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

Techniques d’enquête

Estimation polynomiale locale pour une moyenne de petit domaine sous échantillonnage informatif

par Marius Stefan et Michael A. Hidiroglou

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. 31-52
Statistique Canada, nº 12-001-X au catalogue

Estimation polynomiale locale pour une moyenne de petit domaine sous échantillonnage informatif

Marius Stefan et Michael A. Hidiroglou[1]

Résumé

On a besoin de méthodes par modèle pour estimer des paramètres d'intérêt de petit domaine, comme les totaux et les moyennes, là où les méthodes classiques d'estimation directe ne peuvent garantir une précision suffisante. Les modèles au niveau des unités et au niveau des domaines sont les plus répandus dans la pratique. S'il s'agit d'un modèle au niveau des unités, il est possible d'obtenir des estimateurs efficaces par modèle si le plan de sondage est tel que les modèles d'échantillon et de population coïncident, c'est-à-dire que le plan d'échantillonnage n'est pas informatif pour le modèle en question. Si en revanche le plan de sondage est informatif pour le modèle, les probabilités de sélection seront liées à la variable d'intérêt même après conditionnement par les données auxiliaires disponibles, d'où l'implication que le modèle de la population ne vaut plus pour l'échantillon. Pfeffermann et Sverchkov (2007) se sont reportés aux relations entre les distributions de population et d'échantillon de la variable étudiée pour obtenir des prédicteurs semi-paramétriques approximativement sans biais des moyennes de domaine dans des plans d'échantillonnage informatifs. La procédure qu'ils ont employée est applicable aux domaines avec et sans échantillon. Verret, Rao et Hidiroglou (2015) ont étudié d'autres méthodes utilisant une fonction appropriée des probabilités de sélection d'unités comme variable auxiliaire supplémentaire. Leur technique a donné des estimateurs *Empirical Best Linear Unbiased Prediction* (EBLUP) approximativement sans biais pour les moyennes de petit domaine. Dans le présent exposé, nous étendons la méthode de Verret et coll. (2015) en ne formant aucune hypothèse au sujet des probabilités d'inclusion. Nous nous contentons d'intégrer ces dernières au modèle au niveau des unités en utilisant une fonction lisse des probabilités d'inclusion. C'est une fonction que nous estimons par une approximation locale donnant un estimateur polynomial local. Nous proposons une méthode bootstrap conditionnelle pour l'estimation de l'erreur quadratique moyenne (EQM) des estimateurs polynomiaux locaux et des estimateurs EBLUP. Nous examinons par simulation le biais et les propriétés d'efficacité de l'estimateur polynomial local. Nous présentons enfin les résultats de l'estimateur bootstrap de l'EQM.

Mots-clés : Estimation polynomialelocale; estimation EBLUP; modèle augmenté; modèle à erreur emboîtée; échantillonnage informatif; bootstrap conditionnel.

1 Introduction

On a souvent besoin pour de petites sous-populations (ou domaines) de totaux et de moyennes de population. Lorsque l'inférence repose sur des données d'échantillon par domaine, les estimateurs obtenus des paramètres de petit domaine (ou estimateurs directs) ne sont pas d'une précision suffisante en raison de la petite taille d'échantillon par domaine. Il devient donc nécessaire d'emprunter de la puissance à l'échelle des domaines. On tire de la puissance par des estimateurs indirects (prédicteurs) quand un modèle est exploité pour la population de petits domaines. Ce modèle fait le lien avec les petits domaines apparentés et, par conséquent, un estimateur indirect de petit domaine par modèle se trouve à exploiter toutes les observations de l'échantillon national, tout comme les observations du petit domaine considéré.

Posons que la population d'intérêt, U de taille N, consiste en M domaines non chevauchants avec N_i unités dans le i^e petit domaine U_i $(i = 1, \ldots, M)$. Nous prélevons d'abord un échantillon s de m domaines à l'aide d'un plan de sondage spécifié où les probabilités d'inclusion sont $\pi_i = mp_i$ $(i = 1, \ldots, M)$ et où p_i désigne la probabilité de sélection du petit domaine i. Nous tirons

1. Marius Stefan, Faculté des sciences appliquées, Université polytechnique de Bucarest, Splaiul Independentei, nr. 313. Courriel : mastefan@gmail.com; Michael A. Hidiroglou, ancien employé de Statistique Canada. Courriel : hidirog@yahoo.ca.

indépendamment des sous-échantillons s_i de tailles spécifiées n_i de chaque petit domaine U_i en application du plan d'échantillonnage spécifié avec des probabilités de sélection $p_{j|i}$ $\left(\sum_{j=1}^{N_i} p_{j|i} = 1\right)$. Les probabilités d'inclusion sont $\pi_{j|i} = n_i p_{j|i}$ et les poids d'échantillonnage, $w_{j|i} = \pi_{j|i}^{-1}$. Nous considérons les probabilités de sélection $p_{j|i}$ proportionnelles à une mesure de taille, c_{ij}, reliée à la variable réponse y_{ij}; en d'autres termes, $p_{j|i} = c_{ij} \Big/ \sum_{k=1}^{N_i} c_{ik}$. Nous posons que tous les petits domaines sont échantillonnés, c'est-à-dire que $m = M$. La taille résultante est $n = \sum_{i=1}^{M} n_i$ pour l'ensemble de l'échantillon.

Le modèle de régression à erreur emboîtée de base de la population, qui vient de Battese, Harter et Fuller (1988), est donné par

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij}, \; j = 1, \ldots, N_i; \; i = 1, \ldots, M, \tag{1.1}$$

où y_{ij} est la valeur de la variable réponse pour l'unité j dans le petit domaine i, où $\mathbf{x}_{ij} = \left(1, \; x_{ij1}, \ldots, x_{ijp}\right)^T$ est le vecteur de covariables, où $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_p\right)^T$ est le vecteur d'effets fixes et où $v_i \overset{\text{iid}}{\sim} N\left(0, \sigma_v^2\right)$ correspond aux effets aléatoires de petit domaine indépendants des erreurs au niveau des unités $e_{ij} \overset{\text{iid}}{\sim} N\left(0, \sigma_e^2\right)$. L'estimation des moyennes de petit domaine, $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$, est d'un intérêt premier.

Si le plan de sondage n'est pas informatif pour le modèle, c'est-à-dire que le modèle en (1.1) tient pour l'échantillon, il est alors possible d'obtenir des estimateurs efficaces par modèle des moyennes de petit domaine $\bar{Y}_i$ par le meilleur prédicteur linéaire sans biais empirique ou EBLUP (voir Rao et Molina, 2015, chapitre 6, pour un excellent compte rendu de cette méthode). Dans ce cas, les modèles d'échantillon et de population coïncident, d'où la possibilité d'appliquer (1.1) aux données d'échantillon pour estimer $\bar{Y}_i$.

Si la probabilité de sélection $p_{j|i}$ est liée à y_{ij} même après conditionnement par $\mathbf{x}_{ij}$, le plan de sondage est informatif et le modèle en (1.1) ne tient plus pour l'échantillon. La conséquence est que l'estimateur EBLUP, qui est fondé sur (1.1) pour l'échantillon, risque d'être lourdement entaché d'un biais. Il est donc nécessaire de développer des estimateurs pouvant tenir compte de la sélection de l'échantillon et ainsi réduire le biais d'estimation. C'est pourquoi Verret et coll. (2015) ont augmenté le modèle en (1.1) en incluant la variable $g\left(p_{j|i}\right)$, où $g\left(p_{j|i}\right)$ est une fonction spécifiée de la probabilité $p_{j|i}$. Le modèle de ces auteurs pour l'échantillon est donné par

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_0 + g\left(p_{j|i}\right) \delta_0 + v_{0i} + e_{0ij}, \; j = 1, \ldots, n_i; \; i = 1, \ldots, M, \tag{1.2}$$

où $v_{0i} \overset{\text{iid}}{\sim} N\left(0, \sigma_{0v}^2\right)$ et est indépendant de $e_{0ij} \overset{\text{iid}}{\sim} N\left(0, \sigma_{0e}^2\right)$ et où $\boldsymbol{\beta}_0 = \left(\beta_{00}, \beta_{01}, \ldots, \beta_{0p}\right)^T$. Verret et coll. (2015) ont vérifié la justesse de (1.2) après avoir ajusté le modèle aux données d'échantillon $\left(y_{ij}, \mathbf{x}_{ij}, p_{j|i}\right)$, $j = 1, \ldots, n_i$; $i = 1, \ldots, M$, pour divers choix de $g(\cdot)$ qui assurent le meilleur ajustement à ces données. Ils ont avancé les quatre possibilités suivantes pour le choix de $g\left(p_{j|i}\right)$: $p_{j|i}$, $\log\left(p_{j|i}\right)$, $w_{j|i} = \left(n_i p_{j|i}\right)^{-1}$ et $n_i w_{j|i} = p_{j|i}^{-1}$. Comme leur modèle d'échantillon est paramétrique, la théorie EBLUP peut servir à l'estimation des paramètres d'intérêt à l'aide du modèle en (1.2).

Verret et coll. (2015) ont montré par une simulation que l'estimateur EBLUP résultant, désigné par $\hat{\bar{Y}}_i^{\text{VRH}}$ et obtenu en (1.2), performe bien avec un plan de sondage informatif en réduisant tant le biais que

l'erreur quadratique moyenne si on le compare à l'estimateur EBLUP, $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$, tiré des données d'échantillon dans le modèle non augmenté (1.1). Dans leur étude de simulation, ils ont comparé leur méthode à celle de Pfeffermann et Sverchkov (2007). Leurs résultats font voir que l'estimateur corrigé pour le biais de Pfeffermann et Sverchkov (2007) performe bien avec un plan d'échantillonnage informatif pour le biais, mais que son EQM est significativement supérieure à l'EQM correspondante de l'estimateur EBLUP sur modèle augmenté.

Dans notre exposé, nous ne formulons aucune hypothèse quant à la forme de la fonction $g\left(p_{j|i}\right)$. Nous intégrons plutôt les $p_{j|i}$ au modèle en (1.1) par une fonction lisse inconnue $m_0\left(p_{j|i}\right)$. Notre fonction lisse $m_0(\cdot)$ n'a pas de forme paramétrique comme celle de Verret et coll. (2015). Nous posons que $m_0(\cdot)$ peut être localement approximée par un polynôme d'ordre q. Pour chaque point l du petit domaine U_k, le polynôme correspondant s'obtient par le développement en séries de Taylor de $m_0\left(p_{j|i}\right)$ dans un voisinage de $p_{l|k}$. Pour chaque point (l, k) dans la population, nous remplaçons $m_0\left(p_{j|i}\right)$ par l'approximation paramétrique correspondante et ajustons le modèle résultant comme dans l'ajustement paramétrique. C'est la méthode que nous qualifions de localisation polynomiale paramétrique.

L'approximation locale donne un modèle augmenté qui est semi-paramétrique. Opsomer, Claeskens, Ranalli, Kauermann et Breidt (2008) ont employé de tels modèles dans des estimations de petit domaine. Ces auteurs retiennent une technique par splines pénalisés pour estimer la partie non paramétrique de leurs modèles. Breidt et Opsomer (2000) et Breidt, Opsomer, Johnson et Ranalli (2007) ont utilisé la technique polynomiale locale dans la théorie de l'échantillonnage d'enquête pour élaborer des estimateurs par modèle. De tels estimateurs font appel à des modèles non paramétriques sans effets aléatoires. Autant que nous sachions, on n'a guère étudié jusqu'à présent tout ce qui est estimation de moyenne de petit domaine $\bar{Y}_i$ par une technique de localisation polynomiale assortie de modèles semi-paramétriques.

Voici comment nous avons structuré notre propos. À la section 2, nous examinons deux méthodes donnant des estimateurs tenant compte de la sélection de l'échantillon, lesquelles ont été conçues par Pfeffermann et Sverchkov (2007) et Verret et coll. (2015). À la section 3, nous exposons une procédure en trois étapes permettant d'estimer le modèle augmenté semi-paramétrique qui est proposé et la moyenne de petit domaine $\bar{Y}_i$ par voie d'approximation polynomiale locale. Nous désignons par $\hat{\bar{Y}}_i^{\mathrm{PL}}$ l'estimateur ainsi obtenu de moyenne de petit domaine. L'erreur quadratique moyenne (EQM) de $\hat{\bar{Y}}_i^{\mathrm{PL}}$ est estimée à la section 4 par une méthode bootstrap conditionnelle paramétrique. Nous employons aussi cette méthode pour estimer l'EQM des estimateurs EBLUP sur modèle augmenté (1.2). À la section 5, nous faisons une étude de simulation dans le cadre plan-modèle (ou pm) pour comparer le biais et l'EQM du nouvel estimateur $\hat{\bar{Y}}_i^{\mathrm{PL}}$ à ceux de l'estimateur EBLUP ainsi qu'aux deux estimateurs examinés dans Verret et coll. (2015). Nous étudions également avec quelle efficacité la méthode bootstrap conditionnelle estime l'EQM du polynôme local proposé et des estimateurs EBLUP dans Verret et coll. (2015). Nous évaluons le rendement de ces estimateurs en biais relatif moyen et en intervalle de confiance moyen. Nous livrons nos observations en conclusion à la section 6.

2 Méthodes existantes

Posons que le modèle de population (1.1) vaut pour l'échantillon. Soit $\bar{\mathbf{X}}_i$ la moyenne de domaine des valeurs $\mathbf{x}_{ij}$ de population. L'estimateur EBLUP de $\mu_i = \bar{\mathbf{X}}_i^T\boldsymbol{\beta} + v_i$ est alors donné par

$$\hat{\mu}_i^{\text{EBLUP}} = \bar{\mathbf{X}}_i^T\hat{\boldsymbol{\beta}} + \hat{v}_i = \hat{\gamma}_i\bar{y}_i + \left(\bar{\mathbf{X}}_i - \hat{\gamma}_i\bar{\mathbf{x}}_i\right)^T\hat{\boldsymbol{\beta}}, \tag{2.1}$$

où $\hat{\gamma}_i = \hat{\sigma}_v^2 \Big/ \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2/n_i\right)$, $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} \Big/ n_i$, $\bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij} \Big/ n_i$ sont les moyennes d'échantillon non pondérées de la variable réponse y et des covariables $\mathbf{x}$, et où $\hat{v}_i = \hat{\gamma}_i\left(\bar{y}_i - \bar{\mathbf{x}}_i^T\hat{\boldsymbol{\beta}}\right)$. L'estimateur du vecteur de régression $\boldsymbol{\beta}$ en (1.1) est

$$\hat{\boldsymbol{\beta}} = \left\{\sum_{i=1}^{M}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\left(\mathbf{x}_{ij} - \hat{\gamma}_i\bar{\mathbf{x}}_i\right)^T\right\}^{-1}\left\{\sum_{i=1}^{M}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij} - \hat{\gamma}_i\bar{\mathbf{x}}_i\right)y_{ij}\right\}. \tag{2.2}$$

Nous obtenons les composantes estimées de la variance $\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right)$ par la méthode d'Henderson qui consiste en un ajustement de constantes (HFC) ou en un calcul de maximum de vraisemblance avec contrainte (MVC) (voir Battese et coll., 1988, et Rao et Molina, 2015, chapitre 7). L'estimateur EBLUP de la moyenne de domaine $\bar{Y}_i$ peut s'écrire sous la forme $\hat{\mu}_i^{\text{EBLUP}}$ comme

$$\hat{\bar{Y}}_i^{\text{EBLUP}} = \frac{1}{N_i}\left[\left(N_i - n_i\right)\hat{\mu}_i^{\text{EBLUP}} + n_i\left\{\bar{y}_i + \left(\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i\right)^T\hat{\boldsymbol{\beta}}\right\}\right]. \tag{2.3}$$

À noter que $\hat{\bar{Y}}_i^{\text{EBLUP}} \approx \hat{\mu}_i^{\text{EBLUP}}$ si le taux d'échantillonnage n_i/N_i est suffisamment petit. L'estimateur EBLUP $\hat{\bar{Y}}_i^{\text{EBLUP}}$ s'accorde avec le plan dans le cas d'un échantillonnage aléatoire simple (EAS) ou avec stratification (EASS) avec répartition proportionnelle à l'intérieur du petit domaine U_i et donc en équiprobabilité des $p_{j|i}$.

Pfeffermann et Sverchkov (2007) ont étudié l'estimation de moyenne de petit domaine dans un échantillonnage informatif en posant le modèle suivant pour les données d'échantillon :

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\alpha} + u_i + h_{ij};\ j = 1, \ldots, n_i;\ i = 1, \ldots, M, \tag{2.4}$$

où $u_i \overset{\text{iid}}{\sim} N\left(0, \sigma_u^2\right)$ et $h_{ij}\left|\, j \in s_i \right. \overset{\text{iid}}{\sim} N\left(0, \sigma_h^2\right)$. Ils ont supposé que le poids des unités selon le plan $w_{j|i} = \pi_{j|i}^{-1}$ est aléatoire avec une espérance conditionnelle

$$\begin{aligned} E_{si}\left(w_{j|i}\left|\, \mathbf{x}_{ij}, y_{ij}, v_i\right.\right) &= E_{si}\left(w_{j|i}\left|\, \mathbf{x}_{ij}, y_{ij}\right.\right) \\ &= k_i\exp\left(\mathbf{x}_{ij}^T\mathbf{a} + dy_{ij}\right), \end{aligned} \tag{2.5}$$

où $\mathbf{a}$ et d sont des constantes fixes inconnues et où

$$k_i = \frac{N_i}{n_i}\left\{\sum_{j=1}^{N_i}\exp\left(-\mathbf{x}_{ij}^T\mathbf{a} - dy_{ij}\right) \Big/ N_i\right\}.$$

L'estimateur de $\bar{Y}_i$ de Pfeffermann et Sverchkov (2007) protège contre l'échantillonnage informatif dans l'éventualité que cette hypothèse se vérifie. L'estimateur est donné par

$$\hat{\bar{Y}}_i^{\text{PS}} = \frac{1}{N_i}\left[\left(N_i - n_i\right)\hat{\mu}_{iu}^{\text{EBLUP}} + n_i\left\{\bar{y}_i + \left(\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i\right)^T\hat{\boldsymbol{\alpha}}\right\} + \left(N_i - n_i\right)\hat{d}\hat{\sigma}_h^2\right], \tag{2.6}$$

où $\hat{\mu}_{iu}^{\text{EBLUP}} = \bar{\mathbf{X}}_i^T \hat{\boldsymbol{\alpha}} + \hat{u}_i$ est l'estimateur EBLUP de $\mu_{iu} = \bar{\mathbf{X}}_i^T \boldsymbol{\alpha} + u_i$ dans le modèle d'échantillon en (2.4) et où $\hat{d}$ est un estimateur de d dans le modèle en (2.5) pour les poids $w_{j|i}$. Le dernier terme en (2.6) corrige tout biais dû à l'échantillonnage informatif en (2.5). Pfeffermann et Sverchkov (2007) ont obtenu l'estimateur $\hat{d}$ de d en (2.5) par une régression des poids d'échantillonnage $w_{j|i}$ sur $k_i \exp\left(\mathbf{x}_{ij}^T \mathbf{a} + d y_{ij}\right)$. Nous pouvons estimer les coefficients k_i, $\mathbf{a}$ et d par ajustement du modèle (2.5) à l'aide de la procédure NLIN en SAS ou de la fonction `nls` en Splus. Les calculs sont itératifs et les valeurs initiales de $\mathbf{a}$ et d s'obtiennent par une régression de $\log\left(w_{j|i}\right)$ sur $\mathbf{x}_{ij}$ et y_{ij}. Les valeurs initiales pour $\hat{k}_i$, $i = 1, \ldots, M$ se prennent comme $k_i = N_i / n_i$.

Nous obtenons l'estimateur de Verret et coll. (2015) lorsque nous appliquons la théorie EBLUP au modèle en (1.2). Soit $\mathbf{x}_{ij}^{\text{aug}} = \left(\mathbf{x}_{ij}^T, g\left(p_{j|i}\right)\right)^T$ le vecteur $\mathbf{x}_{ij}$ augmenté de la variable $g\left(p_{j|i}\right)$ et soit $\bar{G}_i$ la moyenne de domaine des valeurs de population $g\left(p_{j|i}\right)$ et $\mu_{0i} = \bar{\mathbf{X}}_i^T \boldsymbol{\beta}_0 + \bar{G}_i \delta_0 + v_{0i}$. L'estimateur EBLUP de μ_{0i} est donné par

$$\hat{\mu}_{0i}^{\text{EBLUP}} = \bar{\mathbf{X}}_i^T \hat{\boldsymbol{\beta}}_0 + \bar{G}_i \hat{\delta}_0 + \hat{v}_{0i} = \hat{\gamma}_{0i} \bar{y}_i + \left(\bar{\mathbf{X}}_i - \hat{\gamma}_{0i} \bar{\mathbf{x}}_i\right)^T \hat{\boldsymbol{\beta}}_0 + \left(\bar{G}_i - \hat{\gamma}_{0i} \bar{g}_i\right) \hat{\delta}_0, \tag{2.7}$$

où $\hat{\gamma}_{0i} = \hat{\sigma}_{0v}^2 / \left(\hat{\sigma}_{0v}^2 + \hat{\sigma}_{0e}^2 / n_i\right)$, $\bar{g}_i = \sum_{j=1}^{n_i} g\left(p_{j|i}\right) / n_i$ et $\hat{v}_{0i} = \hat{\gamma}_{0i}\left(\bar{y}_i - \bar{\mathbf{x}}_i^T \hat{\boldsymbol{\beta}}_0 - \bar{g}_i \hat{\delta}_0\right)$. Nous estimons les paramètres $(\boldsymbol{\beta}_0, \delta_0)$ par

$$\left(\hat{\boldsymbol{\beta}}_0^T, \hat{\delta}_0\right)^T = \left\{\sum_{i=1}^{M} \sum_{j=1}^{n_i} \mathbf{x}_{ij}^{\text{aug}} \left(\mathbf{x}_{ij}^{\text{aug}} - \hat{\gamma}_{0i} \bar{\mathbf{x}}_i^{\text{aug}}\right)^T\right\}^{-1} \left\{\sum_{i=1}^{M} \sum_{j=1}^{n_i} \left(\mathbf{x}_{ij}^{\text{aug}} - \hat{\gamma}_{0i} \bar{\mathbf{x}}_i^{\text{aug}}\right) y_{ij}\right\}, \tag{2.8}$$

avec $\bar{\mathbf{x}}_i^{\text{aug}} = \sum_{j=1}^{n_i} \mathbf{x}_{ij}^{\text{aug}} \Big/ n_i = \left(\bar{\mathbf{x}}_i^T, \bar{g}_i\right)^T$. Nous estimons les paramètres de modèle $\left(\hat{\sigma}_{0v}^2, \hat{\sigma}_{0e}^2\right)$ par la méthode HFC ou MVC. L'estimateur de la moyenne de domaine $\bar{Y}_i$, qui est désigné par $\hat{\bar{Y}}_i^{\text{VRH}}$, peut s'écrire sous la forme $\hat{\mu}_{0i}^{\text{EBLUP}}$ comme

$$\hat{\bar{Y}}_i^{\text{VRH}} = \frac{1}{N_i}\left[\left(N_i - n_i\right) \hat{\mu}_{0i}^{\text{EBLUP}} + n_i \left\{\bar{y}_i + \left(\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i\right)^T \hat{\boldsymbol{\beta}}_0 + \left(\bar{G}_i - \bar{g}_i\right)^T \hat{\delta}_0\right\}\right]. \tag{2.9}$$

3 Estimateur polynomial local

3.1 Estimation d'une moyenne de petit domaine

Le but est d'estimer la moyenne $\bar{Y}_i$ du petit domaine U_i pour $i = 1, \ldots, M$. Si nous divisons la population U_i en unités observées dans l'échantillon s_i de taille n_i et en unités inobservées dans la partie non échantillonnée $\bar{s}_i = U_i / s_i$ de taille $N_i - n_i$, nous pouvons formuler $\bar{Y}_i$ comme

$$\bar{Y}_i = \frac{1}{N_i}\left(\sum_{j \in s_i} y_{ij} + \sum_{j \in \bar{s}_i} y_{ij}\right). \tag{3.1}$$

Comme nous ignorons les valeurs y des unités inobservées dans les ensembles $\bar{s}_i$ pour $i = 1, \ldots, M$, nous devons les estimer. Si nous désignons par $\hat{y}_{ij}$ l'estimateur de y_{ij} pour ces unités, l'estimateur résultant de la moyenne $\bar{Y}_i$ est

$$\hat{\bar{Y}}_i = \frac{1}{N_i}\left(\sum_{j \in s_i} y_{ij} + \sum_{j \in \bar{s}_i} \hat{y}_{ij}\right). \tag{3.2}$$

Nous obtenons les estimateurs $\hat{y}_{ij}$ de y_{ij} pour $j \in \overline{s}_i$ en nous fondant sur un modèle augmenté comprenant une fonction lisse inconnue des probabilités de sélection $p_{j|i}$, ce que nous désignons par $m_0\left(p_{j|i}\right)$. Le modèle d'échantillon semi-paramétrique augmenté que nous proposons est donné par

$$y_{ij} = \tilde{\mathbf{x}}_{ij}^T \boldsymbol{\beta}_1 + m_0\left(p_{j|i}\right) + v_{1i} + e_{1ij},\ j = 1, \ldots, n_i;\ i = 1, \ldots, M, \tag{3.3}$$

où $v_{1i} \overset{\text{iid}}{\sim} N\left(0, \sigma_{1v}^2\right)$ et est indépendant de $e_{1ij} \overset{\text{iid}}{\sim} N\left(0, \sigma_{1e}^2\right)$. Le vecteur $\tilde{\mathbf{x}}_{ij} = \left(x_{ij1}, \ldots, x_{ijp}\right)^T$ dans le modèle (3.3) représente les covariables $\mathbf{x}_{ij}$ sans une constante (l'ordonnée à l'origine en l'occurrence) et $\boldsymbol{\beta}_1 = \left(\beta_{11}, \ldots, \beta_{1p}\right)^T$, un vecteur d'effets fixes. Le modèle (3.3) est semi-paramétrique, car la variable réponse y_{ij} dépend linéairement du vecteur de variables auxiliaires, $\tilde{\mathbf{x}}_{ij}$, et la probabilité de sélection $p_{j|i}$ s'ajoute non paramétriquement par la fonction lisse $m_0(\cdot)$.

Nous posons que le modèle en (3.3) est d'une structure des covariances semblable à celle du modèle en (1.2); les effets de petit domaine v_{1i} et les erreurs aléatoires e_{1ij} sont i.i.d., à distribution normale et indépendants les uns des autres. Toutefois, le modèle semi-paramétrique (3.3) est plus souple que le modèle paramétrique (1.2), puisque la fonction $m_0\left(p_{j|i}\right)$ n'a pas à être d'une forme particulière. Il y a un inconvénient à ce paramétrage. Comme le modèle en (3.3) n'est pas un modèle mixte linéaire, la théorie générale EBLUP à la section 2 ne peut directement servir à dégager des estimateurs de $m_0\left(p_{j|i}\right)$, $\boldsymbol{\beta}_1$ et v_{1i}. Nous proposons donc de procéder à l'estimation en (3.3) en combinant la théorie EBLUP des modèles mixtes linéaires et la technique d'estimation polynomiale locale (Fan et Gijbels, 1996).

Nous estimons (3.3) en trois étapes. D'abord, nous obtenons des estimations de $m_0\left(p_{j|i}\right)$, $\hat{m}_0\left(p_{j|i}\right)$, $j = 1, \ldots, N_i$, $i = 1, \ldots, M$, pour toutes les unités de la population. Ces estimations sont d'un caractère local, car elles reposent sur la technique d'estimation polynomiale locale. En deuxième lieu, nous prenons les estimations $\hat{m}_0\left(p_{j|i}\right)$, $j \in s_i$, des unités observées pour obtenir des estimateurs globaux de $\boldsymbol{\beta}_1$ et v_{1i}, $i = 1, \ldots, M$. Nous désignons ces estimateurs par $\hat{\boldsymbol{\beta}}_{\text{glo},1}$ et $\hat{v}_{\text{glo},1i}$, $i = 1, \ldots, M$. En troisième étape, nous utilisons les estimateurs locaux $\hat{m}_0\left(p_{j|i}\right)$ pour les unités inobservées en première étape et les estimateurs globaux $\hat{\boldsymbol{\beta}}_{\text{glo},1}$ et $\hat{v}_{\text{glo},1i}$ en deuxième étape afin d'estimer y_{ij} pour $j \in \overline{s}_i$ et $i = 1, \ldots, M$. Les estimateurs de y_{ij} ainsi obtenus, qui sont désignés par $\hat{y}_{ij}$, sont

$$\hat{y}_{ij} = \tilde{\mathbf{x}}_{ij}^T \hat{\boldsymbol{\beta}}_{\text{glo},1} + \hat{m}_0\left(p_{j|i}\right) + \hat{v}_{\text{glo},1i},\ j \in \overline{s}_i. \tag{3.4}$$

Nous intégrons les $\hat{y}_{ij}$ à l'équation (3.2) pour dégager l'estimateur de la moyenne de petit domaine $\hat{\bar{Y}}_i$.

Il s'agit maintenant de décrire la première étape plus en détail. À la suite de Ruppert et Matteson (2015), nous estimons les valeurs de la fonction inconnue $m_0\left(p_{l|k}\right)$ pour toutes les unités $l \in U_k$ et les petits domaines k, avec $k = 1, \ldots, M$, en procédant par régression polynomiale locale. Cette régression repose sur le principe selon lequel une fonction lisse peut être approximée localement par un polynôme de faible degré. Nous approximons $m_0\left(p_{j|i}\right)$ dans le modèle en (3.3) par un polynôme de q^{e} degré, disons $m_1\left(p_{j|i}\right)$, par un développement en séries de Taylor autour de $p_{l|k}$. L'approximation est donnée par

$$m_1\left(p_{j|i}\right) = m_0\left(p_{l|k}\right) + \sum_{a=1}^{q} \frac{1}{a!} m_0\left(p_{l|k}\right)^{(a)} \left(p_{j|i} - p_{l|k}\right)^a,\ j \in s_i;\ i = 1, \ldots, M, \tag{3.5}$$

où $m_0\left(p_{l|k}\right)^{(a)}$ est la a^{e} dérivée de $m_0\left(p_{j|i}\right)$ en évaluation à $p_{l|k}$. La fonction $m_1\left(p_{j|i}\right)$ dépend de $l \in U_k$, mais nous écartons cette dépendance pour simplifier la notation.

Pour chaque point $p_{l|k}, l \in U_k;\ k = 1, \ldots, M,$ dans le modèle en (3.3), nous remplaçons $m_0\left(p_{j|i}\right)$ par son approximation $m_1\left(p_{j|i}\right)$ en (3.5). Le modèle résultant est donné par

$$y_{ij} = \tilde{\mathbf{x}}_{ij}^T\boldsymbol{\beta}_1 + m_0\left(p_{l|k}\right) + \sum_{a=1}^{q}\frac{1}{a!}m_0\left(p_{l|k}\right)^{(a)}\left(p_{j|i} - p_{l|k}\right)^a + v_{1i} + e_{1ij},\ \ j \in s_i;\ i = 1, \ldots, M. \quad (3.6)$$

Le modèle (3.6) est un modèle à approximation locale pour (3.3) qui dépend du point $l \in U_k$ de la population. Nous désignerons par $\boldsymbol{\beta}_1$ et v_{1i} les estimations de $\hat{\boldsymbol{\beta}}_{\text{loc},1}$ et $\hat{v}_{\text{loc},1i}$ en (3.6). Il convient de noter que (3.6) permet l'estimation de $m_0\left(p_{l|k}\right)$, la valeur de la fonction lisse $m_0(\cdot)$ en un point $p_{l|k}$. Nous formulons (3.6) sous la forme

$$y_{ij} = \tilde{\mathbf{x}}_{ij}^T\boldsymbol{\beta}_1 + u_0 + \sum_{a=1}^{q}u_a\left(p_{j|i} - p_{l|k}\right)^a + v_{1i} + e_{1ij} : j \in s_i;\ i = 1, \ldots, M, \quad (3.7)$$

où $u_a = m_0\left(p_{l|k}\right)^{(a)} \big/ a!$ pour $a = 0, \ldots, q$. Le modèle en (3.7) est un modèle mixte linéaire avec paramètres fixes $\left(\boldsymbol{\beta}_1,\ u_0, \ldots, u_q\right)$ et effets aléatoires de petit domaine $v_{1i},\ i = 1, \ldots, M$.

Soit $\hat{u}_0$ un estimateur de u_0 obtenu par ajustement de modèle en (3.7). Un estimateur approximé de $m_0\left(p_{l|k}\right) = u_0$ est donné par $\hat{m}_0\left(p_{l|k}\right) = \hat{u}_0$. Comme nous voulons des estimateurs de $m_0\left(p_{l|k}\right)$ pour $l \in U_k$ et $k = 1, \ldots, M$, nous utilisons $N = \sum_{i=1}^{M}N_i$ modèles (3.7). Comme l'a fait remarquer un corédacteur, si N est grand, l'estimation des valeurs de $m_0(\cdot)$ pour tous les points de la population peut être vorace en calcul.

Il est plus commode de travailler en notation matricielle. C'est pourquoi nous définissons $\mathbf{y}_i = \left(y_{i1}, \ldots, y_{in_i}\right)^T$, $\tilde{\mathbf{X}}_i = \left(\tilde{\mathbf{x}}_{i1}^T, \ldots, \tilde{\mathbf{x}}_{in_i}^T\right)^T$, $\mathbf{m}_{0,i} = \left(m_0\left(p_{1|i}\right), \ldots, m_0\left(p_{n_i|i}\right)\right)^T$, $\mathbf{v}_1 = \left(v_{11}, \ldots, v_{1M}\right)^T$ et $\mathbf{e}_{1i} = \left(e_{1i1}, \ldots, e_{1in_i}\right)^T$. Le modèle en (3.3) peut s'exprimer sous une forme matricielle par empilement des observations. L'équation résultante est

$$\mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\beta}_1 + \mathbf{m}_0 + \mathbf{Z}\mathbf{v}_1 + \mathbf{e}_1, \quad (3.8)$$

où $\mathbf{y} = \text{col}_{1\le i\le M}\left(\mathbf{y}_i\right)$, $\tilde{\mathbf{X}} = \text{col}_{1\le i\le M}\left(\tilde{\mathbf{X}}_i\right)$, $\mathbf{m}_0 = \text{col}_{1\le i\le M}\left(\mathbf{m}_{0,i}\right)$, $\mathbf{Z} = \text{diag}_{1\le i\le M}\left\{\mathbf{1}_{n_i}\right\}$ et $\mathbf{e}_1 = \text{col}_{1\le i\le M}\left(\mathbf{e}_{1i}\right)$.

Pour l'unité l du petit domaine U_k, nous définissons la $n \times (q + 1)$ matrice :

$$\mathbf{Q} = \begin{pmatrix} 1 & \left(p_{1|1} - p_{l|k}\right) & \cdots & \left(p_{1|1} - p_{l|k}\right)^q \\ \vdots & \vdots & \cdots & \vdots \\ 1 & \left(p_{n_M|M} - p_{l|k}\right) & \cdots & \left(p_{n_M|M} - p_{l|k}\right)^q \end{pmatrix},$$

où $n = \sum_{i=1}^{M}n_i$ est la taille totale d'échantillon. Soit $\mathbf{u} = (m_0(p_{l|k}), m_0^{(1)}(p_{l|k})/1!, \ldots, m_0^{(q)}\left(p_{l|k}\right)/q!)^T$ représentant le vecteur des dérivées de la fonction $m_0(\cdot)$ en évaluation à $p_{l|k}$. Les termes $\mathbf{Q}$ et $\mathbf{u}$ dépendent de l'unité $l \in U_k$ où la localisation se fait. Nous n'avons pas parlé de la dépendance à l'égard de l'unité l du petit domaine U_k pour ne pas alourdir la notation. Nous définissons le vecteur $\mathbf{m}_1$ obtenu par empilement des n valeurs de la fonction $m_1(\cdot)$ en (3.5). Ainsi, $\mathbf{m}_1 = \text{col}_{1\le i\le M}\left(\mathbf{m}_{1,i}\right)$ avec $\mathbf{m}_{1,i} = \left(m_1\left(p_{1|i}\right), \ldots, m_1\left(p_{n_i|i}\right)\right)^T$. Cela permet d'approximer $\mathbf{m}_0$ par

$\mathbf{m}_0 \approx \mathbf{m}_1$. Le vecteur $\mathbf{m}_1$ est donné par $\mathbf{m}_1 = \mathbf{Qu}$. Il s'ensuit qu'une approximation en (3.8) dans un voisinage de $l \in U_k$ est

$$\mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\beta}_1 + \mathbf{Qu} + \mathbf{Z}\mathbf{v}_1 + \mathbf{e}_1. \tag{3.9}$$

Les équations (3.8) et (3.9) sont les équivalents en forme matricielle des équations (3.3) et (3.7) respectivement. La matrice $\tilde{\mathbf{X}}$ en (3.9) ne comprend pas le terme constant représentant l'ordonnée à l'origine, puisque ce terme est déjà contenu dans $\mathbf{Q}$. L'équation (3.9) est un modèle mixte linéaire type avec des paramètres fixes $\boldsymbol{\beta}_{\text{fixe}} = (\boldsymbol{\beta}_1^T, \mathbf{u}^T)^T$ et des effets aléatoires de petit domaine $\mathbf{v}_1$. Nous employons $V(\mathbf{v}_1) = \mathbf{G} = \sigma_{1v}^2\mathbf{I}_M$, $V(\mathbf{e}_{1i}) = \mathbf{R}_i = \sigma_{1e}^2\mathbf{I}_{n_i}$ et $V(\mathbf{e}_1) = \mathbf{R} = \text{diag}_{1\le i\le M}\{\mathbf{R}_i\}$ comme matrices respectives des covariancesde $\mathbf{v}_1$, $\mathbf{e}_{1i}$ et $\mathbf{e}_1$. La matrice des covariancesde $\mathbf{y}_i$ est donnée par $V(\mathbf{y}_i) = \mathbf{V}_i = \sigma_{1v}^2\mathbf{J}_{n_i} + \sigma_{1e}^2\mathbf{I}_{n_i}$. Les matrices $\mathbf{I}_M$ et $\mathbf{I}_{n_i}$ sont les matrices identité d'ordre M et n_i respectivement, tandis que $\mathbf{J}_{n_i}$ est la matrice carrée d'ordre n_i dont tous les éléments sont égaux à 1. Il s'ensuit que $V(\mathbf{y}) = \mathbf{V} = \text{diag}_{1\le i\le M}\{\mathbf{V}_i\}$.

Posons que $\mathbf{V}$ est connu et que $\mathbf{v}_1$ et $\mathbf{e}_1$ sont en distribution normale. Par la théorie EBLUP classique, nous pouvons obtenir des estimateurs de $\boldsymbol{\beta}_{\text{fixe}}$ et $\mathbf{v}_1$ en minimisant

$$\Phi = (\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_1 - \mathbf{Qu} - \mathbf{Z}\mathbf{v}_1)^T \mathbf{R}^{-1}(\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_1 - \mathbf{Qu} - \mathbf{Z}\mathbf{v}_1) + \mathbf{v}_1^T\mathbf{G}^{-1}\mathbf{v}_1.$$

À noter que toutes les observations comprises dans Φ sont en équipondération. Il nous faut toutefois modifier Φ pour nous aligner sur la façon dont se fait l'estimation polynomialelocale. Nous nous reportons à cette fin à l'équation en (3.7) et estimons ses paramètres en associant des poids noyau $K\left((p_{j|i} - p_{l|k})/h\right)/h$ à chaque unité échantillonnée $j \in s_i$; $i = 1, \ldots, M$. Nous choisissons des valeurs de pondération noyau qui sont plus grandes pour les points d'échantillon proches de $l \in U_k$ et plus petites pour les points d'échantillon qui s'en éloignent. Le poids $K(\cdot)$ est une fonction de densité de probabilité et h est une largeur de bande qui tient compte de la taille du voisinage local. Nous expliquons à la section 3.2 comment on peut en arriver à une largeur de bande optimale. Soit $\mathbf{W}$ la matrice diagonale $n \times n$ de poids noyau par

$$\mathbf{W} = \underset{\substack{1\le j\le n_i \\ 1\le i\le M}}{\text{diag}} \left\{ \frac{1}{h} K\left(\frac{p_{j|i} - p_{l|k}}{h} \right) \right\}.$$

La matrice $\mathbf{W}$ dépend de l'unité l du petit domaine U_k et de la largeur de bande h. Nous excluons les indices $l \in U_k$ et h de la définition de la matrice $\mathbf{W}$ pour ne pas alourdir la notation. D'après Wu et Zhang (2002), l'intégration de la pondération noyau dans Φ nous amène à minimiser Φ_W, où

$$\Phi_W = (\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_1 - \mathbf{Qu} - \mathbf{Z}\mathbf{v}_1)^T \mathbf{W}^{1/2}\mathbf{R}^{-1}\mathbf{W}^{1/2}(\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_1 - \mathbf{Qu} - \mathbf{Z}\mathbf{v}_1) + \mathbf{v}_1^T\mathbf{G}^{-1}\mathbf{v}_1,$$

et où $\mathbf{W}^{1/2}$ est la racine carrée de la matrice $\mathbf{W}$.

Estimer les paramètres en (3.9) en minimisant Φ_W équivaut à estimer les paramètres donnés par

$$\mathbf{W}^{1/2}\mathbf{y} = \mathbf{W}^{1/2}\tilde{\mathbf{X}}\boldsymbol{\beta}_1 + \mathbf{W}^{1/2}\mathbf{Qu} + \mathbf{W}^{1/2}\mathbf{Z}\mathbf{v}_1 + \mathbf{e}_1. \tag{3.10}$$

L'estimation EBLUP pondérée en (3.9) avec la matrice de pondération donnée par $\mathbf{W}$ correspond à une estimation EBLUP classique venant du modèle en (3.10). Définissons $\mathbf{y}_w = \mathbf{W}^{1/2}\mathbf{y}$, $\mathbf{X}_w = [\mathbf{W}^{1/2}\tilde{\mathbf{X}},\ \mathbf{W}^{1/2}\mathbf{Q}]$ et $\mathbf{Z}_w = \mathbf{W}^{1/2}\mathbf{Z}$. L'équation (3.10) peut se reformuler comme

$$\mathbf{y}_w = \mathbf{X}_w\boldsymbol{\beta}_{\text{fixe}} + \mathbf{Z}_w\mathbf{v}_1 + \mathbf{e}_1. \tag{3.11}$$

Soit $\hat{\boldsymbol{\beta}}_{\text{loc,fixe}} = \left(\hat{\boldsymbol{\beta}}_{\text{loc},1}^T, \hat{\mathbf{u}}^T\right)^T$ et $\hat{\mathbf{v}}_{\text{loc},1} = \left(\hat{v}_{\text{loc},11}, \ldots, \hat{v}_{\text{loc},1M}\right)^T$ les estimateurs EBLUP des effets fixes et aléatoires en (3.11). Les estimateurs $\hat{\boldsymbol{\beta}}_{\text{loc,fixe}}$ et $\hat{\mathbf{v}}_{\text{loc},1}$ sont fondés sur les estimateurs locaux des composantes de la variance $\left(\sigma_{1v}^2, \sigma_{1e}^2\right)$. Les estimateurs de ces composantes, désignés par $\left(\hat{\sigma}_{\text{loc},1v}^2, \hat{\sigma}_{\text{loc},1e}^2\right)$, s'obtiennent par la méthode HFC ou MVC avec le modèle en (3.11). Comme $\mathbf{u} = \left(m_0\left(p_{l|k}\right), m_0^{(1)}\left(p_{l|k}\right)/1!, \ldots, m_0^{(q)}\left(p_{l|k}\right)/q!\right)^T$, un estimateur $\hat{m}_0\left(p_{l|k}\right)$ de $m_0\left(p_{l|k}\right)$ est la première composante $\hat{u}_0$ de $\hat{\mathbf{u}}$.

Notons que $\hat{\boldsymbol{\beta}}_{\text{loc},1}$, $\hat{m}_0\left(p_{l|k}\right)$ et $\hat{v}_{\text{loc},1k}$ pourraient servir à l'obtention d'estimations locales $\hat{y}_{\text{loc},kl}$ pour la valeur inconnue y_{kl}, où $\hat{y}_{\text{loc},kl} = \tilde{\mathbf{x}}_{kl}^T\hat{\boldsymbol{\beta}}_{\text{loc},1} + \hat{m}_0\left(p_{l|k}\right) + \hat{v}_{\text{loc},1k}$ pour $l \in \overline{s}_k$. Toutefois, un examinateur a signalé que, dans la pratique, ce cadre méthodologique ne serait sans doute pas d'un bon comportement, parce qu'il faut un solide équilibre des petits domaines sur tout l'éventail des probabilités $p_{l|k}$. Si l'équilibre n'est pas sauvegardé, l'estimation obtenue souffrirait grandement de cette localisation. C'est pourquoi nous avons opté pour une estimation globale de $\boldsymbol{\beta}_1$ et $\mathbf{v}_1$.

Expliquons maintenant la deuxième étape de notre procédure. Il est possible d'estimer globalement les paramètres $\boldsymbol{\beta}_1$ et $\mathbf{v}_1$ selon les estimations $\hat{m}_0\left(p_{j|i}\right)$ et les données auxiliaires $\tilde{\mathbf{x}}_{ij}$ liées aux unités de l'échantillon. Pour $j \in s_i$ et $i = 1, \ldots, M$, définissons une nouvelle variable, disons ξ, comme

$$\xi_{ij} = y_{ij} - \hat{m}_0\left(p_{j|i}\right),\ j \in s_i;\ i = 1, \ldots, M.$$

Les n valeurs ξ_{ij} représentent les différences entre les y_{ij} observés et leurs estimateurs locaux $\hat{m}_0\left(p_{j|i}\right)$. Avec le modèle en (3.3), ξ satisfait le modèle suivant

$$\xi_{ij} = \tilde{\mathbf{x}}_{ij}^T\boldsymbol{\beta}_{\text{glo},1} + v_{\text{glo},1i} + e_{\text{glo},1ij},\ j \in s_i;\ i = 1, \ldots, M, \tag{3.12}$$

où $v_{\text{glo},1i} \sim N\left(0, \sigma_{\text{glo},1v}^2\right)$ et $e_{\text{glo},1ij} \sim N\left(0, \sigma_{\text{glo},1e}^2\right)$. La mention glo en indice indique que (3.12) est un modèle global.

Comme (3.12) représente un modèle mixte linéaire paramétrique, nous pouvons prendre l'estimation EBLUP classique (hors pondération) pour estimer ses paramètres. Soit $\hat{\boldsymbol{\beta}}_{\text{glo},1}$ et $\hat{v}_{\text{glo},1i}$ les meilleurs estimateurs linéaires sans biais empiriques de $\boldsymbol{\beta}_{\text{glo},1}$ et $v_{\text{glo},1i}$. Soit $\left(\hat{\sigma}_{\text{glo},1v}^2, \hat{\sigma}_{\text{glo},1e}^2\right)$ les estimateurs des composantes de la variance $\left(\sigma_{\text{glo},1v}^2, \sigma_{\text{glo},1e}^2\right)$, où la méthode HFC ou MVC peut servir à l'estimation de ces mêmes paramètres. Nous estimons $\left(\boldsymbol{\beta}_1, v_{1i}, \sigma_{1v}^2, \sigma_{1e}^2\right)$ du modèle en (3.3) par $(\hat{\boldsymbol{\beta}}_{\text{glo},1}, \hat{v}_{\text{glo},1i}, \hat{\sigma}_{\text{glo},1v}^2, \hat{\sigma}_{\text{glo},1e}^2)$ et le modèle en (3.12). Les estimateurs globaux $\hat{\boldsymbol{\beta}}_{\text{glo},1}$, $\hat{v}_{\text{glo},1i}$ et $\left(\hat{\sigma}_{\text{glo},1v}^2, \hat{\sigma}_{\text{glo},1e}^2\right)$ sont exempts de tout biais causé par un plan de sondage informatif, parce que ξ_{ij} n'est plus lié aux $p_{j|i}$ après conditionnement par $\mathbf{x}_{ij}$.

En troisième étape, nous estimons les valeurs y_{ij} inobservées pour $j \in \overline{s}_i$ et $i = 1, \ldots, M$ par insertion dans l'équation (3.4); il s'agit i. des estimateurs locaux $\hat{m}_0\left(p_{j|i}\right)$ pour $j \in \overline{s}_i$ en première étape et ii. des estimateurs globaux $\hat{\boldsymbol{\beta}}_{\text{glo},1}$ et $\hat{v}_{\text{glo},1i}$ en deuxième étape. Les $\hat{y}_{ij}$ dégagés pour $j \in \overline{s}_i$ sont

insérés dans (3.2) pour le calcul de l'estimateur $\hat{\bar{Y}}_i$. À noter que $\hat{\bar{Y}}_i$ demande que $\tilde{\mathbf{x}}_{ij}$ et $p_{j|i}$ soient connus pour toutes les unités de la population. Un examinateur a fait observer que, dans la pratique, cette hypothèse pourrait venir limiter l'applicabilité de la méthode proposée, ce qui pourrait se résoudre comme problème si les organismes statistiques nationaux donnaient accès aux probabilités de sélection de toutes les unités, de telles valeurs pouvant être nécessaires à des applications comme la nôtre.

3.2 Sélection de largeur de bande

Les polynômes locaux exigent que soient spécifiés le noyau $K(\cdot)$, l'ordre de l'ajustement polynomial q et la largeur de bande h. Fan et Gijbels (1996) indiquent que les valeurs de q supérieures à l'unité n'apportent pas une amélioration significative par rapport à l'ajustement linéaire $(q = 1)$. Fan et Gijbels (1996) indiquent en outre que le choix de h est bien plus important que le degré du polynôme. Dans ce qui suit, nous emploierons un noyau de densité normale avec q égal à un, puisque cela mène à des résultats satisfaisants dans la plupart des applications.

Nous établissons le h optimal par la méthode de validation croisée (CV). Pour un h donné, calculons l'estimateur de y_{ij} en (3.4) à l'aide de l'échantillon qui reste une fois la j^{e} unité retranchée de s_i. Si nous désignons l'estimateur résultant de y_{ij} par $\tilde{y}_{ij}$, nous définissons à la suite de Wu et Zhang (2002) le critère CV comme

$$\text{CV}(h) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{n_i}\sum_{j\in s_i}(y_{ij} - \tilde{y}_{ij})^2 .$$

Le terme $1/n_i$ tient compte du nombre d'observations dans le petit domaine U_i. Nous obtenons la largeur de bande optimale h_{opt} en minimisant le $\text{CV}(h)$. Étant donné h_{opt}, l'estimateur polynomial local de la moyenne de petit domaine $\bar{Y}_i$ en (3.2) est désigné par $\hat{\bar{Y}}_i^{\text{PL}}$.

4 Estimation de l'EQM par le bootstrap

L'estimation de l'EQM des estimateurs de petit modèle est un problème épineux même avec des estimateurs EBLUP classiques. La théorie générale EBLUP prévoit une approximation finie de $\text{EQM}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right)$ par voie de linéarisation. Un estimateur peut s'obtenir par cette approximation pour $\text{EQM}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right)$ (voir les détails dans Prasad et Rao, 1990). Verret et coll. (2015) ont procédé par approximation finie pour dégager l'estimateur d'erreur quadratique moyenne pour $\hat{\bar{Y}}_i^{\text{VRH}}$ en (2.9), chose possible parce que l'estimateur $\hat{\bar{Y}}_i^{\text{VRH}}$ est un estimateur EBLUP type sur modèle mixte linéaire assorti de la variable supplémentaire connue $g(p_{j|i})$. On n'a besoin d'aucune théorie nouvelle pour estimer l'EQM de $\hat{\bar{Y}}_i^{\text{VRH}}$. Dans notre cas et compte tenu pour l'estimation locale répétée du modèle en (3.6), il est impossible d'obtenir une approximation finie de l'erreur quadratique moyenne de $\hat{\bar{Y}}_i^{\text{PL}}$, $\text{EQM}\left(\hat{\bar{Y}}_i^{\text{PL}}\right)$, ni pour son estimateur $\text{eqm}\left(\hat{\bar{Y}}_i^{\text{PL}}\right)$. Nous avons employé deux versions de la méthode bootstrap pour estimer l'EQM des estimateurs de petit domaine dont il a été question jusqu'ici. Pour estimer l'EQM de $\hat{\bar{Y}}_i^{\text{EBLUP}}$, nous avons opté pour un bootstrap *inconditionnel*, alors que, pour $\hat{\bar{Y}}_i^{\text{PL}}$, $\hat{\bar{Y}}_i^{\text{VRH1}}$ et $\hat{\bar{Y}}_i^{\text{VRH2}}$, notre bootstrap était *conditionnel*. Nous allons décrire comment se calcule chaque type de bootstrap.

Décrivons d'abord le bootstrap inconditionnel. C'est là une variante du bootstrap paramétrique de Hall et Maiti (2006) qui a été proposé par González-Manteiga, Lombardia, Molina, Morales et Santamaria (2008). Cette méthode peut servir à l'estimation de l'EQM de $\hat{\bar{Y}}_i^{\text{EBLUP}}$ avec le modèle en (1.1), parce que les estimations des divers paramètres du modèle (1.1) ne dépendent pas des probabilités de sélection $p_{j|i} : j \in s_i;\ i = 1, \ldots, M$. Nous prédisons les valeurs y en générant $v_i^* \sim N\left(0, \hat{\sigma}_v^2\right)$ et $e_{ij}^* \sim N\left(0, \hat{\sigma}_e^2\right)$, où $\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right)$ sont les estimateurs HFC ou MVC de $\left(\sigma_v^2, \sigma_e^2\right)$. Par l'estimateur EBLUP $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$, nous obtenons les valeurs bootstrap de y_{ij} comme

$$y_{ij}^* = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + v_i^* + e_{ij}^*,\ j \in U_i;\ i = 1, \ldots, M. \tag{4.1}$$

La version bootstrap du paramètre cible $\bar{Y}_i$ se calcule comme $\bar{Y}_i^* = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}^*$. La version bootstrap de l'estimateur EBLUP $\hat{\bar{Y}}_i^{\text{EBLUP}}$ est donnée par

$$\hat{\bar{Y}}_i^{\text{EBLUP*}} = \frac{1}{N_i}\left(\sum_{j \in s_i} y_{ij}^* + \sum_{j \in \bar{s}_i} \hat{y}_{ij}^*\right),$$

où $\hat{y}_{ij}^* = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}^* + \hat{v}_i^*$ et où $\left(\hat{\boldsymbol{\beta}}^*, \hat{v}_i^*\right)$ sont les estimateurs EBLUP de $(\boldsymbol{\beta}, v_i)$ selon $\left(y_{ij}^*,\ \mathbf{x}_{ij}\right),\ j \in s_i$, pour $i = 1, \ldots, M$. Si nous reprenons B fois cette procédure, l'estimateur bootstrap de $\text{EQM}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right)$ est

$$\text{eqm}_{\text{boot}}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\bar{Y}}_i^{\text{EBLUP*}}(b) - \bar{Y}_i^*(b)\right)^2, \tag{4.2}$$

où $\hat{\bar{Y}}_i^{\text{EBLUP*}}(b)$ et $\bar{Y}_i^*(b)$ sont les valeurs de $\hat{\bar{Y}}_i^{\text{EBLUP*}}$ et $\bar{Y}_i^*$ pour la b^{e} itération bootstrap. Comme les estimateurs $\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}_v^2, \hat{\sigma}_e^2\right)$ sont sérieusement entachés d'un biais à cause du plan de sondage informatif, nous prévoyons que $\text{eqm}_{\text{boot}}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right)$ sera un estimateur biaisé de $\text{EQM}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right)$, et ce, parce qu'il est fondé sur le modèle de population en (1.1) et que ce modèle ne vaut pas pour l'échantillon.

Passons maintenant à l'estimation de $\text{EQM}\left(\hat{\bar{Y}}_i^{\text{PL}}\right)$ par le bootstrap conditionnel. Rappelons-nous que $\hat{\bar{Y}}_i^{\text{PL}}$ repose sur le modèle augmenté en (3.3). Il est donc naturel d'utiliser ce modèle au moment de juger de la précision de l'estimateur polynomial local. Il est impossible d'employer le bootstrap inconditionnel paramétrique, car il faudrait produire des valeurs bootstrap $\left(y_{ij}^*, p_{j|i}^*\right)$ tant pour y_{ij} que pour $p_{j|i}$, d'où l'implication que nous devrions savoir comment les y_{ij} sont liés aux probabilités de sélection $p_{j|i}$. Comme l'a fait remarquer le corédacteur, la relation entre y_{ij} et $p_{j|i}$ n'est pas précisément connue dans la pratique. Nous avons donc choisi de garder les probabilités de sélection $p_{j|i}$ de l'échantillon initial et de produire des valeurs bootstrap uniquement pour la variable réponse y_{ij}. Le bootstrap ainsi obtenu est conditionnel à $p_{j|i},\ j \in U_i;\ i = 1, \ldots, M$; c'est la raison pour laquelle nous parlons ici de *bootstrap conditionnel paramétrique*. Rao, Sinha et Dumitrescu (2014) s'en sont déjà servis et Chatrchi (2018) l'a fait plus récemment à son tour pour estimer l'EQM d'un modèle mixte spline pénalisé.

Dans notre contexte, nous procédons de la manière suivante pour estimer $\text{EQM}\left(\hat{\bar{Y}}_i^{\text{PL}}\right)$. Nous générons $v_{1i}^* \sim N\left(0, \hat{\sigma}_{\text{glo},1v}^2\right)$ et $e_{1ij}^* \sim N\left(0, \hat{\sigma}_{\text{glo},1e}^2\right)$ et obtenons les réponses bootstrap

$$y_{1ij}^* = \tilde{\mathbf{x}}_{ij}^T \hat{\boldsymbol{\beta}}_{\text{glo},1} + \hat{m}_0\left(p_{j|i}\right) + v_{1i}^* + e_{1ij}^*,\ j \in U_i;\ i = 1, \ldots, M. \tag{4.3}$$

Nous avons estimé les $\hat{m}_0\left(p_{j|i}\right)$ par le modèle local en (3.6). Nous avons estimé le triplet $(\hat{\boldsymbol{\beta}}_{\text{glo},1}, \hat{\sigma}^2_{\text{glo},1v}, \hat{\sigma}^2_{\text{glo},1e})$ par le modèle global en (3.12) et les données de l'échantillon $\left(y_{ij}, \tilde{\mathbf{x}}_{ij}, p_{j|i}\right)$, $j \in s_i$; $i = 1, \ldots, M$. La moyenne bootstrap de population est $\bar{Y}^*_{1i} = N_i^{-1} \sum_{j=1}^{N_i} y^*_{1ij}$. Soit $\hat{\boldsymbol{\beta}}^*_{\text{glo},1}, \hat{m}^*_0\left(p_{j|i}\right)$ et $\hat{v}^*_{\text{glo},1i}$ les versions bootstrap des estimateurs $\hat{\boldsymbol{\beta}}_{\text{glo},1}, \hat{m}_0\left(p_{j|i}\right)$ et $\hat{v}_{\text{glo},1i}$ selon les données bootstrap $\left(y^*_{1ij}, \tilde{\mathbf{x}}_{ij}, p_{j|i}\right)$, $j \in s_i$; $i = 1, \ldots, M$, et le h_{opt} tiré de l'ensemble de données initial $\left(y_{ij}, \tilde{\mathbf{x}}_{ij}, p_{j|i}\right)$, $j \in s_i$; $i = 1, \ldots, M$. Nous n'avons pas recalculé le h^*_{opt} optimal lié à $\left(y^*_{1ij}, \tilde{\mathbf{x}}_{ij}, p_{j|i}\right)$, $j \in s_i$; $i = 1, \ldots, M$, parce que trop de calculs devraient s'ensuivre dans l'étude de Monte Carlo. La procédure bootstrap est donc conditionnelle à $p_{j|i}$, $j \in U_i$; $i = 1, \ldots, M$, tout comme à h_{opt} tiré de l'échantillon initial. Comme $\bar{s}_i$ est l'ensemble d'unités non échantillonnées dans le domaine i, les valeurs bootstrap prédites $\hat{y}^*_{1ij}$ pour $j \in \bar{s}_i$ s'obtiennent comme

$$\hat{y}^*_{1ij} = \tilde{\mathbf{x}}^T_{ij} \hat{\boldsymbol{\beta}}^*_{\text{glo},1} + \hat{m}^*_0\left(p_{j|i}\right) + \hat{v}^*_{\text{glo},1i}. \tag{4.4}$$

L'estimateur résultant de $\bar{Y}^*_{1i}$ est

$$\hat{\bar{Y}}^*_{1i} = \frac{1}{N_i}\left(\sum_{j \in s_i} y^*_{1ij} + \sum_{j \in \bar{s}_i} \hat{y}^*_{1ij}\right).$$

Si nous répétons cette procédure B fois, l'estimateur bootstrap conditionnel de l'EQM de l'estimateur polynomial local de $\bar{Y}_i$ est donné par

$$\text{eqm}_{\text{boot}}\left(\hat{\bar{Y}}^{\text{PL}}_i\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\bar{Y}}^*_{1i}(b) - \bar{Y}^*_{1i}(b)\right)^2, \tag{4.5}$$

où $\hat{\bar{Y}}^*_{1i}(b)$ et $\bar{Y}^*_{1i}(b)$ sont les valeurs de $\hat{\bar{Y}}^*_{1i}$ et $\bar{Y}^*_{1i}$ pour la b^{e} itération bootstrap.

Le bootstrap conditionnel peut aussi servir à l'estimation de l'erreur quadratique moyenne d'un estimateur EBLUP, $\hat{\bar{Y}}^{\text{VRH}}_i$, avec le modèle augmenté (1.2) proposé par Verret et coll. (2015). Nous avons inclus cette procédure dans la simulation de la section 5 pour donner une idée de la façon dont les estimateurs résultants de l'EQM se comparent aux estimateurs obtenus pour $\hat{\bar{Y}}^{\text{PL}}_i$. Les étapes du calcul de $\text{eqm}\left(\hat{\bar{Y}}^{\text{VRH}}_i\right)$ sont semblables à celles de l'obtention de l'erreur quadratique moyenne de l'estimateur polynomial local $\hat{\bar{Y}}^{\text{PL}}_i$. Dans ce cas, les valeurs bootstrap des réponses y_{ij} reposent sur le modèle augmenté en (1.2) et les estimateurs $\left(\hat{\boldsymbol{\beta}}_0, \hat{\delta}_0\right)$ et $\left(\hat{\sigma}^2_{0v}, \hat{\sigma}^2_{0e}\right)$ obtenus lorsque la théorie EBLUP classique s'utilise avec les données d'échantillon $\left(y_{ij}, \mathbf{x}_{ij}, g\left(p_{j|i}\right)\right)$, $j \in s_i$; $i = 1, \ldots, M$.

5 Étude de simulation

Le paramétrage de cette étude par simulation suit celui de Verret et coll. (2015). Nous avons considéré une population comptant $M = 15$ petits domaines et $N_i = 15$ unités par petit domaine. Nous avons opté pour ces valeurs relativement faibles de domaines et d'unités pour alléger les calculs. Nous avons employé une seule variable auxiliaire x. Nous avons généré les valeurs x de population à partir d'une distribution gamma à moyenne 10 et à variance 50. Nous avons produit les valeurs y_{ij} de population par le modèle

$$y_{ij} = 4 + x_{ij} + v_i + e_{ij};\ i = 1, \ldots, 15;\ j = 1, \ldots, 15, \tag{5.1}$$

où $v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$ et $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$ avec $\sigma_v^2 = 0{,}5$ et $\sigma_e^2 = 2$.

Nous avons pris en compte une taille unique d'échantillon, $n_i = 3$, à l'intérieur d'un petit domaine. Nous avons procédé par échantillonnage de Poisson conditionnel (EPC) pour prélever des échantillons non équiprobables dans les petits domaines, les probabilités étant proportionnelles aux tailles spécifiées c_{ij} (voir Tillé, 2006, chapitre 5). Nous avons examiné deux choix de tailles c_{ij} dans l'étude de simulation. Le premier choix était

$$c_{ij} = \exp\left[\frac{1}{3}\left\{-\frac{(v_i + e_{ij})}{\sigma_e} + \frac{\delta_{ij}}{5}\right\}\right], \tag{5.2}$$

où $\delta_{ij} \overset{\text{iid}}{\sim} N(0, 1)$. Les mesures de taille en (5.2) équivalent à celles de Pfeffermann et Sverchkov (2007) dans leur propre étude de simulation et satisfont la relation en (2.5) pour les poids $w_{j|i} = \pi_{j|i}^{-1}$.

Dans un second choix de mesures de taille d'après Asparouhov (2006), nous prenons deux mesures invariante (I) et non invariante (NI). Dans la mesure invariante, c_{ij} est indépendant de v_i étant donné $\mathbf{x}_{ij}$, sinon il s'agit de la mesure non invariante. Les mesures de taille invariantes sont données par

$$c_{ij} = \left(1 + \exp\left\{-\tau\left(\frac{1}{\alpha}e_{ij} + \sqrt{1 - \frac{1}{\alpha^2}}\, e_{ij}^*\right)\right\}\right)^{-1}. \tag{5.3}$$

Les mesures non invariantes le sont par

$$c_{ij} = \left(1 + \exp\left\{-\tau\left[\frac{1}{\alpha}(v_i + e_{ij}) + \sqrt{1 - \frac{1}{\alpha^2}}\,(v_i^* + e_{ij}^*)\right]\right\}\right)^{-1}, \tag{5.4}$$

où la paire aléatoire (v_i^*, e_{ij}^*) se génère indépendamment de (v_i, e_{ij}) par les mêmes distributions comme v_i et e_{ij}. Ce sont les mesures de taille retenues par Asparouhov (2006). Le coefficient τ permet de tenir compte de la variation des poids et la valeur α, du degré de contenu informatif du plan de sondage. Nous avons choisi $\tau = 0{,}5$ et $\alpha = 1; 2; 3$ et ∞ correspondant à la pluralité de degrés de contenu informatif venant de c_{ij} en (5.3) et (5.4). Si α augmente, le contenu informatif diminue; $\alpha = \infty$ correspond à un échantillonnage non informatif. Si certains $\pi_{j|i}$ dépassaient l'unité, nous les avons fixés à un et recalculé les probabilités pour les unités restantes.

5.1 Rendement de l'estimateur polynomial local de $\bar{Y}_i$

Nous avons comparé le biais et l'erreur quadratique moyenne des estimateurs $\hat{\bar{Y}}_i^{\text{EBLUP}}$, $\hat{\bar{Y}}_i^{\text{VRH}}$ et $\hat{\bar{Y}}_i^{\text{PL}}$. L'estimateur EBLUP $\hat{\bar{Y}}_i^{\text{EBLUP}}$ fondé sur (1.1) présuppose que le modèle d'échantillon coïncide avec le modèle de population, faisant ainsi abstraction du contenu informatif du plan d'échantillonnage. Nous avons étudié deux versions de $\hat{\bar{Y}}_i^{\text{VRH}}$ examinées par Verret et coll. (2015) pour divers choix de $g(\cdot)$ rendant compte du contenu informatif. Il s'agit d'estimateurs EBLUP fondés sur le modèle d'échantillon augmenté en (1.2). Ils sont désignés par $\hat{\bar{Y}}_i^{\text{VRH1}}$ quand $g(p_{j|i}) = p_{j|i}$ et par $\hat{\bar{Y}}_i^{\text{VRH2}}$ quand $g(p_{j|i}) = \log(p_{j|i})$. Nous présentons les résultats seulement pour ces fonctions g, car elles sont d'un meilleur rendement que les autres fonctions dans Verret et coll. (2015). Précisons enfin que $\hat{\bar{Y}}_i^{\text{PL}}$ représente notre nouvel estimateur polynomial local.

Nous avons calculé le biais et l'erreur quadratique moyenne des estimateurs à l'aide de $R = 1\,000$ échantillons en simulation prélevés dans un traitement plan-modèle. Dans chaque passage, $r = 1, \ldots, R$, nous avons d'abord produit les valeurs y_{ij} de population selon le modèle de population en (5.1) et calculé $\bar{Y}_i^{(r)}$, la moyenne du petit domaine i dans la r^{e} population générée. Nous avons ensuite tiré des échantillons de taille $n_i = 3$ à l'intérieur des petits domaines dans un échantillonnage de Poisson conditionnel où les probabilités étaient proportionnelles aux tailles spécifiées $c_{ij}^{(r)}$ en (5.2) pour les mesures de taille PS de Pfeffermann et Sverchkov (2007) et en (5.3) et (5.4) pour les mesures invariantes et non invariantes AP d'Asparouhov (2006). Avec chaque échantillon r en simulation $(r = 1, \ldots, R)$, nous avons établi les estimations $\hat{\bar{Y}}_i^{\text{EBLUP}(r)}$, $\hat{\bar{Y}}_i^{\text{VRH1}(r)}$, $\hat{\bar{Y}}_i^{\text{VRH2}(r)}$ et $\hat{\bar{Y}}_i^{\text{PL}(r)}$ pour chaque petit domaine U_i. Nous avons trouvé une largeur de bande optimale $h_{\text{opt}}^{(r)}$ pour $\hat{\bar{Y}}_i^{\text{PL}(r)}$ en appliquant le critère de validation croisée. Une grille de la forme (0,01; 0,02; 0,03; …; 0,15) nous a donné les valeurs possibles de $h_{\text{opt}}^{(r)}$ dans les populations générées en (5.1).

Pour un estimateur donné de la moyenne de petit domaine $\bar{Y}_i$, nous avons considéré les mesures de rendement suivantes :

Biais absolu moyen

$$\overline{\text{BA}} = \frac{1}{M}\sum_{i=1}^{M} \text{BA}_i,$$

où

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

Racine de l'erreur quadratique moyenne (REQM)

$$\overline{\text{REQM}} = \frac{1}{M}\sum_{i=1}^{M} \sqrt{\frac{1}{R}\sum_{r=1}^{R} \left(\hat{\bar{Y}}_i^{(r)} - \bar{Y}_i^{(r)} \right)^2}.$$

Le tableau 5.1 présente le biais absolu moyen $\left(\overline{\text{BA}}\right)$ des estimateurs $\hat{\bar{Y}}_i^{\text{EBLUP}}$, $\hat{\bar{Y}}_i^{\text{VRH1}}$, $\hat{\bar{Y}}_i^{\text{VRH2}}$ et $\hat{\bar{Y}}_i^{\text{PL}}$ avec les mesures de taille PS en (5.2) et AP en (5.3 et 5.4) pour $\alpha = 1; 2; 3$ et ∞.

Tableau 5.1
Biais absolu moyen $\left(\overline{\text{BA}}\right)$ des mesures de taille PS et AP

Génération de $p_{j\|i}$ / Estimateur			$\hat{\bar{Y}}_i^{\text{EBLUP}}$ sans $g(p_{j\|i})$	$\hat{\bar{Y}}_i^{\text{VRH1}}$ $g(p_{j\|i}) = p_{j\|i}$	$\hat{\bar{Y}}_i^{\text{VRH2}}$ $g(p_{j\|i}) = \log(p_{j\|i})$	$\hat{\bar{Y}}_i^{\text{PL}}$ $m_0(p_{j\|i})$
PS			0,309	0,020	0,004	0,011
AP	$\alpha = 1$	I	0,431	0,002	0,036	0,004
		NI	0,425	0,010	0,035	0,005
	$\alpha = 2$	I	0,206	0,017	0,022	0,024
		NI	0,219	0,019	0,016	0,016
	$\alpha = 3$	I	0,139	0,005	0,012	0,033
		NI	0,137	0,008	0,013	0,019
	$\alpha = \infty$	I	0,008	0,008	0,008	0,026
		NI	0,006	0,006	0,006	0,021

Comme on le fait observer dans Verret et coll. (2015), le $\overline{\text{BA}}$ de l'estimateur EBLUP $\hat{\bar{Y}}_i^{\text{EBLUP}}$ avec la seule variable auxiliaire x est bien plus élevé que les estimateurs fondés sur les modèles augmentés ($p_{j|i}$ et $\log(p_{j|i})$) et la méthode polynomiale locale. Cela se vérifie quelle que soit la façon dont les mesures de taille sont générées (PS ou AP). Le $\overline{\text{BA}}$ de $\hat{\bar{Y}}_i^{\text{EBLUP}}$ prend sa plus grande valeur (0,431) lorsque le plan de sondage est très informatif ($\alpha = 1$) et diminue à mesure que α augmente. Cette observation vaut également pour les estimateurs fondés sur les modèles augmentés. L'inclusion de $p_{j|i}$ ou de $\log(p_{j|i})$ en variable d'augmentation dans le modèle donne de petites valeurs $\overline{\text{BA}}$, la plus haute étant de 0,036. Si nous comparons les $\overline{\text{BA}}$ de l'estimateur polynomial local $\hat{\bar{Y}}_i^{\text{PL}}$ aux $\overline{\text{BA}}$ des modèles augmentés VRH, nous constatons qu'ils sont comparables pour $\alpha = 1$ et $\alpha = 2$ et un peu plus élevés pour $\alpha \geq 3$.

Le tableau 5.2 présente les données de simulation de la racine de l'erreur quadratique moyenne $\left(\overline{\text{REQM}}\right)$ des estimateurs pour les mesures de taille PS en (5.2) et AP en (5.3 et 5.4) et pour $\alpha = 1; 2; 3$ et ∞. L'estimateur EBLUP, $\hat{\bar{Y}}_i^{\text{EBLUP}}$, avec le modèle (1.1) et sans la variable d'augmentation $g(p_{j|i})$ a les valeurs $\overline{\text{REQM}}$ les plus hautes (0,740 pour I et 0,752 pour NI) avec les mesures de taille AP correspondant à $\alpha = 1$; la valeur est de 0,685 pour les mesures de taille PS. Les valeurs $\overline{\text{REQM}}$ décroissent à mesure que croît α avec 0,608 pour I et 0,610 pour NI dans le cas d'un échantillonnage non informatif ($\alpha = \infty$). Les valeurs $\overline{\text{REQM}}$ pour $\hat{\bar{Y}}_i^{\text{VRH1}}$, $\hat{\bar{Y}}_i^{\text{VRH2}}$ et $\hat{\bar{Y}}_i^{\text{PL}}$ sont significativement moindres que celles de $\hat{\bar{Y}}_i^{\text{EBLUP}}$ dans le cas d'un échantillonnage très informatif ($\alpha = 1$) et pour la mesure de taille PS. Il existe de légères différences de $\overline{\text{REQM}}$ entre la méthode non paramétrique et la méthode paramétrique dans Verret et coll. (2015).

Tableau 5.2
Racine de l'erreur quadratique moyenne $\left(\overline{\text{REQM}}\right)$ pour les mesures de taille PS et AP

Génération de $p_{j\|i}$ \ Estimateur			$\hat{\bar{Y}}_i^{\text{EBLUP}}$ sans $g(p_{j\|i})$	$\hat{\bar{Y}}_i^{\text{VRH1}}$ $g(p_{j\|i}) = p_{j\|i}$	$\hat{\bar{Y}}_i^{\text{VRH2}}$ $g(p_{j\|i}) = \log(p_{j\|i})$	$\hat{\bar{Y}}_i^{\text{PL}}$ $m_0(p_{j\|i})$
PS			0,685	0,229	0,200	0,200
AP	$\alpha = 1$	I	0,740	0,089	0,170	0,087
		NI	0,752	0,158	0,200	0,149
	$\alpha = 2$	I	0,644	0,562	0,568	0,557
		NI	0,650	0,557	0,555	0,555
	$\alpha = 3$	I	0,617	0,588	0,591	0,612
		NI	0,619	0,587	0,589	0,607
	$\alpha = \infty$	I	0,608	0,619	0,621	0,626
		NI	0,610	0,622	0,625	0,629

Quand l'échantillonnage est moins informatif ($\alpha = 3$), l'estimateur linéaire local $\hat{\bar{Y}}_i^{\text{PL}}$ est meilleur que l'estimateur $\hat{\bar{Y}}_i^{\text{EBLUP}}$, mais ses valeurs $\overline{\text{REQM}}$ sont un peu plus élevées que celles des estimateurs

paramétriques $\hat{\bar{Y}}_i^{\text{VRH1}}$ et $\hat{\bar{Y}}_i^{\text{VRH2}}$. Nous observons dans ce cas que la fonction estimée $m_0\left(p_{j|i}\right)$ est proche d'un tracé plat, d'où l'implication que l'approximation linéaire locale ne convient pas aussi bien. Cela explique que $\hat{\bar{Y}}_i^{\text{PL}}$ soit légèrement pire que $\hat{\bar{Y}}_i^{\text{VRH1}}$ et $\hat{\bar{Y}}_i^{\text{VRH2}}$ avec un faible degré de contenu informatif de l'échantillonnage. Un estimateur polynomial local est d'un bon rendement lorsque la fonction $m_0(\cdot)$ est significativement non constante.

Avec un échantillon non informatif $(\alpha = \infty)$, $\hat{\bar{Y}}_i^{\text{EBLUP}}$ l'emporte sur $\hat{\bar{Y}}_i^{\text{VRH1}}$, $\hat{\bar{Y}}_i^{\text{VRH2}}$ et $\hat{\bar{Y}}_i^{\text{PL}}$ dans les mesures invariantes ou non. Cette conclusion s'écarte quelque peu de celle de Verret et coll. (2015), là où, pour $\alpha = \infty$, leurs estimateurs $\hat{\bar{Y}}_i^{\text{EBLUP}}$, $\hat{\bar{Y}}_i^{\text{VRH1}}$ et $\hat{\bar{Y}}_i^{\text{VRH2}}$ ont des valeurs égales $\overline{\text{BA}}$ et $\overline{\text{REQM}}$. Verret et coll. (2015) employaient des populations et des échantillons plus grands, ce qui pourrait expliquer que leurs modèles augmentés aient produit des estimations aussi bonnes que le modèle de population avec des plans d'échantillonnage non informatifs. Dans notre paramétrage de simulation, nous avons constaté que $\overline{\text{BA}}$ et $\overline{\text{REQM}}$ pour l'EBLUP sont petits lorsque les valeurs α sont de plus de 6, ce qui correspond à un plan de sondage presque non informatif. Nous recommandons en pareil cas d'utiliser l'estimateur EBLUP.

5.2 Rendement des estimateurs de l'EQM

Considérons maintenant le rendement des méthodes bootstrap d'estimation de l'EQM des estimateurs EBLUP et VRH et de l'estimateur polynomial local. Soit $\hat{\bar{Y}}_i$ un estimateur de $\bar{Y}_i$ et $\text{eqm}_{\text{boot}}\left(\hat{\bar{Y}}_i\right)$ l'estimateur bootstrap de $\text{EQM}\left(\hat{\bar{Y}}_i\right)$. En prévoyant $R = 1\,000$ populations et échantillons en simulation, nous avons d'abord pris la mesure des valeurs EQM comme

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

où $\bar{Y}_i^{(r)}$ est la moyenne réelle et où $\hat{\bar{Y}}_i^{(r)}$ est la valeur de l'estimateur pour la r^{e} population. Soit $\text{eqm}_{\text{boot}}\left(\hat{\bar{Y}}_i\right)$ l'estimateur bootstrap de $\text{EQM}\left(\hat{\bar{Y}}_i\right)$. Il est désigné par $\text{eqm}_{\text{boot}}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right)$ pour l'estimateur EBLUP $\hat{\bar{Y}}_i^{\text{EBLUP}}$ et correspond à la méthode bootstrap paramétrique (inconditionnelle) à l'équation (4.2). Pour notre estimateur polynomial local $\hat{\bar{Y}}_i^{\text{PL}}$ et les estimateurs de Verret et coll. (2015), $\hat{\bar{Y}}_i^{\text{VRH1}}$ et $\hat{\bar{Y}}_i^{\text{VRH2}}$, les valeurs d'erreur quadratique moyenne, désignées par $\text{eqm}_{\text{boot}}\left(\hat{\bar{Y}}_i^{\text{PL}}\right)$ et $\text{eqm}_{\text{boot}}\left(\hat{\bar{Y}}_i^{\text{VRH}j}\right)$, pour $j = 1$ et $j = 2$ respectivement, se calculent par la méthode bootstrap paramétrique conditionnelle à la section 4. Pour chaque échantillon prélevé sur la r^{e} population en simulation $(r = 1, \ldots, R)$, nous avons pris $B = 400$ bootstraps pour calculer la r^{e} valeur de $\text{eqm}_{\text{boot}}\left(\hat{\bar{Y}}_i\right)$ que nous désignons par $\text{eqm}_{\text{boot}}^{(r)}\left(\hat{\bar{Y}}_i\right)$. Nous avons envisagé deux mesures pour évaluer le rendement de $\text{eqm}_{\text{boot}}\left(\hat{\bar{Y}}_i\right)$, à savoir le biais relatif absolu et l'intervalle de confiance moyens. Ces mesures se définissent ainsi :

Biais relatif absolu moyen :

$$\overline{\text{BRA}} = \frac{1}{M}\sum_{i=1}^{M}\left|\frac{E\left(\text{eqm}_{\text{boot}}\left(\hat{\bar{Y}}_i\right)\right)}{\text{EQM}\left(\hat{\bar{Y}}_i\right)} - 1\right|,$$

où

$$E\left(\text{eqm}_{\text{boot}}\left(\hat{\bar{Y}}_i\right)\right) = \frac{1}{R}\sum_{r=1}^{R} \text{eqm}_{\text{boot}}^{(r)}\left(\hat{\bar{Y}}\right).$$

Niveau de confiance moyen :

$$\overline{\text{NC}} = \frac{1}{M}\sum_{i=1}^{M} \text{NC}_i,$$

où $\text{NC}_i = R^{-1}\sum_{r=1}^{R} \text{I}\left(\bar{Y}_i^{(r)} \in \text{IC}^{(r)}\right)$ et $\text{IC}^{(r)} = \left[\hat{\bar{Y}}_i^{(r)} \pm 1{,}96\sqrt{\text{eqm}_{\text{boot}}^{(r)}\left(\hat{\bar{Y}}_i\right)}\right]$.

Le tableau 5.3 présente les données de simulation du biais relatif moyen $\left(\overline{\text{BRA}}\right)$ des estimateurs de l'EQM tant pour les deux mesures de taille PS (5.2) que pour les mesures d'Asparouhov (5.3 et 5.4) et avec $\alpha = 1; 2; 3$ et ∞.

Tableau 5.3
Biais relatif moyen (%) de l'erreur quadratique moyenne $\left(\overline{\text{BRA}}\right)$ pour les mesures de taille PS et AP

Génération de $p_{j\|i}$ \ Estimateur			$\hat{\bar{Y}}_i^{\text{EBLUP}}$ sans $g\left(p_{j\|i}\right)$	$\hat{\bar{Y}}_i^{\text{VRH1}}$ $g\left(p_{j\|i}\right) = p_{j\|i}$	$\hat{\bar{Y}}_i^{\text{VRH2}}$ $g\left(p_{j\|i}\right) = \log\left(p_{j\|i}\right)$	$\hat{\bar{Y}}_i^{\text{PL}}$ $m_0\left(p_{j\|i}\right)$
PS			25,4	3,9	3,4	7,7
AP	$\alpha = 1$	I	39,9	9,7	14,4	7,5
		NI	46,6	4,1	8,7	10,0
	$\alpha = 2$	I	16,0	2,9	3,8	5,9
		NI	21,4	3,8	3,5	5,8
	$\alpha = 3$	I	13,4	6,1	6,4	5,8
		NI	15,4	7,3	7,4	8,8
	$\alpha = \infty$	I	4,6	4,2	4,5	6,2
		NI	6,1	6,4	6,3	6,9

Le $\overline{\text{BRA}}$ de $\hat{\bar{Y}}_i^{\text{EBLUP}}$ selon le modèle sans la variable d'augmentation $g\left(p_{j|i}\right)$ est très élevé pour un échantillonnage très informatif $(\alpha = 1)$; nous obtenons 39,9 % pour I et 46,6 % pour NI. Le $\overline{\text{BRA}}$ diminue progressivement pour s'établir à 5 % environ avec un échantillonnage non informatif $(\alpha = \infty)$. En général, le $\overline{\text{BRA}}$ des estimateurs paramétriques ou non est inférieur à 10 %, la seule exception étant les 14,4 % de l'estimateur $\hat{\bar{Y}}_i^{\text{VRH2}}$ assorti de la variable d'augmentation $\log\left(p_{j|i}\right)$.

Le tableau 5.4 présente les données de simulation du niveau de confiance moyen $\left(\overline{\text{NC}}\right)$ lié aux estimateurs de l'EQM pour les mesures de taille PS (5.2) et AP (5.3 et 5.4) et pour $\alpha = 1; 2; 3$ et ∞ et un niveau nominal de 0,95.

Tableau 5.4
Niveau de confiance moyen de l'erreur quadratique moyenne $\left(\overline{\text{NC}}\right)$ pour les mesures de taille PS et AP

Génération de $p_{j\|i}$ \ Estimateur			$\hat{\bar{Y}}_i^{\text{EBLUP}}$ sans $g(p_{j\|i})$	$\hat{\bar{Y}}_i^{\text{VRH1}}$ $g(p_{j\|i}) = p_{j\|i}$	$\hat{\bar{Y}}_i^{\text{VRH2}}$ $g(p_{j\|i}) = \log(p_{j\|i})$	$\hat{\bar{Y}}_i^{\text{PL}}$ $m_0(p_{j\|i})$
PS			0,898	0,937	0,941	0,936
AP	$\alpha = 1$	I	0,856	0,918	0,908	0,928
		NI	0,834	0,930	0,920	0,934
	$\alpha = 2$	I	0,916	0,937	0,936	0,932
		NI	0,907	0,936	0,933	0,936
	$\alpha = 3$	I	0,922	0,927	0,926	0,934
		NI	0,918	0,930	0,933	0,926
	$\alpha = \infty$	I	0,937	0,935	0,935	0,938
		NI	0,934	0,934	0,933	0,931

L'estimateur EBLUP $\hat{\bar{Y}}_i^{\text{EBLUP}}$ accuse le pire taux de couverture lorsque le plan de sondage est très informatif. Le taux s'améliore à mesure que diminue le contenu informatif. Le taux de couverture des autres estimateurs s'établit entre 93 % et 95 % sauf pour $\hat{\bar{Y}}_i^{\text{VRH2}}$ (assorti de $\log(p_{j|i})$) dont le taux est légèrement moindre.

5.3 Inclusion d'une variable d'augmentation

L'estimation polynomiale locale nous donne un mode automatique d'obtention d'un modèle augmenté raisonnable en fonction des probabilités de sélection $p_{j|i}$. Toutefois, comme nous ignorons si le plan d'échantillonnage est informatif ou non, ne devrions-nous pas toujours prévoir une variable d'augmentation dans le modèle ? Si le plan de sondage n'est pas informatif, il est raisonnable de choisir le modèle en (1.1). À noter que, dans ce cas, l'inclusion de la variable d'augmentation, $p_{j|i}$ ou $\log(p_{j|i})$, influe très peu sur le biais relatif absolu soit de l'estimateur soit de l'EQM estimé. La conclusion est semblable dans Verret et coll. (2015) avec leurs tailles supérieures de population et d'échantillon.

La même question se pose à propos de l'application de la procédure d'estimation polynomiale locale, mais la conclusion n'est pas aussi nette. Si le plan est très informatif, l'estimation polynomiale locale gagne pour le biais absolu et l'erreur quadratique moyenne lorsque $\alpha = 1$ ou $\alpha = 2$. Si le plan se fait moins informatif $(\alpha = 3)$, le traitement paramétrique de Verret et coll. (2015) représente un meilleur choix, mais par une très faible marge.

Dans la pratique, la valeur de α est inconnue et la décision est à prendre d'employer la variable d'augmentation dans un modèle paramétrique ou non. Nous appliquons à cette fin la procédure proposée par Verret et coll. (2015) et voyons quelque peu comment doit s'orienter le choix pour un ensemble quelconque de données. Définissons $u_{ij} = v_i + e_{ij}$ et ajustons le modèle suivant $y_{ij} = \beta_0 +$

$\beta_1 x_{ij} + u_{ij}$ aux données d'échantillon par les moindres carrés ordinaires (MCO). Les résidus sont $\tilde{u}_{ij} = y_{ij} - \tilde{\beta}_0 - \tilde{\beta}_1 x_{ij}$, où $\tilde{\beta}_0$ et $\tilde{\beta}_1$ sont les estimateurs MCO de β_0 et β_1 respectivement. La figure 5.1 présente les courbes des résidus de $\left(\hat{m}_0\left(p_{j|i}\right), \tilde{u}_{ij}\right)$, $j = 1, \ldots, N_i$; $i = 1, \ldots, M$ pour les mesures AP et pour $\alpha = 1$, 2, 3 et ∞ dans le cas de l'invariance. Quand $\alpha = 1$, la relation entre $\tilde{u}_{ij}$ et $\hat{m}_0\left(p_{j|i}\right)$ est clairement linéaire, d'où l'idée que le plan est informatif. Quand α augmente, le plan devient moins informatif. Il convient de noter que $m_0\left(p_{j|i}\right)$ est une constante lorsque $\alpha = \infty$. Les mêmes observations s'imposent dans le cas des mesures non invariantes. Avec les mesures de taille PS, le tracé ressemble à celui de la figure 5.1 quand $\alpha = 1$.

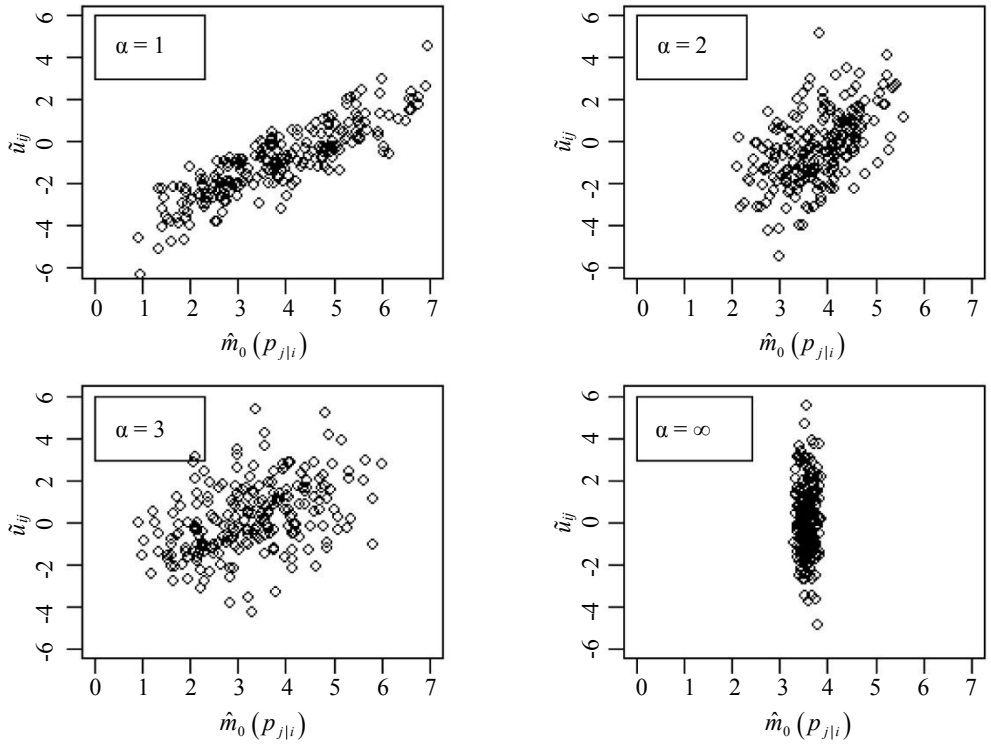

Figure 5.1 Courbes des résidus pour la population : mesures de taille AP invariantes.

Le tableau 5.5 présente les coefficients de corrélation estimés $\hat{\rho} = \text{cor}\left(\tilde{u}_{ij}, \hat{m}_0\left(p_{j|i}\right)\right)$ pour les mesures de taille PS et AP et pour $\alpha = 1$; 2; 3 et ∞.

Tableau 5.5
Coefficient de corrélation estimé $\hat{\rho} = \text{cor}\left(\tilde{u}_{ij}, \hat{m}_0\left(p_{j|i}\right)\right)$ pour les mesures de taille PS et AP

Coefficient de corrélation estimé	AP								PS
	$\alpha = 1$		$\alpha = 2$		$\alpha = 3$		$\alpha = \infty$		
	I	NI	I	NI	I	NI	I	NI	
$\hat{\rho}$	0,870	0,850	0,450	0,510	0,240	0,210	0,007	0,001	0,800

Pour ce qui est de $\overline{\text{REQM}}$, nous avons vu à la section 5.1 que $\hat{\bar{Y}}_i^{\text{EBLUP}}$ l'emporte sur les estimateurs fondés sur des modèles augmentés pour $\alpha \geq 6$. Des résultats non mentionnés au tableau 5.5 indiquent que, pour $\alpha \geq 6$, la valeur absolue du coefficient de corrélation est de moins de 0,1. En s'appuyant sur cette simulation limitée, l'utilisateur pourrait arrêter son choix de l'estimateur à utiliser pour un ensemble de données réelles : i. si $|\hat{\rho}|$ est supérieur à 0,5, il emploiera $\hat{\bar{Y}}_i^{\text{PL}}$; ii. s'il est inférieur à 0,1, il optera pour $\hat{\bar{Y}}_i^{\text{EBLUP}}$; iii. dans les autres cas, il choisira $\hat{\bar{Y}}_i^{\text{VRH1}}$ ou $\hat{\bar{Y}}_i^{\text{VRH2}}$.

6 Observations en conclusion

Nous avons étudié l'estimation d'une moyenne de petit domaine avec un échantillonnage informatif en prenant uneapproche par modèle augmenté où la variable d'augmentation est une fonction lisse $m_0\left(p_{j|i}\right)$ des probabilités de sélection $p_{j|i}$. Notre modèle augmenté est semi-paramétrique. Il diffère de ce que proposent Verret et coll. (2015), car aucune hypothèse n'est formulée au sujet de la fonction d'augmentation $m_0(\cdot)$.

Nous avons proposé une démarche en trois étapes pour estimer le modèle semi-paramétrique augmenté. D'abord, nous avons estimé un ajustement polynomial local pour chaque unité de la population (échantillonnée ou non). Ensuite, nous avons défini, compte tenu de ces valeurs d'ajustement local, une nouvelle variable dépendante pour dégager des estimateurs globaux des paramètres de régression et des effets de petit domaine. Les estimateurs ainsi obtenus ont servi à calculer les valeurs prédites de la variable dépendante y pour l'ensemble des unités non échantillonnées. En dernier lieu et à l'aide des valeurs d'échantillon observées et des valeurs prédites de y, nous avons calculé l'estimateur polynomial local $\hat{\bar{Y}}_i^{\text{PL}}$ de la moyenne de petit domaine $\bar{Y}_i$.

Nous avons adopté la méthode bootstrap paramétrique conditionnelle pour estimer l'erreur quadratique moyenne de l'estimateur nouvellement proposé. Le bootstrap conditionnel est une version modifiée du bootstrap paramétrique de Hall et Maiti (2006).

Nous avons mené une étude de simulation permettant de comparer la performance en termes de biais et d'erreur quadratique moyenne de l'estimateur EBLUP classique, $\hat{\bar{Y}}_i^{\text{EBLUP}}$, de l'estimateur EBLUP augmenté de Verret et coll. (2015), $\hat{\bar{Y}}_i^{\text{VRH}}$, et de l'estimateur polynomial local proposé, $\hat{\bar{Y}}_i^{\text{PL}}$. Comme on pouvait s'y attendre, $\hat{\bar{Y}}_i^{\text{EBLUP}}$ est entaché d'un grand biais quand l'échantillonnage est informatif. Le nouvel estimateur $\hat{\bar{Y}}_i^{\text{PL}}$ présente une EQM égale ou inférieure à celle de $\hat{\bar{Y}}_i^{\text{VRH}}$ avec un plan de sondage hautement informatif. Si le plan d'échantillonnage est moins informatif, il est préférable de recourir à l'un des deux estimateurs de Verret et coll. (2015), c'est-à-dire d'augmenter le modèle de base avec $p_{j|i}$ ou $\log\left(p_{j|i}\right)$. On notera que les gains ainsi réalisés sont des plus modestes. Si le plan d'échantillonnage est très peu informatif ou ne l'est pas du tout, on devrait employer l'estimateur $\hat{\bar{Y}}_i^{\text{EBLUP}}$ fondé sur le modèle de population.

Nous avons également évalué la performance de l'estimation bootstrap de l'erreur quadratique moyenne des estimateurs $\hat{\bar{Y}}_i^{\text{EBLUP}}$, $\hat{\bar{Y}}_i^{\text{VRH}}$ et $\hat{\bar{Y}}_i^{\text{PL}}$ pour ce qui est du biais relatif absolu moyen $\left(\overline{\text{BRA}}\right)$ et du

niveau de confiance moyen $\left(\overline{\text{NC}}\right)$. Le bootstrap conditionnel est un bon moyen d'estimation des erreurs quadratiques moyennes.

L'avantage avec l'estimateur polynomial local est qu'il nous offre un moyen automatique d'augmenter le modèle en cas de plan informatif. Son plus grand inconvénient est la charge de calcul qu'il impose tant pour l'estimation des paramètres que pour la fiabilité du traitement. La procédure décrite à la section 5.3 nous suggère une façon de déterminer si l'estimation polynomiale locale en vaut la peine ou non. Une autre approche consisterait à augmenter le modèle au niveau des unités par un terme spline-P des probabilités de sélection permettant de tenir compte du contenu informatif du plan de sondage. C'est une orientation qui a récemment été étudiée par Chatrchi (2018).

Remerciements

Nous remercions J.N.K. Rao d'avoir proposé le bootstrap conditionnel pour l'estimation de l'erreur quadratique moyenne de l'estimateur polynomial. Nous le remercions également de ses observations sur notre article. Nos remerciements vont enfin au corédacteur et à un examinateur pour leurs commentaires constructifs qui ont amélioré notre exposé.

Bibliographie

Asparouhov, T. (2006). General multi-level modelling with sampling weights. *Communication in Statistics, Theory and Methods*, 439-460.

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

Breidt, F.J., et Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28, 4, 1026-1053.

Breidt, F.J., Opsomer, J.D., Johnson, A.A. et Ranalli, M.G. (2007). Estimation assistée par un modèle semi-paramétrique pour les enquêtes sur les ressources naturelles. *Techniques d'enquête*, 33, 1, 41-51. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2007001/article/9850-fra.pdf.

Chatrchi, G. (2018). *Small Area Estimation: Informative Sampling and Two-Fold Models*. Thèse de doctorat non publiée, Carleton University, Ottawa, Canada.

Fan, J., et Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Londres: Chapman and Hall.

González-Manteiga, W., Lombardia, M.J., Molina, I., Morales, D. et Santamaria, L. (2008). Bootstrap mean squared error of a small area EBLUP. *Journal of Statistical Computation and Simulation*, 78, 5, 443-462.

Hall, P., et Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 68, 221-238.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. et Breidt, F.J. (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1), 265-286.

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

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

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

Rao, J.N.K., Sinha, S.K. et Dumitrescu, L. (2014). Robust small area estimation under semi-parametric mixed models. *The Canadian Journal of Statistics*, 42, 126-141.

Ruppert, D., et Matteson, D.E. (2015). *Statistics and Data Analysis for Financial Engineering with R Examples: 2nd Ed.*, New York: Springer.

Tillé, Y. (2006). *Sampling Algorithms:*New York: Springer.

Verret, F., Rao, J.N.K. et Hidiroglou, M.A. (2015). Estimation sur petits domaines fondée sur un modèle sous échantillonnage informatif. *Techniques d'enquête*, 41, 2, 353-368. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2015002/article/14248-fra.pdf.

Wu, H., et Zhang, J.T. (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association*, 97, 459, 883-897.