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

Techniques d’enquête

Stratégies modifiées de meilleure prédiction observée pour l’estimation sur petits domaines avec des données au niveau de l’unité

par Jiangshan Zhang et Jiming Jiang

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. 413-432
Statistique Canada, nº 12-001-X au catalogue

Stratégies modifiées de meilleure prédiction observée pour l'estimation sur petits domaines avec des données au niveau de l'unité

Jiangshan Zhang et Jiming Jiang[1]

Résumé

La meilleure prédiction observée (MPO) dans le cadre d'un modèle de régression à erreurs emboîtées (REE) a déjà été proposée en utilisant une erreur quadratique moyenne de prédiction (EQMP) fondée sur le plan comme outil pour calculer le meilleur estimateur prédictif (MEP). Une étude récente a montré que la MPO selon le modèle de REE pouvait être associée à une instabilité numérique lors du calcul du MEP. Nous proposons plusieurs modifications de la MPO en vertu du modèle de REE, y compris celles utilisant une EQMP fondée sur un modèle pour calculer le MEP, afin d'améliorer la stabilité numérique et le rendement prédictif. Nous comparons le rendement des stratégies modifiées de MPO avec celui des méthodes existantes dans le cadre d'une étude par simulation. Un exemple de données réelles est examiné.

Mots-clés : Erreur de spécification du modèle; estimation sur petits domaines; niveau de l'unité; robustesse.

1. Introduction

Des méthodes d'estimation sur petits domaines (EPD) fondées sur un modèle sont habituellement élaborées selon certaines hypothèses de modèle. Comme l'a déclaré George Box, « tous les modèles sont incorrects ». Il est très probable qu'une erreur de spécification du modèle soit souvent inévitable dans les applications pratiques. Bien sûr, il existe des moyens d'utiliser des modèles plus complexes, comme les termes d'ordre plus élevés et la modélisation non paramétrique, afin de rendre le modèle supposé plus flexible. En revanche, il existe des raisons pratiques pour lesquelles un modèle plus simple et plus facile à interpréter peut être privilégié dans la pratique. Par exemple, un modèle linéaire est intéressant en raison de sa simplicité et de sa facilité d'interprétation et il utilise des variables auxiliaires de manière simple. Il convient de noter que les données auxiliaires peuvent être recueillies à l'aide de l'argent des contribuables; par conséquent, il peut être « politiquement incorrect » de ne pas utiliser ces renseignements. D'un point de vue statistique, il est possible d'exécuter une procédure de sélection de modèle, qui peut aboutir à un modèle plus sophistiqué en éliminant certaines des variables auxiliaires. Mais, en fin de compte, personne n'utilisera un tel modèle dans la pratique, en raison de sa complexité et du fait qu'il n'utilise pas les renseignements recueillis par l'intermédiaire de l'argent des contribuables. Ainsi, dans de tels cas, il n'est pas possible de modifier un modèle déjà adopté; tout ce que l'on peut faire, c'est trouver une meilleure façon d'ajuster le modèle supposé, de sorte que les prédicteurs de petits domaines qui en résultent soient plus robustes face à d'éventuelles erreurs de spécification du modèle.

Jiang, Nguyen et Rao (2011) se sont engagés dans une telle situation en proposant une méthode d'EPD appelée meilleure prédiction observée (MPO), qui diffère considérablement de la méthode traditionnelle du

1. Jiangshan Zhang et Jiming Jiang, Department of Statistics, University of California, Davis, CA 95616. Courriel : jimjiang@ucdavis.edu.

meilleur prédicteur linéaire sans biais empirique (MPLSBE). L'idée est de tenir compte de deux modèles : l'un est le modèle supposé, utilisé pour calculer le meilleur prédicteur (MP) lorsque le modèle supposé est retenu; l'autre est un modèle plus large utilisé pour obtenir l'estimation du paramètre selon une mesure de l'erreur quadratique moyenne de prédiction (EQMP) observée, permettant d'obtenir le meilleur estimateur prédictif (MEP). La MPO est le MP avec les paramètres inconnus concernés, qui sont remplacés par leur MEP. Dans Jiang et coll. (2011), les auteurs ont calculé la MPO selon le modèle de Fay-Herriot au niveau du domaine (Fay et Herriot, 1979). Dans ce cas, une EQMP fondée sur un modèle est utilisée pour calculer le MEP, ce qui suppose que les moyennes des petits domaines sont totalement non précisées. Nous l'appellerons MPO du modèle par domaine (MD) dans le présent article. Les auteurs ont montré, tant sur le plan théorique qu'empirique, que la MPO MD est plus robuste que la méthode traditionnelle du MPLSBE (par exemple Rao et Molina, 2015) sous le même modèle.

Dans Jiang et coll. (2011), les auteurs ont également tenu compte du modèle de régression à erreurs emboîtées (REE) au niveau de l'unité (Battese, Harter et Fuller, 1988), auquel cas une EQMP fondée sur le plan est utilisée pour calculer le MEP. C'est ce que nous appellerons MPO au niveau de l'unité fondée sur le plan (NU/FP) dans le présent article. Les auteurs n'ont pas étudié le rendement de la MPO NU/FP proposée. Jiang, Nguyen et Rao (2015) ont montré de façon empirique que la MPO NU/FP était plus robuste que le MPLSBE selon le modèle de REE. Toutefois, dans une étude récente, Chen, Lahiri et Salvati (2025) ont présenté des résultats de simulation surprenants, qui démontrent le contraire des conclusions de Jiang et coll. (2015) : la MPO NU/FP peut enregistrer un faible rendement en ce qui concerne l'EQMP comparativement au MPLSBE et à certaines méthodes de MPO modifiées proposées par Chen et coll. (2025). Il semble que la stabilité numérique de la MPO NU/FP soit la cause de son piètre rendement, en particulier en ce qui concerne l'obtention du MPE des composantes de la variance dans le cadre du modèle de REE.

L'objectif du travail en cours d'exécution est double. Premièrement, nous proposons de modifier la MPO NU/FP en imposant des limites à la recherche sur l'espace des paramètres dans l'optimisation de l'EQMP fondée sur le plan pour calculer le MEP. Deuxièmement, nous avons proposé une MPO au niveau de l'unité fondée sur un modèle, appelée MPO NU/BM (basée sur un modèle), dans le cadre du modèle de REE. Nous nous rappelons que la MPO NU/FP est obtenue à l'aide d'une EQMP fondée sur le plan. La nouvelle MPO NU/BM se distingue par le fait qu'elle repose sur une EQMP basée sur un modèle par rapport à un modèle plus large, similaire à celui de la MPO MD (Jiang et coll., 2011; voir ci-dessus). Nous proposons également une autre approche qui utilise des estimateurs « plug-in » des composantes de la variance, calculés à partir d'un modèle de Fay-Herriot induit, de sorte que l'optimisation de l'EQMP basée sur un modèle ne concerne que le MEP des effets fixes.

Les progrès méthodologiques sont décrits à la section suivante. Dans la section 3, nous menons une étude par simulation comparant le rendement de différents types de MPO et de MPLSBE en vertu du modèle de REE, notamment les modifications que nous proposons et certaines modifications proposées par Chen et coll. (2025). Un exemple de données réelles est examiné à la section 4.

2. Méthodes

Nous décrivons d'abord brièvement le modèle de REE et la MPO NU/FP, avant de proposer plusieurs stratégies de MPO modifiées dans le même contexte de modèle.

2.1 Le modèle de REE et la MPO NU/FP

Envisageons l'échantillonnage à partir de sous-populations finies, qui sont considérées comme des échantillons de certaines super-populations. Supposons que les sous-populations de résultats, $\{Y_{ik}, k = 1,\ldots, N_i\}$, et les données auxiliaires, $\{X_{ikl}, k = 1,\ldots, N_i\}$, $l = 1,\ldots, p$, sont des réalisations des super-populations correspondantes satisfaisant au modèle de REE suivant,

$$Y_{ik} = X'_{ik}\beta + v_i + e_{ik},\ i = 1,\ldots, m,\ k = 1,\ldots, N_i, \tag{2.1}$$

où β est un vecteur de coefficients de régression inconnus, v_i est un effet aléatoire propre à la sous-population et e_{ik} est une erreur. On suppose que les v_i et e_{ik} sont indépendants avec $v_i \sim N(0, \sigma_v^2)$ et $e_{ik} \sim N(0, \sigma_e^2)$.

Dans le cadre d'une population finie, la véritable moyenne de petits domaines est

$$\theta_i = \bar{Y}_i = \frac{1}{N_i}\sum_{k=1}^{N_i} Y_{ik}$$

pour $1 \le i \le m$. De plus, écrivez $r_i = n_i / N_i$. Supposons que des échantillons aléatoires simples (EAS, par exemple Lohr, 2021) $(y_{ij}, x_{ij}), 1 \le j \le n_i$ sont tirés de la population finie, $\{(Y_{ik}, X_{ik}), 1 \le k \le N_i\}$, $1 \le i \le m$. Ensuite, une version de population finie du meilleur prédicteur (MP), au sens de l'EQMP minimale, a l'expression suivante :

$$\tilde{\theta}_i = \mathrm{E}_M(\theta_i \mid y_i) = \bar{X}'_i\beta + \left\{r_i + (1 - r_i)\frac{n_i\rho}{1 + n_i\rho}\right\}(\bar{y}_{i\cdot} - \bar{x}'_{i\cdot}\beta) \tag{2.2}$$

où $y_i = (y_{ij})_{1 \le j \le n_i}$, $\bar{y}_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}$, $\bar{x}_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i} x_{ij}$. Ici, E_M désigne l'espérance (conditionnelle) selon le modèle de REE supposé, β et $\rho = \sigma_v^2 / \sigma_e^2$ sont les paramètres réels selon le modèle de REE. Veuillez noter que le MP dépend du modèle.

L'étape suivante consiste à calculer le MEP de $\psi = (\beta', \rho)'$. Soit $\theta = (\theta_i)_{1 \le i \le m}$, le vecteur des moyennes de petits domaines, et $\tilde{\theta} = [\tilde{\theta}_i]_{1 \le i \le m}$, le vecteur des MP, considérées comme des fonctions de ψ, c'est-à-dire $\tilde{\theta}_i = \tilde{\theta}_i(\psi)$. Les EQMP fondées sur le plan (par exemple Fuller, 2009) peuvent être exprimées sous la forme

$$\mathrm{EQMP}(\tilde{\theta}) = \mathrm{E}(|\tilde{\theta} - \theta|^2) = \sum_{i=1}^{m} \mathrm{E}\{\tilde{\theta}_i(\psi) - \theta_i\}^2 = \mathrm{E}\{Q(\psi) + \cdots\} \tag{2.3}$$

où $\cdots$ ne dépend pas de ψ. Il convient de noter que le E en (2.3) est différent du E_M en (2.2) en ce sens que E est entièrement sans modèle; à savoir, l'espérance en (2.3) concerne l'EAS par rapport aux

populations finies, ce qui n'a rien à voir avec le modèle supposé. De plus, la fonction $Q(\psi)$ a l'expression suivante :

$$Q(\psi) = \sum_{i=1}^{m} \left\{ \tilde{\theta}_i^2(\psi) - 2\frac{1-r_i}{1+n_i\rho}\bar{y}_{i\cdot}\bar{X}_i'\beta + b_i(\rho)\hat{\mu}_i^2 \right\} = \sum_{i=1}^{m} Q_i \tag{2.4}$$

où $b_i(\rho) = 1 - 2a_i(\rho)$ avec $a_i(\rho) = r_i + (1-r_i)n_i\rho(1+n_i\rho)^{-1}$ et Q_i est défini de façon évidente. De plus, $\hat{\mu}_i^2$ est un estimateur sans biais selon le plan de $\bar{Y}_i^2$ sous l'EAS, qui a l'expression suivante :

$$\hat{\mu}_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{N_i - 1}{N_i(n_i-1)}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 \tag{2.5}$$

en supposant que $n_i > 1$. Le MEP de ψ, $\hat{\psi}$ est le minimiseur de $Q(\psi)$ par rapport à ψ.

La MPO NU/FP de θ_i est obtenue par (2.2) avec ψ remplacé par son MEP.

2.2 MPO restreinte pour le modèle de REE

Bien que, intuitivement, on s'attende à ce que la MPO NU/FP soit plus robuste que le MPLSBE lorsque le modèle de REE est mal défini, Chen et coll. (2025) ont fait état de résultats de simulation surprenants qui démontrent le contraire. Il semble que le problème soit lié à l'optimisation numérique de (2.4) sur un espace de paramètres non restreint de ψ.

Il convient de noter que, bien que (2.3) soit valable pour tout ψ fixe, cela pourrait ne plus être le cas lorsque ψ est aléatoire, c'est-à-dire, $\psi = \tilde{\psi}$, qui dépend des x et y échantillonnés. Toutefois, il s'avère que l'équation (2.3) est toujours à peu près valable si $\tilde{\psi}$ ne varie pas trop. On a observé de façon empirique que l'estimateur REML (maximum de vraisemblance restreint) était « toujours » stable. L'instabilité numérique est attribuable au MEP selon le modèle de REE. Cela entraîne la modification suivante de la MPO NU/FP.

Soit $\tilde{\psi}$ l'estimateur REML de ψ. Nous proposons d'optimiser (2.4) sur un petit quartier de $\tilde{\psi}$, c'est-à-dire, $\psi_k \in [(1-\delta)\tilde{\psi}_k, (1+\delta)\tilde{\psi}_k]$ pour la k la composante de ψ, $1 \le k \le p+1$, où $p = \dim(\beta)$, et δ est un petit nombre positif, tel que $\delta = 0{,}05$. Cela mène au MEP restreint (MEPR). Pour le choix de δ, il peut être obtenu par des techniques axées sur les données. Voir la section 3.3 pour de plus amples renseignements.

La MPO NU/FP restreinte (MPO NU/FPR) est obtenue de la même façon que la MPO NU/FP, mais en utilisant le MEPR au lieu du MEP.

2.3 Une MOP basée sur un modèle pour le modèle de REE

Jiang et coll. (2011) ont obtenu la MPO NU/FP à l'aide d'une EQMP fondée sur le plan (EAS). Bien que cette dernière soit presque dépourvue d'hypothèses de modèle, tout dépend du plan d'échantillonnage. Par exemple, si un échantillonnage stratifié (par exemple Lohr, 2021) est utilisé, la MPO NU/FP ne sera pas valide. Bien sûr, on pourrait modifier le calcul de la MPO NU/FP pour intégrer le plan d'échantillonnage. Mais cela nécessiterait une approche au cas par cas, plutôt qu'une formule unique et universelle.

En revanche, bien qu'il s'agisse d'une EQMP fondée sur un modèle, comme celle appliquée dans le cadre du modèle plus large de Jiang et coll. (2011) pour calculer la MPO MD (voir ci-dessus), il peut quand même y avoir une hypothèse de modèle faible, cela ne dépend pas du plan d'échantillonnage. En fait, le modèle plus large est presque exempt de modèle, dont l'extension au niveau de l'unité peut être exprimée sous la forme

$$Y_{ik} \;=\; \mu_{ik} \;+\; v_i \;+\; e_{ik},\; i=1,\ldots,m,\; k=1,\ldots,N_i, \tag{2.6}$$

où μ_{ik} est la moyenne inconnue de l'unité k^{e} dans la i^{e} sous-population, et v_i et e_{ik} sont les mêmes que dans (2.1). Nous formulons l'hypothèse suivante.

A1. La moyenne par domaine-échantillon, $\overline{y}_{i\cdot}$, est un estimateur sans biais sous le modèle de la moyenne par domaine de la population, $\overline{Y}_i = N_i^{-1}\sum_{k=1}^{N_i} Y_{ik}$, c'est-à-dire $\mathrm{E}(\overline{y}_{i\cdot}) = \mathrm{E}(\overline{Y}_i) \equiv \overline{\mu}_i$, $1 \le i \le m$.

Le modèle plus large suppose que chaque unité a sa propre moyenne au niveau de l'unité, ce qui est très peu restrictif. Pour ce qui est de l'hypothèse supplémentaire *A1*, elle suppose que la moyenne de l'échantillon du domaine est un estimateur sans biais de la moyenne de la population du domaine, ce qui est raisonnable. Il convient de noter qu'il y a un compromis entre la simplicité et la robustesse du modèle. En général, un modèle plus flexible, comme le modèle plus large supposé ci-dessus, est plus robuste face à la violation des hypothèses de modèle (parce qu'il n'y a pas beaucoup d'hypothèses, après tout). Toutefois, le modèle plus large n'utilise pas l'information provenant des variables auxiliaires; en ce sens, il n'est pas utile en ce qui concerne l'élaboration de prédicteurs des moyennes de petits domaines. Le modèle plus large n'est utilisé que pour obtenir des estimateurs des paramètres du modèle afin que les prédicteurs résultants soient plus résistants à une erreur de spécification du modèle. Il s'agit de la même idée que pour la MPO (Jiang et coll., 2011). Par contre, un modèle plus simple, comme un modèle linéaire, est plus attrayant pour les formateurs et formatrices; en ce sens, il est plus utile. Néanmoins, un modèle plus simple est plus susceptible d'être mal défini, ce qui le rend moins résistant à la violation des hypothèses de modèle.

Selon le modèle plus large et *A1*, la moyenne sur petits domaines θ_i peut être exprimée sous la forme

$$\begin{aligned} \theta_i \;&=\; \mathrm{E}(\overline{Y}_i \,|\, v_i) \;=\; \frac{1}{N_i}\sum_{k=1}^{N_i} \mathrm{E}(Y_{ik} \,|\, v_i) \;=\; \frac{1}{N_i}\sum_{k=1}^{N_i} \mu_{ik} + v_i \\ &=\; \frac{1}{N_i}\sum_{k=1}^{N_i} \mathrm{E}(Y_{ik}) + v_i \;=\; \mathrm{E}(\overline{Y}_i) + v_i \;=\; \mathrm{E}(\overline{y}_{i\cdot}) + v_i. \end{aligned} \tag{2.7}$$

Soit $\overline{x}_{i\cdot}^{c}$ la moyenne de $x_{ik}, n_i < k \le N_i$ (rappelons que $1 \le j \le n_i$ correspond aux unités d'échantillonnage de la sous-population i). Alors, nous avons

$$\overline{x}_{i\cdot}^{c} \;=\; \frac{\overline{X}_i}{1-r_i} - \frac{r_i \overline{x}_{i\cdot}}{1-r_i}. \tag{2.8}$$

Avec (2.8), le MP dans (2.2) peut être écrit sous la forme

$$\tilde{\theta}_i(\psi) \;=\; r_i \overline{y}_{i\cdot} + (1-r_i)\, a_i(\sigma_v^2, \sigma_e^2) \left[\overline{y}_{i\cdot} - \left\{ \overline{x}_i - \frac{\overline{x}_{i\cdot}^{c}}{a_i(\sigma_v^2, \sigma_e^2)} \right\}' \beta \right] \tag{2.9}$$

où $a_i(\sigma_v^2,\sigma_e^2) = n_i\sigma_v^2 / (\sigma_e^2 + n_i\sigma_v^2)$. Dans le cadre du modèle plus large, nous pouvons évaluer les EQMP comme suit :

$$\text{EQMP}\{\tilde{\theta}(\psi)\} \;=\; \text{E}_\text{m}\{\tilde{\theta}(\psi)-\theta\}^2 = \sum_{i=1}^{m} \text{E}_\text{m}\{\tilde{\theta}_i(\psi)-\theta_i\}^2 \tag{2.10}$$

où $\tilde{\theta}(\psi) = [\tilde{\theta}_i(\psi)]_{1\le i\le m}$ et E_m désigne l'espérance fondée sur le modèle plus large. En substituant (2.9) dans (2.10), l'EQMP peut être exprimée sous la forme

$$\text{EQMP}\{\tilde{\theta}(\psi)\} \;=\; \sum_{i=1}^{m} \text{E}_\text{m}\left[r_i\overline{y}_{i\cdot} - \left(\overline{\mu}_i + v_i\right) + (1-r_i)\, a_i \left\{ \overline{y}_{i\cdot} - \left(\overline{x}_i - \frac{\overline{x}_{i\cdot}^c}{a_i} \right)' \beta \right\} \right]^2, \tag{2.11}$$

où a_i désigne $a_i(\sigma_v^2,\sigma_e^2)$. En suivant des calculs semblables à ceux de Jiang et coll. (2011), nous avons, selon l'hypothèse *A1* et après une certaine simplification, l'expression suivante :

$$\text{EQMP}\{\tilde{\theta}(\psi)\} \;=\; \text{E}_m(I_1 + 2I_2 + I_3), \tag{2.12}$$

où $I_1 = \sum_{i=1}^{m} \left(\overline{y}_i - \overline{\mu}_i - v_i\right)^2$,

$$I_2 \;=\; \sum_{i=1}^{m} (1-r_i)\, a_i \left(\overline{y}_i - \overline{\mu}_i - v_i\right) \left[(1-1/a_i)\, \overline{y}_i - (\overline{x}_{i\cdot}' - \overline{x}_{i\cdot}^c / a_i)'\beta \right]$$

et $I_3 = \sum_{i=1}^{m} (1-r_i)^2 a_i^2 \{(1-1/a_i)\, \overline{y}_i - (\overline{x}_{i\cdot}' - \overline{x}_{i\cdot}^c / a_i)'\beta\}^2$. Comme on peut le voir, I_1 ne dépend pas de β et, dans le cadre du modèle plus large, nous avons $\text{E}_\text{m}(I_1) = \sum_{i=1}^{m} \sigma_e^2 / n_i$. En ce qui concerne I_2, l'espérance dans le cadre du modèle plus large est

$$\begin{aligned}
\text{E}_\text{m}(I_2) &= \sum_{i=1}^{m} \text{E}_m \left[(1-r_i)\, a_i \left(\overline{y}_i - \overline{\mu}_i - v_i\right) \left\{ (1-1/a_i)\, \overline{y}_i - (\overline{x}_{i\cdot}' - \overline{x}_{i\cdot}^c / a_i)'\beta \right\} \right] \\
&= \sum_{i=1}^{m} (1-r_i)\, a_i \text{E}_\text{m} \left[\overline{e}_i \left\{ (1-1/a_i)\, \overline{y}_i - (\overline{x}_{i\cdot}' - \overline{x}_{i\cdot}^c / a_i)'\beta \right\} \right] \\
&= \sum_{i=1}^{m} (1-r_i)\, a_i \text{E}_\text{m} \left\{ \overline{e}_i (1-1/a_i)\, \overline{e}_i \right\} \\
&= \sum_{i=1}^{m} (1-r_i)\, (a_i - 1)\, \sigma_e^2 / n_i .
\end{aligned} \tag{2.13}$$

Il s'ensuit que (2.11) peut être exprimé sous la forme

$$\text{EQMP}\{\tilde{\theta}(\psi)\} \;=\; \text{E}_\text{m} \left(\sum_{i=1}^{m} \left[\frac{\sigma_e^2}{n_i} \left\{ 2(1-r_i)\, (a_i - 1) + 1 \right\} + (1-r_i)^2 a_i^2 \left\{ \left(1 - \frac{1}{a_i}\right) \overline{y}_i - \left(\overline{x}_{i\cdot}' - \frac{\overline{x}_{i\cdot}^c}{a_i} \right)' \beta \right\}^2 \right] \right). \tag{2.14}$$

Si nous paramétrons de nouveau ψ en $\psi = (\beta, \rho, \sigma_e^2)$ (rappel, $\rho = \sigma_v^2 / \sigma_e^2$), nous pouvons réécrire a_i comme $a_i = \rho n_i / (1 + \rho n_i)$, ce qui ne dépend pas de σ_e^2. Ainsi, on peut voir que (2.14) est une fonction qui augmente de manière monotone avec σ_e^2. Cela signifie qu'il est impossible d'estimer σ_e^2 en optimisant

(2.14). L'estimation de σ_e^2 sera abordée plus tard. Pour l'instant, supposons que σ_e^2 est connu. Supposons que $\gamma = (\beta, \rho)$. Ensuite, en suivant l'idée de la MPO, nous estimons γ en minimisant l'expression à l'intérieur de l'espérance du côté droit de (2.14), c'est-à-dire,

$$\begin{aligned} Q(\gamma) &= \sum_{i=1}^{m} \left[\frac{\sigma_e^2}{n_i} \{2(1-r_i)(a_i-1)\} + (1-r_i)^2 a_i^2 \left\{ \left(1-\frac{1}{a_i}\right)\bar{y}_i - \left(\bar{x}_{i\cdot}' - \frac{\bar{x}_{i\cdot}^{c\prime}}{a_i}\right)\beta \right\}^2 \right] \\ &= \sum_{i=1}^{m} \left[\frac{\sigma_e^2}{n_i} \{2(1-r_i)(a_i-1)\} + \{(g_i-1)\bar{y}_i - (g_i\bar{x}_{i\cdot} - \bar{X}_i)'\beta\}^2 \right], \end{aligned} \tag{2.15}$$

en utilisant (2.8), où $g_i = r_i + (1-r_i)a_i$. Il convient de noter que $Q(\gamma)$ est lié à la différence entre $g_i\bar{x}_{i\cdot}$ et $\bar{X}_i$. Étant donné ρ et un σ_e^2 prédéfini, le β qui minimise $Q(\gamma)$ est

$$\tilde{\beta}\left(\rho, \sigma_e^2\right) = (g_i - 1)(H_i'H_i)^{-1}H_i\bar{y}_i \tag{2.16}$$

avec $H_i = H_i\left(\rho, \sigma_e^2\right) = (g_i\bar{x}_{i\cdot} - \bar{X}_i)$. Soit $H = (H_i)_{1\le i\le m}$ et $L = (g_i\bar{y}_i)_{1\le i\le m}$. Nous intégrons ensuite (2.16) dans (2.15), puis nous minimisons l'expression suivante par rapport à $\rho \ge 0$:

$$\tilde{Q}(\rho) = Q(\gamma)|_{\beta=\tilde{\beta}(\rho,\sigma_e^2)} = \sum_{i=1}^{m} \frac{2\sigma_e^2}{n_i}(1-r_i)(a_i-1) + L'\left[I_m - H'(H'H)^{-1}H\right]L. \tag{2.17}$$

Le minimiseur est alors noté $\tilde{\rho}$. Le MEP de β est ensuite obtenu par (2.16) avec $\rho = \tilde{\rho}$.

Quant au cas où σ_e^2 est inconnu, il peut, en pratique, être estimé à l'aide d'une technique de lissage, telle que celle démontrée dans Otto et Bell (1995). Ici, nous recommandons d'utiliser l'estimateur REML standard de σ_e^2 (par exemple Jiang et Nguyen, 2021, section 1.3.2) à la place de σ_e^2 dans le calcul ci-dessus.

2.4 Modèle induit au niveau du domaine

Chen et coll. (2025) ont également considéré un modèle au niveau du domaine induit par le modèle de REE :

$$\bar{y}_{i\cdot} = \bar{X}_i'\beta + v_i + \bar{e}_{i\cdot},\ i = 1,\ldots,m, \tag{2.18}$$

où $\bar{e}_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i} e_{ij}$. On suppose que le modèle induit satisfait à l'hypothèse du modèle de Fay-Herriot, selon laquelle la variance échantillonnale pour $\bar{e}_{i\cdot}$ est estimée comme $D_i = s^2/n_i$ et s^2 est la variance de l'échantillon regroupée fondée sur les données de tous les m petits domaines. Ces derniers auteurs ont considéré la MPO MD basée sur (2.18) et ont montré qu'elle était plus performante que la MPO NU/FP et le MPLSBE dans leur étude par simulation.

Ici, nous proposons une approche différente en utilisant les données au niveau de l'unité. Premièrement, nous utilisons le modèle de Fay-Herriot induit, (2.18), pour obtenir l'estimateur de σ_v^2. L'estimateur de σ_e^2 est le même que ci-dessus, c'est-à-dire s^2. Nous estimons ensuite β en minimisant (2.15) sur β, où σ_v^2 et σ_e^2 sont remplacés par ces estimateurs ci-dessus. Nous appelons cette méthode MPO NU/FH.

La motivation derrière la MPO NU/FH est similaire à celle de la MPO restreinte abordée à la section 2.2. Notre étude empirique a montré que l'instabilité numérique de la MPO NU/FP est principalement attribuable aux estimations erratiques des composantes de variance. Par ailleurs, le rendement de l'estimateur β est relativement stable une fois que les composantes de variance sont connues.

3. Études de simulation

Pour examiner le rendement des MPO modifiées proposées dans le cadre du modèle de REE ainsi que leur comparaison avec les méthodes existantes, nous envisageons un contexte de simulation semblable à celui de Jiang et coll. (2015). Pour être plus précis, nous considérons un modèle avec une seule variable auxiliaire, présumée être associée linéairement à la variable réponse y_{ij} comme

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

où x_{ij} est la valeur de la variable auxiliaire pour la j^{e} unité du i^{e} domaine, β_1 est un coefficient inconnu associé à la variable auxiliaire, v_i et e_{ij} sont comme en (2.1).

Nous avons comparé le rendement des prédicteurs suivants :

1) l'estimateur direct (dir);
2) MPLSBE selon le modèle de REE (mplsbe);
3) MPLSBE en contexte unitaire, calculé sous (2.1) avec $X'_{ik}\beta$ remplacé par $\beta_1 \overline{X}_i$ (mplsbe-cu);
4) la MPO NU/FH (section 2.4) (mpo-nu/fh);
5) MPO MD selon le modèle de Fay-Herriot induit (section 2.4) (mpo-fh);
6) la MPO en contexte unitaire, l'équivalent de la MPO NU/FP pour (3) (mpo-cu);
7) la MPO NU/BM (section 2.3) (mpo-nu/bm);
8) la MPO NU/FPR (section 2.2) (mpo-nu/FPR).

Les deux prédicteurs de contexte unitaire, 3) et 6), en plus de 5), ont été proposés par Chen et coll. (2025).

Nous examinons ci-dessous deux scénarios d'erreur de spécification de modèle par rapport au modèle de REE proposé.

3.1 Erreur de spécification complète

Nous commençons par une situation d'erreur de spécification complète. La réponse y pour la population finie a été générée à partir du modèle de REE de superpopulation :

$$y_{ij} = b + v_i + e_{ij},\ i = 1,\ldots,m,\ j = 1,\ldots,N_i. \tag{3.2}$$

La taille de la population de tous les domaines m a été fixée à $N_1 = \cdots = N_m =$ 1 000. De même, les tailles d'échantillon de tous les domaines ont été établies comme $n_i = 4,$ et des échantillons d'EAS ont été tirés. Deux valeurs constantes différentes $b =$ 5, 10 ont été prises en compte et des simulations ont été effectuées

avec le nombre total de petits domaines $m =$ 40, 100 ou 400. L'effet propre au domaine v_i a été généré à partir de $N(0,1)$, et l'erreur e_{ij} a été générée à partir de $N(0,\sigma_i^2)$, où σ_i^2 sont générés indépendamment par la distribution gamma Γ (3; 0,5). La variable auxiliaire x de la population finie a été générée indépendamment à partir d'une distribution normale logarithme logNormal(1; $0{,}5^2$). Pour chaque scénario, $K =$ 1 000 réplications ont été effectuées.

Comme mesures de rendement, nous avons tenu compte des EQMP empiriques globales et propres au domaine :

$$\text{EQMP}_i \approx \frac{1}{K}\sum_{k=1}^{K}\left\{\hat{\theta}_i^{(k)} - \theta_i^{(k)}\right\}^2$$

$$\text{EQMP} \approx \frac{1}{K}\sum_{k=1}^{K}\frac{1}{m}\sum_{i=1}^{m}\left\{\hat{\theta}_i^{(k)} - \theta_i^{(k)}\right\}^2, \tag{3.3}$$

où $\hat{\theta}_i^{(k)}$ et $\theta_i^{(k)}$ sont les moyennes prédites de petits domaines, respectivement, pour le domaine i dans la simulation k^{e}. Nous tenons également compte du biais de prédiction global :

$$\text{Biais} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{m}\sum_{i=1}^{m}\left\{\hat{\theta}_i^{(k)} - \theta_i^{(k)}\right\} \tag{3.4}$$

et l'erreur-type (ET) moyenne :

$$\text{ET}_i = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left\{\hat{\theta}_i^{(k)} - \bar{\theta}_i^{(k)}\right\}^2} \quad \text{et} \quad \text{ET} = \frac{1}{m}\sum_{i=1}^{m}\text{ET}_i\,. \tag{3.5}$$

Le tableau 3.1 présente les EQMP globales simulées, le biais et l'ET pour différents nombres de domaines et combinaisons de tailles d'effet, dans le cas d'une erreur de spécification complète du modèle, c'est-à-dire (3.2). Plusieurs observations s'imposent. En ce qui concerne l'EQMP globale, premièrement, les estimateurs directs ne changent pas beaucoup entre les scénarios de simulation. C'est parce qu'elle n'est pas basée sur un modèle et que les paramètres des composantes de variance n'ont pas été modifiés. Deuxièmement, lorsque $b = 10$, certains prédicteurs peuvent avoir un rendement inférieur à l'estimateur direct; en fait, les MPLSBE ont obtenu le pire rendement dans ce cas. Par contre, lorsque $b = 5$, l'estimateur direct a obtenu le pire rendement. Cela donne à penser que le modèle est utile tant que son erreur de spécification n'est pas trop grave. Troisièmement, les deux prédicteurs en contexte unitaire et les deux MPO associées au modèle de Fay-Herriot ont été plus performants que la MPO NU/BM et la MPO NU/FPR dans tous les scénarios de simulation. Parmi celles-ci, la MPO basée sur le modèle de Fay-Herriot induit présente la plus petite EQMP globale. Enfin, les trois MPO au niveau unitaire ont dépassé le MPLSBE au niveau unitaire dans tous les scénarios; parmi ces trois MPO au niveau unitaire, la MPO NU/FH a obtenu les meilleurs résultats, suivie de la MPO NU/BM et de la MPO NU/FPR. En fait, dans certains cas, l'indice MPO NU/FH a obtenu de meilleurs résultats que la MPO en contexte unitaire et le MPLSBE en contexte unitaire.

Tableau 3.1
EQMP globale simulée des prédicteurs : cas d'erreur de spécification complète

	(m = 40, b = 10)			(m = 100, b = 10)			(m = 400, b = 10)		
Type	**EQMP**	**Biais**	**ET**	**EQMP**	**Biais**	**ET**	**EQMP**	**Biais**	**ET**
dir	1,502	0,005	1,214	1,494	-0,003	1,225	1,494	-0,002	1,223
mplsbe	1,715	0,353	1,259	1,628	0,245	1,250	1,573	0,186	1,241
mplsbe-cu	0,823	0,006	0,900	0,793	-0,002	0,890	0,784	-0,001	0,881
mpo nu/fh	1,162	0,098	1,081	1,156	0,093	1,080	1,169	0,094	1,078
mpo-fh	0,710	0,006	0,840	0,643	-0,002	0,808	0,615	-0,001	0,784
mpo-cu	0,813	0,006	0,900	0,792	-0,002	0,890	0,779	-0,001	0,881
mpo nu/bm	1,509	0,178	1,247	1,508	0,185	1,239	1,499	0,186	1,220
mpo nu/fpr	1,598	0,266	1,222	1,583	0,226	1,239	1,562	0,201	1,224
	(m = 40, b = 5)			**(m = 100, b = 5)**			**(m = 400, b = 5)**		
Type	**EQMP**	**Biais**	**ET**	**EQMP**	**Biais**	**ET**	**EQMP**	**Biais**	**ET**
dir	1,497	0,003	1,229	1,507	0,005	1,219	1,509	0,005	1,226
mplsbe	1,092	0,475	0,937	1,092	0,496	0,889	1,018	0,498	0,876
mplsbe-cu	0,812	0,003	0,905	0,784	0,006	0,886	0,773	-0,001	0,880
mpo nu/fh	0,831	0,052	0,914	0,770	0,050	0,874	0,749	0,048	0,866
mpo-fh	0,777	0,004	0,849	0,648	0,006	0,802	0,613	-0,001	0,784
mpo-cu	0,810	0,004	0,906	0,782	0,006	0,886	0,770	-0,001	0,881
mpo nu/bm	0,918	0,105	0,956	0,857	0,090	0,915	0,829	0,081	0,881
mpo nu/fpr	0,936	0,200	0,952	0,871	0,168	0,914	0,842	0,118	0,907

Note : dir = direct; mplsbe = meilleur prédicteur linéaire sans biais empirique; EQMP = erreur quadratique moyenne de prédiction; ET = erreur type; mplsbe-cu = meilleur prédicteur linéaire sans biais empirique en contexte unitaire; mpo-cu = meilleure prédiction observée en contexte unitaire; mpo-fh = meilleure prédiction observée selon le modèle de Fay-Herriot; mpo nu/bm = meilleure prédiction observée au niveau de l'unité/basée sur un modèle; mpo nu/fh = meilleure prédiction observée au niveau de l'unité/Fay-Herriot; mpo nu/fpr = meilleure prédiction observée au niveau de l'unité/fondée sur le plan restreinte.

En ce qui concerne le biais, le MPLSBE présente le biais le plus élevé pour tous les scénarios, ce qui peut être attribué à l'erreur de spécification complète. Les MPO au niveau de l'unité ont un biais plus important que les autres prédicteurs en contexte unitaire et les MPO fondées sur le modèle de Fay-Herriot, mais les biais sont plus faibles que ceux des MPLSBE. La MPO NU/FH présente le biais le plus faible parmi les MPO au niveau de l'unité.

Enfin, en ce qui concerne l'ET, l'estimateur direct, le MPLSBE et les trois MPO au niveau de l'unité ont des valeurs plus élevées que les autres, le MPLSBE ayant l'ET la plus élevée lorsque $b = 10$. Toutefois, lorsque $b = 5$, l'estimateur direct a l'ET la plus élevée par rapport à tous les autres, et les différences entre les autres prédicteurs sont faibles.

3.2 Erreur de spécification partielle

En pratique, l'erreur de spécification complète peut être un peu extrême, puisque le modèle supposé est complètement différent du véritable modèle sous-jacent; une telle divergence peut être détectée statistiquement. Par conséquent, une situation plus réaliste est envisagée, où le modèle présumé est en partie correct comparativement au modèle réel. Dans ce cas, nous générons la population finie de y selon le modèle de REE suivant :

$$y_{ij} = b_0 + b_1 x_{ij} + \nu_i + e_{ij},\ i = 1,...,m,\ j = 1,...,N_i \tag{3.6}$$

où $b_0 = b$, comme dans le cas précédent, $b_1 = 4$, et les autres paramètres sont les mêmes que dans le cas d'erreur de spécification complète. Les EQMP globales simulées sont présentées dans le tableau 3.2.

Tableau 3.2
EQMP globale simulée des prédicteurs : cas d'erreur de spécification partielle

	(m = 40, b = 10)			(m = 100, b = 10)			(m = 400, b = 10)		
Type	**EQMP**	**Biais**	**ET**	**EQMP**	**Biais**	**ET**	**EQMP**	**Biais**	**ET**
dir	9,939	-0,002	3,145	9,914	-0,010	3,141	9,862	-0,003	3,136
mplsbe	1,752	0,397	1,271	1,637	0,288	1,252	1,582	0,191	1,241
mplsbe-cu	1,428	-0,001	1,191	1,148	-0,008	1,075	1,148	-0,002	1,011
mpo nu/fh	1,568	0,107	1,230	1,170	0,069	1,077	0,969	0,058	0,984
mpo-fh	1,598	-0,001	1,256	1,211	-0,009	1,104	1,008	-0,002	1,004
mpo-cu	1,427	-0,001	1,190	1,148	-0,009	1,075	1,019	-0,002	1,010
mpo nu/bm	1,591	0,314	1,213	1,539	0,267	1,212	1,509	0,195	1,208
mpo nu/fpr	1,681	0,334	1,243	1,623	0,267	1,246	1,596	0,199	1,243
	(m = 40, b = 5)			**(m = 100, b = 5)**			**(m = 400, b = 5)**		
Type	**EQMP**	**Biais**	**ET**	**EQMP**	**Biais**	**ET**	**EQMP**	**Biais**	**ET**
dir	9,939	0,010	3,158	9,848	0,002	3,141	9,906	-0,004	3,141
mplsbe	1,109	0,464	0,941	1,041	0,487	0,898	1,016	0,498	0,877
mplsbe-cu	1,019	0,011	1,191	1,400	0,003	1,068	1,005	-0,004	1,005
mpo nu/fh	1,436	0,114	1,195	1,116	0,073	1,049	0,927	0,055	0,963
mpo-fh	1,568	0,011	1,261	1,207	0,003	1,090	1,003	-0,004	1,000
mpo-cu	1,400	0,011	1,191	1,146	0,003	1,068	1,005	-0,004	1,005
mpo nu/bm	1,045	0,465	0,931	0,955	0,477	0,892	0,874	0,475	0,785
mpo nu/fpr	1,064	0,341	0,974	0,982	0,320	0,935	0,899	0,241	0,915

Note : dir = direct; mplsbe = meilleur prédicteur linéaire sans biais empirique; EQMP = erreur quadratique moyenne de prédiction; ET = erreur type; mplsbe-cu = meilleur prédicteur linéaire sans biais empirique en contexte unitaire; mpo-cu = meilleure prédiction observée en contexte unitaire; mpo-fh = meilleure prédiction observée selon le modèle de Fay-Herriot; mpo nu/bm = meilleure prédiction observée au niveau de l'unité/basée sur un modèle; mpo nu/fh = meilleure prédiction observée au niveau de l'unité/Fay-Herriot; mpo nu/fpr = meilleure prédiction observée au niveau de l'unité/fondée sur le plan restreinte.

On constate que, en termes d'EQMP globale, dans ce cas, chaque prédicteur basé sur un modèle a clairement surpassé l'estimateur direct. De plus, lorsque $b = 10$, il ne semblait pas y avoir de meilleur performant clair. Cependant, globalement, les deux MPO associées au modèle de Fay-Herriot (MPO NU/FH et MPO-FH) et les deux prédicteurs en contexte unitaire ont été plus performants que le MPLSBE, la MPO NU/BM et la MPO NU/FPR. Par contre, lorsque $b = 5$, la MPO NU/BM semblait être le meilleur performant, suivie de la MPO NU/FPR.

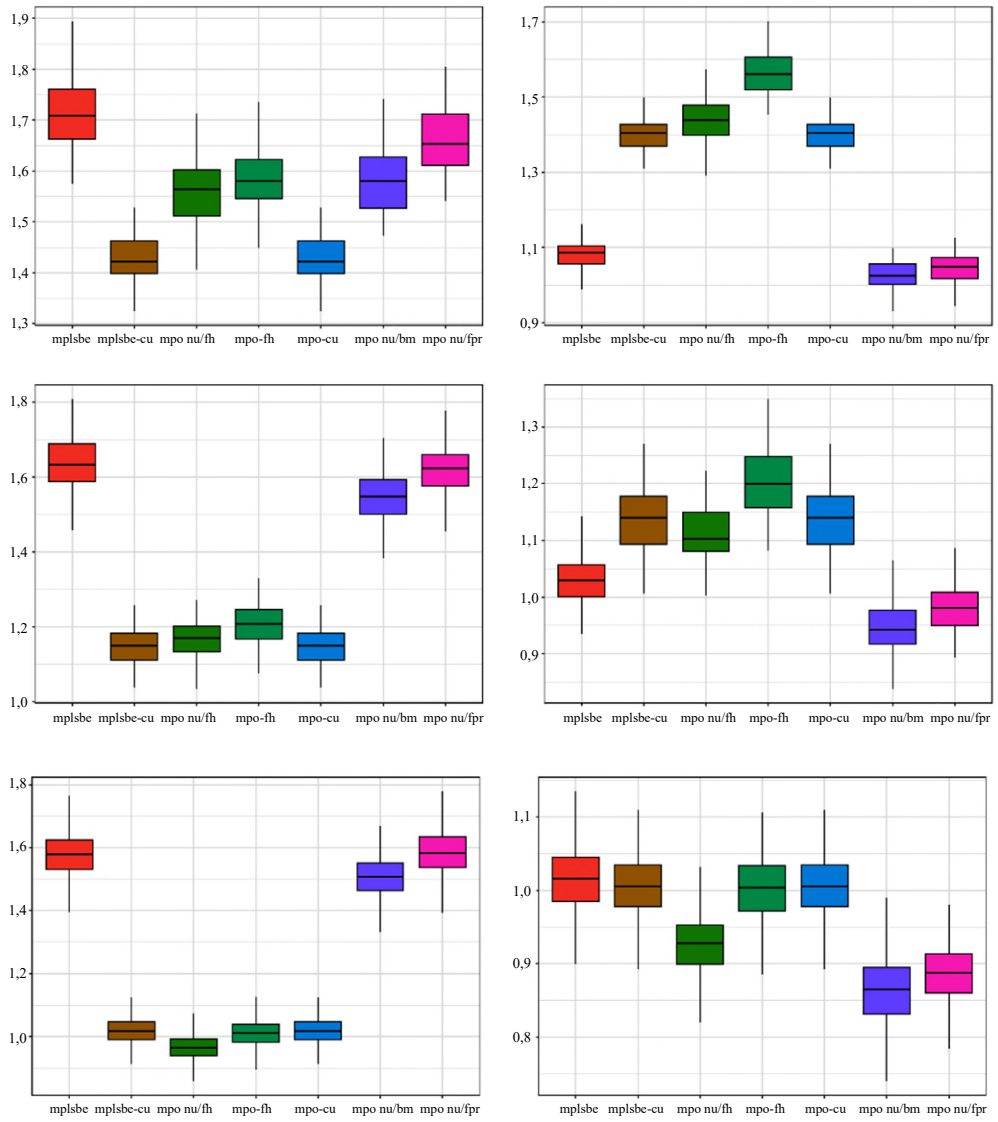
En ce qui concerne le biais, les prédicteurs basés sur le contexte unitaire et les estimateurs directs présentent toujours un biais plus faible, tandis que les méthodes basées sur le niveau unitaire affichent maintenant de meilleures performances par rapport au cas d'erreur de spécification complète. Encore une fois, le MPLSBE a le biais le plus élevé.

En ce qui concerne l'ET, l'estimateur direct a la valeur la plus élevée, car il ne tire aucun avantage du modèle (voir ci-dessous pour de plus de détails). Lorsque $b = 10$, les ET ne diffèrent pas beaucoup entre les autres prédicteurs. Cependant, lorsque $b = 5$, les prédicteurs basés sur le niveau unitaire, notamment le MPLSBE, ont une valeur plus faible par rapport aux prédicteurs en contexte unitaire et aux deux MPO associées au modèle de Fay-Herriot.

On constate que, comparativement au scénario d'erreur de spécification complète (voir le tableau 3.1), l'estimateur direct s'en tire beaucoup moins bien, comparativement à d'autres estimateurs, dans le scénario d'erreur de spécification partielle. En effet, dans le scénario d'erreur de spécification partielle, une relation linéaire existe toujours entre Y et X. Cela permet aux méthodes fondées sur un modèle d'en tirer parti, tandis que l'estimateur direct ne peut pas en profiter.

Les diagrammes en boîte des EQMP empiriques propres au domaine des sept prédicteurs basés sur un modèle sont présentés dans la figure 3.1, ce qui fournit des détails supplémentaires sur leur rendement relatif.

Figure 3.1 Diagrammes en boîte des EQMP empiriques propres au domaine : erreur de spécification du modèle partielle



Notes : - Colonne de gauche : $b = 10$; colonne de droite : $b = 5$. Haut : $m = 40$; milieu : $m = 100$; bas : $m = 400$.

- EQMP = erreur quadratique moyenne de prédiction; mplsbe = meilleur prédicteur linéaire sans biais empirique; mplsbe-cu = meilleur prédicteur linéaire sans biais empirique en contexte unitaire; mpo-cu = meilleure prédiction observée en contexte unitaire; mpo-fh = meilleure prédiction observée selon le modèle de Fay-Herriot; mpo nu/bm = meilleure prédiction observée au niveau de l'unité/basée sur un modèle; mpo nu/fh = meilleure prédiction observée au niveau de l'unité selon le modèle de Fay-Herriot; mpo nu/fpr = meilleure prédiction observée au niveau de l'unité/fondée sur le plan restreinte.

Résumé : En général, nous suggérons d'utiliser la MPO que le modèle soit spécifié de manière incorrecte ou non. Les résultats théoriques et empiriques (par exemple Jiang et coll., 2011, 2015) ont montré, ce qui est également confirmé par la tendance observée dans les études de simulation actuelles, que, si le modèle sous-jacent est correctement spécifié, ou presque correctement spécifié, la MPO et le MPLSBE obtiennent des résultats semblables; d'autre part, en cas d'erreur de spécification du modèle importante, la MPO peut obtenir un rendement nettement supérieur à celui du MPLSBE.

Cette différence semble indiquer en fait un moyen de diagnostiquer l'existence d'une grave erreur de spécification du modèle, que nous explorerons dans nos travaux futurs, mais dont l'idée de base est la suivante. Soit $\hat{\theta}_i$ et $\tilde{\theta}_i$ désignent respectivement la MPO et le MPLSBE pour la i^{e} moyenne de petits domaines, $1 \le i \le m$. Il est possible d'élaborer une répartition asymptotique pour une version normalisée de la différence euclidienne au carré, $|\hat{\theta} - \tilde{\theta}|^2 = \sum_{i=1}^{m} (\hat{\theta}_i - \tilde{\theta}_i)^2$, selon l'hypothèse nulle selon laquelle il n'y a pas d'erreur de spécification du modèle. Si l'hypothèse nulle est rejetée, cela suggère une erreur de spécification du modèle grave. Dans ce dernier cas, la MPO serait la méthode privilégiée. De plus, dans ce cas, les MPO basées sur un modèle au niveau unitaire pourraient ne pas être aussi performantes que les MPO basées sur un modèle en contexte unitaire ou sur le modèle de Fay-Herriot.

Cette différence peut également être explorée au moyen d'une sélection axée sur les données de δ avec la MPO NU/FPR; voir la sous-section suivante. Veuillez noter que, lorsque $\delta = 0$, la MPO est identique au MPLSBE.

Enfin, en ce qui concerne l'efficacité des calculs, nous n'avons observé aucune différence significative entre les différentes méthodes dans nos études de simulation.

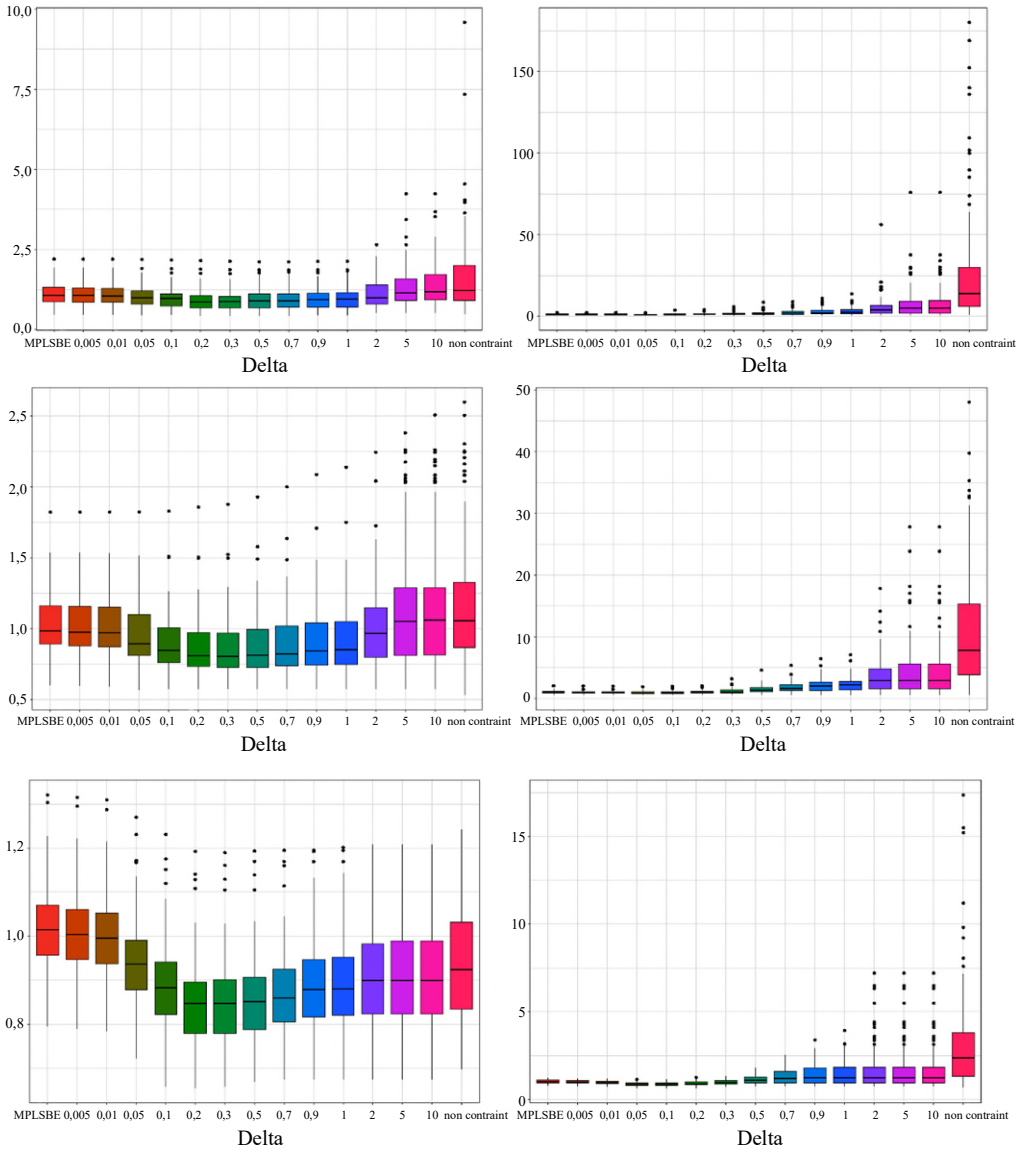
3.3 EQMP de la MPO NU/FPR restreinte vs δ

Pour approfondir la relation entre l'intervalle d'optimisation, δ, et le rendement de la MPO NU/FPR, nous avons mené une autre étude par simulation avec les mêmes paramètres que la section 3.2. Nous examinons la MPO NU/FP restreinte pour différentes valeurs de δ allant de 0 à 10, ainsi que la MPO NU/FP non restreinte, qui correspond à $\delta = \text{Inf}$. Il convient de noter que lorsque $\delta = 0$, aucune optimisation n'est effectuée, et le résultat est égal à la valeur initiale, ce qui mène au MPLSBE.

Les diagrammes en boîte des EQMP empiriques propres au domaine simulées sont présentés dans la figure 3.2. Ils montrent que, lorsque δ est grand, les EQMP propres au domaine de la MPO NU/FPR sont relativement grandes, et des valeurs aberrantes significatives sont observées lorsque $\delta = \text{Inf}$. Avec des valeurs de δ relativement faibles, mais pas trop faibles, comme de 0,05 à 0,5, les EQMP de la MPO NU/FPR sont plus petites que celles du MPLSBE, avec tous les paramètres de simulation. Lorsque l'erreur de spécification du modèle est modérée, $\delta =$ 0,01 à 0,2 est une bonne région pour obtenir des EQMP plus faibles, tandis que pour une erreur de spécification du modèle complète, la fourchette optimale de δ est plus grande, soit de 0,1 à 0,5. Dans l'ensemble, nos résultats de simulation donnent à penser que $\delta =$ 0,1 est un choix simple qui assure l'amélioration de la MPO NU/FPR par rapport au MPLSBE dans tous les cas.

De plus, à mesure que m augmente, l'intervalle optimal de δ est plus proche de 0. Le résultat semble indiquer que, lorsque m est grand, une optimisation plus affinée avec une valeur de δ plus petite est nécessaire.

Figure 3.2 Diagrammes en boîte des EQMP empiriques propres au domaine avec différentes plages du paramètre de contrainte δ pour $b = 5$ $b = 5$



Notes : - Colonne de gauche : erreur de spécification du modèle complète; colonne de droite : erreur de spécification du modèle partielle. Haut : $m = 40$; milieu : $m = 100$; bas : $m = 400$.

- EQMP = erreur quadratique moyenne de prédiction; MPLSBE = meilleur prédicteur linéaire sans biais empirique.

4. Exemple de données réelles : les données du TVSFP

Jiang et coll. (2015) ont utilisé les données du Television School and Family Smoking Prevention and Cessation Project (TVSFP; par exemple Hedeker, Gibbons et Flay, 1994) pour illustrer leur méthode de MPO dans le cadre du modèle de REE. Jiang, Rao, Fan et Nguyen (2018) ont utilisé les mêmes données pour démontrer leur méthode de prédiction par modèle mixte classifié. L'étude initiale visait à examiner les effets indépendants et combinés d'un programme scolaire de résistance sociale en milieu scolaire et d'un programme télévisé en termes de prévention du tabagisme et d'abandon du tabac.

Les sujets étaient des élèves de septième année de 28 établissements scolaires de Los Angeles (LA), en Californie, aux États-Unis. Les participants ont été testés en janvier 1986 dans le cadre d'une étude initiale. Les mêmes participants ont répondu à un questionnaire immédiatement après l'intervention en avril 1986, à un questionnaire de suivi un an plus tard (en avril 1987) et à un questionnaire de suivi deux ans plus tard (en avril 1988). Les établissements scolaires ont été répartis dans l'une des quatre conditions d'étude suivantes selon un processus de randomisation : a) un programme en classe (PC) axé sur la résistance sociale; b) une intervention dans les médias (TV); c) une combinaison de conditions PC et TV; d) un groupe témoin sans traitement. Une cote obtenue sur une échelle des connaissances sur le tabac et la santé (THKS pour *tobacco and health knowledge scale*) était l'une des principales variables de résultats de l'étude. Le questionnaire THKS comprenait sept questions utilisées pour évaluer les connaissances des élèves en matière de tabac et de santé. La cote THKS des participants a été définie comme la somme des éléments auxquels ils ont répondu correctement. Les données sont disponibles à l'adresse www.hsph.harvard.edu/fitzmaur/ala/tvsfp.txt.

La variable de résultat de cette analyse est la différence entre les cotes THKS avant et après l'intervention (après l'intervention moins le prétest). Au total, 1 600 élèves provenant des 28 écoles ont participé; le nombre d'élèves provenant de chaque école variait, de 18 à 137. Selon Jiang et coll. (2011), les petits domaines associés aux 28 écoles de Los Angeles sont définis comme les 28 sous-populations de participants, chaque sous-population correspondant aux caractéristiques de l'une des 28 écoles. Ainsi, les données des 28 écoles sont considérées comme des échantillons au niveau de l'unité provenant des 28 petits domaines. Jiang et coll. (2015) ont examiné le modèle de REE suivant pour les données du TVSFP :

$$y_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + v_i + e_{ij} \tag{4.1}$$

$i = 1,\ldots,28,\ j = 1,\ldots,n_i,$ où y_{ij} est le résultat au niveau de l'unité; x_{i1}, x_{i2} sont des indicateurs (1 ou 0) du PC et des programmes TV, respectivement; $\beta_j, j = 0,1,2,3$ sont des effets fixes inconnus, avec β_3 correspondant à l'interaction entre PC et TV; v_i est un effet aléatoire propre au domaine et e_{ij} est une erreur supplémentaire. On suppose que les effets aléatoires et les erreurs sont indépendants, avec $v_i \sim N(0,\sigma_v^2)$ et $e_{ij} \sim N(0,\sigma_e^2),$ où σ_v^2, σ_e^2 sont des variances inconnues.

Jiang et coll. (2015) ont fait remarquer que le modèle de REE proposé était probablement mal défini. Par exemple, le résultat n'est clairement pas normal. Les auteurs ont fait valoir que la MPO était appropriée dans cette situation. Ici, nous tenons compte des différentes modifications de la MPO étudiée à la section précédente, et de la méthode du MPLSBE, appliquée à cet ensemble de données. Les résultats sont présentés dans les tableaux 4.1 et 4.2. La MPO NU/FP d'origine de Jiang et coll. (2015), copiée des tableaux 4.1 et

4.2 de cette dernière référence, est également incluse à des fins de comparaison. Il semble que les différentes méthodes aient produit des valeurs prédites similaires pour les moyennes des petits domaines. En particulier, dans ce cas, la MPO NU/FH et la MPO FH ont produit les mêmes résultats. C'est parce que les variables auxiliaires sont au niveau du domaine et que le ratio n_i / N_i est traité comme nul, auquel cas les deux prédicteurs sont identiques. Pour des raisons semblables, les MPO NU/BM et MPO NU/FPR sont très semblables, mais pas identiques. Les coefficients de régression estimés (effets fixes) par les différentes méthodes sont présentés au tableau 4.3. Ces estimations paraissent en outre semblables et affichent la même tendance (par exemple positive ou négative).

Tableau 4.1
Moyennes de petits domaines prédites à partir des données du TVSFP (partie 1)

ID	PC	TV	mplsbe	mplsbe-cu	mpo nu/fh	mpo-fh	mpo-cu	mpo nu/bm	mpo nu/fpr	mpo nu/fp
403	1	0	0,913	0,918	0,887	0,887	0,918	0,890	0,890	0,886
404	1	1	0,855	0,835	0,841	0,841	0,844	0,835	0,836	0,844
193	0	0	0,216	0,212	0,215	0,215	0,209	0,213	0,212	0,215
194	0	0	0,220	0,214	0,220	0,220	0,209	0,217	0,217	0,221
196	1	0	0,908	0,915	0,880	0,880	0,918	0,885	0,884	0,878
197	0	0	0,222	0,213	0,223	0,223	0,209	0,219	0,219	0,225
198	1	1	0,809	0,834	0,777	0,777	0,843	0,789	0,787	0,771
199	0	1	0,453	0,449	0,426	0,426	0,452	0,427	0,427	0,426
401	1	1	0,844	0,837	0,825	0,825	0,843	0,824	0,824	0,826
402	0	1	0,200	0,212	0,190	0,190	0,209	0,195	0,194	0,188
405	0	1	0,433	0,450	0,398	0,398	0,452	0,406	0,405	0,394
407	0	0	0,504	0,448	0,500	0,500	0,453	0,480	0,483	0,508
408	1	0	0,904	0,919	0,874	0,874	0,918	0,881	0,880	0,871
409	0	0	0,226	0,215	0,228	0,228	0,209	0,223	0,223	0,230

Note : ID = numéro d'identification des écoles; mplsbe = meilleur prédicteur linéaire sans biais empirique; mplsbe-cu = meilleur prédicteur linéaire sans biais empirique en contexte unitaire; mpo-cu = meilleure prédiction observée en contexte unitaire; mpo-fh = meilleure prédiction observée selon le modèle de Fay-Herriot; mpo nu/bm = meilleure prédiction observée au niveau de l'unité/basée sur un modèle; mpo nu/fh = meilleure prédiction observée au niveau de l'unité/Fay-Herriot; mpo nu/fpr = meilleure prédiction observée au niveau de l'unité/fondée sur le plan restreinte; PC = programme en classe; TV = télévision; TVSFP = Television School and Family Smoking Prevention and Cessation Project.

Tableau 4.2
Moyennes de petits domaines prédites à partir des données du TVSFP (partie 2)

ID	PC	TV	mplsbe	mplsbe-cu	mpo nu/fh	mpo-fh	mpo-cu	mpo nu/bm	mpo nu/fpr	mpo nu/fp
410	1	1	0,815	0,834	0,783	0,783	0,843	0,794	0,793	0,778
411	0	1	0,444	0,450	0,411	0,411	0,452	0,416	0,416	0,409
412	1	0	0,929	0,916	0,911	0,911	0,919	0,907	0,908	0,913
414	1	0	0,939	0,916	0,926	0,926	0,919	0,918	0,919	0,929
415	1	1	0,870	0,834	0,864	0,864	0,844	0,853	0,854	0,869
505	1	1	0,820	0,834	0,793	0,793	0,843	0,801	0,800	0,790
506	0	1	0,429	0,450	0,393	0,393	0,452	0,402	0,401	0,389
507	0	1	0,452	0,450	0,426	0,426	0,452	0,426	0,427	0,426
508	0	1	0,442	0,449	0,413	0,413	0,452	0,416	0,416	0,411
509	1	0	0,929	0,917	0,913	0,913	0,919	0,910	0,910	0,915
510	1	0	0,906	0,918	0,882	0,882	0,918	0,886	0,885	0,880
513	0	0	0,198	0,213	0,188	0,188	0,209	0,193	0,192	0,185
514	1	0	0,868	0,834	0,862	0,862	0,844	0,851	0,853	0,866
515	0	0	0,193	0,212	0,183	0,183	0,209	0,189	0,188	0,180

Note : ID = numéro d'identification des écoles; mplsbe = meilleur prédicteur linéaire sans biais empirique; mplsbe-cu = meilleur prédicteur linéaire sans biais empirique en contexte unitaire; mpo-cu = meilleure prédiction observée en contexte unitaire; mpo-fh = meilleure prédiction observée selon le modèle de Fay-Herriot; mpo nu/bm = meilleure prédiction observée au niveau de l'unité/basée sur un modèle; mpo nu/fh = meilleure prédiction observée au niveau de l'unité/Fay-Herriot; mpo nu/fpr = meilleure prédiction observée au niveau de l'unité/fondée sur le plan restreinte; PC = programme en classe; TV = télévision; TVSFP = Television School and Family Smoking Prevention and Cessation Project.

Tableau 4.3
Coefficients de régression estimés pour les données du TVSFP

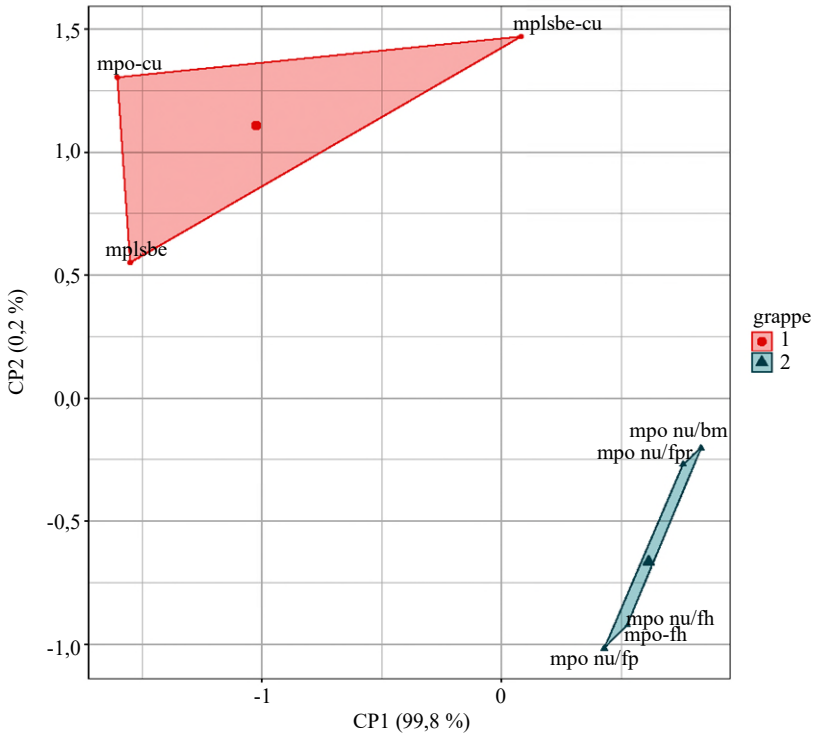
Méthodes	β_0	β_1	β_2	β_3
mplsbe	0,211	0,707	0,241	-0,319
mplsbe-cu	0,212	0,705	0,237	-0,319
mpo nu/fh	0,206	0,687	0,214	-0,289
mpo-fh	0,206	0,687	0,214	-0,289
mpo-cu	0,209	0,709	0,243	-0,318
mpo nu/bm	0,206	0,688	0,216	-0,291
mpo nu/fpr	0,206	0,689	0,216	-0,292
mpo nu/fp	0,206	0,687	0,213	-0,288

Note : mplsbe = meilleur prédicteur linéaire sans biais empirique; mplsbe-cu = meilleur prédicteur linéaire sans biais empirique en contexte unitaire; mpo-cu = meilleure prédiction observée en contexte unitaire; mpo-fh = meilleure prédiction observée selon le modèle de Fay-Herriot; mpo nu/bm = meilleure prédiction observée au niveau de l'unité/basée sur un modèle; mpo nu/fh = meilleure prédiction observée au niveau de l'unité/Fay-Herriot; mpo nu/fpr = meilleure prédiction observée au niveau de l'unité/fondée sur le plan restreinte; TVSFP = Television School and Family Smoking Prevention and Cessation Project.

Malgré cette similitude apparente, il existe des différences lorsqu'on l'examine de plus près. Veuillez noter que nous prédisons non pas une, mais 28 moyennes de petits domaines, que nous devons examiner conjointement. De plus, nous comparons 8 méthodes différentes pour voir si certaines méthodes sont plus similaires que d'autres lorsqu'il s'agit de prédire un vecteur à 28 dimensions de moyennes de petits domaines. Pour faciliter une telle comparaison, nous utilisons des techniques modernes d'analyse de données de grande dimension au moyen du regroupement des k-moyennes (k-moyennes) et l'analyse en composantes principales (ACP). Les k-moyennes regroupent les prédicteurs en 28 dimensions avec 2 grappes potentielles. Ici, le nombre de grappes est déterminé à l'aide de la méthode du coude (Thorndike, 1953), avec le pourcentage de variance expliquée comme métrique. Le résultat a classé les MPLSBE, MPLSBE-CU et MPO-CU dans une même classe, tandis que les MPO-NU/FH, MPO-FH, MPO-NU/BM, MPO-NU/FPR et MPO-NU/BM ont été regroupées dans l'autre.

Pour saisir les différences et les similarités les plus importantes parmi les prédicteurs en 28 dimensions, nous avons appliqué l'analyse en composantes principales (ACP). Ensuite, nous avons utilisé les deux premières composantes principales (CP), qui expliquent la majeure partie de la variation des prédicteurs en 28 dimensions, pour ajuster un autre regroupement de k-moyennes. Le résultat est présenté à la figure 4.1. On observe que la première CP expliquait environ 99,8 % de la variation, tandis que la seconde CP expliquait environ 0,2 % de la variation. Cela signifie que les deux plus importantes CP représentent très bien les prédicteurs en 28 dimensions. Les k-moyennes fondées sur les deux plus importantes CP ont donné le même résultat de catégorisation des résultats que ce qui précède. Pour la grappe contenant le MPLSBE, la variance à l'intérieur de la grappe est plus grande que celle de l'autre grappe, bien que la variance entre les grappes explique 84,3 % de la variance totale.

Figure 4.1 Catégorisation des résultats des k-moyennes des 2 principales composantes des prédicteurs de petits domaines en 28 dimensions



Note : CP = composante principale; mplsbe = meilleur prédicteur linéaire sans biais empirique; mplsbe-cu = meilleur prédicteur linéaire sans biais empirique en contexte unitaire; mpo-cu = meilleure prédiction observée en contexte unitaire; mpo-fh = meilleure prédiction observée selon le modèle de Fay-Herriot; mpo nu/bm = meilleure prédiction observée au niveau de l'unité/basée sur un modèle; mpo nu/fh = meilleure prédiction observée au niveau de l'unité/Fay-Herriot; mpo nu/fpr = meilleure prédiction observée au niveau de l'unité/fondée sur le plan restreinte.

On peut se demander si le résultat de l'ACP est interprétable. Nous pensons que la réponse dépend assez étroitement de la caractéristique de la structure de données particulière. Il convient de noter qu'ici, nous essayons de comparer différentes méthodes sur leurs valeurs prédites d'un vecteur en 28 dimensions. En général, nous essayons de comparer les valeurs de différents prédicteurs d'un vecteur de dimension m de moyennes de petits domaines. Lorsque m est grand, l'interprétation devient difficile. En effet, ces méthodes basées sur l'apprentissage automatique ont pour caractéristique de ne pas être entièrement interprétables. Cette caractéristique, selon laquelle tout n'est pas interprétable, est aujourd'hui progressivement acceptée par la communauté de la science des données, surtout lorsque la prédiction est l'intérêt principal, comme c'est le cas dans l'estimation sur petits domaines (EPD).

Cela dit, en tant que statisticiens, nous préférons toujours interpréter ce qui est observé. À cette occasion, nous pouvons voir que toutes les méthodes de MPO, sauf une, ont été regroupées dans une grappe et que toutes les méthodes de MPLSBE, plus la MPO-CU, ont été regroupées dans l'autre. Ainsi, dans un certain sens, les résultats laissent supposer que les méthodes de MPO sont plus proches les unes des autres et que les méthodes de MPLSBE sont plus proches les unes des autres, ce qui n'est pas étonnant. L'ACP a

également regroupé les méthodes fondées sur le CU en une seule grappe, ce qui semble également raisonnable.

En résumé, l'analyse des k-moyennes et l'ACP ont révélé des renseignements supplémentaires sur les différents prédicteurs de petits domaines dans le cadre du modèle de REE. Alors que les études de simulation de la section 3 se concentraient sur les EQMP globales ou spécifiques à chaque domaine, les analyses de grande dimension ont examiné le prédicteur multidimensionnel pour petits domaines dans son ensemble et ont exposé les différences ou les similitudes entre les différentes méthodes.

Malgré les différences potentielles dans les caractéristiques des petits domaines, les programmes en classe et de télévision semblent améliorer efficacement les cotes THKS des participants. Toutefois, la question de savoir si cela se traduit par une amélioration de la prévention du tabagisme et de l'arrêt de la consommation de tabac n'est pas claire dans la pratique. Le programme en classe était relativement plus efficace que l'émission de télévision. Sans aucune intervention, les cotes THKS n'ont montré aucune amélioration dans les différents établissements scolaires.

Code logiciel. Accessible à l'adresse https://github.com/Celaeno1017/obpner.

Remerciements

Les auteurs sont reconnaissants envers un rédacteur associé et deux examinateurs dont les commentaires éclairés ont contribué à améliorer le travail et la présentation. Les recherches de Jiming Jiang sont en partie soutenues par la subvention DMS-2210569 de la National Science Foundation des États-Unis.

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. *J. Amer. Statist. Assoc.*, 80, 28-36.

Chen, Y., Lahiri, P. et Salvati, N. (2025). Effets d'erreurs de spécification de modèle sur les estimateurs sur petits domaines. *Techniques d'enquête*, 51(2), 707-719. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2025002/article/00001-fra.pdf.

Fay, R.E. et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.*, 74, 269-277.

Fuller, W.A. (2009). *Sampling Statistics*, NJ, Hoboken: John Wiley & Sons, Inc.

Hedeker, D., Gibbons, R.D. et Flay, B.R. (1994). Random-effects regression models for clustered data with an example from smoking prevention research. *J. Consulting Clinical Psych.*, 62, 757-765.

Jiang, J. et Nguyen, T. (2021). *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer.

Jiang, J., Nguyen, T. et Rao, J.S. (2011). Best predictive small area estimation. *J. Amer. Statist. Assoc.*, 106, 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. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2015001/article/14200-fra.pdf.

Jiang, J., Rao, J.S., Fan, J. et Nguyen, T. (2018). Classified mixed model prediction. *J. Amer. Statist. Assoc.*, 113, 269-279.

Lohr, S. (2021). *Sampling: Design and Analysis*. Chapman & Hall/CRC.

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.

Thorndike, R.L. (1953). Who belongs in the family? *Psychometrika*, 18, 267-276.