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

Techniques d’enquête

Méthodes d’estimation sur petits domaines avec échantillonnage défini par un seuil d’inclusion

par María Guadarrama, Isabel Molina et Yves Tillé

Date de diffusion : le 30 juin 2020

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

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

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

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

Programme des services de dépôt

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Techniques d'enquête, juin 2020
Vol. 46, nº 1, pp. 53-80
Statistique Canada, nº 12-001-X au catalogue

Méthodes d'estimation sur petits domaines avec échantillonnage défini par un seuil d'inclusion

María Guadarrama, Isabel Molina et Yves Tillé[1]

Résumé

L'échantillonnage défini par un seuil d'inclusion est appliqué quand il est trop coûteux ou difficile d'obtenir les informations requises pour un sous-ensemble d'unités de la population et que, par conséquent, ces unités sont délibérément exclues de la sélection de l'échantillon. Si les unités exclues sont différentes des unités échantillonnées pour ce qui est des caractéristiques d'intérêt, les estimateurs naïfs peuvent être fortement biaisés. Des estimateurs par calage ont été proposés aux fins de réduction du biais sous le plan. Toutefois, dans les estimations sur petits domaines, ils peuvent être inefficaces y compris en l'absence d'échantillonnage défini par un seuil d'inclusion. Les méthodes d'estimation sur petits domaines fondées sur un modèle peuvent servir à réduire le biais causé par l'échantillonnage défini par un seuil d'inclusion si le modèle supposé se vérifie pour l'ensemble de la population. Parallèlement, pour les petits domaines, ces méthodes fournissent des estimateurs plus efficaces que les méthodes de calage. Étant donné qu'on obtient les propriétés fondées sur un modèle en supposant que le modèle se vérifie, mais qu'aucun modèle n'est exactement vrai, nous analysons ici les propriétés de plan des procédures de calage et des procédures fondées sur un modèle pour l'estimation de caractéristiques sur petits domaines sous échantillonnage défini par un seuil d'inclusion. Nos conclusions confirment que les estimateurs fondés sur un modèle réduisent le biais causé par un échantillonnage défini par un seuil d'inclusion et donnent des résultats significativement meilleurs en matière d'erreur quadratique moyenne du plan.

Mots-clés : Estimateurs par calage; échantillonnage défini par un seuil d'inclusion; meilleur prédicteur linéaire sans biais empirique (EBLUP pour *empirical best linear unbiased predictor*); meilleur prédicteur empirique ou estimateur bayésien empirique (MPE ou EBE); modèle à erreurs emboîtées; modèles au niveau de l'unité.

1 Introduction

Haziza, Chauvet et Deville (2010) décrivent l'échantillonnage défini par un seuil d'inclusion comme une technique dans laquelle un ensemble d'unités est délibérément exclu d'une sélection possible dans l'échantillon. Selon l'Organisation de coopération et de développement économiques (OCDE), cette procédure d'échantillonnage consiste à établir un seuil permettant d'exclure de la sélection d'un échantillon toutes les unités situées au-dessus ou en dessous de ce seuil. Selon Särndal, Swensson et Wretman (1992, pages 531-533), cette technique d'échantillonnage est généralement utilisée quand la distribution de la variable étudiée est fortement asymétrique et qu'il n'y a pas de base de sondage fiable couvrant les petits éléments. Benedetti, Bee et Espa (2010) reconnaissent l'avantage apporté par un échantillonnage défini par un seuil d'inclusion pour ce qui est de la réduction des coûts d'une enquête. Cette procédure est souvent utilisée dans les enquêtes-entreprises, pour lesquelles les petites entreprises sont délibérément exclues de l'échantillon en raison de la difficulté à obtenir des renseignements de leur part. Le coût de l'obtention et du maintien d'une base de sondage fiable pour l'ensemble de la population ne compense pas le gain d'exactitude qui s'ensuit.

L'enquête mensuelle sur les industries manufacturières réalisée par Statistique Canada est un exemple d'échantillonnage défini par un seuil d'inclusion (Benedetti et coll., 2010). En Espagne, l'enquête

1. María Guadarrama, Luxembourg Institute of Socio-Economic Research (LISER), 11, Porte des Sciences, Campus Belval L-4366 Esch-sur-Alzette, Luxembourg. Courriel : maria.guadarrama@liser.lu; Isabel Molina, Universidad Carlos III de Madrid, C/Madrid 126, 28903, Getafe, Madrid, Espagne. Courriel : isabel.molina@uc3m.es; Yves Tillé, Institut de Statistique, Université de Neuchâtel, 51, Av. de Bellevaux, 2000 Neuchâtel, Suisse. Courriel : yves.tille@unine.ch.

mensuelle sur l'indice de production industrielle (IPI) réalisée par l'Institut national de statistique espagnol (INE en espagnol) collecte des données auprès d'entreprises qui produisent un volume important de produits d'après l'enquête annuelle sur les produits industriels (EIAP en espagnol), voir INE (2018). Des enquêtes connexes, comme l'enquête sur l'indice des prix industriels (IIP) et l'indice du chiffre d'affaires des entreprises (IBT), utilisent également une forme d'échantillonnage défini par un seuil d'inclusion. Étant donné que les probabilités d'inclusion des unités exclues sont de zéro, cette procédure donne des estimateurs fondés sur le plan de sondage biaisés, voir notamment Särndal et coll. (1992) ou Haziza et coll. (2010). Afin de réduire le biais d'échantillonnage défini par un seuil d'inclusion, Haziza et coll. (2010) proposent d'utiliser de l'information auxiliaire à l'étape du plan ou à celle de l'estimation. Concrètement, ils proposent d'utiliser un échantillonnage équilibré ou un calage.

Dans le travail présenté ici, nous nous limitons à l'étape de l'estimation et nous étudions les effets de l'échantillonnage défini par un seuil d'inclusion sur l'estimation des paramètres de domaine (ou de région). Nous analysons certaines des méthodes de calage que proposent Haziza et coll. (2010) pour diminuer ce problème. Dans le cas de domaines à petite taille d'échantillon (petits domaines ou petites régions), il se peut que les estimateurs par calage soient inefficaces même en l'absence d'échantillonnage défini par un seuil d'inclusion. Pour une plus grande efficacité, nous envisageons des méthodes d'estimation sur petits domaines. Pour l'estimation de paramètres linéaires, nous considérons le meilleur prédicteur linéaire sans biais empirique (EBLUP) et, pour les paramètres non linéaires généraux, nous examinons le meilleur prédicteur empirique/l'estimateur bayésien empirique (MPE ou EBE). Nous appliquons les méthodes étudiées à l'estimation des ventes totales de certains produits du tabac dans des provinces espagnoles.

En l'absence d'échantillonnage défini par un seuil d'inclusion, les estimateurs fondés sur un modèle examinés sont approximativement optimaux quand le modèle se vérifie pour toutes les unités de population. Cependant, comme aucun modèle ne se vérifie exactement, nous nous demanderons si les estimateurs fondés sur un modèle donnent toujours de meilleurs résultats que les estimateurs fondés sur le plan de base (qui ne dépendent pas de modèles) et les estimateurs par calage avec le mécanisme de rééchantillonnage; c'est-à-dire sans hypothèses de modèle et en présence d'un échantillonnage défini par un seuil d'inclusion.

L'article est organisé comme suit. La section 2 décrit le cadre théorique. Les quatre sections qui suivent décrivent les méthodes d'estimation examinées, à savoir les estimateurs directs de base (section 3), les différentes méthodes de calage (section 4), les EBLUP aux fins de l'estimation des paramètres linéaires (section 5) et le MPE/EBE pour l'estimation de paramètres plus généraux dans de petits domaines (section 6). La section 7 décrit une procédure bootstrap permettant d'estimer l'erreur quadratique moyenne des estimateurs pour petits domaines proposés. La section 8 compare, au moyen d'expériences de simulation, l'efficacité de plusieurs estimateurs sur petits domaines selon un échantillonnage défini par un seuil d'inclusion. La section 9 décrit l'application et, enfin, la section 10 présente les conclusions de l'étude.

2 Échantillonnage défini par un seuil d'inclusion pour petits domaines

Nous considérons une population U partitionnée en m sous-ensembles U_i, $i = 1, \ldots, m$, ci-après nommés domaines ou régions, de tailles N_i, $i = 1, \ldots, m$, avec $N = \sum_{i=1}^{m} N_i$. Nous limitons notre étude au cas où les domaines agissent comme des strates d'échantillonnage. Ensuite, on tire des échantillons indépendants des différents domaines, où l'échantillon s_i de taille n_i du domaine i est supposé tiré d'un échantillonnage défini par un seuil d'inclusion, $i = 1, \ldots, m$. Pour cela, on exclut un sous-ensemble d'unités $U_{iE} \subseteq U_i$ de la sélection. En d'autres termes, le domaine U_i est partitionné en deux sous-ensembles, U_{iI} et U_{iE}, de tailles connues N_{iI} et N_{iE} respectivement, avec $N_i = N_{iI} + N_{iE}$. L'ensemble U_{iI} contient les unités qui peuvent être sélectionnées pour l'échantillon, il est appelé ici ensemble des unités incluses, tandis que U_{iE} contient les unités exclues.

Soit y_{ij} la valeur de la variable cible y pour l'unité j^{e} dans le domaine i^{e}. Nous nous intéressons à l'estimation des totaux de domaine $Y_i = \sum_{i=1}^{N_i} y_{ij}$ ou des moyennes $\bar{Y}_i = Y_i / N_i$, $i = 1, \ldots, m$. Selon un échantillonnage défini par un seuil d'inclusion dans chaque domaine, l'échantillon s_i est supposé tiré du sous-ensemble d'individus inclus, U_{iI}, du domaine i. Ensuite, les probabilités d'inclusion des individus inclus $(j \in U_{iI})$ sont $\pi_{j|i} = \Pr(j \in s_i) > 0$ et $w_{j|i} = \pi_{j|i}^{-1}$ sont les poids d'échantillonnage correspondants. Pour les unités exclues $(j \in U_{iE})$, les probabilités d'inclusion sont nulles et, par conséquent, les poids d'échantillonnage correspondants ne sont pas définis. Par conséquent, pour les domaines i avec $U_{iE} \neq \varnothing$, les estimateurs fondés sur le plan de base de Y_i ou $\bar{Y}_i$ sont biaisés et il n'existe pas d'estimateur sans biais par rapport au plan de sondage.

3 Estimateurs directs de base

Nous examinons d'abord les estimateurs directs de base, obtenus uniquement à l'aide des observations n_i de la variable d'intérêt de la région cible. En l'absence d'échantillonnage défini par un seuil d'inclusion, ces estimateurs sont convergents par rapport au plan de sondage à mesure que la taille de l'échantillon du domaine n_i augmente. De plus, ils sont non paramétriques dans le sens qu'ils ne nécessitent aucune hypothèse de modèle. Toutefois, il peut y avoir des erreurs d'échantillonnage inacceptables dans des petits domaines. De plus, comme nous le verrons plus bas, selon un échantillonnage défini par un seuil d'inclusion, leur biais de plan pourrait être important.

L'estimateur par dilatation habituel (Horvitz et Thompson, 1952) de Y_i qu'on obtient en ignorant que l'échantillon s_i est tiré uniquement de U_{iI} est donné par $\hat{Y}_i = \sum_{j \in s_i} w_{ij} y_{ij}$. Selon un échantillonnage défini par un seuil d'inclusion, $\hat{Y}_i$ estime en fait le total dans les strates incluses, $Y_{iI} = \sum_{i \in U_{iI}} y_{ij}$, plutôt que le total global $Y_i = Y_{iI} + Y_{iE}$, où $Y_{iE} = \sum_{i \in U_{iE}} y_{ij}$. En effet, $E_\pi(\hat{Y}_i) = Y_{iI}$, où E_π désigne une espérance dans un échantillonnage répété, puisque les poids d'échantillonnage $w_{j|i} = \pi_{j|i}^{-1}$ dans $\hat{Y}_i$ se dilatent à U_{iI} au lieu de U_i. Personne n'utiliserait cet estimateur, car son biais, $B_\pi(\hat{Y}_i) = E_\pi(\hat{Y}_i) - Y_i = -Y_{iE}$, donné en termes relatifs par la proportion du total représentée par la population exclue, $\mathrm{BR}_\pi(\hat{Y}_i) = -Y_{iE} / Y_i$, peut être important.

Quand on ne dispose pas d'information auxiliaire, il est plus judicieux d'utiliser l'estimateur de Hájek (Hájek, 1971) pour la moyenne $\bar{Y}_i$, donnée par $\hat{\bar{Y}}_i^{\mathrm{HA}} = \hat{Y}_i / \hat{N}_i$, où $\hat{N}_i = \sum_{j \in s_i} w_{ij}$. L'estimateur correspondant pour le total est $\hat{Y}_i^{\mathrm{HA}} = N_i \hat{\bar{Y}}_i^{\mathrm{HA}}$, si l'on considère que les moyennes dans les strates incluses et exclues sont égales. En effet, si on ignore le biais de ratio (d'ordre inférieur) et qu'on note que $E_\pi(\hat{N}_i) = N_{iI}$, le biais de plan asymptotique (en tant que $n_i \to \infty$) de $\hat{Y}_i^{\mathrm{HA}}$ est donné en termes absolus et relatifs par

$$B_\pi\left(\hat{Y}_i^{\mathrm{HA}}\right) \cong N_{iE}\left(\bar{Y}_{iI} - \bar{Y}_{iE}\right), \qquad \mathrm{BR}_\pi\left(\hat{Y}_i^{\mathrm{HA}}\right) \cong \frac{N_{iE}}{N_i} \frac{\bar{Y}_{iI} - \bar{Y}_{iE}}{\bar{Y}_i}, \tag{3.1}$$

où $\bar{Y}_{iI} = Y_{iI} / N_{iI}$ et $\bar{Y}_{iE} = Y_{iE} / N_{iE}$ sont respectivement les véritables moyennes des ensembles d'unités incluses et exclues de la région i (Haziza et coll., 2010). Pour la moyenne, le biais de $\hat{\bar{Y}}_i^{\mathrm{HA}}$ est obtenu en divisant par N_i dans (3.1). Pour un domaine i avec $U_{iE} \neq \varnothing$, le biais ci-dessus disparaît seulement quand $\bar{Y}_{iI} = \bar{Y}_{iE}$, ce qui est improbable dans les cas réels où l'échantillonnage défini par un seuil d'inclusion est appliqué, voir par exemple Haziza et coll. (2010) ou la section 9. Dans la section qui suit, nous décrivons brièvement les techniques de calage comme moyen de réduire le biais de l'échantillonnage défini par un seuil d'inclusion.

Remarque 3.1. L'estimateur de Hájek de $\bar{Y}_i$ est un cas particulier de l'estimateur par le ratio habituel. Dans de nombreuses enquêtes-entreprises mensuelles, les paramètres d'intérêt sont en fait les changements de certains totaux dans le temps, comme $\theta_{it} = Y_i(t)/Y_i(t-1)$, où $Y_i(t)$ est le total de la variable cible au temps t dans le domaine i. Les estimations de la variation par le ratio sont rapportées au lieu des totaux réels, car on croit souvent que ces ratios ne sont pas touchés par le biais d'échantillonnage défini par un seuil d'inclusion. Soit $\hat{\theta}_{it} = \hat{Y}_i(t)/\hat{Y}_i(t-1)$ l'estimateur direct de base de θ_{it}. Comme nous l'avons vu ci-dessus, le biais de l'estimateur par le ratio attribuable à l'échantillonnage défini par un seuil d'inclusion a tendance à être beaucoup plus faible que celui des totaux absolus $\hat{Y}_i(t)$ et $\hat{Y}_i(t-1)$. Cependant, comme nous l'avons également vu, le biais d'échantillonnage défini par un seuil d'inclusion des estimateurs par le ratio disparaît seulement en cas d'hypothèses solides. En effet, si l'on ignore le biais du ratio, qui est négligeable pour les grandes valeurs, n_i, le biais de $\hat{\theta}_{it}$ est donné par

$$B_\pi\left(\hat{\theta}_{it}\right) \cong \frac{Y_{iI}(t)}{Y_{iI}(t-1)} - \frac{Y_i(t)}{Y_i(t-1)},$$

où $Y_{iI}(t)$ désigne le total correspondant uniquement pour les unités incluses. Ce biais est nul seulement si les ratios pour la population $Y_i(t)/Y_i(t-1)$ sont les mêmes que ceux des unités incluses $Y_{iI}(t)/Y_{iI}(t-1)$.

4 Estimateurs par calage

De façon classique, on applique le calage quand on connaît les totaux vrais de certaines variables auxiliaires, susceptibles d'être corrélées à la variable étudiée. L'intention du calage est d'ajuster les poids de sondage $w_{j|i}$ de façon à ce que les estimateurs par dilatation correspondants des totaux vrais

disponibles n'aient aucune erreur. Si les poids ajustés fournissent des estimateurs des totaux disponibles des variables auxiliaires qui ne comportent pas d'erreur, on s'attend à ce qu'ils réduisent également l'erreur dans l'estimation du total de la variable étudiée, à condition qu'il soit linéairement lié aux variables auxiliaires. Même en présence d'un modèle linéaire sous-jacent, en l'absence d'échantillonnage défini par un seuil d'inclusion, les estimateurs par calage sont convergents par rapport au plan de sondage à mesure que la taille d'échantillon de domaine n_i augmente y compris si le modèle ne se vérifie pas. En ce sens, ils sont assistés par un modèle et leurs propriétés sont généralement évaluées dans le cadre de la configuration fondée sur le plan. Toutefois, si les valeurs n_i sont petites, les estimations peuvent souffrir d'un biais de petit échantillon.

Comme nous le verrons ci-dessous, les estimateurs par calage réduisent le biais causé par l'échantillonnage défini par un seuil d'inclusion si le modèle linéaire sous-jacent se vérifie pour l'ensemble de la population (unités incluses et exclues). Toutefois, pour les petits domaines, ils peuvent comporter des erreurs d'échantillonnage d'une ampleur inacceptable, hormis un biais de petit échantillon non négligeable.

Soit $\mathbf{x}_{ij}$ le vecteur des variables auxiliaires pour l'unité j dans le domaine i. Selon qu'on dispose des totaux de domaine ou seulement des totaux de population de ces variables auxiliaires, on peut appliquer différentes méthodes de calage. Tout d'abord, examinons le cas où le vecteur des totaux de domaine $\mathbf{X}_i = \sum_{j=1}^{N_i} \mathbf{x}_{ij}$ est disponible. Notons que $\mathbf{X}_i$ le total dans l'ensemble du domaine $U_i = U_{iI} \cup U_{iE}$. Ensuite, une des méthodes de calage consiste à déterminer les poids de calage $h_{j|i},\ j \in s_i,$ qui minimisent

$$\sum_{j \in s_i} \left(h_{j|i} - w_{j|i}\right)^2 \Big/ w_{j|i} \tag{4.1}$$
$$\text{s.c.} \sum_{j \in s_i} h_{j|i} \mathbf{x}_{ij} = \mathbf{X}_i.$$

Les poids de calage qui en résultent $h_{j|i}$ sont donnés par

$$h_{j|i} = w_{j|i} \left\{1 + \left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right)' \left(\sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} \mathbf{x}'_{ij}\right)^{-1} \mathbf{x}_{ij}\right\},\ j \in s_i, \tag{4.2}$$

sous réserve de la non-singularité de $\sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} \mathbf{x}'_{ij}$. L'estimateur par calage du total du domaine Y_i est ensuite donné par

$$\hat{Y}_i^{\text{LCAL}} = \sum_{j \in s_i} h_{j|i} y_{ij} = \hat{Y}_i + \left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right)' \hat{\mathbf{B}}_i, \tag{4.3}$$

qui est l'estimateur par la régression généralisée (GREG) bien connu de $Y_i,$ où

$$\hat{\mathbf{B}}_i = \left(\sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} \mathbf{x}'_{ij}\right)^{-1} \sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} y_{ij}.$$

L'estimateur de Hájek $\hat{Y}_i^{\text{HA}}$ est un cas particulier de (4.3), avec $\mathbf{x}_{ij} = 1,\ j = 1, \ldots, N_i.$ En l'absence d'échantillonnage défini par un seuil d'inclusion, l'estimateur GREG ci-dessus est convergent par rapport au plan de sondage quand la taille de l'échantillon de domaine n_i augmente, bien qu'il puisse présenter un

biais de petit échantillon. Il réduit la variance si les variables de calage sont corrélées linéairement avec le résultat et que la corrélation est forte. Selon un échantillonnage défini par un seuil d'inclusion, le deuxième terme du deuxième membre de (4.3) corrige le biais de l'estimateur par dilatation de base $\hat{Y}_i$ en tant qu'estimateur de Y_i à l'aide des totaux du domaine connus dans $\mathbf{X}_i$. Toutefois, pour la petite taille d'échantillon de domaine n_i, cette réduction du biais d'échantillonnage défini par un seuil d'inclusion pourrait être transférée sur une augmentation de la variance.

Dans la procédure ci-dessus, un problème de calage différent se pose pour chaque domaine. Dans le cas où l'on dispose seulement de la population totale $\mathbf{X} = \sum_{i=1}^{m} \sum_{j=1}^{N_i} \mathbf{x}_{ij}$, on peut chercher des poids de calage pour tous les domaines simultanément, $g_{j|i}$, $j \in s_i$, $i = 1, \ldots, m$, en résolvant un seul problème de calage :

$$\min_{\{g_{j|i}:\, j \in s_i,\, i = 1, \ldots, m\}} \sum_{i=1}^{m} \sum_{j \in s_i} \left(g_{j|i} - w_{j|i}\right)^2 \Big/ w_{j|i} \tag{4.4}$$

$$\text{s.c.} \sum_{i=1}^{m} \sum_{j \in s_i} g_{j|i} \mathbf{x}_{ij} = \mathbf{X}.$$

Dans ce cas, les poids de calage $g_{j|i}$ sont donnés par

$$g_{j|i} = w_{j|i} \left\{1 + \left(\mathbf{X} - \hat{\mathbf{X}}\right)' \left(\sum_{i=1}^{m} \sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} \mathbf{x}'_{ij}\right)^{-1} \mathbf{x}_{ij}\right\}, \quad j \in s_i,\, i = 1, \ldots, m, \tag{4.5}$$

sous réserve de la non-singularité de $\sum_{i=1}^{m} \sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} \mathbf{x}'_{ij}$. L'estimateur par calage qui en résulte du total de domaine Y_i est ensuite obtenu sous la forme :

$$\hat{Y}_i^{\text{LCALN}} = \sum_{j \in s_i} g_{j|i} y_{ij} = \hat{Y}_i + \left(\mathbf{X} - \hat{\mathbf{X}}\right)' \hat{\mathbf{B}}_i^N, \tag{4.6}$$

où

$$\hat{\mathbf{B}}_i^N = \left(\sum_{\ell=1}^{m} \sum_{j \in s_\ell} w_{j|\ell} \mathbf{x}_{\ell j} \mathbf{x}'_{\ell j}\right)^{-1} \sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} y_{ij}.$$

Contrairement à l'estimateur GREG, la correction de $\hat{Y}_i$ dans $\hat{Y}_i^{\text{LCALN}}$ utilise le total de la population globale $\mathbf{X}$ et l'estimateur par dilatation correspondant.

L'estimateur par calage linéaire LCAL (ou GREG) (4.3) devrait avoir un biais d'échantillonnage défini par un seuil d'inclusion plus petit que (4.6), car il utilise de l'information auxiliaire de chaque domaine particulier i. Par ailleurs, pour les domaines ayant de petites tailles d'échantillon, sa variance (et le biais de petit échantillon) peut être importante puisqu'elle utilise seulement des données propres à un domaine. L'autre estimateur par calage donné dans (4.6) devrait présenter un biais d'échantillonnage défini par un seuil d'inclusion légèrement plus grand parce qu'il utilise seulement de l'information auxiliaire agrégée au niveau national, mais la variance sous le plan devrait être plus petite. Nous étudions ensuite les propriétés de (4.3). À cette fin, nous examinerons la version théorique de l'estimateur LCAL (4.3), donnée par

$$\tilde{Y}_i^{\text{LCAL}} = \hat{Y}_i + \left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right)' \mathbf{B}_{il}. \tag{4.7}$$

Ici, $\mathbf{B}_{iI} = \left(\sum_{j\in U_{iI}} \mathbf{x}_{ij}\mathbf{x}'_{ij}\right)^{-1} \sum_{j\in U_{iI}} \mathbf{x}_{ij} y_{ij}$ est la version du recensement de $\hat{\mathbf{B}}_i$ fondée sur l'ensemble d'unités incluses du domaine i. Notons que l'échantillon s_i est tiré seulement de U_{iI} et par conséquent $\hat{\mathbf{B}}_i$ estime $\mathbf{B}_{iI}$. Nous décomposons le biais de $\hat{Y}_i^{\mathrm{LCAL}}$ comme étant

$$\begin{aligned} B_\pi\left(\hat{Y}_i^{\mathrm{LCAL}}\right) &= E_\pi\left(\hat{Y}_i^{\mathrm{LCAL}} - \tilde{Y}_i^{\mathrm{LCAL}}\right) + B_\pi\left(\tilde{Y}_i^{\mathrm{LCAL}}\right), \\ &= E_\pi\left\{\left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right)'\left(\hat{\mathbf{B}}_i - \mathbf{B}_{iI}\right)\right\} + B_\pi\left(\tilde{Y}_i^{\mathrm{LCAL}}\right). \end{aligned} \tag{4.8}$$

Le terme y tend vers zéro quand $n_i \to \infty$ qu'on applique ou pas un échantillonnage défini par un seuil d'inclusion, étant donné que $\hat{\mathbf{B}}_i$ tend vers $\mathbf{B}_{iI}$. Cependant, pour les petites valeurs n_i ce terme peut ne pas être négligeable, ce qui signifie que l'estimateur LCAL souffre d'un biais de petit échantillon même si $U_{iE} = \varnothing$. En l'absence d'échantillonnage défini par un seuil d'inclusion, le terme du biais $B_\pi\left(\tilde{Y}_i^{\mathrm{LCAL}}\right)$ dans (4.8) est exactement égal à zéro. Selon un échantillonnage défini par un seuil d'inclusion, nous savons que $E_\pi\left(\hat{Y}_i\right) = Y_{iI}$ et $E_\pi\left(\hat{\mathbf{X}}_i\right) = \mathbf{X}_{iI}$, où $\mathbf{X}_{iI} = \sum_{j\in U_{iI}} \mathbf{x}_{ij}$. En notant que $\mathbf{X}_i - \mathbf{X}_{iI} = \mathbf{X}_{iE}$, pour $\mathbf{X}_{iE} = \sum_{j\in U_{iE}} \mathbf{x}_{ij}$, nous obtenons le biais de plan de cet estimateur théorique LCAL, donné en termes absolus et relatifs par

$$B_\pi\left(\tilde{Y}_i^{\mathrm{LCAL}}\right) = -N_{iE}\left(\bar{Y}_{iE} - \bar{\mathbf{X}}'_{iE}\mathbf{B}_{iI}\right), \quad \mathrm{BR}_\pi\left(\tilde{Y}_i^{\mathrm{LCAL}}\right) = -\frac{N_{iE}}{N_i}\,\frac{\bar{Y}_{iE} - \bar{\mathbf{X}}'_{iE}\mathbf{B}_{iI}}{\bar{Y}_i}. \tag{4.9}$$

Ce biais est faible quand le même modèle se vérifie pour les individus inclus et exclus.

Étant donné que l'estimateur par calage $\hat{Y}_i^{\mathrm{LCAL}}$ doit servir à estimer Y_i (et non pas Y_{iI}), pour la moyenne de domaine $\bar{Y}_i = Y_i / N_i$ nous examinons l'estimateur obtenu simplement par la division de y par N_i (au lieu de N_{iI}), $\hat{\bar{Y}}_i^{\mathrm{LCAL}} = \hat{Y}_i^{\mathrm{LCAL}} / N_i$. Le biais asymptotique de $\hat{\bar{Y}}_i^{\mathrm{LCAL}}$ est donné par (4.9) divisé par N_i.

Nous analysons maintenant les propriétés selon le modèle et le mécanisme de rééchantillonnage. Notons que $\hat{\mathbf{B}}_i$ dans l'estimateur GREG est l'estimateur des moindres carrés pondérés du vecteur des coefficients de régression $\boldsymbol{\beta}_i$ dans le modèle de régression linéaire suivant pour les unités du domaine i :

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}_i + \varepsilon_{ij},\ E_m\left(\varepsilon_{ij}\right) = 0,\ E_m\left(\varepsilon_{ij}^2\right) = \sigma_\varepsilon^2, \quad j = 1,\ldots,N_i, \tag{4.10}$$

où les erreurs de modèle ε_{ij} sont toutes mutuellement indépendantes. Nous souhaitons connaître la valeur ajoutée par le modèle aux propriétés du plan des estimateurs, c'est-à-dire que nous voulons savoir quel serait le gain si les données étaient effectivement produites (au moins approximativement) par le modèle supposé. Soit E_m l'espérance sous le modèle (4.10). Si le modèle de régression linéaire (4.10) se vérifie véritablement pour toutes les unités du domaine (incluses et exclues), alors $E_m(\mathbf{B}_{iI}) = \boldsymbol{\beta}_i$ et si nous supposons l'espérance du terme de biais dans (4.9) sous le modèle (4.10), nous obtenons le biais par rapport au plan et au modèle,

$$B_{m,\pi}\left(\tilde{Y}_i^{\mathrm{LCAL}}\right) = -N_{iE}\left\{E_m\left(\bar{Y}_{iE}\right) - \bar{\mathbf{X}}'_{iE}E_m\left(\mathbf{B}_{iI}\right)\right\} = -N_{iE}\left(\bar{\mathbf{X}}'_{iE}\boldsymbol{\beta}_i - \bar{\mathbf{X}}'_{iE}\boldsymbol{\beta}_i\right) = 0. \tag{4.11}$$

En revanche, si l'on suppose exactement le même modèle de régression, le biais de l'estimateur direct de base $\hat{\bar{Y}}_i^{\mathrm{HA}}$ selon un échantillonnage défini par un seuil d'inclusion n'est pas nul, à moins que les moyennes des variables auxiliaires pour les unités exclues et incluses soient égales. En effet,

$$B_{m,\pi}\left(\hat{Y}_i^{\mathrm{HA}}\right) = N_{iE} E_m\left(\bar{Y}_{iI} - \bar{Y}_{iE}\right) = N_{iE}\left(\bar{\mathbf{X}}_{iI} - \bar{\mathbf{X}}_{iE}\right)' \boldsymbol{\beta}_i. \tag{4.12}$$

Ainsi, la condition dans laquelle l'estimateur LCAL est sans biais par rapport au plan, à savoir celle où le modèle linéaire (4.10) se vérifie sans erreur pour toutes les unités du domaine, est beaucoup plus faible que les conditions requises pour que l'estimateur direct de base soit sans biais par rapport au plan. Cela signifie que les estimateurs par calage auront tendance à être moins biaisés que l'estimateur direct de base et peuvent réduire considérablement le biais d'échantillonnage défini par un seuil d'inclusion si le résultat est généré par le modèle de régression linéaire propre au domaine ci-dessus.

Passons maintenant à l'estimateur LCALN (4.6) pour définir la version théorique correspondante

$$\tilde{Y}_i^{\mathrm{LCALN}} = \hat{Y}_i + \left(\mathbf{X} - \hat{\mathbf{X}}\right)' \mathbf{B}_{iI}^N, \tag{4.13}$$

où $\mathbf{B}_i^N$ est la version du recensement pour les unités incluses,

$$\mathbf{B}_i^N = \left(\sum_{\ell=1}^{m} \sum_{j \in U_{\ell I}} \mathbf{x}_{\ell j} \mathbf{x}'_{\ell j}\right)^{-1} \sum_{j \in U_{iI}} \mathbf{x}_{ij} y_{ij}.$$

En décomposant le biais de la même façon que dans (4.8), nous obtenons

$$B_\pi\left(\hat{Y}_i^{\mathrm{LCALN}}\right) = E_\pi\left\{\left(\mathbf{X} - \hat{\mathbf{X}}\right)' \left(\hat{\mathbf{B}}_i^N - \mathbf{B}_{iI}^N\right)\right\} + B_\pi\left(\tilde{Y}_i^{\mathrm{LCALN}}\right). \tag{4.14}$$

Encore une fois, $E_\pi\left\{\left(\mathbf{X} - \hat{\mathbf{X}}\right)\left(\hat{\mathbf{B}}_i^N - \mathbf{B}_{iI}^N\right)\right\}/N_i$ n'est pas nul pour les petites valeurs n_i mais tend vers zéro quand $n_i \to \infty$ y compris selon un échantillonnage défini par un seuil d'inclusion, tandis que $B_\pi\left(\tilde{Y}_i^{\mathrm{LCALN}}\right) = 0$ seulement en l'absence de biais d'échantillonnage défini par un seuil d'inclusion. En général, si l'on utilise la décomposition $\mathbf{X} = \mathbf{X}_I + \mathbf{X}_E$, où $\mathbf{X}_I$ et $\mathbf{X}_E$ sont respectivement les totaux nationaux pour les unités incluses et exclues, le biais de plan de $\tilde{Y}_i^{\mathrm{LCALN}}$ est donné par

$$B_\pi\left(\tilde{Y}_i^{\mathrm{LCALN}}\right) = -\left(Y_{iE} - \mathbf{X}'_E \mathbf{B}_{iI}^N\right). \tag{4.15}$$

Considérons maintenant le modèle linéaire avec des coefficients de régression constante pour toutes les unités de population, qu'on appellera modèle m_2 :

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \varepsilon_{ij},\ E_{m_2}\left(\varepsilon_{ij}\right) = 0,\ E_{m_2}\left(\varepsilon_{ij}^2\right) = \sigma_\varepsilon^2, \quad j = 1, \ldots, N_i,\ i = 1, \ldots, m, \tag{4.16}$$

où, encore une fois, les erreurs de modèle ε_{ij} sont mutuellement indépendantes. Notons que, selon ce modèle, $E_{m_2}\left(\mathbf{B}_{iI}^N\right) \neq \boldsymbol{\beta}$ en général, mais si nous considérons plutôt la somme $\mathbf{B}_I = \sum_{i=1}^{m} \mathbf{B}_{iI}^N$, nous obtenons $E_{m_2}\left(\mathbf{B}_I\right) = \boldsymbol{\beta}$. Cela signifie que l'estimateur LCALN théorique pour un domaine particulier, $\tilde{Y}_i^{\mathrm{LCALN}}$, n'est pas sans biais par rapport au plan et au modèle, parce que

$$B_{m_2,\pi}\left(\tilde{Y}_i^{\mathrm{LCALN}}\right) = -\left\{\mathbf{X}'_{iE}\boldsymbol{\beta} - \mathbf{X}'_E E_{m_2}\left(\mathbf{B}_{iI}^N\right)\right\},$$

n'est pas nécessairement égal à zéro. Cependant, l'estimateur national obtenu par l'addition de ceux des domaines, $\tilde{Y}^{\mathrm{LCALN}} = \sum_{i=1}^m \tilde{Y}_i^{\mathrm{LCALN}} = \hat{Y} + \left(\mathbf{X} - \hat{\mathbf{X}}\right)'\mathbf{B}_I$, est en fait sans biais par rapport au plan et au modèle, parce que

$$B_{m_2,\pi}\left(\tilde{Y}^{\mathrm{LCALN}}\right) = -\left\{\mathbf{X}'_E\boldsymbol{\beta} - \mathbf{X}'_E E_{m_2}\left(\mathbf{B}_I\right)\right\} = 0.$$

Ainsi, selon le modèle (4.16) avec des coefficients de régression constants pour toutes les unités de population, l'estimateur LCALN n'est pas sans biais par rapport au plan et au modèle pour un domaine particulier, mais il est sans biais lors de l'agrégation pour tous les domaines, à condition que le même modèle se vérifie pour les unités incluses et exclues dans tous les domaines. Pour la moyenne $\bar{Y}_i$, le biais de l'estimateur théorique $\tilde{\bar{Y}}_i^{\mathrm{LCALN}} = \tilde{Y}_i^{\mathrm{LCALN}} / N_i$ est donné par (4.15) divisé par N_i.

Étudions maintenant les variances. Pour l'estimateur LCAL théorique (4.7), la variance sous le plan est donnée par

$$V_\pi\left(\tilde{Y}_i^{\mathrm{LCAL}}\right) = V_\pi\left(\hat{Y}_i - \hat{\mathbf{X}}'_i\mathbf{B}_{iI}\right) = V_\pi\left(\sum_{j\in s_i} w_{j|i}E_{ij}\right), \tag{4.17}$$

où $E_{ij} = y_{ij} - \mathbf{x}'_{ij}\mathbf{B}_{iI}$, $j \in U_{iI}$. Nous pouvons ensuite appliquer les estimateurs de la variance habituels pour les estimateurs par dilatation. Dans le cas de l'estimateur LCALN de (4.13), la variance est donnée par

$$V_\pi\left(\tilde{Y}_i^{\mathrm{LCALN}}\right) = V_\pi\left(\hat{Y}_i - \hat{\mathbf{X}}'\mathbf{B}_{iI}^N\right).$$

Notons que $\hat{\mathbf{X}}$ est fondé sur les unités d'échantillon n, tandis que $\hat{\mathbf{X}}_i$ utilise seulement les unités n_i du domaine i. Par conséquent, la contribution de $\hat{\mathbf{X}}$ à la variance de LCALN doit être nettement inférieure à la contribution de $\hat{\mathbf{X}}_i$ dans (4.17). Cela signifie que, dans la mesure où les lignes de régression nationale et de domaine sont semblables, la variance de l'estimateur LCALN, obtenue à partir du calage au niveau national, doit être plus petite que celle de l'estimateur par calage LCAL propre au domaine.

5 EBLUP selon le modèle à erreurs emboîtées

Les estimateurs décrits jusqu'à maintenant utilisent uniquement l'information sur les résultats provenant du domaine. Cela signifie que, quand la taille d'échantillon de domaine n_i est petite, ces estimateurs peuvent être inefficaces y compris en l'absence d'échantillonnage défini par un seuil d'inclusion. Les méthodes d'estimation sur petits domaines (ou indirectes) sont conçues pour réduire la variance en augmentant la taille réelle de l'échantillon. À ce propos, voir le compte rendu exhaustif des méthodes d'estimation sur petits domaines dans Rao et Molina (2015). Dans la présente section, nous nous intéressons aux méthodes fondées sur un modèle, qui fournissent aux estimateurs de bonnes propriétés dans la distribution induite par le modèle. Étant donné que les propriétés fondées sur un modèle sont connues, nous souhaitons analyser si les estimateurs ont de bonnes propriétés sous le mécanisme de rééchantillonnage, qui ne suppose pas que le modèle se vérifie.

Pour cela, nous examinerons un modèle au niveau de l'unité très répandu, qui a été introduit par Battese, Harter et Fuller (1988) et est souvent dit modèle à erreurs emboîtées. Comme pour le modèle m_2 dans (4.16), ce modèle suppose une régression linéaire constante pour toutes les unités de population, mais permet une hétérogénéité inexpliquée entre les domaines en incluant les effets de domaine aléatoires u_i hormis les erreurs de modèle e_{ij}. Ce modèle, le modèle noté m_3, suppose

$$\begin{aligned} y_{ij} &= \mathbf{x}'_{ij}\boldsymbol{\beta} + u_i + e_{ij},\ u_i \overset{\text{iid}}{\sim} N\left(0, \sigma_u^2\right), \\ e_{ij} &\overset{\text{iid}}{\sim} N\left(0, \sigma_e^2\right),\ j = 1, \ldots, N_i,\ i = 1, \ldots, m, \end{aligned} \tag{5.1}$$

où les effets de domaine u_i et les erreurs e_{ij} sont tous mutuellement indépendants. Les vecteurs $\boldsymbol{\beta}$ et $\boldsymbol{\theta} = \left(\sigma_u^2, \sigma_e^2\right)'$ sont inconnus. En établissant $\sigma_u^2 = 0$ dans (5.1), nous obtenons le modèle m_2 donné dans (4.16). Si $\mathbf{y}_i = \left(y_{i1}, \ldots, y_{iN_i}\right)'$ désigne le vecteur des résultats pour le domaine i et $\mathbf{X}_i = \left(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iN_i}\right)'$ la matrice de plan correspondante, le modèle dans la notation de la matrice se lit

$$\mathbf{y}_i \overset{\text{ind}}{\sim} N\left(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i\right),\ \mathbf{V}_i = \sigma_u^2\mathbf{1}_{N_i}\mathbf{1}'_{N_i} + \sigma_e^2\mathbf{I}_{N_i},\ i = 1, \ldots, m, \tag{5.2}$$

où $\mathbf{1}_k$ est un vecteur de uns de taille k et $\mathbf{I}_k$ est la matrice identité $k \times k$.

Nous considérons les paramètres de domaine linéaires définis comme étant $H_i = \mathbf{b}'_i\mathbf{y}_i$, où $\mathbf{b}_i$ est un vecteur non stochastique d'éléments connus. La moyenne de domaine $H_i = \bar{Y}_i = N_i^{-1}\sum_{j=1}^{N_i} y_{ij}$ est obtenue au moyen de $\mathbf{b}_i = N_i^{-1}\mathbf{1}_{N_i}$.

On suppose qu'un échantillon s_i est tiré de l'ensemble des unités incluses dans le domaine i, à savoir $s_i \subset U_{iI}$. Nous désignons par $r_i = (U_{iI} - s_i) \cup U_{iE}$ l'ensemble des unités non échantillonnées du domaine U_i, qui comprend les unités non échantillonnées de U_{iI} et toutes les unités de U_{iE}. Notons que $U_i = s_i \cup r_i = U_{iI} \cup U_{iE}$. Alors, l'échantillon global s est composé des échantillons s_i tirés des ensembles d'unités incluses dans chaque domaine U_{iI}, $i = 1, \ldots, m$, à savoir $s = s_1 \cup \cdots \cup s_m$.

Nous décomposons le vecteur de domaine $\mathbf{y}_i$ et les matrices de plan et de covariance $\mathbf{X}_i$ et $\mathbf{V}_i$ dans les sous-vecteurs et les sous-matrices correspondants pour les unités échantillonnées et non échantillonnées, indiqués par les indices s et r respectivement, comme suit :

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_{is} \\ \mathbf{y}_{ir} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_{is} \\ \mathbf{X}_{ir} \end{pmatrix}, \quad \mathbf{V}_i = \begin{pmatrix} \mathbf{V}_{is} & \mathbf{V}_{isr} \\ \mathbf{V}_{irs} & \mathbf{V}_{ir} \end{pmatrix}.$$

Le paramètre linéaire $H_i = \mathbf{b}'_i\mathbf{y}_i$ peut alors être exprimé comme étant $H_i = \mathbf{b}'_{is}\mathbf{y}_{is} + \mathbf{b}'_{ir}\mathbf{y}_{ir}$. Selon le modèle (5.1), le meilleur prédicteur linéaire sans biais (BLUP) de H est la fonction linéaire sans biais par rapport au modèle des données d'échantillon $\hat{H}_i = \boldsymbol{\alpha}'_{is}\mathbf{y}_{is}$, qui réduit au minimum l'erreur quadratique moyenne (EQM) du modèle, $\text{EQM}_{m_3}\left(\hat{H}_i\right) = E_{m_3}\left(\hat{H}_i - H_i\right)^2$. Le BLUP de $H_i = \mathbf{b}'_{is}\mathbf{y}_{is} + \mathbf{b}'_{ir}\mathbf{y}_{ir}$ est alors

$$\hat{H}_i^{\text{BLUP}}(\boldsymbol{\theta}) = \mathbf{b}'_{is}\mathbf{y}_{is} + \mathbf{b}'_{ir}\left[\mathbf{X}_{ir}\tilde{\boldsymbol{\beta}}_s + \mathbf{V}_{irs}\mathbf{V}_{is}^{-1}\left(\mathbf{y}_{is} - \mathbf{X}'_{is}\tilde{\boldsymbol{\beta}}_s\right)\right], \tag{5.3}$$

où $\tilde{\boldsymbol{\beta}}_s$ est l'estimateur des moindres carrés pondérés de $\boldsymbol{\beta}$, donné par

$$\tilde{\boldsymbol{\beta}}_s = \tilde{\boldsymbol{\beta}}_s(\boldsymbol{\theta}) = \left(\sum_{i=1}^{m} \mathbf{X}_{is}' \mathbf{V}_{is}^{-1} \mathbf{X}_{is}\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_{is}' \mathbf{V}_{is}^{-1} \mathbf{y}_{is}. \tag{5.4}$$

Le BLUP de H_i donné dans (5.3) dépend des vraies valeurs des composantes de variance $\boldsymbol{\theta} = (\sigma_u^2, \sigma_e^2)'$, qui sont généralement inconnues. En les remplaçant par des estimateurs convergents par rapport au modèle correspondants $\hat{\boldsymbol{\theta}} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)'$, nous obtenons le BLUP dit empirique (EBLUP), noté $\hat{H}_i^{\text{EBLUP}} = \hat{H}_i^{\text{BLUP}}(\hat{\boldsymbol{\theta}})$.

Si la fraction de sondage du domaine, n_i / N_i, est négligeable, le BLUP de $\bar{Y}_i$ peut être exprimé comme étant la moyenne pondérée

$$\hat{\bar{Y}}_i^{\text{BLUP}} \cong \gamma_{is} \left[\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})' \tilde{\boldsymbol{\beta}}_s\right] + (1 - \gamma_{is}) \bar{\mathbf{X}}_i' \tilde{\boldsymbol{\beta}}_s, \tag{5.5}$$

où $\gamma_{is} = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2 / n_i)$ est dans l'intervalle $(0, 1)$ et tend vers 1 quand $n_i \to \infty$ (Rao et Molina, 2015). Par conséquent, pour les domaines ayant une grande taille d'échantillon n_i, $\hat{\bar{Y}}_i^{\text{BLUP}}$ s'approche de l'estimateur par la régression de l'enquête $\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})' \tilde{\boldsymbol{\beta}}_s$, tandis que pour les domaines ayant une petite taille d'échantillon n_i, $\hat{\bar{Y}}_i^{\text{BLUP}}$ emprunte de l'information des autres domaines en s'approchant de l'estimateur synthétique de type régression $\bar{\mathbf{X}}_i' \tilde{\boldsymbol{\beta}}_s$. En remplaçant les composantes de variance dans $\boldsymbol{\theta} = (\sigma_u^2, \sigma_e^2)'$ par des estimateurs convergents $\hat{\boldsymbol{\theta}} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)'$ dans le BLUP, désignant $\hat{\gamma}_{is} = \hat{\sigma}_u^2 / (\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_i)$ et $\hat{\boldsymbol{\beta}}_s = \tilde{\boldsymbol{\beta}}_s(\hat{\boldsymbol{\theta}})$, nous obtenons l'EBLUP de $\bar{Y}_i$, donné par

$$\hat{\bar{Y}}_i^{\text{EBLUP}} \cong \hat{\gamma}_{is} \left[\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})' \hat{\boldsymbol{\beta}}_s\right] + (1 - \hat{\gamma}_{is}) \bar{\mathbf{X}}_i' \hat{\boldsymbol{\beta}}_s. \tag{5.6}$$

Le BLUP est sans biais et optimal selon le modèle m_3 dans le sens qu'il minimise l'EQM selon ce modèle. Nous étudions maintenant ses propriétés de plan, qui ne supposent pas que le modèle est correct et qui tiennent par conséquent compte du biais des écarts par rapport au modèle. À cette fin, nous considérons le paramètre de régression du recensement pour les unités incluses, défini comme $\mathbf{B}_I = \left(\sum_{i=1}^{m} \mathbf{X}_{iI}' \mathbf{V}_{iI}^{-1} \mathbf{X}_{iI}\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_{iI}' \mathbf{V}_{iI}^{-1} \mathbf{y}_{iI}$, où $\mathbf{y}_{iI}$, $\mathbf{X}_{iI}$ et $\mathbf{V}_{iI}$ sont le sous-vecteur et les sous-matrices correspondants de $\mathbf{y}_i$, $\mathbf{X}_i$ et $\mathbf{V}_i$, pour les unités incluses $(j \in U_{iI})$. Encore une fois, nous considérons la version théorique du BLUP définie sous la forme $\mathbf{B}_I$,

$$\tilde{\bar{Y}}_i^{\text{BLUP}} = \gamma_{is} \left[\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})' \mathbf{B}_I\right] + (1 - \gamma_{is}) \bar{\mathbf{X}}_i' \mathbf{B}_I. \tag{5.7}$$

Si chaque échantillon s_i est tiré du domaine correspondant U_{iI} par échantillonnage aléatoire simple sans remise (EASSR), alors $E_\pi(\bar{y}_{is}) = \bar{Y}_{iI}$ et $E_\pi(\bar{\mathbf{x}}_{is}) = \bar{\mathbf{X}}_{iI}$. À partir de ces faits, on peut facilement calculer le biais de plan $\tilde{\bar{Y}}_i^{\text{BLUP}}$ selon un EASSR, qui est donné par

$$B_\pi\left(\tilde{\bar{Y}}_i^{\text{BLUP}}\right) = \gamma_{is} \frac{N_{iE}}{N_{iI}} \left[\left(\bar{Y}_i - \bar{\mathbf{X}}_i' \mathbf{B}_I\right) - \left(\bar{Y}_{iE} - \bar{\mathbf{X}}_{iE}' \mathbf{B}_I\right)\right] + (1 - \gamma_{is})\left(\bar{\mathbf{X}}_i' \mathbf{B}_I - \bar{Y}_i\right). \tag{5.8}$$

Ce biais sera faible si (5.1) se vérifie pour l'ensemble de la population, et dans ce cas $E_{m_3}(\bar{Y}_i) = \bar{\mathbf{X}}_i \boldsymbol{\beta}$ et $E_{m_3}(\bar{Y}_{iE}) = \bar{\mathbf{X}}_{iE} \boldsymbol{\beta}$. En utilisant ces résultats quand nous prenons l'espérance selon le modèle m_3 dans (5.8), nous obtenons $B_{m_3,\pi}\left(\tilde{\bar{Y}}_i^{\text{BLUP}}\right) = 0$. En fait, ce résultat se vérifie aussi selon le modèle m_2.

En ce qui concerne la variance, si s_i est obtenu par EASSR dans U_{iI}, la variance sous le plan de l'estimateur BLUP théorique est donnée par

$$V_\pi\left(\tilde{\bar{Y}}_i^{\mathrm{BLUP}}\right) = \gamma_{is}^2 V_\pi\left(\bar{y}_{is} - \bar{\mathbf{x}}_{is}\mathbf{B}_I\right) = \frac{\gamma_{is}^2}{N_i^2} V_\pi\left(\hat{Y}_i - \hat{\mathbf{X}}_i'\mathbf{B}_I\right).$$

Par conséquent, si les droites de régression par les moindres carrés (MC) du recensement pour les domaines du modèle (4.10) sont similaires à la droite de régression par les moindres carrés pondérés (MCP) du modèle (5.1), c'est-à-dire si $\mathbf{B}_I \approx \mathbf{B}_{iI}$, alors la variance du BLUP de $\bar{Y}_i$ diminue jusqu'à celle de l'estimateur LCAL de $\bar{Y}_i$ obtenu à partir de (4.17), multipliée par le facteur $\gamma_{is}^2 \in (0, 1)$.

Selon des plans d'échantillonnage plus généraux dans U_{iI}, nous considérons le meilleur prédicteur linéaire sans biais pseudo-empirique (pseudo-EBLUP) de $\bar{Y}_i$ proposé par You et Rao (2002) au lieu de l'EBLUP. En définissant l'estimateur théorique analogue qui utilise les moyennes de l'échantillon pondérées $\bar{y}_{iw} = \left(\sum_{j\in s_i} w_{j|i}\right)^{-1} \sum_{j\in s_i} w_{j|i} y_{ij}$ et $\bar{\mathbf{x}}_{iw} = \left(\sum_{j\in s_i} w_{j|i}\right)^{-1} \sum_{j\in s_i} w_{j|i}\mathbf{x}_{ij}$ au lieu des moyennes non pondérées $\bar{y}_{is}$ et $\bar{\mathbf{x}}_{is}$ dans (5.7), nous obtenons les mêmes expressions pour le biais par rapport au plan et la variance, avec γ_{is} changé en $\gamma_{iw} = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2 \delta_{iw})$, pour $\delta_{iw} = \left(\sum_{j\in s_i} w_{j|i}\right)^{-2} \sum_{j\in s_i} w_{j|i}^2$.

6 Meilleur prédicteur empirique selon le modèle à erreurs emboîtées

L'estimation des paramètres de domaine non linéaires nécessite des méthodes plus générales d'estimation sur petits domaines, comme le meilleur prédicteur/l'estimateur bayésien (MP ou EBE). À ce propos, voir Molina et Rao (2010). Les paramètres non linéaires particuliers sont les indicateurs de pauvreté et d'inégalité définis selon une mesure du bien-être, comme la famille des indicateurs de pauvreté introduits par Foster, Greer et Thorbecke (1984). Le meilleur prédicteur peut aussi servir à estimer d'autres caractéristiques comme la médiane ou les quantiles, ou encore toute la fonction de distribution empirique de la variable d'intérêt, voir Pratesi (2016). De plus, on peut l'utiliser pour estimer les totaux et les moyennes d'une variable cible donnée, quand la variable dépendante dans le modèle considéré est une transformation bijective (par exemple, des transformations logarithmiques ou de type Box-Cox plus générales) de cette variable cible. Ces transformations sont généralement appliquées en cas de non-normalité ou d'hétéroscédasticité.

Dans la présente section, la variable cible (par exemple, la mesure du bien-être) pour l'unité j^{e} dans le domaine i^{e} est notée comme étant v_{ij} et $y_{ij} = T\left(v_{ij}\right)$, où T est une transformation bijective. Nous supposons que y_{ij} suit le modèle à erreurs emboîtées (5.1). Par la transformation inverse $v_{ij} = T^{-1}\left(y_{ij}\right)$, nous pouvons exprimer notre paramètre cible (défini à l'origine en termes de variables cibles v_{ij}) comme une fonction du vecteur $\mathbf{y}_i = \left(y_{i1}, \ldots, y_{iN_i}\right)'$ des réponses du modèle pour les unités de domaine, $H_i = h\left(\mathbf{y}_i\right)$. Le meilleur prédicteur (MP) de $H_i = h\left(\mathbf{y}_i\right)$ est défini comme la fonction des données d'échantillon $\mathbf{y}_{is}$ qui minimise l'EQM du modèle, et qui est

$$\hat{H}_i^{\mathrm{MP}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = E_{m_3}\left[h\left(\mathbf{y}_i\right) \middle| \mathbf{y}_{is}; \boldsymbol{\beta}, \boldsymbol{\theta}\right], \tag{6.1}$$

où l'espérance est prise par rapport à la distribution du modèle de $\mathbf{y}_{ir} \mid \mathbf{y}_{is}$, qui dépend des valeurs vraies de $\boldsymbol{\beta}$ et $\boldsymbol{\theta}$. Le MP de H_i est sans biais par rapport au modèle (5.1), quelle que soit la complexité de la fonction $h(\cdot)$ définissant le paramètre cible. Toutefois, il ne peut pas être calculé en pratique, car les paramètres du modèle $\boldsymbol{\beta}$ et $\boldsymbol{\theta}$ sont généralement inconnus. Un meilleur prédicteur empirique (MPE) de H_i, noté $\hat{H}_i^{\mathrm{MPE}}$, est ensuite obtenu par le remplacement de $\boldsymbol{\beta}$ et $\boldsymbol{\theta}$ dans $\hat{H}_i^{\mathrm{MP}}(\boldsymbol{\beta}, \boldsymbol{\theta})$ par les estimateurs convergents $\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{\theta}}$, quand $\hat{H}_i^{\mathrm{MPE}} = \hat{H}_i^{\mathrm{MP}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$. Le MPE n'est pas exactement sans biais, mais le biais découlant de l'estimation de $\boldsymbol{\beta}$ et $\boldsymbol{\theta}$ est généralement négligeable quand la taille globale de l'échantillon n est grande. Dans le cas d'un paramètre linéaire $H_i = \mathbf{b}_i' \mathbf{y}_i$, le MPE selon le modèle à erreurs emboîtées avec normalité obtenu au moyen de $\hat{\boldsymbol{\beta}}_s = \tilde{\boldsymbol{\beta}}_s(\hat{\boldsymbol{\theta}})$ pour estimer $\boldsymbol{\beta}$ est égal à $\hat{H}_i^{\mathrm{EBLUP}}$.

Quand $h(\cdot)$ est si complexe que l'espérance définissant le MPE dans (6.1) ne peut pas être calculée analytiquement, on peut appliquer les méthodes Monte-Carlo pour obtenir une approximation de $\hat{H}_i^{\mathrm{MPE}}$ comme le proposent Molina et Rao (2010). Pour cela, on simule, à partir du modèle (5.1) ajusté aux données d'échantillon originales, des répliques L $y_{ij}^{(\ell)}$; $\ell = 1, \ldots, L$ de y_{ij}, $j \in r_i$, où r_i sont les unités non échantillonnées du domaine i, en attribuant les éléments d'échantillon y_{ij}, $j \in s_i$ pour former le vecteur de population $\mathbf{y}_i^{(\ell)}$, en calculant le paramètre cible correspondant $H_i^{(\ell)} = h(\mathbf{y}_i^{(\ell)})$ pour chaque $\ell = 1, \ldots, L$ et, enfin, en établissant la moyenne des répliques L sous la forme $\hat{H}_i^{\mathrm{MPE}} = L^{-1} \sum_{\ell=1}^{L} H_i^{(\ell)}$. Il faut noter que le MPE nécessite les valeurs $\mathbf{x}_{ij}$ pour toutes les unités de population, et non pas seulement pour les unités incluses. Pour en savoir plus, voir Molina et Rao (2010).

7 Estimation de l'EQM

L'EBLUP de la section 5 ou le MPE décrit à la section 6 sont fondés sur le modèle à erreurs emboîtées (5.1). Les estimateurs par calage décrits à la section 4 sont également assistés par un modèle de régression linéaire. Si nous voulons avoir des mesures d'exactitude comparables, il semble raisonnable d'obtenir les EQM de tous les estimateurs selon un modèle de régression donné (EQM de modèle), en supposant que le modèle se vérifie pour toutes les unités de population (incluses et exclues). Nous estimons ici l'EQM du modèle au moyen de la méthode bootstrap proposée dans Molina et Rao (2010), fondée sur la méthode bootstrap paramétrique originale pour les populations finies de González-Manteiga, Lombardia, Molina, Morales et Santamaría (2008). Dans cette procédure, l'EQM par la méthode bootstrap de $\hat{H}_i^{\mathrm{MPE}}$ selon le modèle à erreurs emboîtées (5.1) est obtenue comme suit : (i) on ajuste le modèle (5.1) aux données d'échantillon $\{(\mathbf{y}_{is}, \mathbf{X}_{is});\ i = 1, \ldots, m\}$, pour obtenir les estimateurs $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ et $\hat{\sigma}_e^2$ de $\boldsymbol{\beta}$, σ_u^2 et σ_e^2 respectivement; (ii) pour $b = 1, \ldots, B$, on produit indépendamment $u_i^{*(b)} \overset{\mathrm{iid}}{\sim} N(0, \hat{\sigma}_u^2)$ et $e_{ij}^{*(b)} \overset{\mathrm{iid}}{\sim} N(0, \hat{\sigma}_e^2)$, $j = 1, \ldots, N_i$, $i = 1, \ldots, m$; (iii) pour $b = 1, \ldots, B$, on construit des vecteurs bootstrap de domaine $\mathbf{y}_i^{*(b)} = (y_{i1}^{*(b)}, \ldots, y_{iN_1}^{*(b)})'$, dont les éléments sont générés en tant que

$$y_{ij}^{*(b)} = \mathbf{x}_{ij}' \hat{\boldsymbol{\beta}} + u_i^{*(b)} + e_{ij}^{*(b)}, \qquad j = 1, \ldots, N_i,\ i = 1, \ldots, m.$$

à partir du vecteur bootstrap de domaine $\mathbf{y}_i^{*(b)}$, on calcule le paramètre bootstrap cible $H_i^{*(b)} = h(\mathbf{y}_i^{*(b)})$, pour $b = 1, \ldots, B$; (iv) à partir de chaque vecteur bootstrap de population $\mathbf{y}_i^{*(b)}$, on prend la partie de l'échantillon $\mathbf{y}_{is}^{*(b)}$, où les indices d'échantillon s_i sont exactement ceux de l'échantillon original tiré de

U_{iI}, pour $i = 1, \ldots, m$. À l'aide des données de l'échantillon bootstrap global $\mathbf{y}_s^{*(b)} = \left(\mathbf{y}_{1s}^{*(b)}, \ldots, \mathbf{y}_{ms}^{*(b)}\right)'$ et des vecteurs de population $\mathbf{x}_{ij}$, $j = 1, \ldots, N_i$, supposés connus pour toutes les unités de population, on calcule l'EBE bootstrap de H_i, noté $\hat{H}_i^{\mathrm{MPE}*(b)}$, $b = 1, \ldots, B$; (v) un estimateur de l'EQM bootstrap pour l'EBE selon le modèle (5.1), $\mathrm{EQM}_{m_3}\left(\hat{H}_i^{\mathrm{MPE}}\right)$, est obtenu sous la forme

$$\mathrm{eqm}_B\left(\hat{H}_i^{\mathrm{MPE}}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{H}_i^{\mathrm{MPE}*(b)} - H_i^{*(b)}\right)^2. \tag{7.1}$$

On peut obtenir de la même façon les estimateurs bootstrap de l'EQM selon le même modèle d'estimateurs par calage. Dans le cas particulier d'un paramètre linéaire, $H_i = \mathbf{b}_i'\mathbf{y}_i$, si $\hat{\boldsymbol{\beta}}_s$ est l'estimateur des moindres carrés pondérés (5.4), alors (7.1) est un estimateur de $\mathrm{EQM}_{m_3}\left(\hat{H}_i^{\mathrm{EBLUP}}\right)$. Cet estimateur bootstrap naïf de l'EQM du modèle est sans biais au premier ordre, dans le sens que son biais de modèle est $O\left(m^{-1}\right)$, et non pas $o\left(m^{-1}\right)$. Les corrections de biais décrites dans la littérature augmentent la variance et peuvent produire des estimations de l'EQM négatives. En effet, on ne trouve pas dans la littérature d'estimateurs bootstrap de l'EQM qui soient à la fois strictement positifs et sans biais au deuxième ordre. C'est pourquoi, par souci de simplicité, nous considérons l'estimateur bootstrap naïf (7.1), qui ne peut pas produire de valeurs négatives et qui a de bonnes performances pour un nombre moyen de régions m.

8 Expériences de simulation

8.1 Objectifs et description générale

Dans la présente section, nous décrirons des expériences de simulation conçues pour comparer les propriétés des petits échantillons des estimateurs de $\bar{Y}_i$ présentés plus haut dans le contexte de l'échantillonnage défini par un seuil d'inclusion. Plus précisément, nous comparons l'estimateur direct naïf $\hat{\bar{Y}}_i^{\mathrm{HA}}$, les estimateurs par calage $\hat{\bar{Y}}_i^{\mathrm{LCAL}}$ et $\hat{\bar{Y}}_i^{\mathrm{LCALN}}$, et l'EBLUP selon le modèle à erreurs emboîtées $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$, dans deux scénarios différents. Dans le premier scénario, les valeurs de la variable cible pour toutes les unités de population sont générées à partir du même modèle. Dans le deuxième, les unités incluses et exclues sont générées à partir de modèles différents.

En l'absence d'échantillonnage défini par un seuil d'inclusion, les estimateurs par calage sont convergents par rapport au plan de sondage à mesure que la taille du domaine n_i augmente même si le modèle correspondant ne se vérifie pas, mais cela n'est pas le cas pour les estimateurs fondés sur un modèle. D'autre part, selon le modèle correspondant, l'EBLUP d'un paramètre linéaire est approximativement l'estimateur linéaire et sans biais le plus efficace, de sorte que la réalisation de simulations selon un modèle n'apporterait pas de nouvelles connaissances. L'objectif ici est de déterminer si les prédicteurs fondés sur un modèle ont également de bonnes performances pour ce qui est du plan (d'échantillonnage défini par un seuil d'inclusion). C'est pourquoi, nous exécutons des simulations fondées sur le plan de sondage en générant un vecteur de population $\mathbf{y} = \left(\mathbf{y}_1', \ldots, \mathbf{y}_m'\right)'$ à partir du modèle à erreurs emboîtées dans (5.1), en le maintenant fixe et en tirant à répétition un nouvel échantillon défini

par un seuil d'inclusion dans chaque simulation Monte-Carlo. On répartit les unités aux ensembles d'unités incluses ou exclues en générant une variable binaire aléatoire c_{ij} pour chaque unité $j = 1, \ldots, N_i$ et chaque domaine $i = 1, \ldots, m$. Les unités j avec $c_{ij} = 1$ sont attribuées à U_{iI} et celles avec $c_{ij} = 0$ sont attribuées à U_{iE}. Dans chaque répétition Monte-Carlo, des échantillons sont tirés, indépendamment pour chaque domaine i, à partir des unités U_{iI} $i = 1, \ldots, m$.

8.2 Modèle de régression commun

Nous considérons une population de $N = 20\,000$ individus divisés en $m = 80$ domaines d'une même taille $N_i = 250$, $i = 1, \ldots, m$. Nous considérons trois variables auxiliaires, avec des valeurs générées sous la forme $x_{ij\kappa} \overset{\text{iid}}{\sim} N(3, 2)$, $\kappa = 1, 2, 3$. Les variables binaires c_{ij} déterminant la répartition des unités dans U_{iI} ou U_{iE} pour chaque domaine i sont générées indépendamment en tant que $c_{ij} \overset{\text{ind}}{\sim} \text{Bern}\left(p_{j|i}\right)$, où les probabilités $p_{j|i} = \Pr\left(c_{ij} = 1\right)$ sont liées au vecteur des variables auxiliaires $\mathbf{x}_{ij} = \left(x_{ij1}, x_{ij2}, x_{ij3}\right)'$ sous la forme

$$p_{j|i} = \frac{\exp\left(\mathbf{x}_{ij}'\boldsymbol{\zeta}\right)}{1 + \exp\left(\mathbf{x}_{ij}'\boldsymbol{\zeta}\right)}, \quad j = 1, \ldots, N_i, \quad i = 1, \ldots, m.$$

Nous prenons $\boldsymbol{\zeta} = (0{,}75;\ 1;\ 1)'$. À partir de cette valeur, le nombre total d'unités incluses (avec $c_{ij} = 1$) de tous les domaines représente environ la moitié de la population.

On génère les valeurs de la variable cible y_{ij} à partir du modèle à erreurs emboîtées (5.1) au moyen de $\mathbf{x}_{ij} = \left(x_{ij1}, x_{ij2}, x_{ij3}\right)'$ et en prenant $\boldsymbol{\beta} = (1;\ 1{,}5;\ 1)'$, $\sigma_u^2 = (0{,}75)^2$ et $\sigma_e^2 = 4^2$, ce qui donne un coefficient de détermination $R^2 \approx 0{,}5$. Alors, si l'on garde les valeurs de population $\left\{\left(\mathbf{x}_{ij}, y_{ij}, c_{ij}\right);\ j = 1, \ldots, N_i,\ i = 1, \ldots, m\right\}$ fixes, on tire $K = 1\,000$ échantillons Monte-Carlo $s^{(k)}$, $k = 1, \ldots, K$. On obtient chacun de ces échantillons en tirant des sous-échantillons indépendants $s_i^{(k)}$ de taille n_i des unités dans U_{iI} par échantillonnage aléatoire simple sans remise, $i = 1, \ldots, m$. On suppose les tailles d'échantillon de domaine $n_i \in \{5;\ 10;\ 30;\ 50\}$, chaque taille d'échantillon étant répétée pour 20 domaines subséquents. Au moyen des données de l'échantillon k^{e} nous calculons l'estimateur direct de base, les estimateurs par calage au niveau du domaine (LCAL) et au niveau de la population (LCALN), ainsi que l'EBLUP. Les poids, $h_{j|i}$ et $g_{j|i}$, dans les estimateurs par calage (4.3) et (4.6) respectivement sont obtenus au moyen de la fonction calib du module `sampling` (Tillé et Matei, 2016) de R (R Development Core Team, 2016). Les EBLUP sont obtenus au moyen du module de R `sae` (Molina et Marhuenda, 2015), qui, par défaut, estime les paramètres du modèle σ_u^2, σ_e^2 et $\boldsymbol{\beta}$ au moyen du maximum de vraisemblance restreint (ou REML, pour *restricted maximum likelihood*).

Supposons que $\hat{\bar{Y}}_i$ est un estimateur générique de $\bar{Y}_i$ et $\hat{\bar{Y}}_i^{(k)}$ sa valeur obtenue avec l'échantillon k^{e}. Nous évaluons les performances des estimateurs en termes de biais relatif (BR) et de racine carrée de l'EQM relative (REQMR) selon le plan, dont on obtient une approximation empirique comme suit

$$\text{BR}_\pi\left(\hat{\bar{Y}}_i\right) = 100\,\frac{K^{-1}\sum_{k=1}^{K}\left(\hat{\bar{Y}}_i^{(k)} - \bar{Y}_i\right)}{\bar{Y}_i}, \quad \text{REQMR}_\pi\left(\hat{\bar{Y}}_i\right) = 100\,\frac{\sqrt{K^{-1}\sum_{k=1}^{K}\left(\hat{\bar{Y}}_i^{(k)} - \bar{Y}_i\right)^2}}{\bar{Y}_i}.$$

Nous calculons en outre les moyennes sur les domaines du BR absolu et de la REQMR comme suit

$$\overline{\text{BRA}} = m^{-1}\sum_{i=1}^{m}\left|\text{BR}_{\pi}\left(\hat{\bar{Y}}_i\right)\right|, \quad \overline{\text{REQMR}} = m^{-1}\sum_{i=1}^{m}\text{REQMR}_{\pi}\left(\hat{\bar{Y}}_i\right).$$

La figure 8.1 présente des diagrammes de quartiles du pourcentage de BR pour les estimateurs de la moyenne considérés $\bar{Y}_i$, où chaque diagramme de quartile correspond aux 20 domaines de chaque groupe des tailles d'échantillon $n_i = 5; 10; 30; 50$. Nous observons le biais important de l'échantillonnage défini par un seuil d'inclusion de l'estimateur direct de base, le BR médian étant supérieur à 20 % pour toutes les tailles d'échantillon de domaine. Ce biais de l'échantillonnage défini par un seuil d'inclusion est corrigé par tous les autres estimateurs. Néanmoins, l'estimateur LCALN donne des diagrammes de quartiles plus larges. Cet estimateur obtient un biais important pour certains domaines, probablement parce que le modèle l'assistant ne tient pas compte des effets de domaine. L'estimateur LCAL est fondé sur un modèle qui tient compte des effets de domaine et a de bonnes performances pour ce qui est du biais de plan uniformément pour toutes les tailles d'échantillon de domaine, mais l'EBLUP donne également d'assez bons résultats concernant le biais de plan.

Si nous observons maintenant la REQMR à la figure 8.2, nous pouvons voir que les REQMR des EBLUP sont nettement plus petites pour toutes les tailles d'échantillon de domaine. L'estimateur LCAL obtient des REQMR plus proches à mesure que la taille de l'échantillon de domaine augmente, mais pour $n_i = 5$ il obtient des REQMR très grandes. Nous avons constaté que l'estimateur LCALN peut être fortement biaisé pour certains domaines et qu'il a aussi de grandes REQMR pour toutes les tailles d'échantillon de domaine. En résumé, EBLUP donne la REQMR de plan la plus basse tout en maîtrisant le biais de plan.

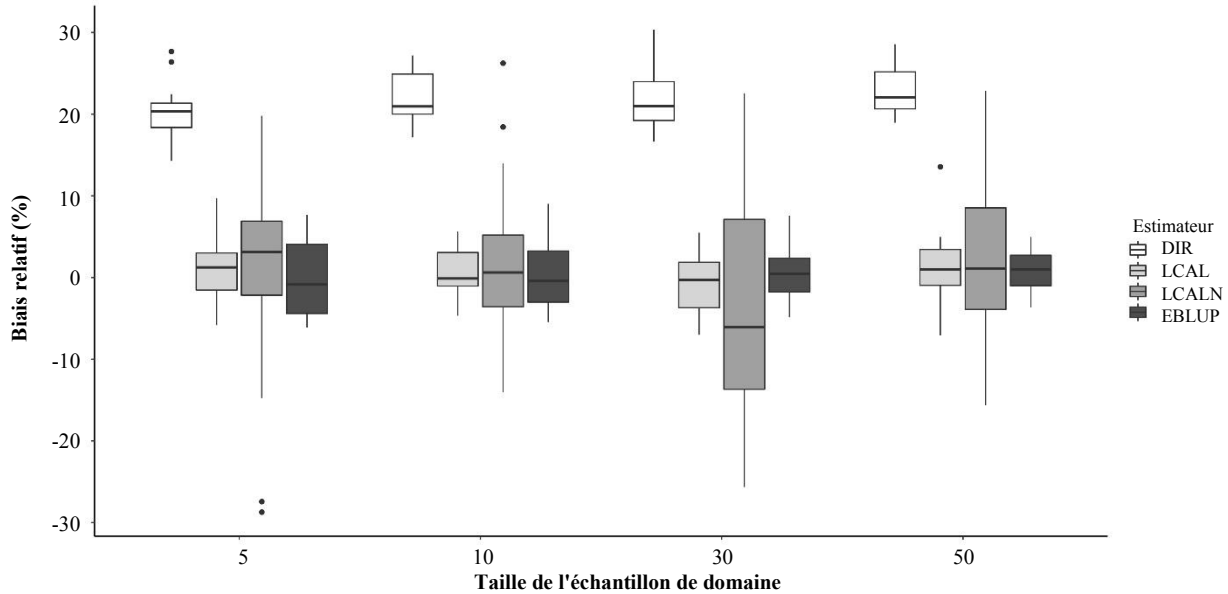

Figure 8.1 Diagrammes de quantiles du BR de domaine (%) de l'estimateur direct de base et des estimateurs LCAL, LCALN et EBLUP pour $n_i = 5; 10; 30; 50$.

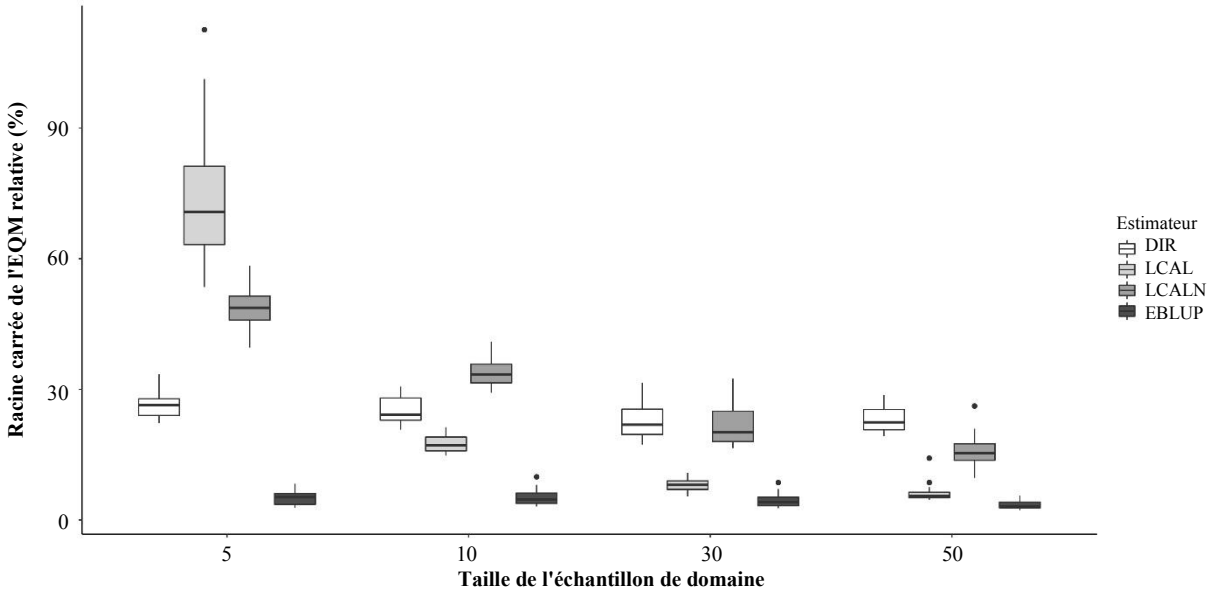

Figure 8.2 Diagrammes de quantiles de domaines REQMR (%) de l'estimateur direct de base et des estimateurs LCAL, LCALN et EBLUP pour n_i = 5; 10; 30; 50.

Le tableau 8.1 présente les moyennes pour tous les domaines du BR absolu et de la REQMR, ainsi que la part en pourcentage du carré du biais provenant de l'EQM totale du plan. Encore une fois, nous observons le biais important de l'échantillonnage défini par un seuil d'inclusion de l'estimateur direct de base, avec une proportion de biais de $B_\pi^2 / \mathrm{EQM}_\pi \approx 100\,\%$, contrairement à tous les autres estimateurs. L'estimateur LCAL a le plus petit biais relatif absolu moyen, et il est suivi de près par l'estimateur EBLUP. Le LCALN obtient les meilleures performances pour ce qui est du ratio du biais en raison de son EQM importante. C'est pourquoi nous considérons que l'estimateur LCAL donne de meilleures performances. Comme cela a été dit, l'estimateur EBLUP est nettement plus performant si l'on examine à la fois le biais et l'EQM.

Tableau 8.1
Moyennes pour tous les domaines du BR absolu, de la REQMR et $B_\pi^2 / \mathrm{EQM}_\pi$ pour l'estimateur direct de base et les estimateurs LCAL, LCALN et EBLUP (en pourcentage)

Méthode	$\overline{\text{BRA}}$	$\overline{\text{REQMR}}$	$B_\pi^2 / \mathrm{EQM}_\pi$
DIR	21,82	24,45	98,32
LCAL	2,96	27,33	2,48
LCALN	8,97	30,44	0,04
EBLUP	3,13	4,56	0,18

8.3 Différents modèles de régression

Dans la présente expérience de simulation, nous conservons les mêmes valeurs de population et plan d'échantillonnage qu'auparavant, mais les valeurs de la variable cible pour les unités incluses et exclues sont générées à partir de modèles ayant des valeurs de paramètres différentes. Il est entendu que ce

scénario n'est pas favorable pour les estimateurs fondés sur un modèle considérés ici, mais il peut être réaliste, car en pratique, le modèle supposé ne peut pas être vérifié pour les unités exclues. Par conséquent, au lieu d'une valeur $\boldsymbol{\beta}$ constante pour toutes les unités de population, nous supposons $\boldsymbol{\beta}_I = (1; 1{,}5; 1)'$ pour les unités incluses et $\boldsymbol{\beta}_E = (0{,}5; 1{,}6; 0{,}5)'$ pour les unités exclues. On suppose que les valeurs des variables explicatives et des composantes de variance σ_u^2 et σ_e^2 sont exactement les mêmes qu'auparavant. Encore une fois, nous tirons $K = 1\,000$ échantillons $s^{(k)}$ par EASSR indépendant dans les unités du domaine i avec $c_{ij} = 1$, avec les mêmes tailles d'échantillon de domaine n_i qu'auparavant. Au moyen des données d'échantillon provenant de l'échantillon k^{e} nous calculons l'estimation par estimateur direct de base, LCAL, LCALN et EBLUP de $\bar{Y}_i$.

La figure 8.3 montre les diagrammes de quantiles des biais relatifs correspondants en pourcentage pour chaque taille d'échantillon de domaine. Dans ce cas, tous les estimateurs sont biaisés, mais le biais de l'estimateur direct de base devient très grand, atteignant plus de 40 % pour certains domaines. Le biais des estimateurs LCAL et EBLUP demeure relativement faible pour tous les domaines, mais celui de LCALN reste très grand en valeur absolue pour certains domaines. En l'absence d'échantillonnage défini par un seuil d'inclusion, les estimateurs par calage sont asymptotiquement sans biais par rapport au plan de sondage à mesure que la taille de l'échantillon de domaine n_i augmente, même si le modèle considéré ne se vérifie pas. Toutefois, cela n'est pas vrai en cas d'échantillonnage défini par un seuil d'inclusion et c'est pourquoi les biais relatifs des estimateurs par calage ne diminuent pas quand n_i croît. Y compris dans ce scénario défavorable comprenant différents modèles générateurs pour les unités incluses et exclues, les EBLUP présentent un biais modéré, qui est comparable à celui de l'estimateur LCAL et qui donne des performances nettement meilleures pour ce qui est de la REQMR.

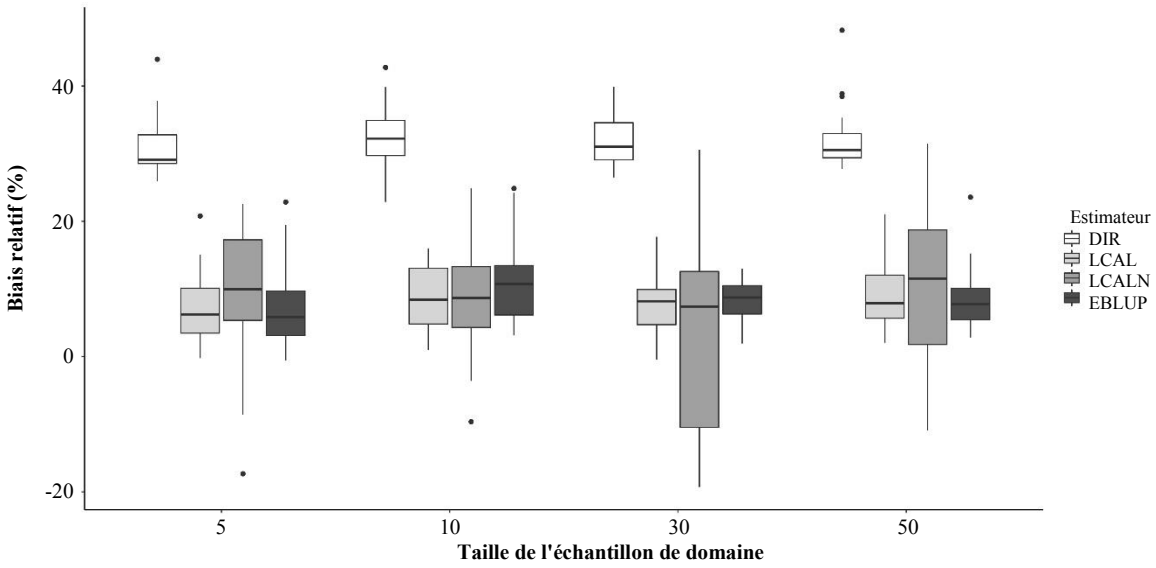

Figure 8.3 Diagrammes de quantiles du BR de domaine (%) de l'estimateur direct de base et des estimateurs LCAL, LCALN et EBLUP pour $n_i = 5; 10; 30; 50$, quand $\boldsymbol{\beta}_I = (1; 1{,}5; 1)'$ pour les unités incluses et $\boldsymbol{\beta}_E = (0{,}5; 1{,}6; 0{,}5)'$ pour les unités exclues.

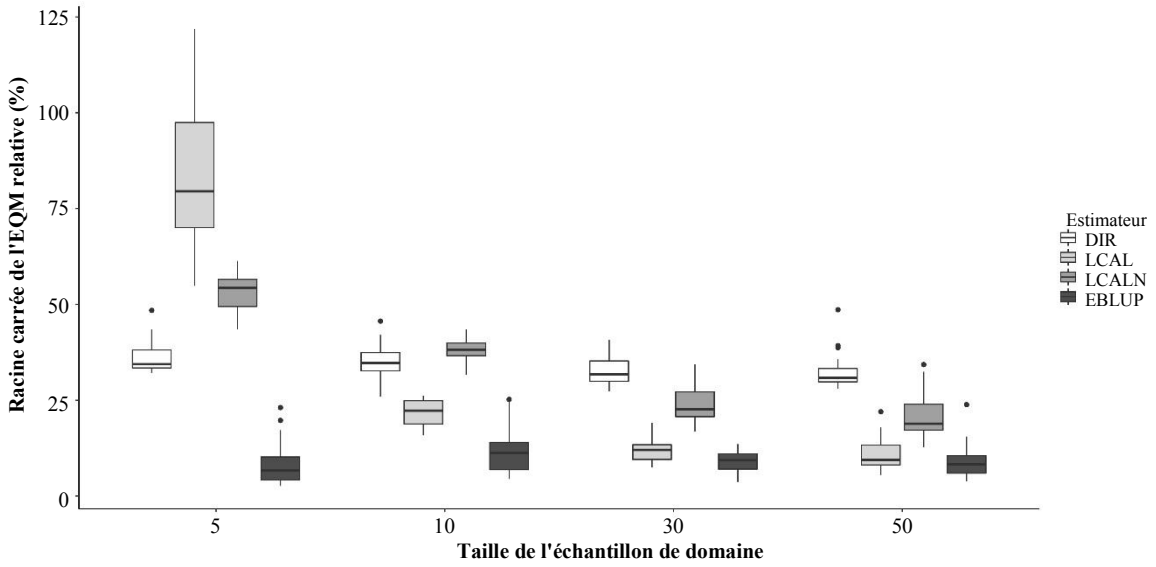

Figure 8.4 Diagrammes de quantiles des REQMR de domaine (%) de l'estimateur direct de base et des estimateurs LCAL, LCALN et EBLUP pour n_i = 5; 10; 30; 50, quand $\boldsymbol{\beta}_I = (1; 1{,}5; 1)'$ pour les unités incluses et $\boldsymbol{\beta}_E = (0{,}5; 1{,}6; 0{,}5)'$ pour les unités exclues.

Encore une fois, les moyennes pour tous les domaines du BR absolu et de la REQMR sont présentées au tableau 8.2, ainsi que le ratio du carré du biais. Comme nous l'avons indiqué, l'estimateur direct de base souffre d'un biais considérable, tandis que les estimateurs LCAL et EBLUP gardent un $\overline{\text{BRA}}$ inférieur à 10 %. L'estimateur LCALN affiche le plus faible ratio de biais en raison d'une EQM plus grande. Encore une fois, l'estimateur EBLUP est le plus efficace, avec une REQMR moyenne également inférieure à 10 %.

Tableau 8.2
Moyennes pour tous les domaines du BR absolu, de la REQMR et B_π^2/EQM_π pour l'estimateur direct de base et les estimateurs LCAL, LCALN et EBLUP, quand $\boldsymbol{\beta}_I = (1; 1{,}5; 1)'$ pour les unités incluses et $\boldsymbol{\beta}_E = (0{,}5; 1{,}6; 0{,}5)'$ pour les unités exclues (en pourcentage)

Méthode	$\overline{\text{BRA}}$	$\overline{\text{REQMR}}$	B_π^2/EQM_π
DIR	31,78	34,11	99,87
LCAL	8,47	30,83	77,43
LCALN	12,75	34,49	29,56
EBLUP	8,73	9,48	75,78

Nous avons répété l'expérience de simulation en supposant une valeur de $\boldsymbol{\beta}_E$ plus éloignée de $\boldsymbol{\beta}_I$, pour que les deux modèles de régression soient sensiblement différents. Nous n'avons pas inclus les

résultats en raison de contraintes d'espace, cependant, comme on pouvait s'y attendre, les valeurs du BR et de la REQMR augmentent pour tous les estimateurs, mais les conclusions sont semblables à celles de la dernière expérience. L'estimateur direct de base obtient le plus grand BR, les estimateurs par calage et EBLUP réduisent nettement le biais d'échantillonnage défini par un seuil d'inclusion de l'estimateur direct de base et EBLUP obtient une REQMR plus petite, particulièrement pour les domaines ayant les plus petites tailles d'échantillon.

9 Estimation des ventes totales dans les provinces espagnoles

Nous décrivons ici une application à l'estimation des ventes totales d'un produit du tabac dans les provinces espagnoles. L'ensemble de données disponibles contient, pour $N = 12\,791$ bureaux de tabac (presque tous) dans $m = 48$ provinces espagnoles (îles Canaries, Ceuta et Melilla non comprises), le volume des achats de ce produit effectués par chaque établissement au cours des trois mois précédant novembre 2016 (z_{ij}, en euros). Il contient également une variable indiquant si un appareil enregistrant tous les renseignements requis au sujet de chaque vente a été fourni à l'établissement. Seuls les établissements aux ventes les plus importantes reçoivent l'appareil en question. Ces établissements (au total $n = 1\,842$) sont en mesure de déclarer précisément les données relatives à leurs ventes. Par conséquent, les données de ces établissements donnent le volume de vente (v_{ij}, en euros) du produit examiné en novembre 2016.

Nous estimons les ventes totales $V_i = \sum_{j=1}^{N_i} v_{ij}$ dans chacune des $m = 48$ provinces incluses dans les données au moyen de l'estimateur de base direct, des estimateurs par calage sélectionnés et d'un estimateur fondé sur un modèle. Les établissements j pour lesquels z_{ij} et v_{ij} sont disponibles pour une province i composent l'ensemble des unités incluses U_{iI}, qui est égal à l'échantillon s_i dans ce cas (il n'y a pas d'échantillonnage dans U_{iI}). Alors, les estimateurs directs de base sont donnés ici par $\hat{V}_i^{\mathrm{HA}} = N_i \bar{V}_{iI}$, $i = 1, \ldots, m$, qui ont une variance nulle, mais qui pourraient être fortement biaisés. Étant donné que les valeurs vraies dans les applications réelles ne sont pas disponibles et que, par conséquent, les biais réels ne peuvent pas être évalués (nous n'avons pas d'information de U_{iE}), nous comparerons ici les estimateurs en considérant l'ensemble des établissements de chaque province dont les ventes sont enregistrées comme EASSR de cette province. Notons qu'il s'agit du meilleur scénario pour l'estimateur direct de base. Par conséquent, pour l'estimateur direct de base $\hat{V}_i^{\mathrm{HA}}$ si l'on considère que l'échantillon réel $s_i = U_{iI}$ est un EASSR de U_i, la variance est égale à l'EQM (nous ignorons le biais). L'estimateur sans biais par rapport au plan de l'EQM est alors

$$\mathrm{eqm}_{\pi}\left(\hat{V}_i\right) = N_i^2 \frac{s_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right), \qquad i = 1, \ldots, m,$$

où $s_i^2 = (n_i - 1)^{-1} \sum_{j \in s_i} \left(v_{ij} - \bar{v}_{is}\right)^2$ est la variance de l'échantillon des ventes de la province i et ici $n_i = N_{iI}$, $i = 1, \ldots, m$.

Pour les estimateurs qui considèrent un modèle de régression, nous effectuons d'abord une analyse descriptive préliminaire des variables. Les histogrammes des ventes v_{ij} et des achats z_{ij} montrent des distributions asymétriques à droite pour les deux variables. De plus, un diagramme de dispersion des résidus des moindres carrés ordinaires à partir d'un modèle linéaire pour v_{ij} en termes de z_{ij}, contre z_{ij} révèle une légère tendance à l'hétéroscédasticité. Il semblerait que la transformation des ventes par la racine carrée, c'est-à-dire en supposant $y_{ij} = v_{ij}^{1/2}$ comme variable réponse et $\mathbf{x}_{ij} = \left(1, x_{ij}\right)'$, avec $x_{ij} = z_{ij}^{1/2}$ comme covariable réduise le problème. Par conséquent, nous examinerons un modèle à erreurs emboîtées (5.1) pour les ventes transformées y_{ij} en termes d'achats transformés x_{ij}, et les MPE ou EBE des ventes totales de chaque province, $V_i = \sum_{j=1}^{N_i} v_{ij}$, seront calculés à partir de ce modèle. Notons que, pour ce qui est des réponses du modèle y_{ij}, les ventes totales sont données par $V_i = \sum_{j=1}^{N_i} y_{ij}^2 = h\left(\mathbf{y}_i\right)$. Alors, le MPE/EBE de $V_i = h\left(\mathbf{y}_i\right)$ est donné par $\hat{V}_i^{\text{MPE}} = E_{m_3}\left[h\left(\mathbf{y}_i\right) \mid \mathbf{y}_{is}; \hat{\boldsymbol{\theta}}\right]$, $i = 1, \ldots, m$, qui peut être calculé de façon analytique ou approximative par simulation de Monte-Carlo. Nous estimons l'EQM du modèle du MPE/EBE au moyen du bootstrap paramétrique décrit à la section 7 pour $H_i = V_i$, en supposant $H_i^{*(b)} = V_i^{*(b)}$ et $\hat{H}_i^{\text{MPE}*(b)} = \hat{V}_i^{\text{MPE}*(b)}$ et en considérant que le modèle se vérifie pour les unités incluses et exclues. Les résidus du modèle sont décrits ci-dessous.

Notons que l'estimateur LCAL (ou GREG) n'est pas défini pour une fonction non linéaire des valeurs de la variable de réponse dans les unités de population, comme le total des ventes $V_i = \sum_{j=1}^{N_i} y_{ij}^2$ après la transformation par la racine carrée. Nous calculons ici alors l'estimateur GREG selon (4.3) au moyen de v_{ij} au lieu de y_{ij} et z_{ij} au lieu de x_{ij}, qui est assisté par le modèle linéaire (4.10) pour les ventes non transformées v_{ij} en termes d'achats z_{ij}. Afin de mesurer l'incertitude de l'estimateur GREG, et pour la rendre comparable à celle du MPE/EBE, nous avons estimé son EQM de modèle au moyen de la même procédure bootstrap, en remplaçant $\hat{H}_i^{\text{MPE}*(b)}$ par $\hat{V}_i^{\text{GREG}*(b)}$. L'estimateur bootstrap obtenu de l'EQM comprend en fait l'erreur causée par le fait que le bon modèle est celui comportant des variables transformées.

Avant de comparer les estimations, nous analysons les résidus du modèle à erreurs emboîtées (5.1), donnés par $\hat{e}_{ij} = y_{ij} - \mathbf{x}_{ij}'\hat{\boldsymbol{\beta}} - \hat{u}_i$. La figure 9.1 montre le diagramme de dispersion de ces résidus par rapport aux valeurs prédites $\hat{y}_{ij} = \mathbf{x}_{ij}'\hat{\boldsymbol{\beta}} + \hat{u}_i$ (à gauche) et un histogramme des résidus (à droite). Nous pouvons voir quelques valeurs aberrantes négatives sur le graphique de gauche, qui concordent avec la queue légèrement plus grande à gauche dans l'histogramme. Les résidus ne présentent aucune autre tendance remarquable. Nous constatons que dans l'histogramme, ils semblent être surtout concentrés autour de zéro, ce qui indique la puissance prédictive élevée du modèle.

La figure 9.2 montre le diagramme Q-Q normal des effets de domaine prédits $\hat{u}_i$. Ce diagramme supporte la normalité de $\hat{u}_i$ sauf pour une valeur aberrante apparaissant sur la queue gauche de la distribution. Ce point correspond à la province ayant la plus petite taille d'échantillon (n_i = 3 observations), ce qui donne à penser que l'effet aléatoire estimé pour cette province, $\hat{u}_i$, n'est pas très fiable. Par conséquent, nous considérons que le modèle à erreurs emboîtées est raisonnablement bien ajusté aux données disponibles.

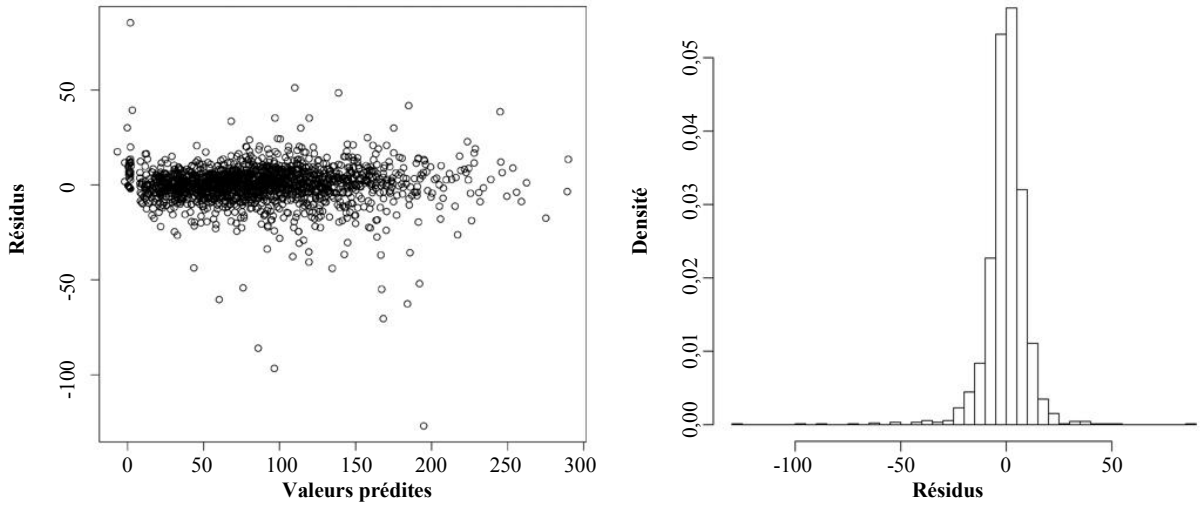

Figure 9.1 Résidus du MPE/EBE par rapport aux valeurs prédites (à gauche) et histogramme des résidus du MPE/EBE (à droite).

<

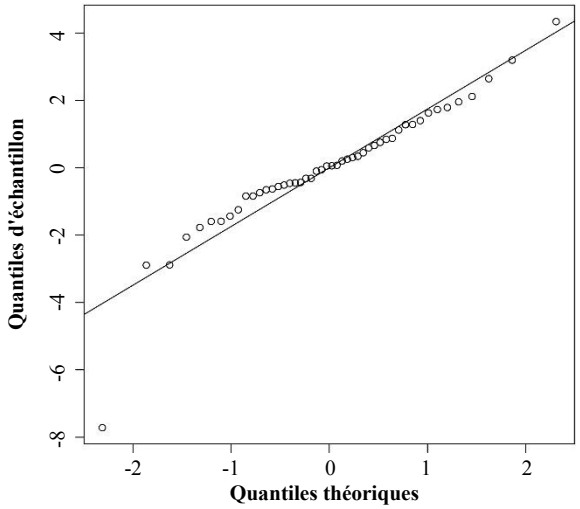

Figure 9.2 Diagramme Q-Q normal des effets prédits pour la province $\hat{u}_i$.

Nous comparerons maintenant les estimations obtenues. La figure 9.3 à gauche montre les MPE/EBE des ventes totales du produit du tabac considéré pour chaque province par rapport aux estimations directes. Les tailles d'échantillon des provinces sont utilisées comme étiquettes de points. Ce diagramme indique une grande similitude entre les deux types d'estimations, sauf pour les deux provinces ayant les plus grandes tailles d'échantillon, où les MPE/EBE sont légèrement plus grands que les estimations directes, ce qui pourrait être dû au biais d'échantillonnage défini par un seuil d'inclusion de l'estimateur

direct. La figure 9.3 à droite présente les MPE/EBE par rapport aux estimations par régression généralisée. La grande similitude entre les estimations GREG et EBE illustrées dans le graphique confirme le fait que les estimateurs directs pourraient en fait sous-estimer les ventes totales dans cette application.

Enfin, nous comparons les trois types d'estimations des ventes totales pour chaque province dans la figure 9.4 à gauche, qui montre les estimations ponctuelles pour chaque province (axe des x); les provinces sont triées par tailles d'échantillon, de la plus petite à la plus grande, et les tailles d'échantillon sont indiquées sur les étiquettes de l'axe des x. Les conclusions sont identiques aux conclusions précédentes, à savoir que les trois types d'estimations prennent des valeurs très semblables pour toutes les provinces, sauf pour un petit nombre de provinces ayant les tailles d'échantillon les plus grandes, où l'estimateur direct de base prend des valeurs légèrement plus petites (qui sous-estiment peut-être les ventes totales). La figure 9.4 (à droite) montre les coefficients de variation (CV) estimés qu'on obtient si l'on ignore le biais causé par l'échantillonnage défini par un seuil d'inclusion. Les estimateurs par MPE/EBE ont des performances uniformément supérieures à celles des autres estimateurs pour ce qui est des CV estimés, dont les valeurs sont maintenues sous 10 % pour presque toutes les provinces alors que l'estimateur GREG obtient des valeurs de CV supérieures à 10 % pour les provinces ayant les plus petites tailles d'échantillon. Nous observons des sommets dans les CV estimés pour certaines provinces qui n'ont pas nécessairement les plus petites tailles d'échantillon. Ces plus grandes valeurs de CV s'expliquent par la présence d'achats et de ventes nuls du produit considéré dans de nombreux bureaux de tabac de ces provinces (le produit examiné n'y est pas acquis chaque mois). Il apparaît clairement que l'estimateur direct est le moins efficace de tous.

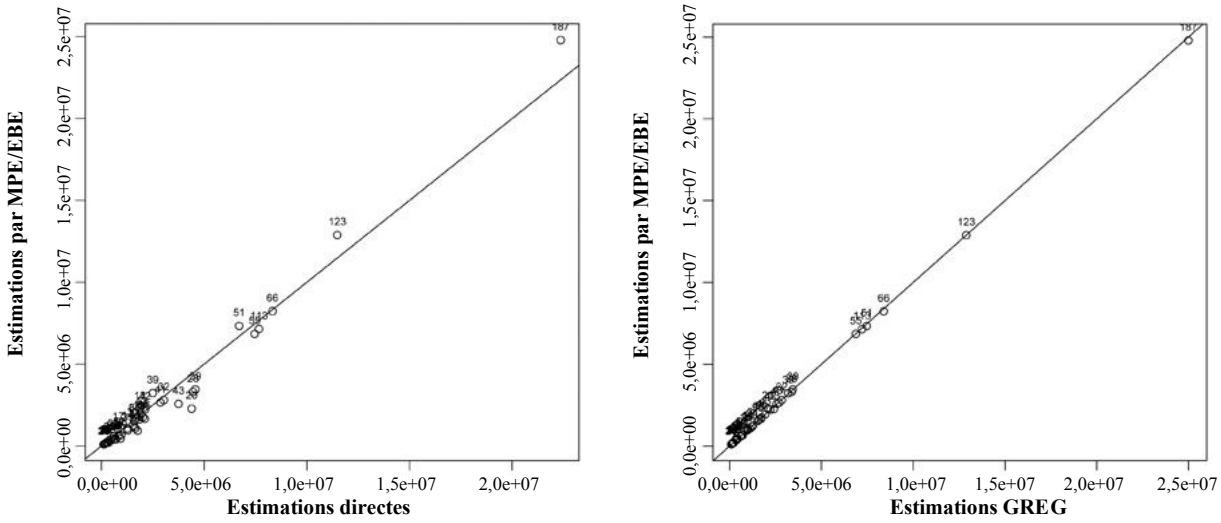

Figure 9.3 Les MPE/EBE des ventes totales de chaque province par rapport aux estimations directes (à gauche) et aux estimations GREG (à droite).

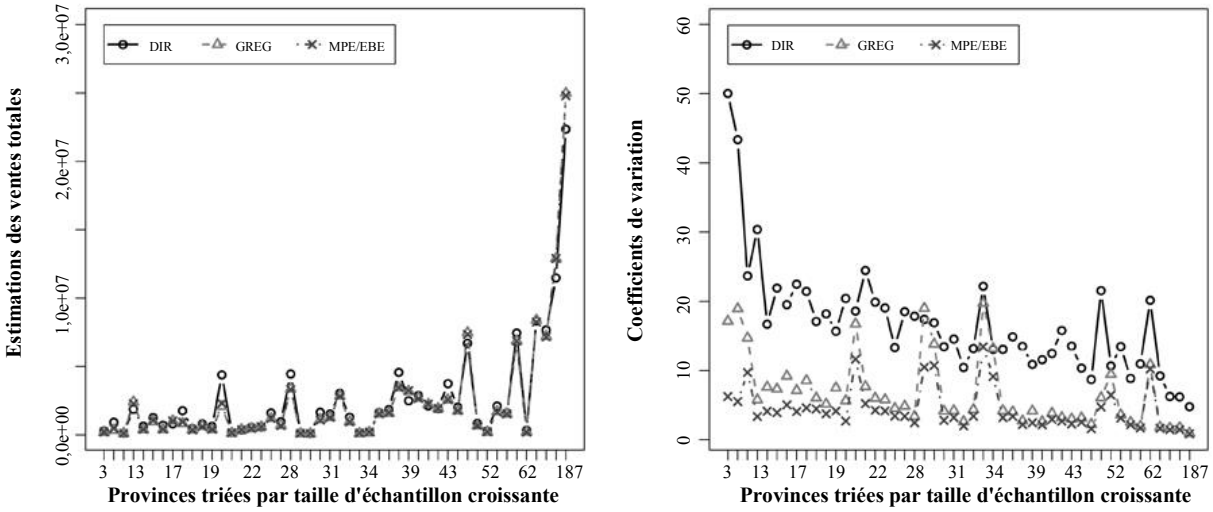

Figure 9.4 Estimations directes, par calage et par MPE/EBE des ventes totales pour chaque province (à gauche) et coefficients de variation estimés correspondants (à droite).

Le tableau A.1 de l'annexe donne les estimations par l'estimateur direct, LCAL et MPE/EBE des ventes totales de la province pour le produit, complétées par l'estimation de leurs CV. Le tableau confirme les meilleures performances du MPE/EBE pour ce qui est de l'estimation du CV dans le modèle à erreurs emboîtées, particulièrement pour les provinces ayant de petites tailles d'échantillon. Enfin, l'estimateur direct a des performances médiocres en termes de CV même si l'on ne tient pas compte du biais attribuable à l'échantillonnage défini par un seuil d'inclusion.

10 Conclusions

L'échantillonnage défini par un seuil d'inclusion est souvent utilisé dans les enquêtes-entreprises, quand le prélèvement d'un échantillon représentatif de l'ensemble de la population entraîne un coût qui ne compense pas vraiment le gain d'exactitude subséquent. Cependant, dans certaines enquêtes, une partie de la population cible peut ne pas être disponible aux fins d'échantillonnage, c'est-à-dire que certains secteurs de population pourraient ne pas être représentés dans l'échantillon. Ces situations se présentent plus souvent qu'on ne le souhaiterait, ce qui donne des estimations directes biaisées comme nous l'avons vu tout au long de l'article.

Nous avons étudié les propriétés de plan théoriques des estimateurs directs de base, par calage et fondés sur un modèle sous un échantillonnage défini par un seuil d'inclusion sur des petits domaines. Nos résultats montrent que l'estimateur EBLUP pour un paramètre linéaire, de la même façon que les estimateurs par calage, réduit considérablement le biais dû à l'échantillonnage défini par un seuil d'inclusion si les modèles pour les individus inclus et exclus sont raisonnablement semblables. Pour ce qui

est de l'EQM, l'estimateur EBLUP est nettement plus performant que les estimateurs par calage, surtout pour les domaines ayant une petite taille d'échantillon.

Dans nos études par simulations et dans l'application, nous avons comparé les méthodes proposées en supposant que le modèle est identique pour toutes les unités de la population (incluses ou exclues). L'hypothèse du modèle est discutable, car aucun moyen ne permet de vérifier le modèle pour les unités exclues. Dans les cas où l'estimation pour l'ensemble du domaine (et non pas seulement pour U_{iI}) est nécessaire, comme dans le présent travail, on doit s'appuyer sur de l'information antérieure subjective concernant la validité du modèle supposé pour les unités exclues. Dans tous les cas, les estimations doivent être considérées seulement comme une indication de ce que pourraient être les valeurs vraies si un même modèle se vérifiait pour toutes les unités de domaine. En fait, nous avons aussi analysé dans des simulations l'utilisation de modèles différents pour les unités incluses et exclues. Dans ce cas, les estimateurs fondés sur un modèle se sont encore montrés les plus efficaces et leur biais n'était pas beaucoup plus important que celui des estimateurs par calage.

Les EQM des estimateurs par calage et fondés sur un modèle sont obtenues sous le modèle. Les instituts nationaux de statistique préfèrent les EQM de plan, car ils ne supposent pas qu'un modèle est correct et tiennent par conséquent compte des défaillances du modèle. On rencontre toutefois les mêmes problèmes pour trouver des estimateurs sans biais par rapport au plan de sondage pour l'EQM de plan selon un échantillonnage défini par un seuil d'inclusion que pour trouver des estimateurs sans biais par rapport au plan de sondage des indicateurs de domaine cibles H_i. Nous envisageons d'appliquer les idées de Strzalkowska-Kominiak et Molina (2019), qui proposent d'emprunter de l'information d'autres domaines aussi pour estimer l'EQM de plan dans un domaine donné, pour trouver des estimateurs de l'EQM du plan ayant un biais moindre d'échantillonnage défini par un seuil d'inclusion.

Enfin, nous avons considéré que les domaines agissent comme des strates d'échantillonnage et que l'échantillonnage défini par un seuil d'inclusion est appliqué dans chaque domaine. Étant donné que les strates sont différentes des domaines (en général, elles recoupent plusieurs domaines), l'application d'un échantillonnage défini par un seuil d'inclusion dans chaque strate donne des tailles d'échantillon aléatoires de domaine. La littérature a rarement étudié l'estimation sur petits domaines dans cette situation. Néanmoins, en rassemblant les sous-échantillons des différentes strates correspondant à un même domaine, nous obtenons un échantillon de chaque domaine. On peut alors réaliser une inférence conditionnellement sur les tailles d'échantillon du domaine observées (Rao, 1985), ce qui réduirait le problème considéré ici.

Remerciements

Les travaux de M. Guadarrama et I. Molina sont soutenus par le ministère espagnol de l'Économie et de la Compétitivité, par les bourses MTM2015-69638-R (MINECO/FEDER, UE) et MTM2015-72907-EXP.

Annexe

Estimations des ventes totales par province

Tableau A.1
Estimations par l'estimateur direct de base, GREG et MPE/EBE des ventes totales pour le produit sélectionné et coefficients de variation estimés (%) pour chaque province espagnole (en augmentant la taille de l'échantillon)

PROVINCE	n_i	$\hat{Y}_i^{HA}$	$\hat{Y}_i^{GREG}$	$\hat{Y}_i^{MPE}$	cv($\hat{Y}_i^{HA}$)	cv($\hat{Y}_i^{GREG}$)	cv($\hat{Y}_i^{MPE}$)
SORIA	3	293 020,0	187 824,9	213 325,0	50,0	17,1	6,2
ZAMORA	7	932 520,0	345 095,8	454 657,0	43,3	18,9	5,5
ALAVA	11	130 083,6	119 918,5	118 835,3	23,7	14,7	9,7
ALMERIA	13	1 870 104,6	2 407 333,1	2 272 051,4	30,4	5,8	3,4
PALENCIA	14	626 340,0	380 367,4	409 775,4	16,7	7,6	4,1
SALAMANCA	14	1 265 580,0	966 094,1	1 068 230,6	21,9	7,3	3,9
AVILA	15	708 696,0	392 474,1	418 917,2	19,5	9,2	5,0
LERIDA	17	817 817,6	1 011 032,3	1 014 770,2	22,5	7,1	4,1
CIUDAD REAL	18	1 764 000,0	841 228,2	939 994,9	21,4	8,6	4,6
GUADALAJARA	18	463 047,8	362 148,3	363 856,9	17,1	6,0	4,5
RIOJA	18	809 900,0	622 488,3	595 178,6	18,2	5,2	3,7
SEGOVIA	19	610 370,5	386 734,4	402 324,0	15,7	7,5	4,2
CACERES	20	4 391 826,0	2 081 619,7	2 286 462,0	20,4	5,6	2,7
GUIPUZCOA	20	181 634,0	136 700,0	156 311,8	18,6	16,7	11,6
HUESCA	22	377 954,5	372 101,3	371 246,5	24,5	7,7	5,2
TERUEL	22	534 417,3	446 565,7	465 643,3	19,9	6,0	4,3
CUENCA	23	588 464,3	587 005,5	586 347,5	19,0	5,8	4,2
VALLADOLID	24	1 609 875,0	1 210 132,8	1 188 336,1	13,3	4,5	3,4
BURGOS	28	961 645,7	708 510,0	666 698,1	18,5	4,9	3,4
CORDOBA	28	4 457 614,3	3 367 169,5	3 312 801,5	17,9	3,4	2,4
ORENSE	28	148 577,1	88 104,6	108 428,9	17,4	19,0	10,5
LUGO	30	107 213,3	92 938,7	104 233,7	16,9	13,8	10,7
ALBACETE	31	1 654 606,5	1 115 182,2	1 073 719,8	13,4	4,2	2,8
LEON	31	1 528 254,2	1 274 531,6	1 270 341,6	14,5	4,2	3,2
PROVINCE	n_i	$\hat{Y}_i^{DIR}$	$\hat{Y}_i^{GREG}$	$\hat{Y}_i^{MPE}$	cv($\hat{Y}_i^{DIR}$)	cv($\hat{Y}_i^{GREG}$)	cv($\hat{Y}_i^{MPE}$)
HUELVA	32	3 031 328,1	2 838 874,0	2 816 281,3	10,5	2,6	2,0
NAVARRA	33	1 291 343,0	956 737,9	957 660,4	13,2	4,4	3,4
PONTEVEDRA	33	159 229,1	107 198,9	138 367,4	22,2	19,7	13,4
VIZCAYA	34	228 618,8	183 267,3	206 304,6	13,1	13,2	9,1
TOLEDO	35	1 619 939,4	1 529 104,8	1 539 799,3	13,1	4,2	3,2
CADIZ	38	1 851 521,1	1 585 755,9	1 620 844,2	14,9	4,0	3,4
BADAJOZ	39	4 571 743,6	3 439 625,5	3 457 692,5	13,5	2,7	2,2
MALAGA	39	2 499 392,3	3 188 031,1	3 237 081,8	10,9	4,2	2,5
TARRAGONA	41	2 872 882,0	2 690 969,7	2 656 117,8	11,6	2,6	2,2
GRANADA	42	2 123 693,3	2 221 155,1	2 241 916,2	12,5	3,8	2,9
JAEN	43	1 928 229,8	1 940 379,2	1 943 101,0	15,8	3,2	2,7
ZARAGOZA	43	3 750 210,7	2 564 909,0	2 578 011,3	13,5	3,0	2,3
GERONA	45	2 029 222,2	1 748 165,7	1 767 490,3	10,4	3,2	2,5
MURCIA	51	6 700 070,6	7 467 465,0	7 341 434,6	8,7	2,2	1,6
BALEARES	52	849 950,8	650 012,6	694 416,3	21,5	6,1	4,7
CANTABRIA	52	285 632,3	204 947,7	226 163,1	10,7	9,5	6,4
ASTURIAS	55	2 113 034,5	1 702 020,8	1 661 932,8	13,5	3,6	3,1
CASTELLON	55	1 605 604,4	1 526 618,1	1 530 394,2	8,9	2,5	2,2
SEVILLA	55	7 458 078,2	6 878 368,2	6 857 368,8	11,0	2,0	1,7
CORUNA	62	340 200,0	217 028,5	206 041,8	20,2	10,9	10,2
ALICANTE	66	8 324 589,1	8 390 895,3	8 240 996,9	9,2	1,8	1,6
VALENCIA	113	7 671 137,7	7 209 128,2	7 153 290,2	6,3	1,7	1,4
MADRID	123	11 483 342,8	12 892 853,8	12 892 305,0	6,2	1,7	1,5
BARCELONA	187	22 356 500,5	24 990 558,9	24 797 372,9	4,8	1,0	0,9

Bibliographie

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

Benedetti, R., Bee, M. et Espa, G. (2010). A framework for cut-off sampling in business survey design. *Journal of Official Statistics*, 26, 651-671.

Foster, J., Greer, J. et Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica: Journal of the Econometric Society*, 761-766.

González-Manteiga, W., Lombardia, 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, 443-462.

Hájek, J. (1971). Discussion of an essay on the logical foundations of survey sampling, comment on a paper by D. Basu. Dans *Foundations of Statistical Inference*, (Éds., V.P. Godambe et D.A. Sprott), Toronto: Holt, Rinehart, Winston.

Haziza, D., Chauvet, G. et Deville, J.-C. (2010). Sampling estimation in presence of cut-off sampling. *Australian & New Zealand Journal of Statistics*, 52, 303-319.

Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

INE (2018). Índices de producción industrial (IPI) base 2015. Rapport technique, Instituto Nacional de Estadística, España.

Molina, I., et Marhuenda, Y. (2015). sae: An R package for small area estimation. *R. Journal*, 1, 81-98.

Molina, I., et Rao, J.N.K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38, 369-385.

Pratesi, M. (2016). *Analysis of Poverty Data by Small Area Estimation*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Rao, J.N.K. (1985). Inférence conditionnelle dans les enquêtes par sondage. *Techniques d'enquête*, 11, 1, 17-35. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1985001/article/14364-fra.pdf.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation*. Hoboken, New Jersey: John Wiley & Sons, Inc.

R Development Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienne.

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

Strzalkowska-Kominiak, E., et Molina, I. (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)*.

Tillé, Y., et Matei, A. (2016). *Sampling: Survey Sampling*. R package version 2.8.

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