 0$ et $E(\hat{\sigma}_v^2 - \sigma_v^2) \approx 0$, nous obtenons

$$E[g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) - g_{1i}(\sigma_v^2, \sigma^2)] \approx \frac{1}{2}tr[D_i(\sigma_v^2, \sigma^2)B_i(\sigma_v^2, \sigma^2)],$$

où $D_i(\sigma_v^2, \sigma^2)$ est la matrice 2 × 2 des dérivées secondes de $g_{1i}(\sigma_v^2, \sigma^2)$ relativement à $\sigma_v^2$ et à $\sigma^2$. Il est facile de vérifier que

$$\frac{1}{2}tr[D_i(\sigma_v^2, \sigma^2)B_i(\sigma_v^2, \sigma^2)] = g_{3i}(\sigma_v^2, \sigma^2).$$

Or en notant que $E[g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}^2)] \approx g_{3i}(\sigma_v^2, \sigma^2)$ nous obtenons le résultat souhaité.

## Annexe 3

L'estimateur à base de plan de la variance de $\bar{y}_{iw}$ dans le cadre d'un échantillonnage avec PPT est donné d'après

$$v(\bar{y}_{iw}) = \frac{m}{m-1}\sum_j w_{ij}^2(y_{ij} - \bar{y}_{iw})^2.$$

L'estimateur modélisé de la variance de Kott (1989) est

$$v^*(\bar{y}_{iw}) = \{V(\bar{y}_{iw})/Ev(\bar{y}_{iw})\}v(\bar{y}_{iw}) = \frac{\left(\sum_j w_{ij}^2\right)\sum_j w_{ij}^2(y_{ij} - \bar{y}_{iw})^2}{\sum_j w_{ij}^2\left(1 - 2w_{ij} + \sum_j w_{ij}^2\right)},$$

où $E$ et $V$ désignent l'espérance et la variance pour ce qui est du modèle au niveau de l'unité (3).

## Bibliographie

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

Battese, G.E., Harter, R. et Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Datta, G.S., et Lahiri, P. (1997). A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictor in Small-Area Estimation Problems. Rapport technique, University of Nebraska-Lincoln.

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

Ghosh, M., et Meeden, G. (1986). Empirical Bayes estimation in finite population sampling. *Journal of the American Statistical Association*, 81, 1058-1069.

Ghosh, M., et Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

Kackar, R.N., et Harville, D.A. (1984). Approximations for standard errors of estimators for fixed and random effects models. *Journal of the American Statistical Association*, 79, 853-862.

Kleffe, J., et Rao, J.N.K. (1992). Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model. *Journal of Multivariate Analysis*, 43, 1-15.

Kott, P. (1989). Estimation robuste pour petits domaines à l'aide du modèle des effets aléatoires. *Techniques d'enquête*, 15, 3-13.

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, 163-171.

Searle, S.R., Casella, G. et McCulloch, C.E. (1992). *Variance Components*. New York : John Wiley & Sons, Inc.