N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

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

2003

Statistique Canada Statistics Canada

Canada

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

ESTIMATION DES EFFETS FIXES ET DES COMPOSANTES DE LA VARIANCE PAR UN MODÈLE À VALEUR ALÉATOIRE À L'ORIGINE EN UTILISANT DES DONNÉES D'ENQUÊTE

Yong You, J.N.K. Rao et Milorad Kovacevic[1]

RÉSUMÉ

Le modèle à valeur aléatoire à l'origine est souvent utilisé dans l'analyse des données d'enquête fondée sur un modèle. Dans le présent document, nous décrivons l'élaboration d'une méthode itérative à équations d'estimations pondérées (IEEP) pour estimer les effets fixes et les composantes de la variance du modèle à valeur aléatoire à l'origine à l'aide des poids de sondage. La méthode IEEP se situe dans le prolongement de la méthode proposée par You et Rao (2002) d'estimation des effets fixes et de la méthode proposée par Waclawiw et Liang (1993) d'estimation des composantes de la variance. Nous présentons une étude de simulation et un exemple reposant sur des données réelles d'enquête pour illustrer l'application de la méthode que nous proposons. Notre étude montre que la méthode IEEP proposée donne de bons résultats et converge rapidement. L'un de ses avantages est que la procédure d'estimation ne nécessite que les poids d'échantillonnage finaux, contrairement aux autres méthodes décrites dans la littérature qui demandent des poids d'échantillonnage de plus haut niveau, comme les poids des unités primaires d'échantillonnage, dont les utilisateurs ne disposent pas nécessairement en pratique.

MOTS CLÉS : Équations d'estimation, modèle hiérarchisé d'erreurs de régression, poids d'échantillonnage.

1. INTRODUCTION

Le modèle à valeur aléatoire à l'origine sert souvent à l'analyse de données d'enquête fondée sur un modèle, et notamment à l'estimation régionale (petites régions). Dans les méthodes classiques d'estimation des paramètres de régression (effets fixes) et des composantes de variance dans un tel modèle, on néglige les poids d'échantillonnage. Tel est le cas pour les méthodes d'ajustement de constantes, de maximum de vraisemblance (MV) et de maximum de vraisemblance en valeur résiduelle (MVR). Dans le présent document, nous proposons un système itératif d'équations d'estimations pondérées (IEEP) pour estimer les effets fixes et les composantes de la variance dans un modèle à valeur aléatoire à l'origine à l'aide des poids d'échantillonnage. Cette méthode se situe dans le prolongement de la méthode de You et Rao (2002) d'estimation des effets fixes et de la méthode de Waclawiw et Liang (1993) d'estimation des composantes de la variance. Elle actualise à tour de rôle les estimations des effets fixes et les estimations des composantes de variance, jusqu'à ce que ces dernières convergent. On peut donc simultanément obtenir des estimations pondérées des effets fixes et des composantes de la variance. Nous présentons une petite étude de simulation et un exemple reposant sur des données réelles d'enquête pour illustrer l'application de la méthode que nous proposons.

Soit y_{ij} la valeur de la variable d'intérêt pour la j^{e} unité du i^{e} groupe (région) et soit $x_{ij} = (x_{ij1},...,x_{ijp})'$ avec $x_{ij1} = 1$ le vecteur de variables auxiliaires liées à y_{ij} ($i = 1,...,m; j = 1,...,N_i$). Un modèle de population à valeur aléatoire à l'origine qui fait intervenir $\{y_{ij}, x_{ij}\}$ nous est donné par

$$y_{ij} = x_{ij}'\beta + v_i + e_{ij}, \quad j = 1,...,N_i, i = 1,...,m, \tag{1}$$

[1] Yong You, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Ottawa, Canada K1A 0T6; J.N.K. Rao, École de mathématiques et de statistique, Université Carleton, Ottawa, Canada K1S 5B6; Milorad Kovacevic, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Canada K1A 0T6.

où $\beta = (\beta_0,...,\beta_{p-1})'$ est le vecteur $p \times 1$ d'effets fixes ou de paramètres de régression, v_i un effet aléatoire lié au i^e groupe et N_i le nombre d'unités de population dans le i^e groupe. Nous posons que les effets aléatoires v_i sont iid $N(0, \sigma_v^2)$ et indépendants des erreurs e_{ij} au niveau des unités, qui sont iid $N(0, \sigma_e^2)$ par hypothèse.

Nous supposons en outre que les échantillons sont indépendamment tirés dans chaque groupe en fonction d'un plan d'échantillonnage spécifié. Nous posons enfin que les données d'échantillon $\{y_{ij}, x_{ij}, j = 1,...,n_i; i = 1,...,m\}$ sont conformes au modèle de population, c'est-à-dire que

$$y_{ij} = x_{ij}'\beta + v_i + e_{ij}, \quad j = 1,...,n_i, i = 1,...,m, \tag{2}$$

où n_i est la taille d'échantillon du i^e groupe. Cela implique qu'on n'a pas à tenir compte du plan d'échantillonnage de chaque groupe ou qu'il n'y a pas de biais de sélection. Le modèle (2) est aussi connu comme modèle hiérarchisé d'erreurs de régression (Battese, Harter et Fuller, 1988). Notre propos sera surtout d'estimer les paramètres de régression β et les composantes de la variance σ_e^2 et σ_v^2 à l'aide du modèle (2) et des poids d'échantillonnage.

Le modèle par unité (2) peut faire l'objet d'une agrégation au niveau des groupes par des estimateurs directs d'enquête. Soit $\tilde{w}_{ij}$ le poids d'échantillonnage de base de y_{ij}. Un estimateur direct d'échantillonnage de la moyenne de population au niveau des groupes nous est donné par

$$\bar{y}_{iw} = \frac{\sum_{j=1}^{n_i} \tilde{w}_{ij} y_{ij}}{\sum_{j=1}^{n_i} \tilde{w}_{ij}} = \sum_{j=1}^{n_i} w_{ij} y_{ij},$$

où $w_{ij} = \tilde{w}_{ij} / \sum_{j=1}^{n_i} \tilde{w}_{ij} = \tilde{w}_{ij} / \tilde{w}_{i\cdot}$ et $\sum_{j=1}^{n_i} w_{ij} = 1$. Si nous suivons You et Rao (2002, 2003), nous tirons le modèle agrégé suivant du modèle (2) au niveau des unités :

$$\bar{y}_{iw} = \bar{x}_{iw}'\beta + v_i + \bar{e}_{iw}, \quad i = 1,...,m, \tag{3}$$

où $\bar{e}_{iw} = \sum_{j=1}^{n_i} w_{ij} e_{ij}$ avec $E(\bar{e}_{iw}) = 0$ et $\text{var}(\bar{e}_{iw}) = \sigma_e^2 \sum_{j=1}^{n_i} w_{ij}^2 \equiv \sigma_e^2 \delta_i^2$, et où $\bar{x}_{iw} = \sum_{j=1}^{n_i} w_{ij} x_{ij}$.

Voici comment se présente le reste de notre exposé. La section 2 indique certaines méthodes types d'estimation de β, σ_e^2 et σ_v^2 en ne considérant pas les poids d'échantillonnage $\tilde{w}_{ij}$. À la section 3, nous décrivons un système itératif d'équations d'estimations pondérées (IEEP). À la section 4, nous évaluons, par une petite étude de simulation, l'application de la méthode proposée. À la section 5, nous comparons la méthode que nous proposons aux méthodes types avec des données réelles. À la section 6 enfin, nous livrons nos conclusions avec des observations.

2. ESTIMATION SANS LES POIDS D'ÉCHANTILLONNAGE

Pour estimer les paramètres de régression β, nous posons d'abord que les composantes de la variance σ_e^2 et σ_v^2 sont connues dans le modèle (2) par unité. Nous estimons ensuite β par l'estimateur des moindres carrés généralisés (MCG)

$$\tilde{\beta}_{MCG} = \left(\sum_{i=1}^{m} x_i' V_i^{-1} x_i \right)^{-1} \left(\sum_{i=1}^{m} x_i' V_i^{-1} y_i \right) \equiv \tilde{\beta}(\sigma_e^2, \sigma_v^2), \tag{4}$$

où $x_i = (x_{i1},...,x_{in_i})'$, $y_i = (y_{i1},...,y_{in_i})'$ et $V_i = \sigma_e^2 I_{n_i} + \sigma_v^2 1_{n_i} 1_{n_i}'$ avec 1_{n_i} et I_{n_i} désignant respectivement le vecteur unitaire et la matrice d'identité d'ordre n_i. L'estimateur $\tilde{\beta}_{MCG}$ dépend des composantes de variance σ_e^2 et σ_v^2. La variance de $\tilde{\beta}_{MCG}$ est $\text{var}(\tilde{\beta}_{MCG}) = (\sum_{i=1}^{m} x_i' V_i^{-1} x_i)^{-1}$.

Dans l'application d'une méthode simple d'estimation des composantes de la variance σ_e^2 et σ_v^2, nous procédons à deux régressions par moindres carrés ordinaires, puis employons la méthode des moments pour obtenir des estimateurs sans biais de σ_e^2 et σ_v^2 (Fuller et Battese, 1973; Stukel et Rao, 1997). Un estimateur sans biais de σ_e^2, désigné par $\hat{\sigma}_{eM}^2$, est donné par

$$\hat{\sigma}_{eM}^2 = (n-m-p+1)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{\varepsilon}_{ij}^2 , \tag{5}$$

où les $\{\hat{\varepsilon}_{ij}\}$ sont les résidus de la régression par moindres carrés ordinaires (MCO) de $y_{ij} - \bar{y}_i$ en fonction de $\{x_{ij1} - \bar{x}_{i\cdot 1}, ..., x_{ijp} - \bar{x}_{i\cdot p}\}$ et où $(\bar{y}_i, \bar{x}_{i\cdot 1}, ..., \bar{x}_{i\cdot p})$ sont les moyennes d'échantillon du i^{e} groupe. Un estimateur sans biais de σ_v^2 est

$$\tilde{\sigma}_{vM}^2 = n_*^{-1} \left[\sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{u}_{ij}^2 - (n-p)\hat{\sigma}_e^2 \right], \tag{6}$$

où $n_*^{-1} = n - tr[(X'X)^{-1} \sum_{i=1}^{m} n_i^2 \bar{x}_i \bar{x}_i']$ avec $X' = (x_1', ..., x_m')$ et où les $\{\hat{u}_{ij}\}$ sont les résidus de la régression MCO de y_{ij} sur $\{x_{ij1}, ..., x_{ijp}\}$. Comme $\tilde{\sigma}_{vM}^2$ peut prendre des valeurs négatives, un estimateur tronqué de σ_v^2 s'obtient par $\hat{\sigma}_{vM}^2 = \max(\tilde{\sigma}_{vM}^2, 0)$. À noter que $\hat{\sigma}_{vM}^2$ n'est plus sans biais, mais convergent à mesure qu'augmente m, qui est le nombre de groupes. Les estimateurs $\hat{\sigma}_{eM}^2$ et $\hat{\sigma}_{vM}^2$ équivalent à ceux obtenus par l'application de la méthode bien connue d'ajustement de constantes (« Fitting-of-constants » ou F-C) de Henderson (1953). On qualifie donc d'estimateurs d'ajustement de constantes les estimateurs $\hat{\sigma}_{eM}^2$ et $\tilde{\sigma}_{vM}^2$ par la méthode des moments. Une fois σ_e^2 et σ_v^2 estimés, nous utilisons l'estimateur MCG (4) et obtenons $\hat{\beta}_{MCG} = \tilde{\beta}(\hat{\sigma}_{eM}^2, \hat{\sigma}_{vM}^2)$ comme étant l'estimation de β.

Goldstein (1995) a proposé une méthode itérative par moindres carrés généralisés (IMCG) pour estimer le paramètre fixe de régression β et les composantes de variance σ_e^2 et σ_v^2. De telles méthodes IMCG comportent une double application de la méthode MCG. En première étape, nous obtenons l'estimation MCG $\tilde{\beta}_{MCG}$ de β en posant que σ_e^2 et σ_v^2 sont connues. En seconde étape, nous prenons l'estimation MCG $\tilde{\beta}_{MCG}$ donnée par (4) pour former les résidus « bruts » $\tilde{y}_{ij} = y_{ij} - x_{ij}' \tilde{\beta}_{MCG}$. L'estimation de σ_e^2 et σ_v^2 est alors une application de la méthode MCG de forme vectorielle de la matrice à produits croisés des résidus $\tilde{y}_{ij}$ dans une hypothèse de normalité. La méthode IMCG comprend une actualisation itérative jusqu'à convergence entre l'estimation MCG de β et les estimations correspondantes de σ_e^2 et σ_v^2.

3. ESTIMATION AVEC LES POIDS D'ÉCHANTILLONNAGE

Nous présentons maintenant un système itératif d'équations d'estimations pondérées (IEEP) permettant d'estimer β et les composantes de la variance σ_e^2 et σ_v^2 à l'aide des poids d'échantillonnage $\tilde{w}_{ij}$.

3.1 Estimation de β

D'abord, nous tirons le meilleur estimateur linéaire sans biais de prévision (estimation BLUP) de l'effet aléatoire v_i du modèle agrégé (3) pour des paramètres donnés β, σ_e^2 et σ_v^2 sous la forme suivante :

$$\tilde{v}_{iw}(\beta,\sigma_e^2,\sigma_v^2) = \gamma_{iw}(\bar{y}_{iw} - \bar{x}'_{iw}\beta)\,. \tag{7}$$

Ensuite, nous résolvons une équation d'estimation pondérée pour β en appliquant la méthode de You et Rao (2002, 2003) :

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\tilde{w}_{ij}x_{ij}[y_{ij} - x'_{ij}\beta - \tilde{v}_{iw}(\beta,\sigma_e^2,\sigma_v^2)] = 0\,. \tag{8}$$

Un estimateur sans biais de β fondé sur le modèle s'obtient à partir de (8) sous la forme suivante :

$$\tilde{\beta}_w = \left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\tilde{w}_{ij}x_{ij}(x_{ij} - r_{iw}\bar{x}_{iw})'\right]^{-1}\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\tilde{w}_{ij}(x_{ij} - r_{iw}\bar{x}_{iw})y_{ij}\right] \equiv \tilde{\beta}_w(\sigma_e^2,\sigma_v^2)\,. \tag{9}$$

Il convient de noter que $\tilde{\beta}_w$ s'établit à l'aide du modèle (2) par unité, du modèle agrégé (3) et de la pondération d'enquête $\tilde{w}_{ij}$. En nous fondant sur (9) et le modèle (2) par unité, nous constatons que $\tilde{\beta}_w \mid \beta,\sigma_e^2,\sigma_v^2 \sim N(\beta,\Phi_w)$, où la matrice des covariances Φ_w est donnée dans You et Rao (2002, 2003) comme

$$\Phi_w = \sigma_e^2(\sum_{i=1}^{m}\sum_{j=1}^{n_i}x_{ij}z'_{ij})^{-1}(\sum_{i=1}^{m}\sum_{j=1}^{n_i}z_{ij}z'_{ij})[(\sum_{i=1}^{m}\sum_{j=1}^{n_i}x_{ij}z'_{ij})^{-1}]' + \sigma_v^2(\sum_{i=1}^{m}\sum_{j=1}^{n_i}x_{ij}z'_{ij})^{-1}[\sum_{i=1}^{m}(\sum_{j=1}^{n_i}z_{ij})(\sum_{j=1}^{n_i}z_{ij})'][(\sum_{i=1}^{m}\sum_{j=1}^{n_i}x_{ij}z'_{ij})^{-1}]'$$

avec $z_{ij} = \tilde{w}_{ij}(x_{ij} - \gamma_{iw}\bar{x}_{iw})$. Cette matrice Φ_w dépend de σ_e^2 et de σ_v^2.

3.2 Estimation de σ_e^2

Pour estimer σ_e^2, nous obtenons d'abord les résidus intragroupes comme $e_{ij} - \bar{e}_{iw} = y_{ij} - \bar{y}_{iw} - (x_{ij} - \bar{x}_{iw})'\beta$, puis prenons l'espérance de la somme pondérée des carrés des résidus

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\tilde{w}_{ij}(e_{ij} - \bar{e}_{iw})^2$$

en tenant compte du modèle, ce qui nous mène à l'estimateur suivant sans biais de σ_e^2 basé sur un modèle:

$$\tilde{\sigma}_{ew}^2 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\tilde{w}_{ij}\left[y_{ij} - \bar{y}_{iw} - (x_{ij} - \bar{x}_{iw})'\beta\right]^2}{\sum_{i=1}^{m}\left[(1-\delta_i^2)\sum_{j=1}^{n_i}\tilde{w}_{ij}\right]} \equiv \tilde{\sigma}_{ew}^2(\beta)\,, \tag{10}$$

où $\delta_i^2 = \sum_{j=1}^{n_i}w_{ij}^2$. À noter que l'estimateur $\tilde{\sigma}_{ew}^2$ dépend de β.

3.3 Estimation de σ_v^2

Pour estimer σ_v^2, nous notons d'abord que l'estimateur BLUP $\tilde{v}_{iw}$ donné par (7) est aussi la moyenne postérieure de v_i compte tenu de $\bar{y}_{iw}$ selon le modèle agrégé (3), c'est-à-dire que $\tilde{v}_{iw} = E(v_i \mid \bar{y}_{iw})$, en supposant que les paramètres β, σ_e^2 et σ_v^2 sont connus (You et Rao, 2003). Nous notons également que $E(\tilde{v}_{iw}) = E(E(v_i \mid \bar{y}_{iw})) = E(v_i) = 0$ et

$$E(V(v_i \mid \bar{y}_{iw})) = E(E(\tilde{v}_{iw} - v_i)^2 \mid \bar{y}_{iw}) = E(\tilde{v}_{iw} - v_i)^2,$$

de même que

$$E(\tilde{v}_{iw} - v_i)^2 = \sigma_v^2(\gamma_{iw} - 1)^2 + \sigma_e^2\delta_i^2\gamma_{iw}^2,$$

où $\gamma_{iw} = \sigma_v^2/(\sigma_v^2 + \sigma_e^2\delta_i^2)$. Ensuite, par la décomposition $V(v_i) = V(E(v_i \mid \bar{y}_{iw})) + E(V(v_i \mid \bar{y}_{iw}))$, la composante de variance σ_v^2 peut s'exprimer sous la forme suivante :

$$\sigma_v^2 = V(\tilde{v}_{iw}) + E(\tilde{v}_{iw} - v_i)^2 = E(\tilde{v}_{iw}^2) + E(\tilde{v}_{iw} - v_i)^2. \tag{12}$$

Nous prenons alors la moyenne (12) sur les groupes i et employons (11) pour obtenir l'estimateur suivant de σ_v^2 :

$$\tilde{\sigma}_{vw}^2 = \frac{1}{m}\sum_{i=1}^{m}\tilde{v}_{iw}^2 + \frac{\sigma_v^2}{m}\sum_{i=1}^{m}(\gamma_{iw} - 1)^2 + \frac{\sigma_e^2}{m}\sum_{i=1}^{m}\delta_i^2\gamma_{iw}^2 = \frac{1}{m}\sum_{i=1}^{m}\tilde{v}_{iw}^2 + \frac{1}{m}\sum_{i=1}^{m}\frac{\sigma_e^2\sigma_v^2\delta_i^2}{\sigma_v^2 + \sigma_e^2\delta_i^2} \equiv \tilde{\sigma}_{vw}^2(\tilde{v}_w, \sigma_e^2, \sigma_v^2), \tag{13}$$

où $\tilde{v}_w = (\tilde{v}_{1w}, ..., \tilde{v}_{mw})'$. On remarquera que $\tilde{\sigma}_{vw}^2$ dépend de σ_e^2 et σ_v^2. Dans un contexte d'échantillonnage aléatoire simple (EAS) avec $w_{ij} = 1/n_i$, l'estimateur (13) se ramène à l'estimateur de σ_v^2 présenté par Waclawiw et Liang (1993) :

$$\tilde{\sigma}_{vw}^2 = \frac{1}{m}\sum_{i=1}^{m}\tilde{v}_i^2 + \frac{1}{m}\sum_{i=1}^{m}\frac{\sigma_e^2\sigma_v^2}{n_i\sigma_v^2 + \sigma_e^2}$$

avec $\tilde{v}_i = \gamma_i(\bar{y}_i - \bar{x}_i'\beta)$ et $\gamma_i = \sigma_v^2/(\sigma_v^2 + \sigma_e^2/n_i)$.

3.4 Procédure itérative

Nous proposons maintenant une procédure itérative d'actualisation par étapes qui permet d'obtenir simultanément des estimations pondérées de β, σ_e^2 et σ_v^2. À partir des valeurs initiales $\hat{\sigma}_e^{2(0)}$ et $\hat{\sigma}_v^{2(0)}$ et pour $k = 0,1,2,...$, nous actualisons les paramètres de la manière suivante :

(1) nous calculons $\hat{\beta}^{(k+1)} = \tilde{\beta}_w(\hat{\sigma}_e^{2(k)}, \hat{\sigma}_v^{2(k)})$, où $\tilde{\beta}_w$ est donné par (9);

(2) nous calculons $\hat{\sigma}_e^{2(k+1)} = \tilde{\sigma}_{ew}^2(\hat{\beta}^{(k+1)})$, où $\tilde{\sigma}_{ew}^2$ est donnée par (10);

(3) nous calculons $\hat{v}_{iw}^{(k+1)} = \tilde{v}_{iw}(\hat{\beta}^{(k+1)}, \hat{\sigma}_e^{2(k+1)}, \hat{\sigma}_v^{2(k)})$, où $\tilde{v}_{iw}$ est donnée par (7);

(4) nous calculons $\hat{\sigma}_v^{2(k+1)} = \tilde{\sigma}_{vw}^2(\hat{v}_w^{(k+1)}, \hat{\sigma}_e^{2(k+1)}, \hat{\sigma}_v^{2(k)})$, où $\tilde{\sigma}_{vw}^2$ est donnée par (13) et où $\tilde{v}_w^{(k+1)} = (\tilde{v}_{1w}^{(k+1)}, ..., \tilde{v}_{mw}^{(k+1)})'$.

Les étapes (1) à (4) forment un cycle complet. Nous continuons par cycles itératifs jusqu'à convergence afin d'obtenir les estimations IEEP de β, σ_e^2 et σ_v^2. Nous pouvons prendre les estimations d'ajustement de constantes (F-C) $\hat{\sigma}_{eM}^2$ et $\hat{\sigma}_{vM}^2$ comme valeurs initiales $\hat{\sigma}_e^{2(0)}$ et $\hat{\sigma}_v^{2(0)}$ respectivement.

4. ÉTUDE DE SIMULATION

Pour évaluer la méthode IEEP proposée, nous avons procédé à une petite étude de simulation. Nous avons construit une population synthétique finie avec $m = 30$ groupes (grappes). Chaque groupe comprenait $N_i = 500$ unités de population et la population synthétique était tirée des modèles complets au niveau des unités (1) par $x_{ij} = (1, x_{1ij})'$, $\beta_0 = 50$, $\beta_1 = 10$, $\sigma_e^2 = 225$ et $\sigma_v^2 = 100$. La variable auxiliaire x_{1ij} était tirée d'une distribution exponentielle de moyenne 200. Sur cette population synthétique, nous avons prélevé indépendamment des échantillons PPT (échantillonnage avec probabilité proportionnelle à la taille) dans les divers groupes. Pour l'exécution de cet échantillonnage PPT, nous nous sommes reportés aux x_{1ij} comme valeurs de mesure de taille de chaque y_{ij}. À l'aide de ces valeurs x, nous avons calculé les probabilités de sélection $p_{ij} = x_{1ij} / \sum_j x_{1ij}$ de chaque unité y_{ij} et les valeurs ainsi obtenues ont servi à la sélection PPT d'échantillons avec remise de même taille, $n_i = n$, dans chaque groupe pour des valeurs *n* respectives de 5 et 20. Les poids d'échantillonnage de base sont donnés par $\tilde{w}_{ij} = n^{-1} p_{ij}^{-1}$, de sorte que $w_{ij} = p_{ij}^{-1} / \sum_j p_{ij}^{-1}$.

Nous avons repris toute cette procédure $R = 500$ fois et, dans chaque passage $r\,(r = 1,...,R)$, nous avons établi l'estimation MCG de β et les estimations F-C de σ_e^2 et σ_v^2. Avec ces estimations comme valeurs initiales dans l'algorithme IEEP, nous avons respectivement obtenu les estimations IEEP de β, σ_e^2 et σ_v^2. La convergence a été très rapide et quelques itérations ont suffi.

Notre but est d'évaluer le rendement de la méthode IEEP proposée et de comparer celle-ci à l'estimation MCG de β et aux estimations F-C de σ_e^2 et σ_v^2. C'est pourquoi nous avons établi des estimations IEEP en échantillonnage aléatoire simple (EAS) et pour des probabilités inégales de sélection PPT. Nous avons calculé le biais relatif en valeur absolue (BRA) et l'erreur relative (ER) des estimateurs. Ainsi, le BRA de l'estimateur $\hat{\sigma}_e^2$ de σ_e^2 se calcule comme $BRA(\hat{\sigma}_e^2) = |E^*(\hat{\sigma}_e^2)/\sigma_e^2 - 1|$, où E^* désigne la moyenne sur les $R = 500$ passages. En d'autres termes, $E^*(\hat{\sigma}_e^2) = \sum_{r=1}^R \hat{\sigma}_e^2(r)/R$, où $\hat{\sigma}_e^2(r)$ est l'estimation fondée sur le r^{e} passage de simulation. L'ER de $\hat{\sigma}_e^2$ se calcule comme $ER(\hat{\sigma}_e^2) = \sqrt{\sum_{r=1}^R (\hat{\sigma}_e^2(r) - \sigma_e^2)^2 / R}\,/\,\sigma_e^2$. Le tableau 1 livre les valeurs BRA en simulation. Pour les effets fixes β_0 et β_1, les deux méthodes MCG et IEEP donnent de très bons résultats en terme de BRA. Le biais relatif en valeur absolue est de moins de 1 % pour β_0 et de moins de 0,02 % pour β_1, ce qui révèle une absence de biais. Dans le cas de l'estimation des composantes de la variance, le BRA est de moins de 2 % pour σ_e^2 et de moins de 4 % pour σ_v^2 avec les deux méthodes F-C et IEEP, indice que l'une et l'autre de ces techniques mènent à des estimations sans biais des composantes visées. Ajoutons que les estimateurs de σ_e^2 ont un biais inférieur à celui des estimateurs de σ_v^2. Le tableau 2 compare les valeurs ER. Pour β_0 et β_1, la méthode IEEP(EAS) donne une ER inférieure à celle de la méthode MCG, mais la méthode générale IEEP donne, elle, une ER supérieure à cause du recours à un échantillonnage avec probabilités inégales de sélection. À mesure que la taille d'échantillon *n* augmente, l'ER diminue. Pour σ_e^2, les méthodes IEEP(EAS) et F-C présentent la même ER. Pour σ_v^2, l'ER est moins élevée avec la méthode IEEP(EAS) qu'avec la méthode F-C. Dans un échantillonnage à probabilités inégales, l'IEEP a une ER supérieure à l'ER de la F-C tant pour σ_e^2 que pour σ_v^2, comme on pouvait s'y attendre. De plus, les estimateurs de σ_v^2 donnent une ER supérieure à celle des estimateurs de σ_e^2. Nous en concluons que la méthode IEEP proposée

mène à des estimateurs sans biais des effets fixes et des composantes de la variance. Dans un échantillonnage aléatoire simple, la méthode IEEP est très efficace. En général, l'IEEP donne une ER plus élevée que les méthodes courantes, parce qu'elle tient compte des probabilités inégales de sélection dans la procédure d'estimation.

Tableau 1 : Comparaison en pourcentage des biais relatifs en valeur absolue (BRA)

	n = 5			n = 20		
	MCG	IEEP(EAS)	IEEP	MCG	IEEP(EAS)	IEEP
β_0	0,64	0,55	0,24	0,71	0,51	0,32
β_1	0,005	0,007	0,013	0,007	0,007	0,008
	n = 5			n = 20		
	F-C	IEEP(EAS)	IEEP	F-C	IEEP(EAS)	IEEP
σ_e^2	1,5	2,3	1,9	1,3	2,0	1,7
σ_v^2	3,2	3,6	3,8	3,3	3,3	3,5

Tableau 2 : Comparaison en pourcentage des erreurs relatives (ER)

	n = 5			n = 20		
	MCG	IEEP(EAS)	IEEP	MCG	IEEP(EAS)	IEEP
β_0	5,1	4,2	8,5	2,7	2,0	6,5
β_1	0,048	0,043	0,082	0,025	0,021	0,042
	n = 5			n = 20		
	F-C	IEEP(EAS)	IEEP	F-C	IEEP(EAS)	IEEP
σ_e^2	12,6	12,8	16,7	6,0	6,0	8,8
σ_v^2	30,7	28,2	38,2	15,0	13,3	28,8

5. APPLICATION À DES DONNÉES RÉELLES

Dans cette section, nous examinerons un ensemble de données réelles étudié par Battese, Harter et Fuller (1988) dans le contexte des estimations régionales (petites régions). Battese et coll. (1988) ont étudié les estimations du nombre moyen d'hectares ensemencés en maïs et en soya par segment dans le cas de 12 comtés du centre - nord de l'Iowa. Le nombre total de segments échantillonnés (taille globale d'échantillon) est de 36 pour ces comtés; la taille d'échantillon n_i $(i = 1,...,12)$ dans chaque comté varie de 1 à 5. Le nombre total de segments N_i (taille de population) dans chaque comté va de 402 à 965.

Dans nos calculs comme dans You et Rao (2002), nous avons supposé un échantillonnage aléatoire simple (EAS) à l'intérieur des régions. Ainsi, le poids d'échantillonnage de base pour chaque unité échantillonnée de la région i est $\tilde{w}_{ij} = N_i / n_i$ et $w_{ij} = n_i^{-1}$. Le modèle d'échantillonnage est

$$y_{ij} = \beta_0 + x_{1ij}\beta_1 + x_{2ij}\beta_2 + v_i + e_{ij}, \quad j = 1,...,n_i, i = 1,...,12,$$

où y_{ij} est le nombre d'hectares ensemencés en maïs (ou en soya) dans le j^e segment du i^e comté et où x_{1ij} et x_{2ij} sont respectivement le nombre de pixels caractérisés comme en maïs ou en soya dans le j^e segment du i^e comté. Notre but principal est d'estimer les effets fixes β et les composantes de variance σ_e^2 et σ_v^2.

Pour l'estimation des composantes de la variance, nous avons mis cinq méthodes en comparaison, à savoir les méthodes F-C (ajustement de constantes), MV (maximum de vraisemblance), MV restreinte ou en valeur résiduelle (MVR), IMCG et IEEP (méthode proposée). Le tableau 3 présente les estimations de σ_e^2 et σ_v^2 pour le maïs et le

soya pour ces cinq méthodes. Au tableau 3, les méthodes F-C et MVR livrent des estimations analogues, les méthodes MV et IMCG des estimations identiques et la méthode IEEP donne des résultats acceptables. Les estimations IEEP se situent entre les estimations F-C (MVR) et MV (IMCG) sauf pour σ_v^2 dans le cas du soya. Pour la méthode IEEP, nous avons pris les estimations F-C comme valeurs initiales. L'algorithme itératif IEEP converge très rapidement. Dans cet exemple, il n'a fallu que quelques itérations.

Tableau 3 : Estimation des composantes de la variance σ_e^2 et σ_v^2

	Paramètre	F-C	MVR	MV	IMCG	IEEP
Maïs	σ_e^2	149,6	147,3	137,4	137,3	139,4
	σ_v^2	139,7	139,9	120,9	121,1	130,2
Soya	σ_e^2	195,2	190,4	177,0	177,0	187,9
	σ_v^2	261,8	247,3	217,6	217,6	207,2

Le tableau 4 indique les estimations pondérées des effets fixes β à l'aide de (9) et en fonction d'estimateurs différents de σ_e^2 et σ_v^2, incluant les méthodes FC, MVR, IMCG et IEEP. Il ressort de ce tableau que les estimations ponctuelles se ressemblent fort. Les méthodes F-C et MVR donnent des erreurs-types légèrement supérieures à celles des méthodes IMCG et IEEP. Les résultats montrent que, en échantillonnage aléatoire simple (EAS), les estimations IEEP sont efficaces, ce qui concorde avec les résultats de simulation présentés à la section 4.

Tableau 4 : Estimation des effets fixes β

		Estimations				Erreurs-types			
	Coefficient	F-C	MVR	IMCG	IEEP	F-C	MVR	IMCG	IEEP
Maïs	β_0	58,491	58,481	58,526	58,492	27,122	26,933	25,939	26,185
	β_1	0,316	0,316	0,316	0,316	0,054	0,054	0,052	0,052
	β_2	-0,160	-0,160	-0,159	-0,160	0,062	0,061	0,059	0,060
Soya	β_0	-14,483	-14,388	-14,227	-13,907	31,396	30,974	29,805	30,588
	β_1	0,005	0,005	0,004	0,003	0,062	0,061	0,059	0,060
	β_2	0,514	0,514	0,515	0,515	0,072	0,071	0,068	0,070

6. CONCLUSION

Dans cet exposé, nous avons proposé une nouvelle méthode d'estimation des effets fixes et des composantes de la variance dans un modèle multiniveaux à valeur aléatoire à l'origine (modèle hiérarchisé d'erreurs de régression) à l'aide des poids d'échantillonnage. Nous avons comparé la méthode IEEP proposée à certaines méthodes existantes par une étude de simulation et une application à un ensemble de données réelles. Notre méthode IEEP est d'un calcul simple et converge rapidement. Elle fait appel tant à des données d'enquête qu'aux poids d'échantillonnage. En échantillonnage aléatoire simple, elle se révèle très efficace. En décrivant cette méthode, nous élargissons les travaux de You et Rao (2002) et de Waclawiw et Liang (1993) en ce qui concerne l'estimation des effets fixes et des composantes de la variance. La méthode que nous proposons peut servir à l'analyse de données d'enquête fondée sur un modèle en général et à l'estimation basée sur un modèle pour de petites régions en particulier.

RÉFÉRENCES

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

Fuller, W.A. et Battese, G.E. (1973) Transformation for estimation of linear models with nested error structure. *Journal of the American Statistical Association*, 68, 626-632.

Goldstein, H. (1995) Multilevel Statistical Models. London, Edward Arnold: New York, Wiley.

Henderson, C.R. (1953) Estimation of variance and covariance components. *Biometrics*, 9, 226-252.

Stukel, D. M. et Rao, J. N. K. (1997) Estimation of regression models with nested error structure and unequal error variances under two and three stage cluster sampling. *Statistics & Probability Letters*, 35, 401-407.

Waclawiw, M. A. et Liang, K. Y. (1993) Prediction of random effects in the generalized linear model. *Journal of the American Statistical Association*, 88, 171-178.

You, Y. et Rao, J.N.K. (2002) A pseudo empirical best linear unbiased prediction approach to small area estimation using survey weights. *La revue canadienne de statistique*, 30, 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*, 111, 197-208.