

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

Ingénierie inverse d'un processus de réconciliation pour estimer l'erreur quadratique moyenne d'estimations sur petits domaines réconciliés

par François Verret et Braedan Walker

Date de diffusion : le 8 septembre 2025



Statistique
Canada

Statistics
Canada

Canada

Ingénierie inverse d'un processus de réconciliation pour estimer l'erreur quadratique moyenne d'estimations sur petits domaines réconciliés

François Verret et Braedan Walker¹

Résumé

L'approche de l'estimation sur petits domaines est fréquemment utilisée pour produire des estimations à un niveau désagrégé lorsque l'estimation directe à partir d'une enquête ne peut s'appuyer sur un échantillon suffisamment grand pour produire des estimations précises. Cela se fait souvent à l'aide du modèle au niveau du domaine de Fay-Herriot, en supposant que les estimations directes sont indépendantes sous le plan de sondage et ont une variance connue, et en appliquant un processus de lissage aux estimations de la variance des estimations directes afin de mieux répondre à cette dernière hypothèse. Il n'est pas rare que l'on compare les estimations sur petits domaines aux estimations directes à un niveau agrégé. Dans le présent article, il est démontré qu'une hypothèse d'indépendance erronée peut avoir une incidence importante sur l'erreur quadratique moyenne (EQM) des estimations réconciliées. Les valeurs des covariances entre les estimations directes sont donc nécessaires pour obtenir de bonnes estimations ponctuelles et d'EQM. Il est difficile d'obtenir de bonnes estimations de ces covariances en raison de la petite taille des échantillons dans certains domaines. Une manière originale de dériver les valeurs de ces covariances, en procédant à l'ingénierie inverse d'un processus de réconciliation hypothétique, est présentée.

Mots clés : Calage; modèle de Fay-Herriot; matrice de covariances de travail.

1. Introduction

Ces dernières années, Statistique Canada a déployé des efforts considérables pour produire des estimations à un niveau plus désagrégé. Son Plan d'action sur les données désagrégées traite les différentes expériences de vie de divers groupes en décomposant les données en sous-catégories selon les peuples autochtones, le genre, les populations racisées et les personnes ayant une incapacité. Les données sont également ventilées au plus fin niveau géographique possible. Statistique Canada utilise donc de plus en plus les méthodes d'estimation sur petits domaines (EPD) pour produire des statistiques, notamment des statistiques sur l'industrie manufacturière et le travail, et envisage d'utiliser ces méthodes dans d'autres domaines. La modélisation au niveau du domaine à l'aide du modèle de Fay-Herriot (Fay et Herriot, 1979) est la méthode la plus souvent utilisée, car elle intègre efficacement et facilement les mécanismes aléatoires du modèle et du plan. Pour l'estimation de totaux, les praticiens des enquêtes pourraient souhaiter faire correspondre les estimations sur petits domaines à certaines estimations directes (potentiellement publiées) à un niveau plus agrégé. À cette fin, il est possible d'appliquer un processus tel que la réconciliation (Dagum et Cholette, 2006).

C'est le cas de l'EPD mensuelle du total de l'emploi dans l'Enquête sur la population active (EPA), laquelle fait l'objet du présent document, où les estimations sont obtenues pour un partitionnement géographique de chacune des dix provinces défini par les régions métropolitaines de recensement (RMR), les agglomérations de recensement (AR) et les zones de travail autonomes (ZTA). Les RMR correspondent aux plus grandes villes du pays, soit celles dont la population est supérieure ou égale à 100 000 habitants et dont le noyau est supérieur ou égal à 50 000 habitants. Les AR correspondent aux villes plus petites. Elles ne font pas partie d'une RMR et ont une population d'au moins 10 000 habitants. Les ZTA désignent des zones fonctionnelles composées de subdivisions de recensement en dehors des RMR et des AR, regroupées en fonction des tendances du navettage (OCDE, 2020). Bien que l'échantillon mensuel de l'EPA soit important (plus de 53 000 ménages échantillonnés; Statistique Canada, 2020) et que l'enquête produise de bonnes estimations pour la plupart des RMR, celle-ci n'est pas conçue pour produire des estimations précises pour toutes les régions chaque mois. Par conséquent, de nombreuses petites régions n'auront que peu ou pas

¹François Verret, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6 (francois.verret@statcan.gc.ca); Braedan Walker, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6 (braedan.walker@statcan.gc.ca)

d'échantillon. Pour ces domaines, l'estimation peut être égale à 0 et la qualité serait susceptible d'être mauvaise. L'estimation de la variance correspondante sera également de mauvaise qualité pour la même raison. C'est pour cette raison, et parce que d'excellentes informations auxiliaires sont disponibles au niveau du domaine, que nous nous appuyons sur des méthodes d'estimation sur petits domaines.

On utilise sur une base mensuelle et transversale le modèle de Fay-Herriot (FH) pour mener à bien l'EPD et on applique de façon transversale un processus de réconciliation aux totaux d'emploi provinciaux. Le modèle de FH de base suppose l'indépendance basée sur le plan entre les estimations directes au niveau des domaines (ou, de manière équivalente, l'indépendance entre les erreurs d'échantillonnage du modèle). L'indépendance n'existe toutefois pas dans le contexte du total de l'emploi de l'EPA en raison du calage des poids. Une extension du modèle de FH où les erreurs d'échantillonnage sont dépendantes doit donc être ajustée. Pour ajuster le modèle de FH de base, les variances de l'estimation directe sur petits domaines doivent être connues. Il arrive fréquemment que l'on améliore ces estimations en appliquant un processus de lissage de la variance comportant un modèle de fonction de variance généralisée. Lorsque l'on étend le modèle de FH pour tenir compte des erreurs d'échantillonnage corrélées entre les domaines, une difficulté supplémentaire apparaît dans l'estimation des termes de covariance en plus des termes de variance, étant donné les petites tailles d'échantillon dans les domaines. Le présent article décrit une méthode dans laquelle un processus de réconciliation hypothétique (note : ce n'est pas celui appliqué aux EPD du modèle de FH) est inversé pour obtenir des termes de covariance tels que les variances lissées soient préservées et que les éléments de la matrice de variance résultante, lorsque l'on fait leur somme, correspondent à des estimations de variance de bonne qualité d'estimations directes à un niveau agrégé.

Le document est structuré comme suit. La section 2 présente le modèle de FH d'EPD étendu (lequel comprend le modèle de base lorsque tous les termes de covariance sont nuls) et l'application initiale du modèle de base pour estimer l'emploi total dans l'EPA. La section 3 décrit la réconciliation, son application aux estimations sur petits domaines du total de l'emploi de l'EPA et indique où l'hypothèse d'indépendance montre des signes d'échec. La section 4 décrit l'ingénierie inverse d'un processus de réconciliation hypothétique pour obtenir une matrice de variance de travail sous le plan complète et comment les estimations ponctuelles et d'EQM sont améliorées grâce à l'utilisation de cette matrice de variance de travail non diagonale. La section 5 comporte une conclusion.

2. Le modèle (de base et) étendu de Fay-Herriot

Dans le cadre du modèle de FH étendu (avec des erreurs d'échantillonnage possiblement dépendantes), le vecteur de paramètres de la population finie qui nous intéresse (de dimension M par 1, où M est le nombre de domaines) est :

$$\theta = X\beta + Bv,$$

où B est une matrice diagonale de constantes connues et $v \sim N(0, \sigma^2 I)$. Dans le cadre de notre application, nous travaillons avec des estimations de totaux et $B = \text{diag}(N_1, \dots, N_M)$, où N_m est le nombre de personnes âgées de 15 ans et plus dans le domaine m . L'estimateur direct est donné par :

$$\hat{\theta} = X\beta + Bv + e,$$

où $e \sim N(0, \Psi)$ est indépendant de v . L'estimateur composite (ou meilleur prédicteur linéaire sans biais empirique, EBLUP pour *empirical best linear unbiased predictor*) est donné par :

$$\begin{aligned}\hat{\theta}^{\text{EPD}} &= X\hat{\beta} + \hat{\sigma}^2 B^2 \hat{\Sigma}_{zz}^{-1} (\hat{\theta} - X\hat{\beta}) \\ &= \hat{H}^T X\hat{\beta} + (I - \hat{H}^T) \hat{\theta},\end{aligned}$$

où $\hat{\beta} = (X^T \hat{\Sigma}_{zz}^{-1} X)^{-1} X^T \hat{\Sigma}_{zz}^{-1} \hat{\theta}$, $\hat{\Sigma}_{zz} = \hat{\sigma}^2 B^2 + \Psi$ et $\hat{H}^T = I - \hat{\sigma}^2 B^2 \hat{\Sigma}_{zz}^{-1}$. À des fins d'estimation, on suppose que la matrice complète Ψ est connue. Le modèle de FH de base correspond à une matrice Ψ diagonale. Il est le modèle le plus souvent utilisé dans la pratique et est celui qui est programmé dans la plupart des logiciels, y compris le système généralisé d'estimation (G-Est) de Statistique Canada. (Estevao et coll., 2023). Le modèle de base a été le premier à être ajusté pour cette application de l'EPA.

Pour cette application, les informations auxiliaires sont constituées des dénombrements de la population en âge de travailler (15 à 64 ans) et des bénéficiaires de l'assurance-emploi (une régression spline quadratique a été effectuée pour la première variable). On suppose que la variance du plan des estimations directes, ψ_m , est connue, mais elle ne l'est manifestement pas dans la pratique. Les estimations $\hat{\psi}_m$ sont obtenues à l'aide de méthodologies traditionnelles

fondées sur le plan (en utilisant la méthode bootstrap de Rao-Wu [1988]) et un processus de lissage leur est appliqué pour compenser leur manque de précision causé par la petite taille des échantillons en ajustant une fonction de variance généralisée (Beaumont et Bocci, 2016). Nous obtenons ainsi des valeurs lissées $\tilde{\psi}_m$, qui sont supposées être suffisamment proches de ψ_m . La variance des erreurs du modèle (σ^2) est estimée à l'aide du maximum de vraisemblance avec contrainte (MVC), compte tenu des variances connues des erreurs d'échantillonnage. L'EBLUP du domaine m sous le modèle de base est donnée par :

$$\hat{\theta}_m^{\text{EPD}} = \hat{\gamma}_m \hat{\theta}_m + (1 - \hat{\gamma}_m) \mathbf{x}_m^T \hat{\boldsymbol{\beta}},$$

où $\hat{\gamma}_m = b_m^2 \hat{\sigma}^2 / (b_m^2 \hat{\sigma}^2 + \tilde{\psi}_m)$ (b_m étant le m^{e} élément de la diagonale de B) et $\mathbf{x}_m^T \hat{\boldsymbol{\beta}}$ est l'estimateur synthétique du domaine m . L'estimateur composite sera donc plus près de l'estimateur direct si la variance du modèle est importante par rapport à la variance du plan et vice-versa.

La matrice des EQM de l'estimateur composite est donnée en annexe (comme cas spécial de la matrice des EQM de l'estimateur composite réconcilié) et a pour composantes $G_1(\sigma^2)$, $G_2(\sigma^2)$ et $G_3(\sigma^2)$. $G_1(\sigma^2)$ est le terme principal représentant l'EQM si on avait connu $\boldsymbol{\beta}$ et σ^2 , $G_2(\sigma^2)$ est la variance due à l'estimation de $\boldsymbol{\beta}$ et $G_3(\sigma^2)$ est la variance due à l'estimation de σ^2 . Que les erreurs d'échantillonnage soient indépendantes ou non (c'est-à-dire que Ψ soit diagonale ou non), lors de l'ajustement du modèle de FH de base, nous constatons que le m^{e} élément de la diagonale de $G_1(\sigma^2)$ est $g_{1m}(\sigma^2) = \gamma_m^2 \psi_m + (1 - \gamma_m)^2 N_m^2 \sigma^2$. Dans tous les cas, on obtient une estimation de l'EQM à l'aide de la somme $G_1(\hat{\sigma}^2) + G_2(\hat{\sigma}^2) + 2G_3(\hat{\sigma}^2)$.

3. Réconciliation appliquée aux estimations sur petits domaines

Les estimations provinciales directes de l'EPA du total de l'emploi ont une bonne précision et sont diffusées. Lorsque l'on agrège les estimations sur petits domaines (obtenues à l'aide du modèle de base) au niveau provincial, on constate dans le tableau 3-1 qu'elles peuvent être très différentes des estimations directes provinciales en termes relatifs, en particulier pour l'Île-du-Prince-Édouard et la Saskatchewan. Les différences sont de même ampleur et systématiques lorsque l'on examine les autres mois (non présentés). Cela indique que le modèle n'est pas parfait et explique le choix de réconcilier les estimations sur petits domaines.

Tableau 3-1

Différences provinciales entre l'estimation directe de l'emploi total et l'agrégation des estimations sur petits domaines en supposant des erreurs d'échantillonnage indépendantes pour le mois de novembre 2021

Province	Est. directe	Est. sur petits domaines	Différence	Différence relative (%)
T.-N.-L.	229 601	224 525	5 076	2,2
Î.-P.-É.	81 098	73 215	7 883	9,7
N.-É.	477 326	483 855	-6 529	-1,4
N.-B.	360 082	360 306	-224	-0,1
Qc	4 380 216	4 361 616	18 600	0,4
Ont.	7 660 158	7 778 329	-118 171	-1,5
Man.	669 344	652 554	16 790	2,5
Sask.	556 206	509 706	46 500	8,4
Alb.	2 298 973	2 277 357	21 616	0,9
C.-B.	2 708 386	2 701 059	7 327	0,3
Canada	19 421 390	19 422 521	-1 131	0,0

La réconciliation est conçue pour rétablir les contraintes d'agrégation dans les systèmes de séries chronologiques, par exemple, après la désaisonnalisation. Bien que nous ne soyons pas dans un contexte de séries chronologiques, nous pouvons utiliser un cadre transversal de réconciliation pour que la somme provinciale des EPD corresponde à

l'estimation directe provinciale. En réconciliant nos estimations sur petits domaines $\hat{\theta}_m^{\text{EPD}}$, nous recherchons de nouvelles estimations $\hat{\theta}_m^{\text{EPD, réconciliée}}$ qui minimisent la distance suivante :

$$\sum_{m=1}^M \frac{(\hat{\theta}_m^{\text{EPD}} - \hat{\theta}_m^{\text{EPD, réconciliée}})^2}{|c_m \hat{\theta}_m^{\text{EPD}}|},$$

sous la contrainte que la somme provinciale des nouvelles estimations corresponde aux estimations provinciales directes de l'EPA. Les changements relatifs dans les estimations individuelles de la réconciliation peuvent être contrôlés à l'aide des « coefficients d'altérabilité », c_m . Afin de limiter les changements aux estimations sur petits domaines les plus précises, nous avons choisi de définir c_m comme étant la racine de l'erreur quadratique moyenne relative (REQMR) estimée de $\hat{\theta}_m^{\text{EPD}}$: $c_m = \sqrt{\varphi_m / \hat{\theta}_m^{\text{EPD}}}$, où $\varphi_m = \text{eqm}(\hat{\theta}_m^{\text{EPD}})$.

Afin de mesurer les avantages de la réconciliation aux estimations provinciales, nous avons appliqué nos procédures à l'EPA de mai 2016 et comparé nos estimations à celles du questionnaire détaillé du Recensement de 2016 en utilisant la différence relative absolue moyenne (DRAM) entre les deux ensembles d'estimations. En ce qui concerne les RMR et les AR, la DRAM des estimations directes s'élève à 23,9 %. La DRAM des estimations du modèle de FH (non réconciliées) s'élève à 6,3 %. La DRAM des estimations réconciliées est de 6,5 %. L'estimation sur petits domaines a donc un effet très bénéfique sur les estimations des RMR et des AR, tandis que la réconciliation a un effet négatif marginal. Pour les ZTA, la DRAM des estimations directes est de 70,4 %. La DRAM de l'estimation du modèle de FH s'élève à 26,6 %. Celle des estimations réconciliées est de 22,2 %. L'estimation sur petits domaines et la réconciliation ont donc une incidence bénéfique pour ces domaines.

Tableau 3-2

Différences provinciales entre le coefficient de variation (c.v.) de l'estimation directe de l'emploi total et la REQMR de l'agrégation des estimations sur petits domaines en supposant des erreurs d'échantillonnage indépendantes pour le mois de novembre 2021

Province	C.v. direct	REQMR des agrégats des estimations réconciliées	Quotient
T.-N.-L.	1,5 %	5,4 %	3,6
Î.-P.-É.	1,5 %	8,0 %	5,3
N.-É.	1,2 %	3,9 %	3,3
N.-B.	1,3 %	3,8 %	2,9
Qc	0,6 %	1,4 %	2,3
Ont.	0,5 %	1,0 %	2,0
Man.	0,9 %	3,7 %	4,1
Sask.	1,1 %	4,8 %	4,4
Alb.	0,8 %	2,1 %	2,6
C.-B.	0,8 %	1,8 %	2,3

Pour estimer l'EQM des estimations réconciliées, nous avons tout d'abord essayé d'appliquer une méthode de bootstrap paramétrique qui suppose que les erreurs d'échantillonnage sont indépendantes ($\mathbf{e} \sim N(\mathbf{0}, \Psi)$, avec Ψ diagonale). Les étapes suivantes ont été répétées pour chaque réplique de la méthode bootstrap :

1. Générer des valeurs de recensement : $\theta_{m*} \sim_{\text{ind}} N(\mathbf{x}_m^T \hat{\boldsymbol{\beta}}, N_m^2 \hat{\sigma}^2)$
2. Générer des estimations directes : $\hat{\theta}_{m*} \sim_{\text{ind}} N(\theta_{m*}, \hat{\psi}_m)$
3. Appliquer l'estimation sur petits domaines du modèle de FH de base.
4. Appliquer la réconciliation aux estimations résultant de l'étape 3 à l'estimation directe provinciale bootstrap correspondante (dérivée de l'agrégation des valeurs d'estimation directe bootstrap appropriées $\hat{\theta}_{m*}$).

L'EQM des agrégats provinciaux $\hat{\theta}_m^{\text{EPD, réconciliée}}$ est théoriquement égale à la variance de l'estimation directe provinciale en raison de la réconciliation. Les estimations de la variance du bootstrap paramétrique correspondantes devraient donc être proches des estimations de la variance des estimations provinciales directes. Lorsque l'on fait

l'hypothèse d'indépendance, nous constatons de grandes différences, comme le montre le tableau 3-2, en particulier chez les plus petites provinces. Ainsi, contrairement au cas des estimations non-réconciliées, il est important de ne pas supposer que Ψ est diagonale pour les estimations réconciliées. En outre, le danger est que les estimations de la REQMR des estimations sur petits domaines réconciliées risquent également d'être gonflées.

Avec les coefficients d'altérabilité choisis (et en supposant que la REQMR des estimations composites est connue plutôt qu'estimée), on peut démontrer que le premier terme de l'EQM des estimations réconciliées est :

$$g_{1m}^{\text{réconcilié}} = \gamma_m^2 \psi_m + N_m^2 (1 - \gamma_m)^2 \sigma^2 + \left(\frac{\sqrt{\varphi_m}}{\sum_{j \in \Omega_m} \sqrt{\varphi_j}} \right)^2 \sum_{j \in \Omega_m} N_j^2 (1 - \gamma_j) \psi_j \\ + 2N_m \gamma_m \frac{\sqrt{\varphi_m}}{\sum_j \sqrt{\varphi_j}} \sum_{j \neq m, j \in \Omega_m} N_j (1 - \gamma_j) \psi_{jm} + \left(\frac{\sqrt{\varphi_m}}{\sum_j \sqrt{\varphi_j}} \right)^2 \sum_{j \neq k; j, k \in \Omega_m} N_j N_k (1 - \gamma_j)(1 - \gamma_k) \psi_{jk},$$

où Ω_m est l'ensemble des domaines situés dans la même province que m . Les deux premiers termes de l'équation sont ceux associés à l'EBLUP. Le troisième terme est toujours positif et représente la variance supplémentaire due à la réconciliation si les erreurs d'échantillonnage étaient indépendantes. Les deux derniers termes sont négatifs si les erreurs d'échantillonnage sont négativement corrélées et expliquent en partie la surestimation observée dans notre application de la méthode bootstrap paramétrique naïve/standard.

La covariance d'échantillonnage présente dans les estimations directes de l'EPA est due aux procédures d'estimation. Les strates de l'EPA sont nombreuses et petites, ce qui garantirait l'indépendance des estimateurs de Horvitz-Thompson (c'est-à-dire pondérés par le poids de sondage). Cependant, les estimateurs finaux calés de l'EPA sont dépendants, car l'ajustement par calage des poids de l'EPA est effectué au niveau provincial. La dépendance est forte pour le total de l'emploi parce qu'il est relativement proche de certains totaux de calage (c'est-à-dire le nombre de personnes âgées de 15 ans et plus par province). Le modèle de FH étendu aux erreurs d'échantillonnage dépendantes devrait donc être ajusté à la place. Le défi consiste à obtenir les termes de covariance de Ψ en plus des termes de variance compte tenu de la petite taille des échantillons.

4. Ingénierie inverse d'un processus de réconciliation hypothétique pour obtenir la matrice complète de variance des erreurs d'échantillonnage

Nous avons tenté tout d'abord de supposer que toutes les estimations directes au niveau des domaines dans la même province sont également corrélées. Cependant, la matrice résultante n'est pas nécessairement une matrice de covariance appropriée car elle peut ne pas être semi-définie positive (elle ne l'a jamais été dans notre application). La décomposition spectrale de cette matrice et le remplacement des valeurs propres négatives par 0 ont permis d'obtenir une matrice semi-définie positive, mais au détriment des termes de diagonale/variance qui ne sont pas restés intacts.

Les propriétés souhaitées de la matrice $\tilde{\Psi}$ sont les suivantes :

1. la diagonale devrait correspondre aux estimations de la variance directe lissée $\tilde{\psi}_m$;
2. la structure de covariance devrait respecter la structure de covariance de certains totaux agrégés. Dans le cas qui nous préoccupe, nous respecterons la structure de variance-covariance des estimations provinciales de l'EPA, puisque le calage des poids de l'EPA et la réconciliation des estimations du modèle de FH sont effectués à ce niveau;
3. la matrice devrait être une matrice de covariance appropriée. Elle devrait notamment être symétrique et semi-définie positive.

L'idée clé pour construire une telle matrice est que des matrices de variance apparaissent naturellement comme le résultat d'un processus statistique. Inspirés par l'échantillonnage de l'EPA et le calage provincial des poids de l'EPA, nous supposons donc que nos estimations directes résultent de la réconciliation de variables indépendantes hypothétiques au niveau provincial et nous utilisons la matrice de variance de ces estimations réconciliées pour définir $\tilde{\Psi}$.

Soient $X_m \sim_{\text{ind}}(\mu_{X,m}, \sigma_{X,m}^2)$, $m = 1, \dots, M$ et $\underline{Y} \sim (\mu_Y, \Sigma_Y)$, des ensembles de variables indépendantes l'un de l'autre et respectivement les variables normales hypothétiques indépendantes au niveau des petits domaines et le vecteur de totaux agrégés pour lesquels nous disposons d'une bonne estimation de leur matrice de covariance (par exemple, des estimations provinciales). Nous présumons que nos estimations directes $\hat{\theta}$ sont le résultat de la réconciliation de $\underline{X}$ avec $\underline{Y}$. À savoir, lors de la réconciliation, nous avons obtenu $\hat{\theta}$ en minimisant

$$\sum_{m=1}^M \frac{(X_m - \hat{\theta}_m)^2}{|c_m X_m|}$$

sous la contrainte $G\hat{\theta} = \underline{Y}$. La solution explicite au problème de la réconciliation (c'est-à-dire le vecteur des estimations directes) est donnée par

$$\hat{\theta} = \underline{X} + V_e G^T (G V_e G^T)^{-1} (\underline{Y} - G \underline{X}),$$

où $V_e = \text{diag}(|c_1 X_1|, \dots, |c_M X_M|)$.

Il faut choisir le coefficient d'altérabilité c_m de la réconciliation hypothétique. En général, lors de la réconciliation, on n'est pas limité dans ce choix. Toutefois, dans la situation présente, on voudra utiliser une valeur qui garantisse que le problème d'ingénierie inverse puisse être résolu. Dans le contexte habituel de la réconciliation, on définit le plus souvent $c_m = \sqrt{V(X_m)}/X_m = \text{CV}(X_m)$, mais pour simplifier la résolution des équations qui s'ensuivent, nous avons défini $c_m = \sqrt{V(\hat{\theta}_m)}/X_m$. Ainsi $V_e = \text{diag}(\sqrt{\tilde{\psi}_1}, \dots, \sqrt{\tilde{\psi}_M})$.

Tableau 4-1

Différences provinciales entre l'estimation directe de l'emploi total et l'agrégation des estimations sur petits domaines en ne supposant pas des erreurs d'échantillonnage indépendantes pour le mois de novembre 2021

Province	Est. directe	Est. sur petits domaines	Différence	Différence relative (%)	Différence relative (%) du tableau 3-1 lorsqu'on a supposé l'indépendance
T.-N.-L.	229 601	231 945	-2 344	-1,0	2,2
Î.-P.-É.	81 098	78 809	2 289	2,8	9,7
N.-É.	477 326	482 921	-5 595	-1,2	-1,4
N.-B.	360 082	363 074	-2 992	-0,8	-0,1
Qc	4 380 216	4 396 197	-15 981	-0,4	0,4
Ont.	7 660 158	7 722 312	-62 154	-0,8	-1,5
Man.	669 344	667 764	1 580	0,2	2,5
Sask.	556 206	546 322	9 884	1,8	8,4
Alb.	2 298 973	2 295 764	3 209	0,1	0,9
C.-B.	2 708 386	2 722 514	-14 128	-0,5	0,3
Canada	19 421 390	19 507 622	-86 232	-0,4	0,0

Nous avons également

$$\begin{aligned} \tilde{\Psi} &= V(\hat{\theta}) = V(\underline{X} + V_e G^T (G V_e G^T)^{-1} (\underline{Y} - G \underline{X})) \\ &= V([I - V_e G^T (G V_e G^T)^{-1} G] \underline{X} + V_e G^T (G V_e G^T)^{-1} \underline{Y}) \\ &= V(A \underline{X} + C \underline{Y}) \\ &= \text{Adiag}(\sigma_{X,1}^2, \dots, \sigma_{X,M}^2) A^T + C \Sigma_Y C^T. \end{aligned}$$

Les valeurs de la diagonale de $V(\hat{\theta})$ sont connues, ainsi que A et Σ_Y . En utilisant les éléments diagonaux de la dernière équation, on peut donc résoudre les équations pour trouver les valeurs de $\sigma_{X,1}^2, \dots, \sigma_{X,M}^2$. Désignons A^2 comme la matrice des éléments de A mis au carré. En utilisant la théorie de la régression linéaire, nous pouvons montrer que pour minimiser les changements des éléments diagonaux de $V(\hat{\theta})$, nous devrions choisir :

$$\underline{\sigma}_X^2 = (A^{2T} A^2)^{-1} A^{2T} [D(V(\hat{\theta})) - D(C \Sigma_Y C^T)],$$

où D est l'opérateur qui prend la diagonale d'une matrice et la transforme en vecteur. Ces valeurs peuvent ensuite être utilisées dans l'équation précédente pour obtenir la matrice $\tilde{\Psi}$ complète. La matrice est, de par sa construction, une matrice de covariance proprement dite. Elle possède par ailleurs la diagonale souhaitée (ou minimise les modifications de celle-ci), ainsi que la matrice de covariance intégrée Σ_Y des agrégats ciblés.

En utilisant cette matrice de covariance dans l'estimation ponctuelle, la nécessité de procéder à une réconciliation n'est pas aussi frappante tel que le montre le tableau 4-1, même si les différences demeurent cohérentes d'un mois à l'autre (non présentées). En l'utilisant dans les équations de l'EQM, nous obtenons naturellement des REQMR pour les agrégats provinciaux d'estimations du modèle de FH obtenues par réconciliation qui concordent avec les c.v. des estimations provinciales directes, comme le montre le tableau 4-2. Plus important encore, les indicateurs de qualité des estimations sur petits domaines réconciliés ne sont plus artificiellement gonflés, ce qui est la principale préoccupation.

Tableau 4-2

REQMR de l'agrégation des estimations sur petits domaines et REQMR moyenne des estimations sur petits domaines en supposant l'indépendance et la non-indépendance pour le mois de novembre 2021

Province	C.v. de l'estimation provinciale directe	REQMR des agrégats des estimations réconciliées		REQMR moyenne des estimations réconciliées	
		en supposant que Ψ est diagonale	en ne supposant pas que Ψ est diagonale	en supposant que Ψ est diagonale	en ne supposant pas que Ψ est diagonale
T.-N.-L.	1,5 %	5,4 %	1,5 %	14,5 %	9,4 %
Î.-P.-É.	1,5 %	8,0 %	1,5 %	16,0 %	7,0 %
N.-É.	1,2 %	3,9 %	1,2 %	8,8 %	6,3 %
N.-B.	1,3 %	3,8 %	1,3 %	10,2 %	7,3 %
Qc	0,6 %	1,4 %	0,6 %	8,9 %	7,1 %
Ont.	0,5 %	1,0 %	0,5 %	7,0 %	6,1 %
Man.	0,9 %	3,7 %	0,9 %	12,7 %	7,0 %
Sask.	1,1 %	4,8 %	1,1 %	12,7 %	6,8 %
Alb.	0,8 %	2,1 %	0,8 %	7,8 %	6,5 %
C.-B.	0,8 %	1,8 %	0,8 %	9,2 %	7,3 %

5. Conclusion

Lorsque la covariance des erreurs d'échantillonnage est forte, il convient d'utiliser le modèle de FH étendu pour les erreurs d'échantillonnage non indépendantes au lieu du modèle de FH de base. L'utilisation d'une structure de covariance appropriée est essentielle pour obtenir les véritables EBLUP et pour avoir des composantes de l'EQM qui ont le comportement décrit dans Prasad et Rao (1990) (c.-à-d. G_1 étant le terme principal de l'EQM si le nombre de domaines est suffisamment élevé). L'une des données clés est la matrice de variance des erreurs d'échantillonnage Ψ . Nous avons démontré la façon d'effectuer une ingénierie inverse d'un processus de réconciliation hypothétique pour obtenir une matrice de variance de travail appropriée qui présente les variances souhaitées au niveau des petits domaines et de certains niveaux agrégés.

Bien que les estimations de l'EQM soient améliorées par rapport à celles obtenues sous l'hypothèse d'indépendance, leur qualité dépend de la qualité de l'approximation des termes de covariance dans Ψ . Ce résultat peut être amélioré en ajoutant des contraintes dans la réconciliation hypothétique (par exemple, pour les petits domaines avec une taille d'échantillon ou une fraction d'échantillonnage suffisamment grande). L'effet des divergences pourrait être étudié dans un contexte de simulation.

Remerciements

Nous tenons à remercier Jean-François Beaumont, Keven Bosa et Cynthia Bocci pour leur soutien et leur contribution.

Bibliographie

- Beaumont, J.-F., et Bocci, C. (2016), “Small Area Estimation in the Labour Force Survey”, Article présenté au Comité consultatif sur les méthodes statistiques de Statistique Canada, 31 mars 2016.
- Dagum, E.B., et P. Cholette (2006), *Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series*, New York: Springer-Verlag, Lecture Notes in Statistics, 186.
- Estevao, V., You, Y., Hidioglou, M., Beaumont, J.-F. et Rubin-Bleuer, S. (2023), “Small Area Estimation-Area Level Model with EBLUP Estimation- Methodology Specifications”, Document de Statistique Canada.
- Isaki, C.T., Tsay, J.H. et Fuller, W.A. (2000), “Estimation des facteurs de correction au recensement”, *Techniques d’enquête*, 26, p. 31-42.
- OECD (2020), *Delineating Functional Areas in All Territories*, OECD Territorial Reviews, OECD Publishing, Paris, <https://doi.org/10.1787/07970966-en>.
- Prasad, N.G.N. et Rao, J.N.K. (1990), “The Estimation of the Mean Squared Error of Small Area Estimators.” *Journal of the American Statistical Association*, 85, p. 163-171.
- Rao, J.N.K., et Molina, I. (2015), *Small Area Estimation*. New York : John Wiley & Sons, Inc.
- Rao, J.N.K., et Wu, C.F.J. (1988), “Resampling inference with complex survey data”, *Journal of the American Statistical Association*, 83, p. 231-241.
- Statistique Canada (2020). Guide de l'Enquête sur la population active, Catalogue no. 71-543-G. <https://www150.statcan.gc.ca/n1/pub/71-543-g/71-543-g2020001-fra.htm>.

Annexe

L'EQM de l'estimateur composite réconcilié peut être calculée en suivant les étapes d'Isaki, Tsay et Fuller (2000) :

1. Exprimons $\hat{\theta}^{\text{réconcilié}} - \theta$ comme une fonction de $\hat{\beta}$, de $\hat{\sigma}^2$, de v et de e
2. Linéarisons le résultat en termes de $\hat{\beta}$ et $\hat{\sigma}^2$ et exprimons-le comme une fonction de e et de $z = Bv + e$
3. Supposons que $\hat{\sigma}^2$ et que z sont indépendants et dérivons la variance approximative de $\hat{\theta}^{\text{réconcilié}} - \theta$

C'est ce qui sera fait par la suite, en supposant que la matrice V_e liée aux coefficients d'altérabilité de la réconciliation n'est pas aléatoire. L'EQM de l'estimateur composite (non réconcilié) correspond à $A = I$ dans les dérivations, où $A = I - V_e G^T (G V_e G^T)^{-1} G$.

Étape 1 : Exprimons $\hat{\theta}^{\text{réconcilié}} - \theta$ comme une fonction de $\hat{\beta}$, de $\hat{\sigma}^2$, de v et de e

$$\begin{aligned}\hat{\theta}^{\text{réconcilié}} &= \hat{\theta}^{\text{EPD}} + V_e G^T (G V_e G^T)^{-1} G [\hat{\theta} - \hat{\theta}^{\text{EPD}}] \\ &= A \hat{\theta}^{\text{EPD}} + (I - A) \hat{\theta} \\ &= A \hat{H}^T X \hat{\beta} + (I - A \hat{H}^T) (X \beta + Bv + e).\end{aligned}$$

Par conséquent,

$$\begin{aligned}\hat{\theta}^{\text{réconcilié}} - \theta &= A \hat{H}^T X \hat{\beta} + (I - A \hat{H}^T) (X \beta + Bv + e) - (X \beta + Bv) \\ &= A \hat{H}^T [X(\hat{\beta} - \beta) - z] + e.\end{aligned}$$

Étape 2 : Linéarisons le résultat en termes de $\hat{\beta}$ et $\hat{\sigma}^2$ et exprimons-le comme une fonction de e et de $z = Bv + e$

Le terme constant de la linéarisation de Taylor est $e - A H^T z$, où $H^T = I - \sigma^2 B^2 \Sigma_{zz}^{-1}$.

Le terme en $(\hat{\beta} - \beta)$ est $\left. \frac{\partial \hat{\theta}^{\text{réconcilié}} - \theta}{\partial \hat{\beta}} \right|_{\hat{\beta}=\beta, \hat{\sigma}^2=\sigma^2} (\hat{\beta} - \beta) = A H^T X (\hat{\beta} - \beta)$.

Le terme en $(\hat{\sigma}^2 - \sigma^2)$ est

$$\begin{aligned}\left. \frac{\partial \hat{\theta}^{\text{réconcilié}} - \theta}{\partial \hat{\sigma}^2} \right|_{\hat{\beta}=\beta, \hat{\sigma}^2=\sigma^2} (\hat{\sigma}^2 - \sigma^2) &= -A \left. \frac{\partial \hat{H}^T}{\partial \hat{\sigma}^2} \right|_{\hat{\sigma}^2=\sigma^2} z (\hat{\sigma}^2 - \sigma^2) \\ &= A H^T B^2 \Sigma_{zz}^{-1} z (\hat{\sigma}^2 - \sigma^2).\end{aligned}$$

Nous avons donc $\hat{\theta}^{\text{réconcilié}} - \theta \cong e - A H^T z + A H^T X (\hat{\beta} - \beta) + A H^T B^2 \Sigma_{zz}^{-1} z (\hat{\sigma}^2 - \sigma^2)$.

Étape 3 : Supposons que $\hat{\sigma}^2$ et que z sont indépendants et dérivons la variance approximative de $\hat{\theta}^{\text{réconcilié}} - \theta$

$$\begin{aligned}G_1(\sigma^2) &= V(e - A H^T z) \\ &= \Psi - \Psi \Sigma_{zz}^{-1} \Psi + (I - A) \Psi \Sigma_{zz}^{-1} \Psi (I - A)^T \\ G_2(\sigma^2) &= A H^T X V(\hat{\beta}) X^T H A^T \\ &= A H^T X (X^T \Sigma_{zz}^{-1} X)^{-1} X^T H A^T \\ G_3(\sigma^2) &= A H^T B^2 \Sigma_{zz}^{-1} B^2 H A^T V_{\sigma\sigma}.\end{aligned}$$

Si l'on obtient $\hat{\sigma}^2$ à l'aide du MVC, alors $V_{\sigma\sigma} = \frac{2}{\text{trace}(P^2)}$, où $P = \Sigma_{zz}^{-1} - \Sigma_{zz}^{-1} X (X^T \Sigma_{zz}^{-1} X)^{-1} X^T \Sigma_{zz}^{-1}$ (Rao et Molina, 2015).

On remarque que puisque $GA = 0$ nous avons comme prévu $V\left(G(\hat{\theta}^{\text{réconcilié}} - \theta)\right) = G\Psi G^T + 0 + 0$, c'est-à-dire la matrice de la variance sous le plan des totaux de la réconciliation. Des dérivations similaires lorsque l'indépendance de l'erreur d'échantillonnage est supposée à tort donnent

$$G_1(\sigma^2) = (I - AH_d^T)\Psi(I - AH_d^T)^T + \sigma^2 AH_d^T H_d^{\square\square\square} A^T,$$

où $H_d^T = I - \sigma^2 B^2 D_{zz}^{-1} = \text{diag}(1 - \gamma_1, \dots, 1 - \gamma_M)$ et D_{zz} est une matrice diagonale avec les mêmes éléments diagonaux que $\sigma^2 B^2 + \Psi$. Sans réconciliation (c.-à-d. lorsque $A = I$), nous avons ainsi $g_{1m}(\sigma^2) = \gamma_m^2 \psi_m + (1 - \gamma_m)^2 N_m^2 \sigma^2$, comme lorsque les erreurs d'échantillonnage sont supposées indépendantes à juste titre.