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

Techniques d’enquête

Mesure de l’incertitude associée aux estimateurs pour petits domaines basés sur un modèle

par J.N.K. Rao, Susana Rubin-Bleuer et Victor M. Estevao

Date de diffusion : le 20 décembre 2018

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 suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

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)

Publication autorisée par le ministre responsable de Statistique Canada

© Sa Majesté la Reine du chef du Canada, représentée par le ministre de l'Industrie 2018

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada.

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2018
Vol. 44, nº 2, pp. 163-180
Statistique Canada, nº 12-001-X au catalogue

Mesure de l'incertitude associée aux estimateurs pour petits domaines basés sur un modèle

J.N.K. Rao, Susana Rubin-Bleuer et Victor M. Estevao[1]

Résumé

Les domaines (ou sous-populations) pour lesquels les échantillons sont de petite taille sont appelés petits domaines. Les estimateurs directs classiques ne sont pas suffisamment précis pour ces petits domaines, en raison de la petite taille des échantillons. Or, la demande de statistiques fiables pour les petits domaines a augmenté considérablement. On utilise à l'heure actuelle des estimateurs indirects des moyennes ou des totaux de petits domaines basés sur un modèle pour résoudre les difficultés que pose l'estimation directe. Ces estimateurs reposent sur des modèles de liaison qui empruntent de l'information aux divers domaines pour accroître l'efficacité. En particulier, beaucoup d'attention a été accordée dans la littérature aux meilleurs estimateurs empiriques ou estimateurs EB (pour *Empirical Best*) sous des modèles de régression linéaires au niveau du domaine et au niveau de l'unité contenant des effets aléatoires de petit domaine. L'erreur quadratique moyenne (EQM) des estimateurs EB sous le modèle sert fréquemment à mesurer la variabilité des estimateurs. Les estimateurs par linéarisation, ainsi que les estimateurs jackknife et bootstrap de l'EQM sous le modèle sont d'usage très répandu. Toutefois, les organismes statistiques nationaux s'intéressent souvent à l'estimation de l'EQM des estimateurs EB sous le plan de sondage, pour cadrer avec les estimateurs classiques de l'EQM sous le plan associés aux estimateurs directs pour les grands domaines dont les tailles d'échantillon sont adéquates. Les estimateurs de l'EQM sous le plan des estimateurs EB peuvent être obtenus pour les modèles au niveau du domaine, mais ils ont tendance à être instables quand la taille de l'échantillon du domaine est petite. Des estimateurs composites de l'EQM, obtenus en prenant une somme pondérée de l'estimateur de l'EQM sous le plan et de l'estimateur de l'EQM sous le modèle, sont proposés dans le présent article. Les propriétés des estimateurs de l'EQM sous le modèle au niveau du domaine sont étudiées en examinant le biais, la racine carrée de l'erreur quadratique moyenne relative et le taux de couverture des intervalles de confiance sous le plan de sondage. Le cas d'un modèle au niveau de l'unité est également examiné sous échantillonnage aléatoire simple dans chaque domaine. Les résultats d'une étude en simulation montrent que les estimateurs composites proposés de l'EQM offrent un bon compromis pour l'estimation de l'EQM sous le plan.

Mots-clés : Modèles au niveau du domaine et au niveau de l'unité; estimateurs composites de l'erreur quadratique moyenne sous le plan; meilleur prédicteur linéaire sans biais empirique; estimation de l'erreur quadratique moyenne sous le plan.

1 Introduction

Les données d'enquête sont souvent utilisées pour produire des estimations de totaux ou de moyennes de domaine (sous-population). Les estimateurs directs classiques pour les domaines, y compris les estimateurs par calage utilisant les totaux de population connus de variables auxiliaires, sont conçus pour fournir des estimateurs fiables pour les domaines dont la taille d'échantillon est grande. En revanche, la précision des estimateurs directs n'est pas suffisante pour les domaines dont la taille d'échantillon est petite (petits domaines). Toutefois, la demande de statistiques fiables pour les petits domaines a beaucoup augmenté ces dernières années. Il faut donc recourir à des estimateurs indirects qui empruntent de l'information à des domaines apparentés grâce à de l'information auxiliaire connue, comme les données des recensements et des dossiers administratifs, pour accroître l'efficacité. Les estimateurs indirects basés sur des modèles de liaison explicites sont d'usage très répandu, notamment, les meilleurs estimateurs empiriques (estimateurs EB pour *Empirical Best*) basés sur des modèles de régression linéaires au niveau

1. J.N.K. Rao, professeur émérite de recherche, Université Carleton, Ottawa, (Ontario). Courriel : jrao34@rogers.com; Susana Rubin-Bleuer, professeure adjointe de recherche, Université Carleton, Ottawa, (Ontario); Victor M. Estevao, méthodologiste principal, Statistique Canada, Immeuble R.-H.-Coats, 25e étage, 100 promenade Tunney's Pasture, Ottawa (Ontario), K1A 0T6. Courriel : victor.estevao@canada.ca.

du domaine ou au niveau de l'unité avec effets aléatoires de domaine. Une description détaillée de l'estimation EB sous ces modèles est donnée dans Rao et Molina (2015), chapitres 6 et 7. À la section 2, nous présentons les estimateurs EB des moyennes de petit domaine sous des modèles de base au niveau du domaine et au niveau de l'unité.

Après des évaluations externes minutieuses, les organismes statistiques nationaux considèrent souvent les estimateurs basés sur un modèle de type EB comme étant appropriés pour produire les statistiques officielles. Beaumont et Bocci (2016) ont comparé les estimations EB et les estimations directes du taux de chômage pour des petits domaines calculées d'après Enquête sur la population active (EPA) du Canada à des estimations « de référence » obtenues d'après l'Enquête nationale auprès des ménages, dont la portée est beaucoup plus grande (comparable à celle du questionnaire détaillé du recensement), et constaté que l'erreur relative des estimations EB est nettement plus petite que celle des estimations directes correspondantes. Les auteurs ont utilisé un modèle de régression linéaire de base au niveau du domaine avec effets aléatoires de domaine pour produire les estimations EB. Des évaluations externes ont été utilisées pour la première fois dans l'article fondamental de Fay et Herriot (1979) sous un modèle de base au niveau du domaine pour produire des estimations du revenu moyen pour de petites localités aux États-Unis.

L'erreur quadratique moyenne (EQM) sous le modèle des estimateurs EB est souvent utilisée pour mesurer la variabilité des estimateurs. En particulier, les estimateurs par linéarisation, de même que les estimateurs jackknife et bootstrap de l'EQM sous le modèle sont d'usage très répandu. À la section 3, nous décrivons brièvement l'estimation de l'EQM basée sur un modèle, y compris les estimateurs s'appuyant sur un cadre inconditionnel ou conditionnel.

La littérature sur l'estimation de l'EQM sous le modèle est très abondante, mais les organismes statistiques nationaux s'intéressent souvent à l'estimation de l'EQM des estimateurs EB sous le plan de sondage, pour cadrer avec les estimateurs classiques de l'EQM sous le plan des estimateurs directs pour les grands domaines dont la taille d'échantillon est adéquate (Pfeffermann et Gilboa, 2017). Il est possible d'obtenir les estimateurs de l'EQM sous le plan des estimateurs EB pour le modèle de base au niveau du domaine, mais ils ont tendance à être instables quand la taille de l'échantillon de domaine est petite. Pour résoudre ce problème, nous proposons à la section 4 des estimateurs composites de l'EQM calculés en prenant une somme pondérée de l'estimateur de l'EQM sous le plan et de l'estimateur de l'EQM sous le modèle. Nous étudions également le cas des modèles au niveau de l'unité sous échantillonnage aléatoire simple dans les domaines. À la section 5, nous présentons les résultats d'études en simulation sur la performance des estimateurs composites proposés de l'EQM, évaluée au moyen du biais relatif absolu (BRA), de la racine carrée de l'erreur quadratique moyenne relative (REQMR) et de la couverture des intervalles de confiance. Tant les modèles au niveau du domaine que ceux au niveau de l'unité sont pris en considération dans l'étude en simulation. Enfin, nous présentons certaines conclusions à la section 6.

2 Estimateurs EB

À la présente section, nous présentons les estimateurs EB pour les moyennes ou totaux de petit domaine, notés θ_i, pour m domaines dont les échantillons sont de petite taille. Pour les modèles au niveau du

domaine, nous supposons que les estimateurs directs $\hat{\theta}_i$ et les covariables au niveau du domaine associées $\mathbf{z}_i$ sont disponibles pour les m domaines, où $\mathbf{z}_i$ est un vecteur de dimension $p \times 1$. Dans le cas des modèles au niveau de l'unité, nous supposons que les données au niveau de l'unité $\{(y_{ij}, \mathbf{x}_{ij}),\ j = 1, \ldots, n_i;\ i = 1, \ldots, m\}$ sont disponibles pour les domaines échantillonnés, où n_i est la taille d'échantillon dans le domaine i et $\mathbf{x}_{ij}$ est un vecteur de dimension $p \times 1$ de covariables qui peuvent inclure des covariables au niveau du domaine. Nous supposons que les moyennes de population de domaine $\bar{\mathbf{X}}_i$ sont connues.

2.1 Modèle de base au niveau du domaine

Nous supposons que l'estimateur direct $\hat{\theta}_i$ est sans biais sous le plan (soit exactement ou approximativement pour une grande taille d'échantillon global n). Par exemple, les estimateurs calés sur des moyennes globales connues de variables auxiliaires sont approximativement sans biais. Nous exprimons cette hypothèse sous forme d'un modèle d'échantillonnage $\hat{\theta}_i = \theta_i + e_i$, où l'erreur d'échantillonnage e_i est de moyenne nulle et de variance ψ_i. Nous supposons en outre que la variance d'échantillonnage ψ_i est connue et non aléatoire. En pratique, les estimateurs des variances d'échantillonnage sont lissés et l'estimateur lissé résultant est considéré comme une approximation de ψ_i. Beaumont et Bocci (2016) proposent une méthode de lissage des variances d'échantillonnage dans le contexte de l'EPA du Canada. Le modèle reliant les domaines repose sur l'hypothèse que les θ_i sont aléatoires et obéissent au modèle de liaison « correspondant » $\theta_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i$, où l'effet aléatoire de domaine v_i est de moyenne nulle et de variance σ_v^2 et est indépendant de l'erreur d'échantillonnage e_i. Nous supposons en outre que v_i et e_i suivent des lois normales.

La combinaison du modèle d'échantillonnage avec le modèle de liaison aboutit au modèle de base au niveau du domaine

$$\hat{\theta}_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i + e_i, \quad v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2), \quad e_i \overset{\text{id}}{\sim} N(0, \psi_i), \quad i = 1, \ldots, m. \tag{2.1}$$

Les principaux avantages du modèle (2.1) par rapport aux estimateurs directs découlent du fait qu'il tient compte du plan d'échantillonnage grâce au modèle d'échantillonnage et qu'il ne requiert que des covariables au niveau du domaine, qui s'obtiennent plus facilement que des covariables au niveau de l'unité.

Pour des paramètres du modèle connus $(\boldsymbol{\beta}, \sigma_v^2)$, le « meilleur » (noté *B* pour *Best*) estimateur de θ_i est donné par

$$\tilde{\theta}_i^B = \tilde{E}\left(\theta_i \mid \hat{\theta}_i, \boldsymbol{\beta}, \sigma_v^2\right) = \gamma_i \hat{\theta}_i + (1 - \gamma_i)\, \mathbf{z}_i'\boldsymbol{\beta}, \tag{2.2}$$

où $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \psi_i)$. Le meilleur estimateur (2.2) est sans biais pour θ_i en ce sens que $E\left(\tilde{\theta}_i^B - \theta_i\right) = 0$, où l'espérance est prise par rapport au modèle présumé (2.1), qui est l'espérance conjointe par rapport au modèle et au plan (Rubin-Bleuer et Schiopu-Kratina, 2005). Il découle de (2.2) que plus de poids est attribué à l'estimateur direct $\hat{\theta}_i$ si la variance sous le modèle σ_v^2 est grande comparativement à la variance d'échantillonnage ψ_i, et que plus de poids est accordé à l'estimateur synthétique $\mathbf{z}_i'\boldsymbol{\beta}$ si la variance d'échantillonnage est grande.

L'erreur quadratique moyenne (EQM) du meilleur estimateur sous le modèle (2.1) est donnée par

$$\text{EQM}\left(\tilde{\theta}_i^B\right) = E\left(\tilde{\theta}_i^B - \theta_i\right)^2 = \gamma_i \psi_i, \tag{2.3}$$

où le terme $\gamma_i \psi_i$ est souvent noté $g_{1i}\left(\sigma_v^2\right)$. Il découle de (2.3) que l'estimateur optimal donne lieu à une réduction importante de l'EQM par rapport à l'estimateur direct si γ_i est petit ou que la variance sous le modèle est relativement faible comparativement à la variance totale $\sigma_v^2 + \psi_i$. Ce résultat fournit une justification convaincante de l'utilisation de l'approche basée sur un modèle pour produire des estimations sur petits domaines.

En pratique, les paramètres du modèle sont inconnus et nous remplaçons les paramètres figurant dans (2.2) par les estimateurs du maximum de vraisemblance restreint (REML) $\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}_v^2\right)$ pour obtenir le meilleur estimateur empirique (estimateur EB) :

$$\hat{\theta}_i^{\text{EB}} = \hat{\gamma}_i \hat{\theta}_i + \left(1 - \hat{\gamma}_i\right) \mathbf{z}_i' \hat{\boldsymbol{\beta}}. \tag{2.4}$$

Rao et Molina (2015), chapitre 6, décrivent en détail l'estimation du REML des paramètres du modèle.

2.2 Modèle de base au niveau de l'unité

Considérons maintenant un modèle de base au niveau de l'unité qui utilise des données d'échantillon au niveau de l'unité $\left\{\left(y_{ij}, \mathbf{x}_{ij}\right),\ j = 1, \ldots, n_i;\ i = 1, \ldots, m\right\}$, où n_i est la taille de l'échantillon dans le domaine i. Nous supposons que les moyennes de population de domaine $\overline{\mathbf{X}}_i$ sont connues. Nous émettons en outre l'hypothèse d'un modèle de régression linéaire à erreurs emboîtées de base au niveau de l'unité pour la population et supposons que le même modèle tient pour l'échantillon (Battese, Harter et Fuller, 1988). Le modèle d'échantillonnage est donné par

$$y_{ij} = \mathbf{x}_{ij}' \boldsymbol{\beta} + v_i + e_{ij},\quad j = 1, \ldots, n_i;\ \ i = 1, \ldots, m, \tag{2.5}$$

où les effets aléatoires de domaine $v_i \overset{\text{iid}}{\sim} N\left(0, \sigma_v^2\right)$ sont présumés être indépendants des erreurs au niveau de l'unité $e_{ij} \overset{\text{iid}}{\sim} N\left(0, \sigma_e^2\right)$. Les modèles au niveau de l'unité peuvent mener à d'importants gains d'efficacité par rapport aux modèles au niveau du domaine, parce que les paramètres du modèle peuvent être estimés avec plus de précision en utilisant toutes les observations dans l'échantillon global, contrairement aux modèles au niveau du domaine.

Pour les paramètres connus $\left(\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2\right)$, le « meilleur » estimateur de la moyenne de domaine $\overline{Y}_i$ est donné par

$$\hat{\overline{Y}}_i^B = E\left[\overline{Y}_i \,\middle|\, \left(y_{ij},\ j = 1, \ldots, n_i;\ \mathbf{x}_{ij},\ j = 1, \ldots, N_i\right), \boldsymbol{\beta}, \sigma_v^2, \sigma_e^2\right] = \overline{\mathbf{X}}_i' \boldsymbol{\beta} + a_i\left(\overline{y}_i - \overline{\mathbf{x}}_i' \boldsymbol{\beta}\right), \tag{2.6}$$

où $\overline{y}_i$ et $\overline{\mathbf{x}}_i$ sont les moyennes d'échantillon, $a_i = \left(1 - f_i\right)\gamma_i + f_i$ avec la fraction d'échantillonnage $f_i = n_i / N_i$ et $\gamma_i = \sigma_v^2 / \left(\sigma_v^2 + \sigma_e^2 / n_i\right)$, et N_i est le nombre d'unités de la population dans le domaine i (Rao et Molina, 2015, chapitre 7). Si la taille de la population du domaine N_i est grande et que $f_i \approx 0$, alors (2.6) se réduit à une combinaison pondérée de l'estimateur « par la régression sur l'échantillon » $\overline{y}_i + \left(\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_i\right)' \boldsymbol{\beta}$ et de l'estimateur par la régression synthétique $\overline{\mathbf{X}}_i' \boldsymbol{\beta}$ avec les poids γ_i et $1 - \gamma_i$,

respectivement. Nous désignons cette approximation de $\hat{\bar{Y}}_i^B$ par $\hat{\mu}_i^B$. À mesure qu'augmente la taille d'échantillon de domaine n_i, l'estimateur optimal donne plus de poids à l'estimateur par la régression sur l'échantillon. En pratique, nous remplaçons les paramètres du modèle par les estimateurs du REML $\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}_v^2, \hat{\sigma}_e^2\right)$ pour obtenir l'estimateur EB $\hat{\bar{Y}}_i^{\text{EB}}$ ou $\hat{\mu}_i^{\text{EB}}$.

L'estimateur EB sous le modèle au niveau de l'unité (2.5) ne tient pas compte des poids de sondage w_{ij}, contrairement au modèle au niveau du domaine. Par conséquent, l'estimateur EB n'est pas convergent sous le plan quand la taille d'échantillon de domaine augmente, à moins que les poids ne soient tous égaux à l'intérieur du domaine.

L'EQM de $\hat{\mu}_i^B$ est égale à $g_{1i}(\sigma_v^2, \sigma_e^2) = \gamma_i(\sigma_e^2/n_i)$, tandis que l'EQM de l'estimateur par la régression sur l'échantillon est égal à σ_e^2/n_i. Il s'ensuit maintenant que l'estimateur optimal donne lieu à une réduction importante de l'EQM par rapport à l'estimation par la régression sur l'échantillon si γ_i est petit ou que la variance du modèle σ_v^2 est petite comparativement à la variance totale $\sigma_v^2 + \sigma_e^2/n_i$.

3 Estimateurs de l'EQM basée sur le modèle

À la présente section, nous nous concentrons sur l'EQM basée sur le modèle des estimateurs EB sous les modèles de base au niveau du domaine et au niveau de l'unité. Il n'existe aucune expression analytique de l'EQM, sauf dans quelques cas particuliers. Ce problème a fait couler beaucoup d'encre dans la littérature traitant de l'estimation sur petits domaines (EPD), et a mené à des approximations d'ordre deux de l'EQM qui, à leur tour, sont utilisées pour obtenir des estimateurs sans biais d'ordre deux de l'EQM sous les modèles hypothétiques.

3.1 Modèle de base au niveau du domaine

Nous nous concentrons sur les estimateurs du REML des paramètres du modèle, notés $\hat{\boldsymbol{\beta}}$ et $\hat{\sigma}_v^2$. Un estimateur d'ordre deux sans biais inconditionnel de l'EQM sous le modèle de l'estimateur EB est donné par

$$\text{eqm}\left(\hat{\theta}_i^{\text{EB}}\right) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2). \tag{3.1}$$

Ici, le terme principal dans (3.1) est donné par (2.3) avec σ_v^2 remplacé par $\hat{\sigma}_v^2$, et les deux autres termes dans (3.1) sont d'ordre inférieur et rendent compte de l'estimation de $\boldsymbol{\beta}$ et σ_v^2, respectivement (voir Rao et Molina, 2015, chapitre 6 pour des précisions). L'estimateur de l'EQM (3.1) est positif et sans biais à l'ordre deux en ce sens que son biais est d'ordre inférieur à $1/m$ pour m grand. Des méthodes bootstrap paramétriques ont également été utilisées pour obtenir un estimateur de l'EQM. Cependant, l'estimateur résultant n'est pas sans biais à l'ordre deux et un ajustement supplémentaire du biais est effectué pour s'assurer de l'absence de biais à l'ordre deux. Ces ajustements nécessitent habituellement des méthodes bootstrap doubles et certains des estimateurs bootstrap ajustés de l'EQM peuvent prendre des valeurs négatives; voir Rao et Molina (2015), chapitre 6.

3.2 Modèle de base au niveau de l'unité

Nous nous concentrons de nouveau sur l'estimation du REML des paramètres du modèle dans le modèle au niveau de l'unité (2.5). Un estimateur inconditionnel sans biais à l'ordre deux positif de l'EQM de l'estimateur EB $\hat{\mu}_i^{\text{EB}}$ est donné par

$$\text{eqm}\left(\hat{\mu}_i^{\text{EB}}\right) = g_{1i}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) + g_{2i}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) + 2g_{3i}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right), \tag{3.2}$$

où le premier terme est le terme principal donné à la section 2.2, le deuxième terme est dû à l'estimation de $\boldsymbol{\beta}$ et le dernier terme est dû à l'estimation de σ_v^2 et σ_e^2. L'estimateur EB $\hat{\mu}_i^{\text{EB}}$ et l'estimateur de l'EQM inconditionnelle associée (3.2) sont valides quand la fraction d'échantillonnage f_i est négligeable. Le lecteur est invité à consulter (Rao et Molina, 2015, section 7.2.3) pour l'estimation de l'EQM dans le cas de fractions d'échantillonnage non négligeables.

4 Estimation de l'EQM sous le plan de sondage

À la présente section, nous commençons par étudier l'estimation de l'EQM sous le plan, puis nous proposons une estimation composite de l'EQM qui offre un équilibre entre le biais sous le plan et le coefficient de variation.

4.1 Modèle au niveau du domaine

Examinons maintenant l'estimation de l'EQM sous le plan de l'estimateur EB en traitant les paramètres de petit domaine θ_i comme des paramètres inconnus fixes. Comme il est mentionné dans l'introduction, les statisticiens d'enquête cherchent souvent à estimer l'EQM sous le plan des estimateurs EB, pour cadrer avec les estimateurs classiques de l'EQM sous le plan des estimateurs directs pour les grands domaines pour lesquels la taille d'échantillon est adéquate. L'EQM sous le plan est donnée par $\text{EQM}_d\left(\hat{\theta}_i^{\text{EB}}\right) = E\left[\left(\hat{\theta}_i^{\text{EB}} - \theta_i\right)^2 \middle| \boldsymbol{\theta}\right]$, où $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ est le vecteur des moyennes de domaine.

Si nous exprimons $\hat{\theta}_i^{\text{EB}}$ sous la forme $\hat{\theta}_i + h_i\left(\hat{\boldsymbol{\theta}}\right)$ avec $h_i\left(\hat{\boldsymbol{\theta}}\right) = -(1-\hat{\gamma}_i)\left(\hat{\theta}_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}}\right)$, un estimateur exactement sans biais de l'EQM sous le plan est donné par

$$\text{eqm}_d\left(\hat{\theta}_i^{\text{EB}}\right) = \psi_i + 2\psi_i\left[\partial h_i\left(\hat{\boldsymbol{\theta}}\right)\Big/\partial\hat{\theta}_i\right] + h_i^2\left(\hat{\boldsymbol{\theta}}\right). \tag{4.1}$$

Datta, Kubokawa, Molina et Rao (2011) donnent une expression explicite pour la dérivée figurant dans le deuxième terme de (4.1) dans le cas des estimateurs du REML des paramètres du modèle. L'estimateur (4.1) peut prendre des valeurs négatives et être très instable en ce qui concerne la racine carrée de l'erreur quadratique moyenne relative (REQMR), comme le montrent Datta et coll. (2011). Il s'ensuit que (4.1) n'est pas un estimateur fiable de l'EQM sous le plan, même s'il est sans biais par rapport au plan. Les résultats de nos simulations présentés à la section 5 visent à étudier les propriétés conditionnelles des estimateurs (3.1) et (4.1) de l'EQM dans le cadre de l'estimation basée sur le plan de sondage.

Certains éclaircissements théoriques peuvent être obtenus en examinant le cas où les paramètres du modèle sont connus et en considérant le meilleur estimateur (2.2) de la moyenne de domaine θ_i. Dans ce cas, Rivest et Belmonte (2000) ont obtenu un estimateur sans biais sous le plan donné par

$$\text{eqm}_d\left(\tilde{\theta}_i^B\right) = \gamma_i\psi_i + (1-\gamma_i)^2\left[\left(\hat{\theta}_i - \mathbf{z}_i'\boldsymbol{\beta}\right)^2 - (\psi_i + \sigma_v^2)\right]. \tag{4.2}$$

Notons que, pour une grande variance d'échantillonnage ψ_i nous avons $\gamma_i \approx 0$ et (4.2) se réduit à

$$\text{eqm}_d\left(\tilde{\theta}_i^B\right) \approx \left(\hat{\theta}_i - \mathbf{z}_i'\boldsymbol{\beta}\right)^2 - \psi_i. \tag{4.3}$$

Il découle de (4.3) que l'estimateur de l'EQM peut prendre des valeurs négatives et, en fait, la probabilité d'obtenir une valeur négative est proche de 0,5 quand γ_i est proche de zéro ou que la variance d'échantillonnage ψ_i est grande. Dans ce cas particulier où les paramètres du modèle sont connus, nous pouvons étudier le biais sous le plan de l'estimateur sous le modèle de l'EQM de (2.2), donné par $\text{eqm}\left(\tilde{\theta}_i^B\right) = \gamma_i\psi_i$ lorsque l'on prend la moyenne sur l'ensemble des domaines. On peut montrer que le biais moyen sous le plan converge en probabilité sous le modèle vers zéro quand $m \to \infty$ (Rao et Molina, 2015, page 287). Ce résultat donne à penser que l'estimateur de l'EQM sous le modèle devrait avoir de bonnes propriétés en ce qui concerne le biais moyen sous le plan, à condition que le modèle présumé soit valide.

L'estimateur sans biais sous le plan (4.1) n'est pas utilisable en pratique quand il prend une valeur négative pour l'échantillon pris en considération. Donc, nous proposons une modification de (4.1) qui mène à un estimateur positif de l'EQM. Nous désignons l'estimateur de l'EQM modifié par $\text{mod-eqm}_d\left(\hat{\theta}_i^{\text{EB}}\right)$. Il utilise (4.1) quand celui-ci prend une valeur positive pour l'échantillon pris en considération et remplace (4.1) par l'estimation de l'EQM sous le modèle (3.1) quand (4.1) prend une valeur négative. Il est possible d'utiliser une autre estimation positive de l'EQM, par exemple un estimateur de l'EQM sous le plan positif naïf proposé par Pfeffermann et Gilboa (2017). Nous n'avons pas étudié cette modification dans notre étude en simulation.

Nous proposons maintenant des estimateurs composites de l'EQM sous le plan visant à offrir un équilibre entre le biais sous le plan et la REQMR. Nous obtenons un estimateur composite en prenant une moyenne pondérée de l'estimateur de l'EQM sous le plan (4.1) et de l'estimateur inconditionnel de l'EQM sous le modèle (3.1) avec les poids $\hat{\gamma}_i$ et $(1-\hat{\gamma}_i)$, respectivement. Cet estimateur composite de l'EQM peut s'écrire sous la forme

$$\text{eqm}_{c1}\left(\hat{\theta}_i^{\text{EB}}\right) = \hat{\gamma}_i\,\text{eqm}_d\left(\hat{\theta}_i^{\text{EB}}\right) + (1-\hat{\gamma}_i)\,\text{eqm}\left(\hat{\theta}_i^{\text{EB}}\right). \tag{4.4}$$

Il découle de (4.4) que moins de poids est accordé à l'estimateur de l'EQM sous le plan quand la variance d'échantillonnage est grande, ce qui contrôle la REQMR de l'estimateur composite de l'EQM. En outre, le biais par rapport au plan de l'estimateur composite de l'EQM est toujours plus petit que celui de l'estimateur de l'EQM sous le modèle. Quand la valeur de $\hat{\gamma}_i$ (ou de la taille de l'échantillon de domaine) est très petite, un autre choix des poids pour la construction de l'estimateur composite consiste à remplacer $\hat{\gamma}_i$ par $\sqrt{\hat{\gamma}_i}$ et $1-\hat{\gamma}_i$ par $1-\sqrt{\hat{\gamma}_i}$ dans (4.4). L'estimateur composite de l'EQM résultant

$$\text{eqm}_{c2}\left(\hat{\theta}_i^{\text{EB}}\right) = \sqrt{\hat{\gamma}_i}\ \text{eqm}_d\left(\hat{\theta}_i^{\text{EB}}\right) + \left(1 - \sqrt{\hat{\gamma}_i}\right)\text{eqm}\left(\hat{\theta}_i^{\text{EB}}\right) \tag{4.5}$$

donne plus de poids à $\text{eqm}_d\left(\hat{\theta}_i^{\text{EB}}\right)$ que (4.4) et donne donc de meilleurs résultats en ce qui concerne le biais sous le plan au prix d'une augmentation de l'EQM. Comme l'estimateur (4.4), l'estimateur composite de l'EQM de rechange (4.5) possède toujours un biais sous le plan plus petit que l'estimateur de l'EQM sous le modèle. Les estimateurs (4.4) et (4.5) peuvent aussi prendre tous deux des valeurs négatives, mais vraisemblablement moins souvent en raison de leur construction. Pour être certains d'obtenir des estimateurs composites de l'EQM positifs, nous faisons une modification similaire à $\text{mod-eqm}_d\left(\hat{\theta}_i^{\text{EB}}\right)$ et remplaçons (4.4) et (4.5) par l'estimation de l'EQM sous le modèle (3.1) quand ils prennent des valeurs négatives pour l'échantillon pris en considération. Nous désignons les estimateurs modifiés par $\text{mod-eqm}_{c1}\left(\hat{\theta}_i^{\text{EB}}\right)$ et $\text{mod-eqm}_{c2}\left(\hat{\theta}_i^{\text{EB}}\right)$, respectivement. À la section 5, nous examinons la performance des deux estimateurs composites de l'EQM modifiés comparativement à l'estimateur de l'EQM sous le modèle (3.1) et à l'estimateur de l'EQM sous le plan modifié en ce qui concerne le BRA, la REQMR et le taux de couverture des intervalles de confiance.

4.2 Modèle au niveau de l'unité

Nous nous concentrons sur l'échantillonnage aléatoire simple (EAS) sans remise dans chaque domaine. Même pour ce plan de sondage particulier, on ne trouve dans la littérature aucune expression analytique pour l'EQM sous le plan de l'estimateur EB $\hat{\bar{Y}}_i^{\text{EB}}$ ni pour son estimateur, contrairement au cas du modèle au niveau du domaine. Par conséquent, nous proposons une méthode heuristique qui consiste à évaluer l'EQM sous le plan du meilleur estimateur $\hat{\bar{Y}}_i^{B}$ donné par (2.6) sous EAS en supposant que tous les paramètres du modèle sont connus, puis à estimer l'EQM sous le plan. L'estimateur de l'EQM sans biais sous le plan du meilleur estimateur ainsi obtenu dépend des paramètres du modèle et nous remplaçons ces derniers par les estimateurs du REML correspondants. L'estimateur de l'EQM résultant n'est pas sans biais sous le plan pour l'EQM sous le plan de l'estimateur EB, et il sous-estime vraisemblablement l'EQM sous le plan réel, parce qu'il n'est pas tenu compte de la variabilité associée aux paramètres estimés du modèle. Nous étudions la performance de cet estimateur dans une étude en simulation.

Sous EAS sans remise dans le domaine i, nous avons

$$\hat{\bar{Y}}_i^{B} - \bar{Y}_i = a_i\left(\bar{u}_i - \bar{U}_i\right) - (1 - a_i)\bar{U}_i, \tag{4.6}$$

où $\bar{u}_i$ est la moyenne d'échantillon de domaine et $\bar{U}_i$ est la moyenne de population de domaine des valeurs $u_{ij} = y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\beta}$. Il découle de (4.6) que l'EQM sous le plan du meilleur estimateur est donnée par

$$\text{EQM}_d\left(\hat{\bar{Y}}_i^{B}\right) = E_d\left(\hat{\bar{Y}}_i^{B} - \bar{Y}_i\right)^2 = a_i^2\, V_d\left(\bar{u}_i\right) + (1 - a_i)^2\, \bar{U}_i^2, \tag{4.7}$$

où

$$V_d\left(\bar{u}_i\right) = n_i^{-1}\left(1 - f_i\right)S_{ui}^2, \quad \text{et} \quad S_{ui}^2 = (N_i - 1)^{-1}\sum_{j=1}^{N_i}(u_{ij} - \bar{U}_i)^2, \tag{4.8}$$

en notant que le terme correspondant au produit croisé est nul sous EAS.

Il découle maintenant de (4.7) et (4.8) qu'un estimateur de l'EQM sans biais sous le plan du meilleur estimateur est donné par

$$\text{eqm}_d\left(\hat{\bar{Y}}_i^B\right) = a_i^2\, n_i^{-1}\,(1-f_i)\, s_{ui}^2 + (1-a_i)^2\, \hat{\bar{U}}_i^{2D}, \tag{4.9}$$

où $\hat{\bar{U}}_i^{2D} = n_i^{-1}\sum_{j=1}^{n_i} u_{ij}^2 - N_i^{-1}(N_i-1)\, s_{ui}^2$ et $s_{ui}^2 = (n_i-1)^{-1}\sum_{j=1}^{n_i}\left(u_{ij}-\bar{u}_i\right)^2$. En remplaçant les paramètres du modèle dans (4.9) par leurs estimateurs du REML, on obtient un estimateur de l'EQM sous le plan de l'estimateur EB, désigné par $\text{eqm}_d^*\left(\hat{\bar{Y}}_i^{\text{EB}}\right)$. Cet estimateur de l'EQM sous-estime vraisemblablement l'EQM sous le plan de l'estimateur EB, parce que le meilleur estimateur (2.6) ne tient pas compte de la variabilité des estimateurs des paramètres du modèle.

Un estimateur composite de l'EQM, $\text{eqm}_c^*\left(\hat{\bar{Y}}_i^{\text{EB}}\right)$, s'obtient maintenant en prenant une combinaison pondérée de $\text{eqm}_d^*\left(\hat{\bar{Y}}_i^{\text{EB}}\right)$ et de l'estimateur de l'EQM sous le modèle $\text{eqm}\left(\hat{\bar{Y}}_i^{\text{EB}}\right)$ avec les poids $\hat{\gamma}_i$ et $1-\hat{\gamma}_i$, respectivement. Il est donné par

$$\text{eqm}_c^*\left(\hat{\bar{Y}}_i^{\text{EB}}\right) = \hat{\gamma}_i\, \text{eqm}_d^*\left(\hat{\bar{Y}}_i^{\text{EB}}\right) + (1-\hat{\gamma}_i)\, \text{eqm}\left(\hat{\bar{Y}}_i^{\text{EB}}\right). \tag{4.10}$$

Molina et Kominiak (2017) ont proposé des estimateurs bootstrap paramétriques et non paramétriques de l'EQM sous le plan de $\hat{\bar{Y}}_i^{\text{EB}}$. Ils ont également obtenu un estimateur composite de l'EQM, similaire à (4.10), en utilisant l'estimateur bootstrap non paramétrique (BNP) de l'EQM et l'estimateur bootstrap paramétrique (BP) de l'EQM comme composantes de l'estimateur composite de l'EQM associé à $\hat{\gamma}_i$ et $(1-\hat{\gamma}_i)$, respectivement. Comme l'ont fait remarquer ces auteurs, un inconvénient de cet estimateur composite de l'EQM est qu'il faut exécuter à la fois la procédure BP et la procédure BNP pour chaque domaine, ce qui ralentit les calculs. Molina et Kominiak (2017) ont également proposé un estimateur composite bootstrap paramétrique sous le plan (BPP) de l'EQM. L'estimateur BPP évite d'exécuter à la fois la procédure BP et la procédure BNP pour chaque domaine. Les deux estimateurs composites bootstrap de l'EQM ont donné de bons résultats dans une étude en simulation basée sur le plan de sondage.

5 Étude en simulation

À la présente section, nous donnons les résultats d'études en simulation limitées des propriétés sous le plan des estimateurs composites proposés de l'EQM. La section 5.1 présente les résultats pour le modèle au niveau du domaine, et la section 5.2, pour le modèle au niveau de l'unité.

5.1 Modèle au niveau du domaine

En adoptant la configuration de simulation utilisée par Datta et coll. (2011), nous employons le modèle (2.1) avec $m = 30$ domaines, $\mathbf{z}_i = (1, z_{i1})'$ pour $i = 1, \ldots, m$, où les valeurs des covariables $z_{i1}, \ldots, z_{im}$ sont générées indépendamment à partir de la loi $N(-1, 1)$ et maintenues fixes au cours des simulations. En outre, $\boldsymbol{\beta} = (1, 1)'$, $\sigma_v^2 = 1$ et les valeurs de la variance d'échantillonnage sont

(2,0; 0,6; 0,5; 0,4; 0,2), avec chaque valeur différente de ψ_i attribuée à six domaines consécutifs. En notant que $v_i \sim N(0, 1)$, nous générons $\{\theta_i; i = 1, \ldots, m\}$ à partir du modèle de liaison $\theta_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i$ et maintenons les valeurs fixes au cours des simulations pour refléter l'approche basée sur le plan avec conditionnement sur les moyennes de domaine θ_i. Puis, $R = 100\,000$ échantillons simulés $\left\{\hat{\theta}_i^{(r)}: \ i = 1, \ldots, m\right\}$, $r = 1, \ldots, R$ sont générés à partir du modèle d'échantillonnage $\hat{\theta}_i = \theta_i + e_i$ avec l'erreur d'échantillonnage e_i générée à partir de la loi $N(0, \psi_i)$ pour la variance d'échantillonnage spécifiée ψ_i, qui est supposée fixe et connue. Nous notons que notre configuration de simulation n'est pas exactement basée sur le plan, mais qu'elle « s'en approche suffisamment » pour les besoins de notre étude.

Partant des données simulées $\left\{\left(\hat{\theta}_i^{(r)}, \mathbf{z}_i\right): \ i = 1, \ldots, m\right\}$, nous calculons les estimations EB $\hat{\theta}_i^{\mathrm{EB}(r)}$ et obtenons une approximation de l'EQM de $\hat{\theta}_i^{\mathrm{EB}}$ sous la forme

$$\mathrm{EQM}_i^{\mathrm{EB}} = R^{-1}\sum_{r=1}^{R}\left(\hat{\theta}_i^{\mathrm{EB}(r)} - \theta_i\right)^2. \tag{5.1}$$

Nous calculons les estimateurs de l'EQM pour chaque échantillon simulé et prenons leur moyenne sur les 100 000 simulations. Nous désignons les moyennes des estimateurs de l'EQM sur les simulations par $\mathrm{eqm}_i^{\mathrm{EB}}$ $\mathrm{eqm}_{di}^{\mathrm{EB}}$, $\text{mod-eqm}_{di}^{\mathrm{EB}}$, $\text{mod-eqm}_{c1i}^{\mathrm{EB}}$ et $\text{mod-eqm}_{c2i}^{\mathrm{EB}}$, ce qui correspond aux estimateurs de l'EQM sous le modèle, sans biais sous le plan, sans biais sous le plan modifié, composite modifié 1 et composite modifié 2, respectivement. Le biais relatif (BR) de $\mathrm{eqm}_i^{\mathrm{EB}}$ est donné par

$$\mathrm{BR}_i^{\mathrm{EB}} = \left(\mathrm{eqm}_i^{\mathrm{EB}} - \mathrm{EQM}_i^{\mathrm{EB}}\right)/\mathrm{EQM}_i^{\mathrm{EB}} \tag{5.2}$$

où $\mathrm{EQM}_i^{\mathrm{EB}}$ est donné par (5.1). Le biais relatif absolu (BRA) est simplement défini comme étant $\mathrm{BRA}_i^{\mathrm{EB}} = \left|\mathrm{BR}_i^{\mathrm{EB}}\right|$. Les termes $\mathrm{BRA}_{di}^{\mathrm{EB}}$, $\mathrm{BRA}_{di-\mathrm{mod}}^{\mathrm{EB}}$, $\mathrm{BRA}_{c1i-\mathrm{mod}}^{\mathrm{EB}}$ et $\mathrm{BRA}_{c2i-\mathrm{mod}}^{\mathrm{EB}}$ sont définis de la même manière.

Nous calculons aussi la racine carrée de l'erreur quadratique moyenne relative (REQMR) des estimateurs de l'EQM sur les simulations. Nous désignons ces valeurs par $\mathrm{REQMR}_i^{\mathrm{EB}}$, $\mathrm{REQMR}_{di}^{\mathrm{EB}}$, $\mathrm{REQMR}_{di-\mathrm{mod}}^{\mathrm{EB}}$, $\mathrm{REQMR}_{c1i-\mathrm{mod}}^{\mathrm{EB}}$ et $\mathrm{REQMR}_{c2i-\mathrm{mod}}^{\mathrm{EB}}$ pour les estimateurs de l'EQM sous le modèle, sans biais sous le plan, sans biais sous le plan modifié, composite modifié 1 et composite modifié 2, respectivement. Ici, la REQMR de l'estimateur de l'EQM sous le modèle est définie comme étant

$$\mathrm{REQMR}_i^{\mathrm{EB}} = \left\{R^{-1}\sum_{r=1}^{R}\left(\mathrm{eqm}_i^{\mathrm{EB}(r)} - \mathrm{EQM}_i^{\mathrm{EB}}\right)^2\right\}^{1/2} \Big/ \mathrm{EQM}_i^{\mathrm{EB}}. \tag{5.3}$$

Les REQMR des autres estimateurs de l'EQM sont définies de manière similaire.

Nous commençons par comparer la moyenne des $\mathrm{eqm}_i^{\mathrm{EB}}$ sur l'ensemble des domaines à la moyenne des $\mathrm{eqm}_{di}^{\mathrm{EB}}$ sur l'ensemble des domaines. Nous obtenons les valeurs 0,42 et 0,35, respectivement, ce qui montre que la moyenne de l'estimateur de l'EQM sous le modèle, 0,42, est assez proche de la moyenne des EQM sous le plan des estimateurs EB, 0,35, ce qui confirme le résultat théorique mentionné à la section 4.1. Le résultat théorique s'appuie sur l'hypothèse que les paramètres du modèle sont connus, tandis que la simulation traite le cas général de paramètres du modèle inconnus.

Nous examinons ensuite la probabilité d'obtenir une valeur négative pour les trois estimateurs de l'EQM suivants : sans biais sous le plan, composite 1 et composite 2. La figure 5.1 montre le pourcentage de valeurs négatives sur l'ensemble des simulations pour chacun des 30 domaines. Cette figure montre clairement que la probabilité d'obtenir une valeur négative pour l'estimateur de l'EQM sans biais sous le plan peut être aussi élevée que 50 % pour les six premiers domaines (groupe 1), avec une variance d'échantillonnage beaucoup plus grande que pour les domaines restants (groupe 2). Par ailleurs, cette probabilité est négligeable pour les domaines du groupe 2. La probabilité moyenne sur les domaines du groupe 1 est de 45,67 % comparativement à 0,03 % pour le groupe 2. La probabilité d'obtenir une valeur négative pour l'estimateur composite 1 de l'EQM est nulle sur l'ensemble des 30 domaines, tandis que la probabilité moyenne pour l'estimateur composite 2 de l'EQM est de 9,15 % sur les domaines du groupe 1 et nulle sur les domaines du groupe 2. Les résultats susmentionnés donnent à penser que l'estimateur composite 1 de l'EQM pourrait ne pas nécessiter de modification, même pour des domaines pour lesquels la variance d'échantillonnage est grande. Notons que, dans la présente étude en simulation, les estimateurs composite 1 et composite modifié 1 de l'EQM sont identiques, parce qu'aucune valeur nulle n'a été décelée pour l'estimateur composite 1 de l'EQM.

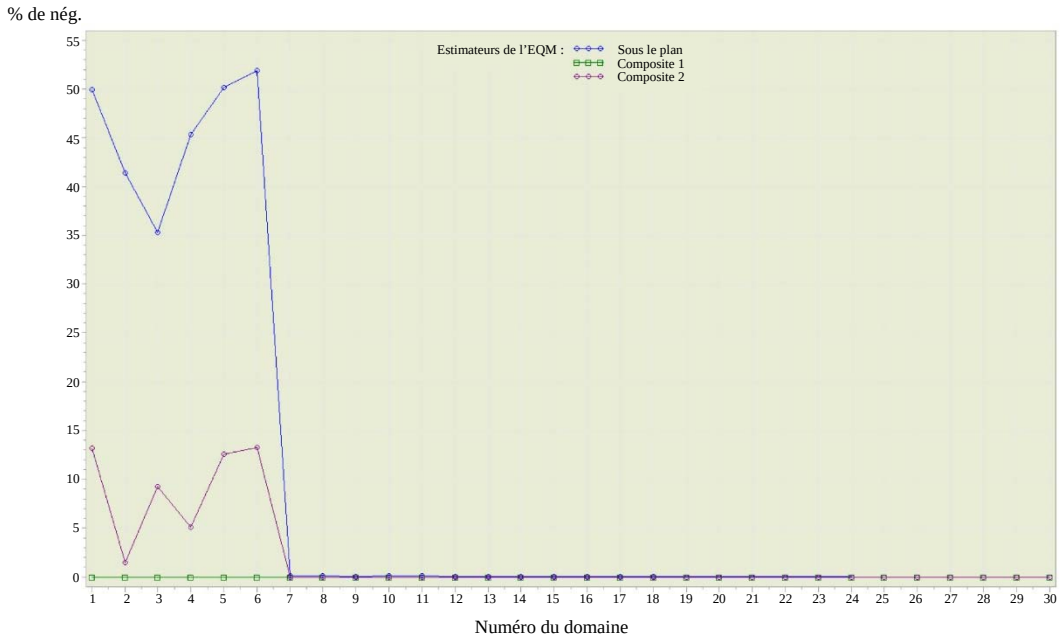

Figure 5.1 Tracé du pourcentage de valeurs négatives des estimateurs de l'EQM : modèle au niveau du domaine.

Examinons maintenant le BRA des estimateurs de l'EQM. La figure 5.2 donne les valeurs du BRA pour chacun des 30 domaines pour les estimateurs de l'EQM suivants : sous le modèle, sans biais sous le plan, sans biais sous le plan modifié, composite 1 modifié et composite 2 modifié. Le tableau 5.1 donne le BRA sous le plan moyen en %, ainsi que la REQMR sous le plan moyenne en % sur les domaines du groupe 1 et du groupe 2.

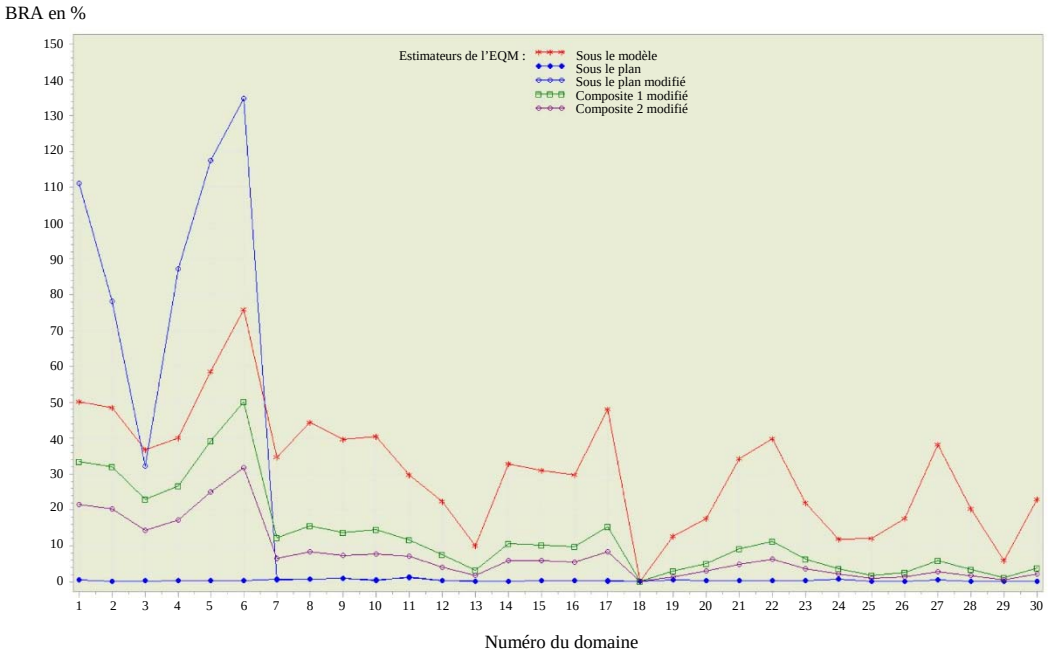

Figure 5.2 Tracé du BRA en % des estimateurs de l'EQM : modèle au niveau du domaine.

Tableau 5.1
BRA sous le plan moyen en % et REQMR sous le plan moyenne en % des estimateurs de l'EQM : modèle au niveau du domaine

Estimateur de l'EQM	BRA sous le plan moyen en %		REQMR sous le plan moyenne en %	
	Domaines 1 à 6	Domaines 7 à 30	Domaines 1 à 6	Domaines 7 à 30
Sous le plan	0,33	0,39	246,71	33,62
Sous le plan modifié	93,49	0,38	221,86	33,58
Sous le modèle	51,66	25,76	54,98	26,61
Composite 1 modifié	34,08	7,60	96,98	24,70
Composite 2 modifié	32,00	4,13	146,31	28,20

Comme prévu, la figure 5.2 montre que le BRA de l'estimateur sans biais sous le plan est nul (hormis les erreurs de simulation) dans tous les domaines. Par ailleurs, étonnamment, le BRA de l'estimateur sans biais sous le plan modifié de l'EQM est grand pour les six premiers domaines, la valeur moyenne étant de 93,49 %, mais il est négligeable pour les autres domaines (0,38 %). L'estimateur de l'EQM sous le modèle présente aussi un grand BRA pour les six premiers domaines, la moyenne étant de 51,66 %, mais passe à 25,76 % pour les autres domaines. Dans le cas de l'estimateur composite 1 de l'EQM, la valeur moyenne du BRA est réduite à 34,08 % pour le groupe 1 et est petite pour le groupe 2 (7,60 %). L'estimateur composite 2 modifié de l'EQM, qui donne plus de poids à l'estimateur sans biais sous le plan de l'EQM, réduit le BRA moyen à 32,00 % pour le groupe 1 et à 4,13 % pour le groupe 2.

La figure 5.3 donne le tracé de la REQMR des estimateurs de l'EQM sur l'ensemble des 30 domaines et le tableau 5.1 donne les valeurs de la REQMR moyenne en % pour les domaines du groupe 1 et du groupe 2. Comme prévu, l'estimateur de l'EQM sans biais sous le plan affiche une très grande REQMR pour le groupe 1, la valeur moyenne étant de 246,71 %. L'estimateur sans biais sous le plan modifié de l'EQM est tout aussi instable pour le groupe 1 (REQMR moyenne de 221,86 %) en plus de présenter un grand BRA. L'estimateur de l'EQM sous le modèle est celui dont la REQMR est la plus petite, comme prévu, la valeur moyenne étant de 54,98 % pour le groupe 1 comparativement à 96,98 % pour l'estimateur composite 1 de l'EQM et à 146,31 % pour l'estimateur composite 2 modifié de l'EQM. Par ailleurs, pour les domaines du groupe 2 dont les variances d'échantillonnage sont plus petites, la REQMR moyenne est à peu près la même pour les trois estimateurs de l'EQM, soit 24,70 % pour l'estimateur composite 1, 26,61 % pour l'estimateur sous le modèle et 28,20 % pour l'estimateur composite 2 modifié de l'EQM. La REQMR moyenne pour les estimateurs de l'EQM sans biais sous le plan et sans biais sous le plan modifié est à peine plus grande pour le groupe 2, les valeurs étant de 33,62 % et 33,58 %, respectivement.

Enfin, nous nous tournons vers les taux de couverture des intervalles de confiance pour une valeur nominale de 95 %. Les taux de couverture selon la théorie normale pour l'estimateur de l'EQM sous le modèle se calculent comme il suit

$$\mathrm{TC}\left[\mathrm{eqm}\left(\hat{\theta}_i^{\mathrm{EB}}\right)\right] = R^{-1}\sum_{r=1}^{R} I\left[\hat{\theta}_i^{\mathrm{EB}(r)} - 1{,}96\left(\mathrm{eqm}_i^{\mathrm{EB}(r)}\right)^{1/2} \le \theta_i \le \hat{\theta}_i^{\mathrm{EB}(r)} + 1{,}96\left(\mathrm{eqm}_i^{\mathrm{EB}(r)}\right)^{1/2}\right] \quad (5.4)$$

où $I[\bullet]$ est une fonction indicatrice prenant la valeur 1 si θ_i est compris dans l'intervalle calculé et 0 autrement. Les taux de couverture pour les autres estimateurs de l'EQM sont définis de manière similaire. La figure 5.4 donne le tracé des taux de couverture en % pour les estimateurs de l'EQM. La courbe associée à l'estimateur de l'EQM sans biais sous le plan n'est pas incluse dans le graphique parce qu'il est impossible de calculer les taux de couverture des intervalles de confiance en raison d'estimations négatives de l'EQM pour certaines simulations. Le fait d'écarter ces simulations et de calculer les intervalles pour les simulations restantes pourrait fausser le taux de couverture.

Le graphique révèle une importante sous-couverture pour les domaines du groupe 1 dont la variance d'échantillonnage est grande. En particulier, le taux de couverture moyen pour les estimateurs sous le modèle, composite 1 modifié et composite 2 modifié est de 68,53 %, 78,43 % et 72,87 %, respectivement, tandis que l'estimateur sous le plan modifié de l'EQM présente une certaine amélioration, le taux étant de 85,82 %. Par ailleurs, pour les domaines du groupe 2 dont les variances d'échantillonnage sont plus petites, le taux de couverture moyen augmente pour atteindre 91,73 %, 91,74 %, 90,89 % et 89,85 % pour les estimateurs de l'EQM sous le modèle, composite 1 modifié, composite 2 modifié et sous le plan modifié, respectivement. La figure 5.4 donne à penser que les taux de couverture pour les estimateurs sous le modèle et composites modifiés de l'EQM sont comparables pour tous les domaines, ceux du groupe 1 présentant une sous-couverture importante en raison des petites tailles d'échantillon ou des grandes variances d'échantillonnage dans ces domaines.

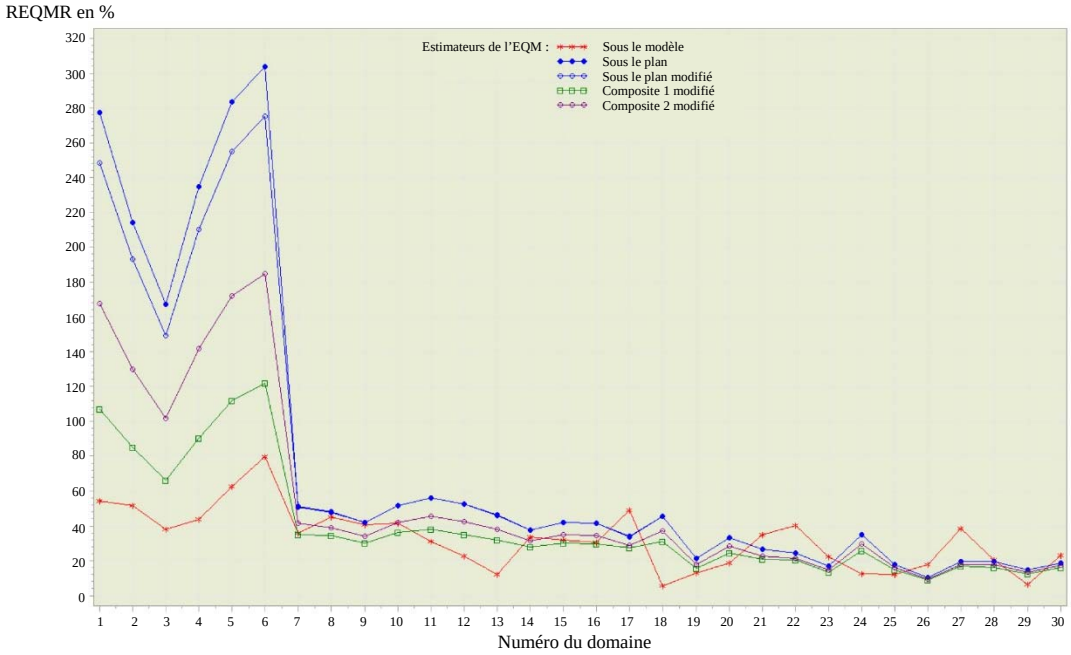

Figure 5.3 Tracé de la REQMR en % des estimateurs de l'EQM : modèle au niveau du domaine.

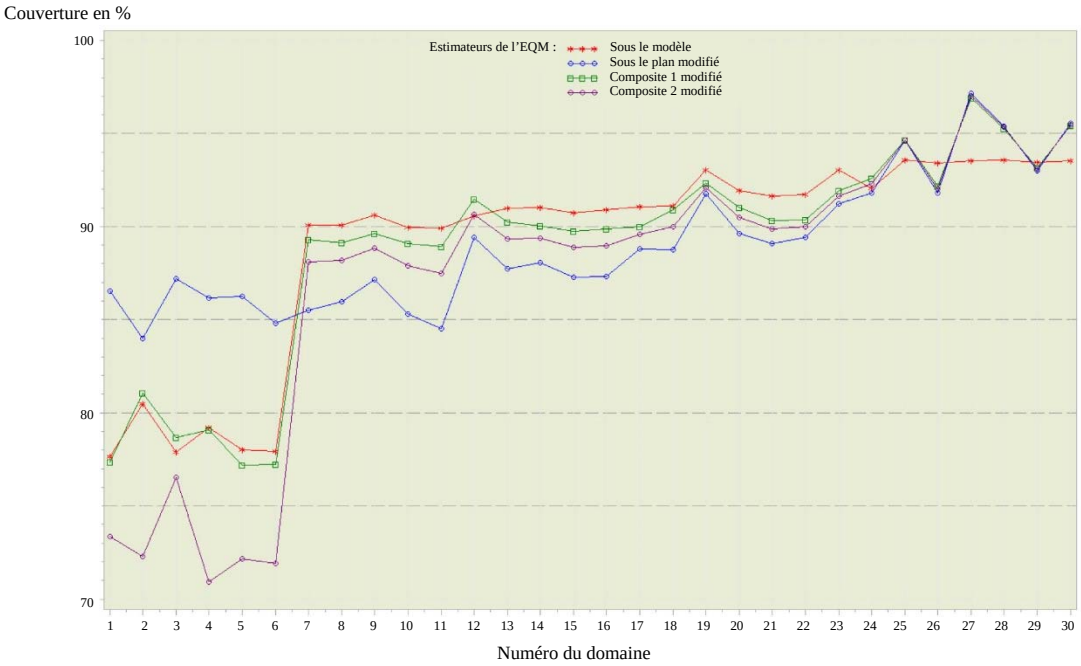

Figure 5.4 Tracé des taux de couverture en % pour les estimateurs de l'EQM : modèle au niveau du domaine.

5.2 Modèle au niveau de l'unité

À la présente section, nous présentons certains résultats d'une étude en simulation limitée des propriétés par rapport au plan de sondage de quatre estimateurs de l'EQM sous un simple modèle de moyenne au niveau de l'unité donné par

$$y_{ij} = \beta + v_i + e_{ij}, \quad j = 1, \ldots, N_i; \quad i = 1, \ldots, m \tag{5.5}$$

où les effets aléatoires de domaine $v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$ sont indépendants des erreurs au niveau de l'unité $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$. Les estimateurs de l'EQM étudiés comprennent l'estimateur de l'EQM sous le modèle $\text{eqm}\left(\hat{\bar{Y}}_i^{\text{EB}}\right)$ de l'estimateur EB $\hat{\bar{Y}}_i^{\text{EB}}$ (Rao et Molina, 2015, section 7.2.3), l'estimateur par substitution (*plug-in*) de l'EQM sous le plan $\text{eqm}_d^*\left(\hat{\bar{Y}}_i^{\text{EB}}\right)$ obtenu à partir de (4.9) en remplaçant les paramètres du modèle β, σ_v^2 et σ_e^2 par les estimateurs de leur REML, l'estimateur de l'EQM composite donné par (4.10), et un estimateur « conditionnel » de l'EQM, $\text{eqm}_{\text{CH}}\left(\hat{\bar{Y}}_i^{\text{EB}}\right)$, proposé par Chambers, Chandra et Tzavidis (2001, section 2.2.2).

Pour la simulation basée sur le plan de sondage, nous utilisons $m = 30$ petits domaines et commençons par générer les tailles de population de domaine N_i, à partir d'une loi uniforme $U[443, 542]$ et les maintenons fixes pour toutes les simulations, à l'instar de Chambers et coll. (2011). Nous générons deux populations finies fixes $\{y_{ij},\ j = 1, \ldots, N_i;\ i = 1, \ldots, m\}$ à partir du modèle de moyenne (5.5) pour le paramètre moyen spécifié $\beta = 500$ et les paramètres de variance $\sigma_v^2 = 10{,}40$, $\sigma_e^2 = 94{,}09$ pour la première population finie (désignée population A) et $\beta = 500$, $\sigma_v^2 = 40{,}32$, $\sigma_e^2 = 94{,}09$ pour la deuxième population finie (désignée population B). Notons que le ratio des variances $\delta = \sigma_v^2 / \sigma_e^2$ est égal à 0,11 pour la population A et est inférieur à 0,43 pour la population B. Ensuite, nous tirons des échantillons aléatoires simples stratifiés $\{y_{ij},\ j = 1, \ldots, n_i;\ \ i = 1, \ldots, 30\}$ sans remise de chaque population finie, en traitant chaque domaine comme une strate, où les tailles d'échantillon de domaine sont choisies égales, soit $n_i = 5$ ou $n_i = 20$. En tout, nous tirons $S = 10\ 000$ échantillons aléatoires simples stratifiés et calculons les estimations de l'EQM pour chaque échantillon. Indépendamment, nous tirons aussi $R = 30\ 000$ échantillons aléatoires stratifiés et calculons les estimations EB d'après chaque échantillon. Une approximation de l'EQM de l'estimateur EB pour chaque domaine est obtenue de manière analogue à (5.1) en utilisant les 30 000 simulations. En utilisant un grand nombre de simulations, $R = 30\ 000$, l'EQM empirique représente une bonne approximation de l'EQM réelle de l'estimateur EB. Par ailleurs, un plus petit nombre de simulations, tel que $S = 10\ 000$, est utilisé pour étudier la performance des quatre estimateurs de l'EQM, afin de réduire les calculs. Cette configuration de simulation en deux étapes est souvent utilisée pour le modèle au niveau de l'unité (voir, par exemple, González-Manteiga, Lombardia, Molina, Morales et Santamaria, 2008). Habituellement, le calcul de l'EQM est beaucoup plus rapide que le calcul du BR et de la REQMR de plusieurs estimateurs de l'EQM, en particulier les estimateurs bootstrap de l'EQM.

En utilisant les estimations de l'EQM simulées et l'EQM de l'estimateur EB simulé, nous calculons le biais relatif (BR), le biais relatif absolu (BRA) et la racine carrée de l'erreur quadratique moyenne relative

(REQMR) des estimateurs de l'EQM de manière analogue à (5.2) et (5.3). Dans le cas de la population A et d'une taille d'échantillon de domaine de 5, l'estimateur par substitution de l'EQM sous le plan donne lieu à une sous-estimation dans tous les domaines, le BR variant de -87,0 % à -18,1 %. Cette sous-estimation est due au fait qu'il n'est pas tenu compte de la variabilité des estimations des paramètres. Par ailleurs, l'estimateur de l'EQM sous le modèle surestime généralement l'EQM sous le plan, le BR variant de -66,4 % à 150,1 %. D'où, l'estimateur composite de l'EQM réduit la sous-estimation causée par l'estimateur par substitution de l'EQM sous le plan, le BR variant de -55,0 % à 115,4 %. L'estimateur de l'EQM conditionnel surestime systématiquement l'EQM sous le plan, le BR variant de 31,7 % à 316,1 %. La performance des estimateurs de l'EQM en ce qui concerne le BR s'améliore à mesure que le ratio δ s'élève vers 0,43 ou que la taille de l'échantillon de domaine augmente pour atteindre 20.

Le tableau 5.2 donne les valeurs médiane et moyenne du BRA pour les deux populations et les deux tailles d'échantillon. Nous constatons que l'estimateur composite de l'EQM donne de meilleurs résultats que les autres estimateurs pour la population A et une taille de domaine de 5, les BRA médian et moyen valant 53 %. Par contre, l'estimateur conditionnel de l'EQM présente de grandes valeurs du BRA médian, soit 208 %, et du BRA moyen, soit 191 %. Les valeurs médiane et moyenne du BRA pour tous les estimateurs de l'EQM diminuent à mesure que le ratio δ augmente pour atteindre 0,43 ou que la taille d'échantillon diminue pour atteindre 20.

Tableau 5.2
BRA sous le plan médian et moyen en % des estimateurs de l'EQM : modèle au niveau de l'unité

Estimateurs de l'EQM	Population A				Population B			
	$n_i = 5$		$n_i = 20$		$n_i = 5$		$n_i = 20$	
	Médian	Moyen	Médian	Moyen	Médian	Moyen	Médian	Moyen
Sous le plan	60,7	54,4	11,2	11,1	8,9	8,9	1,8	2,0
Conditionnel	207,9	190,7	23,2	19,9	9,4	8,3	0,7	1,0
Sous le modèle	77,4	81,7	44,0	38,8	29,6	28,4	6,8	8,6
Composite	52,9	53,3	13,1	14,0	7,1	8,8	1,3	1,8

Le tableau 5.3 donne les valeurs médiane et moyenne en % de la REQMR sous le plan de sondage pour les deux populations et les deux tailles d'échantillon. Il montre que l'estimateur de l'EQM sous le modèle et l'estimateur de l'EQM composite donnent de meilleurs résultats que les autres estimateurs de l'EQM, surtout pour la population A et la taille de domaine de 5. Dans ce dernier cas, l'estimateur par substitution de l'EQM sous le plan et l'estimateur conditionnel de l'EQM présentent de grandes valeurs médianes et moyennes de la REQMR, soit environ 400 % contre 110 % pour l'estimateur de l'EQM sous le modèle et l'estimateur de l'EQM composite. La performance de tous les estimateurs de l'EQM s'améliore en ce qui concerne la REQMR à mesure que le ratio δ augmente ou que la taille d'échantillon de domaine augmente. Dans le cas de la population B et de la taille de domaine de 20, l'estimateur de l'EQM sous le modèle présente les plus petites valeurs médiane et moyenne de la REQMR, soit environ 10 % contre 30 % pour les autres estimateurs de l'EQM.

Tableau 5.3
REQMR sous le plan médiane et moyenne en % des estimateurs de l'EQM : modèle au niveau de l'unité

Estimateurs de l'EQM	Population A				Population B			
	$n_i = 5$		$n_i = 20$		$n_i = 5$		$n_i = 20$	
	Médian	Moyen	Médian	Moyen	Médian	Moyen	Médian	Moyen
Sous le plan	414,5	382,0	62,1	60,3	57,6	57,6	29,3	29,0
Conditionnel	416,5	384,5	64,1	62,2	63,9	64,3	28,4	28,1
Sous le modèle	107,8	108,5	45,4	41,6	31,6	31,7	8,9	10,8
Composite	113,7	112,9	37,8	38,1	40,7	41,5	26,6	26,4

6 Conclusion

Dans le présent article, nous étudions les propriétés de divers estimateurs de l'EQM en vue d'évaluer l'EQM sous le plan de sondage des estimateurs EB de moyennes de petits domaines. Nous examinons les modèles au niveau du domaine ainsi qu'au niveau de l'unité.

Dans le cas du modèle au niveau du domaine, nous proposons deux estimateurs composites de l'EQM en prenant une moyenne pondérée d'un estimateur de l'EQM sans biais sous le plan et d'un estimateur de l'EQM sous un modèle. Des modifications afin de s'assurer d'obtenir des estimateurs positifs de l'EQM sont également présentées. La performance des divers estimateurs de l'EQM est étudiée au moyen de simulations en examinant le biais relatif absolu, la racine carrée de l'erreur quadratique moyenne relative et le taux de couverture des intervalles de confiance. Nos résultats pour le modèle au niveau du domaine donnent à penser que l'estimateur de l'EQM sans biais sous le plan n'est pas utilisable en pratique quand la taille d'échantillon de domaine est très petite, étant donné la forte probabilité d'obtenir une valeur négative. Par ailleurs, cette probabilité pour l'estimateur composite 1 de l'EQM (utilisant les mêmes poids que l'estimateur EB) est nulle ou essentiellement négligeable. Pour les domaines dont la taille d'échantillon est très petite, nos simulations pour le modèle au niveau du domaine laissent entendre que l'estimateur composite 1 de l'EQM produit un BRA plus petit que l'estimateur de l'EQM sous le modèle, au prix d'une augmentation de la REQMR. Pour les domaines dont la taille d'échantillon est plus grande, le BRA de l'estimateur de l'EQM sous le modèle persiste, contrairement au BRA de l'estimateur composite 1 de l'EQM. Pour ce qui est des taux de couverture des intervalles de confiance, l'estimateur de l'EQM sous le modèle et l'estimateur composite 1 de l'EQM sont comparables dans tous les domaines, mais tous deux peuvent donner lieu à une sous-couverture importante pour les domaines dont la taille d'échantillon est très petite. Dans l'ensemble, l'estimateur composite 1 de l'EQM offre un bon compromis pour estimer l'EQM sous le plan.

Dans l'étude en simulation du modèle au niveau de l'unité, nos résultats donnent à penser que l'estimateur composite de l'EQM offre généralement un bon compromis entre le BRA et la REQMR. Cependant, l'estimateur par substitution de l'EQM sous le plan utilisé dans l'estimateur composite doit être modifié afin de tenir compte de la variabilité des estimateurs des paramètres du modèle, afin d'éviter ou de réduire la sous-estimation de l'EQM sous le plan de sondage de l'estimateur EB.

Remerciements

Les auteurs remercient les rédacteurs adjoints et associé, et les deux examinateurs de leurs commentaires et suggestions qui leur ont permis d'améliorer leur examen des divers estimateurs de l'EQM.

Bibliographie

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association,* 83, 401, 28-36.

Beaumont, J.-F., et Bocci, C. (2016). Small area estimation in the Labour Force Survey. Manuscrit non publié.

Chambers, R., Chandra, H. et Tzavidis, N. (2011). Estimation de l'erreur quadratique moyenne robuste au biais pour les estimateurs sur petits domaines pseudo linéaires. *Techniques d'enquête*, 37, 2, 163-181. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2011002/article/11604-fra.pdf.

Datta, G., Kubokawa, T., Molina, I. et Rao, J.N.K. (2011). Estimation of mean squared error of model-based small area estimators. *Test*, 20, 367-388.

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

González-Manteiga, W., Lombardia, M.J., Molina, I., Morales, D. et Santamaria, L. (2008). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, 78, 443-462.

Molina, I., et Kominiak, E.S. (2017). Estimation of proportions in small areas: Application to the labour force using the Swiss Census Structural Survey. Rapport technique non publié.

Pfeffermann, D., et Gilboa, A. (2017). Estimation of randomization MSE in small area estimation. Document présenté à la conférence SAE 2017, Paris.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation, Second Edition.* Hoboken, New Jersey: Wiley.

Rivest, L.-P., et Belmonte, E. (2000). Une erreur quadratique moyenne conditionnelle des estimateurs régionaux. *Techniques d'enquête*, 26, 1, 79-90. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2000001/article/5179-fra.pdf.

Rubin-Bleuer, S., et Schiopu-Kratina, I. (2005). On the two-phase framework for joint model and design-based inference. *The Annals of Statistics*, 33, 6, 2789-2810.