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

Techniques d’enquête

Effets d’erreurs de spécification de modèle sur les estimateurs sur petits domaines

par Yuting Chen, Partha Lahiri et Nicola Salvati

Date de diffusion : le 23 décembre 2025

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Techniques d'enquête, décembre 2025
Vol. 51, nº 2, pp. 707-719
Statistique Canada, nº 12-001-X au catalogue

Effets d'erreurs de spécification de modèle sur les estimateurs sur petits domaines

Yuting Chen, Partha Lahiri et Nicola Salvati[1]

Résumé

Des modèles de régression à erreurs emboîtées sont couramment utilisés pour intégrer des variables auxiliaires propres aux unités afin d'améliorer les estimations sur petits domaines. En cas d'erreur de spécification relative à la structure de la moyenne du modèle, l'erreur quadratique moyenne de prédiction (EQMP) fondée sur le plan des meilleurs prédicteurs linéaires sans biais empiriques (MPLSBE) augmente généralement. La méthode de meilleure prédiction observée (MPO) a été proposée dans le but d'améliorer l'EQMP fondée sur le plan par rapport à la méthode MPLSBE. Dans la présente étude, nous menons des expériences de simulation de Monte Carlo pour comprendre l'effet d'erreurs de spécification de la structure de la moyenne sur différents estimateurs sur petits domaines. Nos résultats laissent entendre que la méthode MPO reposant sur des variables auxiliaires au niveau de l'unité ne fournit pas de meilleurs résultats que la méthode MPLSBE pour ce qui est de l'EQMP fondée sur le plan, sauf si le nombre de petits domaines m est extrêmement grand. À l'inverse, les rendements de la méthode MPO s'améliorent nettement lorsque des variables auxiliaires au niveau du domaine sont utilisées. La présente étude inclut à la fois des données analytiques et numériques pour illustrer ces observations; elle fournit en outre des renseignements pratiques pour faire face aux erreurs de spécification de modèle dans le contexte d'estimations sur petits domaines (EPD).

Mots-clés : Erreur quadratique moyenne de prédiction (EQMP); meilleure prédiction observée (MPO); méthode du meilleur prédicteur linéaire sans biais empirique (MPLSBE); prédicteur sur petits domaines.

1. Introduction

Des estimations directes pondérées par les poids d'enquête (par exemple Cochran, 1977) sont couramment utilisées pour produire un vaste éventail de statistiques socioéconomiques, environnementales et relatives à la santé ainsi que d'autres statistiques officielles pour des régions nationales et de grandes régions infranationales. Lorsque les cibles sont des régions géographiques plus petites, le problème de disposer d'un échantillon limité ou de ne disposer d'aucun échantillon peut nuire à la fiabilité des estimations directes. Par conséquent, l'estimation sur petits domaines (EPD) a suscité un intérêt croissant ces dernières décennies, car elle permet de tirer profit de la force de sources de données connexes, comme des données du recensement et des données administratives et géospatiales dans le cadre de la modélisation statistique. Pour obtenir un examen complet de diverses méthodes et divers modèles d'EPD, veuillez consulter Rao et Molina (2015).

Battese, Harter et Fuller (1988) ont introduit un modèle de régression à erreurs emboîtées (REE) qui lie la variable réponse aux variables auxiliaires au niveau de l'unité :

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + v_i + e_{ij}, \quad i = 1,\ldots,m, \quad j = 1,\ldots,N_i, \tag{1.1}$$

où m est le nombre de domaines; N_i est la taille connue de la population du domaine i; les v_i sont les effets aléatoires propres au domaine; les e_{ij} sont les erreurs d'échantillonnage. Nous supposons que $\{v_i, i = 1,\ldots,m\}$ et $\{e_{ij}, i = 1,\ldots,m;\ j = 1,\ldots,N_i\}$ sont indépendants avec $v_i \sim N(0, \sigma_v^2)$ et $e_{ij} \sim N(0, \sigma_e^2)$.

1. Yuting Chen, Partha Lahiri et Nicola Salvati, University of Maryland College Park, University of Maryland College Park et Università di Pisa. Courriel : ychen215@umd.edu.

Les coefficients de régression $\boldsymbol{\beta}$ et les composantes de la variance σ_v^2 et σ_e^2 sont généralement inconnus. Dans la présente section, pour simplifier l'exposé, nous supposons qu'un échantillonnage aléatoire simple (EAS) de taille n_i est tiré pour un petit domaine i.

Dans de nombreuses applications, les variables auxiliaires au niveau de l'unité ne sont pas disponibles ou sont obsolètes. Dans de tels cas, des modèles comme le cas spécial du modèle REE suivant (Newhouse, Merfeld, Ramakrishnan, Swartz et Lahiri, 2022), souvent appelé « le modèle du contexte de l'unité », sont utilisés :

$$y_{ij} = \bar{\mathbf{X}}_i'\boldsymbol{\beta} + v_i + e_{ij}, \quad i = 1,\ldots,m, \quad j = 1,\ldots,N_i, \tag{1.2}$$

où $\bar{\mathbf{X}}_i = N_i^{-1}\sum_{j=1}^{N_i}\mathbf{x}_{ij}$ est un vecteur de moyennes de population connues des variables auxiliaires pour le domaine i. Les mêmes hypothèses relatives aux effets aléatoires v_i et aux erreurs d'échantillonnage e_{ij} s'appliquent.

Le modèle au niveau du domaine, introduit par Fay et Herriot (1979), peut être une autre possibilité de gestion du manque de variables auxiliaires fiables propres à l'unité. Un modèle au niveau du domaine correspondant au modèle au niveau de l'unité (1.1) est fourni par :

$$\bar{y}_i = \bar{\mathbf{X}}_i'\boldsymbol{\beta} + v_i + \bar{e}_i, \quad i = 1,\ldots,m, \tag{1.3}$$

où $\bar{y}_i$ est la moyenne d'échantillon non pondérée; les v_i sont les effets aléatoires propres au domaine; les $\bar{e}_i$ sont les erreurs d'échantillonnage. On suppose que v_i et $\bar{e}_i$ sont indépendants avec $v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$ et $\bar{e}_i \overset{\text{iid}}{\sim} N(0, D_i)$, où D_i est la variance échantillonnale connue de $\bar{y}_i$. En pratique, les D_i sont estimés à l'aide d'une technique de lissage comme celle fournie par Otto et Bell (1995). Dans ce cas, il est possible de proposer un estimateur lissé de D_i sous la forme s^2/n_i, où s^2 est la variance de l'échantillon regroupé à l'aide de données provenant de tous les m domaines.

Le meilleur prédicteur (MP) d'un effet mixte dans un modèle linéaire mixte est simplement la moyenne conditionnelle de l'effet mixte en fonction des données. La formule explicite du MP d'une moyenne de petits domaines peut être explicitement obtenue à la fois pour des modèles au niveau de l'unité et du domaine. Comme dans Jiang, Nguyen et Rao (2015), nous supposons la normalité pour le modèle et les erreurs d'échantillonnage. Toutefois, il convient de mentionner que la normalité n'est pas nécessaire pour obtenir le MP de moyennes de petits domaines. Le MP peut, par exemple, être obtenu en partant de l'hypothèse selon laquelle les moyennes conditionnelles des effets aléatoires en fonction des données sont une fonction linéaire des observations d'échantillon; voir, par exemple, Ghosh et Lahiri (1987). Le meilleur prédicteur linéaire sans biais (MPLSB) est obtenu à partir du meilleur prédicteur (MP) dans le cadre du modèle linéaire mixte théorique, lorsque les coefficients de régression inconnus dans le MP sont remplacés par les estimateurs par les moindres carrés pondérés. Le MPLSB estimé (MPLSBE) est obtenu lorsque les composantes de la variance inconnues du modèle sont remplacées par des estimateurs standard (par exemple le maximum de vraisemblance restreint [REML]). Le MPLSB et le MPLSBE peuvent être considérés comme étant un meilleur prédicteur estimé (MPE) des effets mixtes.

Jiang et coll. (2011) ont proposé la méthode de la meilleure prédiction observée (méthode MPO) pour des moyennes de petits domaines dans le cadre d'un modèle linéaire mixte en tentant de réduire les effets d'erreurs de spécification de la structure de la moyenne dans le modèle théorique. Nous notons que leur méthode MPO peut être motivée comme étant un MPE lorsque les meilleurs estimateurs prédictifs (MEP) des paramètres de modèle, qui minimisent l'erreur quadratique moyenne de prédiction observée, sont utilisés à la place d'estimateurs de paramètres de modèle standard. Dans le contexte de modèles au niveau du domaine, Jiang, Nguyen et Rao (2011) ont utilisé à la fois des études théoriques et empiriques pour démontrer que la méthode MPO produit de meilleurs résultats que le MPLSBE en ce qui concerne l'EQMP fondée sur le plan lorsque le modèle linéaire mixte sous-jacent fait l'objet d'erreurs de spécification. Par la suite, Jiang et coll. (2015) ont considéré la méthode MPO pour un modèle de régression à erreurs emboîtées, dans le cadre duquel la fonction de moyenne et les composantes de la variance faisaient l'objet d'erreurs de spécification. Leurs simulations ont indiqué que la méthode MPO pouvait fournir des résultats nettement supérieurs à ceux de la méthode MPLSBE en matière d'EQMP fondée sur le plan globale et propre au domaine. Toutefois, nos études par simulations indiquent qu'utiliser la méthode MPO avec des variables auxiliaires au niveau de l'unité (MPO-UNIT) ne fournit pas de meilleurs résultats que la méthode MPLSBE, sauf si le nombre de domaines m est extrêmement grand, ce qui est inhabituel dans une estimation sur petits domaines. Nous avons constaté que l'utilisation de variables auxiliaires au niveau du domaine, comme dans le modèle du contexte de l'unité ou le modèle au niveau du domaine, peut améliorer les rendements de la méthode MPO.

Dans le présent document, nous étudions les effets d'erreurs de spécification en ce qui concerne la fonction de moyenne et les composantes de la variance sur les rendements prédictifs d'estimateurs sur petits domaines existants. Nous présentons également des données analytiques et numériques expliquant les raisons pour lesquelles la méthode MPO-UNIT fournit des résultats inférieurs en matière d'EQMP fondée sur le plan et dans quelle mesure l'utilisation de variables auxiliaires au niveau du domaine peut améliorer son efficacité. Le reste du présent document est structuré de la manière suivante. Dans la section 2, nous fournissons de plus amples précisions sur la méthodologie MPO et les raisons pour lesquelles la méthode MPO avec variables auxiliaires au niveau du domaine produit de meilleurs résultats que celle avec variables auxiliaires au niveau de l'unité. Dans la section 3, nous présentons des études numériques et les résultats de l'évaluation en ce qui concerne l'EQMP fondée sur le plan globale et propre au domaine. Dans la section 4, nous concluons et fournissons des lignes directrices pratiques pour mettre en œuvre la méthode MPO en présence d'erreurs de spécification de modèle.

2. Comparaison analytique entre la méthode de la meilleure prédiction observée avec variables auxiliaires au niveau de l'unité et celle avec variables auxiliaires au niveau du domaine

Soit $\boldsymbol{\psi} = (\boldsymbol{\beta}', \sigma_v^2, \sigma_e^2)$, le vecteur de paramètres de modèle du modèle de régression à erreurs emboîtées, et $\tilde{\boldsymbol{\theta}}(\boldsymbol{\psi}) = [\tilde{\theta}_i(\boldsymbol{\psi})]_{1 \le i \le m}$, le vecteur des MP pour les moyennes réelles de petits domaines $\boldsymbol{\theta} = [\theta_i]_{1 \le i \le m}$. L'EQMP fondée sur le plan de $\tilde{\boldsymbol{\theta}}(\boldsymbol{\psi})$ est donnée par

$$\mathrm{EQMP}(\tilde{\boldsymbol{\theta}}(\boldsymbol{\psi})) = \mathrm{E}(|\tilde{\boldsymbol{\theta}}(\boldsymbol{\psi}) - \boldsymbol{\theta}|^2) = \sum_{i=1}^{m} \mathrm{E}(\tilde{\theta}_i(\boldsymbol{\psi}) - \theta_i)^2, \tag{2.1}$$

où l'espérance se rapporte au plan d'échantillonnage. Selon Jiang et coll. (2011) et Jiang et coll. (2015), l'EQMP dans (2.1) peut être exprimée autrement, ce qui est une idée clé de la méthode MPO. Cela mène ainsi à l'équation fondamentale de la MPO :

$$\mathrm{EQMP}(\tilde{\boldsymbol{\theta}}(\boldsymbol{\psi})) = \mathrm{E}\{Q(\boldsymbol{\psi})\}, \tag{2.2}$$

et $Q(\cdot)$ désigne la fonction d'EQMP observée. Selon le modèle de régression à erreurs emboîtées, la fonction Q peut être exprimée comme suit :

$$\begin{aligned} Q = {} & \boldsymbol{\beta}'(\bar{\mathbf{X}} - \mathbf{G}\bar{\mathbf{x}})'(\bar{\mathbf{X}} - \mathbf{G}\bar{\mathbf{x}})\,\boldsymbol{\beta} - 2\bar{\mathbf{y}}'\{(\mathbf{I}_m - 2\mathbf{G})\,\bar{\mathbf{X}} + \mathbf{G}^2\bar{\mathbf{x}}\}\,\boldsymbol{\beta} \\ & + \bar{\mathbf{y}}'\mathbf{G}^2\bar{\mathbf{y}} + \mathbf{1}_m'(\mathbf{I}_m - 2\mathbf{G})\,\hat{\boldsymbol{\mu}}^2, \end{aligned} \tag{2.3}$$

où $\bar{\mathbf{X}} = (\bar{\mathbf{X}}_i')_{1\le i\le m}$, $\bar{\mathbf{x}} = (\bar{\mathbf{x}}_i')_{1\le i\le m}$, $\bar{\mathbf{x}}_i$ est le vecteur des moyennes d'échantillon non pondérées des variables auxiliaires pour le domaine i, $\bar{\mathbf{y}} = (\bar{y}_i)_{1\le i\le m}$, $\mathbf{G} = \mathrm{diag}\{n_i/N_i + (1-n_i/N_i)\,n_i\sigma_v^2/(n_i\sigma_v^2+\sigma_e^2), 1\le i\le m\}$ et $\hat{\boldsymbol{\mu}}^2 = (\hat{\mu}_i^2)_{1\le i\le m}$ est l'estimateur sans biais de $(\bar{Y}_i^2)_{1\le i\le m}$ par rapport au plan. Le MEP de $\boldsymbol{\psi}$, désigné par $\hat{\boldsymbol{\psi}}$, est le minimiseur de $Q(\boldsymbol{\psi})$ par rapport à $\boldsymbol{\psi}$.

Pour comprendre les raisons pour lesquelles la méthode MPO donne des résultats optimaux lors de l'utilisation de variables auxiliaires au niveau du domaine, il est instructif de prendre en compte le cas où tous les paramètres de modèle $\boldsymbol{\psi}$ au niveau de l'unité sont connus. Puisque $\mathrm{E}(\bar{\mathbf{x}} - \bar{\mathbf{X}}) = 0$ et $\mathrm{E}\{\bar{\mathbf{y}}'\mathbf{G}^2(\bar{\mathbf{x}} - \bar{\mathbf{X}})\,\boldsymbol{\beta}\} = \mathrm{cov}(\bar{\mathbf{y}}'\mathbf{G}^2, \boldsymbol{\beta}'\bar{\mathbf{x}}')$, nous obtenons

$$\mathrm{E}Q = \mathrm{E}\{\boldsymbol{\beta}'(\bar{\mathbf{x}} - \bar{\mathbf{X}})'\mathbf{G}^2(\bar{\mathbf{x}} - \bar{\mathbf{X}})\,\boldsymbol{\beta}\} + \cdots \tag{2.4}$$

où $\cdots$ sont les termes faisant intervenir des paramètres de population finie et les paramètres de modèle connus. Puisque $\mathbf{G}^2$ est défini positif, l'expression est minimisée pour $\bar{\mathbf{x}} - \bar{\mathbf{X}} = 0$, pour toute valeur $\boldsymbol{\beta}$ fixe, ce qui indique que l'utilisation de variables auxiliaires au niveau du domaine fournit des rendements de MPO optimaux.

3. Simulations

Les calculs plus détaillés de la procédure de la MPO à la fois pour les modèles au niveau du domaine et de l'unité figurent dans la documentation supplémentaire de Jiang et coll. (2011). De plus, puisque le modèle du contexte de l'unité est un type particulier de modèle au niveau de l'unité, nous pouvons calculer directement la MPO dans un modèle du contexte de l'unité en remplaçant $\mathbf{x}_{ij}$ par $\bar{\mathbf{X}}_i$.

Les paramètres de simulation sont semblables à ceux de Jiang et coll. (2015). Pour simplifier, nous considérons un cas de variable auxiliaire unique que l'on suppose associée linéairement à la variable réponse y_{ij} selon le modèle suivant :

$$y_{ij} = \beta_1 x_{ij} + v_i + e_{ij} \quad i = 1,\ldots,m,\ j = 1,\ldots,N_i, \tag{3.1}$$

où les x_{ij} sont des valeurs connues d'une variable auxiliaire pour la j^{e} unité du i^{e} domaine; β_1 est un coefficient de régression inconnu; v_i et e_{ij} sont les mêmes que dans (1.1). Nous supposons que les x_{ij} ne sont pas tous identiques dans un domaine. Dans le contexte envisagé, (1.2) devient :

$$y_{ij} = \beta_1 \bar{X}_i + v_i + e_{ij} \quad i = 1,\ldots,m,\ j = 1,\ldots,N_i. \tag{3.2}$$

Le modèle correspondant au niveau du domaine est fourni par :

$$\bar{y}_i = \beta_1 \bar{X}_i + v_i + \bar{e}_i \quad i = 1,\ldots,m. \tag{3.3}$$

Pour la configuration des simulations, nous tirons des échantillons par échantillonnage aléatoire simple sans remplacement dans la population de chaque petit domaine et considérons les estimateurs suivants des moyennes de petits domaines dans le reste de la présente section :

A) Estimateur direct (moyenne de l'échantillon);

B) MPO selon le modèle théorique du contexte de l'unité (3.2) (MPO-CU);

C) MPO selon le modèle de base de Fay-Herriot (au niveau du domaine) (3.3) avec estimations de la variance directes lissées : $\hat{D}_i = s^2/n_i$, $s^2 = (n-1)^{-1}\sum_{i=1}^{m}\sum_{i=1}^{n_i}(y_{ij} - \bar{y})^2$ et $n = \sum_{i=1}^{m} n_i$;

D) MPO selon le modèle théorique au niveau de l'unité (3.1) (MPO-UNIT);

E) MPLSBE selon le modèle théorique au niveau de l'unité (3.1) (MPLSBE-UNIT);

F) MPLSBE selon le modèle théorique du contexte de l'unité (3.2) (MPLSBE-CU).

Pour introduire des erreurs de spécification de modèle, nous générons d'abord y pour la population finie à partir du modèle suivant de régression à erreurs emboîtées hétéroscédastiques de superpopulation :

$$y_{ij} = b + v_i + e_{ij} \quad i = 1,\ldots,m,\ j = 1,\ldots,N_i. \tag{3.4}$$

Les tailles de population et d'échantillon sont les mêmes pour tous les domaines et sont respectivement fixées à N_i = 1 000 et n_i = 4. Nous considérons ensuite deux valeurs différentes de b : b = 5,10 et trois valeurs du nombre de petits domaines : m = 40, 100 ou 400, v_i étant généré à partir de la distribution normale $N(0,1)$, et e_{ij} généré à partir de la distribution normale $N(0, \sigma_{ei}^2)$, où σ_{ei}^2 est généré indépendamment de la distribution gamma $\Gamma(3, 0{,}5)$. Dans chaque cas, x pour la population finie est généré à partir d'une distribution de superpopulation log-normale avec une moyenne de 1 et un écart-type de 0,5.

Chaque scénario est simulé indépendamment K = 1 000 fois. Les rendements des estimateurs (A)-(F), selon les paramètres de simulation susmentionnés, sont évalués à la fois en fonction de l'EQMP globale et de celle fondée sur le plan propre au domaine. L'EQMP fondée sur le plan propre au domaine est définie sous la forme $\text{EQMP}(\hat{\bar{Y}}_i) = \text{E}(\hat{\bar{Y}}_i - \bar{Y}_i)^2$, où $\bar{Y}_i = N_i^{-1}\sum_{i=1}^{N_i} y_{ij}$ est la moyenne de petits domaines réelle, et $\hat{\bar{Y}}_i$ est la valeur prédite pour le i^{e} domaine, soit par la méthode MPO ou la méthode MPLSBE, et E se rapporte

au plan de sondage. Dans des simulations de Monte Carlo, des estimations de l'EQMP pour le domaine i et de l'EQMP générale sont respectivement fournies par :

$$\text{EQMP}_i \approx \frac{1}{K}\sum_{k=1}^{K}(\hat{\bar{Y}}_i^{(k)} - \bar{Y}_i^{(k)})^2, \tag{3.5}$$

$$\text{EQMP} \approx \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{K}\sum_{k=1}^{K}(\hat{\bar{Y}}_i^{(k)} - \bar{Y}_i^{(k)})^2\right], \tag{3.6}$$

où $\bar{Y}_i^{(k)}$ et $\hat{\bar{Y}}_i^{(k)}$ sont respectivement la moyenne réelle et la moyenne estimée pour le domaine i dans la k^{e} simulation.

Le tableau 3.1 fait état des résultats pour l'EQMP simulée générale fondée sur le plan pour les diverses conditions de simulation et les divers estimateurs lorsque la population finie est générée en fonction du modèle réel sous-jacent (3.4). L'erreur de spécification du modèle n'a pas d'incidence sur l'estimateur direct puisque celui-ci n'est pas fondé sur le modèle. Le comportement de la méthode MPLSBE-UNIT est semblable à celui de l'estimateur direct, car on attribue automatiquement davantage de poids à la moyenne d'échantillon lorsque le modèle n'est pas robuste. Alors que la méthode MPO-UNIT visait à gérer les erreurs de spécification de modèle, ses résultats sont étonnamment médiocres, voire inférieurs à ceux de la métode de l'estimateur direct et de la méthode MPLSBE-UNIT lorsque $b = 10$. Les méthodes MPO-CU et MPLSBE-CU fournissent des résultats semblables et meilleurs que ceux de la méthode MPLSBE-UNIT. La raison pour laquelle la méthode MPLSBE-CU fournit également de bons résultats pourrait être que, du fait de la grande taille de population N_i et du fait que les valeurs de population proviennent d'une distribution commune à tous les domaines, les moyennes de population $\bar{X}_i$ sont à peu près égales pour tous les domaines. Ainsi, le modèle du contexte de l'unité et le modèle au niveau du domaine fournissent des approximations proches du modèle réel. Par conséquent, les prédicteurs reposant sur des variables auxiliaires au niveau du domaine tendent à fournir de meilleurs résultats que ceux reposant sur des variables auxiliaires au niveau de l'unité. Globalement, nos résultats laissent entendre que dans le cas d'erreurs de spécification du modèle importantes, la méthode MPO-UNIT peut ne pas être un choix efficace.

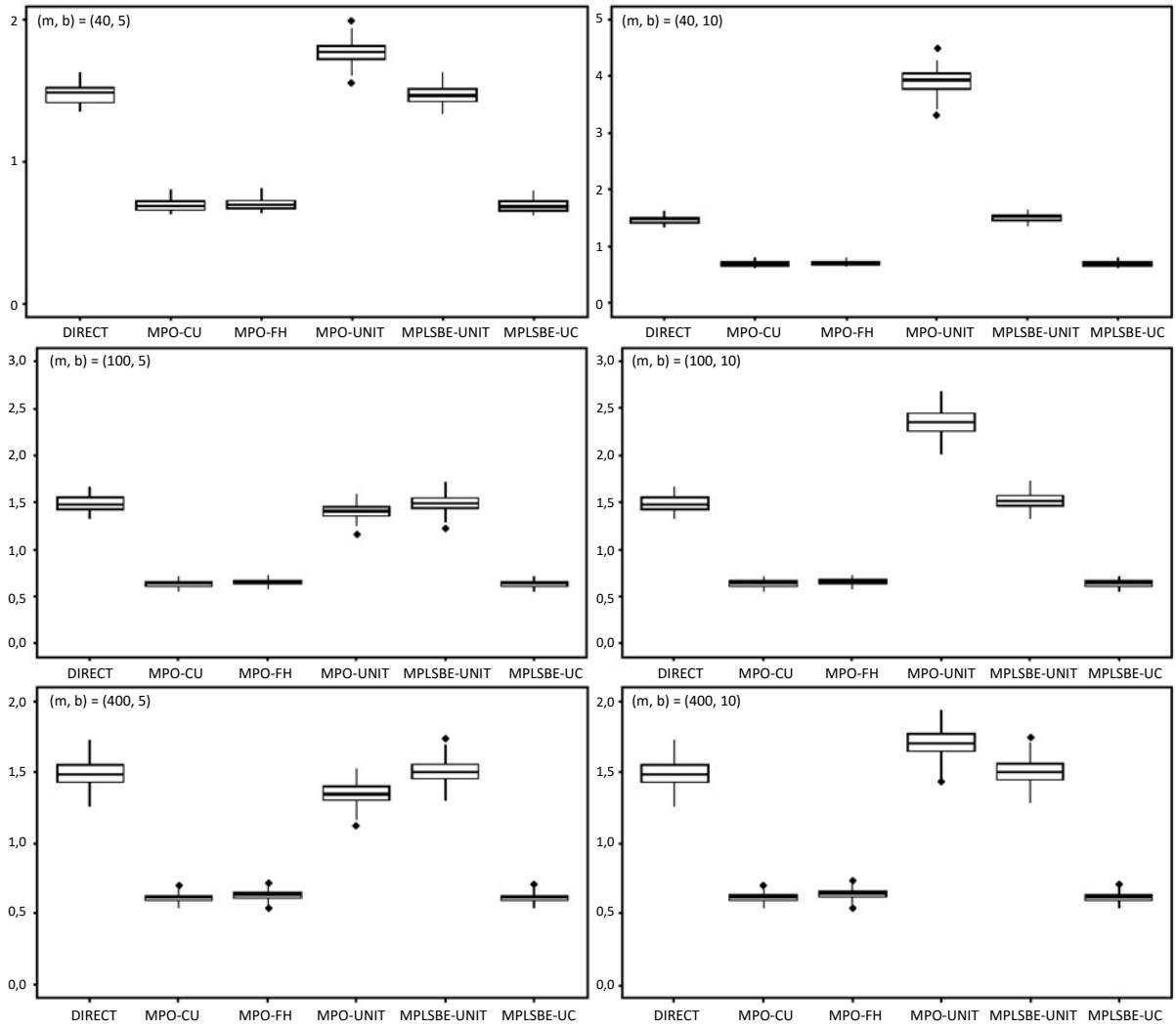
Pour les EQMP propres au domaine, nous utilisons des diagrammes de quartiles pour présenter les distributions des EQMP fondées sur le plan propres au domaine associées à tous les estimateurs. Voir la figure 3.1. Le diagramme de quartiles de la méthode MPO-UNIT présente une EQMP médiane et une variabilité bien plus élevées que celles des méthodes MPO-CU et MPO-FH et dans certains cas moins bonnes que celles de la méthode de l'estimateur direct et de la méthode MPLSBE-UNIT. La méthode MPO-CU donne une variabilité légèrement supérieure à la méthode MPO-FH. C'est peut-être parce que, contrairement aux modèles au niveau du domaine, les modèles de contexte de l'unité intègrent l'incertitude découlant de l'estimation des paramètres de modèle, comme les variances échantillonnales.

Tableau 3.1
Erreur quadratique moyenne de prédiction simulée globale fondée sur le plan lorsque la population finie est générée à partir du modèle de régression à erreurs emboîtées hétéroscédastiques (3.4)

(*m*,*b*)	DIRECT	MPO-CU	MPO-FH	MPO-UNIT	MPLSBE-UNIT	MPLSBE-CU
(40, 5)	1,475	0,689	0,698	1,775	1,474	0,687
(100, 5)	1,488	0,638	0,657	1,409	1,495	0,638
(400, 5)	1,493	0,614	0,637	1,352	1,508	0,613
(40, 10)	1,475	0,696	0,705	3,919	1,522	0,694
(100, 10)	1,488	0,646	0,664	2,364	1,513	0,645
(400, 10)	1,493	0,621	0,644	1,711	1,510	0,621

Note : MPO-CU = meilleure prédiction observée selon le modèle théorique du contexte de l'unité; MPO-FH = meilleure prédiction observée selon le modèle de base de Fay-Herriot (au niveau du domaine); MPO-UNIT = meilleure prédiction observée selon le modèle théorique au niveau de l'unité; MPLSBE-UNIT = meilleur prédicteur linéaire sans biais empiriques selon le modèle théorique au niveau de l'unité; MPLSBE-CU = meilleur prédicteur linéaire sans biais empirique selon le modèle théorique du contexte de l'unité.

Figure 3.1 Distributions des erreurs quadratiques moyennes de prédiction simulée propres au domaine fondées sur le plan lorsque la population finie est générée à partir du modèle de régression à erreurs emboîtées hétéroscédastiques (3.4)



Note : MPO-CU = meilleure prédiction observée selon le modèle théorique du contexte de l'unité; MPO-FH = meilleure prédiction observée selon le modèle de base de Fay-Herriot (au niveau du domaine); MPO-UNIT = meilleure prédiction observée selon le modèle théorique au niveau de l'unité; MPLSBE-UNIT = meilleur prédicteur linéaire sans biais empiriques selon le modèle théorique au niveau de l'unité; MPLSBE-CU = meilleur prédicteur linéaire sans biais empirique selon le modèle théorique du contexte de l'unité.

Comme il en a été question dans Jiang et coll. (2015), les conditions de simulation susmentionnées peuvent être un peu extrêmes, puisque les modèles théoriques sont complètement différents du modèle sous-jacent réel. Cela nous motive à prendre en compte des cas modérés selon lesquels le modèle théorique est partiellement correct par rapport au modèle réel. En conservant les modèles théoriques (3.1)-(3.3), nous générons la population finie pour y à partir du modèle de superpopulation sous-jacent :

$$y_{ij} = b_0 + b_1 x_{ij} + v_i + e_{ij} \quad i = 1,\ldots,m,\ j = 1,\ldots,N_i \tag{3.7}$$

où $b_0 = 10$, $b_1 = 5$. Il convient de mentionner que la pente de (3.7) n'est pas nulle et qu'elle correspond à la partie de la relation linéaire du modèle théorique (3.1). Nous avons toutefois de légères erreurs de spécification, puisque le modèle réel a ici également une ordonnée à l'origine non nulle. De plus, e_{ij} est généré à partir de la même distribution normale que celle décrite précédemment, nous avons donc également le problème d'hétéroscédasticité. Nous considérons trois valeurs différentes de m : $m =$ 40, 100 ou 400, et v_i est généré à partir d'une distribution normale $N(0,1)$.

Les résultats fondés sur $K =$ 1 000 simulations sont présentés dans le tableau 3,2. Comme on s'y attendait, dans ce cas, la méthode MPLSBE-UNIT fournit de meilleurs résultats que l'estimateur direct puisque le modèle théorique est plus proche du modèle réel et que la méthode MPLSBE peut être renforcée par d'autres domaines connexes. Alors que l'on s'attendait à ce que la méthode MPO-UNIT fournisse des résultats comparables à ceux de la méthode MPLSBE-UNIT dans ce scénario, nos résultats de simulation avec la méthode MPO-UNIT sont en fait moins bons que ceux de la méthode MPLSBE-UNIT pour ce qui est de l'EQMP simulée. Lorsque $m =$ 40, l'EQMP simulée pour MPO-UNIT est semblable à celle de la méthode d'estimateur DIRECT, ce qui indique un piètre rendement. En revanche, les méthodes MPO-CU et MPLSBE-CU présentent des résultats semblables, légèrement supérieurs à ceux de la méthode MPLSBE-UNIT lorsque $m =$ 100 et $m =$ 400. La méthode MPO-FH présente un rendement stable et de meilleurs résultats que les méthodes MPO-CU et MPLSBE-CU lorsque $m =$ 40.

Les résultats présentés dans les tableaux 3.1 et 3.2 indiquent que la méthode MPO-UNIT peut ne pas réduire efficacement l'incidence d'erreurs de spécification du modèle, par rapport à la méthode MPLSBE-UNIT. Ce résultat étonnant a donné lieu à un examen approfondi, ce qui fait l'objet des remarques 1 et 2 suivantes.

Tableau 3.2
Erreur quadratique moyenne de prédiction simulée globale fondée sur le plan lorsque la population finie est générée à partir du modèle de régression à erreurs emboîtées hétéroscédastiques (3.7)

m	DIRECT	MPO-CU	MPO-FH	MPO-UNIT	MPLSBE-UNIT	MPLSBE-CU
40	18,243	1,798	1,588	18,230	1,532	1,872
100	18,339	1,341	1,248	7,958	1,514	1,358
400	18,422	1,085	1,059	3,122	1,509	1,087

Note : MPO-CU = meilleure prédiction observée selon le modèle théorique du contexte de l'unité; MPO-FH = meilleure prédiction observée selon le modèle de base de Fay-Herriot (au niveau du domaine); MPO-UNIT = meilleure prédiction observée selon le modèle théorique au niveau de l'unité; MPLSBE-UNIT = meilleur prédicteur linéaire sans biais empiriques selon le modèle théorique au niveau de l'unité; MPLSBE-CU = meilleur prédicteur linéaire sans biais empirique selon le modèle théorique du contexte de l'unité.

Remarque 1 : Erreur quadratique moyenne de prédiction de la méthode MPO-UNIT ou $|\bar{\mathbf{X}} - \bar{\mathbf{x}}|$

Pour étudier plus en profondeur la relation entre les rendements de la méthode MPO-UNIT et la différence absolue entre les moyennes d'échantillon et de population, nous avons mené une étude numérique à l'aide du modèle de superpopulation par régression à erreurs emboîtées (REE) suivant :

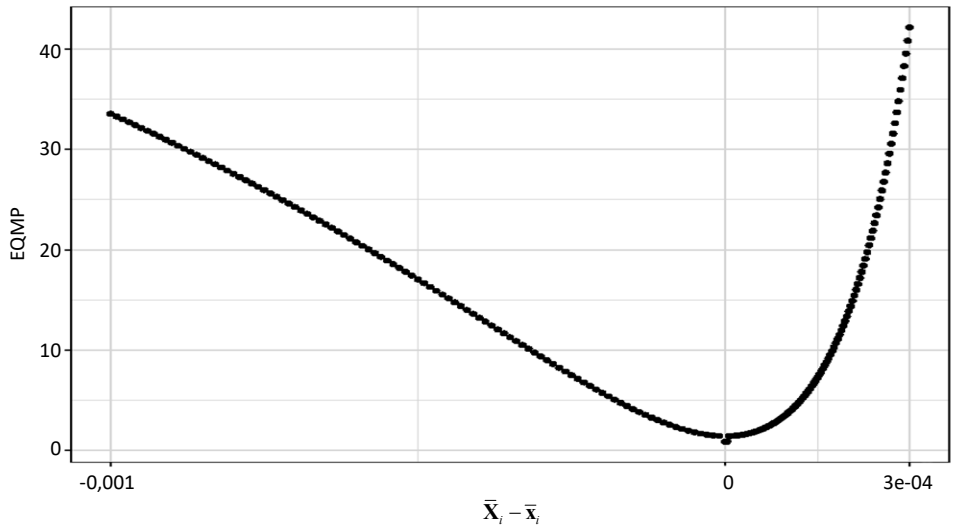
$$y_{ij} = 10 + v_i + e_{ij} \quad i = 1,\ldots,50,\ j = 1,\ldots,1\,000. \tag{3.8}$$

Les processus de génération de la population finie pour x, v_i et e_{ij} demeurent semblables à ceux du modèle susmentionné. Nous avons ensuite tiré des échantillons EAS de taille $n_i = 4$ pour chaque petit domaine.

Dans un cadre de simulation, lorsque le nombre d'unités au sein d'un domaine particulier est relativement petit, il est souvent impossible d'obtenir une moyenne d'échantillon $\bar{\mathbf{x}}$ égale, voire proche de la moyenne de population $\bar{\mathbf{X}}$. Pour étudier la façon dont des calculs de moyennes d'échantillon $\bar{\mathbf{x}}$ à partir de moyennes de population $\bar{\mathbf{X}}$ influent sur les résultats de la méthode MPO, nous avons introduit un terme de biais dans les moyennes d'échantillon. Plus précisément, pour chaque domaine, les moyennes d'échantillon ont été corrigées en définissant: $\bar{\mathbf{x}} = \bar{\mathbf{X}} + \text{biais}$, où le terme de biais variait pour prendre des valeurs négatives, positives et nulles.

L'EQMP fondée sur le plan pour la méthode MPO reposant sur des variables auxiliaires au niveau de l'unité (EQMP_i) a été calculée pour chaque domaine; ces valeurs ont été comparées avec la grandeur de la différence absolue $|\bar{\mathbf{X}} - \bar{\mathbf{x}}|$. Les résultats ont été visualisés dans la figure 3.2, qui montre que l'EQMP_i simulée pour la méthode MPO-UNIT augmente avec la différence entre $\bar{X}_i$ et $\bar{x}_i$ pour le domaine $i = 1$. Plus particulièrement, lorsque $\bar{X}_i - \bar{x}_i = 0$ (ce qui correspond au modèle du contexte de l'unité), l'EQMP_i simulée est inférieure à celle des autres scénarios de différence non nulle. Ces résultats reflètent notre espérance théorique de Q fondée sur le plan de la section 2.

Figure 3.2 Relation entre $|\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i|$ et l'EQMP_i pour le domaine $i = 1$



Note : EQMP = Erreur quadratique moyenne de prédiction.

Remarque 2 : Minimisation de $Q(\psi)$ ou de l'EQMP$(\tilde{\theta}(\psi))$

Une autre explication possible des mauvais résultats de la méthode MPO-UNIT peut être que l'équation de base de la méthode MPO, valide pour des ψ fixes, peut ne pas correspondre à des ψ aléatoires, comme c'est le cas pour le meilleur estimateur prédictif (MEP) ou l'estimation du maximum de vraisemblance (EMV). Pour étudier cela, nous commençons par l'équation de la MPO de base :

$$\mathrm{E}(|\tilde{\boldsymbol{\theta}}(\boldsymbol{\psi})-\boldsymbol{\theta}|^2)=\mathrm{E}\{Q(\boldsymbol{\psi})\}. \tag{3.9}$$

Soit $\hat{\boldsymbol{\psi}}$, le MEP de ψ obtenu en minimisant l'EQMP observée; et soit $\tilde{\boldsymbol{\psi}}$, désignant l'EMV. Les méthodes MPO-UNIT et MPLSBE-UNIT correspondantes sont respectivement $\tilde{\boldsymbol{\theta}}(\hat{\boldsymbol{\psi}})$ et $\tilde{\boldsymbol{\theta}}(\tilde{\boldsymbol{\psi}})$. Si l'équation (3.9) s'applique à des ψ aléatoires, nous obtenons

$$\mathrm{E}(|\tilde{\boldsymbol{\theta}}(\hat{\boldsymbol{\psi}})-\boldsymbol{\theta}|^2)=\mathrm{E}\{Q(\hat{\boldsymbol{\psi}})\}\le\mathrm{E}\{Q(\tilde{\boldsymbol{\psi}})\}=\mathrm{E}(|\tilde{\boldsymbol{\theta}}(\tilde{\boldsymbol{\psi}})-\boldsymbol{\theta}|^2). \tag{3.10}$$

L'inégalité découle de la définition du MEP. Pour vérifier cela, nous comparons les quatre termes de l'inégalité en utilisant les mêmes paramètres de simulation que dans les analyses précédentes. Les résultats sont présentés dans les tableaux 3.3 et 3.4.

Lorsque le modèle théorique est complètement différent du modèle réel, le tableau 3.3 indique que $\mathrm{E}(|\tilde{\boldsymbol{\theta}}(\hat{\boldsymbol{\psi}})-\boldsymbol{\theta}|^2)$, le premier terme de (3.10), serait bien supérieur à $\mathrm{E}\{Q(\hat{\boldsymbol{\psi}})\}$, le deuxième terme de (3.10), sauf si m était extrêmement grand. Cependant $\mathrm{E}\{Q(\tilde{\boldsymbol{\psi}})\}$, le troisième terme de (3.10), est à peu près égal à $\mathrm{E}(|\tilde{\boldsymbol{\theta}}(\tilde{\boldsymbol{\psi}})-\boldsymbol{\theta}|^2)$, le troisième terme de (3.10), même pour une valeur m inférieure. Lorsque le modèle théorique est partiellement correct, le tableau 3.4 montre que ni $\mathrm{E}(|\tilde{\boldsymbol{\theta}}(\hat{\boldsymbol{\psi}})-\boldsymbol{\theta}|^2)=\mathrm{E}\{Q(\hat{\boldsymbol{\psi}})\}$ ni $\mathrm{E}\{Q(\tilde{\boldsymbol{\psi}})\}=\mathrm{E}(|\tilde{\boldsymbol{\theta}}(\tilde{\boldsymbol{\psi}})-\boldsymbol{\theta}|^2)$ de (3.10) n'est possible, $\mathrm{E}(|\tilde{\boldsymbol{\theta}}(\hat{\boldsymbol{\psi}})-\boldsymbol{\theta}|^2)$ étant bien supérieur à $\mathrm{E}Q(\hat{\boldsymbol{\psi}})$, en particulier pour des valeurs inférieures de m (par exemple 40). Ainsi, notre simulation étaye l'argument selon lequel l'EQMP de la méthode MPO-UNIT pourrait être supérieure à celle de la méthode MPLSBE-UNIT, sauf si m était extrêmement et déraisonnablement grand.

Un facteur contribuant potentiellement à ces différences entre $\mathrm{E}(|\tilde{\boldsymbol{\theta}}(\hat{\boldsymbol{\psi}})-\boldsymbol{\theta}|^2)$ et $\mathrm{E}\{Q(\hat{\boldsymbol{\psi}})\}$ est la stabilité numérique de $\hat{\boldsymbol{\psi}}$, MEP de ψ. En particulier, lorsque $m=$ 40, près de 20 % des estimations de MEP fournissent une valeur nulle pour le ratio σ_v^2/σ_e^2. Les résultats de la méthode MPO-UNIT semblent dépendre de la valeur de m. Les résultats de la méthode MPO-UNIT s'améliorent pour des valeurs de m extrêmement grandes (par exemple 4 000 ou 40 000), lorsque les deux termes de l'inégalité de (3.10), c'est-à-dire $\mathrm{E}(|\tilde{\boldsymbol{\theta}}(\hat{\boldsymbol{\psi}})-\boldsymbol{\theta}|^2)=\mathrm{E}\{Q(\hat{\boldsymbol{\psi}})\}$ et $\mathrm{E}\{Q(\tilde{\boldsymbol{\psi}})\}=\mathrm{E}(|\tilde{\boldsymbol{\theta}}(\tilde{\boldsymbol{\psi}})-\boldsymbol{\theta}|^2)$, tendent à être à peu près satisfaites; par conséquent, la méthode MPO-UNIT présente finalement des résultats marginalement meilleurs que la méthode MPLSBE-UNIT dans le cas d'erreurs de spécification du modèle. Nous formulons l'hypothèse que les MEP des composantes de variance se stabilisent lorsque m augmente, ce qui indique que de futurs travaux de recherche devraient porter sur l'amélioration des MEP des composantes de la variance avant d'appliquer la méthode MPO dans le cadre de modèles au niveau de l'unité. De plus, il convient de mentionner que le calcul d'estimations du maximum de vraisemblance (MV) ou du maximum de vraisemblance restreint (REML) se fait communément avec des paquets disponibles, comme `lme4` ou `sae`, qui sont optimisés pour

assurer une stabilité numérique. En revanche, le calcul des MEP associés à la méthode MPO est effectué à l'aide du code élaboré par les auteurs. Par conséquent, cette mise en œuvre peut ne pas encore atteindre le même niveau d'optimisation et de stabilité des calculs que les paquets établis. Nous fournissons un lien accessible vers le code et invitons les lecteurs à l'examiner.

Tableau 3.3
Comparaison des quatre termes de (3.10), selon les paramètres du tableau 3.1

(m,b)	EQMP $(\tilde{\theta}(\hat{\psi}))$	E$\{Q(\hat{\psi})\}$	E$\{Q(\tilde{\psi})\}$	EQMP $(\tilde{\theta}(\tilde{\psi}))$
(40, 5)	1,775	0,448	1,479	1,474
(100, 5)	1,409	0,918	1,493	1,495
(400, 5)	1,352	1,211	1,500	1,508
(1 000, 5)	1,384	1,323	1,504	1,513
(2 000, 5)	1,414	1,379	1,521	1,522
(4 000, 5)	1,414	1,398	1,512	1,515
(40, 10)	3,919	-1,189	1,649	1,522
(100, 10)	2,364	0,493	1,551	1,513
(400, 10)	1,711	1,231	1,508	1,510
(1 000, 10)	1,600	1,389	1,530	1,532
(2 000, 10)	1,509	1,412	1,500	1,496
(4 000, 10)	1,485	1,431	1,493	1,493

Note : EQMP = Erreur quadratique moyenne de prédiction.

Tableau 3.4
Comparaison des quatre termes de (3.10), selon les paramètres du tableau 3.2

m	EQMP $(\tilde{\theta}(\hat{\psi}))$	E$\{Q(\hat{\psi})\}$	E$\{Q(\tilde{\psi})\}$	EQMP $(\tilde{\theta}(\tilde{\psi}))$
40	18,230	-14,036	3,247	1,532
100	7,958	-4,809	1,586	1,514
400	3,122	-0,584	1,583	1,508
1 000	2,173	0,828	1,580	1,569
2 000	1,815	1,131	1,546	1,501
4 000	1,650	1,267	1,496	1,505
40 000	1,488	1,408	1,462	1,503

Note : EQMP = Erreur quadratique moyenne de prédiction.

4. Conclusion

Dans la présente étude, nous avons étudié les effets d'erreurs de spécification dans la structure de la moyenne et la variance d'échantillon dans le cadre du modèle de régression à erreurs emboîtées bien connu lors de l'utilisation des méthodes MPLSBE et MOP utilisant i) uniquement des variables auxiliaires au niveau de l'unité et ii) uniquement des variables auxiliaires au niveau du domaine (c'est-à-dire le modèle du contexte de l'unité). Au moyen d'une série de simulations, nous démontrons que dans le cas d'erreurs importantes de spécification du modèle, la procédure de MPO pour le modèle au niveau du domaine et le modèle du contexte de l'unité fournit de meilleurs résultats que le modèle correspondant au niveau de l'unité reposant uniquement sur des variables auxiliaires propres à l'unité.

Nous proposons deux raisons potentielles pouvant expliquer ces moins bons résultats de la méthode MPO-UNIT. D'abord, la différence entre les moyennes d'échantillon et de population des variables auxiliaires peut influer négativement sur les résultats de la méthode MPO-UNIT. Deuxièmement, la stabilité numérique des MEP pour les composantes de la variance reposant sur les variables auxiliaires au niveau de l'unité pourrait être également un facteur contributif. Nos résultats laissent entendre que l'utilisation de la procédure MPO avec des variables auxiliaires au niveau du domaine est une option prometteuse, en particulier lorsque des défis (comme le manque de renseignements de recensement et les erreurs de spécification du modèle) sont une préoccupation.

De plus, au moment de choisir entre le modèle du contexte de l'unité et le modèle au niveau du domaine, nous recommandons le modèle au niveau du domaine, car celui-ci peut présenter une plus grande stabilité. Dans le modèle au niveau du domaine, la variance des erreurs d'échantillonnage peut être estimée à l'aide de méthodes de lissage et dépend moins du modèle théorique pouvant comporter des erreurs de spécification. Enfin, nous recommandons de mener de plus amples travaux de recherche théoriques et empiriques pour étudier et mieux comprendre ces résultats inattendus de la méthode MPO pour les modèles à erreurs emboîtées reposant seulement sur des variables auxiliaires au niveau de l'unité.

Bibliographie

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

Cochran, W.G. (1977). *Sampling Techniques*. New York: John wiley & Sons, Inc.

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

Ghosh, M. et Lahiri, P. (1987). Robust empirical bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 82(400), 1153-1162

Jiang, J., Nguyen, T. et Rao, J.S. (2011). Best predictive small area estimation. *Journal of the American Statistical Association*, 106(494), 732-745.

Jiang, J., Nguyen, T. et Rao, J.S. (2015). Meilleure prédiction observée par régression à erreurs emboîtées sous spécification éventuellement inexacte de la moyenne et de la variance. *Techniques d'enquête*, 41(1), 39-58. Paper available at https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2015001/article/14200-fra.pdf.

Newhouse, D.L., Merfeld, J.D., Ramakrishnan, A., Swartz, T. et Lahiri, P. (2022). Small area estimation of monetary poverty in mexico using satellite imagery and machine learning. Disponible au SSRN 4235976.

Otto, M.C. et Bell, W.R. (1995). Sampling error modelling of poverty and income statistics for states. *Proceedings of the Section on Government Statistics*, American Statistical Association, 160-165.

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