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

Recueil du Symposium de 2022 de Statistique Canada : Désagrégation des données : dresser un portrait de données plus représentatif de la société

Application de méthodes de lissage de la variance due à l'échantillonnage aux fins d'estimation sur petits domaines

par Yong You et Mike Hidiroglou

Date de diffusion : le 25 mars 2024

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2022 de Statistique Canada
Désagrégation des données : Dresser un portrait de données plus représentatif de la société

Application de méthodes de lissage de la variance due à l'échantillonnage aux fins d'estimation sur petits domaines

Yong You et Mike Hidiroglou[1]

Résumé

Le lissage de la variance due à l'échantillonnage est un sujet important dans l'estimation sur petits domaines. Dans le présent article, nous proposons des méthodes de lissage de la variance due à l'échantillonnage aux fins d'estimation sur petits domaines. En particulier, nous considérons la fonction de variance généralisée et les méthodes d'effet de plan aux fins de lissage de la variance due à l'échantillonnage. Nous évaluons et comparons les variances dues à l'échantillonnage lissées et les estimations sur petits domaines fondées sur des estimations de la variance lissées au moyen de l'analyse de données d'enquête de Statistique Canada. Les résultats de l'analyse de données réelles indiquent que les méthodes de lissage de la variance due à l'échantillonnage proposées fonctionnent très bien pour l'estimation sur petits domaines.

Mots clés : coefficient de variation; effet de plan; fonction de variance généralisée; modèle log-linéaire; erreur relative.

1. Introduction

L'estimation sur petits domaines est devenue très populaire et importante dans les organismes publics et privés en raison de la demande croissante d'estimations fiables. L'estimation sur petits domaines est fondée sur des modèles qui donnent des estimations fiables pour de petits domaines d'intérêt. Dans notre article, nous nous concentrons sur les modèles au niveau du domaine qui sont fondés sur des estimations d'enquête directes agrégées à partir des données au niveau de l'unité et des variables auxiliaires au niveau du domaine. Divers modèles au niveau du domaine ont été proposés dans la littérature en vue d'accroître la précision des estimations directes d'enquête : Rao et Molina (2015) proposent une synthèse intéressante de ces méthodes. Le modèle de Fay-Herriot (Fay et Herriot, 1979) est un modèle au niveau du domaine de base largement utilisé dans la pratique. Il comporte deux composantes : un modèle d'échantillonnage pour les estimations directes d'enquête et un modèle de couplage pour les paramètres de petit domaine choisis. Avec le modèle d'échantillonnage, nous posons qu'il existe un estimateur d'enquête direct y_i, qui est habituellement sans biais par rapport au plan, pour le paramètre de petit domaine θ_i, de sorte que

$$y_i = \theta_i + e_i,\ i = 1,\ldots,m, \tag{1}$$

où e_i est l'erreur d'échantillonnage associée à l'estimateur direct y_i et m est le nombre de petits domaines. En pratique, il est courant de supposer que les e_i sont des variables aléatoires normales indépendantes avec une moyenne $E(e_i) = 0$ et une variance due à l'échantillonnage $Var(e_i) = \sigma_i^2$. Avec le modèle de couplage, nous posons que le paramètre de petit domaine d'intérêt θ_i est lié aux variables auxiliaires au niveau du domaine $x_i = (x_{i1}, \ldots, x_{ip})'$ par un modèle de régression linéaire

$$\theta_i = x_i'\beta + v_i,\ \ i = 1,\ldots,m, \tag{2}$$

où $\beta = (\beta_1, \ldots, \beta_p)'$ est un vecteur $p \times 1$ de coefficients de régression et où les v_i sont des effets aléatoires propres au domaine que nous supposons indépendants et identiquement distribués avec $E(v_i) = 0$ et $Var(v_i) = \sigma_v^2$. L'hypothèse

[1]Yong You, Centre de collaboration internationale et d'innovation en méthodologie (CCIIM), Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), K1A 0T6. Adresse courriel de contact : yong.you@statcan.gc.ca. **Avertissement** : Cet article expose les opinions des auteurs qui ne sont pas nécessairement celles de Statistique Canada. Il décrit des méthodes théoriques qui pourraient ne pas correspondre à celles qu'emploie actuellement l'organisme.

de normalité pour v_i est généralement incluse aussi. La variance de modèle σ_v^2 est inconnue et doit être estimée à partir des données. Pour le modèle de Fay-Herriot, on suppose que la variance due à l'échantillonnage σ_i^2 est connue dans le modèle (1). Comme il s'agit d'une hypothèse très forte, une approche de lissage ou de modélisation est habituellement utilisée pour estimer σ_i^2. La variance due à l'échantillonnage peut être lissée ou modélisée directement comme dans Wang et Fuller (2003), You et Chapman (2006), Sugasawa, Tamae et Kubokawa (2017), etc. You (2021) montre que la méthode de lissage peut donner des estimations fondées sur un modèle plus efficaces et plus exactes que la méthode de modélisation pour petits domaines dans un cadre hiérarchique bayésien. Lesage, Beaumont et Bocci (2021) discutent également du lissage de la variance due à l'échantillonnage pour le modèle de Fay-Herriot.

L'objectif de cet article est de comparer différentes méthodes pour lisser les estimations directes des variances dues à l'échantillonnage pour les proportions dans une estimation sur petits domaines au moyen du modèle de Fay-Herriot. À cette fin, nous procédons de la façon suivante. Soit $\hat{p}_{iw}$ l'estimateur direct fondé sur le plan pour la proportion p_i pour une caractéristique donnée dans le i^{e} domaine. En appliquant le modèle de Fay-Herriot à $\hat{p}_{iw}$, nous obtenons

$$\hat{p}_{iw} = p_i + e_i, \tag{3}$$

la variance due à l'échantillonnage $Var(e_i) = \sigma_i^2$ étant inconnue. Soit $\hat{V}_i$ l'estimation directe de la variance due à l'échantillonnage pour σ_i^2 obtenue à partir des données d'enquête. Habituellement, certaines variables $\hat{V}_i$ sont très instables en raison des petites tailles d'échantillon. Par conséquent, nous devons obtenir une estimation lissée, $\tilde{V}_i$, pour σ_i^2, puis traiter l'estimation de la variance lissée $\tilde{V}_i$ dans le modèle d'échantillonnage (3) comme connu. Dans cet article, nous comparons deux méthodes de lissage. Une des méthodes est fondée sur la fonction de variance généralisée (FVG), l'autre sur les effets de plan (*deff*). Nous proposons ensuite un estimateur de la variance moyen lissé (VML) basé sur les estimateurs lissés de la FVG et de *deff*. Le principal objectif de l'article est de promouvoir les méthodes de FVG et *deff* proposées. Le VML est utilisé comme un choix supplémentaire, car il regroupe les estimations par FVG et *deff* en prenant leur moyenne.

Il existe de nombreuses applications de la FVG dans l'estimation sur petits domaines : voir, par exemple, les premiers travaux de Dick (1995) et l'application récente dans Hidiroglou, Beaumont et Yung (2019). L'effet de plan (*deff*) peut également servir dans la modélisation de la variance et le lissage pour l'estimation sur petits domaines. Par exemple, You (2008) a utilisé les effets de plan lissés dans le temps pour obtenir les matrices de variance et de covariance lissées. Liu, Lahiri et Kalton (2014) ont également appliqué des modèles au niveau du domaine à des proportions en utilisant les effets de plan pour lisser et modéliser la variance due à l'échantillonnage. Dans l'article, nous présentons une méthode générale pour calculer l'effet de plan et proposons un estimateur de la variance lissé fondé sur les effets moyens de plan sur les domaines. Nous montrerons également que l'estimateur de la variance lissé par l'effet de plan et l'estimateur de la variance lissé par la FVG sont à peu près équivalents dans certaines conditions. Nous illustrerons les méthodes de lissage de la variance due à l'échantillonnage par des applications à partir des données de l'Enquête canadienne sur la population active (EPA).

Le présent article est organisé comme suit. À la section 2, nous proposons des méthodes de lissage de la variance due à l'échantillonnage, y compris les méthodes FVG et *deff*. À la section 3, nous comparons les estimations fondées sur un modèle à partir de différentes estimations de la variance due à l'échantillonnage lissées au moyen des données sur le taux de chômage de l'EPA. À la section 4, en guise de conclusion nous proposons quelques observations.

2. Méthodes de lissage de la variance due à l'échantillonnage

2.1. Lissage au moyen de modèles log-linéaires

Dans cette section, nous construirons un modèle de FVG pour obtenir des variances d'échantillonnage lissées. Dans la pratique, cette procédure est largement utilisée pour modéliser la variance. Nous appliquons un modèle de régression log-linéaire sur la variance due à l'échantillonnage direct $\hat{V}_i$ en utilisant la taille d'échantillon n_i comme variable auxiliaire dans le modèle comme suit :

$$log(\hat{V}_i) = \beta_0 + \beta_1 log(n_i) + \varepsilon_i,\ i = 1, \ldots, m, \tag{4}$$

où le terme d'erreur du modèle est $\varepsilon_i \sim N(0, \tau^2)$, et la variance d'erreur du modèle τ^2 est inconnue. Notons que le modèle de régression proposé (4) est l'équivalent du modèle suivant :

$$log(\hat{V}_i) = \beta_0 + \beta_1 log(\frac{1}{n_i}) + \varepsilon_i,\ i = 1, \ldots, m, \tag{5}$$

où $log(1/n_i)$ sert de variable auxiliaire. Les modèles de FVG proposés (4) ou (5) sont les mêmes que ceux utilisés dans You (2021) pour la modélisation hiérarchique bayésienne (HB) de la variance due à l'échantillonnage. Ce modèle de FVG étend également le modèle proposé par Souza, Moura et Migon (2009) aux variances dues à l'échantillonnage au moyen de $log(1/n_i)$ et par l'ajout d'un effet aléatoire normal (ε_i) à la partie régression du modèle.

Supposons que $\hat{\beta}_0$ et $\hat{\beta}_1$ désignent les estimateurs par les moindres carrés ordinaires des coefficients de régression β_0 et β_1. On obtient un estimateur lissé par la FVG naïf de la variance due à l'échantillonnage en prenant l'exponentielle de la valeur ajustée :

$$\tilde{V}_i^{naïf} = exp(\hat{\beta}_0 + \hat{\beta}_1 log(n_i)). \quad (6)$$

Dick (1995) a utilisé l'estimateur lissé naïf $\tilde{V}_i^{naïf}$ dans l'application de l'estimation sur petits domaines du sous-dénombrement du recensement. Comme l'ont fait remarquer Rivest et Belmonte (2000), l'estimateur naïf lissé $\tilde{V}_i^{naïf}$ sous-estime la variance due à l'échantillonnage. On peut considérer cela comme suit. Si Y est une variable aléatoire log-normale avec une moyenne μ et une variance σ^2, la moyenne de Y est $E(Y) = exp(\mu)\, exp(\tau^2/2)$. Il s'ensuit que l'estimateur lissé $\tilde{V}_i^{naïf}$ sous-estime les vraies valeurs en ignorant le deuxième terme $exp(\tau^2/2)$ dans la moyenne de la variable aléatoire log-normale. Notons par $\widehat{\omega}_{RB} = exp(\hat{\tau}^2/2)$ la correction de Rivest et Belmonte (2000), où $\hat{\tau}^2$ est la variance résiduelle estimée du modèle de régression log-linéaire proposé (4). Ensuite, un estimateur lissé par la FVG, noté $\tilde{V}_i^{FVG.RB}$, est donné par

$$\tilde{V}_i^{FVG.RB} = \tilde{V}_i^{naïf} \cdot \widehat{\omega}_{RB} = \tilde{V}_i^{naïf} \cdot exp(\hat{\tau}^2/2). \quad (7)$$

L'estimateur de la FVG naïf $\tilde{V}_i^{naïf}$ dans (6) sous-estime la variance due à l'échantillonnage de $exp(\hat{\tau}^2/2)$. Ce terme, qui est toujours supérieur à 1, peut parfois être grand, dépendamment de la valeur de $\hat{\tau}^2$.

Hidiroglou, Beaumont et Yung (2019) ont proposé un autre terme de correction pour l'estimateur naïf $\tilde{V}_i^{naïf}$. Soit $\tilde{V}^{naïf}$ la somme des estimateurs de la variance lissés naïfs, c'est-à-dire, $\tilde{V}^{naïf} = \sum_{i=1}^{m} \tilde{V}_i^{naïf}$, et soit $\hat{V}^{total}$ la somme des variances dues à l'échantillonnage directes, c'est-à-dire, $\hat{V}^{total} = \sum_{i=1}^{m} \hat{V}_i$. À la suite d'Hidiroglou, Beaumont et Yung (2019), nous définissons un terme de correction, appelé Hidiroglou, Beaumont et Yung (HBY), comme étant $\widehat{\omega}_{HBY} = \hat{V}^{total} / \tilde{V}^{naïf}$. Cela donne un deuxième estimateur de la variance lissé par la FVG, noté $\tilde{V}_i^{FVG.HBY}$.. Il est donné par

$$\tilde{V}_i^{FVG.HBY} = \tilde{V}_i^{naïf} \cdot \widehat{\omega}_{HBY} = \tilde{V}_i^{naïf} \cdot \frac{\hat{V}^{total}}{\tilde{V}^{naïf}}. \quad (8)$$

Notons que l'on obtient un estimateur alternatif de $\widehat{\omega}_{HBY}$ comme estimateur de $exp(\tau^2/2)$ en utilisant la méthode des moments (Beaumont et Bocci, 2016). Cela évite la sensibilité du modèle FVG aux écarts par rapport à l'hypothèse de normalité de ε_i dans le modèle (4). Une des propriétés intéressantes de $\tilde{V}_i^{FVG.HBY}$ est que la somme des estimations lissées de la variance est égale à la somme des estimations directes de la variance due à l'échantillonnage, c'est-à-dire $\sum_{i=1}^{m} \tilde{V}_i^{FVG.HBY} = \sum_{i=1}^{m} \hat{V}_i$. Cette propriété peut garantir que la procédure de lissage ne surestime pas ou ne sous-estime pas systématiquement les variances dues à l'échantillonnage.

2.2. Lissage au moyen des effets de plan

Supposons que $\hat{p}_{iw}$ soit l'estimation directe fondée sur le plan de sondage pour une proportion p_i et $\hat{V}_i$ la variance due à l'échantillonnage directe correspondante sous un plan de sondage complexe pour le i[e] petit domaine. L'effet de plan estimé peut alors être calculé approximativement comme suit :

$$deff_i = \frac{\hat{V}_i}{\hat{p}_{iw}(1-\hat{p}_{iw})/n_i + \hat{V}_i/n_i}, \quad (9)$$

où n_i est la taille de l'échantillon du i[e] petit domaine; voir Gambino (2009). Toutefois, en notant que $deff_i$ dans l'équation (9) n'est pas égal à 1 dans un plan d'échantillonnage aléatoire simple, nous modifions $deff_i$ en le multipliant par un terme de correction $(n_i + 1)/n_i$:

$$deff_i = \frac{\hat{V}_i}{\hat{p}_{iw}(1-\hat{p}_{iw})/n_i + \hat{V}_i/n_i} \cdot \frac{n_i+1}{n_i}. \quad (10)$$

À l'aide de l'équation (10), nous pouvons réécrire la variance due à l'échantillonnage $\hat{V}_i$ fondée sur le plan comme suit :

$$\hat{V}_i = deff_i \cdot \frac{\hat{p}_{iw}(1-\hat{p}_{iw})}{n_i} \cdot \left(1 + \frac{1-deff_i}{n_i}\right)^{-1}. \quad (11)$$

Si la taille d'échantillon n_i est grande, le terme $(1 + (1 - deff_i)/n_i)^{-1}$ peut être négligeable dans (11), l'équation (11) se réduit à

$$\hat{V}_i = deff_i \cdot \frac{\hat{p}_{iw}(1-\hat{p}_{iw})}{n_i}. \quad (12)$$

L'équation (12) est utilisée, par exemple, dans Liu, Lahiri et Kalton (2014) aux fins de lissage et de modélisation de la variance due à l'échantillonnage. Cependant, dans l'estimation sur petits domaines, n_i peut être très petit, et le terme $(1 + (1 - deff_i)/n_i)^{-1}$ peut ne pas être négligeable.

Nous pouvons calculer tous les $deff_i$ des effets de plan en utilisant (10) pour tous les domaines, puis calculer la valeur moyenne sur tous les domaines et ainsi obtenir un effet de plan lissé $\overline{deff} = \frac{1}{m}\sum_{i=1}^{m} deff_i$. L'estimation de la proportion moyenne sur tous les domaines est donnée par $\bar{p}_w = \frac{1}{m}\sum_{i=1}^{m} \hat{p}_{iw}$. Si l'on remplace $deff_i$ par $\overline{deff}$ et $\hat{p}_{iw}$ par $\bar{p}_w$ dans l'équation (11), un estimateur avec *deff* lissé de la variance due à l'échantillonnage pour l'estimation de la proportion $\hat{p}_{iw}$ est :

$$\tilde{V}_i^{DEFF} = \overline{deff} \cdot \frac{\bar{p}_w(1-\bar{p}_w)}{n_i} \cdot \left(1 + \frac{1-\overline{deff}}{n_i}\right)^{-1}. \quad (13)$$

Si la taille de l'échantillon n_i est grande, le terme $(1 + (1 - \overline{deff})/n_i)^{-1}$ dans $\tilde{V}_i^{DEFF}$ peut être négligeable. La variance lissée $\tilde{V}_i^{deff}$ peut ensuite être simplifiée en

$$\tilde{V}_i^{DEFF} = \overline{deff} \cdot \frac{\bar{p}_w(1-\bar{p}_w)}{n_i}. \quad (14)$$

2.3. Comparaison du lissage par FVG et *deff*

Nous montrons maintenant que les estimateurs par FVG et l'estimateur *deff* $\tilde{V}_i^{DEFF}$ peuvent avoir des performances similaires dans certaines conditions. En utilisant $\tilde{V}_i^{FVG.RB}$ comme illustration, nous pouvons exprimer ce terme comme suit :

$$\tilde{V}_i^{FVG.RB} = exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot log(n_i)) \cdot exp(\frac{\hat{\tau}^2}{2}) = C_0 \cdot exp(log(n_i)^{\hat{\beta}_1}) = C_0 \cdot n_i^{\hat{\beta}_1}$$

où $C_0 = exp(\hat{\beta}_0 + \frac{\hat{\tau}^2}{2})$ est une constante. Si la valeur du coefficient de régression $\hat{\beta}_1$ est proche de -1, alors l'estimateur FVG $\tilde{V}_i^{FVG.RB}$ peut s'écrire approximativement comme étant $\tilde{V}_i^{FVG.RB} \approx C_0/n_i$.

L'estimateur *deff* $\tilde{V}_i^{DEFF}$ peut être réécrit comme suit :

$$\tilde{V}_i^{DEFF} = \overline{deff} \cdot \frac{\bar{p}_w(1-\bar{p}_w)}{n_i} \cdot \left(1 + \frac{1-\overline{deff}}{n_i}\right)^{-1} = \frac{C_1}{n_i} \cdot \left(\frac{n_i + 1 - \overline{deff}}{n_i}\right)^{-1} = \frac{C_1}{n_i + 1 - \overline{deff}} \approx \frac{C_1}{n_i},$$

où $C_1 = \overline{deff} \cdot \bar{p}_w(1-\bar{p}_w)$ est une constante. L'estimateur FVG $\tilde{V}_i^{FVG.RB}$ et l'estimateur *deff* $\tilde{V}_i^{DEFF}$ sont proportionnels à n_i^{-1} si le coefficient de régression $\hat{\beta}_1$ est proche de -1 dans le modèle de régression FVG. Par conséquent, dans de telles conditions, les variances lissées par FVG et *deff* devraient avoir des performances similaires.

Dans les applications pratiques, pour utiliser à la fois les estimations lissées par FVG et *deff*, nous pouvons définir un estimateur lissé moyen (VML)) $\tilde{V}_i^{VML}$=($\tilde{V}_i^{FVG.RB}$+$\tilde{V}_i^{FVGVF.HBY}$+$\tilde{V}_i^{DEFF}$)/3 comme méthode simple de regroupement de données pour obtenir l'estimation de la variance lissée finale. Comme nous le verrons dans l'application sur petits domaines de l'EPA à la section 3, l'estimateur lissé moyen $\tilde{V}_i^{VML}$ peut donner de très bons résultats et entraîner une réduction importante du biais et du CV pour les estimations sur petits domaines.

3. Estimation sur petits domaines de l'EPA au moyen de variances d'échantillonnage lissées

Dans cette section, nous appliquons les méthodes de lissage de la variance aux données de l'Enquête canadienne sur la population active (EPA) et comparons les estimations sur petits domaines basées sur les variances dues à

l'échantillonnage lissées. L'EPA produit des estimations mensuelles du taux de chômage aux niveaux national et provincial. L'EPA publie aussi des estimations du chômage pour les régions infraprovinciales comme les régions métropolitaines de recensement (RMR) et les agglomérations de recensement (AR) du Canada. Cependant, les estimations directes ne sont pas fiables pour les régions infraprovinciales en raison de la taille relativement petite de l'échantillon dans certaines régions. Les divers modèles d'estimation sur petits domaines pour l'EPA sont examinés, notamment, dans Hidiroglou, Beaumont et Yung (2019), Lesage, Beaumont et Bocci (2021), You, Rao et Gambino (2003) et You (2008, 2021). Nous appliquons le modèle de Fay-Herriot donné par (1) et (2) aux estimations du taux de chômage de mai 2016 au niveau des RMR/AR. Nous envisageons d'utiliser quatre estimateurs de la variance lissés dans l'application à l'EPA, à savoir $\tilde{V}_i^{FVG.RB}$, $\tilde{V}_i^{FGV.HBY}$, $\tilde{V}_i^{DEFF}$ et l'estimateur lissé moyen $\tilde{V}_i^{VML}$=($\tilde{V}_i^{FVG.RB}$+$\tilde{V}_i^{FVG.HBY}$+$\tilde{V}_i^{DEFF}$)/3. Nous considérons la méthode du meilleur prédicteur linéaire sans biais empirique (MPLSBE ou EBLUP en anglais) dans l'application. Les détails concernant l'estimateur EBLUP et l'estimation de l'EQM connexe basés sur le modèle de Fay-Herriot se trouvent, par exemple, dans Rao et Molina (2015) et You (2021). Le taux mensuel de prestataires d'assurance-emploi dans la région géographique locale est utilisé comme variable auxiliaire dans le modèle de Fay-Herriot, comme dans Hidiroglou, Beaumont et Yung (2019) et You (2008, 2021). Nous comparons les estimations basées sur un modèle et les estimations directes aux estimations du recensement pour évaluer les effets du lissage de la variance due à l'échantillonnage.

Nous obtenons d'abord les variances d'échantillonnage lissées pour l'ensemble des 128 RMR/AR en utilisant les $\tilde{V}_i^{FVG.RB}$, $\tilde{V}_i^{FVG.HBY}$, $\tilde{V}_i^{DEFF}$ et $\tilde{V}_i^{VML}$ proposées. Pour le modèle de FVG (4), les estimations par la régression sont $\hat{\beta}_0 = -3{,}194$ et $\hat{\beta}_1 = -0{,}901$. Le terme de correction résiduelle RB = $exp(\hat{\tau}^2/2)$ est égal à 1,467 et le terme de correction HBY est $\hat{\omega}_{HBY} = \hat{V}^{total}/\tilde{V}^{naïf}$ =1,786. Comme le coefficient de régression $\hat{\beta}_1 = -0.901$ est proche de -1 et que la différence entre les termes des deux corrections n'est pas grande, nous devrions nous attendre à des variances dues à l'échantillonnage lissées semblables pour les données de l'EPA. Nous avons appliqué le modèle de Fay-Herriot à 128 données sur le taux de chômage de l'EPA des RMR et des AR avec les quatre variances d'échantillonnage lissées différentes et nous avons obtenu les estimations EBLUP correspondantes. Des détails concernant l'estimateur EBLUP avec la méthode du maximum de vraisemblance restreint (ou REML, pour *restricted maximum likelihood*) aux fins de l'estimation de la composante de la variance se trouvent, par exemple, dans You (2021) et Rao et Molina (2015). Les estimations sur petits domaines EBLUP sont comparées au moyen de l'erreur relative absolue (ERA) des estimations directes et EBLUP relativement aux estimations du recensement pour chaque RMR ou AR, comme suit : $ERA_i = \left|(\theta_i^{Recens.} - \theta_i^{Est})/\theta_i^{Recens.}\right|$, où θ_i^{Est} est l'estimation directe ou par le meilleur prédicteur linéaire sans biais empirique et $\theta_i^{Recens.}$ est la valeur correspondante du recensement pour le taux de chômage de l'EPA. Très couramment, on évalue les estimations basées sur un modèle avec les valeurs du recensement, comme dans Hidiroglou, Beaumont et Yung (2019) et You (2021). Nous prenons ensuite la moyenne des erreurs relatives absolues sur les RMR ou AR par différents sous-groupes en ce qui a trait à la taille de l'échantillon, comme dans Hidiroglou, Beaumont et Yung (2019). Le tableau 3.1 présente l'erreur relative absolue moyenne pour les estimateurs directs de l'EPA et EBLUP fondés sur différentes estimations en entrée de la variance due à l'échantillonnage. À des fins de comparaison, nous avons également utilisé la variance due à l'échantillonnage directe comme variance due à l'échantillonnage d'entrée dans le modèle de Fay-Herriot. Par exemple, EBLUP(DIR) signifie que l'estimation de la variance due à l'échantillonnage directe (DIR) est utilisée dans le modèle de Fay-Herriot, EBLUP(FVG.RB) signifie que l'estimation de la variance due à l'échantillonnage lissée $\tilde{V}_i^{FVG.RB}$ (FVG.RB) est utilisée, etc.

Tableau 3.1 : Comparaison des erreurs relatives absolues pour les estimations des EBLUP basées sur différentes variances dues à l'échantillonnage en entrée

RMR/AR	EPA directe	EBLUP (DIR)	EBLUP (FVG.RB)	EBLUP (FVG.HBY)	EBLUP (*deff*)	EBLUP (VML)
25 plus petits domaines	0,489	0,279	0,181	0,184	0,180	0,182
25 plus petits domaines suivants	0,338	0,214	0,146	0,147	0,146	0,146
25 plus petits domaines suivants	0,276	0,198	0,138	0,143	0,134	0,138
25 plus petits domaines suivants	0,198	0,161	0,134	0,141	0,130	0,135
28 plus grands domaines	0,132	0,125	0,099	0,108	0,091	0,099
Tous les domaines	0,283	0,194	0,138	0,144	0,135	0,139

Il ressort clairement du tableau 3.1 que les estimations EBLUP améliorent considérablement les estimations directes en réduisant les erreurs relatives absolues. Même en cas d'utilisation des estimations directes de la variance due à l'échantillonnage, EBLUP(DIR) donne une erreur relative absolue beaucoup plus petite que l'estimateur d'enquête direct. Cependant, par son utilisation d'estimations de la variance due à l'échantillonnage lissées, EBLUP a des performances nettement supérieures à celles de l'estimateur direct. Les erreurs relatives absolues sont réduites pour chaque groupe du domaine et pour l'ensemble des domaines. En général, tous les EBLUP avec les quatre variances d'échantillonnage lissées donnent des résultats très semblables. Parmi les estimateurs EBLUP utilisant des variances d'échantillonnage lissées, EBLUP(FVG.HBY) a une erreur relative absolue légèrement plus grande que les autres, et EBLUP(*deff*) a une erreur relative absolue légèrement plus petite. Par exemple, pour toutes les 128 RMR/AR, les erreurs relatives absolues respectives d'EBLUP(FVG.RB), EBLUP(FVG.HBY) et EBLUP(*deff*) sont de 0,138, 0,144 et 0,135. EBLUP(*deff*) est donc le plus performant pour ce qui est de l'erreur relative. Pour la variance due à l'échantillonnage lissée moyenne $\tilde{V}_i^{VML}$, EBLUP(VML) a une erreur relative absolue globale de 0,139, située entre les valeurs d'erreur relative absolue des EBLUP utilisant FVG et *deff*. EBLUP(VML) a d'excellentes performances.
Pour ce qui est du CV moyen, EBLUP réduit aussi considérablement le CV par rapport à l'estimateur direct. L'estimateur direct de l'EPA a un CV moyen de 39,4 %, EBLUP(DIR) a un CV moyen de 24,5 %, tandis qu'EBLUP(FVG.RB) a un CV moyen de 10,3 %, EBLUP(FVG.HBY) a un CV moyen légèrement plus petit à 8,2 %, et EBLUP(*deff*) a une valeur moyenne de CV de 11,8 %. EBLUP(VML) a un CV moyen de 10,2 %. Par conséquent, l'utilisation de variances d'échantillonnage lissées réduit considérablement le CV pour les EBLUP, et encore une fois, le CV pour EBLUP(VML) se situe entre les valeurs du CV des EBLUP utilisant des variances FVG et *deff*. EBLUP(VML) a une erreur relative absolue plus petite qu'EBLUP(FVG.RB) et EBLUP(FVG.HBY), et un CV plus petit qu'EBLUP(FVG.RB) et EBLUP(*deff*). L'utilisation des variances dues à l'échantillonnage lissées moyennes $\tilde{V}_i^{VML}$ donne une réduction équilibrée pour l'erreur relative absolue et le coefficient de variation. Il apparaît clairement que l'estimateur lissé moyen $\tilde{V}_i^{VML}$ donne de très bons résultats.

Lesage, Beaumont et Bocci (2021) ont examiné le modèle de lissage suivant, appelé modèle LBB, aux fins de lissage de la variance due à l'échantillonnage :

$$log(\hat{V}_i) = \beta_0 + \beta_1\, log(z_i) + \beta_2\, log(1 - z_i) + \beta_3\, log(n_i) + \varepsilon_i,\ i = 1, \ldots, m, \qquad (15)$$

où z_i est le taux de prestataires d'assurance-emploi utilisé dans le modèle de Fay-Herriot comme variable auxiliaire pour obtenir les estimateurs EBLUP. En appliquant le modèle de lissage LBB (15) aux données de variance due à l'échantillonnage sur les 128 domaines, nous obtenons les estimations par la régression suivantes $\hat{\beta}_0 = -4{,}443$, $\hat{\beta}_1 = -0{,}486$, $\hat{\beta}_2 = -29{,}139$ and $\hat{\beta}_3 = -0{,}886$. Le terme de correction résiduelle $\hat{\omega}_{RB} = exp(\hat{\tau}^2/2)$ est égal à 1,461 et le terme de correction HBY est $\hat{\omega}_{HBY} = \hat{V}^{total} / \tilde{V}^{naïf}$ =1,782. Nous désignons par $\tilde{V}_i^{LBB.RB}$ l'estimateur de la variance lissée basé sur le modèle LBB (15) en utilisant la formule (7) avec un terme de correction $\hat{\omega}_{RB}$=1,461. De même, soit $\tilde{V}_i^{LBB.HBY}$ l'estimateur de la variance lissée basé sur le modèle LBB (15) au moyen de la formule (8) avec un terme de correction $\hat{\omega}_{HBY}$ ==1,782. Nous comparons maintenant les estimations d'EBLUP fondées sur le modèle de lissage LBB et la méthode de lissage proposée. Plus particulièrement, nous comparons les estimations proposées EBLUP(VML) aux estimations EBLUP en utilisant $\tilde{V}_i^{LBB.RB}$ et $\tilde{V}_i^{LBB.HBY}$, par exemple EBLUP(LBB.RB) et EBLUP(LBB.HBY).

Tableau 3.2 : Comparaison des erreurs relatives absolues basée sur différents modèles de FVG et sur des variances dues à l'échantillonnage lissées

RMR/AR	EPA directe	EBLUP(VML)	EBLUP(LBB.RB)	EBLUP(LBB.HBY)
25 plus petits domaines	0,489	0,182	0,181	0,183
25 plus petits domaines suivants	0,338	0,146	0,144	0,145
25 plus petits domaines suivants	0,276	0,138	0,137	0,142
25 plus petits domaines suivants	0,198	0,135	0,135	0,141
28 plus grands domaines	0,132	0,099	0,099	0,108
Tous les domaines	0,283	0,139	0,138	0,143

Le tableau 3.2 présente l'erreur absolue moyenne aux fins de comparaison des effets du lissage de la variance au moyen du modèle VML et LBB. Le tableau 3.2 montre clairement que toutes les estimations des EBLUP donnent de très bons résultats et améliorent les estimations d'enquêtes directes en réduisant considérablement l'erreur relative absolue par rapport aux valeurs du recensement. EBLUP(VML) et EBLUP(LBB.RB) ont presque les mêmes performances, et EBLUP(LBB.HBY) a une erreur relative absolue légèrement plus grande, comme EBLUP(FVG.HBY) dans le tableau 3.1. EBLUP(LBB.HBY) et EBLUP(FVG.HBY) ont des performances presque identiques si l'on compare les résultats du tableau 3.1 et du tableau 3.2. En ce qui concerne le CV, EBLUP(LBB.RB) et EBLUP(VML) ont le même CV moyen de 10,2 %, et EBLUP(LBB.HBY) a le même CV moyen de 8,2 % qu'EBLUP(FVG.HBY). L'application sur petits domaines de l'EPA montre que le modèle de FVG proposé (4) et les méthodes de lissage de la variance due à l'échantillonnage proposées FVG, *deff* et VML donnent de très bons résultats si l'on compare les estimations EBLUP avec les valeurs du recensement et d'autres modèles de lissage par FVG pour l'application àl'EPA, comme dans Lesage, Beaumont et Bocci. (2021).

4. Conclusion

Nous avons proposé dans l'article des estimateurs de lissage de la variance due à l'échantillonnage qui utilisent la méthode de la fonction de variance généralisée et la méthode de l'effet de plan lissé aux fins d'estimations sur petits domaines. Les modèles et méthodes de lissage proposés nécessitent seulement d'utiliser la taille de l'échantillon dans le modèle et le calcul des effets de plan. Les estimateurs proposés $\tilde{V}_i^{FVG.RB}$, $\tilde{V}_i^{FGB.HBY}$ et $\tilde{V}_i^{DEFF}$ donnent habituellement des estimations de variance lissées semblables. Dans les applications pratiques, nous pourrions utiliser l'estimateur lissé moyen $\tilde{V}_i^{VML}=(\tilde{V}_i^{FVG.RB}+\tilde{V}_i^{FVG.HBY}+\tilde{V}_i^{DEFF})/3$ pour obtenir l'estimation finale de la variance lissée. Les méthodes de lissage proposées simplifient la procédure de lissage dans la pratique, car leurs utilisateurs n'ont pas besoin d'autres modèles complexes de FVG ou de variables auxiliaires pour modéliser la variance due à l'échantillonnage. De plus, la procédure de lissage proposée peut facilement être mise en œuvre en situation réelle.

Bibliographie

Beaumont, J.-F., et C. Bocci (2016), « Estimation sur petits domaines dans l'Enquête sur la population active », document présenté au Comité consultatif des méthodes statistiques, mai 2016, Statistique Canada.

Dick, P. (1995), « Modélisation du sous-dénombrement net dans le recensement du Canada de 1991 », *Techniques d'enquête*, 21, 1, p. 45 à 54.

Fay, 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, p. 269 à 277.

Gambino, J.G. (2009), « Design effect caveats », *The American Statistician*, 63, p. 141 à 146.

Hidiroglou, M.A., Beaumont, J.-F., et W. Yung (2019), « Élaboration d'un système d'estimation sur petits domaines à Statistique Canada », *Techniques d'enquête*, 45, nº 1, p. 101 à 126.

Lesage, E., Beaumont, J.F., et C. Bocci (2021), « Deux diagnostics locaux pour évaluer l'efficacité du meilleur prédicteur empirique issu du modèle de Fay-Herriot », *Techniques d'enquête*, 47, nº 2, p. 279 à 297.

Liu, B., Lahiri, P., et G. Kalton (2014), « Modélisation hiérarchique bayésienne de proportions dans de petits domaines pondérées par les poids de sondage », *Techniques d'enquête*, 40, nº 1, p. 1 à 13.

Rivest, L.P., et E. Belmonte (2000), « Une erreur quadratique moyenne conditionnelle des estimateurs régionaux », *Techniques d'enquête*, 26, p. 67 à 78.

Rivest, L.P., et N. Vandal (2002), « Mean squared error estimation for small areas when the small area variances are estimated », *Proceedings of the International Conference on Recent Advances in Survey Sampling*, 13 juillet, 2002, Ottawa, Canada.

Rao, J.N.K., et I. Molina (2015), *Small Area Estimation*, 2e edition, New York: John Wiley & Sons.

Souza, D.F., Moura, F.A.S., et H.S. Migon (2009), « Prédiction de la population de petits domaines au moyen de modèles hiérarchiques », *Techniques d'enquête*, 35, p. 203 à 214.

Sugasawa, S., Tamae, H., et T. Kubokawa (2017), « Bayesian estimators for small area models shrinking both means and variances », *Scandinavian Journal of Statistics*, 44, p. 150 à 167.

Wang, J., et W. A. Fuller (2003), « The mean square error of small area predictors constructed with estimated area variances », *Journal of the American Statistical Association*, 98, p. 716 à 723.

You, Y. (2008), « Une approche intégrée de modélisation de l'estimation du taux de chômage pour les régions infraprovinciales au Canada », *Techniques d'enquête*, 34, 1, p. 19 à 27.

You, Y. (2021), « Estimation sur petits domaines à l'aide du modèle au niveau de domaine de Fay-Herriot avec lissage et modélisation de variance d'échantillonnage », *Techniques d'enquête*, 47, p. 361 à 370.

You, Y., et B. Chapman (2006), « Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances d'échantillonnage », *Techniques d'enquête*, 32, p. 97 à 103.

You, Y., Rao, J.N.K., et M. Hidiroglou (2013), « De la performance des estimateurs sur petits domaines autocalés sous le modèle au niveau du domaine de Fay-Herriot », *Techniques d'enquête*, 39, p. 217 à 229.