Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Comparaison d'estimateurs sur petits domaines au niveau de l'unité et au niveau du domaine

par Michael A. Hidiroglou et Yong You

Date de diffusion : le 22 juin 2016

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, juin 2016
Vol. 42, nº 1, pp. 45-67
Statistique Canada, nº 12-001-X au catalogue

Comparaison d'estimateurs sur petits domaines au niveau de l'unité et au niveau du domaine

Michael A. Hidiroglou et Yong You[1]

Résumé

Les auteurs comparent les estimateurs EBLUP et pseudo-EBLUP pour l'estimation sur petits domaines en vertu d'un modèle de régression à erreur emboîtée, ainsi que trois autres estimateurs fondés sur un modèle au niveau du domaine à l'aide du modèle de Fay-Herriot. Ils réalisent une étude par simulations fondée sur un plan de sondage pour comparer les estimateurs fondés sur un modèle pour des modèles au niveau de l'unité et au niveau du domaine sous un échantillonnage informatif et non informatif. Ils s'intéressent particulièrement aux taux de couverture des intervalles de confiance des estimateurs au niveau de l'unité et au niveau du domaine. Les auteurs comparent aussi les estimateurs sous un modèle dont la spécification est inexacte. Les résultats de la simulation montrent que les estimateurs au niveau de l'unité sont plus efficaces que les estimateurs au niveau du domaine. L'estimateur pseudo-EBLUP donne les meilleurs résultats à la fois au niveau de l'unité et au niveau du domaine.

Mots-clés : Intervalle de confiance; convergence sous le plan de sondage; modèle de Fay-Herriot; échantillonnage informatif; spécification inexacte du modèle; modèle de régression à erreur emboîtée; racine de l'erreur quadratique moyenne relative (REQMR); poids d'enquête.

1 Introduction

Ces dernières années, les chercheurs ont eu largement recours à des estimateurs sur petits domaines fondés sur des modèles pour obtenir des estimations indirectes fiables sur de petits domaines. Les estimateurs fondés sur des modèles reposent sur des modèles explicites qui établissent des liens avec de petits domaines connexes grâce à des données supplémentaires, comme des données de recensement et des données administratives. On peut classer les modèles d'estimation sur petits domaines en deux grandes catégories : (i) les modèles au niveau de l'unité, qui établissent des liens entre les valeurs unitaires de la variable étudiée et des variables auxiliaires propres à l'unité et (ii) les modèles au niveau du domaine, qui établissent des liens entre les estimateurs directs de la variable étudiée du petit domaine et les variables auxiliaires propres au domaine correspondantes. En général, les modèles au niveau du domaine servent à améliorer les estimateurs directs lorsqu'il n'y a pas de données disponibles au niveau de l'unité. L'échantillonnage est établi selon la méthode de Rao (2003), c'est-à-dire qu'un univers U de taille N est divisé en m petits domaines non chevauchants U_i de taille N_i, où $i = 1,\ldots,m$. L'échantillonnage est réalisé dans chaque petit domaine selon un mécanisme probabiliste pour produire des échantillons s_i de taille n_i. La probabilité de sélection associée à chaque élément $j = 1,\ldots,n_i$ sélectionné dans l'échantillon s_i est désignée par p_{ij}. Les poids de sondage qui en découlent sont donnés par $w_{ij} = n_i^{-1} p_{ij}^{-1}$. En pratique, ces poids peuvent être ajustés pour tenir compte de la non-réponse et de données auxiliaires. Les poids obtenus correspondent aux poids de l'enquête. Dans le présent article, on présume une réponse totale à l'enquête et aucun ajustement pour tenir compte de données auxiliaires. Les estimations directes au niveau du domaine pour chaque domaine sont obtenues à partir des poids de l'enquête et des unités observées dans

1. Michael A. Hidiroglou, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa (Ontario), K1A 0T6, Canada. Courriel : hidirog@yahoo.ca; Yong You, Division de la coopération internationale et des méthodes statistiques institutionnelles, Statistique Canada, Ottawa (Ontario), K1A 0T6, Canada. Courriel : yong.you@canada.ca.

le domaine. Le plan d'enquête peut être intégré de différentes manières aux modèles d'estimation sur petits domaines. Au niveau du domaine, les estimateurs directs fondés sur le plan de sondage sont modélisés directement et la variance de sondage de l'estimateur direct connexe est intégrée au modèle au moyen des erreurs fondées sur le plan de sondage. Au niveau de l'unité, les observations peuvent être pondérées à l'aide du poids de l'enquête. Un certain nombre de facteurs influent sur l'efficacité des estimateurs. Deux facteurs importants sont l'exactitude lors de la spécification du modèle et la corrélation entre la variable d'intérêt et les probabilités de sélection associées au processus d'échantillonnage, c'est-à-dire le caractère informatif du processus d'échantillonnage. Dans le présent article, les auteurs comparent, au moyen d'une étude par simulations, l'incidence de la spécification inexacte du modèle et du caractère informatif du plan d'échantillonnage pour deux méthodes de base utilisées pour l'estimation sur petits domaines au niveau de l'unité et au niveau du domaine en termes de biais, d'erreur quadratique moyenne estimée et de taux de couverture des intervalles de confiance. Un plan d'échantillonnage est informatif si les probabilités de sélection p_{ij} demeurent reliées à la variable d'intérêt y_{ij} même après conditionnement sur les covariables $\mathbf{x}_{ij}$. Dans un tel cas, on dit que l'échantillonnage est informatif parce que le modèle de population n'est plus vérifié pour l'échantillon. Pfeffermann et Sverchkov (2007) tiennent compte de cette possibilité en ajustant la méthode d'estimation sur petits domaines. Verret, Rao et Hidiroglou (2015) ont simplifié la méthode. Dans le présent article, les méthodes d'estimation sur petits domaines ne sont pas ajustées en fonction du caractère informatif; on étudie plutôt leur impact.

La présentation de l'article est la suivante. Les estimateurs ponctuels et les estimateurs de l'erreur quadratique moyenne associés pour les modèles d'estimation au niveau de l'unité et au niveau du domaine sont décrits à la section 2 et à la section 3 respectivement. La simulation et les résultats sont présentés à la section 4. La simulation calcule les estimateurs ponctuels et les erreurs quadratiques moyennes associées pour un plan d'échantillonnage avec probabilités proportionnelles à la taille avec remise (PPTAR) en faisant varier les deux facteurs suivants : (a) le modèle supposé est exact ou inexact, et (b) le plan de sondage est présumé non informatif ou très informatif. La section 5 présente un exemple d'utilisation des données de Battese, Harter et Fuller (1988) pour comparer les estimations au niveau de l'unité et au niveau du domaine. Enfin, les conclusions des travaux sont exposées à la section 6.

2 Modèle d'estimation au niveau de l'unité

L'un des modèles de base pour l'estimation sur petits domaines au niveau de l'unité est le modèle de régression à erreur emboîtée (Battese et coll. 1988) donné par $y_{ij} = \mathbf{x}'_{ij}\beta + v_i + e_{ij}$, $j = 1,\ldots,N_i$, $i = 1,\ldots,m$, où y_{ij} est la variable d'intérêt pour la j^{e} unité de population du i^{e} petit domaine, $\mathbf{x}_{ij} = \left(x_{ij1},\ldots,x_{ijp}\right)'$ est un vecteur $p \times 1$ de variables auxiliaires où $x_{ij1} = 1$, $\boldsymbol{\beta} = \left(\beta_0,\ldots,\beta_{p-1}\right)'$ est un vecteur $p \times 1$ de paramètres de régression et N_i est le nombre d'unités de population dans le i^{e} petit domaine. Les effets aléatoires v_i sont présumés indépendants et identiquement distribués $(i.i.d.)$ $N\left(0,\sigma_v^2\right)$ et indépendants des erreurs au niveau de l'unité e_{ij}, qui sont présumées $i.i.d.$ $N\left(0,\sigma_e^2\right)$. À supposer que N_i est grand, le paramètre d'intérêt correspond à la moyenne pour le i^{e} domaine, $\bar{Y}_i = N_i^{-1}\sum_{j=1}^{N_i} y_{ij}$, qui peut être approximée par :

$$\theta_i = \bar{\mathbf{X}}_i'\boldsymbol{\beta} + v_i, \tag{2.1}$$

où $\bar{\mathbf{X}}_i = \sum_{j=1}^{N_i} \mathbf{x}_{ij} \Big/ N_i$ est le vecteur des moyennes de population connues de $\mathbf{x}_{ij}$ pour le i^{e} domaine. On présume que les échantillons sont tirés indépendamment dans chaque petit domaine selon un plan d'échantillonnage spécifié. Sous un échantillonnage non informatif, les données d'échantillon $\left(y_{ij}, \mathbf{x}_{ij}\right)$ sont présumées obéir au modèle de population, c'est-à-dire

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + v_i + e_{ij}, \quad j = 1,\ldots,n_i, \quad i = 1,\ldots,m, \tag{2.2}$$

où w_{ij} est le poids de sondage de base associé à l'unité (i, j) et n_i est la taille de l'échantillon dans le i^{e} petit domaine.

2.1 Estimation EBLUP

Selon le modèle de régression à erreur emboîtée (2.2), l'estimateur de la meilleure prédiction linéaire sans biais (BLUP) de la moyenne d'un petit domaine, $\theta_i = \bar{\mathbf{X}}_i'\boldsymbol{\beta} + v_i$, est donné par

$$\tilde{\theta}_i = r_i \bar{y}_i + \left(\bar{\mathbf{X}}_i - r_i \bar{\mathbf{x}}_i\right)' \tilde{\boldsymbol{\beta}}, \tag{2.3}$$

où $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} \Big/ n_i$, $\bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij} \Big/ n_i$, $r_i = \sigma_v^2 \Big/ \left(\sigma_v^2 + \sigma_e^2 \Big/ n_i\right)$, et

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \bar{\mathbf{x}}_i' \mathbf{V}_i^{-1} \mathbf{x}_i\right)^{-1} \left(\sum_{i=1}^{m} \bar{\mathbf{x}}_i' \mathbf{V}_i^{-1} y_i\right) \equiv \tilde{\boldsymbol{\beta}}\left(\sigma_e^2, \sigma_v^2\right), \tag{2.4}$$

où $\mathbf{x}_i' = \left(x_{i1},\ldots,x_{in_i}\right)$, $\mathbf{V}_i = \sigma_e^2 \mathbf{I}_{n_i} + \sigma_v^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}'$, $y_i = \left(y_{i1},\ldots,y_{in_i}\right)'$, $i = 1,\ldots,m$. Les deux estimations $\tilde{\theta}_i$ et $\tilde{\boldsymbol{\beta}}$ dépendent des paramètres de variance inconnus σ_e^2 et σ_v^2. On peut utiliser la méthode d'ajustement des constantes pour estimer σ_e^2 et σ_v^2; les estimateurs résultants sont $\hat{\sigma}_e^2 = (n - m - p + 1)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{\varepsilon}_{ij}^2$ et $\hat{\sigma}_v^2 = \max\left(\tilde{\sigma}_v^2, 0\right)$, où $\tilde{\sigma}_v^2 = n_*^{-1}\left[\sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{u}_{ij}^2 - (n-p)\hat{\sigma}_e^2\right]$, $n_* = n - \text{tr}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1} \sum_{i=1}^{m} n_i^2 \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i'\right]$, $\mathbf{X}' = \left(x_1',\ldots,x_m'\right)$ et $n = \sum_{i=1}^{m} n_i$.

Les résidus $\left\{\hat{\varepsilon}_{ij}\right\}$ sont obtenus par la régression par les moindres carrés ordinaires (MCO) de $y_{ij} - \bar{y}_i$ sur $\left\{\mathbf{x}_{ij1} - \bar{\mathbf{x}}_{i\cdot 1},\ldots,\mathbf{x}_{ijp} - \bar{\mathbf{x}}_{i\cdot p}\right\}$ et les résidus $\left\{\hat{u}_{ij}\right\}$, par la régression par les MCO de y_{ij} sur $\left\{x_{ij1},\ldots,x_{ijp}\right\}$. Plus pour de détails, voir Rao (2003, page 138).

En remplaçant σ_e^2 et σ_v^2 par les estimateurs $\hat{\sigma}_e^2$ et $\hat{\sigma}_v^2$ dans l'équation (2.3), on obtient l'estimateur EBLUP de la moyenne de petit domaine θ_i suivant :

$$\hat{\theta}_i^{\text{EBLUP}} = r_i \bar{y}_i + \left(\bar{\mathbf{X}}_i - \hat{r}_i \bar{\mathbf{x}}_i\right)' \hat{\boldsymbol{\beta}}, \tag{2.5}$$

où $\hat{r}_i = \hat{\sigma}_v^2 \Big/ \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \Big/ n_i\right)$ et $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right)$. L'erreur quadratique moyenne (EQM) de l'estimateur EBLUP $\hat{\theta}_i^{\text{EBLUP}}$ est donnée par

$$\text{EQM}\left(\hat{\theta}_i^{\text{EBLUP}}\right) \approx g_{1i}\left(\sigma_e^2, \sigma_v^2\right) + g_{2i}\left(\sigma_e^2, \sigma_v^2\right) + g_{3i}\left(\sigma_e^2, \sigma_v^2\right)$$

voir Prasad et Rao (1990). Les termes g sont

$$\begin{aligned} g_{1i}\left(\sigma_e^2,\sigma_v^2\right) &= (1-r_i)\,\sigma_v^2, \\ g_{2i}\left(\sigma_e^2,\sigma_v^2\right) &= \left(\overline{\mathbf{X}}_i - r_i\overline{\mathbf{x}}_i\right)'\left(\sum_{i=1}^m \mathbf{x}_i'\mathbf{V}_i^{-1}\mathbf{x}_i\right)^{-1}\left(\overline{\mathbf{X}}_i - r_i\overline{\mathbf{x}}_i\right) \end{aligned}$$

et

$$g_{3i}\left(\sigma_e^2,\sigma_v^2\right) = n_i^{-2}\left(\sigma_v^2 + \sigma_e^2 n_i^{-1}\right)^{-3} h\left(\sigma_e^2,\sigma_v^2\right),$$

où $h\left(\sigma_e^2,\sigma_v^2\right) = \sigma_e^4 V\left(\tilde{\sigma}_v^2\right) - 2\sigma_e^2\sigma_v^2\,\mathrm{cov}\left(\hat{\sigma}_e^2,\tilde{\sigma}_v^2\right) + \sigma_v^4 V\left(\hat{\sigma}_e^2\right)$. Les variances et la covariance de $\hat{\sigma}_e^2$ et $\tilde{\sigma}_v^2$ sont données par

$$\begin{aligned} V\left(\hat{\sigma}_e^2\right) &= 2\left(n-m-p+1\right)^{-1}\sigma_e^4 \\ V\left(\tilde{\sigma}_v^2\right) &= 2n_*^{-2}\left[\left(n-m-p+1\right)^{-1}(m-1)(n-p)\,\sigma_e^4 + 2n_*\sigma_e^2\sigma_v^2 + n_{**}\sigma_v^4\right], \end{aligned}$$

et

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

où $n_{**} = \mathrm{tr}\left(\mathbf{Z}'\mathbf{M}\mathbf{Z}\right)^2$, $\mathbf{M} = \mathbf{I}_n - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$,– $\mathbf{Z} = \mathrm{diag}\left(\mathbf{1}_{n_1},\ldots,\mathbf{1}_{n_m}\right)$.

Un estimateur de deuxième ordre sans biais de l'EQM (Prasad et Rao 1990) est donné par

$$\mathrm{eqm}\left(\hat{\theta}_i^{\mathrm{EBLUP}}\right) = g_{1i}\left(\hat{\sigma}_e^2,\hat{\sigma}_v^2\right) + g_{2i}\left(\hat{\sigma}_e^2,\hat{\sigma}_v^2\right) + 2g_{3i}\left(\hat{\sigma}_e^2,\hat{\sigma}_v^2\right). \tag{2.6}$$

Soulignons que l'estimateur EBLUP $\hat{\theta}_i^{\mathrm{EBLUP}}$ donné par (2.5) dépend du modèle d'estimation au niveau de l'unité (2.2). Il est sans biais par rapport au modèle, mais il n'est pas convergent par rapport au plan de sondage sauf si ce dernier repose sur un échantillonnage aléatoire simple. Si le modèle (2.2) n'est plus vérifié pour les données échantillonnées, l'estimateur EBLUP $\hat{\theta}_i^{\mathrm{EBLUP}}$ peut alors être biaisé, c'est-à-dire qu'il comprend un biais additionnel attribuable à la spécification inexacte du modèle.

2.2 Estimation pseudo-EBLUP

You et Rao (2002) ont proposé un estimateur pseudo-EBLUP de la moyenne de petit domaine θ_i combinant les poids de l'enquête et le modèle d'estimation au niveau de l'unité (2.2) afin d'atteindre la convergence par rapport au plan. Soient w_{ij} les poids associés à chaque unité (i,j). Un estimateur direct fondé sur le plan de sondage de la moyenne de petit domaine est donné par

$$\overline{y}_{iw} = \frac{\sum_{j=1}^{n_i} w_{ij}y_{ij}}{\sum_{j=1}^{n_i} w_{ij}} = \sum_{j=1}^{n_i}\tilde{w}_{ij}y_{ij}, \tag{2.7}$$

où $\tilde{w}_{ij} = w_{ij}\Big/\sum_{j=1}^{n_i} w_{ij} = w_{ij}/w_{i\cdot}$ et $\sum_{j=1}^{n_i}\tilde{w}_{ij} = 1$. L'estimateur pondéré $\overline{y}_{iw}$ est aussi appelé « estimateur pondéré de Hájek ». En combinant l'estimateur direct (2.7) et le modèle d'estimation au niveau de l'unité (2.2), on peut obtenir le modèle au niveau du domaine agrégé (pondéré par les poids d'enquête) suivant :

$$\overline{y}_{iw} = \overline{\mathbf{x}}_{iw}'\boldsymbol{\beta} + v_i + \overline{e}_{iw}, \quad i = 1, \ldots, m, \tag{2.8}$$

où $\overline{e}_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{ij} e_{ij}$ avec $E(\overline{e}_{iw}) = 0$, $V(\overline{e}_{iw}) = \sigma_e^2 \sum_{j=1}^{n_i} \tilde{w}_{ij}^2 \equiv \delta_i^2$ et $\overline{\mathbf{x}}_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{ij} \mathbf{x}_{ij}$. Soulignons que le paramètre de régression $\boldsymbol{\beta}$ et les composantes de variance σ_e^2 et σ_v^2 ne sont pas connus dans le modèle (2.8). Selon le modèle (2.8), en supposant que les paramètres $\boldsymbol{\beta}$, σ_e^2 et σ_v^2 sont connus, l'estimateur BLUP de θ_i est donné par

$$\tilde{\theta}_{iw} = r_{iw}\overline{y}_{iw} + \left(\overline{\mathbf{X}}_i - r_{iw}\overline{\mathbf{x}}_{iw}\right)' \boldsymbol{\beta} = \tilde{\theta}_{iw}\left(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2\right), \tag{2.9}$$

où $r_{iw} = \sigma_v^2 / \left(\sigma_v^2 + \delta_i^2\right)$. L'estimateur BLUP $\tilde{\theta}_{iw}$ dépend de $\boldsymbol{\beta}$, σ_e^2 et σ_v^2. Pour estimer le paramètre de régression, You et Rao (2002) ont proposé une méthode d'équation d'estimation pondérée, qui permet d'obtenir un estimateur de $\boldsymbol{\beta}$ comme suit :

$$\tilde{\boldsymbol{\beta}}_w = \left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}\left(\mathbf{x}_{ij} - r_{iw}\overline{\mathbf{x}}_{iw}\right)'\right]^{-1}\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\left(\mathbf{x}_{ij} - r_{iw}\overline{\mathbf{x}}_{iw}\right) y_{ij}\right] \equiv \tilde{\boldsymbol{\beta}}_w\left(\sigma_e^2, \sigma_v^2\right).$$

$\tilde{\boldsymbol{\beta}}_w = \tilde{\boldsymbol{\beta}}_w\left(\sigma_e^2, \sigma_v^2\right)$ dépend de σ_e^2 et σ_v^2. En remplaçant σ_e^2 et σ_v^2 dans $\tilde{\boldsymbol{\beta}}_w$ par les estimateurs d'ajustement des constantes $\hat{\sigma}_e^2$ et $\hat{\sigma}_v^2$, on obtient $\hat{\boldsymbol{\beta}}_w = \tilde{\boldsymbol{\beta}}_w\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right)$; voir Rao (2003, page 149). En remplaçant $\boldsymbol{\beta}$, σ_e^2 et σ_v^2 dans (2.9) par $\hat{\boldsymbol{\beta}}_w$, $\hat{\sigma}_e^2$ et $\hat{\sigma}_v^2$, l'estimateur pseudo-EBLUP de la moyenne de petit domaine θ_i est donné par

$$\hat{\theta}_i^{P-\mathrm{EBLUP}} \triangleq \hat{\theta}_{iw} = \hat{r}_{iw}\overline{y}_{iw} + \left(\overline{\mathbf{X}}_i - \hat{r}_{iw}\overline{\mathbf{x}}_{iw}\right)' \hat{\boldsymbol{\beta}}_w. \tag{2.10}$$

À mesure que la taille de l'échantillon n_i augmente, l'estimateur $\hat{\theta}_i^{P-\mathrm{EBLUP}}$ devient convergent par rapport au plan de sondage. Il a aussi une propriété d'autocalage lorsque les poids w_{ij} sont ajustés de façon à correspondre au total de population connu. Ainsi, si $\sum_{j=1}^{n_i} w_{ij} = N_i$, $\sum_{i=1}^{m} N_i \hat{\theta}_i^{P-\mathrm{EBLUP}}$ correspond à l'estimateur direct de régression du total global

$$\sum_{i=1}^{m} N_i \hat{\theta}_i^{P-\mathrm{EBLUP}} = \hat{Y}_w + \left(\mathbf{X} - \hat{\mathbf{X}}_w\right)' \hat{\boldsymbol{\beta}}_w,$$

où $\hat{Y}_w = \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij} y_{ij}$, et $\hat{\mathbf{X}}_w = \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}$. Pour plus de détails, voir You et Rao (2002).

L'EQM de $\hat{\theta}_i^{P-\mathrm{EBLUP}}$ est donnée par

$$\mathrm{EQM}\left(\hat{\theta}_i^{P-\mathrm{EBLUP}}\right) \approx g_{1iw}\left(\sigma_e^2, \sigma_v^2\right) + g_{2iw}\left(\sigma_e^2, \sigma_v^2\right) + g_{3iw}\left(\sigma_e^2, \sigma_v^2\right),$$

où $g_{1iw}\left(\sigma_e^2, \sigma_v^2\right) = (1 - r_{iw})\sigma_v^2$ et $g_{2iw}\left(\sigma_e^2, \sigma_v^2\right) = \left(\overline{X}_i - r_{iw}\overline{x}_{iw}\right)'\Phi_w\left(\overline{X}_i - r_{iw}\overline{x}_{iw}\right)$. Le terme Φ_w est

$$\Phi_w\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}z_{ij}'\right)^{-1}\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{z}_{ij}\mathbf{z}_{ij}'\right)\left[\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\mathbf{z}_{ij}'\right)^{-1}\right]'\sigma_e^2$$

$$+ \left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\mathbf{z}_{ij}'\right)^{-1}\left[\sum_{i=1}^{m}\left(\sum_{j=1}^{n_i}\mathbf{z}_{ij}\right)\left(\sum_{j=1}^{n_i}\mathbf{z}_{ij}\right)'\right]\left[\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}x_{ij}\mathbf{z}_{ij}\right)^{-1}\right]'\sigma_v^2,$$

où $\mathbf{z}_{ij} = w_{ij}\left(\mathbf{x}_{ij} - r_{iw}\overline{\mathbf{x}}_{iw}\right)$ et $g_{3iw}\left(\sigma_e^2, \sigma_v^2\right) = r_{iw}\left(1 - r_{iw}\right)^2 \sigma_e^{-4}\sigma_v^{-2} h\left(\sigma_e^2, \sigma_v^2\right)$. Le facteur $h\left(\sigma_e^2, \sigma_v^2\right)$ correspond à la même fonction que pour l'EQM de l'estimateur EBLUP donné à la section 2.1. Un estimateur de deuxième ordre presque sans biais de l'EQM peut s'écrire

$$\text{eqm}\left(\hat{\theta}_i^{P-\text{EBLUP}}\right) = g_{1iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + g_{2iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + 2g_{3iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right). \tag{2.11}$$

(Voir Rao 2003, page 150 et You et Rao 2002, page 435). Soulignons que l'estimateur de l'EQM (2.11) ne tient pas compte des termes du produit vectoriel. Torabi et Rao (2010) ont obtenu un estimateur de deuxième ordre de l'EQM exact tenant compte des termes du produit vectoriel à l'aide des méthodes de linéarisation et de « bootstrap ». Le produit vectoriel compte deux termes. Le premier est simple et a une forme explicite. Bien que la méthode de linéarisation fonctionne bien, la forme explicite du deuxième terme du produit vectoriel est très longue; de plus, les formules fondées sur la méthode de linéarisation ne sont pas fournies dans l'article de Torabi et Rao (2010). La méthode « bootstrap » sous-estime toujours l'EQM réelle. Pour obtenir un estimateur non biaisé de l'EQM, il faut appliquer une méthode bootstrap double exigeant beaucoup de calculs. L'estimateur de l'EQM (2.11) se comporte comme l'estimateur par linéarisation de Torabi et Rao (2010) lorsque la variation des poids d'enquête est faible. Dans le cas de l'autopondération à l'intérieur des domaines, l'un des termes du produit vectoriel est zéro et l'autre est de l'ordre $o\left(m^{-1}\right)$. En conséquence, l'estimateur de l'EQM (2.11) est presque sans biais; d'autres détails sont présentés dans Torabi et Rao (2010). C'est pour ces raisons que les termes du produit vectoriel n'ont pas été inclus dans l'estimateur de l'EQM donné en (2.11) dans le cadre de l'étude.

Soulignons qu'en vertu du modèle (2.2), l'estimateur pseudo-EBLUP $\hat{\theta}_i^{P-\text{EBLUP}}$ est légèrement moins efficient que l'estimateur EBLUP $\hat{\theta}_i^{\text{EBLUP}}$. Toutefois, cet estimateur pseudo-EBLUP est convergent par rapport au plan et est donc plus robuste à une spécification inexacte du modèle. L'efficacité des estimateurs EBLUP et pseudo-EBLUP a été évaluée à l'aide d'une étude par simulations.

3 Modèle d'estimation au niveau du domaine

Le modèle de Fay-Herriot (Fay et Herriot 1979) est un modèle d'estimation au niveau du domaine de base couramment utilisé pour l'estimation sur petits domaines afin d'améliorer les estimations d'enquête directes. Le modèle de Fay-Herriot a deux composantes, soit un modèle d'échantillonnage pour les estimations d'enquête directes et un modèle de lien pour les paramètres d'intérêt du petit domaine. Le modèle d'échantillonnage suppose que pour une taille d'échantillon de domaine spécifique $n_i > 1$, il existe un estimateur d'enquête direct $\hat{\theta}_i^{\text{DIR}}$. Cet estimateur d'enquête direct est sans biais sous le plan pour le paramètre de petit domaine θ_i. Le modèle d'échantillonnage est donné par

$$\hat{\theta}_i^{\text{DIR}} = \theta_i + e_i,\ i = 1, \ldots, m, \tag{3.1}$$

où e_i est l'erreur d'échantillonnage associée à l'estimateur direct $\hat{\theta}_i^{\text{DIR}}$ et m est le nombre de petits domaines. Dans la pratique, il est courant de supposer que les variables e_i sont des variables aléatoires normales indépendantes de moyenne $E\left(e_i\right) = 0$ et de variance d'échantillonnage $\text{var}\left(e_i\right) = \sigma_i^2$. Le modèle de lien est obtenu en supposant que le paramètre de petit domaine d'intérêt θ_i est lié aux variables auxiliaires au niveau du domaine $\mathbf{z}_i = \left(\text{z}_{i1}, \ldots, \text{z}_{ip}\right)'$ par le modèle de régression linéaire suivant :

$$\theta_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i,\ i = 1,\ldots,m, \tag{3.2}$$

où $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)'$ est un vecteur $p \times 1$ de coefficients de régression et où les termes v_i sont des effets aléatoires propres au domaine présumés être $i.i.d.$, avec $E(v_i) = 0$ et $\text{var}(v_i) = \sigma_v^2$. On émet aussi généralement une hypothèse de normalité, même si elle est plus difficile à justifier. Une telle hypothèse est nécessaire pour obtenir l'estimation de l'EQM. La variance du modèle σ_v^2 est inconnue et doit être estimée à partir des données. L'effet aléatoire au niveau du domaine v_i rend compte de l'hétérogénéité non structurée entre les domaines que n'expliquent pas les variances d'échantillonnage. La combinaison des modèles (3.1) et (3.2) produit un modèle linéaire mixte au niveau du domaine donné par

$$\hat{\theta}_i^{\text{DIR}} = \mathbf{z}_i'\boldsymbol{\beta} + v_i + e_i. \tag{3.3}$$

Le modèle (3.3) comprend des erreurs aléatoires fondées sur le plan e_i et des effets aléatoires fondés sur le modèle v_i. Aux fins du modèle de Fay-Herriot, la variance d'échantillonnage σ_i^2 est présumée être connue dans le modèle (3.3). Il s'agit d'une hypothèse très forte. On utilise généralement des estimateurs lissés des variances d'échantillonnage dans le modèle de Fay-Herriot; les paramètres σ_i^2 sont ensuite considérés comme connus. Toutefois, si des estimateurs directs des variances d'échantillonnage sont utilisés dans le modèle de Fay-Herriot, il faut ajouter un terme à l'estimateur de l'EQM pour tenir compte de la variation additionnelle (Wang et Fuller 2003).

Si l'on suppose que la variance du modèle σ_v^2 est connue, le meilleur prédicteur linéaire sans biais (BLUP) du paramètre de petit domaine θ_i peut s'écrire

$$\tilde{\theta}_i = \gamma_i \hat{\theta}_i^{\text{DIR}} + (1-\gamma_i)\,\mathbf{z}_i'\tilde{\boldsymbol{\beta}}_{\text{MCP}}, \tag{3.4}$$

où $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \sigma_i^2)$, et $\tilde{\boldsymbol{\beta}}_{\text{MCP}}$ est l'estimateur des moindres carrés pondérés (MCP) de $\boldsymbol{\beta}$ donné par

$$\tilde{\boldsymbol{\beta}}_{\text{MCP}} = \left[\sum_{i=1}^{m} (\sigma_i^2 + \sigma_v^2)^{-1}\mathbf{z}_i\mathbf{z}_i'\right]^{-1}\left[\sum_{i=1}^{m} (\sigma_i^2 + \sigma_v^2)^{-1}\mathbf{z}_i y_i\right] = \left[\sum_{i=1}^{m} \gamma_i \mathbf{z}_i\mathbf{z}_i'\right]^{-1}\left[\sum_{i=1}^{m} \gamma_i \mathbf{z}_i y_i\right].$$

Il existe plusieurs méthodes pour estimer la variance du modèle inconnue σ_v^2; You (2010) présente une vue d'ensemble de ces méthodes. Les auteurs ont choisi la méthode du maximum de vraisemblance restreint (méthode REML) mise au point par Cressie (1992) pour estimer la variance du modèle en vertu du modèle de Fay-Herriot. À l'aide de l'algorithme de score, on obtient l'estimateur REML $\hat{\sigma}_v^2$ suivant :

$$\sigma_v^{2(k+1)} = \sigma_v^{2(k)} + \left[I_R\left(\sigma_v^{2(k)}\right)\right]^{-1} S_R\left(\sigma_v^{2(k)}\right),\ \text{pour}\ k = 1,2,\ldots,$$

où $I_R(\sigma_v^2) = 1/2\,\text{tr}[\mathbf{PP}]$, et $S_R(\sigma_v^2) = 1/2\,\mathbf{y}'\mathbf{PPy} - 1/2\,\text{tr}[\mathbf{P}]$, et $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{V}^{-1}$. Si on utilise une valeur supposée pour $\sigma_v^{2(1)}$ comme valeur de départ, l'algorithme converge très rapidement.

En remplaçant σ_v^2 dans l'équation (3.4) par l'estimateur REML $\hat{\sigma}_v^2$, on obtient l'estimateur EBLUP du paramètre de petit domaine θ_i fondé sur le modèle de Fay-Herriot suivant :

$$\hat{\theta}_i^{\text{FH}} = \hat{\gamma}_i \hat{\theta}_i^{\text{DIR}} + (1-\hat{\gamma}_i)\,\mathbf{z}_i'\hat{\boldsymbol{\beta}}_{\text{MCP}}, \tag{3.5}$$

où $\hat{\gamma}_i = \hat{\sigma}_v^2 \big/ \left(\hat{\sigma}_v^2 + \sigma_i^2\right)$. L'estimateur de l'EQM de $\hat{\theta}_i^{\mathrm{FH}}$ est donné par (voir Rao 2003)

$$\mathrm{eqm}\left(\hat{\theta}_i^{\mathrm{FH}}\right) = g_{1i} + g_{2i} + 2g_{3i}, \tag{3.6}$$

où g_{1i} est le terme principal, g_{2i} rend compte de la variabilité attribuable à l'estimation du paramètre de régression β, et g_{3i} est attribuable à l'estimation de la variance du modèle. Ces termes g sont définis comme suit :

$$g_{1i} = \hat{\gamma}_i \sigma_i^2, g_{2i} = (1-\hat{\gamma}_i)^2 \mathbf{z}_i' \mathrm{var}\left(\hat{\boldsymbol{\beta}}_{\mathrm{MCP}}\right) \mathbf{z}_i = \hat{\sigma}_v^2 (1-\hat{\gamma}_i)^2 \mathbf{z}_i' \left(\sum_{i=1}^{m} \hat{\gamma}_i \mathbf{z}_i \mathbf{z}_i'\right)^{-1} \mathbf{z}_i$$

et $g_{3i} = \left(\sigma_i^2\right)^2 \left(\hat{\sigma}_v^2 + \sigma_i^2\right)^{-3} \mathrm{var}\left(\hat{\sigma}_v^2\right)$.

La variance estimée de $\hat{\sigma}_v^2$ est donnée par $\mathrm{var}\left(\hat{\sigma}_v^2\right) = 2\left(\sum_{i=1}^{m} \left(\hat{\sigma}_v^2 + \sigma_i^2\right)^{-2}\right)^{-1}$; voir Datta et Lahiri (2000).

Jusqu'à maintenant, on a supposé que la variance d'échantillonnage σ_i^2 est présumée connue sous le modèle de Fay-Herriot (3.3). Il s'agit d'une hypothèse très forte. En règle générale, on connaît un estimateur d'enquête direct, disons s_i^2, de la variance d'échantillonnage σ_i^2. Comme ces variances estimées peuvent être très variables, elles sont lissées au moyen de modèles externes et de fonctions généralisées de variance; ces variances lissées sont désignées $\tilde{s}_i^2$. Les estimations lissées de la variance d'échantillonnage $\tilde{s}_i^2$ sont utilisées dans le modèle de Fay-Herriot et considérées comme connues. La valeur $\mathrm{eqm}\left(\hat{\theta}_i^{\mathrm{FH}}\right)$ associée est obtenue en remplaçant σ_i^2 par $\tilde{s}_i^2$ dans l'équation (3.6). Rivest et Vandal (2003) et Wang et Fuller (2003) ont étudié l'estimation de petit domaine sous le modèle de Fay-Herriot à l'aide des estimations directes de la variance d'échantillonnage s_i^2 en vertu de l'hypothèse que les estimateurs s_i^2 sont indépendants des estimateurs d'enquête directs y_i et $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$, où $d_i = n_i - 1$ et n_i est la taille de l'échantillon pour le i^{e} domaine. Quand on utilise l'estimation directe de la variance d'échantillonnage s_i^2 au lieu de la variance d'échantillonnage réelle σ_i^2, un terme additionnel rendant compte de l'incertitude liée à l'utilisation de s_i^2 doit être intégré à l'estimateur de l'EQM (3.6); ce terme, désigné par g_{4i}, est donné par

$$g_{4i} = \frac{4}{n_i - 1} \frac{\hat{\sigma}_v^4 s_i^4}{\left(\hat{\sigma}_v^2 + s_i^2\right)^3};$$

voir Rivest et Vandal (2003) et Wang et Fuller (2003) pour les détails.

Pour appliquer le modèle de Fay-Herriot, il faut obtenir les estimations directes au niveau du domaine et les estimations de la variance d'échantillonnage correspondantes, qui serviront de valeurs d'entrée au modèle de Fay-Herriot. On tient compte de trois estimateurs directs au niveau du domaine, soit l'estimateur direct de la moyenne de l'échantillon en supposant un échantillonnage aléatoire simple (EAS), l'estimateur de Horvitz-Thompson (HT) et l'estimateur pondéré de Hájek (HA). L'estimateur pondéré de Hájek est aussi utilisé dans l'estimateur pseudo-EBLUP pour le modèle d'estimation au niveau de l'unité désigné par $\bar{y}_{iw}$ dans l'équation (2.7). Le tableau 3.1 présente ces trois estimateurs directs au niveau du domaine et les estimateurs de la variance d'échantillonnage correspondants.

Tableau 3.1
Estimateurs directs au niveau du domaine et variances d'échantillonnage

	Estimateur ponctuel	Estimateur de la variance d'échantillonnage
Moyenne directe (EAS)	$\hat{\theta}_i^{\text{EAS}} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$	$\text{var}\left(\hat{\theta}_i^{\text{EAS}}\right) = \frac{1}{n_i\,(n_i-1)}\sum_{j=1}^{n_i}\left(y_{ij}-\hat{\theta}_i^{\text{EAS}}\right)^2$
Estimateur de Horvitz-Thompson (HT)	$\hat{\theta}_i^{\text{HT}} = \frac{1}{N_i}\sum_{j=1}^{n_i} w_{ij}y_{ij} = \frac{1}{N_i}\sum_{j=1}^{n_i}\frac{y_{ij}}{n_i p_{ij}}$	$\text{var}\left(\hat{\theta}_i^{\text{HT}}\right) = \frac{1}{N_i^2 n_i\,(n_i-1)}\sum_{j=1}^{n_i}\left(\frac{y_{ij}}{p_{ij}} - N_i\hat{\theta}_i^{\text{HT}}\right)^2$
Estimateur pondéré de Hájek (HA)	$\hat{\theta}_i^{\text{HA}} = \frac{\sum_{j=1}^{n_i} w_{ij}y_{ij}}{\sum_{j=1}^{n_i} w_{ij}} = \frac{1}{\hat{N}_i}\sum_{j=1}^{n_i}\frac{y_{ij}}{n_i p_{ij}}$	$\text{var}\left(\hat{\theta}_i^{\text{HA}}\right) = \frac{1}{\hat{N}_i^2 n_i\,(n_i-1)}\sum_{j=1}^{n_i}\left(\frac{y_{ij}-\hat{\theta}_i^{\text{HA}}}{p_{ij}}\right)^2$

Ces estimateurs au niveau du domaine servent de valeurs d'entrée au modèle de Fay-Herriot. Les trois estimateurs fondés sur le modèle au niveau du domaine sont désignés comme suit : FH-EAS, FH-HT et FH-HA. On remplace donc $\hat{\theta}_i^{\text{DIR}}$ par $\hat{\theta}_i^{\text{EAS}}$, $\hat{\theta}_i^{\text{HT}}$ ou $\hat{\theta}_i^{\text{HA}}$ dans (3.5) pour obtenir l'estimateur fondé sur le modèle correspondant $\hat{\theta}_i^{\text{FH-EAS}}$, $\hat{\theta}_i^{\text{FH-HT}}$ ou $\hat{\theta}_i^{\text{FH-HA}}$. L'estimateur direct sous EAS $\hat{\theta}_i^{\text{EAS}}$ ne tient pas compte du plan d'échantillonnage et ne converge pas par rapport au plan, sauf si ce dernier repose sur un échantillonnage aléatoire simple. Soulignons que les estimateurs $\hat{\theta}_i^{\text{HT}}$ et $\hat{\theta}_i^{\text{HA}}$ convergent par rapport au plan. Il s'ensuit que les estimateurs fondés sur le modèle correspondants $\hat{\theta}_i^{\text{FH-HT}}$ et $\hat{\theta}_i^{\text{FH-HA}}$ convergent par rapport au plan à mesure que la taille de l'échantillon augmente. En outre, cela signifie que ces estimateurs sont robustes à la spécification inexacte du modèle.

Dans la section qui suit, on compare le modèle au niveau de l'unité avec le modèle de Fay-Herriot au moyen d'une étude par simulations. Les paramètres statistiques utilisés pour ces comparaisons sont le biais, la racine de l'EQM relative et les intervalles de confiance des estimateurs fondés sur le modèle.

4 Étude par simulations

4.1 Génération des données

Pour comparer les estimateurs sur petits domaines au niveau de l'unité et au niveau du domaine, les auteurs ont réalisé une étude par simulations fondée sur le plan. À partir des conditions de simulation de You, Rao et Kovacevic (2003), deux populations finies ont été établies. Chaque population finie comptait $m = 30$ domaines, et chaque domaine consistait en $N_i = 200$ unités de population. Chacune des populations finies a été produite à l'aide du modèle d'estimation au niveau de l'unité $y_{ij} = \beta_0 + x_{1ij}\beta_1 + v_i + e_{ij}$. La variable auxiliaire x_{1ij} a été générée selon une loi exponentielle de moyenne 4 et de variance 8, et les composantes aléatoires ont été générées selon une loi normale avec $v_i \sim N\left(0, \sigma_v^2\right)$ et $e_{ij} \sim N\left(0, \sigma_e^2\right)$, où $\sigma_v^2 = 100$ et $\sigma_e^2 = 225$. Pour la première population, les effets fixes de régression ont été établis à $\beta_0 = 50$ et $\beta_1 = 10$ pour les 30 domaines. Pour la deuxième population, des effets fixes de différentes valeurs ont été utilisés : $\beta_0 = 50$ et $\beta_1 = 10$ pour les domaines $m = 1,\ldots,10$; $\beta_0 = 75$ et $\beta_1 = 15$ pour les domaines $m = 11,\ldots,20$; et $\beta_0 = 100$ et $\beta_1 = 20$ pour les domaines $m = 21,\ldots,30$. Il y avait trois moyennes différentes pour les effets fixes $\beta_0 + x_{1ij}\beta_1$ dans la deuxième population, alors qu'il n'y en avait qu'une dans la première population. Les échantillons PPTAR dans chaque domaine ont été tirés

indépendamment de chaque population construite. Pour mettre en œuvre l'échantillonnage PPTAR, on a d'abord défini une mesure de taille z_{ij} pour une unité donnée (i, j). À l'aide de ces valeurs z_{ij}, on a calculé les probabilités de sélection $p_{ij} = z_{ij} \Big/ \sum_j z_{ij}$ pour chaque unité (i, j), qui ont ensuite servi à sélectionner des échantillons PPTAR de tailles égales $n_i = n$. Dans chaque population générée, on a sélectionné des échantillons de taille $n = 10$ et 30. Le poids de sondage de base est donné par $w_{ij} = n_i^{-1} p_{ij}^{-1}$, de sorte que le poids normalisé correspond à $\tilde{w}_{ij} = p_{ij}^{-1} \Big/ \sum_j p_{ij}^{-1}$. On a choisi la mesure de taille z_{ij} comme une combinaison linéaire de la variable auxiliaire x_{1ij} et des données produites selon une loi exponentielle de moyenne 4 et de variance 16. Le coefficient de corrélation ρ entre y_{ij} et la probabilité de sélection p_{ij} dans chaque domaine variait entre 0,02 et 0,95. La fourchette des probabilités p_{ij} va de la sélection non informative $(\rho = 0{,}02)$ à la sélection fortement informative $(\rho = 0{,}95)$ des échantillons PPTAR. L'échantillonnage est non informatif lorsque la corrélation entre y_{ij} et la probabilité de sélection p_{ij} est très faible, ce qui signifie que l'échantillon et le modèle de population coïncident. Si la probabilité de sélection p_{ij} est fortement corrélée avec l'observation y_{ij}, l'échantillonnage est informatif, et le modèle de population n'est peut-être plus vérifié pour l'échantillon. Pour chaque population, le processus d'échantillonnage PPTAR a été répété $R = 3\,000$ fois. Comme dans Prasad et Rao (1990), l'étude par simulations est fondée sur le plan, puisque les deux populations ont été générées une seule fois, et que des échantillons répétés ont été produits à partir de la même population.

Pour la modélisation au niveau de l'unité, on a ajusté le modèle de régression à erreur emboîtée en fonction des données d'échantillonnage PPTAR générées à partir de chaque population. On a obtenu les estimations EBLUP et pseudo-EBLUP correspondantes ainsi que les estimations de l'EQM à l'aide des formules énoncées à la section 2. On a ensuite établi les estimations des intervalles de confiance en calculant la racine carrée des estimations de l'EQM; les détails du calcul sont présentés à la section 4.2.3. Pour la modélisation au niveau du domaine, on a d'abord calculé les estimations directes au niveau du domaine $\hat{\theta}_i^{\mathrm{EAS}}$, $\hat{\theta}_i^{\mathrm{HT}}$ et $\hat{\theta}_i^{\mathrm{HA}}$ ainsi que les variances d'échantillonnage correspondantes. On a ensuite appliqué le modèle de Fay-Herriot pour obtenir les estimateurs fondés sur le modèle $\hat{\theta}_i^{\mathrm{FH\text{-}EAS}}$, $\hat{\theta}_i^{\mathrm{FH\text{-}HT}}$ et $\hat{\theta}_i^{\mathrm{FH\text{-}HA}}$. La moyenne de population de la variable auxiliaire x_{1ij} de chaque domaine a été utilisée comme variable auxiliaire dans le modèle de Fay-Herriot. On a additionné g_{4i} à l'estimateur de l'EQM pour tenir compte du recours à des variances d'échantillonnage non lissées dans le modèle de Fay-Herriot. Les intervalles de confiance correspondants ont été obtenus de façon similaire pour les estimateurs EBLUP et pseudo-EBLUP au niveau de l'unité.

L'ajustement du modèle aux deux niveaux (unité et domaine) est fondé sur deux scénarios. Le premier (scénario I) suppose que la modélisation est exacte; les données ont été générées à partir de la première population et les modèles d'ajustement étaient le modèle au niveau de l'unité (2.2) et le modèle au niveau du domaine (3.3), tous deux avec le même vecteur $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Le deuxième (scénario II) suppose que la modélisation est inexacte; les données ont été générées à partir de la deuxième population avec des moyennes différentes pour les effets fixes et les mêmes modèles d'ajustement que pour le scénario I, avec un même vecteur $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Soulignons qu'en vertu du scénario I, l'échantillonnage n'est pas informatif lorsque le niveau d'unité exact (2.2) est ajusté en fonction des données de l'échantillon pour obtenir l'estimateur EBLUP; cela est vrai pour tous les coefficient de corrélation ρ entre y_{ij} et p_{ij}.

4.2 Résultats

Dans la section qui suit, on compare certaines données statistiques des estimations au niveau de l'unité et au niveau du domaine en vertu du scénario I (modélisation exacte) et du scénario II (modélisation inexacte).

4.2.1 Comparaison à l'intérieur de chaque petit domaine

À la figure 4.1, on compare les moyennes de population avec les estimations au niveau de l'unité et au niveau du domaine lorsque $n = 10$ pour le scénario I. Les résultats sont fondés sur un plan d'échantillonnage fortement informatif où le coefficient de corrélation entre y_{ij} et la probabilité de sélection p_{ij} est $\rho = 0{,}88$. Les estimations fondées sur le modèle reposent sur la moyenne de $R = 3\,000$ simulations. Les résultats présentés à la figure 4.1 indiquent clairement que les estimateurs EBLUP (équation 2.5) et pseudo-EBLUP (équation 2.10) au niveau de l'unité sont presque sans biais. Les résultats montrent que si la modélisation est exacte, l'échantillonnage n'est pas informatif pour le modèle au niveau de l'unité (2.2) et l'estimateur EBLUP est sans biais. L'estimateur FH-EAS au niveau du domaine surestime systématiquement la moyenne de population, ce qui entraîne un biais important. L'estimateur FH-HT au niveau du domaine sous-estime généralement la moyenne de population et entraîne un biais légèrement plus grand que celui de l'estimateur FH-HA. On obtient des résultats similaires pour $n = 30$.

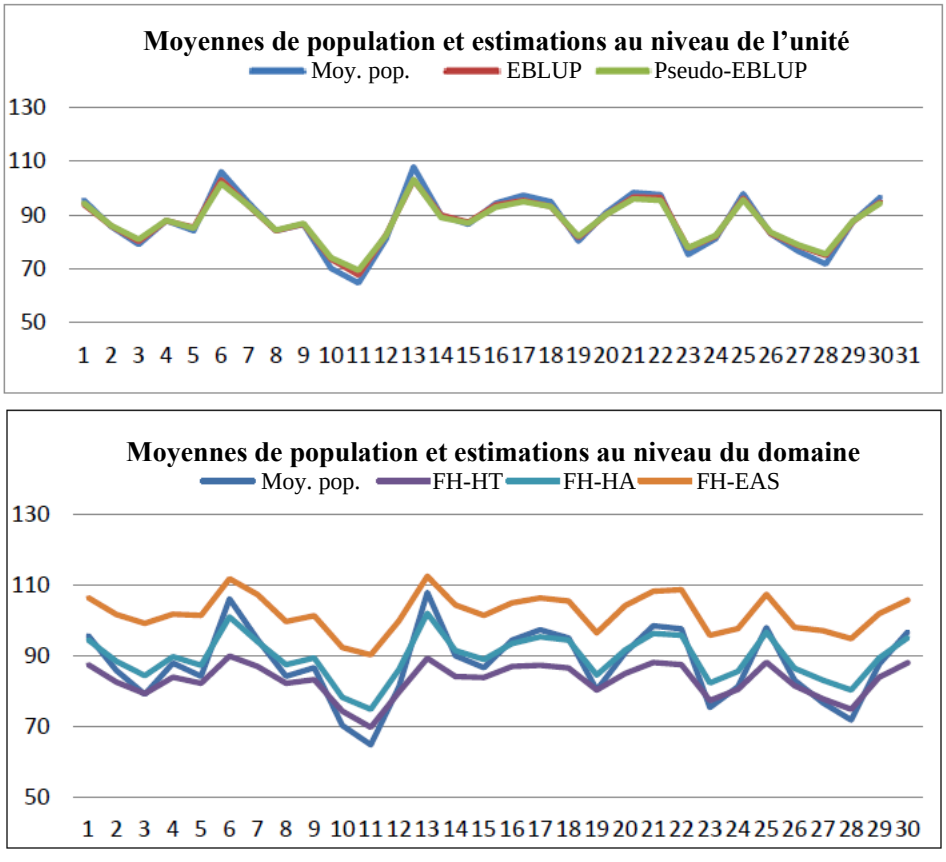

Figure 4.1 Comparaison des moyennes en vertu du scénario I pour $n = 10$.

À la figure 4.2, on compare la racine carrée moyenne de l'eqm pour les estimateurs au niveau de l'unité et au niveau du domaine pour le scénario I lorsque $n = 10$ et $n = 30$. La racine de l'eqm correspond à la racine carrée de l'EQM estimée donnée aux sections 2 et 3 pour les estimateurs au niveau de l'unité et au niveau du domaine. Il est clair que la racine de l'eqm des estimateurs EBLUP et pseudo-EBLUP est beaucoup plus faible que celle des estimateurs FH au niveau du domaine pour $n = 10$ et $n = 30$. Comme on s'y attendait (You et Rao 2002), l'estimateur EBLUP a la plus petite racine de l'eqm et l'estimateur pseudo-EBLUP, une racine de l'eqm légèrement supérieure. Les estimateurs FH-EAS au niveau du domaine ont une racine de l'eqm élevée et affichent des variations importantes. Les estimateurs FH-HT et FH-HA ont en moyenne à peu près la même racine de l'eqm, mais, comme l'illustrent les deux figures, l'estimateur FH-HT est plus variable que l'estimateur FH-HA, particulièrement lorsque $n = 10$. Lorsque $n = 30$, la variabilité de la racine de l'eqm pour les estimateurs FH-HT et FH-HA est considérablement réduite, mais il est clair que l'estimateur FH-HA est plus stable que l'estimateur FH-HT.

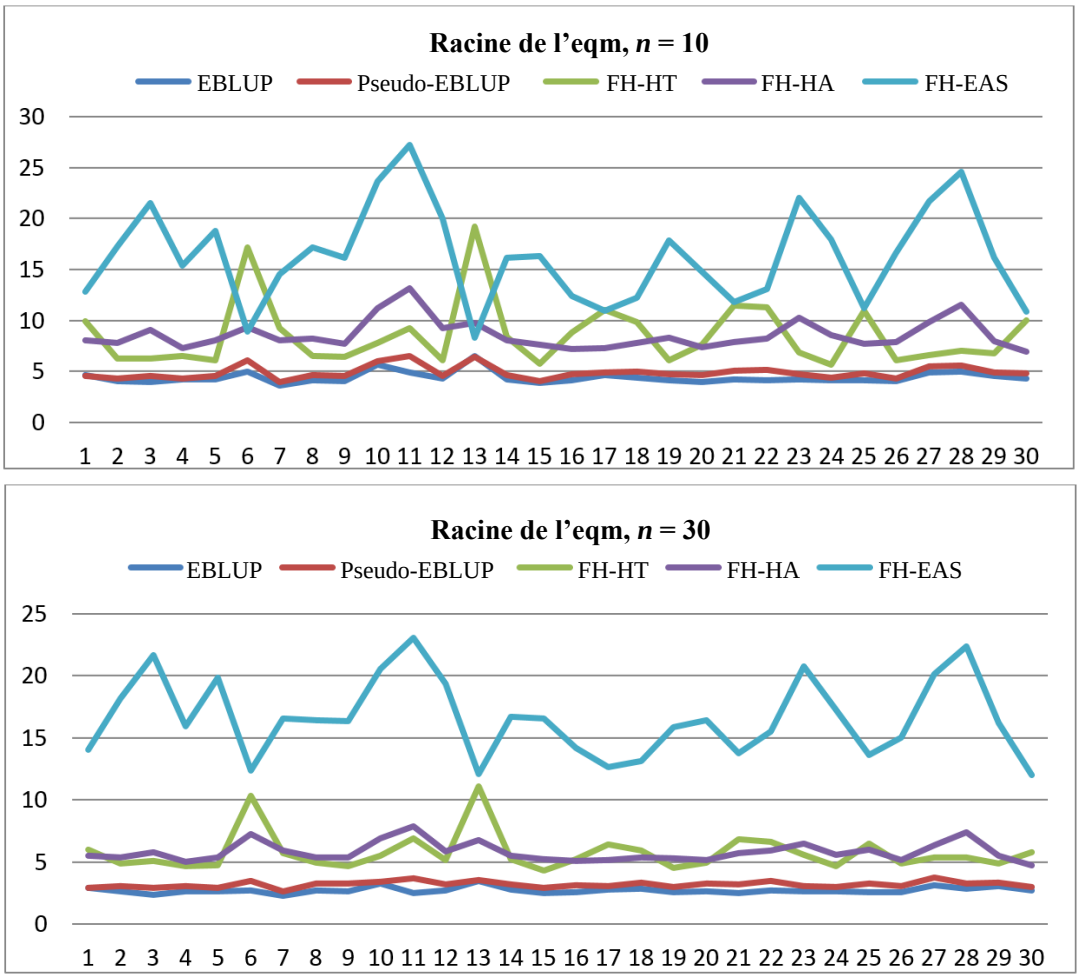

Figure 4.2 Comparaison de la racine de l'eqm en vertu du scénario I pour $n = 10$ et $n = 30$.

À la figure 4.3, on compare les estimations au niveau de l'unité et au niveau du domaine avec les moyennes de population lorsque $n = 10$ en vertu du scénario II. Dans le cas des modèles au niveau de

l'unité, il est clair que l'estimateur EBLUP sous-estime et surestime à la fois la moyenne de population lorsque la spécification du modèle est inexacte, alors que l'estimateur pseudo-EBLUP est sans biais (les estimations pseudo-EBLUP et les moyennes de population sont superposées à la figure 4.3). En ce qui concerne les estimateurs au niveau du domaine, l'estimateur FH-EAS surestime systématiquement les moyennes réelles, alors que l'estimateur FH-HT sous-estime davantage les valeurs que l'estimateur FH-HA lorsque la spécification du modèle est inexacte.

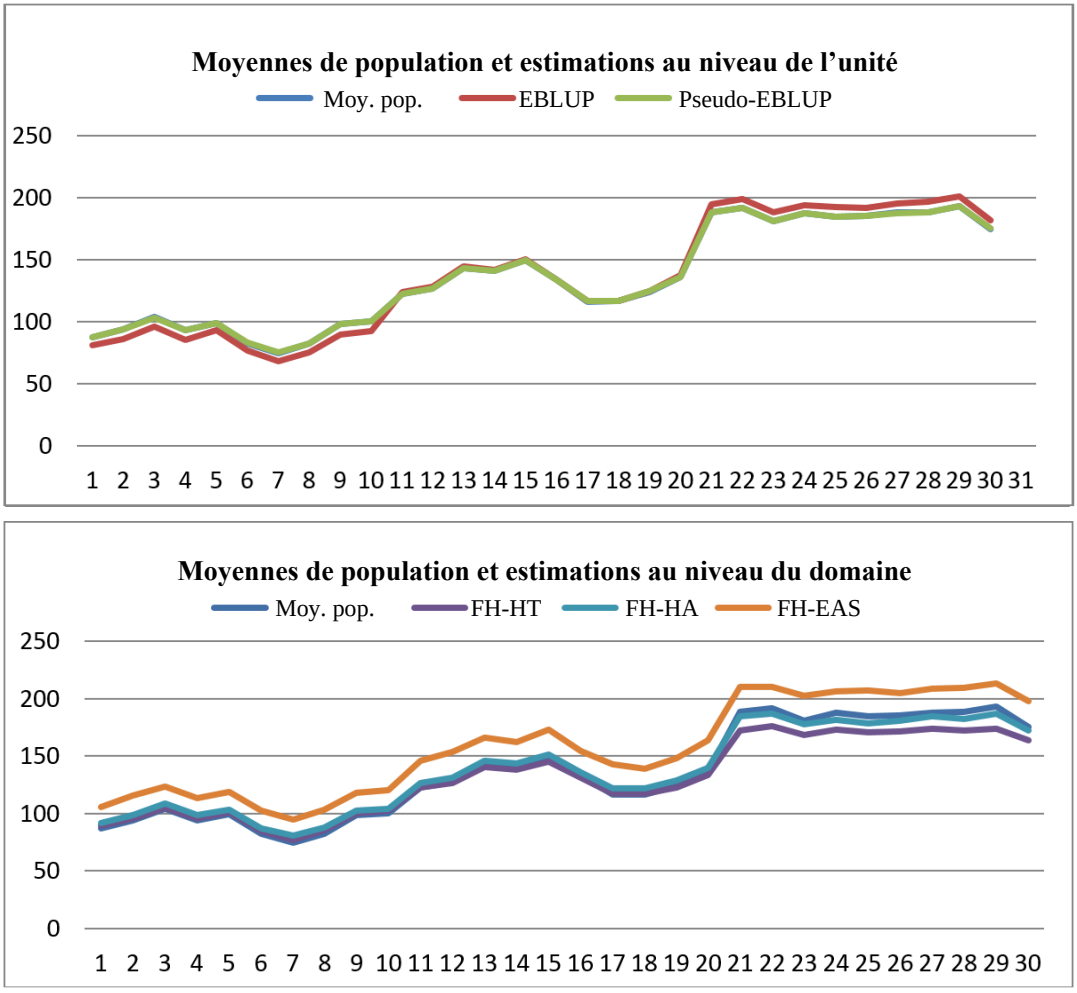

Figure 4.3 Comparaison des moyennes en vertu du scénario II sous une modélisation inexacte pour $n = 10$.

À la figure 4.4, on compare les racines de l'eqm des estimateurs au niveau de l'unité et au niveau du domaine pour les échantillons de taille $n = 10$ et $n = 30$ en vertu d'une modélisation inexacte. On voit bien à la figure 4.4 que l'estimateur pseudo-EBLUP a la plus petite racine de l'eqm lorsque la modélisation est inexacte. L'estimateur EBLUP a une très grande racine de l'eqm lorsque la spécification du modèle est inexacte; de fait, pour les domaines 1 à 10 et 21 à 30, la racine moyenne de l'eqm est 10,01, alors que pour l'estimateur pseudo-EBLUP, la racine correspondante de l'eqm est 7,38 lorsque l'échantillon est de taille $n = 10$. Lorsque l'échantillon est de taille $n = 30$, la racine moyenne de l'eqm est 8,85 pour l'estimateur

EBLUP et seulement 4,38 pour l'estimateur pseudo-EBLUP lorsque le modèle est inexact. En résumé, les résultats montrent que l'estimateur EBLUP donne lieu à des estimations biaisées et à une racine de l' eqm élevée lorsque la modélisation est inexacte.

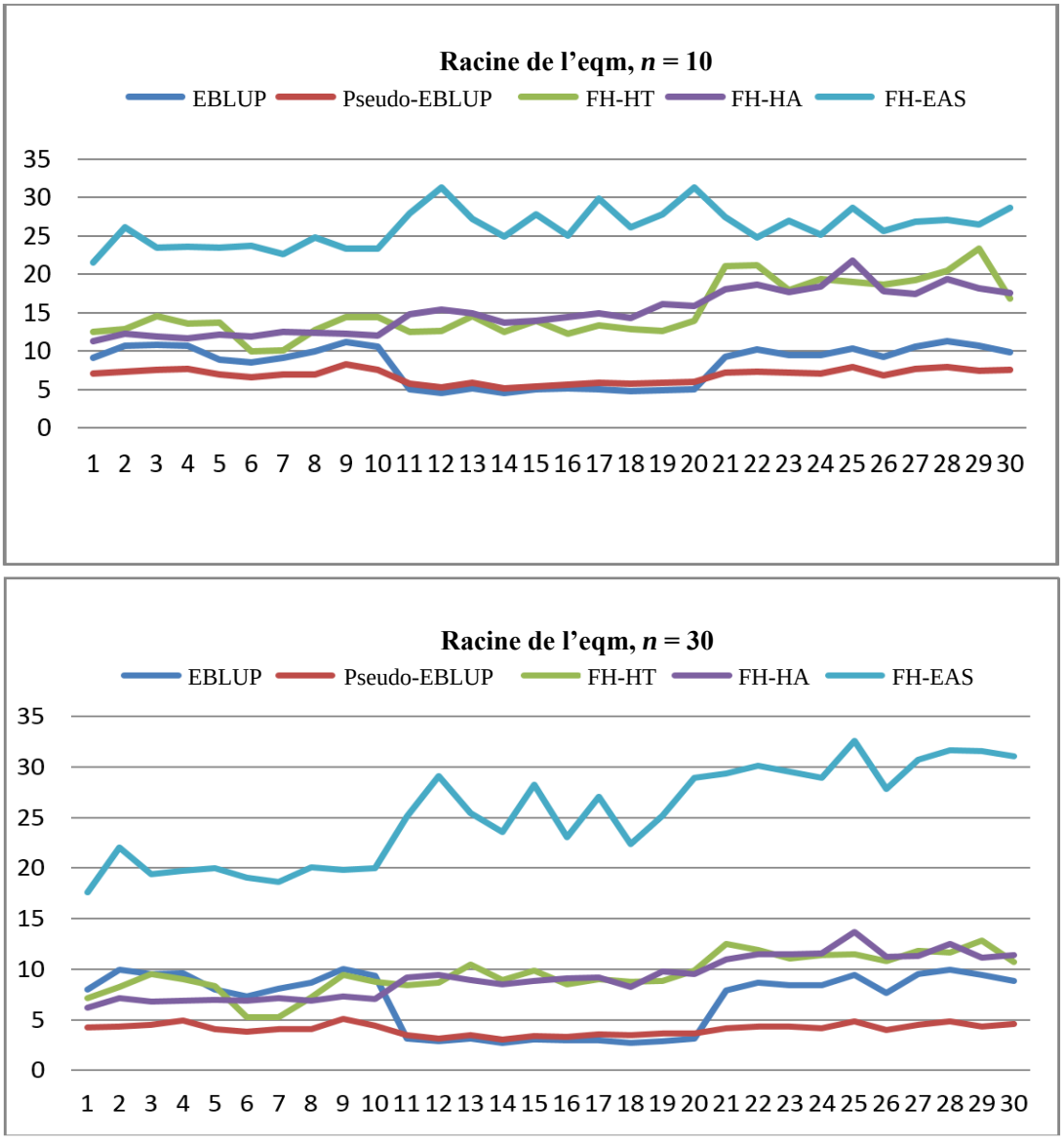

Figure 4.4 Comparaison de la racine de l' eqm en vertu du scénario II pour $n = 10$ et $n = 30$.

4.2.2 Comparaison entre petits domaines

Pour comparer les estimateurs entre domaines, on a examiné le biais relatif absolu (BRA) moyen pour un estimateur spécifié $\hat{\theta}_i$ de la moyenne de population simulée $\bar{Y}_i$ calculé comme suit : $\overline{\text{BRA}} = \left(\sum_{i=1}^{m} \text{BRA}_i\right)\Big/ m$, où

$$\text{BRA}_i = \left| \frac{1}{R} \sum_{r=1}^{R} \frac{\left(\hat{\theta}_i^{(r)} - \bar{Y}_i \right)}{\bar{Y}_i} \right|,$$

et $\hat{\theta}_i^{(r)}$ est l'estimation fondée sur le r^{e} échantillon simulé, $R = 3\,000, m = 30$. Le tableau 4.1 présente le pourcentage du biais relatif absolu moyen $\overline{\text{BRA}}$ des estimateurs au niveau de l'unité et au niveau du domaine pour les 30 domaines en vertu du scénario I. Les résultats sont fondés sur des échantillons sélectionnés de taille 10 et 30 respectivement dans chaque domaine.

Tableau 4.1
Biais relatif absolu moyen $\overline{\text{BRA}}$ en pourcentage en vertu du scénario I

Type	Estimateur	$n = 10$	$n = 30$
Au niveau de l'unité	EBLUP	1,71	0,75
	Pseudo-EBLUP	2,14	0,86
Au niveau du domaine	FH-EAS	17,51	18,64
	FH-HT	6,02	3,12
	FH-HA	4,33	2,59

Dans le cas des modèles au niveau de l'unité, il est clair que si le modèle est exact, l'échantillon devient non informatif en ce qui concerne le modèle au niveau de l'unité (2.2), et les estimateurs EBLUP et pseudo-EBLUP sont sans biais. Le biais relatif absolu moyen $\overline{\text{BRA}}$ pour l'estimateur EBLUP est de 1,71 % lorsque l'échantillon est de taille $n = 10$ et de 0,75 % lorsque l'échantillon est de taille $n = 30$. Pour l'estimateur pseudo-EBLUP, le $\overline{\text{BRA}}$ est de 2,14 % lorsque $n = 10$ et de 0,86 % lorsque $n = 30$. L'estimateur pseudo-EBLUP est associé à un biais légèrement plus élevé que celui de l'estimateur EBLUP. Dans le cas des modèles au niveau du domaine, l'estimateur FH-EAS surestime grandement les moyennes, le $\overline{\text{BRA}}$ atteignant jusqu'à 17,51 % lorsque $n = 10$ et 18,6 % lorsque $n = 30$. Les deux estimateurs au niveau du domaine FH-HT et FH-HA donnent des estimations raisonnables : (i) le $\overline{\text{BRA}}$ pour l'estimateur FH-HT est de 6,02 % lorsque $n = 10$ et de 3,12 % lorsque $n = 30$; (ii) le $\overline{\text{BRA}}$ pour l'estimateur FH-HA est de 4,33 % lorsque $n = 10$ et de 2,59 % lorsque $n = 30$. L'estimateur FH-HA donne de meilleurs résultats que l'estimateur FH-HT. Le biais relatif absolu pour les estimateurs au niveau du domaine est plus grand que celui qui est associé aux estimateurs au niveau de l'unité.

Le tableau 4.2 présente le $\overline{\text{BRA}}$ de divers estimateurs en vertu du scénario II. Il est clair que l'estimateur pseudo-EBLUP est assorti d'un $\overline{\text{BRA}}$ beaucoup plus faible que celui de l'estimateur EBLUP lorsque la modélisation est inexacte. Les $\overline{\text{BRA}}$ de l'estimateur EBLUP en vertu d'une modélisation inexacte sont de 4,31 % $(n = 10)$ et de 4,52 % $(n = 30)$. Pour l'estimateur pseudo-EBLUP, les $\overline{\text{BRA}}$ s'établissent à seulement 0,25 % $(n = 10)$ et 0,12 % $(n = 30)$. Les deux estimateurs FH-HT et FH-HA donnent de très bons résultats. Leurs $\overline{\text{BRA}}$ sont de 3,91 % et 3,48 % respectivement lorsque $n = 10$ et diminuent à 1,51% et à 1,47 % lorsque $n = 30$. L'estimateur FH-EAS donne des résultats médiocres. Les deux estimateurs au niveau du domaine FH-HT et FH-HA donnent de bons résultats, en plus de converger par rapport au plan. Encore une fois, l'estimateur FH-HA est légèrement plus intéressant que l'estimateur FH-HT en termes de $\overline{\text{BRA}}$. Les résultats montrent que le recours à des poids d'enquête dans la modélisation au niveau de l'unité joue un rôle très important lorsque la spécification du modèle au niveau de l'unité est inexacte. L'estimateur

pseudo-EBLUP est sans biais même lorsque la spécification du modèle est inexacte. Il s'agit du meilleur estimateur lorsque le modèle est inexact.

Tableau 4.2
Biais relatif absolu moyen $\overline{\text{BRA}}$ en pourcentage en vertu du scénario II

Type	Estimateur	$n = 10$	$n = 30$
Au niveau de l'unité	EBLUP	4,31	4,52
	Pseudo-EBLUP	0,25	0,12
Au niveau du domaine	FH-EAS	17,11	17,87
	FH-HT	3,91	1,51
	FH-HA	3,48	1,47

On a ensuite comparé la racine de l'EQM relative (REQMR) pour tous les estimateurs. On a notamment calculé la REQMR réelle pour la simulation et la REQMR estimée à partir des estimateurs de l'EQM. La REQMR réelle moyenne pour la simulation est calculée comme suit : $\overline{\text{REQMR}} = \left(\sum_{i=1}^{m} \text{REQMR}_i\right) \Big/ m$, où

$$\text{REQMR}_i = \frac{\sqrt{\text{EQM}_i}}{\bar{Y}_i}, \quad \text{et} \quad \text{EQM}_i = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_i^{(r)} - \bar{Y}_i\right)^2.$$

La racine de l'EQM relative estimée moyenne est calculée comme suit : $\overline{\text{ReqmR}} = \left(\sum_{i=1}^{m} \text{ReqmR}_i\right) \Big/ m$, où

$$\text{ReqmR}_i = \frac{\sqrt{\text{eqm}_i}}{\hat{\theta}_i}, \quad \text{et} \quad \text{eqm}_i = \frac{1}{R}\sum_{r=1}^{R}\text{eqm}_i^{(r)}, \quad \text{et} \quad \hat{\theta}_i = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_i^{(r)}.$$

Le paramètre $\text{eqm}_i^{(r)}$ correspond à l'EQM estimée de $\hat{\theta}_i^{(r)}$ pour le i^{e} domaine. Il est calculé à l'aide des formules énoncées aux sections 2 et 3.

Le tableau 4.3 indique la $\overline{\text{REQMR}}$ et la $\overline{\text{ReqmR}}$ pour les 30 petits domaines. Quand l'échantillon est de taille $n = 10$, la $\overline{\text{REQMR}}$ est de 4,98 % pour l'estimateur EBLUP et de 5,49 % pour l'estimateur pseudo-EBLUP. Comme prévu (You et Rao 2002), l'estimateur pseudo-EBLUP a une REQMR légèrement supérieure à celle de l'estimateur EBLUP. Au niveau de l'unité, les deux estimateurs EBLUP et pseudo-EBLUP ont une REQMR beaucoup plus faible qu'au niveau du domaine. Dans le cas des modèles au niveau du domaine, les estimateurs FH-HT et FH-HA ont un rendement similaire, la REQMR réelle moyenne correspondante s'établissant à 9,72 % et à 9,68 % respectivement lorsque $n = 10$. L'estimateur FH-EAS ne donne pas de bons résultats sous un échantillonnage informatif, la REQMR réelle moyenne s'établissant à 18,89 % lorsque $n = 10$. Même lorsque $n = 30$, la REQMR moyenne pour l'estimateur FH-EAS peut atteindre jusqu'à 18,62 %. Soulignons que la $\overline{\text{ReqmR}}$ est très proche de sa valeur réelle.

En résumé, les résultats présentés au tableau 4.3 montrent que les estimateurs EBLUP et pseudo-EBLUP au niveau de l'unité donnent de meilleurs résultats que les estimateurs FH-HT et FH-HA au niveau du domaine lorsque la modélisation est exacte. Les deux estimateurs FH-HT et FH-HA au niveau du domaine donnent des résultats raisonnablement satisfaisants sous un échantillonnage informatif. Comme prévu, l'estimateur FH-EAS ne donne pas de bons résultats.

Tableau 4.3
REQMR moyenne en pourcentage en vertu du scénario I

Type	Estimateur	$n = 10$		$n = 30$	
		$\overline{\text{REQMR}}$	$\overline{\text{ReqmR}}$	$\overline{\text{REQMR}}$	$\overline{\text{ReqmR}}$
Au niveau de l'unité	EBLUP	4,98	5,09	3,01	3,13
	Pseudo-EBLUP	5,49	5,66	3,58	3,67
Au niveau du domaine	FH-EAS	18,89	17,53	18,62	16,34
	FH-HT	9,72	10,25	6,67	6,69
	FH-HA	9,68	9,71	6,51	6,63

Le tableau 4.4 présente les résultats de la REQMR moyenne en vertu du scénario II. L'estimateur pseudo-EBLUP est le plus robuste et a les plus faibles $\overline{\text{REQMR}}$: les $\overline{\text{REQMR}}$ sont de 5,42 % et de 3,21 % pour $n = 10$ et $n = 30$ respectivement. Les estimateurs au niveau du domaine FH-HT et FH-HA ont un rendement similaire, alors que l'estimateur FH-EAS ne donne pas de bons résultats. Lorsque $n = 10,$ la $\overline{\text{REQMR}}$ est de 11,68 % pour l'estimateur FH-HT et de 11,21 % pour l'estimateur FH-HA. Lorsque $n = 30,$ la $\overline{\text{REQMR}}$ est de 7,24 % pour l'estimateur FH-HT et de 6,79 % pour l'estimateur FH-HA. Comme prévu, l'estimateur FH-EAS a une $\overline{\text{REQMR}}$ élevée sous un échantillonnage informatif. L'estimateur pseudo-EBLUP donne les meilleurs résultats en termes de biais, d'erreur type et de REQMR lorsque la spécification du modèle est inexacte. L'estimateur FH-HA donne des résultats légèrement meilleurs que ceux de l'estimateur FH-HT. La $\overline{\text{ReqmR}}$ estimée est très proche de la $\overline{\text{REQMR}}$ réelle pour tous les estimateurs.

Tableau 4.4
REQMR moyenne en pourcentage en vertu du scénario II

Type	Estimateur	$n = 10$		$n = 30$	
		$\overline{\text{REQMR}}$	$\overline{\text{ReqmR}}$	$\overline{\text{REQMR}}$	$\overline{\text{ReqmR}}$
Au niveau de l'unité	EBLUP	6,78	6,94	5,62	5,81
	Pseudo-EBLUP	5,42	5,45	3,21	3,26
Au niveau du domaine	FH-EAS	19,76	17,43	19,06	16,24
	FH-HT	11,68	11,78	7,24	7,26
	FH-HA	11,21	11,27	6,79	6,91

4.2.3 Comparaison des intervalles de confiance

On a ensuite comparé les intervalles de confiance associés aux estimateurs au niveau de l'unité et au niveau du domaine. L'intervalle de confiance se présente sous la forme $\text{estimateur} \pm z_{\alpha/2}\sqrt{\text{eqm}}$, où $z_{\alpha/2}$ correspond au $100(1-\alpha/2)\%$ centile de la distribution normale centrée réduite. Par exemple, l'intervalle de confiance à 95 % de l'estimateur EBLUP $\hat{\theta}_i^{\text{EBLUP}}$ est obtenu par $\hat{\theta}_i^{\text{EBLUP}} \pm 1{,}96\sqrt{\text{eqm}\left(\hat{\theta}_i^{\text{EBLUP}}\right)}$, où $\text{eqm}\left(\hat{\theta}_i^{\text{EBLUP}}\right)$ est donnée par (2.6). Les intervalles de confiance sont calculés comme ci-dessous. Pour un estimateur donné $\hat{\theta}_i^{(r)},\ r = 1,\ldots,R,\ i = 1,\ldots,m,$ la variable indicatrice $I_i^{(r)}$ est définie comme suit :

$$I_i^{(r)} = \begin{cases} 1 \text{ si } \theta_i \subseteq \left(\hat{\theta}_i^{(r)} - 1{,}96\sqrt{\text{eqm}\left(\hat{\theta}_i^{(r)}\right)}, \hat{\theta}_i^{(r)} + 1{,}96\sqrt{\text{eqm}\left(\hat{\theta}_i^{(r)}\right)}\right) \\ 0 \text{ sinon} \end{cases}.$$

Le taux de couverture des intervalles de confiance correspond à la moyenne des variables $I_i^{(r)}$ pour l'ensemble des $R = 3\,000$ simulations. Les tableaux 4.5 et 4.6 présentent les taux de couverture des intervalles de confiance à 95 % pour les estimateurs au niveau de l'unité et au niveau du domaine en vertu du scénario I. Le coefficient de corrélation ρ entre les probabilités de sélection p_{ij} et y_{ij} est présenté dans la première colonne pour refléter le degré du caractère informatif de l'échantillonnage PPT.

Tableau 4.5
Taux de couverture des intervalles de confiance en vertu du scénario I pour $n = 10$

Coefficient de corrélation (ρ)	EBLUP	Pseudo-EBLUP	FH-EAS	FH-HT	FH-HA
0,95	0,932	0,946	0,618	0,898	0,911
0,88	0,945	0,948	0,649	0,882	0,908
0,75	0,948	0,948	0,705	0,863	0,911
0,51	0,944	0,949	0,825	0,845	0,916
0,28	0,947	0,951	0,901	0,822	0,917
0,12	0,948	0,949	0,924	0,778	0,893
0,02	0,948	0,951	0,925	0,595	0,886
Taux moyen	*0,945*	*0,949*	*0,792*	*0,812*	*0,906*

Discutons d'abord des propriétés de couverture associées aux estimateurs au niveau de l'unité EBLUP et pseudo-EBLUP. Les tableaux montrent que, lorsque le modèle est exact, les taux de couverture des estimateurs EBLUP et pseudo-EBLUP sont assez stables : l'estimateur pseudo-EBLUP a un taux de couverture légèrement meilleur que celui de l'estimateur EBLUP. Lorsque l'échantillon est de taille $n = 10$, le taux de couverture moyen est de 94,5 % pour l'estimateur EBLUP et de 94,9 % pour l'estimateur pseudo-EBLUP. Lorsque l'échantillon est de taille $n = 30$, il est de 93,4 % pour l'estimateur EBLUP et de 94,8 % pour l'estimateur pseudo-EBLUP. Lorsque la taille de l'échantillon passe de $n = 10$ à $n = 30$, les taux de couverture de l'estimateur EBLUP se détériorent légèrement plus que ceux de l'estimateur pseudo-EBLUP. L'estimateur pseudo-EBLUP n'est pas aussi influencé par l'importance du caractère informatif découlant de l'échantillonnage PPT. Les taux de couverture relativement stables de l'estimateur EBLUP montrent que l'échantillon n'est pas informatif en ce qui concerne le modèle au niveau de l'unité exact. Toutefois, lorsque $n = 30$, l'estimateur EBLUP est associé à un taux de couverture légèrement inférieur.

Tableau 4.6
Taux de couverture des intervalles de confiance en vertu du scénario I pour $n = 30$

Coefficient de corrélation (ρ)	EBLUP	Pseudo-EBLUP	FH-EAS	FH-HT	FH-HA
0,95	0,905	0,946	0,265	0,932	0,926
0,88	0,938	0,948	0,286	0,915	0,921
0,75	0,941	0,949	0,377	0,911	0,924
0,51	0,940	0,951	0,625	0,895	0,931
0,28	0,941	0,950	0,806	0,874	0,929
0,12	0,939	0,945	0,923	0,866	0,922
0,02	0,937	0,948	0,937	0,772	0,917
Taux moyen	*0,934*	*0,948*	*0,603*	*0,881*	*0,924*

Passons maintenant aux taux de couverture associés aux estimateurs au niveau du domaine. Comme prévu, les estimateurs FH-EAS ont des taux de couverture faibles lorsque l'échantillonnage est informatif; le taux de couverture augmente à mesure que le plan d'échantillonnage devient non informatif. L'estimateur FH-HA a un meilleur taux de couverture que l'estimateur FH-HT. Le taux de couverture de l'estimateur FH-HT diminue à mesure que le plan d'échantillonnage devient non informatif. Par exemple, lorsque l'échantillon est de taille $n = 10,$ le taux de couverture pour l'estimateur FH-HT n'est que de 59,5 % lorsque l'échantillonnage est non informatif, comparativement à 88,6 % pour l'estimateur FH-HA. À mesure que la taille de l'échantillon augmente, le taux de couverture pour les estimateurs FH-HT et FH-HA s'améliore. Le taux de couverture moyen pour l'estimateur FH-HA est de 90,6 % lorsque $n = 10$ et de 92,4 % lorsque $n = 30.$ L'estimateur FH-HT a un taux de couverture inférieur à celui de l'estimateur FH-HA. Le taux de couverture moyen n'est que de 81,2 % pour l'estimateur FH-HT lorsque $n = 10.$ Le taux de couverture pour l'estimateur FH-EAS est très faible, soit 61,8 % sous un échantillonnage informatif lorsque $n = 10$ et 26,5 % lorsque $n = 30.$ À mesure que la taille de l'échantillon augmente, le taux de couverture diminue pour l'estimateur FH-EAS sous un échantillonnage informatif. Comme prévu, le taux de couverture augmente graduellement pour l'estimateur FH-EAS à mesure que l'échantillonnage devient non informatif. De tous les estimateurs, que l'échantillon soit de taille $n = 10$ ou $n = 30,$ l'estimateur pseudo-EBLUP a le meilleur taux de couverture, et l'estimateur FH-HA, le deuxième meilleur taux de couverture.

Les tableaux 4.7 et 4.8 présentent les taux de couverture en vertu du scénario II. Les résultats montrent que l'estimateur EBLUP a un faible taux de couverture sous un échantillonnage informatif, alors que l'estimateur pseudo-EBLUP a des taux de couverture très stables et élevés (tous autour de 95 % et plus) sous un échantillonnage informatif ou non informatif. Par exemple, lorsque $n = 10,$ l'estimateur EBLUP a un taux de couverture de 84,6 % sous un échantillonnage informatif (coefficient de corrélation de 0,95), qui diminue à 62,9 % lorsque la taille de l'échantillon augmente à $n = 30.$ Sous une modélisation inexacte, le taux de couverture moyen de l'estimateur EBLUP est de 90,4 % pour $n = 10$ et de 79,6 % pour $n = 30.$ Les résultats montrent que l'estimateur EBLUP est sensible à la modélisation lorsque l'échantillonnage est informatif, ce qui s'explique par le fait que cet estimateur repose entièrement sur le modèle et ne dépend pas du tout du plan d'échantillonnage.

Tableau 4.7
Taux de couverture des intervalles de confiance en vertu du scénario II pour $n = 10$

Coefficient de corrélation (ρ)	EBLUP	Pseudo-EBLUP	FH-EAS	FH-HT	FH-HA
0,95	0,846	0,965	0,701	0,865	0,896
0,88	0,855	0,964	0,729	0,887	0,893
0,75	0,881	0,962	0,787	0,873	0,898
0,51	0,921	0,961	0,872	0,848	0,898
0,28	0,936	0,961	0,912	0,843	0,887
0,12	0,945	0,955	0,917	0,765	0,867
0,02	0,943	0,951	0,913	0,592	0,838
Taux moyen	*0,904*	*0,959*	*0,833*	*0,811*	*0,883*

Tableau 4.8
Taux de couverture des intervalles de confiance en vertu du scénario II pour $n = 30$

Coefficient de corrélation (ρ)	EBLUP	Pseudo-EBLUP	FH-EAS	FH-HT	FH-HA
0,95	0,629	0,969	0,239	0,913	0,923
0,88	0,638	0,965	0,275	0,895	0,919
0,75	0,708	0,964	0,406	0,908	0,923
0,51	0,829	0,963	0,701	0,923	0,926
0,28	0,902	0,964	0,854	0,911	0,921
0,12	0,931	0,958	0,921	0,884	0,912
0,02	0,937	0,953	0,918	0,778	0,894
Taux moyen	*0,796*	*0,962*	*0,616*	*0,887*	*0,918*

Des trois estimateurs au niveau du domaine, c'est l'estimateur FH-HA qui donne les meilleurs résultats. Le taux de couverture pour l'estimateur FH-HA est très stable; le taux de couverture moyen est de 88,3 % lorsque $n = 10$ et de 91,8 % lorsque $n = 30$. L'estimateur FH-HT a le taux de couverture le plus faible lorsque l'échantillonnage est très peu informatif, particulièrement lorsque l'échantillon est de taille $n = 10$. Le taux de couverture moyen pour l'estimateur FH-HT n'est que de 81,1 % lorsque $n = 10$ et de 88,7 % lorsque $n = 30$. Les résultats montrent que l'estimateur FH-HA est supérieur à l'estimateur FH-HT. Quant à l'estimateur FH-EAS, il donne des résultats médiocres lorsque l'échantillonnage est informatif, particulièrement lorsque l'échantillon est de taille $n = 30$. Toutefois, l'estimateur FH-EAS donne des résultats relativement bons lorsque l'échantillonnage devient non informatif. Le taux de couverture moyen pour l'estimateur FH-EAS est de 83,3 % lorsque $n = 10$, mais seulement de 61,6 % lorsque l'échantillon est de taille $n = 30$.

Il est clair que l'estimateur pseudo-EBLUP a un taux de couverture très élevé et stable sous une modélisation inexacte. L'estimateur FH-HA a aussi un taux de couverture très stable, mais légèrement inférieur. Les taux de couverture des estimateurs EBLUP et FH-EAS diminuent à mesure que la taille de l'échantillon augmente, particulièrement lorsque l'échantillonnage est informatif.

5 Application aux données réelles

Dans la section qui suit, on compare les estimations au niveau de l'unité et au niveau du domaine au moyen d'une analyse de données réelles. L'ensemble de données étudié est celui présenté par Battese et coll. (1988) dans le cadre d'une étude estimant le nombre moyen d'hectares consacrés à la culture du maïs et du soja par segment dans douze comtés du Centre-Nord de l'Iowa. De ces douze comtés, trois ne comportaient qu'un seul segment échantillonné. Aux fins de la présente étude, les données de ces trois comtés ont été regroupées en un seul, ce qui donne un ensemble de données contenant 10 comtés dont la taille d'échantillon n_i varie entre 2 et 5 dans chaque comté. Le nombre total de segments N_i (taille de la population) dans chaque comté allait de 402 à 1 505. Suivant la méthode de You et Rao (2002), on a présumé un échantillonnage aléatoire simple (EAS) dans chaque comté, et le poids d'enquête de base a été calculé comme suit : $w_{ij} = N_i / n_i$. Pour la modélisation au niveau de l'unité, y_{ij} correspond au nombre d'hectares

de maïs (ou de soja) dans le j^{e} segment du i^{e} comté, les variables auxiliaires étant le nombre de pixels classés comme étant du maïs ou du soja selon Battese et coll. (1988). On a appliqué le modèle au niveau de l'unité à l'ensemble de données modifié et calculé les estimations EBLUP et pseudo-EBLUP. Pour la modélisation au niveau du domaine, on a d'abord calculé les estimations directes sur échantillon $\hat{\theta}_i^{\mathrm{EAS}}$ fondées sur l'EAS. On a ensuite appliqué le modèle de Fay-Herriot aux estimations directes au niveau du domaine et calculé les estimations FH-EAS au niveau du domaine. La figure 5.1 illustre la comparaison entre les estimations directes au niveau du domaine et les estimations fondées sur le modèle au niveau de l'unité et au niveau du domaine. En termes d'estimation ponctuelle, les estimations EBLUP et pseudo-EBLUP sont presque identiques, comme dans You et Rao (2002). Ce résultat s'explique par le fait que le modèle au niveau de l'unité est exact pour ces données (Battese et coll. 1988). Les estimations FH-EAS au niveau du domaine fondées sur le modèle et les estimations directes au niveau du domaine concordent assez bien dans cet exemple.

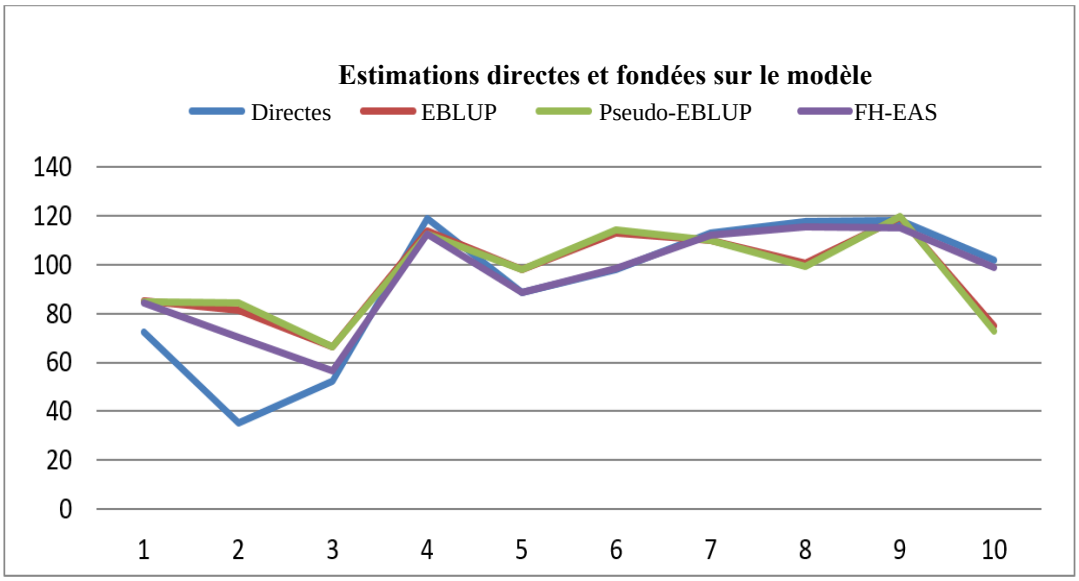

Figure 5.1 Comparaison des estimations directes et des estimations fondées sur le modèle.

La figure 5.2 illustre la comparaison entre les erreurs types des estimateurs directs et fondés sur le modèle. Les erreurs types des estimateurs fondés sur le modèle correspondent à la racine carrée de l'EQM estimée. Les deux estimateurs au niveau de l'unité, EBLUP et pseudo-EBLUP, ont des erreurs types faibles et stables. Comme prévu, l'estimateur pseudo-EBLUP est assorti d'erreurs types légèrement plus grandes que celles de l'estimateur EBLUP. Il est clair que les erreurs types des estimateurs directs et FH-EAS sont très variables et très instables. Cet exemple illustre l'efficacité des estimateurs au niveau de l'unité EBLUP et pseudo-EBLUP.

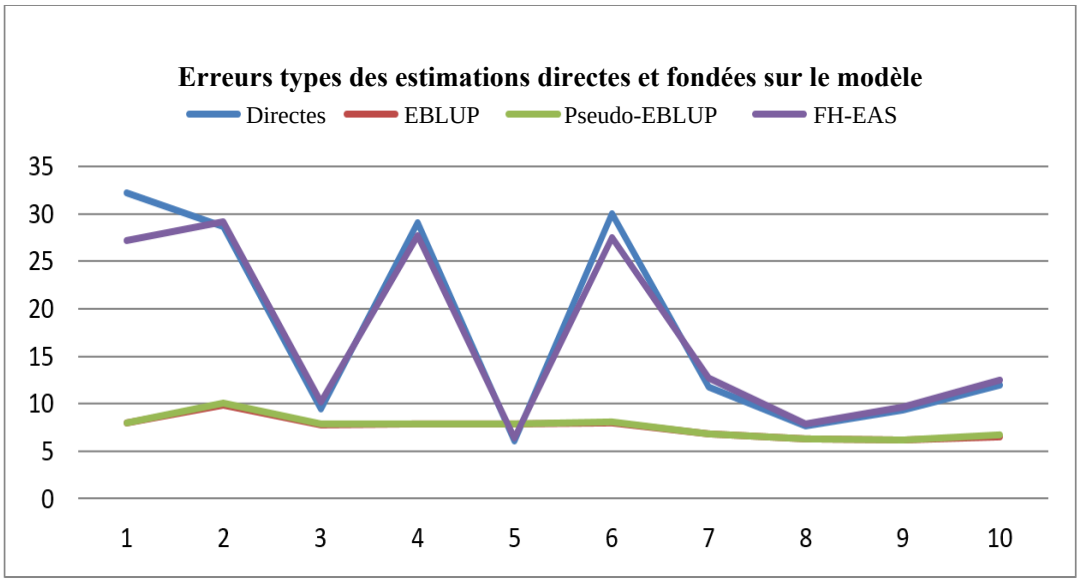

Figure 5.2 Comparaison des erreurs types des estimations directes et des estimations fondées sur le modèle.

6 Conclusions

Dans le présent article, les auteurs ont comparé l'efficacité des estimateurs fondés sur un modèle de régression à erreur emboîtée au niveau de l'unité et sur le modèle de Fay-Herriot au niveau du domaine à l'aide d'une étude par simulations fondée sur le plan. Ils ont comparé les estimations ponctuelles et les taux de couverture des intervalles de confiance des estimateurs au niveau de l'unité et au niveau du domaine. Dans l'ensemble, l'estimateur pseudo-EBLUP au niveau de l'unité est le plus efficace en termes de biais et de taux de couverture, que l'échantillonnage soit ou non informatif. L'estimateur EBLUP est efficace sous une modélisation exacte, puisque l'échantillonnage est non informatif en vertu du modèle au niveau de l'unité exact décrit en (2.2). L'estimateur pseudo-EBLUP est également assez robuste à une spécification inexacte du modèle. En pratique, les auteurs recommandent de construire les estimateurs pseudo-EBLUP à l'aide des poids d'enquête et des observations au niveau de l'unité dont il est question à la section 2.2. Dans le cas des modèles au niveau du domaine, l'estimateur FH-HA donne de meilleurs résultats que l'estimateur FH-HT; l'estimateur FH-EAS donne des résultats médiocres. On recommande donc de construire les estimateurs HA pondérés et d'appliquer ensuite le modèle de Fay-Herriot pour obtenir les estimateurs fondés sur le modèle correspondants si on utilise des estimateurs sur petits domaines au niveau du domaine.

Remerciements

Les auteurs remercient le rédacteur en chef adjoint et deux évaluateurs pour leurs suggestions et commentaires, qui ont permis d'améliorer considérablement la présentation des résultats. Ils tiennent à remercier plus particulièrement l'un des évaluateurs pour ses commentaires fort minutieux et constructifs.

Bibliographie

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.

Cressie, N. (1992). Estimation du maximum de vraisemblance avec contrainte (MVC) dans le lissage des taux de sous-dénombrement du recensement selon l'approche empirique de Baye. *Techniques d'enquête*, 18, 1, 83-103.

Datta, G.S., et Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistics Sinica*, 10, 613-627.

Fay, R.E., et Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 268-277.

Pfeffermann, D., et Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102, 480, 1427-1439.

Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

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

Rivest, L.-P., et Vandal, N. (2003). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, (Éd., J.N.K. Rao).

Torabi, M., et Rao, J.N.K. (2010). The mean squared error estimators of small area means using survey weights. *The Canadian Journal of Statistics*, 38, 598-608.

Verret, F., Rao, J.N.K. et Hidiroglou, M.A. (2015). Estimation sur petits domaines fondée sur un modèle sous échantillonnage informatif. *Techniques d'enquête*, 41, 2, 353-368.

Wang, J., et Fuller, W.A. (2003). The mean squared error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y. (2010). *Small Area Estimation under the Fay-Herriot Model Using Different Model Variance Estimation Methods and Different Input Sampling Variances*. Document de travail de la Direction de la méthodologie, SRID-2010-003E, Statistique Canada, Ottawa, Canada.

You, Y., et Rao, J.N.K. (2002). A pseudo empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, 30, 431-439.

You, Y., Rao, J.N.K. et Kovacevic, M. (2003). 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. Recueil : Symposium 2003, *Défis reliés à la réalisation d'enquêtes pour la prochaine décennie*, Statistique Canada.