Nº 11-522-X au catalogue
ISSN 1709-822X

Recueil du Symposium de 2024 de Statistique Canada : Le futur des statistiques officielles

Exploration d'approches pour l'estimation sur petits domaines en présence d'erreurs de mesure et leur application aux enquêtes-ménages en Indonésie

par Ika Yuni Wulansari, Stephen Woodcock et James Brown

Date de diffusion : le 8 septembre 2025

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2024 de Statistique Canada
Le futur des statistiques officielles

Exploration d'approches pour l'estimation sur petits domaines en présence d'erreurs de mesure et leur application aux enquêtes-ménages en Indonésie

Ika Yuni Wulansari, Stephen Woodcock et James Brown[1]

Résumé

Les objectifs de développement durable des Nations Unies exigent des données détaillées et désagrégées, habituellement obtenues au moyen d'enquêtes-ménages. Toutefois, les enquêtes à elles seules ne peuvent pas répondre à ce besoin de statistiques détaillées. Pour remédier à la situation, les instituts statistiques nationaux adoptent des méthodes sur petits domaines; ils font toutefois face à des défis, car les variables auxiliaires, souvent dérivées d'enquêtes, introduisent des erreurs de mesure dans les modèles. L'objectif est d'appliquer une correction des erreurs de mesure dans le modèle classique au niveau du domaine de Fay-Herriot. Les résultats démontrent la robustesse de l'approche normalisée à faire abstraction des erreurs de mesure, mais montrent que dans des cas particuliers, la correction des erreurs de mesure est bénéfique. Cette approche est appliquée à une étude de cas fondée sur des données d'enquête-ménages en Indonésie.

Mots clés : Estimation sur petits domaines; statistiques désagrégées; erreur de mesure; meilleur prédicteur linéaire sans biais empirique (MPLSBE); données sur les ménages en Indonésie.

1. Introduction et analyse documentaire

1.1 Estimation sur petits domaines

L'élaboration de l'estimation sur petits domaines (EPD) fondée sur un modèle a été très importante ces dernières années. Le modèle le plus populaire utilisé en EPD est le modèle à effet aléatoire, qui comprend un effet de domaine aléatoire pour expliquer les variations entre les domaines au-delà de celles décrites par les variables auxiliaires (FAY III et Herriot, 1979; Rao et Molina, 2015). Selon la disponibilité des variables auxiliaires, il existe deux modèles de base d'EPD : au niveau du domaine et au niveau de l'unité (Rao et Molina, 2015). Le modèle au niveau du domaine est appliqué lorsque des données auxiliaires sont disponibles à un niveau agrégé, comme un district ou une ville. Le modèle classique d'effet aléatoire au niveau du domaine dans le cadre de l'EPD est le modèle de meilleur prédicteur linéaire sans biais empirique (MPLSBE) de Fay-Herriot (FH), qui est une approche paramétrique. L'autre modèle, au niveau de l'unité, est utilisé lorsque des données auxiliaires sont accessibles pour les personnes ou les ménages. Dans un modèle d'EPD standard, les variables auxiliaires doivent être tirées d'un recensement ou d'autres sources qui ne contiennent pas d'erreurs de mesure.

1.2 Estimation sur petits domaines avec erreur de mesure dans les covariables

Le succès de la modélisation de l'EPD dépend considérablement de la disponibilité de données auxiliaires fortement corrélées avec les variables à estimer (Rao et Molina, 2015). Toutefois, dans de nombreuses situations, des renseignements ou covariables auxiliaires ne sont pas disponibles dans les sources de données présentant une couverture complète de la population. Il est très probable que d'autres enquêtes puissent fournir des estimations de covariables utiles, mais avec des erreurs de mesure. Ybarra et Lohr (2008) ont élargi le modèle de FH pour tenir

[1] Ika Yuni Wulansari, University of Technology Sydney, 15 Broadway, Ultimo NSW, Australie, 2007 (ikayuni.wulansari@uts.edu.au), Politeknik Statistika STIS, Indonésie, 13330 (ikayuni@stis.ac.id); Stephen Woodcock, University of Technology Sydney, 15 Broadway, Ultimo NSW, Australie, 2007 (stephen.woodcock@uts.edu.au); James Brown, University of Technology Sydney, 15 Broadway, Ultimo NSW, Australie, 2007 (james.brown@uts.edu.au).

compte de la variabilité d'échantillonnage dans les covariables lors de la production d'estimations sur petits domaines. En connaissant les estimations des variances d'erreur de mesure pour les covariables pour chaque petit domaine, l'incertitude accrue peut être prise en compte tout en obtenant de la précision grâce aux renseignements auxiliaires.

2. Méthodologie

2.1 Meilleur prédicteur linéaire sans biais empirique

Dans la présente étude, nous nous concentrons sur l'approche au niveau du domaine. Ce modèle relie des variables auxiliaires $\boldsymbol{X_i^T} = (X_{1i}, X_{2i}, \dots, X_{pi})$ à des estimateurs directs dans chaque domaine. Le véritable paramètre est θ_i, avec un modèle linéaire comme suit :

$$\theta_i = \boldsymbol{X_i^T \beta} u_i, \qquad i = 1,2, \dots, m \tag{1}$$

où $\boldsymbol{\beta}$ $(\beta_0, \dots, \beta_p)^T$ est le $(p+1) \times 1$ vecteur des coefficients de régression; u_i est l'effet aléatoire sur petit domaine, présumé $u_i \sim iid\ N(0, \sigma_u^2)$, et m est le nombre de petits domaines. Le véritable paramètre θ_i n'est pas observé, mais à partir d'un échantillon aléatoire, nous disposons d'une estimation directe $\widehat{\theta_i}$; le modèle d'échantillonnage peut être formulé comme suit :

$$\widehat{\theta_i} = \theta_i + e_i, \qquad i = 1,2, \dots, m \tag{2}$$

où e_i est l'erreur d'échantillonnage, présumée $e_i \sim iid\ N(0, \psi_i)$. La combinaison de (1) et (2) produit le modèle mixte linéaire qui peut être formulé comme suit :

$$\widehat{\theta_i} = \boldsymbol{X_i^T \beta} u_i + e_i, \qquad i = 1,2, \dots, m \tag{3}$$

Le modèle mixte linéaire de l'équation (3) est appelé le modèle Fay-Herriot (FH) (Fay et Herriot, 1979) et nous pouvons construire l'estimateur FH standard

$$\tilde{\theta}_i^{MPLSB} = \boldsymbol{X_i^T \tilde{\beta}} + \gamma_i(\hat{\theta}_i - \boldsymbol{X_i^T \tilde{\beta}}) = \gamma_i \hat{\theta}_i + (1 - \gamma_i)\boldsymbol{X_i^T \tilde{\beta}}$$

où $\gamma_i = {\sigma_u}^2/({\sigma_u}^2 + \psi_i)$. Cela dépend de la valeur inconnue de ${\sigma_u}^2$, mais en suivant l'approche normalisée d'estimation des paramètres inconnus $\boldsymbol{\beta}$ et ${\sigma_u}^2$ (Rao et Molina, 2015), nous obtenons le meilleur prédicateur linéaire sans biais empirique (MPLSBE) $\hat{\theta}_i^{FH} = \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i)\boldsymbol{X_i^T \hat{\beta}}$, où $\hat{\gamma}_i = \frac{\hat{\sigma}_u{}^2}{\hat{\sigma}_u{}^2 + \psi_i}$. L'estimateur sans biais de l' $EQM\left(\hat{\theta}_i^{FH}\right)$ s'ensuit selon Prasad et Rao (1990).

2.2 Meilleur prédicteur linéaire sans biais empirique avec erreur de mesure dans les covariables

L'estimateur de MPLSBE utilisé dans l'analyse de l'EPD suppose que les variables auxiliaires ne contiennent pas de composante d'erreur. Ybarra et Lohr (2008) élargissent le MPLSBE dans le cadre du modèle FH pour tenir compte des erreurs de mesure dans la variable auxiliaire, ce qui permet d'utiliser des données d'enquête indépendantes de l'enquête d'intérêt comme variables auxiliaires; ce que nous appelons une EPD avec erreur de mesu (EM). Ybarra et Lohr (2008) ont utilisé les données d'enquête, $\boldsymbol{\widehat{X}_i}$, comme estimateur des données sur la population, $\boldsymbol{X_i}$; ce qui, si simplement inséré dans (3), sous-estime l'erreur quadratique moyenne (EQM). Ils élargissent (3) pour $\boldsymbol{X_i}$ avec erreur de mesure, pour obtenir

$$\widehat{\theta_i} = \boldsymbol{\widehat{X}_i^T \beta} + r_i(\boldsymbol{\widehat{X}_i}, \boldsymbol{X_i}) + e_i \tag{4}$$

où $r_i(\boldsymbol{\widehat{X}_i}, \boldsymbol{X_i}) = u_i + (\boldsymbol{X_i} - \boldsymbol{\widehat{X}_i})^T \boldsymbol{\beta}$. Lorsque les paramètres de régression et de covariance sont connus, $C_i = EQM(\boldsymbol{\widehat{X}_i})$ est donné par $E[(\boldsymbol{\widehat{X}_i} - \boldsymbol{X_i})(\boldsymbol{\widehat{X}_i} - \boldsymbol{X_i})^T]$ et $EQM(r_i) = \sigma_u^2 + \boldsymbol{\beta}^T C_i \boldsymbol{\beta}$. Par conséquent, l'équation FH du MPLSB avec erreur de mesure est la suivante :

$$\tilde{\theta}_i^{EM} = \gamma_i \hat{\theta}_i + (1 - \gamma_i)\boldsymbol{\widehat{X}_i^T \beta} \tag{5}$$

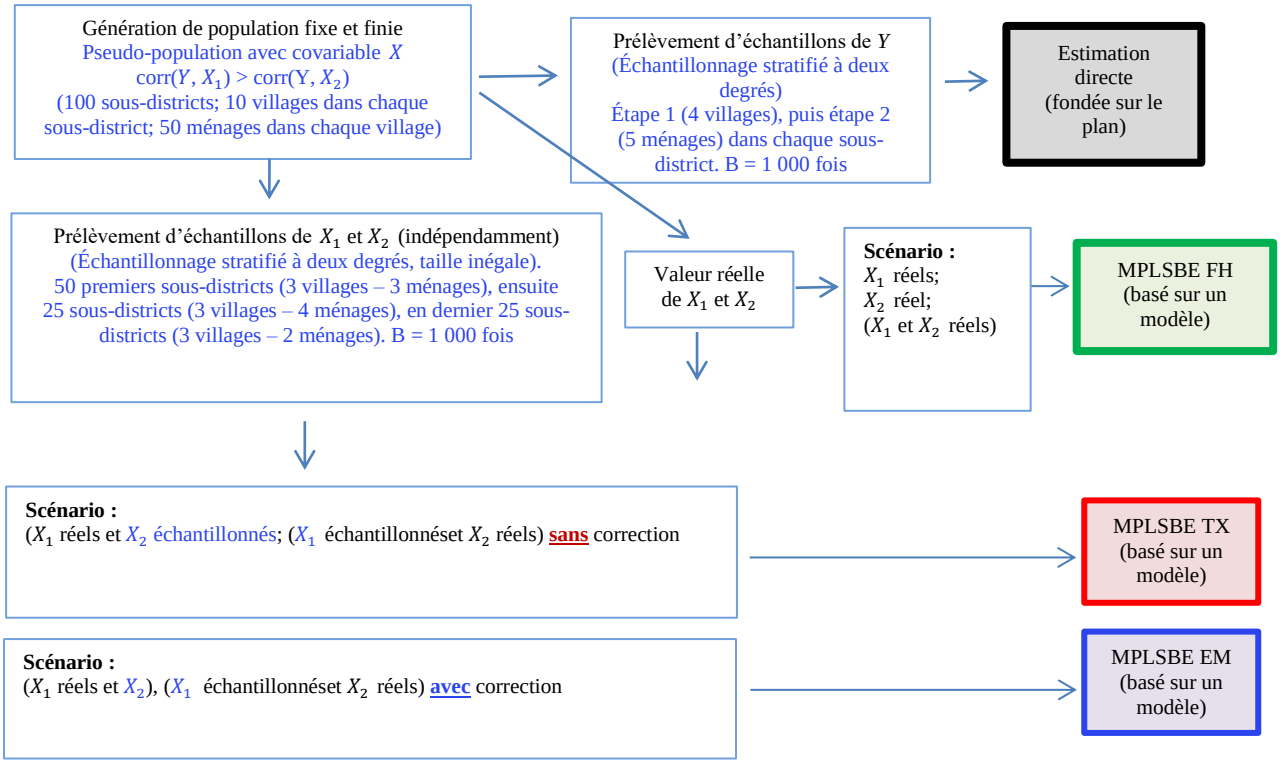
avec $\gamma_i = \frac{EQM(r_i)}{EQM(r_i)+\psi_i}$ et $EQM\left(\tilde{\theta}_i^{EM}\right) = \gamma_i\psi_i$. Comme ci-dessus, $\boldsymbol{\beta}$ et σ_u^2 doivent être estimés à partir des données en présumant que C_i et ψ_i sont connus. Ybarra et Lohr (2008) ont fourni tous les détails de l'estimation et une fois $\boldsymbol{\beta}$ et σ_u^2 remplacés par leur estimateur convergent $\widehat{\boldsymbol{\beta}}$ et $\widehat{\sigma}_u^2$, l'estimateur FH de MPLSBE avec EM est : $\hat{\theta}_i^{EM} = \hat{\gamma}_i\hat{\theta}_i + (1 - \hat{\gamma}_i)\widehat{\boldsymbol{X}}_i^T\boldsymbol{\beta}$, où $\hat{\gamma}_i = \frac{\widehat{\sigma}_u^2+\widehat{\boldsymbol{\beta}}^T C_i\widehat{\boldsymbol{\beta}}}{\widehat{\sigma}_u^2+\widehat{\boldsymbol{\beta}}^T C_i\widehat{\boldsymbol{\beta}}+\psi_i}$. Ybarra et Lohr (2008) ont utilisé la méthode Jackknife proposée par Jiang et coll. (2002) pour estimer $EQM\left(\hat{\theta}_i^{EM}\right) = M_{1i} + M_{2i}$, où $M_{1i} = \gamma_i\psi_i$ est l'EQM de (5) et $M_{2i} = E(\hat{\theta}_i^{EM} - \tilde{\theta}_i^{EM})^2$ est le terme supplémentaire dû à l'estimation des paramètres inconnus.

3. Étude par simulation

3.1 Procédure de simulation

La figure 3.1-1 décrit notre cadre de simulation pour évaluer le rendement de nos méthodes d'estimation dans le cadre de différents scénarios d'échantillonnage et de correction d'erreurs. Nous avons généré une pseudo-population finie fixe, structurée avec 100 sous-districts, 10 villages par sous-district et 50 ménages dans chaque village. Cette population comprenait des covariables bivariées (X_1 et X_2); Y (la variable de réponse) ayant une corrélation plus forte avec X_1 qu'avec X_2. Les échantillons de Y sont tirés à l'aide d'un échantillonnage stratifié à deux degrés : quatre villages sont sélectionnés dans chaque sous-district à la première étape, suivis de cinq ménages de chaque village à la deuxième étape; ce qui est répété 1 000 fois. Les échantillons de X_1 et X_2 sont sélectionnés indépendamment à l'aide d'un échantillonnage stratifié à deux degrés de tailles inégales entre les sous-districts; ce qui est également répété 1 000 fois.

Figure 3.1-1
Cadre de simulation



Ce cadre explore des scénarios où les valeurs réelles de X_1 et X_2 sont connues et les contraste avec des cas comportant des valeurs échantillonnées, avec ou sans correction d'erreurs. Ces scénarios alimentent différentes méthodes d'estimation : l'estimation directe (fondée sur le plan) repose uniquement sur des échantillons Y, tandis que les estimateurs fondés sur un modèle s'appuient sur différentes sources de renseignements auxiliaires : i) MPLSBE FH utilise les valeurs de covariables réelles comme information auxiliaire, ii) MPLSBE TX évalue l'effet de l'utilisation des valeurs échantillonnées X_1 et réelles X_2, et vice versa, sans corrections, et iii) MPLSBE EM intègre des corrections des erreurs de mesure dans l'information auxiliaire.

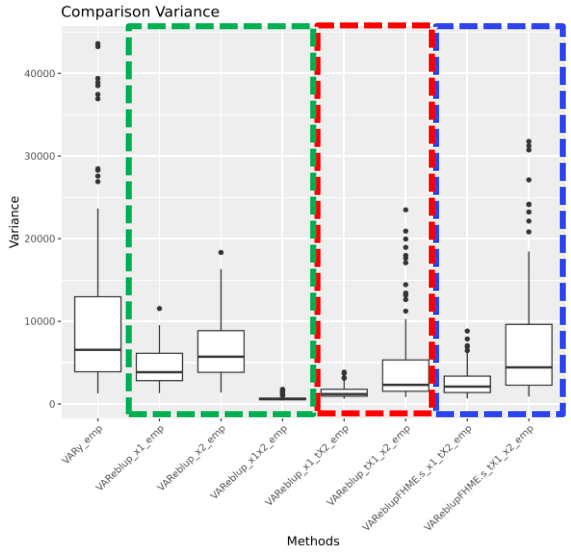
Dans le cadre de leur évaluation, Ybarra et Lohr (2008) ont généré leurs données d'échantillonnage à l'aide d'une approche fondée sur un modèle, tandis que notre approche utilise une génération fondée sur le plan pour imiter les plans d'échantillonnage retrouvés dans les données réelles. Ybarra et Lohr (2008) ont appliqué $k\%$ de C_i (EQM de X_i) choisi aléatoirement pour équivaloir à 3; le reste des C_i étant établi à 0 (comme s'il n'y avait pas d'erreurs) pour $k \in \{0,20,50,80,100\}$, pour simuler divers niveaux d'erreurs de mesure dans de petits domaines. Toutefois, notre approche estimant les covariables à partir d'un échantillon, les erreurs d'échantillonnage induisent intrinsèquement des erreurs de mesure (EM) dans tous les domaines, reflétant ainsi mieux des scénarios concrets. La stratification au moyen de répartitions d'échantillons diverses fait en sorte que les variances des erreurs de mesure réelles C_i varient encore dans les petits domaines. De plus, Ybarra et Lohr (2008) se sont concentrés sur la variable univariée X, tandis que notre travail intègre des covariables bivariées (X_1 et X_2) pour étudier l'incidence des différentes structures de covariables sur la réduction de l'erreur et de la variance.

Ybarra et Lohr (2008) ont simulé des cas univariés dans quatre conditions : sans erreur, sans covariable (modèle avec l'ordonnée à l'origine), avec erreur sans correction, et avec erreur et correction. En revanche, nous simulons des scénarios dans des conditions univariées X sans erreur, de covariables bivariées sans erreur (modèle réel), de covariables bivariées avec erreur sans correction et de covariables bivariées avec erreur et correction. En incluant les covariables bivariées et en mettant l'accent sur la génération de pseudo-population fondée sur le plan, notre approche fournit un cadre complet et réaliste pour évaluer l'incidence des structures de covariables et des plans d'échantillonnage sur l'estimation sur petits domaines.

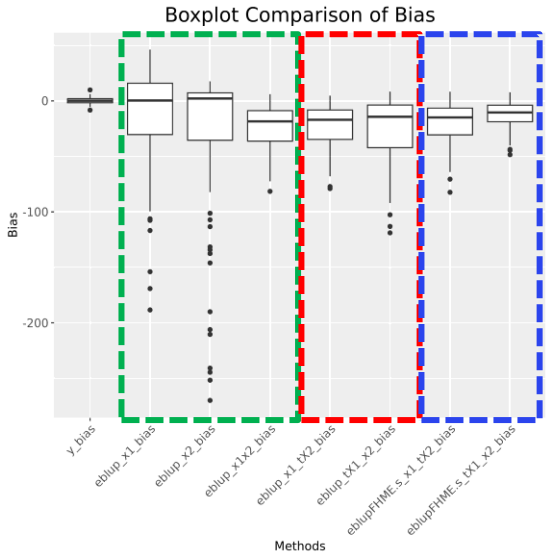
3.2 Résultat de la simulation

La figure 3.2-1 présente une comparaison de la variance et du biais entre différentes approches d'estimation, en démontrant les compromis entre la variabilité et l'exactitude.

Figure 3.2-1 Comparaison de la variance (a) et du biais (b)



a. Comparaison de la variance

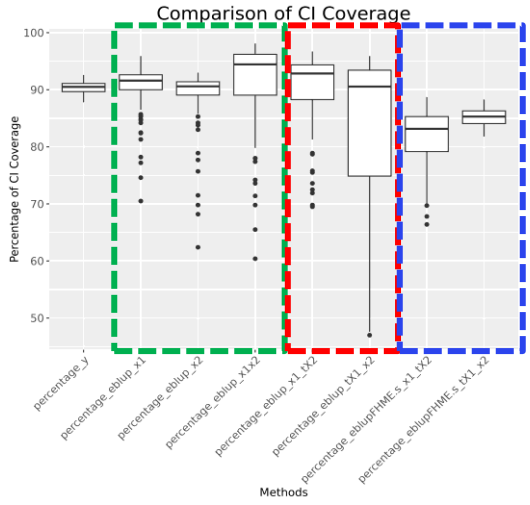


b. Comparaison du biais

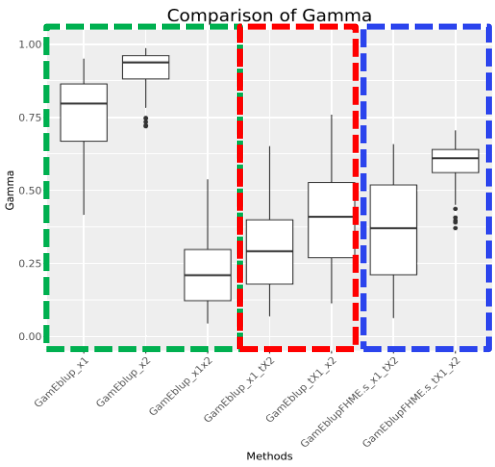
Note :
- Univarié X_1 réels – Univarié X_2 réels – Bivarié (X_1 réels, X_2 réels), MODÈLE RÉEL
- Bivarié (X_1 réels, échantillon X_2) – Bivarié (échantillon X_1, X_2 réels), SANS correction de l'EM
- Bivarié (X_1 réels, $échantillon\ X_2$) – Bivarié ($échantillon\ X_1$, X_2 réels), AVEC correction de l'EM

Comme on pouvait s'y attendre, à la figure 3.2-1, la méthode du MPLSBE avec les X bivariés sans erreur (modèle réel) donne les meilleurs résultats en matière de réduction de la variance entre les sous-districts, comparativement aux estimations directes. Le rendement de la méthode MPLSBE des X bivariés avec erreur (sans correction de l'EM) est meilleur que celui avec les X_1 univariés ou X_2 univariés. Le rendement de la méthode MPLSBE des X bivariés avec erreur (avec correction de l'EM) est également meilleur qu'avec les X_1 univariés ou X_2 univariés, mais il n'est pas meilleur que de simplement ignorer l'erreur de mesure. Toutefois, les tracés du biais à la figure 3.2-1 démontrent que la correction de l'EM offre une certaine protection réduisant la répartition du biais dans les domaines.

Figure 3.2-2
Comparaison de la couverture de l'intervalle de confiance (IC) (a) et gamma (b)



a. Comparaison de la couverture des intervalles de confiance (IC)



b. Comparaison gamma

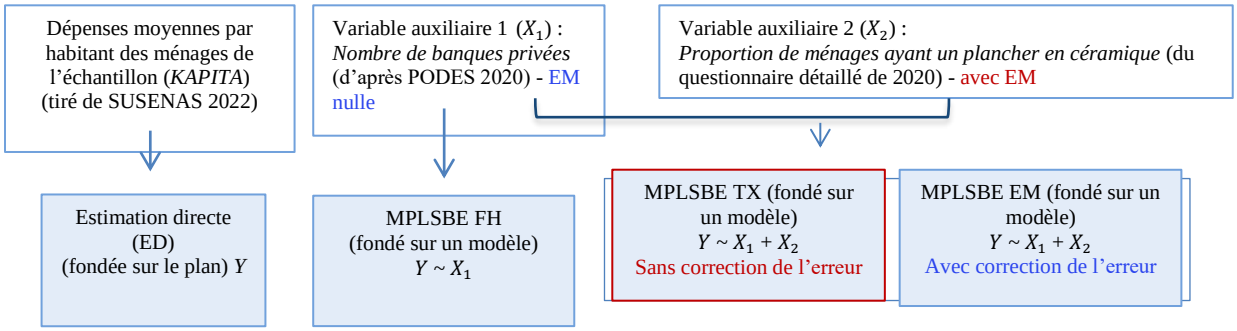
D'après la figure 3.2-2, à mesure que le paramètre de rétrécissement γ des variables bivariées X s'approche de zéro, cela rapproche le MPLSBE du modèle. Toutefois, pour l'EM, le paramètre de rétrécissement γ se rapproche de 1; ce qui reflète l'augmentation de la variance découlant de l'erreur de mesure. Cela entraîne l'amélioration du biais, car les estimations prennent plus d'importance dans l'estimation directe. En ce qui concerne la couverture à 95 % de l'intervalle de confiance (IC), la méthode de MPLSBE TX (sans correction de l'EM) donne un résultat préoccupant, reflétant que l'estimateur standard de l'EQM ne corrige pas l'incertitude accrue attribuable à une erreur de mesure. La méthode de MPLSBE EM donne une meilleure couverture estimée de l'IC que celle de la méthode MPLSBE TX.

4. Étude empirique

4.1 Procédure empirique

Nous avons appliqué l'approche du MPLSBE avec correction des erreurs de mesure aux données d'enquêtes-ménages réalisées en Indonésie. La figure 4.1-1 illustre le déroulement des opérations d'une étude empirique. Nous avons estimé la variable de réponse, Y, soit la moyenne de l'échantillon des dépenses des ménages par habitant (KAPITA), comme une estimation directe fondée sur le plan. Dans le cadre des approches fondées sur un modèle, nous avons estimé le MPLSBE FH en utilisant seulement X_1; ce qui a supposé une erreur de mesure (EM) nulle. Nous avons également estimé le modèle MPLSBE TX à l'aide de X_1 réels et de X_2 échantillonnés, mais sans corriger l'erreur de mesure, et le modèle MPLSBE EM à l'aide de X_1 réels et de X_2 échantillonnés, avec correction de l'erreur de mesure.

Figure 4.1-1
Cadre empirique



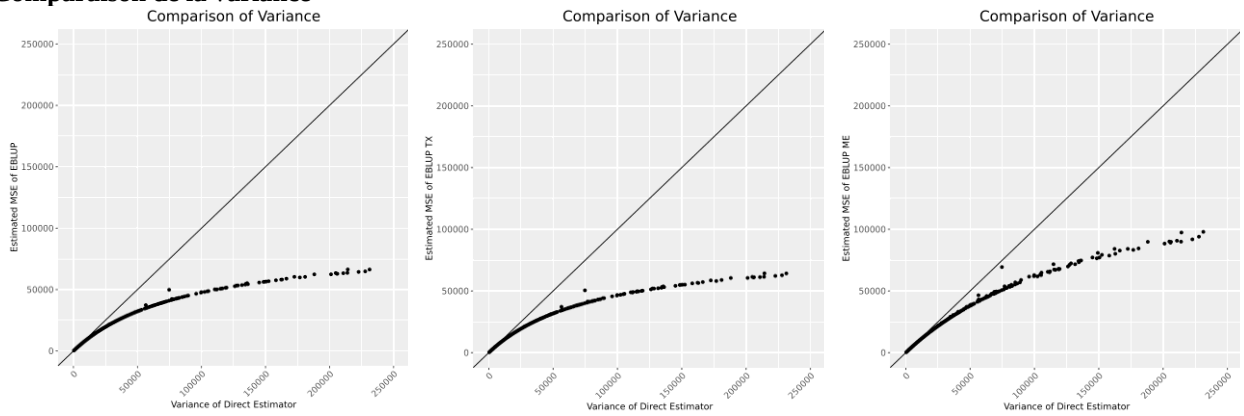
Le processus d'estimation directe a utilisé l'Enquête socioéconomique nationale (SUSENAS) de 2022, qui comprenait 514 districts et 34 468 unités d'échantillonnage primaires (UEP). À partir de cet ensemble de données, un sous-échantillon de 3 447 UEP a été extrait pour reproduire la variance au niveau du sous-district. La variable cible était la moyenne des dépenses des ménages par habitant, appelée KAPITA, exprimée en milliers de rupiahs.
L'information auxiliaire provenait de deux ensembles de données. Le premier portait sur les données sur le potentiel des villages (PODES) de 2020, couvrant 514 districts et fondées sur des dossiers administratifs; ce qui ne supposait aucune erreur de mesure. La variable pertinente était le nombre de banques privées, qui avait une corrélation avec KAPITA. La deuxième source était l'enquête par questionnaire détaillé de 2020, dont la couverture était plus vaste que celle de SUSENAS, et comprenait 514 districts et 268 125 UEP; ce qui a fourni des renseignements auxiliaires avec des erreurs de mesure. D'après cette enquête, la proportion de ménages ayant des types de sols en céramique, qui avait une corrélation avec KAPITA, était la covariable de l'erreur de mesure (EM) dans le modèle. Comme cette procédure reposait sur un sous-échantillonnage, nous avons veillé à ce que la substance de l'échantillon initial ne soit pas modifiée. Nous avons comparé le sous-échantillon à l'échantillon original et vérifié que les caractéristiques essentielles et la structure de variance de l'échantillon original avaient été conservées.

4.2 Résultats empiriques

La figure 4.2-1 présente la comparaison entre les variances des estimateurs directs et celles des estimateurs fondés sur un modèle dans différents scénarios, soulignant l'efficacité des approches fondées sur un modèle pour réduire la

variance ou l'erreur quadratique moyenne (EQM). Les résultats démontrent que l'intégration de la variable supplémentaire X_2 avec correction de l'erreur de mesure (EM), appelée MPLSBE EM, donne une réduction légèrement moindre de l'EQM avec l'estimateur direct que la réduction obtenue en ignorant l'erreur. Ce résultat est conforme à nos conclusions de simulation, soulignant le compromis entre la prise en compte des erreurs dans les variables auxiliaires et les gains d'efficacité qui en découlent.

Figure 4.2-1
Comparaison de la variance



a. ED vs MPLSBE FH b. ED vs MPLSBE TX c. ED vs MPLSBE EM

5. Conclusion

La méthode MPLSBE avec X bivariés et sans erreur est la plus efficace pour ce qui est de la réduction de la variance. La méthode MPLSBE avec X bivariés avec erreur (sans correction de l'EM / MPLSBE TX) enregistre de meilleurs résultats que les X_1 univariés et X_2 univariés en ce qui concerne la réduction de la variance et de l'EQM. La méthode MPLSBE avec X bivariés avec erreur (avec correction de l'EM / MPLSBE EM) a donné également de meilleurs résultats que les X_1 univariés et X_2 univariés. Malheureusement, elle ne donne (habituellement) pas de meilleurs résultats que de simplement ignorer l'erreur, en ce qui concerne la réduction de la variance et de l'EQM. La méthode MPLSBE EM donne un gamma plus élevé, donc pèse davantage sur l'estimation directe (pour protéger contre le biais), mais entraîne une moindre réduction de la variance. La méthode MPLSBE EM fournit en outre une meilleure couverture d'IC que celle du MPLSBE sans correction de l'EM. L'utilisation de covariables avec correction d'erreurs de mesure peut quand même produire une réduction du biais appréciable. En fin de compte, l'EPD fournit des prédictions. On se concentre moins sur les valeurs β réelles, de sorte que le biais d'erreur de mesure traditionnel semble moins important pour les estimations finales.

Remerciements

Les auteurs tiennent à exprimer leur gratitude envers la University of Technology Sydney, pour le soutien financier ayant permis leur participation au Symposium international sur les questions de méthodologie de 2024, organisé par Statistique Canada. Ils souhaitent également souligner le soutien obtenu du ministère des Finances de l'Indonésie, dans le cadre du Indonesia Endowment Fund for Education, sous forme de bourse appuyant ce projet de recherche.

Bibliographie

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

Jiang, J., P. Lahiri et S.-M. Wan. (2002), « A Unified Jackknife Theory for Empirical Best Prediction with M-Estimation », *Annals of Statistics*, 30, p. 1782-1810.

Prasad, N.G.N. et J.N.K. Rao. (1990), « The Estimation of the Mean Squared Error of Small-Area Estimators », *Journal of the American Statistical Association*, 85, p. 163-171.

Rao, J. et I.Molina. (2015), *Small area estimation* (deuxième édition, éd.) Hoboken, New Jersey : Wiley.

Ybarra, L.M. et S.L. Lohr. (2008), « Small area estimation when auxiliary information is measured with error », *Biometrika*, 95(4), p. 919-931.