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

Article

Estimation pour petits domaines sous une contrainte

par Junyuan Wang, Wayne A. Fuller et Yongming Qu

Juin 2008

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2008
Vol. 34, N° 1, pp. 33-40
Statistique Canada, N° 12-001-X au catalogue

Estimation pour petits domaines sous une contrainte

Junyuan Wang, Wayne A. Fuller et Yongming Qu [1]

Résumé

La prédiction sur petits domaines fondée sur des effets aléatoires, appelée (MPLSBE), est une méthode de construction d'estimations pour de petites régions géographiques ou de petites sous-populations en utilisant les données d'enquête existantes. Souvent, le total des prédicteurs sur petits domaines est forcé d'être égal à l'estimation par sondage directe et ces prédicteurs sont alors dits calés. Nous passons en revue plusieurs prédicteurs calés et présentons un critère qui unifie leur calcul. Nous dérivons celui qui est l'unique meilleur prédicteur linéaire sans biais sous ce critère et discutons de l'erreur quadratique moyenne des prédicteurs calés. L'imposition de la contrainte comporte implicitement la possibilité que le modèle de petit domaine soit spécifié incorrectement et que les prédicteurs présentent un biais. Nous étudions des modèles augmentés contenant une variable explicative supplémentaire pour lesquels les prédicteurs sur petits domaines ordinaires présentent la propriété d'autocalage. Nous démontrons à l'aide de simulations que les prédicteurs calés ont un biais un peu plus faible que le prédicteur MPLSBE habituel. Cependant, si le biais est une préoccupation, une meilleure approche consiste à utiliser un modèle augmenté contenant une variable auxiliaire supplémentaire qui est fonction de la taille du domaine. Dans les simulations, les prédicteurs fondés sur le modèle augmenté ont une EQM plus petite que MPLSBE quand le modèle incorrect est utilisé pour la prédiction. De surcroît, l'EQM augmente très légèrement comparativement à celle de MPLSBE si la variable auxiliaire est ajoutée au modèle correct.

Mots clés : Modèle des composantes de la variance; meilleure prédiction linéaire sans biais; calage; convergence sous le plan; modèles linéaires mixtes.

1. Introduction

Dans certaines situations, il est souhaitable d'établir d'après les données d'enquête existantes des estimateurs fiables pour de petites régions géographiques ou de petites sous-populations. Cependant, les tailles des échantillons de domaine peuvent être telles que les estimateurs habituels produisent des erreurs-types inacceptablement grandes et il est alors raisonnable d'utiliser un estimateur fondé sur un modèle. Voir Rao (2003) pour une discussion complète de l'estimation pour petits domaines.

Un modèle pour l'estimation pour petits domaines peut s'écrire

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + b_i, \tag{1}$$

$$Y_i = y_i + e_i, \qquad i = 1, \ldots, n, \tag{2}$$

où les y_i sont les moyennes inobservables de petits domaines, les Y_i sont les estimateurs d'après les données d'enquête observables, les $\boldsymbol{x}_i'$ sont des vecteurs connus, $\boldsymbol{\beta}$ est le vecteur des paramètres de régression, les b_i sont des variables aléatoires indépendantes et de même loi telles que $E(b_i) = 0$ et $V(b_i) = \sigma_b^2$, et les e_i sont les erreurs d'échantillonnage telles que $E(e_i \mid y_i) = 0$ et $V(e_i \mid y_i) = \sigma_{ei}^2$. En combinant (1) et (2), nous obtenons

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + b_i + e_i, \qquad i = 1, \ldots, n, \tag{3}$$

qui est un cas particulier du modèle linéaire mixte.

En supposant que les composantes de la variante σ_b^2 et σ_{ei}^2 sont connues, le meilleur estimateur linéaire sans biais de $\boldsymbol{\beta}$ est

$$\hat{\boldsymbol{\beta}} = \left[\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right]^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$$

$$= \left[\sum_{i=1}^{n}(\sigma_b^2+\sigma_{ei}^2)^{-1}\boldsymbol{x}_i\,\boldsymbol{x}_i'\right]^{-1}\left[\sum_{i=1}^{n}(\sigma_b^2+\sigma_{ei}^2)^{-1}\boldsymbol{x}_i\,Y_i\right], \tag{4}$$

où $\boldsymbol{X}' = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{Y}' = (Y_1, \ldots, Y_n)$ et $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{Y}) = \mathrm{diag}(\sigma_b^2 + \sigma_{e1}^2, \ldots, \sigma_b^2 + \sigma_{en}^2)$. En outre, le meilleur prédicteur linéaire sans biais (MPLSB) de y_i est

$$\tilde{y}_i^H = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + \gamma_i(Y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}), \tag{5}$$

où

$$\gamma_i = (\sigma_b^2 + \sigma_{ei}^2)^{-1}\sigma_b^2. \tag{6}$$

Voir Henderson (1963) et Rao (2003). Quand les composantes de la variance sont inconnues, nous les remplaçons dans (4) et (6) par des estimateurs pour obtenir $\hat{y}_i^H$, c'est-à-dire le MPLSB empirique ou MPLSBE.

Souvent, l'estimateur direct du total de tous les domaines étudiés est considéré comme étant d'une précision adéquate. Si cela est le cas, le praticien pourrait choisir d'utiliser l'estimateur du total convergent sous le plan et exiger que la somme pondérée des prédicteurs sur petits domaines soit égale à cet estimateur. Donc, il est désirable que les prédicteurs sur petits domaines $\hat{y}_i$ satisfassent

1. Junyuan Wang, Directeur de Biostatistics, Global Biostatistics and Programming, Wyeth Research, 35 Cambridge Park Drive, Cambridge, MA 02140; Wayne A. Fuller, Professeur, Department of Statistics, Iowa State University, Ames, IA 50011; Yongming Qu, chercheur scientifique principal, Eli Lilly and Company, Lilly Corporation Center, Indianapolis, IN 26285.

$$\sum_{i=1}^{n} \omega_i \hat{y}_i = \sum_{i=1}^{n} \omega_i Y_i, \tag{7}$$

où les ω_i sont les poids d'échantillonnage, tels que $\sum_{i=1}^{n} \omega_i Y_i$ est un estimateur du total (ou de la moyenne) convergent sous le plan. Un certain nombre de méthodes, souvent appelées « étalonnage » ou « calage », ont été proposées pour construire des prédicteurs qui satisfont (7). Voir, par exemple, Mantel, Singh et Barcau (1993) et You et Rao (2003).

Afin de passer ce genre de méthodes en revue, représentons par $\tilde{\boldsymbol{y}}^H = (\tilde{y}_1^H, ..., \tilde{y}_n^H)'$ le prédicteur MPLSB de $\boldsymbol{y} = (y_1, ..., y_n)'$ défini en (5), où

$$\tilde{\boldsymbol{y}}^H = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{b}}, \tag{8}$$

$\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{b}}$ représentant toute solution de

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1} \\ \boldsymbol{\Sigma}_e^{-1}\boldsymbol{X} & \boldsymbol{\Sigma}_e^{-1} + \boldsymbol{\Sigma}_b^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1}\mathbf{Y} \\ \boldsymbol{\Sigma}_e^{-1}\mathbf{Y} \end{bmatrix}, \tag{9}$$

$\boldsymbol{\Sigma}_b = \sigma_b^2 \boldsymbol{I}_n$, et $\boldsymbol{\Sigma}_e = \text{diag}(\sigma_{e1}^2, ..., \sigma_{en}^2)$. L'équation (9) est appelée l'équation du modèle mixte. Trouver une solution de cette équation équivaut à trouver une solution au problème de minimisation

$$\min_{\boldsymbol{\beta},\boldsymbol{b}} \{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{b})'\boldsymbol{\Sigma}_e^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{b}) + \boldsymbol{b}'\boldsymbol{\Sigma}_b^{-1}\boldsymbol{b}\}. \tag{10}$$

Pfeffermann et Barnard (1991) ont proposé le prédicteur modifié

$$\hat{\boldsymbol{y}}^{\text{PB}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{PB}} + \hat{\boldsymbol{b}}^{\text{PB}}, \tag{11}$$

où $\hat{\boldsymbol{\beta}}^{\text{PB}}$ et $\hat{\boldsymbol{b}}^{\text{PB}}$ représentent toute solution du problème de minimisation (10) avec $\boldsymbol{\beta}$ et $\boldsymbol{b}$ soumis à la contrainte

$$\sum_{i=1}^{n} \omega_i(\boldsymbol{x}_i'\boldsymbol{\beta} + b_i) = \sum_{i=1}^{n} \omega_i Y_i. \tag{12}$$

Cela mène au prédicteur

$$\hat{y}_i^{\text{PB}} = \tilde{y}_i^H + [\text{Var}(\tilde{y})]^{-1}\text{cov}(\tilde{y}_i^H, \tilde{y})\left[\sum_{j=1}^{n} \omega_j Y_j - \tilde{y}\right], \tag{13}$$

où $\tilde{y} = \sum_{i=1}^{n} \omega_i \tilde{y}_i^H$, $\text{cov}(\tilde{y}_i^H, \tilde{y}) = \omega_i \gamma_i \sigma_{ei}^2 + \sum_{j=1}^{n} \omega_j (1 - \gamma_i)(1 - \gamma_j) \boldsymbol{x}_i' V(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_j$, et $\text{Var}(\tilde{y}) = \sum_{i=1}^{n} \omega_i \text{cov}(\tilde{y}_i^H, \tilde{y})$.

Isaki, Tsay et Fuller (2000) ont imposé la contrainte par une méthode qui, approximativement, consiste à construire les meilleurs prédicteurs de $n - 1$ quantités qui ne sont pas corrélées avec $\sum_{i=1}^{n} \omega_i Y_i$. Après certaines opérations matricielles, le prédicteur d'Isaki-Tsay-Fuller (ITF) peut être réécrit sous la forme

$$\hat{y}_i^{\text{ITF}} = \hat{y}_i^H + \left[\sum_{j=1}^{n} \omega_j^2 \widehat{\text{Var}}(Y_j)\right]^{-1} \omega_i \widehat{\text{Var}}(Y_i)\left(\sum_{j=1}^{n} \omega_j Y_j - \sum_{j=1}^{n} \omega_j \hat{y}_j^H\right), \tag{14}$$

où $\widehat{\text{Var}}(Y_i)$ est un estimateur de $\sigma_b^2 + \sigma_{ei}^2$.

Soulignons que le prédicteur de Pfeffermann-Barnard (PB) (13) et le prédicteur ITF (14) sont de la forme

$$\hat{y}_i^{\boldsymbol{a}} = \hat{y}_i + a_i\left(\sum_{j=1}^{n} \omega_j Y_j - \sum_{j=1}^{n} \omega_j \hat{y}_j\right), \tag{15}$$

où $\sum_{i=1}^{n} \omega_i a_i = 1$. Autrement dit, nous pouvons considérer qu'imposer la contrainte (7) revient à un problème d'ajustement. Pour qu'un prédicteur ajusté $\hat{y}_i^{\boldsymbol{a}}$ satisfasse (7), nous allouons la différence $\sum_{j=1}^{n} \omega_j Y_j - \sum_{j=1}^{n} \omega_j \hat{y}_j$ aux prédicteurs sur petits domaines $\hat{y}_i$ en utilisant a_i.

À l'aide du modèle au niveau des unités, You et Rao (2002) ont proposé un estimateur de $\boldsymbol{\beta}$ tel que les prédicteurs résultants satisfont (7). Ils ont dit que ces prédicteurs étaient autocalés. En appliquant leur méthode au modèle au niveau du domaine (3), nous obtenons

$$\hat{y}_i^{\text{YR}} = \hat{\gamma}_i Y_i + (1 - \hat{\gamma}_i)\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\text{YR}}, \tag{16}$$

où

$$\hat{\boldsymbol{\beta}}_{\text{YR}} = \left[\sum_{i=1}^{n} \omega_i(1 - \hat{\gamma}_i)\boldsymbol{x}_i\boldsymbol{x}_i'\right]^{-1} \sum_{i=1}^{n} \omega_i(1 - \hat{\gamma}_i)\boldsymbol{x}_i Y_i. \tag{17}$$

Tout prédicteur qui a la propriété d'autocalage, tel que le prédicteur (16) de You et Rao (YR), est un prédicteur de la forme (15), puisque la différence $\sum_{j=1}^{n} \omega_j Y_j - \sum_{j=1}^{n} \omega_j \hat{y}_j$ est égale à zéro.

À la section 2, nous dérivons le « meilleur » prédicteur de la forme (15). Les résultats conduisent à une vision unifiée de plusieurs prédicteurs fondés sur le MPLSB. À la section 3, nous proposons une autre approche qui possède la propriété d'autocalage. À la section 4, nous décrivons brièvement l'erreur quadratique moyenne (EQM) et à la section 5, nous recourons à des études par simulation pour comparer les prédicteurs. À la section 6, nous présentons nos conclusions et une discussion.

2. « Meilleur » prédicteur linéaire sans biais sous une contrainte

Afin de trouver le « meilleur » prédicteur linéaire sans biais de $\boldsymbol{y}$ qui satisfait la contrainte (7), nous commençons par supposer que les paramètres pour les composantes de la variance sont connus. Selon le lemme 1 de Pfeffermann et Barnard (1991), il est impossible de comparer, composante par composante, des prédicteurs qui satisfont la contrainte (7) pour trouver le meilleur d'entre eux. Par conséquent, un certain critère global est requis. L'un des critères naturels est

$$Q(\hat{\boldsymbol{y}}^{\boldsymbol{a}}) = \sum_{i=1}^{n} \varphi_i E(\hat{y}_i^{\boldsymbol{a}} - y_i)^2, \tag{18}$$

où les φ_i, $i = 1, ..., n$ sont un ensemble choisi de poids positifs.

Théorème 1. *Soit le modèle à effets aléatoires*

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + b_i + e_i, \qquad i = 1, ..., n,$$

où les b_i *sont indépendants et de même loi de moyenne nulle et de variance* σ_b^2, *les* e_i *sont indépendants et de loi de moyenne nulle et de variance* σ_{ei}^2, *et* $\boldsymbol{b} = (b_1, ..., b_n)'$ *est indépendant de* $\boldsymbol{e} = (e_1, ..., e_n)'$. *Supposons que* σ_b^2 *et* σ_{ei}^2 *sont connues, et que* $\boldsymbol{\beta}$ *est inconnu. Soit* $\tilde{y}_i^H$ *le MPLSB de* y_i *défini en* (5). *Soit*

$$\hat{y}_i^{\breve{\boldsymbol{a}}} = \tilde{y}_i^H + \breve{a}_i\left(\sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j \tilde{y}_j^H\right), \qquad (19)$$

où $\breve{a}_i = (\sum_{i=1}^n \varphi_i^{-1}\omega_i^2)^{-1}\varphi_i^{-1}\omega_i$ *et* ω_i *sont les poids fixes de* (7). *Alors* $\hat{\boldsymbol{y}}^{\breve{\boldsymbol{a}}} = (\hat{y}_1^{\breve{\boldsymbol{a}}}, ..., \hat{y}_n^{\breve{\boldsymbol{a}}})'$ *est, parmi tous les prédicteurs linéaires sans biais, le seul qui satisfait* (7) *et minimise le critère* (18).

Preuve : Voir l'annexe A.

Remarque 1. Quand les composantes de la variance sont inconnues, nous les remplaçons dans (6) par des estimateurs appropriés pour obtenir le MPLSB empirique, ou MPLSBE, dénoté par $\hat{y}_i^H$. Donc, nous obtenons le prédicteur modifié

$$\hat{y}_i^{\breve{\boldsymbol{a}}} = \hat{y}_i^H + \breve{a}_i\left(\sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j \hat{y}_j^H\right). \qquad (20)$$

Remarque 2. Le critère (18) définit une « fonction de perte », où le choix des poids φ_i dépend du problème étudié. Par exemple, un statisticien peut décider d'attribuer des poids plus élevés aux domaines « plus importants » et des poids plus faibles aux domaines « moins importants ». Souvent, φ_i est une fonction des composantes de la variance. Dans certains cas, on peut choisir φ_i de sorte que les prédicteurs dérivés possèdent certaines propriétés souhaitables. Par exemple, $\varphi_i = [\widehat{\mathrm{Var}}(Y_i)]^{-1}$ donne les prédicteurs ITF, qui sont les prédicteurs MPLSB (au sens classique) dans le sous-espace qui est orthogonal à $(\omega_1, ..., \omega_n)'$ dans l'espace couvert par $\boldsymbol{Y}$.

Remarque 3. Si $\varphi_i = \omega_i[\mathrm{cov}(\tilde{y}_i^H, \tilde{y})]^{-1}$, où $\tilde{y} = \sum_{j=1}^n \omega_j \tilde{y}_j^H$, nous obtenons le prédicteur (13) dérivé par Pfeffermann et Barnard (1991). Quand $\varphi_i = [\widehat{\mathrm{Var}}(\hat{y}_j^H)]^{-1}$, nous obtenons celui utilisé par Battese, Harter et Fuller (1988).

3. Un autre moyen d'imposer la contrainte

Nous avons discuté d'une famille de prédicteurs pour lesquels le total des prédicteurs sur petits domaines est égal au total des estimations d'enquête directes. L'imposition de la contrainte (7) comporte implicitement la possibilité que le prédicteur sur petits domaines du total présente un biais parce que le modèle (3) est mal spécifié. Dans les applications pratiques, l'erreur de spécification du modèle est une préoccupation valide, car le mécanisme réel de production de $\boldsymbol{Y}$ est inconnu.

Une erreur de spécification courante survient quand les variables explicatives utilisées dans le modèle ne sont pas les mêmes que celles qui ont produit $\boldsymbol{Y}$. Donc, la direction du biais global pourrait ne pas être la même que celle du biais pour un petit domaine particulier. Le cas échéant, les prédicteurs de la forme (15) pourraient accroître le biais pour certains petits domaines comparativement au biais avant ajustement. Mantel et coll. (1993) ont conclu qu'en général, l'effet de l'étalonnage est une légère amélioration du biais global au prix d'une certaine détérioration des autres mesures d'évaluation.

Puisque le biais n'est pas nul s'il existe une corrélation non nulle entre ω_i et $(Y_i - \hat{y}_i)$, on peut le réduire en incluant ω_i dans le modèle. Autrement dit, pour un modèle donné, une approche consiste à utiliser le modèle augmenté

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{e}, \qquad (21)$$

où $\boldsymbol{X}_1 = (\boldsymbol{X}, \boldsymbol{\omega})$ et $\boldsymbol{\omega} = (\omega_1, ..., \omega_n)'$ pour obtenir le MPLSB ou MPLSBE. Quand $\boldsymbol{\omega}$ est dans le modèle, l'ajustement nécessaire pour satisfaire la contrainte (7) sera souvent beaucoup plus faible que celui requis pour le modèle sans $\boldsymbol{\omega}$.

Grâce à l'approche du modèle augmenté, nous pouvons aller une étape plus loin et construire des prédicteurs qui satisfont la contrainte (7). En premier lieu, supposons que les variances σ_{ei}^2 sont connues. Notons que

$$\sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j \hat{y}_j^H = \sum_{i=1}^n \omega_i(1-\gamma_i)(Y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) \qquad (22)$$

et $\omega_i(1-\gamma_i)\mathrm{Var}(Y_i) = \omega_i\sigma_{ei}^2$. En appliquant la théorie du modèle linéaire, nous pouvons montrer que le prédicteur construit avec le modèle augmenté

$$\boldsymbol{Y} = \boldsymbol{X}_2\boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{e}, \qquad (23)$$

où $\boldsymbol{X}_2 = (\boldsymbol{X}, \boldsymbol{\omega}_e)$ et $\boldsymbol{\omega}_e = (\omega_1\sigma_{e1}^2, ..., \omega_n\sigma_{en}^2)'$ possède la propriété d'autocalage quand nous utilisons l'estimateur par les moindres carrés généralisés (MCG) de $\boldsymbol{\beta}$. Notons que cette approche donne un prédicteur différent de celui de You-Rao (16).

Si les σ_{ei}^2 sont inconnues, nous remplaçons σ_{ei}^2 dans $\boldsymbol{\omega}_e$ par son estimateur $\hat{\sigma}_{ei}^2$. À condition que les $\hat{\sigma}_{ei}^2$ dans $\boldsymbol{\omega}_e$ soient les mêmes que les $\hat{\sigma}_{ei}^2$ utilisés pour construire γ_i, les prédicteurs ont la propriété d'autocalage. Si les σ_{ei}^2 sont de la forme $\sigma_e^2 f(\boldsymbol{u}_i)$, où σ_e^2 est inconnue, mais que $\boldsymbol{u}_i$ et $f(.)$

sont connues, nous pouvons construire les variables pour le modèle (23) en utilisant $\boldsymbol{\omega}_e = (\omega_1 f(\boldsymbol{u}_1), \ldots, \omega_n f(\boldsymbol{u}_n))'$ sans estimer σ_e^2. Par exemple, si $\sigma_{ei}^2 = m_i^{-1}\sigma_e^2$, le $\boldsymbol{\omega}_e$ pour le modèle (23) est $(m_i^{-1}\omega_1, \ldots, m_n^{-1}\omega_n)'$.

4. L'EQM des prédicteurs modifiés

Nous pouvons montrer que tout prédicteur de la forme (15) peut s'écrire

$$\tilde{\boldsymbol{y}}^a = \boldsymbol{Y} - \boldsymbol{C}_a^{-1}\boldsymbol{B}\boldsymbol{C}_a(\boldsymbol{I}_n - \boldsymbol{\Gamma})(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \qquad (24)$$

en posant que $\boldsymbol{C}_a = \boldsymbol{A}_a\boldsymbol{T}$, où

$$\boldsymbol{A}_a = \begin{pmatrix} 1 & \boldsymbol{0}'_{n-1} \\ -\boldsymbol{a}_{n-1} & \boldsymbol{I}_{n-1} \end{pmatrix},$$

$$\boldsymbol{a}_{n-1} = (a_2, \ldots, a_n)'.$$

Par conséquent, nous pouvons utiliser l'estimateur de la variance de $\tilde{\boldsymbol{y}}^a$ défini en (14) et proposé dans Isaki et coll. (2000) pour estimer l'EQM de tout prédicteur de la forme (15). Souvent, l'EQM d'un prédicteur ajusté est proche de celle du prédicteur avant l'ajustement.

Comme le modèle augmenté (23) possède la propriété d'autocalage, nous pouvons employer, pour estimer l'EQM, la formule utilisée pour les prédicteurs MPLSBE habituels.

5. Étude par simulation

5.1 Conditions de simulation

Afin d'étudier les propriétés empiriques des prédicteurs sur petits domaines décrits aux sections 2 et 3, nous utilisons des données conçues pour simuler une grande enquête nationale dans laquelle on souhaite produire des estimations par État. Le tableau 1 contient les populations approximatives des 50 États des États-Unis en l'an 2000. Les tailles d'échantillon m_i données dans le tableau sont approximativement proportionnelles aux racines carrées des populations des États.

Nous avons produit en tout 10 000 échantillons pour l'étude par simulation. Chacun était composé d'observations produites à l'aide du modèle

$$Y_{ij} = \boldsymbol{x}_i'\boldsymbol{\beta} + b_i + \varepsilon_{ij}, \qquad (25)$$

où $\boldsymbol{x}_i' = (1, z_i)$, $\boldsymbol{\beta}' = (6{,}0,\ 3{,}0)$, $z_i = \mathrm{Pop}_i^{0,2} - \overline{\mathrm{Pop}^{0,2}}$, Pop_i est la population de l'État i en millions, $\overline{\mathrm{Pop}^{0,2}}$ est la moyenne de $\mathrm{Pop}_i^{0,2}$, $b_i \sim NI(0, 1)$, et $\varepsilon_{ij} \sim NI(0, 16)$. Les b_i et les e_{ij} sont indépendants. Pour les observations au niveau de l'État, le modèle devient

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + b_i + e_i, \qquad (26)$$

où $Y_i = m_i^{-1}\sum_{j=1}^{m_i} Y_{ij}$ et $e_i = m_i^{-1}\sum_{j=1}^{m_i}\varepsilon_{ij}$. Avec les tailles d'échantillon données au tableau 1, le γ_i défini en (6) est 0,784 pour l'État le plus grand (Californie) et 0,304 pour l'État le plus petit (Wyoming).

Tableau 1 Populations et tailles d'échantillon pour la simulation

État	Population (en milliers)	Taille de l'échantillon (m_i)	État	Population (en milliers)	Taille de l'échantillon (m_i)
1	33 640	58	26	4 000	20
2	21 160	46	27	3 610	19
3	19 360	44	28	3 240	18
4	16 000	40	29	3 240	18
5	12 250	35	30	2 890	17
6	12 250	35	31	2 890	17
7	11 560	34	32	2 560	16
8	10 240	32	33	2 560	16
9	8 410	29	34	2 250	15
10	8 410	29	35	1 960	14
11	7 840	28	36	1 690	13
12	7 290	27	37	1 690	13
13	6 250	25	38	1 690	13
14	6 250	25	39	1 210	11
15	5 760	24	40	1 210	11
16	5 760	24	41	1 210	11
17	5 760	24	42	1 210	11
18	5 290	23	43	1 000	10
19	5 290	23	44	810	9
20	5 290	23	45	810	9
21	4 840	22	46	810	9
22	4 410	21	47	640	8
23	4 410	21	48	640	8
24	4 410	21	49	640	8
25	4 000	20	50	490	7

Pour étudier les propriétés des cinq prédicteurs, c'est-à-dire MPLSBE, Pfeffermann-Bernard (PB), Isaki-Tsay-Fuller (ITF), You-Rao (YR) et le modèle augmenté (23) (AUG2), nous nous sommes servis de deux modèles d'estimation. Le premier est un modèle spécifié incorrectement avec $\boldsymbol{x}_i' = 1$ dans la notation de (26). Nous l'appelons modèle (A). De même, nous dénommons modèle (B) le modèle générateur de données (26) avec $\boldsymbol{x}_i' = (1, z_i)$.

Suivant la méthode décrite dans Wang et Fuller (2003), l'estimateur de σ_b^2 est

$$\hat{\sigma}_b^2 = \max\{0{,}5[\hat{V}(\tilde{\sigma}_b^2)]^{0,5},\ \tilde{\sigma}_b^2\}, \qquad (27)$$

où

$$\tilde{\sigma}_b^2 = \sum_{i=1}^{50} c_i\left[\frac{50}{50-k}(Y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\mathrm{MCO}})^2 - \hat{\sigma}_{ei}^2\right], \qquad (28)$$

$$\hat{V}(\tilde{\sigma}_b^2) = \sum_{i=1}^{50} c_i^2\left\{\left[\frac{50}{50-k}(Y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\mathrm{MCO}})^2 - \hat{\sigma}_{ei}^2\right] - \tilde{\sigma}_b^2\right\}^2, \qquad (29)$$

$c_i = (\sum_{i=1}^{50} m_i^{0,5})^{-1} m_i^{0,5}$, $\hat{\boldsymbol{\beta}}_{\mathrm{MCO}}$ est l'estimateur par les moindres carrés ordinaires du coefficient de régression de

Y_i sur $\boldsymbol{x}_i$, k est la dimension du vecteur $\boldsymbol{x}_i$, $\hat{\sigma}_{ei}^2 = m_i^{-1} s_i^2$ et $s_i^2 = (m_i - 1)^{-1} \sum_{j=1}^{m_i} (Y_{ij} - Y_i)^2$ est la variance d'échantillon pour le domaine i.

Le prédicteur MPLSBE est

$$\hat{y}_i = \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_{\mathrm{MCG}} + \hat{\gamma}_i (Y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_{\mathrm{MCG}}), \qquad (30)$$

où

$$\hat{\gamma}_i = (\hat{\sigma}_b^2 + \hat{\sigma}_{ei}^2)^{-1} \hat{\sigma}_b^2, \qquad (31)$$

et

$$\hat{\boldsymbol{\beta}}_{\mathrm{MCG}} = \left[\sum_{i=1}^{n} (\sigma_b^2 + \hat{\sigma}_{ei}^2)^{-1} \boldsymbol{x}_i \boldsymbol{x}_i' \right]^{-1} \left[\sum_{i=1}^{n} (\hat{\sigma}_b^2 + \hat{\sigma}_{ei}^2)^{-1} \boldsymbol{x}_i Y_i \right] \qquad (32)$$

est l'estimateur par les moindres carrés généralisés de $\boldsymbol{\beta}$. La contrainte considérée est $\sum_{i=1}^{50} \omega_i \hat{y}_i^{\boldsymbol{a}} = \sum_{i=1}^{50} \omega_i Y_i$, où

$$\omega_i = \left(\sum_{i=1}^{50} \mathrm{Pop}_i \right)^{-1} \mathrm{Pop}_i. \qquad (33)$$

Avec le prédicteur MPLSBE, nous dérivons les prédicteurs PB et ITF en utilisant (13) et (14). Nous dérivons le prédicteur YR en utilisant (16) avec $\hat{\gamma}_i$ défini en (31) et $\hat{\boldsymbol{\beta}}_{\mathrm{YR}}$ défini en (17). Le prédicteur AUG2 est de la forme (30) avec $\boldsymbol{x}_i' = (1, \omega_i \hat{\sigma}_{ei}^2)$ sous le modèle augmenté (A) et $\boldsymbol{x}_i' = (1, z_i, \omega_i \hat{\sigma}_{ei}^2)$ sous le modèle augmenté (B).

Pour chacun des dix prédicteurs, nous calculons le critère

$$Q(\hat{\boldsymbol{y}}) = 0{,}02 \sum_{i=1}^{50} \varphi_i (\hat{y}_i - y_i)^2 \qquad (34)$$

où $\varphi_i = (\gamma_i m_i^{-1} \sigma_e^2)^{-1}$. Notons que φ_i^{-1} est la variance du prédicteur de y_i construite à l'aide des paramètres connus.

5.2 Résultats de la simulation

L'estimateur de σ_b^2 construit sous le modèle (B) a une moyenne Monte Carlo de 1,001 et un écart-type de 0,386. Quand nous utilisons le modèle (A) pour l'estimation, la moyenne Monte Carlo de l'estimateur de σ_b^2 est 1,720 et l'écart-type est de 0,521. Le carré de la moyenne de $3z_i$ est 0,636. Par conséquent, la méthode d'estimation avec l'utilisation du modèle (A) intègre une part importante de l'effet fixe de domaine du modèle (B) dans l'effet aléatoire. Quand $\hat{\sigma}_b^2$ est plus grand, $\hat{\gamma}_i$ est plus grand et une plus forte proportion de Y_i est utilisée pour construire les prédicteurs, ce qui annule en partie l'effet de l'erreur de spécification du modèle. Sous le modèle augmenté (A), l'estimateur de σ_b^2 est égal à 1,197, c'est-à-dire qu'une part nettement plus faible de l'effet fixe de domaine du modèle (B) est intégrée dans l'effet aléatoire.

Le tableau 2 contient certaines statistiques sommaires pour les prédicteurs fondés sur 10 000 échantillons simulés. Le biais empirique dans $\sum \omega_i (Y_i - \hat{y}_i)$ est nul pour tous les prédicteurs quand le modèle (B) ou sa version augmentée est utilisé, parce que le modèle de prédiction concorde avec le modèle de génération des données. L'écart-type simulé de la différence $\sum \omega_i (Y_i - \hat{y}_i)$ est égal à 0,022 pour le prédicteur MPLSBE habituel. L'écart-type de $\sum \omega_i (Y_i - \hat{y}_i)$ est nul pour les quatre autres prédicteurs, parce que ceux-ci satisfont la contrainte $\sum_{i=1}^{50} \omega_i \hat{y}_i = \sum_{i=1}^{50} \omega_i Y_i$.

Tableau 2
Propriétés Monte Carlo des prédicteurs pour petits domaines (moyenne de 10 000 échantillons générés par le modèle B)

	Quantité	MPLSBE	PB	ITF	YR	AUG2
		Prédicteur construit sous le modèle (A)				
$\sum \omega_i (Y_i - \hat{y}_i)$	Moyenne	-0,100	0,000	0,000	0,000	0,000
	(É.-T.)	(0,027)	(0,000)	(0,000)	(0,000)	(0,000)
$Q(\hat{\boldsymbol{y}})$	Moyenne	1,438	1,446	1,419	1,558	1,298
		Prédicteur construit sous le modèle (B)				
$\sum \omega_i (Y_i - \hat{y}_i)$	Moyenne	-0,000	0,000	0,000	0,000	0,000
	(É.-T.)	(0,022)	(0,000)	(0,000)	(0,000)	(0,000)
$Q(\hat{\boldsymbol{y}})$	Moyenne	1,203	1,202	1,202	1,208	1,219

La prédiction fondée sur le modèle (A) ou sa version augmentée est biaisée, parce que le modèle de génération des données (B) contient une fonction de la taille de population. La moyenne simulée de la différence pondérée $\sum \omega_i (Y_i - \hat{y}_{iA})$ est -0,100, où $\hat{y}_{iA}$ est le prédicteur MPLSBE. La valeur de la statistique t pour le biais pondéré est -3,70. La variance simulée de la moyenne pondérée de la prédiction est égale à

$$V \left\{ \sum_{i=1}^{50} \omega_i \hat{y}_{iA} \right\} = 0{,}060.$$

L'erreur quadratique moyenne estimée de la prédiction sur le modèle (A) de $\sum \omega_i y_i$ pour les données générées par le modèle (B) est

$$\mathrm{EQM} \left\{ \sum_{i=1}^{50} \omega_i (\hat{y}_{iA} - y_i) \right\} = V \left\{ \sum_{i=1}^{50} \omega_i \hat{y}_{iA} \right\} + \mathrm{Biais}^2$$
$$= 0{,}060 + (-0{,}100)^2 = 0{,}070. \qquad (35)$$

La variance de $\sum \omega_i Y_i$ est

$$V \left\{ \sum_{i=1}^{50} \omega_i Y_i \right\} = \sum_{i=1}^{50} \omega_i^2 (\sigma_b^2 + m_i^{-1} \sigma_e^2) = 0{,}0622. \qquad (36)$$

Donc, l'utilisation de $\sum \omega_i Y_i$ comme estimateur de $\sum \omega_i y_i$ est environ 12,5 % plus efficace que celle du prédicteur $\sum \omega_i \hat{y}_{iA}$ fondé sur le modèle (A). Étant donné le calage, les quatre prédicteurs (PB, ITF, YR et AUG2) de $\sum \omega_i y_i$ ont la même EQM que la moyenne estimée directement $\sum \omega_i Y_i$. Le carré du biais représenterait une proportion beaucoup plus importante de l'erreur quadratique moyenne si les petits domaines étaient plus nombreux.

La valeur du critère Q pour MPLSBE est 1,20 sous le modèle (B). Donc, l'estimation des paramètres accroît la variance moyenne des prédicteurs d'environ 20 % comparativement à l'utilisation des paramètres connus. Si les prédictions sont faites en utilisant des valeurs connues de σ_{ei}^2, la valeur du critère Q pour MPLSBE est 1,06. Par conséquent, l'estimation de $\hat{\sigma}_{ei}^2$ est celle qui contribue le plus à l'accroissement de la variabilité. Comme le biais est nul quand le modèle (B) est utilisé pour l'estimation, les ajustements que produisent les prédicteurs contraints sont faibles, si bien que les prédicteurs ajustés donnent des valeurs de critères fort semblables à celles des prédicteurs non ajustés. Les prédicteurs YR ont des valeurs de critère un peu plus grandes que les prédicteurs PB et ITF correspondants, parce que le prédicteur YR utilise un estimateur inefficace de $\boldsymbol{\beta}$. Les prédicteurs reposant sur le modèle augmenté ont une valeur de critère Q légèrement plus grande, parce que le modèle contient une variable redondante. L'accroissement de moins de 2 % de Q est de l'ordre de n^{-1}, c'est-à-dire la perte attendue en raison de l'ajout d'un paramètre inutile dans une prédiction par les moindres carrés.

La valeur du critère Q pour MPLSBE sous le modèle (A) est de 1,438 comparativement à 1,203 pour MPLSBE sous le modèle (B). Il s'agit de la pénalité due à l'erreur de spécification du modèle. Parmi les méthodes d'ajustement fondées sur le modèle (A), la méthode ITF est celle dont la valeur de Q la plus faible. Les différences entre les prédicteurs PB, ITF et MPLSBE sont faibles. La valeur de Q est environ 8 % plus grande pour la méthode You-Rao que pour MPLSBE. Le prédicteur fondé sur le modèle augmenté a une valeur de Q environ 11 % plus petite que le prédicteur MPLSBE. Dans un certain sens, le modèle augmenté est spécifié moins incorrectement.

Non seulement l'approche d'augmentation donne lieu au calage des prédicteurs sur petits domaines, mais elle réduit aussi le biais au niveau du domaine quand le modèle est spécifié incorrectement. Les tableaux 3 et 4 donnent les propriétés Monte Carlo des prédicteurs pour certains domaines. Dans les tableaux, les biais estimés sont normalisés par $(\gamma_i m_i^{-1} \sigma_e^2)^{0,5}$, la racine carrée de l'EQM du prédicteur MPLSB avec paramètres connus, et les EQM estimées sont normalisées par $\gamma_i m_i^{-1} \sigma_e^2$. Quand le modèle correct (B) est utilisé, le biais Monte Carlo au niveau du domaine approche de zéro. Voir le tableau 3. En outre, l'écart est faible entre l'EQM calculée pour les diverses méthodes, le modèle augmenté produisant une EQM un peu plus grande (moins de 2 %). De nouveau, le tableau 3 montre que le processus de calage a peu d'effet sur l'EQM des prédicteurs au niveau du domaine comparativement au prédicteur MPLSBE.

Tableau 3
Propriétés Monte Carlo des prédicteurs individuels de domaine en utilisant des versions du modèle (B) (10 000 échantillons générés par le modèle B)

État	Quantité	EMPLSB	PB	ITF	YR	AUG2
1	Biais	0,011	0,012	0,012	0,013	0,011
	EQM	1,100	1,101	1,105	1,104	1,120
2	Biais	0,000	0,001	0,001	0,002	0,001
	EQM	1,072	1,072	1,072	1,076	1,092
14	Biais	-0,001	-0,001	-0,001	-0,001	-0,001
	EQM	1,058	1,058	1,058	1,058	1,074
26	Biais	0,015	0,015	0,015	0,015	0,018
	EQM	1,078	1,077	1,078	1,079	1,092
38	Biais	-0,005	-0,005	-0,005	-0,006	-0,003
	EQM	1,123	1,122	1,122	1,132	1,135
50	Biais	0,012	0,012	0,012	0,009	0,014
	EQM	1,222	1,222	1,222	1,247	1,246

Si nous utilisons un modèle spécifié incorrectement, tel que le modèle (A), le biais de la somme des prédicteurs MPLSBE en tant qu'estimateur du total est négatif dans notre exemple, parce que les États pour lesquels le biais est négatif ont une pondération ω_i élevée. Voir le tableau 4. Des méthodes d'ajustement, comme PB ou ITF, répartissent le biais entre tous les petits domaines. Donc, l'ajustement réduit le biais négatif des prédicteurs de domaine ayant un grand biais négatif et augmente le biais positif des prédicteurs ayant un grand biais positif. Par conséquent, l'EQM est plus petite pour les grands états et elle est un peu plus grande pour les petits états. Le prédicteur YR produit un biais plus important que le prédicteur ITF. Par ailleurs, les prédicteurs construits avec $\boldsymbol{x}_i' = (1, \omega_i \hat{\sigma}_{ei}^2)$, c'est-à-dire le modèle augmenté (A), sont nettement supérieurs à ceux construits sous le modèle (A). Le biais au niveau des domaines, que ceux-ci soient grands ou petits, est réduit.

Tableau 4
Propriétés Monte Carlo des prédicteurs individuels de domaine en utilisant des versions du modèle (A) (10 000 échantillons générés par le modèle B).

État	Quantité	MPLSBE	PB	ITF	YR	AUG2
1	Biais	-0,597	-0,225	-0,052	-0,473	-0,030
	EQM	1,471	1,165	1,130	1,331	1,139
2	Biais	-0,496	-0,227	-0,170	-0,358	-0,070
	EQM	1,330	1,134	1,115	1,207	1,124
14	Biais	-0,121	0,004	-0,031	0,055	-0,025
	EQM	1,100	1,085	1,086	1,089	1,105
26	Biais	0,057	0,157	0,115	0,249	0,053
	EQM	1,126	1,148	1,136	1,188	1,132
38	Biais	0,380	0,453	0,406	0,601	0,202
	EQM	1,340	1,405	1,361	1,571	1,233
50	Biais	0,922	0,980	0,931	1,178	0,537
	EQM	2,196	2,316	2,215	2,767	1,577

6. Conclusion

Dans le présent article, nous passons en revue plusieurs prédicteurs calés. Nous jetons un nouveau regard sur la contrainte d'étalonnage (7). Nous considérons l'imposition de la contrainte comme un problème d'ajustement et nous présentons un critère qui unifie la dérivation des prédicteurs calés. L'approche du critère pour résoudre le problème permet de considérer d'autres prédicteurs.

L'imposition de la contrainte comprend implicitement la possibilité que le modèle pour petits domaines soit spécifié incorrectement et que les prédicteurs présentent un biais. Quand le modèle est mal spécifié, l'ajustement par calage corrige le biais global, mais non le biais au niveau du petit domaine. L'approche du modèle augmenté aboutit à un prédicteur autocalé et réduit le biais au niveau du petit domaine. En outre, l'estimation de la variance des prédicteurs calés extérieurement est relativement complexe, tandis que l'estimation de la variance des prédicteurs autocalés est simple. Brièvement, si l'on est préoccupé par le biais, l'utilisation du modèle augmenté autocalé est préférable au calage externe.

7. Annexe

Preuve du théorème 1 : Soit $\tilde{y}_i^H$ un prédicteur MPLSB de y_i et soit $\hat{y}_i$ un prédicteur linéaire sans biais de y_i. Selon les résultats standard pour le MPLSB (voir, par exemple, Robinson (1991) et Harville (1976)), nous avons

$$\text{cov}(\tilde{y}_i^H - y_i,\ \hat{y}_j - \tilde{y}_j^H) = 0, \quad \text{si} \quad i \neq j, \qquad \text{(A.1)}$$

pour $i = 1, ..., n$ et $j = 1, ..., n$. Soit $R(\hat{\boldsymbol{y}}^{\boldsymbol{a}})$ la série de prédicteurs linéaires sans biais qui satisfont (7). Pour tout $\hat{\boldsymbol{y}} \in R(\hat{\boldsymbol{y}}^{\boldsymbol{a}})$, selon (A.1), nous avons

$$E\{(\hat{y}_i - y_i)^2\} = E\{(\tilde{y}_i^H - y_i)^2\} + E\{(\hat{y}_i - \tilde{y}_i^H)^2\}. \qquad \text{(A.2)}$$

Par conséquent,

$$\begin{aligned} Q(\hat{\boldsymbol{y}}) &= \sum_{i=1}^n \varphi_i E\{(\hat{y}_i - y_i)^2\} \\ &= \sum_{i=1}^n \varphi_i E\{(\tilde{y}_i^H - y_i)^2\} + \sum_{i=1}^n \varphi_i E\{(\hat{y}_i - \tilde{y}_i^H)^2\}. \end{aligned} \qquad \text{(A.3)}$$

Puisque $\hat{\boldsymbol{y}}$ satisfait (7), nous avons $\sum_{i=1}^n \omega_i \hat{y}_i = \sum_{i=1}^n \omega_i Y_i$ et, pour le $\hat{y}_i^{\bar{a}}$ défini en (19),

$$\begin{aligned} \hat{y}_i^{\bar{a}} &= \tilde{y}_i^H + \breve{a}_i \left[\sum_{j=1}^n \omega_j (Y_j - \tilde{y}_j^H) \right] \\ &= \tilde{y}_i^H + \breve{a}_i \left[\sum_{j=1}^n \omega_j (\hat{y}_j - \tilde{y}_j^H) \right]. \end{aligned}$$

Selon (A.1),

$$\begin{aligned} Q(\hat{y}_i^{\bar{a}}) &= \sum_{i=1}^n \varphi_i E(\hat{y}_i^{\bar{a}} - y_i)^2 \\ &= \sum_{i=1}^n \varphi_i E\left\{ \left[(\tilde{y}_i^H - y_i) + \breve{a}_i \left(\sum_{j=1}^n \omega_j \hat{y}_j - \sum_{j=1}^n \omega_j \tilde{y}_j^H \right) \right]^2 \right\} \\ &= \sum_{i=1}^n \varphi_i E\{(\tilde{y}_i^H - y_i)^2\} + E\left\{ \left[\sum_{j=1}^n \omega_j (Y_j - \tilde{y}_j^H) \right]^2 \right\} \sum_{i=1}^n \varphi_i \breve{a}_i^2. \end{aligned} \qquad \text{(A.4)}$$

Puisque $\breve{a}_i = (\sum_{i=1}^n \varphi_i^{-1} \omega_i^2)^{-1} \varphi_i^{-1} \omega_i$ dans (A.4), nous avons

$$\begin{aligned} Q(\hat{y}_i^{\bar{a}}) &= \sum_{i=1}^n \varphi_i E\{(\tilde{y}_i^H - y_i)^2\} \\ &\quad + E\left\{ \left[\sum_{j=1}^n \omega_j (\hat{y}_j - \tilde{y}_j^H) \right]^2 \right\} \left(\sum_{i=1}^n \varphi_i^{-1} \omega_i^2 \right)^{-1}. \end{aligned} \qquad \text{(A.5)}$$

Notons que

$$\begin{aligned} E\left\{ \left[\sum_{j=1}^n \omega_j (\hat{y}_j - \tilde{y}_j^H) \right]^2 \right\} &\leq \sum_{j=1}^n \sum_{k=1}^n \omega_j \omega_k g_j g_k \\ &= \left(\sum_{i=1}^n \omega_i g_i \right)^2, \end{aligned} \qquad \text{(A.6)}$$

où $g_j = \{E[(\hat{y}_j - \tilde{y}_j^H)^2]\}^{0.5}$. En vertu de l'inégalité de Cauchy,

$$\begin{aligned} \left(\sum_{j=1}^n \omega_j g_j \right)^2 &\leq \left(\sum_{i=1}^n \varphi_i^{-1} \omega_i^2 \right) \left(\sum_{i=1}^n \varphi_i g_i^2 \right) \\ &= \left(\sum_{i=1}^n \varphi_i^{-1} \omega_i^2 \right) \left[\sum_{i=1}^n \varphi_i E\{(\hat{y}_j - \tilde{y}_j^H)^2\} \right]. \end{aligned} \qquad \text{(A.7)}$$

En combinant (A.3), (A.5), (A.6) et (A.7), nous avons $Q(\hat{y}_i^{\bar{a}}) \leq Q(\hat{\boldsymbol{y}})$.

Pour démontrer le caractère unique de $\hat{y}_i^{\bar{a}}$, nous devons vérifier quand les inégalités (A.6) et (A.7) deviennent des égalités. L'inégalité (A.6) devient une égalité si, et uniquement si,

$$\hat{y}_j - \tilde{y}_j^H = c_j^0 + c_j^1 (\hat{y}_1 - \tilde{y}_1^H) \qquad \text{(A.8)}$$

pour certaines constantes c_j^0 et c_j^1, $j = 2, ..., n$. L'inégalité (A.7) devient une égalité si, et uniquement si,

$$\sqrt{\varphi_i^{-1} \omega_i^2} \sqrt{v_j g_j^2} - \sqrt{v_j^{-1} \omega_j^2} \sqrt{\varphi_i g_i^2} = 0, \qquad \text{(A.9)}$$

ou, de manière équivalente,

$$\varphi_i^2 \omega_i^{-2} E\{(\hat{y}_i - \tilde{y}_i^H)^2\} = v_j^2 \omega_j^{-2} E\{(\hat{y}_j - \tilde{y}_j^H)^2\}. \qquad \text{(A.10)}$$

En outre,

$$\sum_{i=1}^{n} \omega_i \hat{y}_i = \sum_{i=1}^{n} \omega_i Y_i. \qquad \text{(A.11)}$$

En combinant (A.8), (A.10) et (A.11), nous obtenons que l'égalité tient si, et uniquement si, $\hat{y}_j = \hat{y}_i^{\bar{a}}$. Donc, nous avons montré que $\hat{y}_i^{\bar{a}}$ est le seul prédicteur linéaire sans biais qui satisfait (7) et minimise le critère (18).

Remerciements

Cette étude a été financée en partie par l'entente de coopération 68-3A75-14 conclue entre le Natural Resources Conservation Service de l'USDA et la Iowa State University.

Bibliographie

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

Harville, D.A. (1976). Extension of the Gauss-Markov Theorem to include the estimation of random effects. *Annals of Statistics*, 384-395.

Henderson, C.R. (1963). Selection index and expected genetic advance. Dans *Statistical Genetics and Plant Breeding*, (Éds., W.D. Hanson et H.F. Robinson), National Academy of Sciences–National Research Council, Washington, Publication 982, 141-163.

Isaki, C.T., Tsay, J.H. et Fuller, W.A. (2000). Estimation des facteurs de correction au recensement. *Techniques d'enquête*, 31-42.

Mantel, H.J., Singh, A.C. et Barcau, M. (1993). Benchmarking of small area estimators. Dans *Proceedings of International Conference on Establishment Surveys*, American Statistical Association, Washington, DC, 920-925.

Pfeffermann, D., et Barnard, C.H. (1991). Some new estimators for small-area means with application to the assessment of Farmland values. *Journal of Business & Economic Statistics*, 73-84.

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

Robinson, G.K. (1991). That MPLSB is a good thing: the estimation of random effects. *Statistical Science*, 15-51.

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

You, Y., et Rao, J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, 431-439.

You, Y., et Rao, J.N.K. (2003). Pseudo hierarchical Bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference*, 197-208.