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

Techniques d’enquête

Estimation sur petits domaines d’indicateurs généraux pendant les années intercensitaires

par William Acero, Isabel Molina et J. Miguel Marín

Date de diffusion : le 29 juin 2026

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 infostats@statcan.gc.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

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.

Techniques d'enquête, juin 2026
Vol. 52, nº 1, pp. 31-66
Statistique Canada, nº 12-001-X au catalogue

Estimation sur petits domaines d'indicateurs généraux pendant les années intercensitaires

William Acero, Isabel Molina et J. Miguel Marín[1]

Résumé

Nous proposons des estimateurs sur petits domaines d'indicateurs généraux pour les années intercensitaires, qui permettent d'éviter l'utilisation de microdonnées de recensement désuètes, tout en étant presque optimaux lors des années de recensement. La procédure consiste à remplacer le fichier de recensement désuet par une enquête au niveau des unités, à plus grande échelle, qui couvre adéquatement les domaines d'intérêt et contient les valeurs des variables auxiliaires pertinentes. Toutefois, l'exigence minimale en matière de données de la méthode proposée se limite à une seule enquête contenant des microdonnées sur la variable cible et les variables auxiliaires appropriées pour la période visée. Nous élaborons également un estimateur de l'erreur quadratique moyenne (EQM) qui tient compte de l'incertitude attribuable à l'enquête à grande échelle utilisée pour remplacer le recensement d'information auxiliaire. Nos résultats empiriques indiquent que les prédicteurs proposés ont nettement un meilleur rendement que les prédicteurs de substitution lorsque les données du recensement sont désuètes et qu'ils sont très proches des prédicteurs optimaux lorsque les données du recensement sont corrects. Par ailleurs, ils montrent que l'estimateur proposé de l'EQM totale corrige le biais des estimateurs de l'EQM purement fondés sur des modèles, qui ne tiennent pas compte de l'incertitude associée à l'enquête à grande échelle.

Mots-clés : Années intercensitaires; données d'enquête au niveau des unités; erreur quadratique moyenne; indicateurs généraux; meilleur prédicteur empirique.

1. Introduction

Les modèles mixtes sont souvent utilisés dans l'estimation sur petits domaines, car ils permettent d'emprunter de l'information de tous les domaines, tout en préservant la spécificité de chacun d'entre eux grâce à l'inclusion d'effets aléatoires pour les domaines dans le modèle. Les prédicteurs optimaux (ou les « meilleurs » prédicteurs) des indicateurs cibles de domaine sont alors obtenus en minimisant l'erreur quadratique moyenne selon le modèle. Étant donné que les meilleurs prédicteurs dépendent des paramètres inconnus du modèle, ces paramètres sont alors remplacés par des estimations basées sur des données échantillonnées, ce qui donne les meilleurs prédicteurs empiriques (EB pour *empirical best* en anglais). Pour obtenir une bonne revue de l'estimation sur petits domaines dans le cadre de modèles mixtes, voir par exemple Jiang et Lahiri (2006).

Les modèles linéaires mixtes (MLM) et les modèles linéaires mixtes généralisés (MLMG) constituent des cas spéciaux de modèles mixtes. Dans le cadre du MLM avec effets aléatoires de domaine proposé par Battese, Harter et Fuller (1988) et en supposant la normalité, Molina et Rao (2010) ont obtenu des prédicteurs EB d'indicateurs généraux de domaine et ont illustré la procédure d'estimation des indicateurs de pauvreté ou d'inégalité définis en fonction d'une seule mesure monétaire du bien-être. On suppose le modèle

1. William Acero, Department of Statistics and Operational Research, Faculty of Mathematics, Complutense University of Madrid, Plaza de las Ciencias 3, Ciudad Universitaria, 28040 Madrid, Espagne. Courriel : wacero@ucm.es; Isabel Molina, Department of Statistics and Operational Research, Faculty of Mathematics, Complutense University of Madrid, Plaza de las Ciencias 3, Ciudad Universitaria, 28040 Madrid, Espagne, et Interdisciplinary Mathematics Institute (IMI), Faculty of Mathematics, Complutense University of Madrid, Plaza de las Ciencias 3, Ciudad Universitaria, 28040 Madrid, Espagne; Miguel Marín, Department of Statistics, Carlos III University of Madrid, Calle Madrid, 126, 28903 Getafe, Madrid, Espagne.

pour une transformation bijective générale (telle qu'une transformation logarithmique) de la mesure du bien-être. Selon le même modèle, le prédicteur EB obtenu par Molina et Rao (2010) est beaucoup plus efficace que l'estimateur ELL d'Elbers, Lanjouw et Lanjouw (2003), notamment lorsque les domaines d'intérêt présentent des effets individuels marqués. Pour estimer les erreurs quadratiques moyennes (EQM) des prédicteurs EB dans le cadre du modèle de Battese et coll. (1988), Molina et Rao (2010) ont proposé d'utiliser la procédure bootstrap paramétrique pour populations finies de González-Manteiga, Lombardía, Molina, Morales et Santamaría (2008).

Lorsque des modèles mixtes sont spécifiés au niveau des unités, la prédiction EB nécessite une enquête au niveau des unités dans laquelle la variable cible et plusieurs variables auxiliaires sont observées, ainsi qu'un recensement contemporain contenant des microdonnées sur les mêmes variables auxiliaires. La prédiction EB exige en outre l'identification des unités de l'enquête dans le fichier du recensement, ce qui n'est pas toujours possible. Pour éviter cette étape, une petite variante du prédicteur EB appelée meilleur prédicteur empirique du recensement (CEB pour *census empirical best* en anglais) a été créée, et la procédure bootstrap a été étendue à l'estimation de l'EQM (Molina, 2019). Selon le MLM de Battese et coll. (1988), Corral, Molina et Nguyen (2020) ont proposé une extension supplémentaire du prédicteur CEB afin de prendre en compte l'hétéroscédasticité et les poids d'enquête, de manière semblable à la procédure pseudo-EB de Guadarrama, Molina et Rao (2018) et à l'aide d'estimateurs pondérés par les poids d'enquête des composantes de variance comme dans Van der Weide (2014). Cette procédure pseudo-EB étendue au recensement a été mise en œuvre dans le cadre de la méthodologie de la Banque mondiale pour la cartographie de la pauvreté. D'autres exemples d'application de la procédure EB originale de Molina et Rao (2010) pour estimer la pauvreté à partir de microdonnées d'enquête et de recensement peuvent être consultés, par exemple, dans Molina et Rao (2010), Molina et Martín (2018), Molina (2019) et Molina et García-Portugués (2019).

Comme nous l'avons déjà mentionné, la prédiction EB nécessite un recensement des variables auxiliaires considérées comme devant être prises en compte dans l'enquête. Parfois, plutôt que le recensement requis, une enquête à plus grande échelle est disponible, couvrant de manière adéquate tous les domaines cibles. Cette enquête contient des variables auxiliaires communes à l'enquête originale à plus petite échelle dans laquelle la variable cible est observée. En outre, même dans les pays où le recensement est toujours effectué, le dernier recensement peut être désuet pendant les années intercensitaires, ce qui donne lieu à des prédicteurs EB biaisés. Différentes approches ont été proposées pour résoudre ce problème. Les modèles au niveau du domaine, comme le modèle de Fay-Herriot (FH) présenté par Fay et Herriot (1979), reposent uniquement sur l'information auxiliaire agrégée (comme les moyennes), qui peut être tirée du dernier recensement. Si aucune donnée de recensement actualisée n'est disponible, une enquête à plus grande échelle pourrait être utilisée pour obtenir des estimations des agrégats de domaine pour chaque variable auxiliaire. Toutefois, l'erreur attribuable à ces estimateurs doit être prise en compte. Ybarra et Lohr (2008) ont proposé un prédicteur sur petits domaines qui tient compte de l'erreur de mesure de l'information

auxiliaire du modèle de FH. Cette méthode nécessite la connaissance des véritables EQM des estimateurs de l'information auxiliaire agrégée (ou de leur variance, s'ils sont non biaisés) pour chaque domaine.

Les modèles au niveau du domaine réduisent la richesse de l'information au niveau des unités provenant de l'enquête à des agrégats de domaine. En outre, ils exigent l'élaboration de modèles appropriés pour chaque indicateur d'intérêt et pour chaque niveau d'agrégation pour lequel des estimations sont souhaitées. Si de nombreux indicateurs présentent un intérêt, il peut s'avérer difficile de trouver des variables auxiliaires agrégées qui sont linéairement liées à tous les indicateurs cibles.

Une autre approche consiste à n'utiliser que l'information auxiliaire agrégée dans un modèle dont la variable réponse est spécifiée au niveau des unités. Il en résulte ce que l'on appelle les modèles du contexte de l'unité, appliqués, par exemple, par Masaki, Newhouse, Silwal, Bedada et Engstrom (2022) au moyen de la prédiction EB et par Cuong (2012) au moyen de l'approche ELL. Corral, Himelein, McGee et Molina (2021) ont comparé différentes approches pour estimer les indicateurs de pauvreté dans des expériences de simulation fondées sur un modèle et sur un plan de sondage, en utilisant l'Enquête intercensitaire mexicaine comme recensement. Les résultats de simulation ont mis en évidence un biais marqué pour les estimateurs obtenus dans le cadre des modèles du contexte de l'unité.

Nous abordons l'estimation d'indicateurs généraux dans le cadre de modèles généraux en deux étapes, lorsque les données de recensement sur les variables auxiliaires sont désuètes ou indisponibles. Le modèle comprend le MLM de Battese et coll. (1988) et des modèles linéaires mixtes généralisés, comme le modèle linéaire mixte logistique étudié par Jiang et Lahiri (2001), bien que la procédure soit extensible à des modèles plus complexes. Au lieu d'un recensement contenant des microdonnées sur les variables auxiliaires, nous supposons qu'une enquête secondaire à plus grande échelle, pour laquelle il y a des valeurs des variables auxiliaires au niveau des unités, est disponible et que ses tailles d'échantillon par domaine sont au moins aussi grandes que celles de l'enquête originale contenant la variable cible. Les deux enquêtes sont censées être contemporaines. Nous proposons une procédure qui permet d'adapter le modèle aux données de l'enquête à plus petite échelle, puis d'utiliser l'approche du prédicteur EB pour la prédiction à partir des valeurs des variables auxiliaires observées dans l'enquête secondaire. Une approche similaire est suivie par Sen et Lahiri (2025) dans un modèle linéaire mixte logistique afin d'estimer les résultats de l'élection présidentielle américaine de 2016 pour les 50 États et le district de Columbia.

Un cas spécial de notre procédure se présente lorsque seule l'enquête contenant la variable cible est disponible. Nos expériences de simulation montrent que, même dans ce cas, les estimateurs proposés donnent quand même de biens meilleurs résultats que les estimateurs directs habituels fondés sur le plan en ce qui concerne le biais et l'EQM, selon le modèle et le plan. En revanche, le meilleur cas spécial correspond à la situation où l'enquête secondaire est en réalité un recensement, auquel cas notre procédure fournit les prédicteurs CEB habituels.

Sen et Lahiri (2025) ont proposé une méthode bootstrap pour estimer l'EQM purement fondée sur le modèle sans tenir compte de l'incertitude découlant de l'utilisation de l'enquête secondaire. Nous proposons

des estimateurs de l'EQM totale qui corrigent les estimateurs de l'EQM fondés sur le modèle en ajoutant l'incertitude liée au plan d'échantillonnage provenant de l'enquête secondaire. Nos résultats de simulation indiquent que cette correction est nécessaire.

La présente étude est structurée de la manière suivante. La section 2 présente le modèle à deux niveaux et définit les indicateurs cibles. À la section 3, on passe en revue le prédicteur (EB) d'un indicateur additif général et l'on décrit le biais résultant de l'utilisation d'un recensement désuet lors de l'estimation des moyennes de domaine. La section 4 présente le nouveau prédicteur qui remédie aux limites de la procédure EB avec des données de recensement désuètes, pour éviter la perte d'information résultant de l'agrégation des modèles au niveau du domaine et le biais attribuable aux modèles du contexte de l'unité. Pour le nouveau prédicteur, on considère une enquête probabiliste secondaire à plus grande échelle comme source plus actuelle de données auxiliaires lorsqu'une telle enquête est disponible, bien que sa disponibilité ne soit pas strictement requise. La section 5 présente un estimateur de l'EQM totale reflétant les deux sources de variabilité inhérentes au prédicteur proposé, à savoir la variabilité fondée sur le modèle et la variabilité fondée sur le plan découlant de l'enquête secondaire. La section 6 compare empiriquement le rendement de différents prédicteurs et estimateurs de l'EQM. À la section 7, la méthode proposée est appliquée à l'estimation des taux de pauvreté par département croisés avec l'origine ethnique autodéclarée en Colombie. Enfin, des conclusions sont présentées à la section 8.

2. Modèle et indicateurs cibles de domaine

Nous considérons une population finie U de taille N, divisée en D sous-populations appelées domaines ou zones, $U_1,\ldots,U_D$, de tailles $N_1,\ldots,N_D$. Soit y_{di}, la valeur de la variable cible, $\mathbf{x}_{di}$, un vecteur de p variables auxiliaires pour l'unité i au sein du domaine d, pour $i=1,\ldots,N_d$, et u_d, un effet aléatoire du domaine d, $d=1,\ldots,D$. Les idées développées dans le présent article sont plutôt générales et peuvent être appliquées à des modèles beaucoup plus complexes. Toutefois, pour simplifier l'exposé, nous considérons un modèle à deux niveaux de la forme

$$\begin{aligned} y_{di}\,|\,u_d &\overset{\text{ind}}{\sim} f_1(y_{di}\,|\,u_d,\mathbf{x}_{di},\boldsymbol{\theta}_1), \quad i=1,\ldots,N_d, \\ u_d &\overset{\text{iid}}{\sim} f_2(u_d;\boldsymbol{\theta}_2), \quad d=1,\ldots,D; \end{aligned} \tag{2.1}$$

voir, par exemple, Pfeffermann et Sverchkov (2007). Dans le modèle ci-dessus, $\boldsymbol{\theta}=(\boldsymbol{\theta}_1^t,\boldsymbol{\theta}_2^t)^t \in \Theta \subset \mathbb{R}^k$ désigne un vecteur de paramètres inconnus du modèle. Ce modèle à deux niveaux englobe plusieurs des modèles utilisés en estimation sur petits domaines, bien que nos résultats s'appliquent à des modèles plus généraux.

Un cas spécial de ce modèle, obtenu en supposant que f_1 et f_2 suivent des distributions normales, est le modèle de régression linéaire à erreurs emboîtées (RLEE) bien connu proposé par Battese et coll. (1988),

$$y_{di} = \mathbf{x}_{di}^t\boldsymbol{\beta} + u_d + e_{di}, u_d \overset{\text{iid}}{\sim} N(0,\sigma_u^2),$$

$$e_{di} \overset{\text{iid}}{\sim} N(0,\sigma_e^2), \quad i=1,\ldots,N_d, d=1,\ldots,D, \tag{2.2}$$

où u_d et e_{di} sont indépendantes, $\boldsymbol{\beta}$ est un vecteur de coefficients de régression inconnus de dimension p, et $\sigma_u^2 > 0$ et $\sigma_e^2 > 0$ sont des variances inconnues. Dans ce modèle, $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}^t, \sigma_e^2)^t$, $\boldsymbol{\theta}_2 = \sigma_u^2$ et, par conséquent, le nombre de paramètres est $k = p + 2$.

Les modèles linéaires mixtes généralisés constituent des cas spéciaux du modèle à deux niveaux (2.1) lorsque f_1 provient de la famille exponentielle naturelle. Par exemple, lorsque $y_{di} \in \{0,1\}$, le modèle linéaire mixte logistique habituel est obtenu en supposant que f_2 suit encore une $N(0,\sigma_u^2)$, et $f_1(y_{di} \mid u_d, \mathbf{x}_{di}, \boldsymbol{\beta}) = p_{di}^{y_{di}}(1-p_{di})^{1-y_{di}}$, pour

$$p_{di} = \frac{\exp(\mathbf{x}_{di}^t\boldsymbol{\beta} + u_d)}{1+\exp(\mathbf{x}_{di}^t\boldsymbol{\beta} + u_d)}, \quad i=1,\ldots,N_d, d=1,\ldots,D.$$

Dans le présent article, nous cherchons à estimer des indicateurs généraux, qui sont censés avoir la forme additive

$$\delta_d = \frac{1}{N_d}\sum_{i=1}^{N_d}\delta_{di}, \quad d=1,\ldots,D, \tag{2.3}$$

où $\delta_{di} = h(y_{di})$ est une fonction mesurable donnée de y_{di}. Un indicateur simple de la forme (2.3) est la moyenne du domaine $\delta_d = \overline{Y}_d$, où $h(y_{di}) = y_{di}$, pour toute i et tout d.

Un cas fréquent consiste à vouloir estimer la moyenne de domaine $\overline{Z}_d$ d'une variable d'intérêt Z prenant des valeurs, z_{di}, $i=1,\ldots,N_d$, mais le modèle à deux niveaux (2.2) est supposé pour une transformation monotone bijective croissante de ces valeurs, $y_{di} = g(z_{di})$, $i=1,\ldots,N_d$. Dans ce cas, $h(y_{di}) = g^{-1}(y_{di})$, pour toute i et tout d. Des indicateurs d'intérêt spéciaux, qui ont la forme (2.3), sont les indicateurs de pauvreté Foster-Greer-Thorbecke (FGT). Ils sont définis en fonction d'une variable du bien-être prenant des valeurs, z_{di}, et un seuil de pauvreté prédéfini z. Pour $\alpha \geq 0$, l'indicateur FGT est

$$F_{\alpha,d} = \frac{1}{N_d}\sum_{i=1}^{N_d}\left(\frac{z-z_{di}}{z}\right)^{\alpha} I(z_{di} < z), \quad d=1,\ldots,D. \tag{2.4}$$

Pour $\alpha = 0$, on obtient le taux de pauvreté du domaine d; pour $\alpha = 1$, on obtient l'écart de pauvreté du domaine. Si le modèle à deux niveaux (2.1) est supposé pour $y_{di} = g(z_{di})$, où $g(\cdot)$ est une transformation monotone bijective croissante, alors, dans ce cas,

$$\delta_{di} = h(y_{di}) = \left(\frac{z-g^{-1}(y_{di})}{z}\right)^{\alpha} I(y_{di} < g(z)), \quad i=1\ldots,N_d, d=1,\ldots,D.$$

Pour estimer δ_d, $d=1,\ldots,D$, un échantillon $s \subset U$ est censé être tiré de la population cible, où la variable cible et les variables auxiliaires sont observées. L'échantillon est censé être stratifié par domaine; c'est-à-dire, un échantillon s_d de taille n_d pour lequel $0 < n_d < N_d$ est censé être tiré indépendamment de chaque domaine U_d, et l'on désigne par $c_d = U_d - s_d$ le complément de l'échantillon de taille $N_d - n_d$,

$d = 1,\ldots,D.$ Alors, $s = s_1 \cup \cdots \cup s_D$ est l'échantillon total de taille $n = \sum_{d=1}^{D} n_d > D.$ Les données observées dans l'échantillon s sont

$$\{(y_{di}, \mathbf{x}_{di}), i \in s_d, d = 1,\ldots,D\}. \tag{2.5}$$

Nous supposons l'absence de biais de sélection, auquel cas les mesures d'échantillon suivent le même modèle à deux niveaux que celui défini dans (2.1).

3. Meilleure prédiction empirique

Soit $\mathbf{y}_d = (y_{d1},\ldots,y_{dN_d})^t,$ le vecteur des valeurs de la variable cible pour les unités du domaine d, $d = 1,\ldots,D,$ et $\mathbf{y} = (\mathbf{y}_1^t,\ldots,\mathbf{y}_D^t)^t,$ le vecteur de population. Chaque vecteur de domaine peut être subdivisé en deux sous-vecteurs, l'un correspondant aux unités échantillonnées et l'autre, aux unités non échantillonnées, par exemple $\mathbf{y}_d = (\mathbf{y}_{ds}^t, \mathbf{y}_{dc}^t)^t$, $d = 1,\ldots,D,$ le vecteur d'échantillon global est alors $\mathbf{y}_s = (\mathbf{y}_{1s}^t,\ldots, \mathbf{y}_{Ds}^t)^t.$ Le meilleur prédicteur de δ_d est le prédicteur $\tilde{\delta}_d$ qui minimise l'EQM fondée sur le modèle, définie par $\mathrm{EQM}_{\mathbf{y}}(\tilde{\delta}_d) = E_{\mathbf{y}}[(\tilde{\delta}_d - \delta_d)^2],$ et il est donné par

$$\tilde{\delta}_d^B = E_{\mathbf{y}_{dc}}(\delta_d \mid \mathbf{y}_{ds}) = \frac{1}{N_d}\left(\sum_{i\in s_d}\delta_{di} + \sum_{i\in c_d}\tilde{\delta}_{di}^B\right), \tag{3.1}$$

où $\tilde{\delta}_{di}^B = E[h(y_{di}) \mid \mathbf{y}_{ds}],\ i = 1,\ldots,N_d.$ Le meilleur prédicteur dépend de $\boldsymbol{\theta},$ qui est inconnu en pratique, c'est-à-dire $\tilde{\delta}_d^B = \tilde{\delta}_d^B(\boldsymbol{\theta}).$ En ajustant le modèle à deux niveaux (2.1) aux données d'enquête (2.5), on obtient un estimateur convergent $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ à mesure que $D \to \infty.$ Le prédicteur EB de δ_d est obtenu en insérant cet estimateur dans le meilleur prédicteur (3.1); c'est-à-dire qu'en prenant $\hat{\delta}_{di}^{EB} = \tilde{\delta}_{di}^B(\hat{\boldsymbol{\theta}}),\ i = 1,\ldots,N_d,$ on calcule

$$\hat{\delta}_d^{EB} = \tilde{\delta}_d^B(\hat{\boldsymbol{\theta}}) = \frac{1}{N_d}\left(\sum_{i\in s_d}\delta_{di} + \sum_{i\in c_d}\hat{\delta}_{di}^{EB}\right). \tag{3.2}$$

Selon le modèle de RLEE avec normalité des effets de domaine et des erreurs comme dans (2.2), la distribution de $\mathbf{y}_{dc} \mid \mathbf{y}_{ds}$ est également normale. Pour la moyenne de domaine $\delta_d = \overline{Y}_d,$ le prédicteur EB reposant sur l'estimateur par les moindres carrés pondérés de $\boldsymbol{\beta}$ est égal au meilleur prédicteur linéaire sans biais empirique (MPLSBE) d'$\overline{Y}_d$ obtenu par Battese et coll. (1988). Molina et Rao (2010) ont appliqué le même modèle à l'estimation d'indicateurs généraux δ_d; ils ont proposé une procédure de simulation de Monte Carlo (MC) pour approximer le prédicteur EB ainsi qu'une procédure bootstrap paramétrique pour estimer l'EQM selon ce modèle. Cho, Guadarrama-Sanz, Molina, Eideh et Berg (2024) ont étendu les procédures de Molina et Rao (2010) au cas d'une sélection informative.

Le meilleur prédicteur de δ_d donné dans (3.1) nécessite l'identification des unités de l'enquête s_d dans le fichier du recensement, ce qui est rarement réalisable. Une solution de rechange permettant d'éviter cette étape et étant de plus en plus populaire dans les applications pratiques est le meilleur prédicteur du recensement (CB pour *census best* en anglais), défini pour δ_d comme suit :

$$\tilde{\delta}_d^{CB} = \frac{1}{N_d}\sum_{i=1}^{N_d}\tilde{\delta}_{di}^{B}. \tag{3.3}$$

De même, en insérant un estimateur convergent $\hat{\boldsymbol{\theta}}$ dans le prédicteur CB (3.3), on obtient le prédicteur empirique CEB de δ_d, donné par

$$\hat{\delta}_d^{CEB} = \tilde{\delta}_d^{CB}(\hat{\boldsymbol{\theta}}) = \frac{1}{N_d}\sum_{i=1}^{N_d}\hat{\delta}_{di}^{EB}. \tag{3.4}$$

Lorsque la fraction de sondage $f_d = n_d / N_d$ est négligeable, le prédicteur CEB est approximativement égal au prédicteur EB de δ_d donné dans (3.2).

À l'exception du cas des moyennes de domaine, les prédicteurs EB et CEB présentés ci-dessus nécessitent un recensement contenant des microdonnées C sur les variables auxiliaires pour chaque unité de la population, lequel est censé être contemporain à l'enquête et mesuré sans erreur; il est désigné dans ce cas-ci par

$$C = \{\mathbf{x}_{di}; i = 1,\ldots,N_d, d = 1,\ldots,D\}. \tag{3.5}$$

Cependant, dans certains pays, aucun recensement n'est effectué désormais. Dans la plupart des pays, ce que l'on appelle le « recensement » correspond en réalité à une enquête de grande envergure, soit un cas que nous examinons à la section suivante. Dans le meilleur des cas, un recensement est réalisé tous les 5 ou 10 ans. Par conséquent, durant les années intercensitaires, les valeurs de recensement disponibles peuvent être complètement désuètes.

Pour simplifier l'exposé, nous illustrons le biais découlant de l'utilisation de données de recensement désuètes dans le cas où les chiffres de domaine N_d, $d = 1,\ldots,D,$ n'ont pas changé. Nous désignons alors l'ensemble des données de recensement désuètes par

$$C^o = \{\mathbf{x}_{di}^o; i = 1,\ldots,N_d, d = 1,\ldots,D\}. \tag{3.6}$$

Soit $\tilde{\delta}_d^{CBo} = \tilde{\delta}_d^{CBo}(\boldsymbol{\theta}),$ le meilleur prédicteur de δ_d obtenu à l'aide de C^o et $\hat{\delta}_d^{CEBo} = \tilde{\delta}_d^{CBo}(\hat{\boldsymbol{\theta}}),$ le prédicteur CEB correspondant. Il convient de mentionner que l'on obtient $\hat{\boldsymbol{\theta}}$ à partir des données d'enquête (2.5), les valeurs de recensement désuètes ne l'influencent donc pas. En revanche, toutes les prédictions d'unité $\hat{\delta}_{di}^{EBo}$, $i = 1,\ldots,N_d,$ et, par conséquent, le prédicteur CEB $\hat{\delta}_d^{CEBo}$ fondé sur $C^o,$ peuvent être fortement biaisés.

Dans le cas spécial $\delta_d = \overline{Y}_d,$ le prédicteur CB obtenu selon le modèle de RLEE à l'aide des données de recensement correctes C est donné par

$$\tilde{\overline{Y}}_d^{CB} = \overline{\mathbf{X}}_d^t\boldsymbol{\beta} + \gamma_d(\overline{y}_d - \overline{\mathbf{x}}_d^t\boldsymbol{\beta}), \tag{3.7}$$

où

$$\gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2 / n_d}, \quad \overline{y}_d = \frac{1}{n_d}\sum_{i\in s_d} y_{di}, \quad \overline{\mathbf{x}}_d = \frac{1}{n_d}\sum_{i\in s_d}\mathbf{x}_{di}.$$

À l'inverse, le prédicteur CB de $\bar{Y}_d$ fondé sur les données de recensement désuètes C^o est

$$\tilde{\bar{Y}}_d^{CBo} = (\bar{\mathbf{X}}_d^o)^t \boldsymbol{\beta} + \gamma_d (\bar{y}_d - \bar{\mathbf{x}}_d^t \boldsymbol{\beta}), \tag{3.8}$$

où $\bar{\mathbf{X}}_d^o = N_d^{-1} \sum_{i=1}^{N_d} \mathbf{x}_{di}^o$. Nous définissons le vecteur des changements subis par les vecteurs de données de recensement désuètes $\mathbf{x}_{di}^o$ par rapport aux vecteurs de données de recensement corrects (non disponibles) $\mathbf{x}_{di}$ de la façon suivante

$$\mathbf{b}_{di} = \mathbf{x}_{di} - \mathbf{x}_{di}^o, \quad i = 1,\ldots,N_d, d = 1,\ldots,D.$$

Les éléments positifs ou négatifs du vecteur $\mathbf{b}_{di}$ indiquent respectivement sous-estimation ou une surestimation des éléments correspondants de $\mathbf{x}_{di}$ lorsque l'on utilise ceux de $\mathbf{x}_{di}^o$. Le résultat suivant fournit le biais selon le modèle et l'EQM selon le modèle de $\tilde{\bar{Y}}_d^{CBo}$ découlant de l'utilisation de valeurs de recensement désuètes, en fonction du vecteur moyen des changements pour le domaine, $\bar{\mathbf{b}}_d = N_d^{-1} \sum_{i=1}^{N_d} \mathbf{b}_{di}$. La démonstration se trouve à la section A de l'annexe.

Proposition 1. *Selon le modèle de RLEE (2.2), il est vérifié que :*

(i) le biais de $\tilde{\bar{Y}}_d^{CBo}$ *donné dans (3.8) est* $B_{\mathrm{y}}(\tilde{\bar{Y}}_d^{CBo}) = -\bar{\mathbf{b}}_d^t \boldsymbol{\beta}$;

(ii) l'EQM de $\tilde{\bar{Y}}_d^{CBo}$ *donnée dans (3.8) est* $EQM_{\mathrm{y}}(\tilde{\bar{Y}}_d^{CBo}) = EQM_{\mathrm{y}}(\tilde{\bar{Y}}_d^{CB}) + (\bar{\mathbf{b}}_d^t \boldsymbol{\beta})^2$, *où*

$$EQM_{\mathrm{y}}(\tilde{\bar{Y}}_d^{CB}) = \gamma_d \frac{\sigma_e^2}{n_d} + (1 - 2\gamma_d) \frac{\sigma_e^2}{N_d}.$$

Le résultat (ii) indique clairement que $\mathrm{EQM}_{\mathrm{y}}(\tilde{\bar{Y}}_d^{CBo}) \geqslant \mathrm{EQM}_{\mathrm{y}}(\tilde{\bar{Y}}_d^{CB})$. Ce résultat montre que, plus généralement, $\hat{\delta}_d^{CEBo}$ est biaisé, ce qui a nécessairement une incidence sur son EQM et, selon l'ampleur des erreurs $\mathbf{b}_{di}$, $i = 1,\ldots,N_d$, et la taille de l'échantillon de domaine n_d, le prédicteur CB $\tilde{\bar{Y}}_d^{CBo}$ pourrait être encore moins efficace qu'un estimateur direct de δ_d fondé sur les données d'enquête (2.5), lequel permet d'éviter de formuler des hypothèses de modèle. En outre, lorsque l'EQM est estimée à l'aide de C^o, on sous-estime cette EQM, ce qui conduit à des gains d'efficacité trompeurs par rapport aux estimateurs directs ou à d'autres estimateurs. À la section 4, nous proposons un estimateur d'un indicateur additif général δ_d qui permet d'éviter l'utilisation de données de recensement désuètes et, à la section 5, nous proposons un estimateur de l'EQM totale de l'estimateur proposé.

4. Meilleur prédicteur empirique d'enquête

Dans la présente section, nous considérons le cas où les microdonnées de recensement sur les variables auxiliaires (3.5) ne sont pas disponibles ou ne sont pas exploitables. Nous supposons plutôt que les mêmes variables auxiliaires sont observées dans un échantillon d'enquête de substitution s'_d, de taille d'échantillon n'_d satisfaisant $n_d \leq n'_d \leq N_d$, pour $d = 1,\ldots,D$. Soit alors $s' = s'_1 \cup \cdots \cup s'_D$, l'échantillon global de cette

enquête de plus grande portée de taille $n' = \sum_{d=1}^{D} n'_d$. Nous désignons les données correspondant à s' de l'enquête de plus grande portée par

$$\{\mathbf{x}_{di}, i \in s'_d, d = 1, \ldots, D\}. \tag{4.1}$$

Dans ce cas-ci, nous remplaçons les données auxiliaires de recensement désuètes C^o par les données auxiliaires (4.1) issues de l'échantillon d'enquête s'. Soit $\pi'_{di} > 0$, la probabilité d'inclusion de l'unité i dans s'_d, et $w'_{di} = (\pi'_{di})^{-1}$, le poids d'enquête correspondant. Au lieu du prédicteur CB $\tilde{\delta}_d^{CBo}$ fondé sur le recensement désuet C^o, nous proposons d'utiliser le meilleur prédicteur d'enquête (SB pour *survey best* en anglais) de δ_d, défini comme suit :

$$\tilde{\delta}_d^{SB} = \frac{1}{w'_{d\cdot}} \sum_{i \in s'_d} w'_{di} \tilde{\delta}_{di}^{B}, \tag{4.2}$$

où $w'_{d\cdot} = \sum_{i \in s'_d} w'_{di}$. Une fois de plus, dans la pratique, $\boldsymbol{\theta}$ est inconnu et, par conséquent, le prédicteur SB en dépend, c'est-à-dire $\tilde{\delta}_d^{SB} = \tilde{\delta}_d^{SB}(\boldsymbol{\theta})$. Soit $\hat{\boldsymbol{\theta}}$, un estimateur convergent de $\boldsymbol{\theta}$ obtenu en ajustant le modèle à deux niveaux (2.1) aux données (2.5) issues de l'enquête s. En remplaçant $\boldsymbol{\theta}$ par $\hat{\boldsymbol{\theta}}$ dans le prédicteur SB $\boldsymbol{\theta}$, on obtient le meilleur prédicteur empirique issu de l'enquête (SEB pour *survey empirical best* en anglais), à savoir

$$\hat{\delta}_d^{SEB} = \tilde{\delta}_d^{SB}(\hat{\boldsymbol{\theta}}) = \frac{1}{w'_{d\cdot}} \sum_{i \in s'_d} w'_{di} \hat{\delta}_{di}^{EB}. \tag{4.3}$$

La taille de l'échantillon de domaine n'_d est censée être suffisamment grande pour que $\hat{\delta}_d^{SEB}$ puisse constituer un bon estimateur de $\hat{\delta}_d^{CEB}$ fondé sur le recensement correct C. Dans ce qui suit, nous présentons une procédure simple permettant de déterminer si n'_d est suffisamment grande en pratique.

Lorsque n'_d augmente, le prédicteur SEB $\hat{\delta}_d^{SEB}$ est un estimateur convergent selon le plan de $\hat{\delta}_d^{CEB}$, lequel est approximativement égal au meilleur prédicteur lorsque la fraction de sondage f_d est négligeable. De plus, si $w'_{d\cdot} = N_d$, ce qui se produit dans le cadre d'un échantillonnage autopondéré au sein des domaines, le prédicteur SEB est sans biais pour le prédicteur CEB selon le plan d'échantillonnage de s', c'est-à-dire

$$E_{s'}(\hat{\delta}_d^{SEB}) = \hat{\delta}_d^{CEB}. \tag{4.4}$$

Même si le plan d'échantillonnage ne satisfait pas $w'_{d\cdot} = N_d$, le biais de ratio de $\hat{\delta}_d^{SEB}$ est négligeable pour de grandes valeurs de n'_d. Étant donné que nous considérons des enquêtes pour lesquelles n'_d est grande, (4.4) est approximativement vérifiée.

Dans la pratique, on peut souhaiter vérifier si une enquête donnée contenant l'information auxiliaire recherchée est satisfaisante. À cette fin, nous déterminons la taille d'échantillon minimale n_d^*, de sorte que l'erreur relative de $\hat{\delta}_d^{SEB}$ en tant qu'estimateur de $\hat{\delta}_d^{CEB}$ soit inférieure à une valeur prédéfinie ϵ_0, avec une probabilité élevée donnée $1-\alpha$, pour $\alpha \in (0,1)$. En appliquant des résultats standards dans le cadre d'un échantillonnage aléatoire simple sans remise (EASSR) et en supposant la normalité du prédicteur SEB, n_d^* se réduit à

$$n_d^* = \frac{k_d N_d}{1+k_d}, \quad k_d = z_{\alpha/2}^2 \frac{cv_d^2}{\epsilon_0^2}, \tag{4.5}$$

où $z_{\alpha/2}$ est le point critique $\alpha/2$ de la distribution normale standard et cv_d est le coefficient de variation de $\{\hat{\delta}_{di}^{EB}; i=1,\ldots,N_d\}$. On peut estimer cv_d à l'aide des données d'enquête $\{\hat{\delta}_{di}^{EB}; i \in s_d'\}$; par exemple, on utilise

$$\widetilde{cv}_d = \frac{\sqrt{\frac{1}{n_d'-1}\sum_{i\in s_d'}\left(\hat{\delta}_{di}^{EB} - \frac{1}{n_d'}\sum_{i\in s_d'}\hat{\delta}_{di}^{EB}\right)^2}}{|\hat{\delta}_d^{SEB}|}$$

et ensuite on prend $\tilde{n}_d^* = \tilde{k}_d N_d/(1+\tilde{k}_d)$, pour $\tilde{k}_d = z_{\alpha/2}^2 \widetilde{cv}_d^2/\epsilon_0^2$. Pour des plans complexes, on peut également intégrer l'effet de plan de $\hat{\delta}_d^{SEB}$ obtenu à partir de s' selon Kish (1965), en prenant la taille d'échantillon souhaitée comme étant $\hat{n}_d^* = \tilde{n}_d^* \, \mathrm{deff}_{s'}(\hat{\delta}_d^{SEB})$, où $\mathrm{deff}_{s'}(\hat{\delta}_d^{SEB}) = \hat{V}_{s'}(\hat{\delta}_d^{SEB} \mid \mathbf{y}) \big/ \hat{V}_{SRS}(\hat{\delta}_d^{SEB} \mid \mathbf{y})$.

Si la taille d'échantillon du domaine observé est $n_d' > \hat{n}_d^*$, alors $\hat{\delta}_d^{SEB}$ estime $\hat{\delta}_d^{CEB}$ selon la précision et la probabilité souhaitées. Pour les domaines où $n_d < n_d' < \hat{n}_d^*$, le prédicteur SEB peut être moins précis que ce que l'on souhaite. Toutefois, ces domaines peuvent néanmoins être pris en compte dans l'analyse, avec prudence. Si $n_d' < n_d$, on établit que $s_d' = s_d$ et l'on utilise les données auxiliaires issues de s_d.

L'expérience de simulation de la section 6 montre que le prédicteur SEB est quasi identique au prédicteur EB lorsque l'enquête de grande envergure s' couvre adéquatement l'ensemble des petits domaines d'intérêt, et qu'il permet d'obtenir de bien meilleurs résultats que l'estimateur direct habituel, même lorsque seule la petite enquête s est disponible; voir la section C de l'annexe.

5. Estimation de l'erreur quadratique moyenne totale

L'incertitude du prédicteur EB est généralement évaluée à l'aide de l'EQM selon le modèle (appelée EQM du modèle), car les estimateurs de l'EQM fondés sur le plan disponibles dans la littérature sont souvent très instables (Rao, Rubin-Bleuer et Estevao, 2018; Stefan et Hidiroglou, 2021). La stabilité des estimateurs de l'EQM du modèle (Molina et Strzalkowska-Kominiak, 2019) le rend comme constituant les mesures d'incertitude privilégiées pour accompagner les estimateurs sur petits domaines fondés sur un modèle. Dans le présent article, nous considérons le cas où le recensement correct C des données auxiliaires n'est pas disponible. Par conséquent, les estimateurs habituels de l'EQM utilisés pour les prédicteurs EB ou CEB, tels que ceux fondés sur des procédures bootstrap (Molina et Rao, 2010), ne peuvent pas être appliqués. Cependant, l'utilisation de C^o mènerait clairement à une sous-estimation de l'EQM réelle.

Par ailleurs, l'incertitude du prédicteur SEB (4.3) est clairement influencée par le plan d'échantillonnage et par les tailles d'échantillon par domaine de l'enquête de grande envergure n_d', $d=1,\ldots,D$. Il est donc raisonnable de tenir compte des deux sources d'incertitude, à savoir l'incertitude liée au modèle et l'incertitude liée au plan attribuable à s'. En conséquence, nous nous intéressons à l'EQM totale du prédicteur SEB, définie comme $\mathrm{EQM}_T(\hat{\delta}_d^{SEB}) = E_{(\mathbf{y},s')}[(\hat{\delta}_d^{SEB} - \delta_d)^2]$.

Définissons le vecteur des estimateurs de Hájek du vecteur moyen $\overline{\mathbf{X}}_d = N_d^{-1}\sum_{i=1}^{N_d}\mathbf{x}_{di}$ reposant sur l'échantillon s' par $\tilde{\overline{\mathbf{X}}}_{ds'} = (w'_{d.})^{-1}\sum_{i\in s'_d} w'_{di}\,\mathbf{x}_{di}$. Lors de l'estimation de la moyenne de domaine $\delta_d = \overline{Y}_d$ selon le modèle de RLEE (2.2), le prédicteur SB $\tilde{\overline{Y}}_d^{SB}$ est donné par

$$\tilde{\overline{Y}}_d^{SB} \;=\; \tilde{\overline{\mathbf{X}}}_{ds'}^t\boldsymbol{\beta} + \gamma_d(\overline{y}_d - \overline{\mathbf{x}}_d^t\boldsymbol{\beta}). \tag{5.1}$$

Le résultat qui s'en suit fournit le biais total du prédicteur SB (5.1), défini comme $B_T(\tilde{\overline{Y}}_d^{SB}) = E_{(\mathbf{y},s')}(\tilde{\overline{Y}}_d^{SB} - \overline{Y}_d)$, ainsi que l'EQM totale. La démonstration se trouve à la section A de l'annexe.

Proposition 2. *Selon le modèle de RLEE (2.2), il est vérifié que :*

(i) le biais total de $\tilde{\overline{Y}}_d^{SB}$ *donné dans (5.1) est* $B_T(\tilde{\overline{Y}}_d^{SB}) = \boldsymbol{\beta}^t B_{s'}(\tilde{\overline{\mathbf{X}}}_{ds'})$, *où* $B_{s'}(\tilde{\overline{\mathbf{X}}}_{ds'}) = E_{s'}(\tilde{\overline{\mathbf{X}}}_{ds'}) - \overline{\mathbf{X}}_d$ *est le biais de* $\tilde{\overline{\mathbf{X}}}_{ds'}$ *selon le plan d'échantillonnage de* s';

(ii) l'EQM totale de $\tilde{\overline{Y}}_d^{SB}$ *donné dans (5.1) est*

$$EQM_T(\tilde{\overline{Y}}_d^{SB}) \;=\; EQM_{\mathbf{y}}(\tilde{\overline{Y}}_d^{CB}) + \boldsymbol{\beta}^t E_{s'}[(\tilde{\overline{\mathbf{X}}}_{ds'} - \overline{\mathbf{X}}_d)(\tilde{\overline{\mathbf{X}}}_{ds'} - \overline{\mathbf{X}}_d)^t]\boldsymbol{\beta}.$$

Si $w'_{d.} = N_d$ est vérifié, alors $B_{s'}(\tilde{\overline{\mathbf{X}}}_{ds'}) = \mathbf{0}_p$, ce qui conduit à un biais total nul pour $\tilde{\overline{Y}}_d^{SB}$. De manière générale, puisque la taille d'échantillon n'_d est censée être grande, le biais de ratio de l'estimateur de Hájek $\tilde{\overline{\mathbf{X}}}_{ds'}$ est négligeable en raison de la convergence de l'estimateur de Hájek, dans des conditions générales. Par conséquent, le biais total de $\tilde{\overline{Y}}_d^{SB}$ pour $\overline{Y}_d$ sera négligeable.

En ce qui concerne l'EQM totale de $\tilde{\overline{Y}}_d^{SB}$, $\mathrm{EQM}_T(\tilde{\overline{Y}}_d^{SB})$, celle-ci se compose de deux termes. Le premier terme correspond à l'EQM du modèle du prédicteur CB quasi optimal $\tilde{\overline{Y}}_d^{CB}$ ayant une $\overline{\mathbf{X}}_d$ connue, qui reflète l'erreur de modèle dans y_{di} ainsi que les effets de domaine u_d. Le second terme est entièrement attribuable à l'erreur d'estimation de $\overline{\mathbf{X}}_d$ à partir de l'échantillon auxiliaire s', qui disparaît lorsque $\tilde{\overline{\mathbf{X}}}_{ds'} = \overline{\mathbf{X}}_d$ (information auxiliaire parfaite). La décomposition montre que toute inexactitude dans $\tilde{\overline{\mathbf{X}}}_{ds'}$ augmente additivement l'EQM totale. Cependant, pour des valeurs élevées de n'_d, par la convergence de l'estimateur de Hájek $\tilde{\overline{\mathbf{X}}}_{ds'}$ vers $\overline{\mathbf{X}}_d$ lorsque n'_d augmente, l'EQM totale se réduit à l'EQM du modèle du prédicteur CB $\tilde{\overline{Y}}_d^{CB}$.

La proposition 2 montre la manière dont l'incertitude liée à l'utilisation de l'information auxiliaire issue de s' a une incidence sur l'EQM totale de $\tilde{\delta}_d^{SB}$ dans le cas de l'estimation des moyennes de domaine $\delta_d = \overline{Y}_d$. Pour des indicateurs plus généraux δ_d, nous obtenons une décomposition différente de l'EQM totale de $\hat{\delta}_d^{SEB}$, laquelle permet de déterminer un estimateur approprié fondé sur s'. Il convient de mentionner que le prédicteur SEB $\hat{\delta}_d^{SEB}$ est en réalité le prédicteur EB de

$$\delta'_d \;=\; \frac{1}{w'_{d.}}\sum_{i\in s'_d} w'_{di}\,\delta_{di}. \tag{5.2}$$

Cependant, (5.2) peut différer considérablement de δ_d si n'_d n'est pas suffisamment grande. Dans le cas $w'_{d.} = N_d$, le résultat suivant décompose l'EQM totale de $\hat{\delta}_d^{SEB}$ en l'EQM totale de $\hat{\delta}_d^{SEB}$ en tant que prédicteur EB de δ'_d ainsi que d'autres termes qui reflètent l'incertitude de δ'_d en tant qu'estimateur de δ_d fondé sur s'. La démonstration se trouve à la section A de l'annexe.

Proposition 3. *Soit* $\hat{\delta}_d^{SEB}$, *le prédicteur de* δ_d *donné dans (4.3). Si* $w'_{d\cdot} = N_d$, *alors l'EQM totale de* $\hat{\delta}_d^{SEB}$ *est donnée par*

$$EQM_T(\hat{\delta}_d^{SEB}) = E_{s'}\left\{E_{\mathbf{y}}\left[(\hat{\delta}_d^{SEB} - \delta'_d)^2 \mid s'\right]\right\} + E_{\mathbf{y}}\left[2Cov_{s'}(\hat{\delta}_d^{SEB}, \delta'_d \mid \mathbf{y}) - V_{s'}(\delta'_d \mid \mathbf{y})\right]. \tag{5.3}$$

Selon la proposition 3, l'EQM totale de $\hat{\delta}_d^{SEB}$ (5.3) se compose de trois termes. Le premier terme correspond à l'incertitude selon le modèle de $\hat{\delta}_d^{SEB}$ en tant que prédicteur de δ'_d (l'estimateur de Horvitz-Thompson [HT] fondé sur s'), dont la moyenne a été calculée sur les échantillons possibles s'. Le deuxième terme ajuste l'EQM totale pour la corrélation entre le prédicteur $\hat{\delta}_d^{SEB}$ et l'estimateur de HT fondé sur s', δ'_d. Si $\hat{\delta}_d^{SEB}$ est positivement corrélé à δ'_d, ce terme augmente l'EQM totale; s'il est négativement corrélé, il la diminue. Le troisième terme est une soustraction de la variance selon le plan de δ'_d, dont la moyenne est calculée sur le modèle. Ce terme apparaît parce que δ'_d est lui-même un estimateur aléatoire de δ_d; il est soustrait afin d'éviter une double comptabilisation de sa variabilité. De manière générale, (5.3) montre que l'erreur totale n'est pas simplement la somme de l'EQM du modèle et de la variance par rapport au plan d'échantillonnage de δ'_d : l'interaction entre $\hat{\delta}_d^{SEB}$ et δ'_d par l'intermédiaire de leur covariance est également concernée.

Selon la proposition 3, un estimateur de l'EQM totale du prédicteur SEB peut être obtenu en supprimant l'espérance extérieure dans le premier terme de (5.3), en remplaçant $\mathrm{Cov}_{s'}(\hat{\delta}_d^{SEB}, \delta'_d \mid \mathbf{y})$ et $V_{s'}(\delta'_d \mid \mathbf{y})$ par leurs estimateurs correspondants fondés sur le plan construits à partir de s', puis en estimant l'espérance par rapport au modèle à partir de s. Cette approche conduit à l'estimateur de l'EQM totale

$$\mathrm{eqm}_T(\hat{\delta}_d^{SEB}) = \hat{E}_{\mathbf{y}}\left[(\hat{\delta}_d^{SEB} - \delta'_d)^2 \mid s'\right] + \hat{E}_{\mathbf{y}}\left[2\widehat{\mathrm{Cov}}_{s'}(\hat{\delta}_d^{SEB}, \delta'_d \mid \mathbf{y}) - \hat{V}_{s'}(\delta'_d \mid \mathbf{y})\right]. \tag{5.4}$$

Pour estimer la covariance et la variance requises, si $w'_{d\cdot} = N_d$, on peut utiliser respectivement

$$\widehat{\mathrm{Cov}}_{s'}(\hat{\delta}_d^{SEB}, \delta'_d \mid \mathbf{y}) = \frac{1}{N_d^2}\sum_{i\in s'_d}\sum_{j\in s'_d}\frac{\pi'_{dij} - \pi'_{di}\pi'_{dj}}{\pi'_{dij}}\frac{\hat{\delta}_{di}^{EB}\delta_{dj}}{\pi'_{di}\pi'_{dj}},$$

$$\hat{V}_{s'}(\delta'_d \mid \mathbf{y}) = \frac{1}{N_d^2}\sum_{i\in s'_d}\sum_{j\in s'_d}\frac{\pi'_{dij} - \pi'_{di}\pi'_{dj}}{\pi'_{dij}}\frac{\delta_{di}\delta_{dj}}{\pi'_{di}\pi'_{dj}},$$

où π'_{dij} désigne la probabilité d'inclusion de deuxième ordre des unités i et j dans s'_d; voir par exemple Särndal, Swensson et Wretman (1992, page 170).

Selon certains modèles à deux niveaux et certains indicateurs cibles δ_d particuliers, des estimateurs analytiques de la forme (5.4) peuvent être obtenus, par exemple, à l'aide d'arguments asymptotiques. Plus précisément, si l'on estime les moyennes de domaine $\delta_d = \overline{Y}_d = N_d^{-1}\sum_{i=1}^{N_d} y_{di}$, on a $\delta'_d = \overline{Y}'_d = N_d^{-1}\sum_{i\in s'_d} w'_{di}\, y_{di}$ et $\hat{\delta}_d^{SEB} = \hat{\overline{Y}}_d^{SEB} = N_d^{-1}\sum_{i\in s'_d} w'_{di}\,\hat{y}_{di}^{EB}$. Dans ce cas, le premier terme du côté droit de (5.4) constitue un estimateur de

$$E_{\mathbf{y}}\left[(\hat{\overline{Y}}_d^{SEB} - \overline{Y}'_d)^2 \mid s'\right] = \frac{1}{N_d^2}\sum_{i\in s'_d}\sum_{j\in s'_d} w'_{di}\, w'_{dj}\, E_{\mathbf{y}}\left[(\hat{y}_{di}^{EB} - y_{di})\,(\hat{y}_{dj}^{EB} - y_{dj})\right].$$

Pour les termes restants de (5.4), on obtient

$$E_{\mathbf{y}}\left[\widehat{\mathrm{Cov}}_{s'}(\hat{\bar{Y}}_d^{SEB}, \bar{Y}'_d \mid \mathbf{y})\right] = \frac{1}{N_d^2}\sum_{i\in s'_d}\sum_{j\in s'_d}\frac{\pi'_{dij}-\pi'_{di}\pi'_{dj}}{\pi'_{dij}}\frac{E_{\mathbf{y}}[\hat{y}_{di}^{EB}y_{dj}]}{\pi'_{di}\pi'_{dj}},$$

$$E_{\mathbf{y}}\left[\hat{V}_{s'}(\bar{Y}'_d \mid \mathbf{y})\right] = \frac{1}{N_d^2}\sum_{i\in s'_d}\sum_{j\in s'_d}\frac{\pi'_{dij}-\pi'_{di}\pi'_{dj}}{\pi'_{dij}}\frac{E_{\mathbf{y}}[y_{di}y_{dj}]}{\pi'_{di}\pi'_{dj}},$$

où, pour le modèle de RLEE (2.2), on a

$$E_{\mathbf{y}}[y_{di}y_{dj}] = \sigma_u^2 + \sigma_e^2 + \boldsymbol{\beta}^t\mathbf{x}_{di}\mathbf{x}_{dj}^t\boldsymbol{\beta}, \quad i,j \in s'_d.$$

Cette espérance peut être estimée en remplaçant les paramètres inconnus $\boldsymbol{\beta}$, σ_u^2 et σ_e^2 par les estimateurs correspondants obtenus en ajustant le modèle de RLEE (2.2) aux données de l'échantillon s. Les espérances $E_{\mathbf{y}}[\hat{y}_{di}^{EB}y_{dj}]$ et $E_{\mathbf{y}}[(\hat{y}_{di}^{EB}-y_{di})(\hat{y}_{dj}^{EB}-y_{dj})]$ peuvent être approximées analytiquement pour un grand nombre de domaines D puis estimées ultérieurement, en utilisant les mêmes arguments que dans Prasad et Rao (1990), Das, Jiang et Rao (2004) ou Baíllo et Molina (2009).

Pour des indicateurs additifs généraux δ_d, nous proposons une méthode bootstrap paramétrique, semblable à celle utilisée par Molina et Rao (2010), applicable (ou applicable par extension) à des modèles très généraux et garantissant un résultat strictement positif.

Il convient de mentionner que, lorsque n'_d est suffisamment grande, l'erreur attribuable à s' peut être considérée comme négligeable. Dans ce cas, le premier terme du côté droit de (5.4) utilisé seul, qui est strictement positif, peut être considéré comme estimateur de l'EQM totale. Néanmoins, en pratique, pour certains domaines d, n'_d peut ne pas être aussi grande que l'on veut, ce qui entraîne un biais non négligeable de cet estimateur naïf. Dans les expériences de simulation présentées à la section 6, où n'_d est modérée, le deuxième terme du côté droit de (5.4) devient nécessaire. Nous considérons une méthode bootstrap paramétrique qui fournit à la fois l'estimateur naïf de l'EQM mentionné et un estimateur corrigé de l'EQM totale fondé sur (5.4).

Bootstrap paramétrique pour l'estimation de l'EQM totale du prédicteur SEB

1. *Ajustement du modèle :* Ajuster le modèle à deux niveaux (2.1) aux données d'enquête au niveau des unités (2.5), afin d'obtenir un estimateur convergent $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1^t, \hat{\boldsymbol{\theta}}_2^t)^t$ de $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^t, \boldsymbol{\theta}_2^t)^t$ lorsque le nombre de domaines $D \to \infty$.
2. *Génération des valeurs bootstrap pour le grand échantillon :* En utilisant $\hat{\boldsymbol{\theta}}$ de l'étape 1 comme valeur « réelle » de $\boldsymbol{\theta}$, générer des effets de domaine bootstrap $u_d^* \overset{\text{iid}}{\sim} f_2(u_d; \hat{\boldsymbol{\theta}}_2)$, $d = 1,\ldots,D$. Maintenant, à l'aide de (4.1) et de u_d^*, $d = 1,\ldots,D$, générer des valeurs de réponse bootstrap pour les unités de s' comme suit :

$$y_{di}'^* \overset{\text{ind}}{\sim} f_1(y_{di} \mid u_d^*, \mathbf{x}_{di}, \hat{\boldsymbol{\theta}}_1), \quad i \in s'_d, d = 1,\ldots,D.$$

   Calculer ensuite $\delta_d'^* = (w'_{d\cdot})^{-1}\sum_{i\in s'_d} w'_{di}\,\delta_{di}^*$, pour $\delta_{di}^* = h(y_{di}'^*)$, $i \in s'_d$, ainsi que la version bootstrap de l'estimateur de variance fondé sur le plan, $\hat{V}_{s'}^* = \hat{V}_{s'}(\delta_d'^*)$.
3. *Génération des valeurs bootstrap pour le petit échantillon :* En utilisant les mêmes effets de domaine bootstrap u_d^* que ceux de l'étape 2, générer

$$y_{di}^{*} \overset{\text{ind}}{\sim} f_1(y_{di} \mid u_d^{*}, \mathbf{x}_{di}, \hat{\boldsymbol{\theta}}_1), \quad i \in s_d, d = 1, \ldots, D.$$

4. *Ajustement du modèle bootstrap et estimation :* Ajuster le modèle à deux niveaux (2.1) aux données de l'échantillon bootstrap $\{y_{di}^{*}, i \in s_d, d = 1, \ldots, D\}$ de l'étape 3 et obtenir les prédicteurs EB bootstrap $\hat{\delta}_{di}^{EB*}$ de $\delta_{di}^{*}, i \in s_d'$. Calculer ensuite le prédicteur SEB bootstrap

$$\hat{\delta}_d^{SEB*} = \frac{1}{w_{d.}'} \sum_{i \in s_d'} w_{di}' \hat{\delta}_{di}^{EB*}.$$

Calculer la version bootstrap de l'estimateur de la covariance fondé sur le plan comme suit :

$$\widehat{\text{Cov}}_{s'}^{*} = \widehat{\text{Cov}}_{s'}(\hat{\delta}_d^{SEB*}, \delta_d'^{*}).$$

5. *Estimateurs bootstrap de l'EQM totale :* Nous considérons deux estimateurs bootstrap différents de l'EQM totale. Le premier est un estimateur naïf donné par

$$\text{EQM}_{T,na}^{*}(\hat{\delta}_d^{SEB}) = E_{\mathbf{y}*}\left[(\hat{\delta}_d^{SEB*} - \delta_d'^{*})^2\right]. \tag{5.5}$$

Corriger ensuite (5.5) pour tenir compte de l'erreur de $\delta_d'^{*}$ en tant qu'estimateur de δ_d^{*} fondé sur s', en prenant

$$\text{EQM}_{T,c}^{*}(\hat{\delta}_d^{SEB}) = \text{EQM}_{T,na}^{*}(\hat{\delta}_d^{SEB}) + E_{\mathbf{y}*}\left[2\widehat{\text{Cov}}_{s'}^{*} - \hat{V}_{s'}^{*}\right]. \tag{5.6}$$

Un inconvénient de l'estimateur corrigé de l'EQM (5.6) est qu'il peut prendre une valeur négative. Afin d'éviter cette situation, nous définissons l'estimateur corrigé positif de l'EQM comme suit :

$$\text{EQM}_{T,cp}(\hat{\delta}_d^{SEB}) = \begin{cases} \text{EQM}_{T,c}(\hat{\delta}_d^{SEB}), \text{ si } \text{EQM}_{T,c}(\hat{\delta}_d^{SEB}) \geq 0, \\ \text{EQM}_{T,na}(\hat{\delta}_d^{SEB}), \text{ si } \text{EQM}_{T,c}(\hat{\delta}_d^{SEB}) < 0. \end{cases} \tag{5.7}$$

Il convient de mentionner que l'espérance $E_{\mathbf{y}*}$ dans ces estimateurs bootstrap est prise par rapport à la distribution du vecteur de population bootstrap $\mathbf{y}*$ utilisé aux étapes 2 et 3 étant donné s' et s, et étant donné les données auxiliaires (4.1) issues de la grande enquête et les données de l'échantillon (2.5).

En pratique, on approxime (5.5) et (5.6) par simulation de Monte Carlo (MC) : répéter les étapes 2 à 4 pour $b = 1, \ldots, B$, avec une grande valeur de B. Soit $\hat{\delta}_d^{SEB*(b)}$, $\delta_d'^{*(b)}$, $\widehat{\text{Cov}}_{s'}^{*(b)}$ et $\hat{V}_{s'}^{*(b)}$, les résultats de la réplique b. L'approximation de MC de l'estimateur naïf est alors

$$\text{eqm}_{T,na}(\hat{\delta}_d^{SEB}) = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\delta}_d^{SEB*(b)} - \delta_d'^{*(b)}\right)^2. \tag{5.8}$$

De même, l'approximation de MC de l'estimateur corrigé s'obtient comme suit :

$$\text{eqm}_{T,c}(\hat{\delta}_d^{SEB}) = \text{eqm}_{T,na}(\hat{\delta}_d^{SEB}) + \frac{1}{B} \sum_{b=1}^{B} \left(2\widehat{\text{Cov}}_{s'}^{*(b)} - \hat{V}_{s'}^{*(b)}\right). \tag{5.9}$$

Enfin, on prend l'approximation de MC de l'estimateur corrigé positif de l'EQM

$$\text{eqm}_{T,cp}(\hat{\delta}_d^{SEB}) = \begin{cases} \text{eqm}_{T,c}(\hat{\delta}_d^{SEB}), \text{ si } \text{eqm}_{T,c}(\hat{\delta}_d^{SEB}) \geq 0, \\ \text{eqm}_{T,na}(\hat{\delta}_d^{SEB}), \text{ autrement.} \end{cases} \tag{5.10}$$

Dans le cas des proportions de petits domaines sous un modèle linéaire mixte logistique, Sen et Lahiri (2025) ont proposé (5.8) comme estimateur de l'EQM du modèle, $E_{\mathbf{y}}\left[(\hat{\delta}_d^{SEB}-\delta_d)^2\right]$. Les études par simulation de la section 6 illustrent le biais potentiel de (5.8) en tant qu'estimateur de l'EQM totale lorsque n'_d n'est pas très grande ainsi que la manière dont l'estimateur corrigé positif (5.10) réduit ce biais.

Enfin, si $\hat{\boldsymbol{\theta}}$ est convergent selon le modèle lorsque $D\to\infty$, et si $\widehat{\mathrm{Cov}}_{s'}(\hat{\delta}_d^{SEB},\delta'_d)$ et $\hat{V}_{s'}(\delta'_d)$ sont convergents selon le plan pour $\mathrm{Cov}_{s'}(\hat{\delta}_d^{SEB},\delta'_d)$ et $V_{s'}(\delta'_d)$, respectivement, lorsque $n'_d\to\infty$, alors $\mathrm{eqm}_{T,c}(\hat{\delta}_d^{SEB})$ de (5.9) doit être convergent pour la valeur réelle $\mathrm{EQM}_T(\hat{\delta}_d^{SEB})$ dans la distribution conjointe de $(\mathbf{y},s')$ lorsque $B\to\infty$, $D\to\infty$ et $n'_d\to\infty$.

6. Expériences de simulation

La présente section décrit une expérience de simulation de MC conçue pour comparer les propriétés de plusieurs prédicteurs des taux et écarts de pauvreté par domaine. Plus précisément, nous comparons les prédicteurs suivants pour chaque indicateur cible $\delta_d\in\{F_{0,d},F_{1,d}\}$:

1. Estimateur direct $\hat{\delta}_d^{DIR}=(w_{d.})^{-1}\sum_{i\in s_d}w_{di}\delta_{di}$;
2. MPLSBE fondé sur le modèle de FH utilisant des données auxiliaires agrégées $\overline{\mathbf{X}}_d^o=N_d^{-1}\sum_{i=1}^{N_d}\mathbf{x}_{di}^o$ provenant du recensement désuet C^o, désigné par $\hat{\delta}_d^{FH}$;
3. Prédicteur EB fondé sur le modèle de RLEE utilisant le recensement désuet C^o, $\hat{\delta}_d^{EBo}$;
4. Prédicteur SEB fondé sur le modèle de RLEE, $\hat{\delta}_d^{SEB}$.

Nous considérons un scénario semblable à celui de Molina et Rao (2010), soit une population composée de $D=80$ domaines et $N_d=2\,500$ unités par domaine $d=1\ldots,D.$ Nous considérons deux variables auxiliaires continues, x_q, $q=1,2,$ dont les valeurs sont générées comme suit : $x_{q,di}\sim\mathrm{Gamma}(k_{qd},t_q)$ avec $k_{1d}=1+5d/D$, $k_{2d}=2$, $t_1=2$, $t_2=3$, $i=1,\ldots,N_d$, $d=1,\ldots,D.$ Le recensement correct C est alors (3.5), avec $\mathbf{x}_{di}=(1,x_{1,di},x_{2,di})^t$, $i=1,\ldots,N_d$, $d=1,\ldots,D.$ Nous prenons les paramètres du modèle $\boldsymbol{\beta}=$ (3, 0,03, -0,04)^t, $\sigma_u^2=0{,}15^2$ et $\sigma_e^2=0{,}5^2$.

L'échantillon s est tiré indépendamment dans chaque domaine d selon un EAS, et les tailles d'échantillon par domaine sont $n_d=25$ pour $1\le d\le 30$, $n_d=50$ pour $31\le d\le 60$ et $n_d=75$ pour $61\le d\le 80$. Le grand échantillon s' est tiré indépendamment de s selon le même plan d'échantillonnage, mais a des tailles d'échantillon par domaine $n'_d=10n_d$, $d=1,\ldots,D.$

À l'aide d'un paramètre de *désuétude* $\lambda\in[0,1]$, où l'on peut interpréter $\lambda\times 100$ comme une baisse ou une hausse relative en pourcentage des valeurs de $x_{q,di}$, un recensement désuet C^o est alors construit comme suit :

$$x_{q,di}^o=\begin{cases}x_{q,di}(1-\lambda), & i=1,\ldots,N_d,\ d=1,\ldots,15,31,\ldots,45,75,\ldots,80;\\ x_{q,di}(1+\lambda), & i=1,\ldots,N_d,\ d=16,\ldots,30,46,\ldots,74.\end{cases}$$

Les simulations sont réalisées en fonction de la distribution conjointe du modèle et du plan d'échantillonnage. Ainsi, pour chaque réplique de MC parmi les $L=1\,000$, nous générons un vecteur de population

de logarithmes des revenus $\mathbf{y}^{(\ell)} = (y_{11}^{(\ell)},\ldots,y_{di}^{(\ell)},\ldots,y_{DN_D}^{(\ell)})^t$ à partir du modèle de RLEE de (2.2), puis nous obtenons les valeurs de revenus selon $z_{di}^{(\ell)} = \exp(y_{di}^{(\ell)})$, $i = 1,\ldots,N_d$, $d = 1,\ldots,D$. Le seuil de pauvreté est fixé à $z =$ 12, ce qui correspond approximativement à 0,6 fois la médiane d'une population préliminaire de valeurs de z_{di} générées selon la procédure décrite ci-dessus. À partir des données de population générées, les valeurs réelles de chaque indicateur par domaine $\delta_d^{(\ell)} \in \{F_{0,d}^{(\ell)}, F_{1,d}^{(\ell)}\}$ sont calculées. Les échantillons $s^{(\ell)}$ et $s'^{(\ell)}$ sont ensuite tirés indépendamment. À l'aide des données issues des échantillons $s^{(\ell)}$ et $s'^{(\ell)}$, nous calculons les valeurs des estimateurs $\hat{\delta}_d^{DIR(\ell)}$, $\hat{\delta}_d^{FH(\ell)}$, $\hat{\delta}_d^{EBo(\ell)}$ et $\hat{\delta}_d^{SEB(\ell)}$. Le vecteur des paramètres du modèle $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, \sigma_u^2, \sigma_e^2)^t$ est estimé à partir de l'échantillon s au moyen de la méthode du maximum de vraisemblance restreint pour chacune des L répliques. Le rendement d'un estimateur $\hat{\delta}_d$ est évalué par rapport au biais relatif (BR) et à la racine de l'erreur quadratique moyenne relative (REQMR), obtenus comme suit :

$$\mathrm{BR}(\hat{\delta}_d) = \frac{L^{-1}\sum_{\ell=1}^{L}(\hat{\delta}_d^{(\ell)} - \delta_d^{(\ell)})}{L^{-1}\sum_{\ell=1}^{L}\delta_d^{(\ell)}}, \quad \mathrm{REQMR}(\hat{\delta}_d) = \frac{\sqrt{L^{-1}\sum_{\ell=1}^{L}(\hat{\delta}_d^{(\ell)} - \delta_d^{(\ell)})^2}}{L^{-1}\sum_{\ell=1}^{L}\delta_d^{(\ell)}},$$

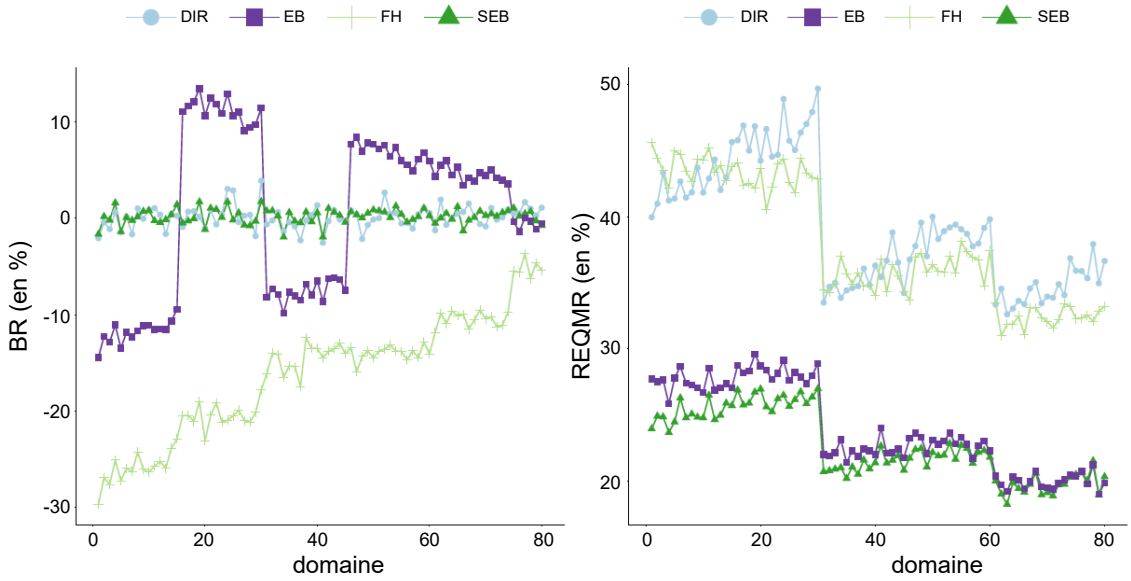
où $\hat{\delta}_d^{(\ell)}$ désigne l'estimation correspondante de $\delta_d^{(\ell)}$ à la ℓ^{e} réplique. En outre, nous avons calculé les moyennes, sur l'ensemble des domaines, du BR absolu (BRA) et de la REQMR :

$$\overline{\mathrm{BRA}} = D^{-1}\sum_{d=1}^{D}|\mathrm{BR}(\hat{\delta}_d)|, \quad \overline{\mathrm{REQMR}} = D^{-1}\sum_{d=1}^{D}\mathrm{REQMR}(\hat{\delta}_d).$$

La figure 6.1 présente le BR (à gauche) et la REQMR (à droite) en pourcentage des quatre estimateurs de l'écart de pauvreté, soit $\hat{\delta}_d^{DIR}$ (estimateur direct), $\hat{\delta}_d^{FH}$ (estimateur de Fay-Herriot), $\hat{\delta}_d^{EBo}$ (meilleur prédicteur empirique) et $\hat{\delta}_d^{SEB}$ (meilleur prédicteur empirique d'enquête), pour chaque domaine $d = 1,\ldots,D$ sur l'axe des x, classé par ordre croissant de la taille d'échantillon n_d, pour $\lambda =$ 0,2. Cette figure révèle un biais important pour les deux estimateurs obtenus en ignorant le caractère désuet du recensement, les estimateurs EB et FH. Cependant, les prédicteurs SEB semblent essentiellement sans biais, de la même façon que les estimateurs directs. Le diagramme de droite montre une REQMR très élevée pour l'estimateur direct et l'estimateur de FH, même si cela s'explique par la petite taille d'échantillon par domaine dans le premier cas et le biais élevé dans le second. Les REQMR des prédicteurs EB fondés sur C^o sont moins élevées, bien qu'ils soient clairement biaisés, tandis que le prédicteur SEB présente systématiquement des REQMR plus faibles. Les gains d'efficacité du prédicteur SEB par rapport au prédicteur EB sont d'autant plus marqués pour les domaines dont la taille d'échantillon est plus faible.

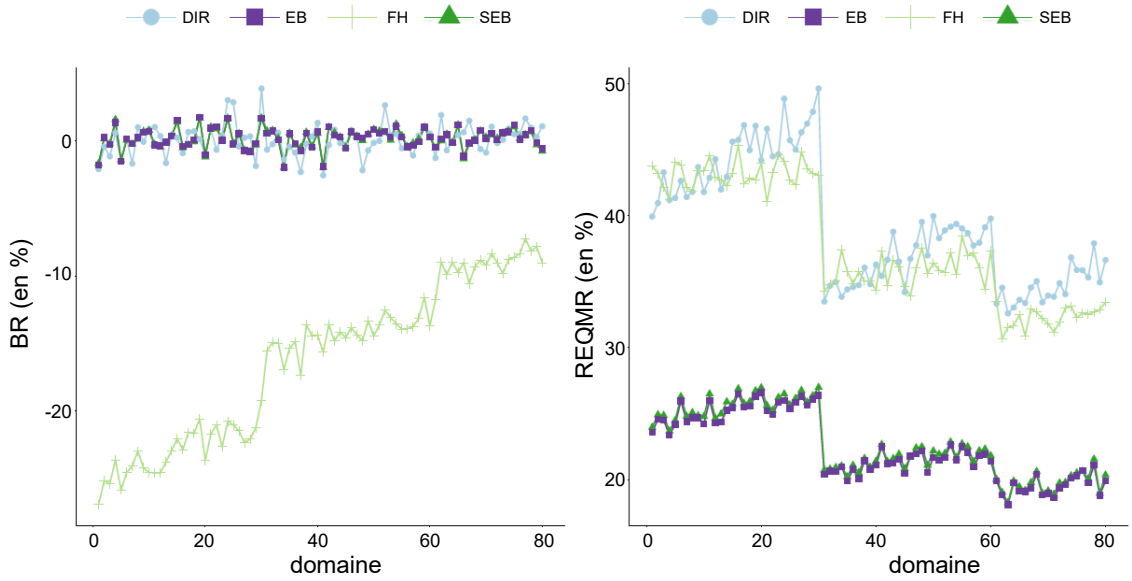
Les résultats analogues obtenus lorsque les données auxiliaires du recensement ne sont pas désuètes $(\lambda = 0)$ sont présentés à la figure 6.2. Dans ce cas, seul l'estimateur de FH semble biaisé, ce biais étant attribuable à des problèmes de non-linéarité, étant donné que les données sont générées au niveau des unités et que les indicateurs cibles δ_d ne sont pas linéairement liés aux moyennes de domaine des variables auxiliaires. Dans le diagramme de droite, on observe que le prédicteur SEB proposé offre un rendement quasi identique à celui du prédicteur EB qui est quasi optimal.

Figure 6.1 BR et REQMR en pourcentage des estimateurs DIR, FH, EB et SEB de l'écart de pauvreté $F_{1,d}$ pour chaque domaine d, avec un paramètre de désuétude $\lambda = 0{,}2$ et $n'_d = 10n_d$



Note : Estimateur direct (DIR); meilleur prédicteur empirique (EB); estimateur de Fay-Herriot (FH); biais relatif (BR); racine de l'erreur quadratique moyenne relative (REQMR); meilleur prédicteur empirique issu de l'enquête (SEB).

Figure 6.2 BR et REQMR en pourcentage des estimateurs DIR, FH, EB et SEB de l'écart de pauvreté $F_{1,d}$ pour chaque domaine d, avec un paramètre de désuétude $\lambda = 0$ et $n'_d = 10n_d$



Note : Estimateur direct (DIR); meilleur prédicteur empirique (EB); estimateur de Fay-Herriot (FH); biais relatif (BR); racine de l'erreur quadratique moyenne relative (REQMR); meilleur prédicteur empirique issu de l'enquête (SEB).

Les moyennes, sur l'ensemble des domaines, du BRA et de la REQMR des estimateurs de l'écart de pauvreté sont présentées au tableau 6.1 pour chaque valeur de λ en pourcentage. Ce tableau met de nouveau

en évidence le biais important de l'estimateur de FH pour toutes les valeurs de λ. Il montre également le comportement optimal de l'estimateur EB par rapport au BRA et à la REQMR lorsque le recensement est convenable $(\lambda = 0)$, suivi de très près par le prédicteur SEB. Toutefois, le BRA moyen de l'estimateur EB augmente à mesure que λ augmente, ce qui entraîne aussi une augmentation de sa REQMR moyenne, contrairement à l'estimateur SEB, qui n'est pas touché par la variation de λ. Pour le taux de pauvreté $F_{0,d}$, toutes les conclusions sont essentiellement les mêmes; voir les figures B.2 et B.1 ainsi que le tableau B.1 de l'annexe (section B). Par conséquent, le prédicteur SEB fondé sur une enquête à plus grande échelle s' semble une solution de rechange concurrentielle pour l'estimation pendant les années intercensitaires.

Tableau 6.1
Moyenne, sur l'ensemble des domaines, du BRA et de la REQMR des estimateurs DIR, FH, EB et SEB de l'écart de pauvreté $F_{1,d}$ selon λ, pour $n'_d = 10n_d$

Indicateur	λ (%)	$\overline{\text{BRA}}$(%)				$\overline{\text{REQMR}}$(%)			
		$\hat{\delta}_d^{DIR}$	$\hat{\delta}_d^{FH}$	$\hat{\delta}_d^{EB}$	$\hat{\delta}_d^{SEB}$	$\hat{\delta}_d^{DIR}$	$\hat{\delta}_d^{FH}$	$\hat{\delta}_d^{EB}$	$\hat{\delta}_d^{SEB}$
$F_{1,d}$	0	0,89	16,25	0,62	0,63	39,17	37,73	22,35	22,66
	10	0,89	16,25	3,99	0,63	39,17	37,76	22,76	22,66
	20	0,89	16,25	7,84	0,63	39,17	37,81	23,91	22,66
	30	0,89	16,29	11,71	0,63	39,17	37,88	25,68	22,66

Note : Biais relatif absolu (BRA); Estimateur direct (DIR); meilleur prédicteur empirique (EB); estimateur de Fay-Herriot (FH); racine de l'erreur quadratique moyenne relative (REQMR); meilleur prédicteur empirique issu de l'enquête (SEB).

Nous analysons maintenant l'évolution des résultats ci-dessus lorsque seule la petite enquête s était disponible. Pour cela, nous avons répété l'expérience de simulation en posant $s' = s$ pour le prédicteur SEB. La figure C.4 de l'annexe C présente les résultats pour l'écart de pauvreté $F_{1,d}$ lorsque le recensement est en fait correct $(\lambda = 0)$, mais que l'on applique néanmoins le prédicteur SEB. Comme on pouvait s'y attendre, le prédicteur EB fondé sur le recensement correct a un meilleur rendement que tous les autres estimateurs. Toutefois, la perte d'efficacité du prédicteur SEB par rapport au prédicteur EB demeure limitée. Il convient de mentionner que, si le recensement correct n'était pas disponible, le prédicteur EB ne serait pas calculable et le prédicteur SEB fondé uniquement sur le petit échantillon s offrirait toujours un meilleur rendement que les estimateurs DIR et FH, même si l'estimateur de FH est fondé sur les données du recensement correct. À l'inverse, lorsque les données de recensement sont désuètes, comme l'illustre la figure C.3 de la même annexe, le prédicteur SEB fondé uniquement sur s n'a pas beaucoup à perdre en ce qui concerne la REQMR par rapport au prédicteur EB, tout en demeurant approximativement sans biais. Les mêmes conclusions s'appliquent au taux de pauvreté $F_{0,d}$, comme l'indiquent le tableau C.1 et les figures C.2 et C.1 de la même annexe.

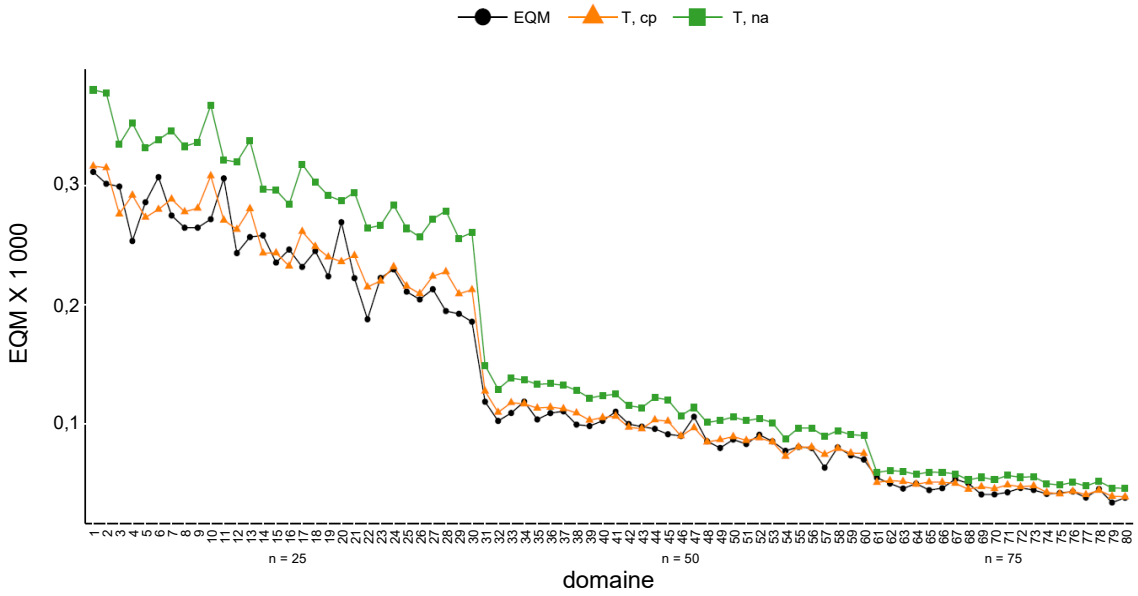
En ce qui concerne différentes valeurs du paramètre de désuétude λ lorsque $s' = s$, des conclusions similaires se dégagent à partir des résultats moyens présentés au tableau C.2 de l'annexe C. Ce tableau montre qu'à mesure que le paramètre de désuétude λ augmente, l'estimateur EB devient plus biaisé, tandis

que l'estimateur SEB fondé uniquement sur s demeure une solution de rechange intéressante, certes légèrement moins efficace, mais nettement moins biaisée que l'estimateur EB fondé sur des données de recensement désuètes.

Dans une nouvelle expérience de simulation, nous analysons maintenant le rendement des estimateurs de l'EQM totale du prédicteur SEB définis à la section 5, notamment $\text{eqm}_{T,na}(\hat{\delta}_d^{SEB})$ et $\text{eqm}_{T,cp}(\hat{\delta}_d^{SEB})$ donnés dans (5.8) et (5.10), respectivement. La véritable EQM totale du prédicteur SEB de l'écart de pauvreté a d'abord été approximée à l'aide de $L = 10\,000$ répliques de MC. Nous avons ensuite calculé l'espérance empirique des estimateurs bootstrap de l'EQM à l'aide de $B = 500$ répliques pour chacune des $L = 500$ simulations de MC.

La figure 6.3 présente, pour l'écart de pauvreté, les moyennes de MC des deux estimateurs bootstrap de l'EQM ainsi que l'EQM totale de MC du prédicteur SEB. Cette figure montre que l'estimateur bootstrap naïf de l'EQM (étiqueté « T, na ») mène systématiquement à une surestimation de l'EQM réelle, et cette surestimation est plus importante pour les domaines dont la taille d'échantillon est la plus faible. L'estimateur corrigé positif de l'EQM (étiqueté « T, cp ») réduit largement ce biais, et il semble proche des valeurs empiriques de l'EQM totale. Il convient de mentionner que ces dernières peuvent être légèrement influencées par l'erreur de MC, nous nous attendons donc à ce qu'elles deviennent plus régulières pour un plus grand nombre de répliques de MC.

Figure 6.3 EQM totale réelle du prédicteur SEB de l'écart de pauvreté, $F_{1,d}$, et espérances empiriques des estimateurs $\text{eqm}_{T,na}(\hat{\delta}_d^{SEB})$ et $\text{eqm}_{T,cp}(\hat{\delta}_d^{SEB})$ obtenues à l'aide de $B = 500$, pour chaque domaine, pour $n_d' = 10n_d$



Note : Erreur quadratique moyenne (EQM); meilleur prédicteur empirique issu de l'enquête (SEB).

Ces résultats indiquent que, même pour des valeurs de n'_d relativement élevées (dans nos simulations, la taille d'échantillon minimale est $n'_d = 250$), le deuxième terme de l'estimateur corrigé de l'EQM (5.9), qui tient compte de l'erreur de δ'_d en tant qu'estimateur de δ_d fondé sur s', demeure nécessaire. Ces conclusions sont valables pour toutes les valeurs de λ, puisque le prédicteur SEB ne repose pas sur le recensement désuet.

Nous soulignons également que, dans ces simulations, l'estimateur corrigé de l'EQM $\text{eqm}_{T,c}(\hat{\delta}_d^{SEB})$ donné dans (5.9) a toujours été positif. Par conséquent, la correction visant à le rendre positif n'a jamais été appliquée.

7. Cartographie de la pauvreté en Colombie

Comme source de données principale s, nous utilisons l'Enquête-ménage intégrée de Colombie de septembre 2023 (en espagnol, *Gran Encuesta Integrada de Hogares* [GEIH]), qui est une enquête-ménage nationale mensuelle menée par le Département administratif national de la statistique (en espagnol, *Departamento Administrativo Nacional de Estadística* [DANE]). Cette enquête vise à recueillir un large éventail de renseignements socioéconomiques auprès des ménages et des personnes. L'un des principaux objectifs de la GEIH est de mesurer les indicateurs de la vie active. Elle fournit également des microdonnées sur le revenu, utiles pour produire des estimations mensuelles des indicateurs de pauvreté et d'inégalité, comme le taux de pauvreté et l'écart de pauvreté. L'échantillon est tiré selon un plan d'échantillonnage à deux degrés. Au premier degré, un échantillon de municipalités est tiré selon un échantillonnage avec probabilité proportionnelle à la taille. Au second degré, des grappes de 10 ménages sont tirées au moyen d'un échantillonnage systématique. Pour les zones urbaines, la taille finale de l'échantillon de la GEIH est de 45 749 personnes. Notre objectif est d'estimer les taux et les écarts de pauvreté pour chacun des 24 départements colombiens, croisés avec l'autodéclaration de l'origine ethnique (en espagnol : Indigena [IND], Negro/Mulato [NM], Gitano/Raizal/Palenquero [GRP] et aucune des ethnies précédentes (NIN). Certaines ethnies ne sont pas représentées dans l'échantillon constitué pour certains départements. Étant donné que le présent article ne traite pas de l'estimation purement synthétique, le nombre de domaines correspond dans ce cas-ci aux $D = 85$ croisements avec l'échantillon. Au moins 32 de ces domaines cibles ont une taille d'échantillon $n_d < 10$, ce qui rend nécessaire le recours à des techniques d'estimation sur petits domaines.

Comme source de données secondaire s', nous utilisons l'Enquête sur les conditions de vie (en espagnol, *Encuesta de Calidad de Vida* [ECV]), également menée par le DANE. Cette enquête permet de quantifier et de caractériser les conditions de vie des ménages en Colombie et comprend des variables relatives au logement (matériaux des murs et des planchers; accès aux services publics, privés et communautaires), aux personnes (éducation, santé, garde d'enfants, utilisation des technologies de l'information et de la communication) et aux ménages (possession de biens et perceptions au sujet des conditions de vie au sein du logement, entre autres). Contrairement à la GEIH, l'ECV est menée annuellement. Son plan d'échantillonnage est très semblable à celui de la GEIH, mais comporte une stratification légèrement différente

mettant davantage l'accent sur les conditions de vie des ménages. Bien que l'ECV vise également à recueillir des microdonnées sur le revenu, son plan et sa fréquence ne permettent pas d'avoir une analyse détaillée de la pauvreté monétaire à court terme. L'ECV partage avec la GEIH des renseignements auxiliaires, comme le niveau de scolarité, les cotisations mensuelles à l'assurance sociale et les revenus provenant de programmes d'aide sociale. La taille de l'échantillon de l'ECV de 2023 pour les domaines communs avec la GEIH est de 148 688, ce qui en fait une enquête à plus grande échelle appropriée s' pour nos besoins.

Nous comparons d'abord les tailles d'échantillon observées par domaine de l'ECV n'_d aux tailles d'échantillon souhaitées $\hat{n}'_{d*}$, $d = 1,\ldots,D$, déterminées en fixant une erreur relative maximale $\epsilon_0 = 0{,}03$ avec une confiance de $1-\alpha = 0{,}95$. En prenant un coefficient de variation commun $cv_d = cv_0 = 0{,}1$, $d = 1,\ldots,D$, nous obtenons 26 domaines dont la taille d'échantillon observée n'_d est inférieure à la taille souhaitée n'_{d*}, comme l'indique le tableau 7.1. Dans ces 26 domaines, la précision du prédicteur SEB peut être inférieure à celle attendue.

Tableau 7.1
Nombre de domaines dont la taille d'échantillon observée de l'ECV n'_d est inférieure ou supérieure à la taille d'échantillon souhaitée $\hat{n}^*_d$, pour $\epsilon_0 = 0{,}03$ et $1-\alpha = 0{,}95$

Cas	$\min(n'_d)$	$\max(n'_d)$	Nombre de domaines
$n'_d < \hat{n}^*_d$	2	42	26
$n'_d \geqslant \hat{n}^*_d$	43	7 638	59

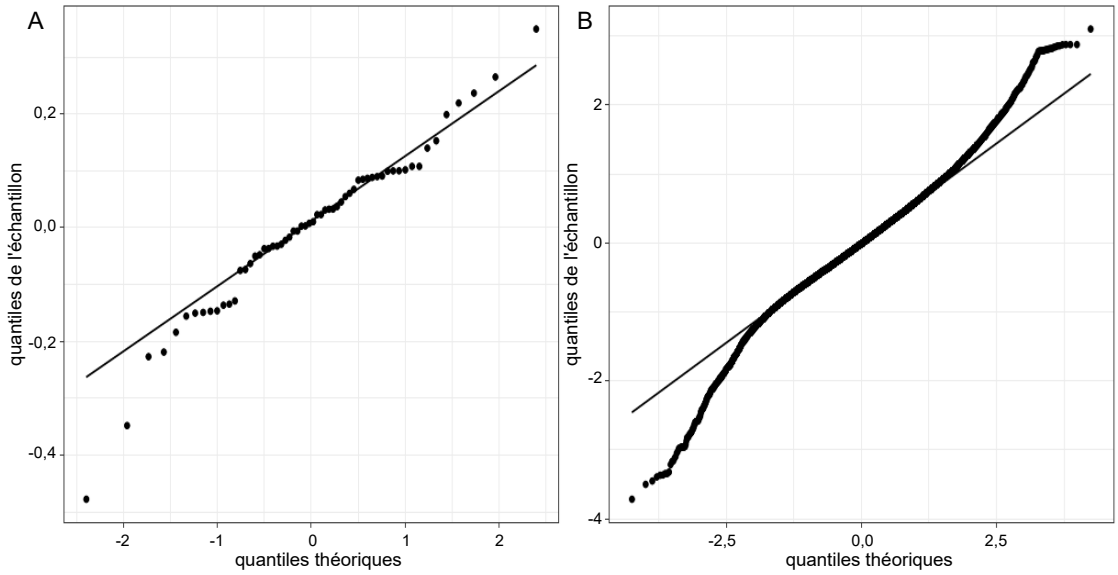
Note : *Encuesta de Calidad de Vida* (ECV).

Nous excluons de notre analyse 25 des 26 domaines pour lesquels $n'_d < \hat{n}^*_d$ et ne conservons qu'un seul domaine de $n_d = 28$, pour lequel nous établissons que $s'_d = s_d$. Parmi les 59 domaines où $n'_d > \hat{n}^*_d$, un seul domaine présentait $n'_d < n_d$, et nous avons également établi que $s'_d = s_d$.

La GEIH vise à déterminer le revenu par habitant z_{di}, $i = 1,\ldots,n_d$, $d = 1,\ldots,D$, mesuré en milliers de pesos colombiens (COP/1 000) et calculé selon la procédure normalisée établie par le DANE. Nous ajustons le modèle de RLEE avec $y_{di} = \log(z_{di} + k)$ comme variable réponse et en prenant $k = 65$ pour obtenir des valeurs positives. Les variables auxiliaires communes à la GEIH et à l'ECV sont le sexe (homme ou femme), le fait d'être prestataire de sécurité sociale (oui ou non), le montant cotisé à la sécurité sociale (en milliers de COP, également transformé au moyen d'un décalage logarithmique) ainsi que le niveau de scolarité (classé en 10 catégories). Le seuil de pauvreté (en milliers de COP) varie selon l'emplacement géographique et est également fourni par le DANE dans les microdonnées de la GEIH.

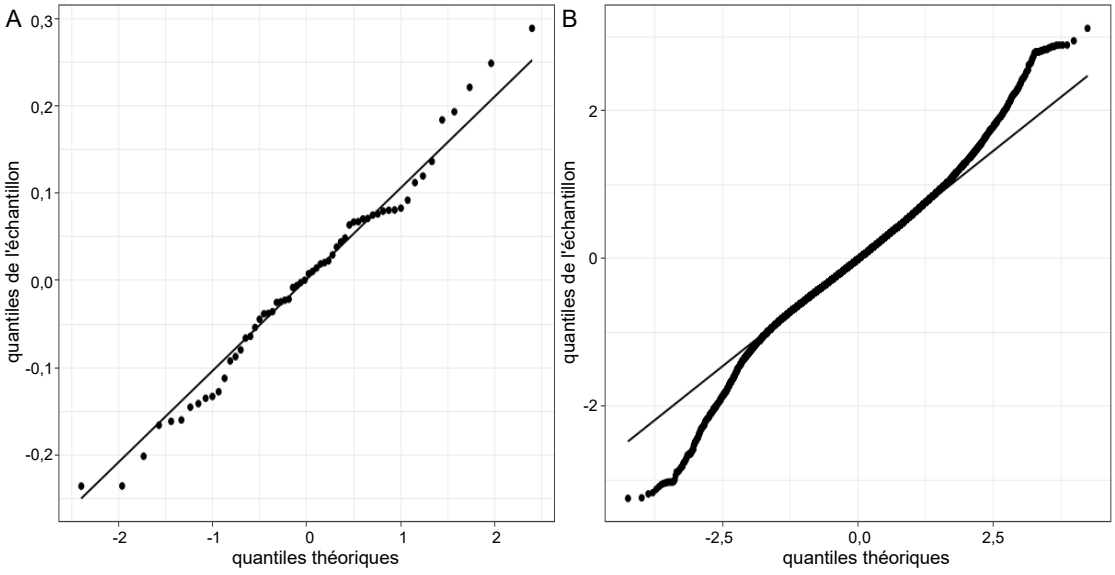
Après l'ajustement de ce modèle initial, nous avons procédé à des diagnostics sur le modèle. Comme on peut le constater au bas de la figure 7.1 (A), le diagramme quantile-quantile pour loi normale des effets de domaine prédits $\hat{u}_d$ présente deux domaines atypiques. Un diagramme quantile-quantile pour loi normale des résidus $\hat{e}_{di} = y_{di} - \mathbf{x}^t_{di}\hat{\boldsymbol{\beta}} - \hat{u}_d$ montre également des queues plus lourdes, comme l'illustre la figure 7.1 (B).

Figure 7.1 Diagrammes quantile-quantile pour loi normale (A) des effets de domaine prédits et (B) des résidus au niveau des unités issus du modèle initial



Afin d'éviter un biais dans les estimations des taux de pauvreté pour ces domaines atypiques (Magdalena - NM et Huila - NIN), nous avons introduit dans le modèle un effet fixe pour ces deux domaines ainsi qu'un autre effet fixe pour les 20 personnes atypiques présentant les résidus estimés $\hat{e}_{di}$ les plus faibles. Après l'ajout de ces deux effets fixes, le diagramme quantile-quantile pour loi normale de $\hat{u}_d$ obtenu ne montre plus de domaines atypiques, comme l'illustre la figure 7.2 (A), et l'on observe une légère déviation par rapport à la normalité dans les queues des résidus au niveau des unités; voir la figure 7.2 (B). Étant donné que les personnes dans l'ECV sont différentes, cet effet fixe a été fixé à zéro pour l'ECV.

Figure 7.2 Diagrammes quantile-quantile pour loi normale (A) des effets de domaine prédits et (B) des résidus au niveau des unités issus du modèle de travail

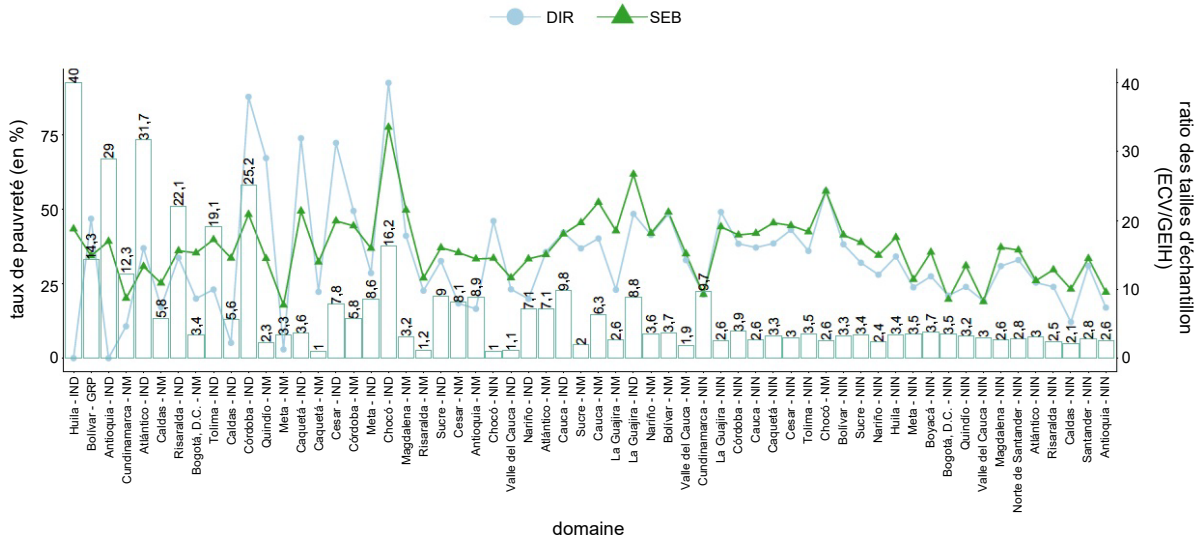


Étant donné que les chiffres estimés $\hat{N}_d$ des deux enquêtes diffèrent, nous avons appliqué une calibration linéaire (Deville et Särndal, 1992) aux poids de la GEIH w_{di}, de manière à ce que les poids calibrés w_{di}^C satisfassent $\sum_{i\in s_d} w_{di}^C = \hat{N}'_d$, pour $\hat{N}'_d = \sum_{i\in s'_d} w'_{di}$.

Puisqu'aucune déviation importante par rapport au modèle de travail n'a été constatée, les estimations directes par extension $\hat{F}_{0,d}^{DIR}$ fondées sur les poids calibrés ainsi que les estimations SEB $\hat{F}_{0,d}^{SEB}$ des taux de pauvreté ont été obtenues à l'aide de ce modèle. La figure 7.3 présente les estimations obtenues pour chaque domaine sur l'axe des x, et les domaines sont classés par ordre croissant de la taille d'échantillon n_d. Un axe vertical secondaire présente le ratio des tailles d'échantillon par domaine entre les enquêtes ECV et GEIH, n'_d / n_d. La figure montre une tendance similaire pour les deux types d'estimations. Toutefois, les estimations directes semblent très instables pour les domaines dont la taille d'échantillon est plus faible. Pour les domaines ayant une taille d'échantillon plus importante, les deux types d'estimations concordent dans une large mesure.

Dans les domaines correspondant aux départements de Huila et d'Antioquia croisés avec l'appartenance ethnique indigène (IND), pour lesquels les tailles d'échantillon de la GEIH sont respectivement de 2 et de 3, les estimations directes du taux de pauvreté sont nulles, des résultats peu plausibles. En revanche, les tailles d'échantillon de l'ECV sont respectivement 40 et 29 fois plus importantes, ce qui conduit à des estimations SEB des taux de pauvreté non nulles et beaucoup plus plausibles. Les résultats relatifs à l'écart de pauvreté $F_{1,d}$ présentent des tendances très similaires; voir les figures D.1 et D.2 de la section D de l'annexe.

Figure 7.3 Estimations DIR et SEB du taux de pauvreté $F_{0,d}$, domaines classés par ordre croissant de la taille d'échantillon de la GEIH n_d



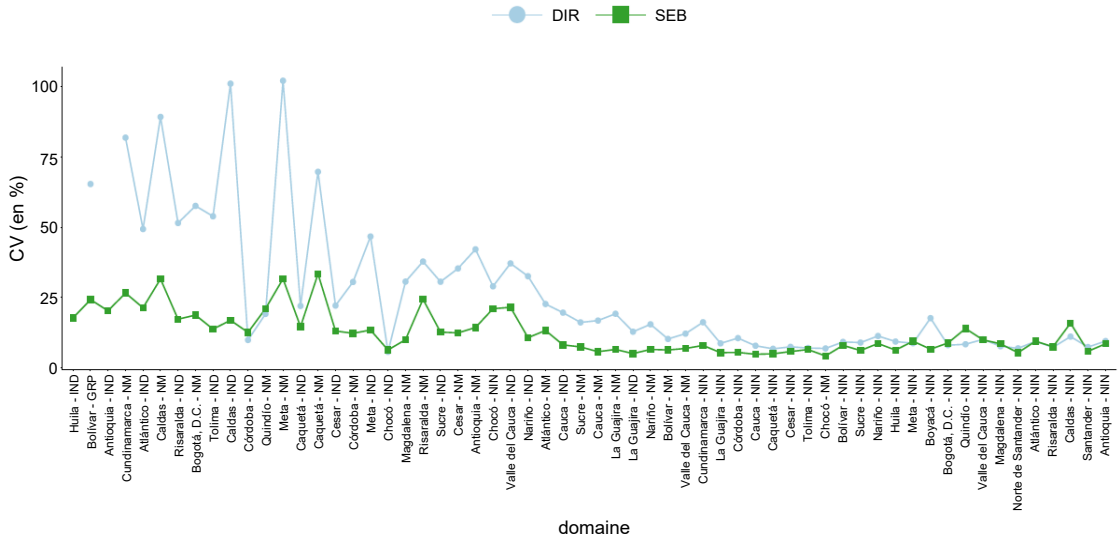
Note : Estimations directes (DIR); *Encuesta de Calidad de Vida* (ECV); *Gran Encuesta Integrada de Hogares* (GEIH); Gitano/Raizal/ Palenquero (GRP); Indigena (IND); Negro/Mulato (NM); estimations du meilleur prédicteur empirique issu de l'enquête (SEB).

En ce qui concerne l'efficacité, la figure 7.4 présente les CV estimés pour tous les domaines, lesquels sont classés par ordre croissant de la taille d'échantillon n_d ; les CV des estimateurs DIR et SEB correspondent en fait aux REQMR estimées (exprimées en pourcentage), lesquelles sont calculées pour l'estimateur SEB à l'aide de l'estimation bootstrap paramétrique corrigée et positive de l'EQM totale donnée dans (5.10), à partir de $B = 1\,000$ répliques bootstrap. Comme dans nos expériences de simulation, l'estimateur corrigé de l'EQM $\text{eqm}_{T,c}(\hat{F}_{0,d}^{SEB})$ s'est de nouveau révélé positif pour tous les domaines, ce qui fait que la correction pour assurer la positivité de $\text{eqm}_{T,cp}(\hat{F}_{0,d}^{SEB})$ est inutile. Il en était de même pour le prédicteur SEB de l'écart de pauvreté dans l'ensemble des domaines.

La figure 7.4 illustre la réduction du CV obtenue avec le prédicteur SEB par rapport à l'estimateur direct. Cette réduction est particulièrement marquée pour les domaines ayant une petite taille d'échantillon. On observe qu'à mesure que la taille d'échantillon du domaine n_d augmente, leurs CV estimés tendent à se rapprocher.

On constate également que, pour certains domaines, tels que Córdoba - IND, ayant une petite taille d'échantillon $(n_d = 23)$, le CV estimé de l'estimateur direct, qui devrait augmenter à mesure que la taille de l'échantillon diminue, s'écarte de la tendance générale. Cet écart par rapport à la tendance escomptée laisse supposer que l'estimation du CV associée à l'estimateur direct pour le domaine Córdoba - IND pourrait être peu fiable. Par ailleurs, dans les domaines où l'estimateur direct est nul, comme Huila - IND et Antioquia - IND, il n'a pas été possible d'estimer le CV de l'estimateur direct. En revanche, le prédicteur SEB fournit une estimation non nulle du taux de pauvreté et un CV estimé connexe, ce qui offre des résultats plus informatifs et plus plausibles pour ces domaines à petite taille d'échantillon.

Figure 7.4 CV estimés des estimateurs DIR et SEB du taux de pauvreté $F_{0,d}$, domaines classés par ordre croissant de la taille d'échantillon de la GEIH n_d



Note : Coefficient de variation (CV); Estimations directes (DIR); *Gran Encuesta Integrada de Hogares* (GEIH); Gitano/Raizal/ Palenquero (GRP); Indigena (IND); Negro/Mulato (NM); estimations du meilleur prédicteur empirique issu de l'enquête (SEB)

8. Conclusions

Le présent article propose des estimateurs sur petits domaines pour des indicateurs additifs pendant les années intercensitaires, lorsque le recensement actuel est soit indisponible, soit désuet. Le meilleur prédicteur empirique issu de l'enquête (SEB) proposé nécessite une enquête secondaire à plus grande échelle s', qui couvre adéquatement l'ensemble des domaines figurant dans l'enquête actuelle s et partage avec celle-ci certaines variables auxiliaires. Dans le pire scénario, le prédicteur SEB peut être calculé uniquement à partir de l'enquête actuelle s. Même dans ce cas, le prédicteur SEB a un meilleur rendement que l'estimateur direct par extension habituel. Par ailleurs, le prédicteur SEB tend vers le meilleur prédicteur empirique pour le recensement lorsque la taille de l'échantillon par domaine n'_d est importante, ce qui est approximativement optimal si la fraction d'échantillonnage f_d est négligeable, et que le recensement correct est disponible.

Les résultats de simulation montrent que le prédicteur SEB demeure sans biais et non influencé par l'utilisation de microdonnées de recensement désuètes, une situation qui entraîne un biais dans le meilleur prédicteur empirique (EB). En outre, l'estimateur corrigé positif de l'erreur quadratique moyenne (EQM) totale $\text{eqm}_{T,cp}(\hat{\delta}_d^{SEB})$ affiche des résultats nettement supérieurs à ceux de l'estimateur bootstrap naïf de l'EQM fondé sur le modèle $\text{eqm}_{T,na}(\hat{\delta}_d^{SEB})$, dont le rendement diminue lorsque n'_d est de petite taille.

La procédure proposée intègre des données provenant des échantillons s et s'. Cette approche d'intégration de données peut être appliquée même lorsque l'échantillon plus grand s' est une enquête non probabiliste. Comme le soulignent Sen et Lahiri (2025), cela est possible grâce à l'utilisation de poids construits, comme ceux proposés par Chen, Li et Wu (2020), qui permettent de corriger le biais de sélection et d'harmoniser l'échantillon non probabiliste avec la population cible. Cette approche permet ainsi à l'analyse combinée de tirer parti de la couverture plus large de s', tout en conservant la validité inférentielle fournie par l'échantillon probabiliste s.

Le prédicteur SEB proposé n'a pas été conçu pour traiter le problème de l'échantillonnage informatif dans l'échantillon s. Cependant, il peut être étendu en recourant au prédicteur pseudo-EB élaboré par Guadarrama et coll. (2018) ou au prédicteur EB dans le cadre d'un échantillonnage informatif élaboré par Cho et coll. (2024).

Le prédicteur SEB proposé peut également être étendu à des indicateurs plus complexes ainsi qu'à des modèles linéaires mixtes paramétriques ou semi-paramétriques plus sophistiqués (voir par exemple Arias-Salazar, Gutiérrez, Guerrero-Gómez, Mancero, Rojas-Perilla et Zhang, 2025; Bikauskaite, Molina et Morales, 2022; Bugallo, Esteban, Hobza, Morales et Pérez, 2024; Chambers, Salvati et Tzavidis, 2015).

Enfin, nous soulignons que le prédicteur SEB et ses estimateurs de l'EQM totale reposent sur un modèle paramétrique. Cela signifie que l'ensemble des hypothèses sous-jacentes au modèle doivent être rigoureusement vérifiées au moyen de diagnostics des modèles et que, en présence de déviations claires par rapport au modèle, celui-ci doit être modifié.

Remerciements

Le présent article a bénéficié du soutien du ministère de la Science et de l'Innovation de l'Espagne [PID2020-115598RB-I0].

Annexe

A. Preuves des résultats

Démonstration de la proposition 1 : Soit $\overline{e}_d = n_d^{-1}\sum_{i\in s_d} e_{di}$ et $\overline{E}_d = N_d^{-1}\sum_{i=1}^{N_d} e_{di}$. D'après le modèle de RLEE,

$$\overline{Y}_d = \overline{\mathbf{X}}_d^t\boldsymbol{\beta} + u_d + \overline{E}_d.$$

Par conséquent, l'erreur de prédiction associée au prédicteur fondé sur un recensement désuet est

$$\begin{aligned}\tilde{\overline{Y}}_d^{CBo} - \overline{Y}_d &= (\overline{\mathbf{X}}_d^o)^t\boldsymbol{\beta} + \gamma_d(\overline{y}_d - \overline{\mathbf{x}}_d^t\boldsymbol{\beta}) - \left[\overline{\mathbf{X}}_d^t\boldsymbol{\beta} + u_d + \overline{E}_d\right] \\ &= \gamma_d(\overline{y}_d - \overline{\mathbf{x}}_d^t\boldsymbol{\beta}) - (u_d + \overline{E}_d) + (\overline{\mathbf{X}}_d^o - \overline{\mathbf{X}}_d)^t\boldsymbol{\beta} \\ &= \gamma_d(u_d + \overline{e}_d) - (u_d + \overline{E}_d) - \overline{\mathbf{b}}_d^t\boldsymbol{\beta}.\end{aligned}$$

En prenant les espérances selon le modèle et en utilisant $E(u_d) = E(\overline{e}_d) = E(\overline{E}_d) = 0,$ on obtient

$$B_{\mathbf{y}}(\tilde{\overline{Y}}_d^{CBo}) = -\overline{\mathbf{b}}_d^t\boldsymbol{\beta},$$

ce qui démontre (i).

Pour (ii), il convient de mentionner que

$$\tilde{\overline{Y}}_d^{CBo} - \overline{Y}_d = (\tilde{\overline{Y}}_d^{CB} - \overline{Y}_d) - \overline{\mathbf{b}}_d^t\boldsymbol{\beta}.$$

Étant donné que $\overline{\mathbf{b}}_d^t\boldsymbol{\beta}$ est constant selon le modèle,

$$V_{\mathbf{y}}(\tilde{\overline{Y}}_d^{CBo} - \overline{Y}_d) = V_{\mathbf{y}}(\tilde{\overline{Y}}_d^{CB} - \overline{Y}_d).$$

Alors,

$$\mathrm{EQM}_{\mathbf{y}}(\tilde{\overline{Y}}_d^{CBo}) = \mathrm{EQM}_{\mathbf{y}}(\tilde{\overline{Y}}_d^{CB}) + (\overline{\mathbf{b}}_d^t\boldsymbol{\beta})^2.$$

Il reste à calculer $\mathrm{EQM}_{\mathbf{y}}(\tilde{\overline{Y}}_d^{CB})$. On observe que,

$$\tilde{\overline{Y}}_d^{CB} - \overline{Y}_d = \gamma_d(u_d + \overline{e}_d) - (u_d + \overline{E}_d) = (\gamma_d - 1)\,u_d + \gamma_d\,\overline{e}_d - \overline{E}_d.$$

Selon le modèle, on a

$$V_{\mathbf{y}}(u_d) = \sigma_u^2, \quad V_{\mathbf{y}}(\overline{e}_d) = \frac{\sigma_e^2}{n_d}, \quad V_{\mathbf{y}}(\overline{E}_d) = \frac{\sigma_e^2}{N_d}, \quad \mathrm{Cov}_{\mathbf{y}}(\overline{e}_d, \overline{E}_d) = \frac{\sigma_e^2}{N_d},$$

et u_d est indépendante des termes d'erreur. Par conséquent,

$$\mathrm{EQM}_{\mathbf{y}}(\tilde{\bar{Y}}_d^{CB}) = (\gamma_d - 1)^2 \sigma_u^2 + \gamma_d^2 \frac{\sigma_e^2}{n_d} + \frac{\sigma_e^2}{N_d} - 2\gamma_d \frac{\sigma_e^2}{N_d}.$$

En réarrangeant les termes et en notant que $\sigma_u^2(1-\gamma_d) = \gamma_d \sigma_e^2 / n_d$, on obtient le résultat recherché.

Démonstration de la proposition 2 : (i) Selon le modèle de RLEE, le prédicteur CB de $\bar{Y}_d$ est donné par

$$\tilde{\bar{Y}}_d^{CB} = \bar{\mathbf{X}}_d^t \boldsymbol{\beta} + \gamma_d (\bar{y}_d - \bar{\mathbf{x}}_d^t \boldsymbol{\beta}).$$

Par ailleurs, le prédicteur SB selon le même modèle est donné par

$$\begin{aligned} \tilde{\bar{Y}}_d^{SB} &= (\tilde{\bar{\mathbf{X}}}_{ds'})^t \boldsymbol{\beta} + \gamma_d (\bar{y}_d - \bar{\mathbf{x}}_d^t \boldsymbol{\beta}) \\ &= (\tilde{\bar{\mathbf{X}}}_{ds'})^t \boldsymbol{\beta} + \gamma_d (\bar{y}_d - \bar{\mathbf{x}}_d^t \boldsymbol{\beta}) + \bar{\mathbf{X}}_d^t \boldsymbol{\beta} - \bar{\mathbf{X}}_d^t \boldsymbol{\beta} \\ &= \tilde{\bar{Y}}_d^{CB} - (\bar{\mathbf{X}}_d - \tilde{\bar{\mathbf{X}}}_{ds'})^t \boldsymbol{\beta}. \end{aligned}$$

L'erreur de prédiction du prédicteur SB $\tilde{\bar{Y}}_d^{SB}$ est alors

$$\tilde{\bar{Y}}_d^{SB} - \bar{Y}_d = \tilde{\bar{Y}}_d^{CB} - \bar{Y}_d - (\bar{\mathbf{X}}_d - \tilde{\bar{\mathbf{X}}}_{ds'})^t \boldsymbol{\beta}. \tag{A.1}$$

En prenant l'espérance par rapport à $\mathbf{y}$ étant donné s', on obtient le biais par rapport au modèle de $\tilde{\bar{Y}}_d^{SB}$,

$$\begin{aligned} B_{\mathbf{y}}(\tilde{\bar{Y}}_d^{SB} \mid s') &= E_{\mathbf{y}}(\tilde{\bar{Y}}_d^{CB} - \bar{Y}_d \mid s') - (\bar{\mathbf{X}}_d - \tilde{\bar{\mathbf{X}}}_{ds'})^t \boldsymbol{\beta} \\ &= (\tilde{\bar{\mathbf{X}}}_{ds'} - \bar{\mathbf{X}}_d)^t \boldsymbol{\beta}. \end{aligned} \tag{A.2}$$

En prenant ensuite l'espérance de (A.2) par rapport à s', on obtient le biais total.

$$B_T(\tilde{\bar{Y}}_d^{SB}) = \boldsymbol{\beta}^t B_{s'}(\tilde{\bar{\mathbf{X}}}_{ds'}). \tag{A.3}$$

(ii) En prenant maintenant la variance du modèle de l'erreur de prédiction (A.1), on obtient

$$V_{\mathbf{y}}(\tilde{\bar{Y}}_d^{SB} - \bar{Y}_d \mid s') = V_{\mathbf{y}}(\tilde{\bar{Y}}_d^{CB} - \bar{Y}_d \mid s') = \mathrm{EQM}_{\mathbf{y}}(\tilde{\bar{Y}}_d^{CB} \mid s'). \tag{A.4}$$

Le modèle de l'EQM de $\tilde{\bar{Y}}_d^{SB}$ est alors

$$\begin{aligned} \mathrm{EQM}_{\mathbf{y}}(\tilde{\bar{Y}}_d^{SB} \mid s') &= V_{\mathbf{y}}(\tilde{\bar{Y}}_d^{SB} - \bar{Y}_d \mid s') + B_{\mathbf{y}}^2(\tilde{\bar{Y}}_d^{SB} \mid s') \\ &= \mathrm{EQM}_{\mathbf{y}}(\tilde{\bar{Y}}_d^{CB} \mid s') + [(\tilde{\bar{\mathbf{X}}}_{ds'} - \bar{\mathbf{X}}_d)^t \boldsymbol{\beta}]^2. \end{aligned}$$

Enfin, en prenant l'espérance par rapport à s' et en notant que $\tilde{\bar{Y}}_d^{CB}$ ne dépend pas de s', on obtient l'EQM totale

$$\mathrm{EQM}_T(\tilde{\bar{Y}}_d^{SB}) = \mathrm{EQM}_{\mathbf{y}}(\tilde{\bar{Y}}_d^{CB}) + \boldsymbol{\beta}^t E_{s'}[(\tilde{\bar{\mathbf{X}}}_{ds'} - \bar{\mathbf{X}}_d)(\tilde{\bar{\mathbf{X}}}_{ds'} - \bar{\mathbf{X}}_d)^t] \boldsymbol{\beta}.$$

Démonstration de la proposition 3 : On remarque que, lorsque $w'_{d\cdot} = N_d$, $\delta'_d = N_d^{-1} \sum_{i \in s'_d} w'_{di} \delta_{di}$ est un estimateur de HT de $\delta_d = N_d^{-1} \sum_{i=1}^{N_d} \delta_{di}$, qui est sans biais selon le plan de sondage s'. Tout d'abord, en soustrayant et en ajoutant δ'_d, on décompose l'EQM totale de $\hat{\delta}_d^{SEB}$ comme suit

$$\begin{aligned}\mathrm{EQM}_T(\hat{\delta}_d^{SEB}) &= E_{(\mathbf{y},s')}\left[(\hat{\delta}_d^{SEB}-\delta_d')^2\right]+E_{(\mathbf{y},s')}\left[(\delta_d'-\delta_d)^2\right]\\ &\quad + 2E_{(\mathbf{y},s')}\left[(\hat{\delta}_d^{SEB}-\delta_d')(\delta_d'-\delta_d)\right].\end{aligned} \tag{A.5}$$

Selon la loi des espérances itérées et le fait que $E_{s'}(\delta_d' \mid \mathbf{y}) = \delta_d$, le deuxième terme du côté droit de (A.5) devient

$$E_{(\mathbf{y},s')}\left[(\delta_d'-\delta_d)^2\right] = E_{\mathbf{y}}\left[V_{s'}(\delta_d' \mid \mathbf{y})\right]. \tag{A.6}$$

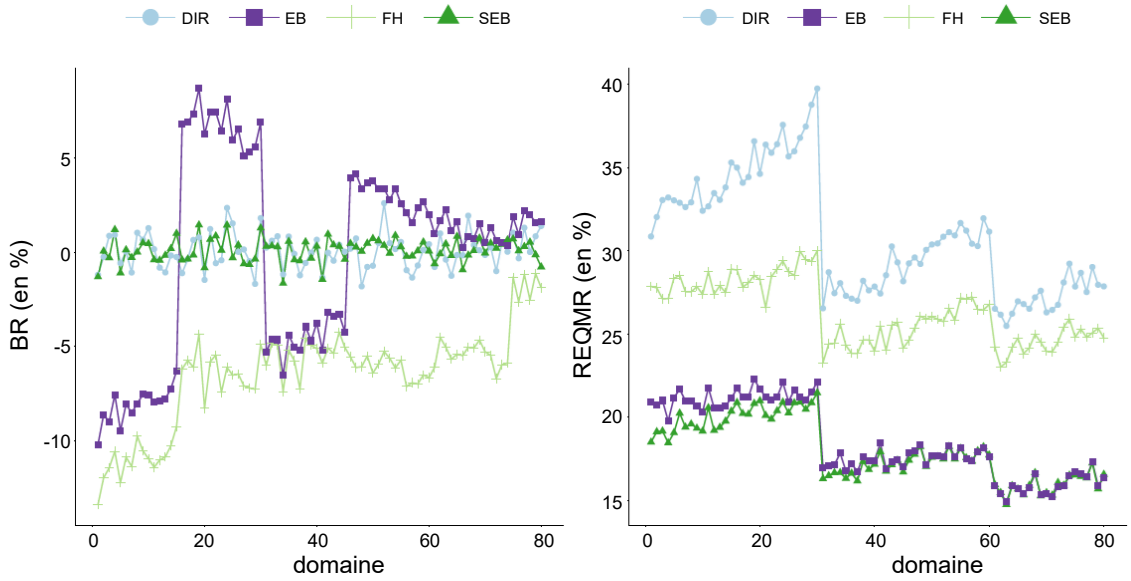
En utilisant les mêmes résultats, le terme du produit croisé de (A.5) peut s'écrire

$$\begin{aligned}E_{(\mathbf{y},s')}\left[(\hat{\delta}_d^{SEB}-\delta_d')(\delta_d'-\delta_d)\right] &= E_{\mathbf{y}}[E_{s'}(\hat{\delta}_d^{SEB}\delta_d' \mid \mathbf{y})-\delta_d E_{s'}(\hat{\delta}_d^{SEB} \mid \mathbf{y})-E_{s'}(\delta_d'^2 \mid \mathbf{y})+\delta_d^2]\\ &= E_{\mathbf{y}}[\mathrm{Cov}_{s'}(\hat{\delta}_d^{SEB},\delta_d' \mid \mathbf{y})-V_{s'}(\delta_d' \mid \mathbf{y})].\end{aligned} \tag{A.7}$$

Le résultat découle alors du remplacement de (A.6) et de (A.7) dans (A.5) et, encore, de l'application de la loi des espérances itérées au premier terme.

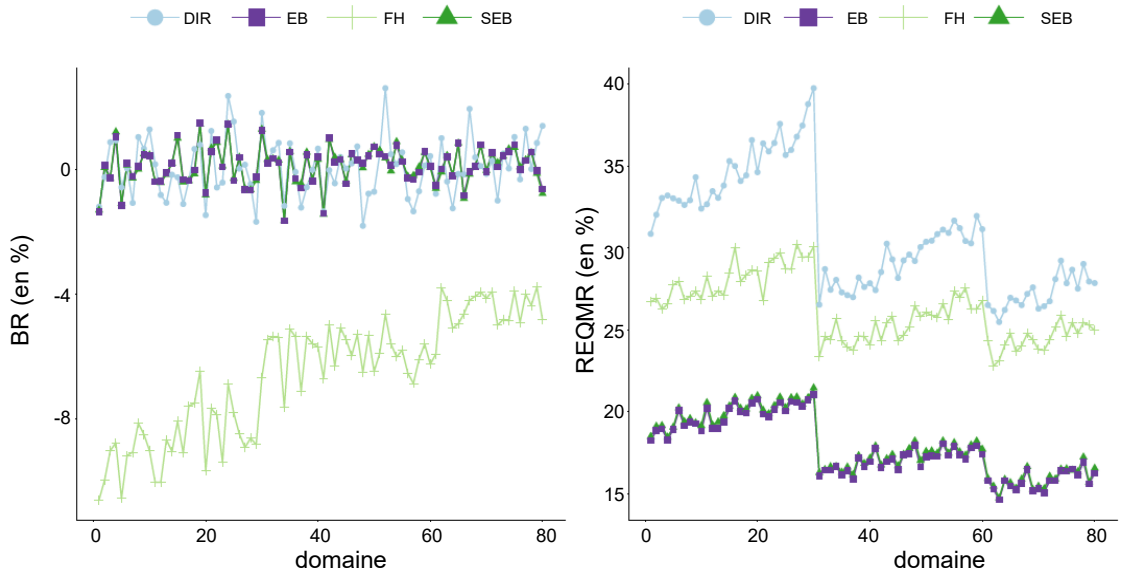
B. Résultats de simulation avec l'enquête à plus grande échelle : taux de pauvreté

Figure B.1 BR et REQMR en pourcentage des estimateurs DIR, FH, EB et SEB du taux de pauvreté $F_{0,d}$ pour chaque domaine d, pour $\lambda = 0{,}2$ et $n_d' = 10n_d$



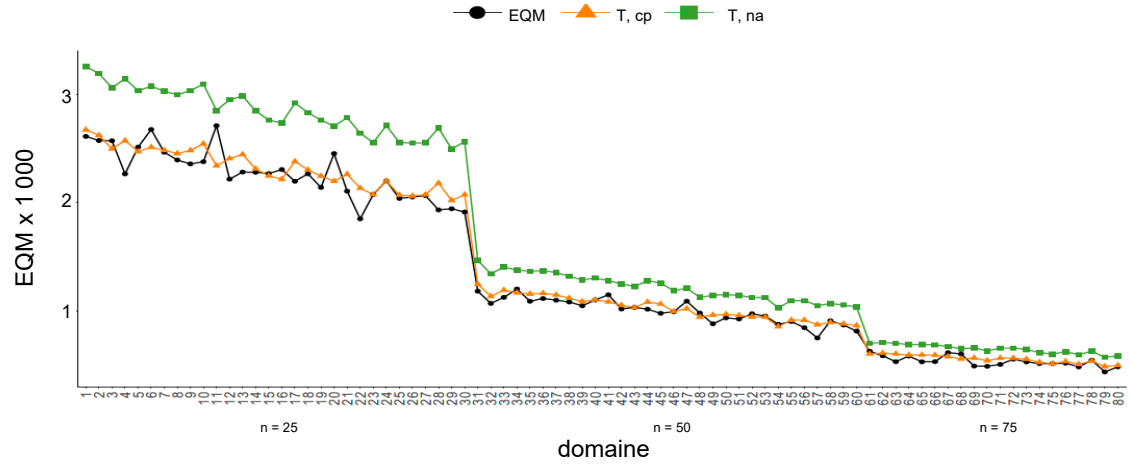
Note : Estimateur direct (DIR); meilleur prédicteur empirique (EB); estimateur de Fay-Herriot (FH); biais relatif (BR); racine de l'erreur quadratique moyenne relative (REQMR); meilleur prédicteur empirique issu de l'enquête (SEB).

Figure B.2 BR et REQMR en pourcentage des estimateurs DIR, FH, EB et SEB du taux de pauvreté $F_{0,d}$ pour chaque domaine d, pour $\lambda = 0$ et $n'_d = 10n_d$



Note : Estimateur direct (DIR); meilleur prédicteur empirique (EB); estimateur de Fay-Herriot (FH); biais relatif (BR); racine de l'erreur quadratique moyenne relative (REQMR); meilleur prédicteur empirique issu de l'enquête (SEB).

Figure B.3 EQM totale réelle en pourcentage du prédicteur SEB du taux de pauvreté $F_{0,d}$ pour chaque domaine d, et espérances empiriques des estimateurs bootstrap naïf et corrigé positif de l'EQM obtenues à l'aide de $B = 500$ répliques bootstrap, pour $n'_d = 10n_d$



Note : Erreur quadratique moyenne (EQM); meilleur prédicteur empirique issu de l'enquête (SEB).

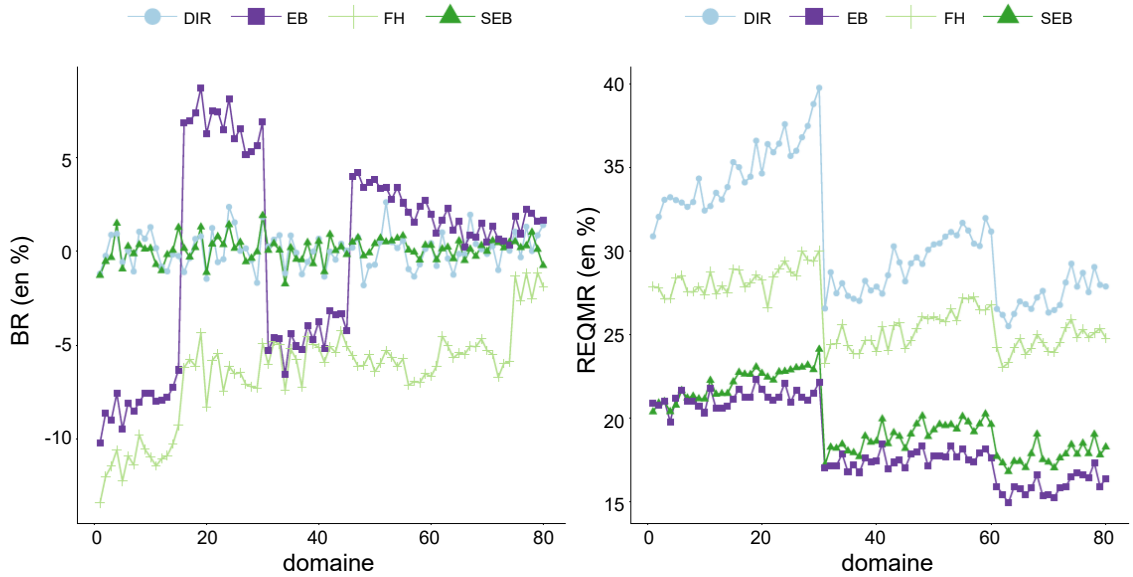
Tableau B.1
Moyenne sur l'ensemble des domaines du BRA et de la REQMR pour les estimateurs DIR, FH, EB et SEB du taux de pauvreté $F_{0,d}$ selon λ, pour $n'_d = 10n_d$

Indicateur	λ (%)	BRA (en %)				REQMR (en %)			
		$\hat{\delta}_d^{DIR}$	$\hat{\delta}_d^{FH}$	$\hat{\delta}_d^{EB}$	$\hat{\delta}_d^{SEB}$	$\hat{\delta}_d^{DIR}$	$\hat{\delta}_d^{FH}$	$\hat{\delta}_d^{EB}$	$\hat{\delta}_d^{SEB}$
$F_{0,d}$	0	0,73	6,56	0,49	0,50	30,81	26,20	17,77	17,97
	10	0,73	6,53	2,31	0,50	30,81	26,21	17,96	17,97
	20	0,73	6,53	4,50	0,50	30,81	26,25	18,50	17,97
	30	0,73	6,56	6,72	0,50	30,81	26,32	19,34	17,97

Note : Biais relatif absolu (BRA); estimateurs directs (DIR); meilleur prédicteur empirique (EB); estimateur de Fay-Herriot (FH); racine de l'erreur quadratique moyenne relative (REQMR); meilleur prédicteur empirique issu de l'enquête (SEB).

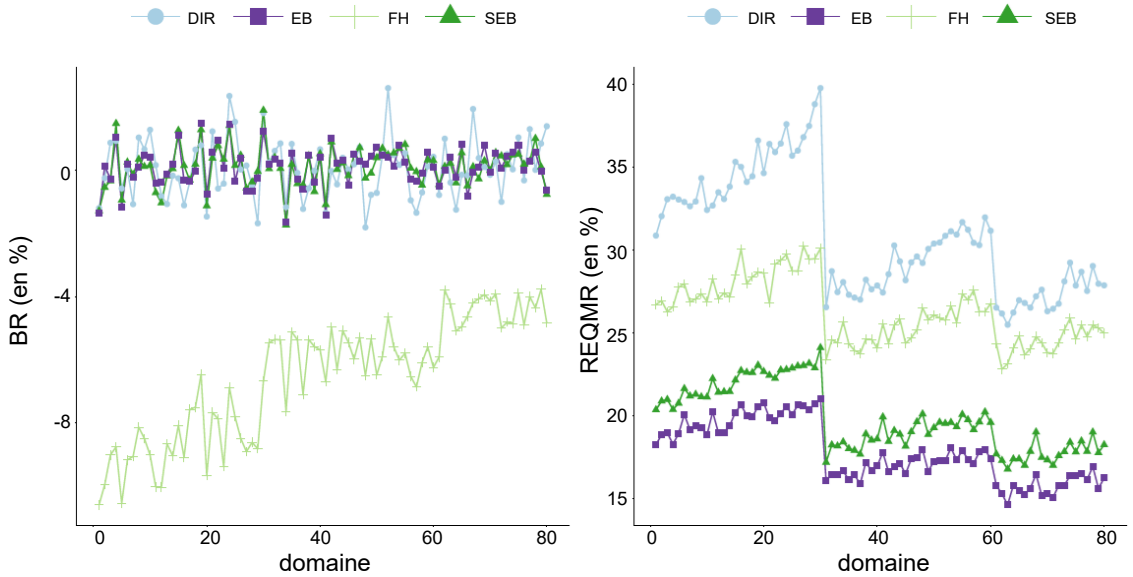
C. Résultats sans l'enquête à plus grande échelle

Figure C.1 BR et REQMR en pourcentage des estimateurs DIR, FH, EB et SEB du taux de pauvreté $F_{0,d}$ pour chaque domaine d, quand $\lambda = 0{,}2$ et $s' = s$



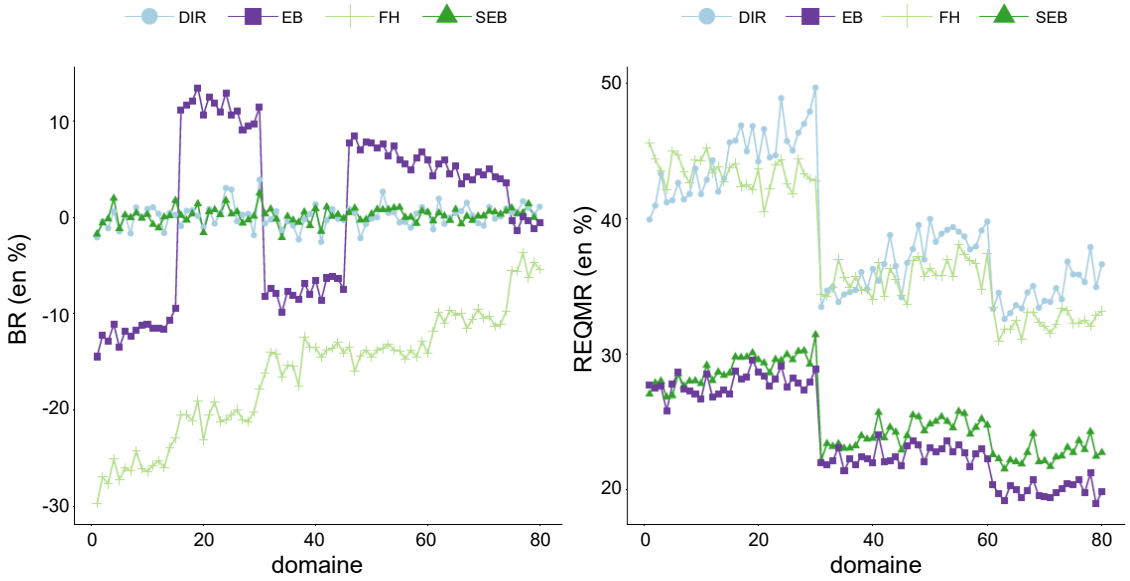
Note : Estimateurs directs (DIR); meilleur prédicteur empirique (EB); estimateur de Fay-Herriot (FH); biais relatif (BR); racine de l'erreur quadratique moyenne relative (REQMR); meilleur prédicteur empirique issu de l'enquête (SEB).

Figure C.2 BR et REQMR en pourcentage des estimateurs DIR, FH, EB et SEB du taux de pauvreté $F_{0,d}$ pour chaque domaine d, quand $\lambda = 0$ et $s' = s$



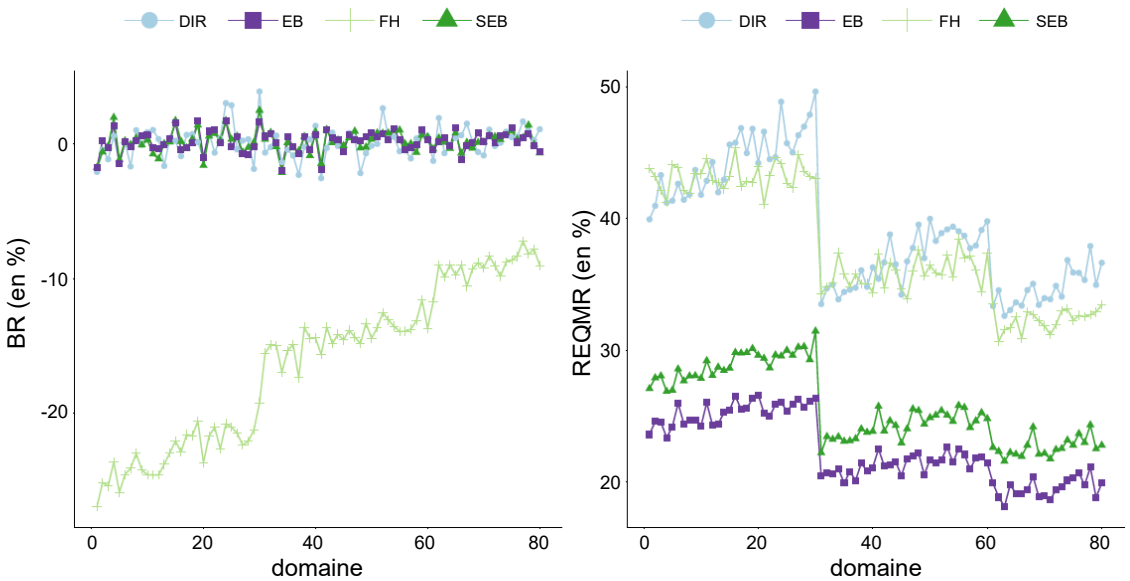
Note : Estimateurs directs (DIR); meilleur prédicteur empirique (EB); estimateur de Fay-Herriot (FH); biais relatif (BR); racine de l'erreur quadratique moyenne relative (REQMR); meilleur prédicteur empirique issu de l'enquête (SEB).

Figure C.3 BR et REQMR en pourcentage des estimateurs DIR, FH, EB et SEB de l'écart de pauvreté $F_{1,d}$ pour chaque domaine d, quand $\lambda = 0{,}2$ et $s' = s$



Note : Estimateurs directs (DIR); meilleur prédicteur empirique (EB); estimateur de Fay-Herriot (FH); biais relatif (BR); racine de l'erreur quadratique moyenne relative (REQMR); meilleur prédicteur empirique issu de l'enquête (SEB).

Figure C.4 BR et REQMR en pourcentage des estimateurs DIR, FH, EB et SEB de l'écart de pauvreté $F_{1,d}$ pour chaque domaine d, quand $\lambda = 0$ et $s' = s$



Note : Estimateurs directs (DIR); meilleur prédicteur empirique (EB); estimateur de Fay-Herriot (FH); biais relatif (BR); racine de l'erreur quadratique moyenne relative (REQMR); meilleur prédicteur empirique issu de l'enquête (SEB).

Tableau C.1
Moyenne sur l'ensemble des domaines du BRA et de la REQMR des estimateurs DIR, FH, EB et SEB du taux de pauvreté, $F_{0,d}$, selon λ, pour $s' = s$

Indicateur	λ (en %)	BRA (en %)				REQMR (en %)			
		$\hat{\delta}_d^{DIR}$	$\hat{\delta}_d^{FH}$	$\hat{\delta}_d^{EB}$	$\hat{\delta}_d^{SEB}$	$\hat{\delta}_d^{DIR}$	$\hat{\delta}_d^{FH}$	$\hat{\delta}_d^{EB}$	$\hat{\delta}_d^{SEB}$
$F_{0,d}$	0	0,73	6,56	0,49	0,49	30,81	26,20	17,77	19,83
	10	0,73	6,53	2,31	0,49	30,81	26,21	17,96	19,83
	20	0,73	6,53	4,50	0,49	30,81	26,25	18,50	19,83
	30	0,73	6,56	6,72	0,49	30,81	26,32	19,34	19,83

Note : Biais relatif absolu (BRA); Estimateurs directs (DIR); meilleur prédicteur empirique (EB); estimateur de Fay-Herriot (FH); racine de l'erreur quadratique moyenne relative (REQMR); meilleur prédicteur empirique issu de l'enquête (SEB).

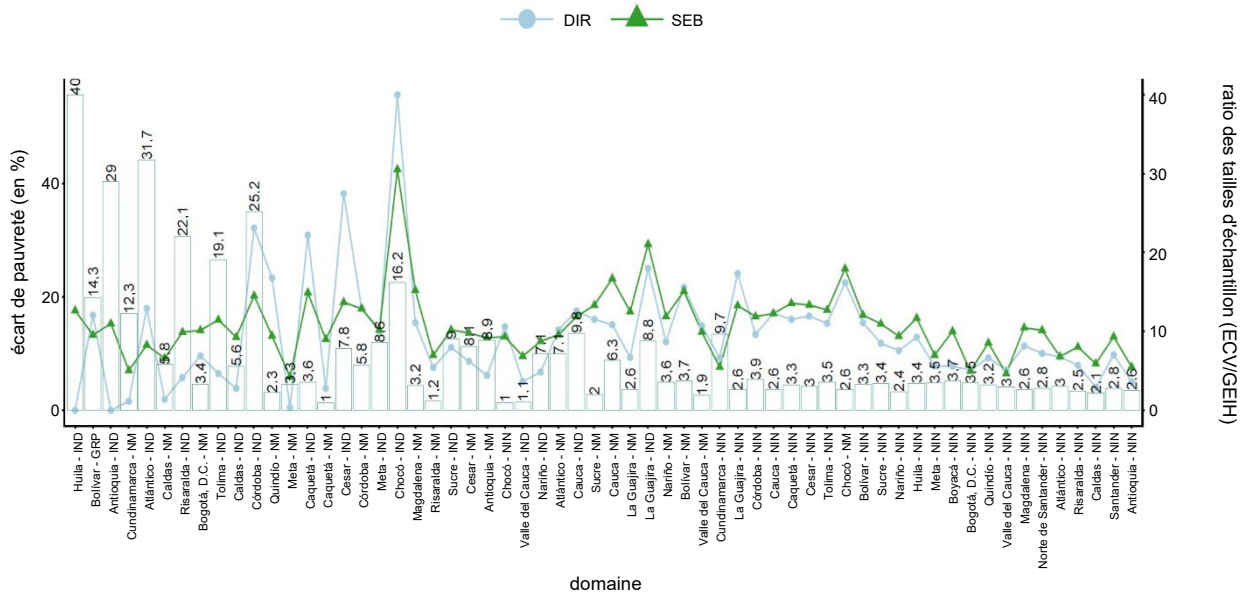
Tableau C.2
Moyenne sur l'ensemble des domaines du BRA et de la REQMR des estimateurs DIR, FH, EB et SEB de l'écart de pauvreté $F_{1,d}$, selon λ, pour $s' = s$

Indicateur	λ (en %)	$\overline{\text{BRA}}$ (%)				$\overline{\text{REQMR}}$ (%)			
		$\hat{\delta}_d^{DIR}$	$\hat{\delta}_d^{FH}$	$\hat{\delta}_d^{EB}$	$\hat{\delta}_d^{SEB}$	$\hat{\delta}_d^{DIR}$	$\hat{\delta}_d^{FH}$	$\hat{\delta}_d^{EB}$	$\hat{\delta}_d^{SEB}$
$F_{1,d}$	0	0,89	16,25	0,62	0,62	39,17	37,73	22,35	25,61
	10	0,89	16,25	3,99	0,62	39,17	37,76	22,76	25,61
	20	0,89	16,25	7,84	0,62	39,17	37,81	23,91	25,61
	30	0,89	16,29	11,71	0,62	39,17	37,88	25,68	25,61

Note : Biais relatif absolu (BRA); Estimateurs directs (DIR); meilleur prédicteur empirique (EB); estimateur de Fay-Herriot (FH); racine de l'erreur quadratique moyenne relative (REQMR); meilleur prédicteur empirique issu de l'enquête (SEB).

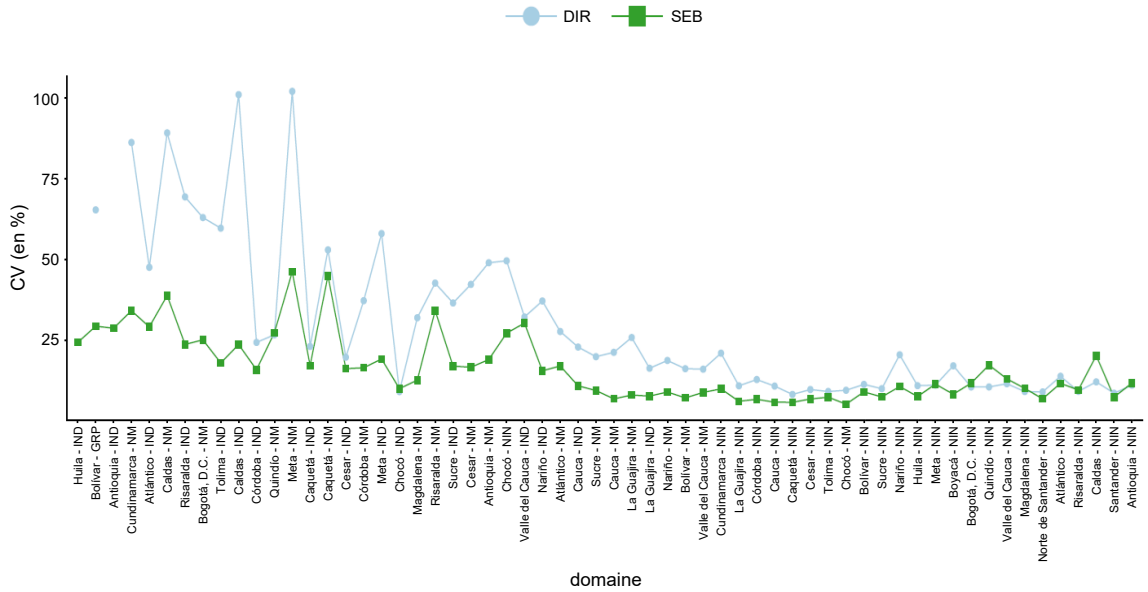
D. Résultats d'applications supplémentaires

Figure D.1 Estimateurs DIR et SEB de l'écart de pauvreté $F_{1,d}$, domaines classés par ordre croissant de la taille d'échantillon de la GEIH n_d



Note : Estimateur direct (DIR); *Encuesta de Calidad de Vida* (ECV); *Gran Encuesta Integrada de Hogares* (GEIH); Gitano/Raizal/Palenquero (GRP); Indigena (IND); Negro/Mulato (NM); meilleur prédicteur empirique issu de l'enquête (SEB).

Figure D.2 CV estimés des estimateurs DIR et SEB de l'écart de pauvreté $F_{1,d}$, domaines classés par ordre croissant de la taille d'échantillon de la GEIH n_d



Note : Coefficient de variation (CV); Estimateur direct (DIR); *Gran Encuesta Integrada de Hogares* (GEIH); Gitano/Raizal/Palenquero (GRP); Indigena (IND); Negro/Mulato (NM); meilleur prédicteur empirique issu de l'enquête (SEB).

Bibliographie

Arias-Salazar, A., Gutiérrez, A., Guerrero-Gómez, S., Mancero, X., Rojas-Perilla, N. et Zhang, H. (2025). Small area estimation for composite indicators: The case of multidimensional poverty incidence. *Journal of Offcial Statistics*, 41(1), 35-59. https://doi.org/10.1177/0282423X241300751.

Baíllo, A. et Molina, I. (2009). Mean-squared errors of small-area estimators under a unit-level multivariate model. *Statistics*, 43(6), 553-569. https://doi.org/10.1080/02331880802605304.

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. https://doi.org/10.1080/01621459.1988.10478561.

Bikauskaite, A., Molina, I. et Morales, D. (2022). Multivariate mixture model for small area estimation of poverty indicators. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(Supplement 2), S724-S755. https://doi.org/10.1111/rssa.12965.

Bugallo, M., Esteban, M.D., Hobza, T., Morales, D. et Pérez, A. (2024). Small area estimation of labour force indicators under unit-level multinomial mixed models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 188(1), 241-270. https://doi.org/10.1093/jrsssa/qnae033.

Chambers, R., Salvati, N. et Tzavidis, N. (2015). Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 179(2), 453-479. https://doi.org/10.1111/rssa.12123.

Chen, Y., Li, P. et Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532), 2011-2021. https://doi.org/10.1080/01621459.2019.1677241.

Cho, Y., Guadarrama-Sanz, M., Molina, I., Eideh, A. et Berg, E. (2024). Optimal predictors of general small area parameters under an informative sample design using parametric sample distribution models. *Journal of Survey Statistics and Methodology*, 12(5), 1430-1463. https://doi.org/10.1093/jssam/smae007.

Corral, P., Himelein, K., McGee, K. et Molina, I. (2021). A map of the poor or a poor map? *Mathematics*, 9(21), 2780. https://doi.org/10.3390/math9212780.

Corral, P., Molina, I. et Nguyen, M. (2020). Pull your small area estimates up by the bootstraps. *World Bank Policy Research Working Paper No. 9256*. Accessible à l'adresse : https://ssrn.com/abstract=3607601.

Cuong, N.V. (2012). A method to update poverty maps. *The Journal of Development Studies*, 48(12), 1844-1863. https://doi.org/10.1080/00220388.2012.682983.

Das, K., Jiang, J. et Rao, J.N.K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, 32(2), 818-840. https://doi.org/10.1214/009053604000000201.

Deville, J.-C. et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376-382. https://doi.org/10.1080/01621459.1992.10475217.

Elbers, C., Lanjouw, J.O. et Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71(1), 355-364. https://doi.org/10.1111/1468-0262.00399.

Fay, R.E. et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366a), 269-277. https://doi.org/10.1080/01621459.1979.10482505.

González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D. et Santamaría, L. (2008). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, 78(5), 443-462. https://doi.org/10.1080/00949650601141811.

Guadarrama, M., Molina, I. et Rao, J.N.K. (2018). Small area estimation of general parameters under complex sampling designs. *Computational Statistics & Data Analysis*, 121, 20-40. https://doi.org/10.1016/j.csda.2017.11.007.

Jiang, J. et Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, 53(2), 217-243. https://doi.org/10.1023/A:1012410420337.

Jiang, J. et Lahiri, P. (2006). Mixed model prediction and small area estimation. *TEST*, 15(1), 1-96. https://doi.org/10.1007/BF02595419.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Masaki, T., Newhouse, D., Silwal, A.R., Bedada, A. et Engstrom, R. (2022). Small area estimation of non-monetary poverty with geospatial data. *Statistical Journal of the IAOS*, 38(3), 1035-1051. https://doi.org/10.3233/SJI-210902.

Molina, I. (2019). *Desagregación de Datos en Encuestas de Hogares: Metodologías de Estimación en Áreas Pequeñas* (Series of the Economic Commission for Latin America and the Caribbean (ECLAC) from United Nations No. Estudios Estadísticos LC/TS.2018/82/Rev.1). CEPAL.

Molina, I. et García-Portugués, E. (2019). *A Short Guide for Small Area Estimation in Household Surveys: Illustration to Poverty Mapping in Palestine with Expenditure Survey and Census Data* (Series of the UN Economic and Social Commission for Western Asia). ESCWA. Accessible à l'adresse : https://www.unescwa.org/node/23858.

Molina, I. et Martín, N. (2018). Empirical best prediction under a nested error model with log transformation. *The Annals of Statistics*, 46(5), 1961-1993. https://doi.org/10.1214/17-AOS1608.

Molina, I. et Rao, J.N.K. (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 38(3), 369-385. https://doi.org/10.1002/cjs.10051.

Molina, I. et Strzalkowska-Kominiak, E. (2019). Estimation of proportions in small areas: Application to the labour force using the Swiss Census Structural Survey. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(1), 281-310. https://doi.org/10.1111/rssa.12498.

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. https://doi.org/10.1198/016214507000001094.

Prasad, N.G.N. et Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85(409), 163-171. https://doi.org/10.1080/01621459.1990.10475320.

Rao, J.N.K., Rubin-Bleuer, S. et Estevao, V.M. (2018). Mesure de l'incertitude associée aux estimateurs pour petits domaines basés sur un modèle. *Techniques d'enquête*, 44(2), 163-180. Accessible à l'adresse : https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2018002/article/54958-fra.pdf.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling* (1st ed.). New York: Springer.

Sen, A. et Lahiri, P. (2025). Estimation of finite population proportions for small areas, a statistical data integration approach. *Journal of Survey Statistics and Methodology*, 13(3), 309-332. https://doi.org/10.1093/jssam/smae049.

Stefan, M. et Hidiroglou, M.A. (2021). Estimation of design-based mean squared error of a small area mean model-based estimator under a nested error linear regression model. *Canadian Journal of Statistics*, 49(4), 1338-1363. https://doi.org/10.1002/cjs.11622.

van der Weide, R. (2014). GLS estimation and empirical bayes prediction for linear mixed models with heteroskedasticity and sampling weights: A background study for the Povmap project. *World Bank Policy Research Working Paper No. 7028*. Accessible à l'adresse : https://ssrn.com/abstract=2495175.

Ybarra, L.M.R. et Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95(4), 919-931. https://doi.org/10.1093/biomet/asn048.