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

Techniques d'enquête

# Élaboration d'un système d'estimation sur petits domaines à Statistique Canada


par Michel A. Hidiroglou, Jean-François Beaumont
et Wesley Yung

Date de diffusion : le 7 mai 2019

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** STATCAN.infostats-infostats.STATCAN@canada.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

**Programme des services de dépôt**

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

## 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.

Publication autorisée par le ministre responsable de Statistique Canada

© Sa Majesté la Reine du chef du Canada, représentée par le ministre de l'Industrie 2019

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada.

**Une version HTML est aussi disponible.**

*This publication is also available in English*.

Techniques d'enquête, 2019 (numéro spécial)
Vol. 45, nº 1, pp. 107-133
Statistique Canada, nº 12-001-X au catalogue


# Élaboration d'un système d'estimation sur petits domaines à Statistique Canada

**Michel A. Hidiroglou, Jean-François Beaumont et Wesley Yung**[1]

Résumé

La demande d'estimations sur petits domaines de la part des utilisateurs des données de Statistique Canada augmente constamment depuis quelques années. Dans le présent document, nous résumons les procédures qui ont été intégrées dans un système de production en SAS permettant d'obtenir des estimations sur petits domaines officielles à Statistique Canada. Ce système comprend : des procédures fondées sur des modèles au niveau de l'unité ou du domaine; l'intégration du plan d'échantillonnage; la capacité de lisser la variance sous le plan pour chaque petit domaine si un modèle au niveau du domaine est utilisé; la capacité de vérifier que les estimations sur petits domaines équivalent à des estimations fiables de niveau plus élevé; et l'élaboration d'outils de diagnostic pour tester la pertinence du modèle. Le système de production a servi à produire des estimations sur petits domaines à titre expérimental pour plusieurs enquêtes de Statistique Canada, notamment : l'estimation des caractéristiques de la santé, l'estimation du sous-dénombrement au recensement, l'estimation des ventes des fabricants et l'estimation des taux de chômage et des chiffres d'emploi pour l'Enquête sur la population active. Certains des diagnostics instaurés dans le système sont illustrés à l'aide des données de l'Enquête sur la population active ainsi que des données administratives auxiliaires.

**Mots-clés :** Estimation sur petits domaines; modèle au niveau du domaine; modèle au niveau de l'unité; EBLUP; méthodes hiérarchiques bayésiennes; statistiques officielles.

## 1 Introduction

De nos jours, les utilisateurs de données sont de plus en plus avertis et ils demandent plus de données et des données à des niveaux plus détaillés. Pour les organismes nationaux de statistique (ONS) qui doivent faire face à la baisse des taux de réponse, la production de données plus détaillées est un défi particulièrement difficile à relever. Les techniques d'estimation sur petits domaines constituent un moyen à envisager pour répondre à cette demande afin de produire des estimations pour des sous-populations ou des petits domaines précis. Un *petit domaine* désigne un sous-groupe de la population pour lequel la taille de l'échantillon est si petite que les estimations directes ne sont pas suffisamment fiables pour être publiées. Les petits domaines comprennent, par exemple, une région géographique (par exemple une province, un comté, une municipalité, etc.), un groupe démographique (par exemple l'âge selon le sexe), un groupe démographique au sein d'une région géographique ou un groupe d'industries détaillé. La demande de données de petits domaines est reconnue depuis des années (voir Brackstone, 1987), mais elle a beaucoup augmenté récemment, comme l'indique le rapport du printemps 2014 du vérificateur général du Canada.

L'étude des procédures d'estimation sur petits domaines ne date pas d'hier à Statistique Canada, car elle a débuté dans les années 1970 avec Singh et Tessier (1976) et Ghangurde et Singh (1977). Drew, Singh et Choudhry (1982) ont proposé une procédure dépendante de l'échantillon pour estimer les caractéristiques de l'emploi sous le niveau provincial. Dick (1995) a modélisé le sous-dénombrement net pour le

1. Michel A. Hidiroglou, Méthodes d'enquêtes auprès des entreprises, Statistique Canada, 22e étage, immeuble R.-H.-Coats, Ottawa (Ontario) K1A 0T6, Canada. Courriel : hidirog@yahoo.ca; Jean-François Beaumont, Division de la coopération internationale et des méthodes statistiques institutionnelles, Statistique Canada, 25e étage, immeuble R.-H.-Coats, Ottawa (Ontario) K1A 0T6, Canada; Wesley Yung, Méthodes d'enquêtes auprès des entreprises, Statistique Canada, 22e étage, immeuble R.-H.-Coats, Ottawa (Ontario) K1A 0T6, Canada.

Recensement de la population canadienne de 1991. L'élaboration d'un système d'estimation sur petits domaines adapté à Statistique Canada arrive au bon moment, car un grand nombre de documents ont été écrits sur le sujet, y compris les ouvrages de Rao (2003) et de Rao et Molina (2015).

Les quatre documents qui ont eu une grande incidence sur l'estimation sur petits domaines (EPD) sont ceux de Gonzalez et Hoza (1978), Fay et Herriot (1979), Battese, Harter et Fuller (1988), et Prasad et Rao (1990). Gonzalez et Hoza (1978) ont été parmi les premiers à proposer des procédures d'estimation sur petits domaines (surtout des méthodes d'estimation synthétique). Fay et Herriot (1979) ont élaboré des procédures pour estimer le revenu pour des petits domaines à l'aide des données du questionnaire détaillé du recensement. Cette méthode et ses variantes comptent parmi les procédures les plus largement utilisées pour produire des estimations sur petits domaines en intégrant des données auxiliaires et des estimations d'enquête directes. Battese et coll. (1988) ont conçu une procédure sur petits domaines pour estimer les superficies en culture à l'aide des données d'enquête et des données satellitaires disponibles pour les unités individuelles. Enfin, Prasad et Rao (1990) ont calculé un estimateur relativement sans biais de l'erreur quadratique moyenne fondé sur le modèle des estimateurs de Fay-Herriot et de Battese-Harter-Fuller.

La théorie statistique de l'EPD fondée sur des modèles est assez complexe et une grande partie des logiciels disponibles dans les bureaux de statistique nationaux ont été programmés de façon ponctuelle et, à ce titre, ne conviennent pas à un environnement de production. On a donc décidé de développer un système qui serait pratique à la fois comme outil de production et comme outil d'apprentissage pour les employés. Au moment où cette décision a été prise, vers 2006, il existait des programmes informatiques développés par le projet EURAREA (2004) pour l'estimation sur petits domaines. Toutefois, cet ensemble de programmes n'était plus en mode de développement et ne représentait pas les dernières percées en matière d'estimation sur petits domaines. Par conséquent, un système souple d'estimation sur petits domaines pouvant répondre aux besoins de production d'estimations sur petits domaines en production a été élaboré à Statistique Canada. Les exigences fondamentales de ce système sur petits domaines comprennent les suivantes : permettre des modèles au niveau de l'unité et du domaine; intégrer le plan d'échantillonnage dans l'estimation des paramètres d'intérêt et l'erreur quadratique moyenne; s'assurer que les estimations sur petits domaines correspondent aux estimations fiables des niveaux plus élevés (c'est-à-dire totaux), et concevoir des outils de diagnostic pour tester la pertinence des modèles employés pour l'estimation sur petits domaines. Estevao, Hidiroglou et You (2015) ont donc développé un système prototype, rédigé en SAS, afin de remplir ces exigences. Ce prototype a été transformé en un système de production que Statistique Canada utilise actuellement.

Le document est organisé de la manière suivante. Dans la section 2, nous présentons la notation employée dans l'article. Dans la section 3, nous discutons des options qu'offre le système de production pour le modèle au niveau du domaine et les méthodes de la meilleure prédiction linéaire sans biais (méthodes EBLUP). Les options pour le modèle au niveau de l'unité à l'aide des méthodes EBLUP sont présentées à la section 4. L'approche hiérarchique bayésienne est présentée à la section 5 pour le modèle au niveau du

domaine. La section 6 illustre le système de production au moyen de l'Enquête sur la population active (EPA) de Statistique Canada. Enfin, certaines conclusions sont présentées à la section 7.

# 2 Notation de base et contexte

Nous présentons d'abord une notation qui définira les divers estimateurs sur petits domaines inclus dans le système de production. Supposons que $U$ représente une population de taille $N$. Cette population est fractionnée en $M$ domaines mutuellement exclusifs et exhaustifs, où chaque domaine $U_i \subset U,\ i = 1, \ldots, M$ a $N_i$ observations. Un échantillon, $s$, de taille $n$ provient de la population à l'aide d'un mécanisme de probabilité $p(s)$ bien défini et l'échantillon ainsi obtenu est divisé en domaines $s_i = s \cap U_i,\ i = 1, \ldots, M$. Il convient de souligner que, pour certains domaines, la taille d'échantillon réalisé $n_i$ peut être de zéro. L'ensemble de $m (m \leq M)$ domaines, où $n_i$ est strictement supérieur à 0, est représentée par $A$. L'ensemble des autres domaines, où $n_i$ est égal à 0, est représenté par $\bar{A}$.

Supposons que $\pi_j = \sum_{\{s:\ j \in s\}} p(s),\ j \in U,$ représente les probabilités d'inclusion où $\{s : j \in s\}$ désigne la sommation de tous les échantillons $s$ qui renferment l'unité $j$. Nous représentons le poids d'échantillonnage de l'unité $j$ par $d_j$, où $d_j = \pi_j^{-1}$. Le poids final associé à l'unité $j$ est représenté par $w_j$. Ce poids est habituellement le produit du poids déterminé par le plan d'échantillonnage original $\left(d_j\right)$ multiplié par un facteur d'ajustement qui représente l'intégration des données auxiliaires disponibles (à l'aide de la régression ou du calage), ainsi que des ajustements pour la non-réponse. À noter que les données auxiliaires utilisées dans le facteur d'ajustement ne sont peut-être pas identiques à celles qui sont employés dans l'estimation sur petits domaines.

L'objectif d'un système d'estimation sur petits domaines est d'estimer un paramètre de population $\theta_i$ (par exemple une moyenne ou un total) pour chaque domaine $i$ d'une variable d'intérêt donnée $y$ lorsque la taille de l'échantillon de certains domaines $n_i$ est trop petite pour qu'on puisse avoir recours à des procédures d'*estimation directe*. Un *estimateur direct* de $\theta_i$ est un estimateur qui utilise les valeurs de la variable d'intérêt, $y$, uniquement dans les unités d'échantillon du domaine $i$. Toutefois, un inconvénient important de ce genre d'estimateurs est qu'ils peuvent produire des erreurs-types inacceptablement grandes, surtout si la taille d'échantillon dans le domaine est petite. Les procédures sur petits domaines utilisent des *estimateurs indirects* qui se renforcent entre les domaines, en se servant de modèles qui relient tous les domaines grâce à certains paramètres communs. Les estimateurs indirects sont efficaces (c'est-à-dire ils augmentent la taille réelle de l'échantillon et diminuent ainsi l'erreur-type) si le modèle vaut pour chaque domaine. Les écarts par rapport au modèle réduisent la précision. Il existe une grande variété d'estimateurs indirects et un bon résumé est fourni dans Rao et Molina (2015).

Les estimateurs sur petits domaines sont classés au niveau du domaine ou de l'unité selon le niveau auquel la modélisation est réalisée. Les estimateurs sur petits domaines au niveau du domaine sont fondés sur des modèles qui établissent un lien entre un paramètre d'intérêt donné et des variables auxiliaires propres au domaine. Les estimateurs sur petits domaines au niveau de l'unité reposent sur des modèles qui établissent un lien entre la variable d'intérêt et les variables auxiliaires propres à l'unité. Les estimateurs sur

petits domaines au niveau du domaine sont calculés si les données sur le domaine au niveau de l'unité ne sont pas disponibles. Ils peuvent également être calculés si les données au niveau de l'unité sont disponibles quand elles sont regroupées au niveau du domaine approprié. Cela pourrait être utile en pratique parce que les estimateurs sur petits domaines au niveau du domaine peuvent être moins susceptibles de donner des valeurs aberrantes que leurs homologues au niveau de l'unité.

# 3 Modèle au niveau du domaine

L'estimateur de petits domaines au niveau du domaine est apparu pour la première fois dans le texte fondateur de Fay et Herriot (1979). Pour donner suite à ce document, supposons que le paramètre d'intérêt est $\theta_i$; des exemples communs sont des totaux, $Y_i = \sum_{j \in U_i} y_j$, ou des moyennes, $\bar{Y}_i = Y_i / N_i$. Tel que susmentionné, le vecteur des variables auxiliaires peut être différent de celui qui est utilisé dans l'estimation directe et désigné par $\mathbf{z}$. Le modèle au niveau du domaine peut s'exprimer à l'aide de deux équations.

La première équation, communément appelée le *modèle d'échantillonnage*, est obtenue comme suit :

$$\hat{\theta}_i = \theta_i + e_i \tag{3.1}$$

et exprime l'estimation directe $\hat{\theta}_i$ en fonction du paramètre inconnu $\theta_i$, plus une erreur aléatoire $e_i$ attribuable à l'échantillonnage. Les erreurs d'échantillonnage $e_i$ sont des erreurs indépendantes et identiquement distribuées de moyenne 0 et de variance $\psi_i$, c'est-à-dire $E_p(e_i \mid \theta_i) = 0$ et $V_p(e_i \mid \theta_i) = \psi_i$, où $p$ désigne l'espérance relative au plan d'échantillonnage. À noter que $\psi_i$ est aussi la variance sous le plan de $\hat{\theta}_i$ et qu'il est habituellement inconnu.

La seconde équation, connue comme le *modèle de lien*, est obtenue comme suit :

$$\theta_i = \mathbf{z}_i^T \boldsymbol{\beta} + b_i v_i \tag{3.2}$$

et exprime le paramètre $\theta_i$ comme un effet fixe $\mathbf{z}_i^T \boldsymbol{\beta}$, plus un effet aléatoire $v_i$ multiplié par $b_i$. Dans le système de production, le terme $b_i$ a une valeur par défaut de un, mais l'utilisateur peut le préciser pour contrôler les erreurs hétéroscédastiques ou l'impact des observations influentes. Les effets aléatoires $v_i$ sont indépendants et identiquement distribués de variance moyenne 0 et de variance du modèle inconnue $\sigma_v^2$, c'est-à-dire $E_m(v_i) = 0$ et $V_m(v_i) = \sigma_v^2$ où $E_m$ désigne l'espérance du modèle et $V_m$, la variance du modèle. Les erreurs aléatoires $e_i$ sont indépendantes des effets aléatoires $v_i$. La combinaison du *modèle d'échantillonnage* et du *modèle de lien* donne un seul modèle linéaire mixte généralisé (MLMG), obtenu comme ceci :

$$\hat{\theta}_i = \mathbf{z}_i^T \boldsymbol{\beta} + b_i v_i + e_i. \tag{3.3}$$

Dans le modèle de Fay-Herriot (3.3), nous constatons que $E_{mp}\left(\hat{\theta}_i\right) = \mathbf{z}_i^T \boldsymbol{\beta}$ et $V_{mp}\left(\hat{\theta}_i\right) = b_i^2 \sigma_v^2 + \tilde{\psi}_i$, où $\tilde{\psi}_i = E_m(\psi_i)$ est la variance sous le plan lisse de $\hat{\theta}_i$. En général, nous ne pouvons pas traiter $\psi_i$ comme étant fixe, parce qu'il n'est pas strictement une fonction des données auxiliaires. Si les $\sigma_v^2$ et $\tilde{\psi}_i$ sont connus, la solution du MLMG donne la meilleure estimation linéaire sans biais empirique (BLUP), $\tilde{\theta}_i^{\mathrm{BLUP}}$

$$\tilde{\theta}_i^{\text{BLUP}} = \begin{cases} \gamma_i \hat{\theta}_i + (1-\gamma_i)\,\mathbf{z}_i^T \tilde{\boldsymbol{\beta}} & \text{pour } i \in A \\ \mathbf{z}_i^T \tilde{\boldsymbol{\beta}} & \text{pour } i \in \bar{A} \end{cases} \tag{3.4}$$

où $\gamma_i = (b_i^2\sigma_v^2)/(\tilde{\psi}_i + b_i^2\sigma_v^2)$ et $\tilde{\boldsymbol{\beta}} = \left(\sum_{i\in A} \mathbf{z}_i \mathbf{z}_i^T \big/ (\tilde{\psi}_i + b_i^2\sigma_v^2)\right)^{-1} \sum_{i\in A} \mathbf{z}_i \hat{\theta}_i \big/ (\tilde{\psi}_i + b_i^2\sigma_v^2)$.

Il existe quatre procédures récursives pour l'estimation de $\sigma_v^2$ et $\boldsymbol{\beta}$ dans le système de production. Les trois premières supposent que $\tilde{\psi}_i$ est connu qu'une version lissée est disponible (voir les détails dans la section suivante). Selon cette hypothèse, les composantes de la variance peuvent être calculées au moyen de la procédure de Fay-Herriot (FH) décrite dans Fay et Herriot (1979), du maximum de vraisemblance restreint (REML) ou de la maximisation de la densité corrigée (ADM) attribuée à Li et Lahiri (2010). La quatrième procédure, WF, attribuée à Wang et Fuller (2003), suppose que nous estimons $\psi_i$ à l'aide de $\hat{\psi}_i$ en sachant que $n_i \geq 2$. La procédure WF ne requiert aucun lissage des valeurs estimées de $\hat{\psi}_i$ avant l'estimation de $\sigma_v^2$. Wang et Fuller (2003) ont effectué des simulations où $n_i$ allait de 9 à 36 et constaté que leur procédure donnait des estimations raisonnables de $\theta_i$ et de son erreur quadratique moyenne estimée.

La principale différence entre ces quatre procédures est la façon dont les $\sigma_v^2$ sont calculés. Ils reposent tous sur un algorithme de notation itératif qui donne $\hat{\sigma}_v^2$ comme une estimation de la variance du modèle $\sigma_v^2$. Les procédures FH, REML et WF peuvent donner des $\hat{\sigma}_v^2$ qui sont inférieurs à zéro. Si cela se produit, les $\hat{\sigma}_v^2$ sont établis à zéro pour les procédures FH et REML. La troncation de $\sigma_v^2$ estimé à zéro a pour inconvénient de rendre l'estimateur sur petits domaines synthétique pour tous les domaines. Li et Lahiri (2010) ont proposé l'ADM afin de régler le problème des $\hat{\sigma}_v^2$ négatifs en maximisant une probabilité corrigée définie comme le produit de la variance du modèle et d'une probabilité standard. Bien que la méthode de l'ADM donne toujours une solution positive pour $\sigma_v^2$, il faudrait l'utiliser avec prudence parce qu'elle surestime la variance du modèle. Les procédures REML, FH et ADM ont recours aux valeurs lissées des valeurs estimées $\hat{\psi}_i$ obtenues à partir de l'échantillon ou d'une estimation fournie par l'utilisateur. Pour la procédure WF, si $\hat{\sigma}_v^2 < 0$, Wang et Fuller (2003) ont suggéré de déterminer $\hat{\sigma}_v^2$ comme $0{,}5\sqrt{\hat{V}(\hat{\sigma}_v^2)}$, où

$$\hat{V}(\hat{\sigma}_v^2) = \sum_{i\in A} 2\kappa_i^2 \left[ (\hat{\psi}_i + b_i^2\hat{\sigma}_v^2)^2 + \frac{(\hat{\psi}_i)^2}{(n_i - 1)} \right]$$

et

$$\kappa_i = \frac{\left[ b_i^2\hat{\sigma}_v^2 + \dfrac{(n_i+1)}{(n_i-1)}\hat{\psi}_i \right]^{-1}}{\sum_{i\in A}\left[ b_i^2\hat{\sigma}_v^2 + \dfrac{(n_i+1)}{(n_i-1)}\hat{\psi}_i \right]^{-1}}.$$

L'ajout de $\hat{\sigma}_v^2$ et d'une estimation des $\tilde{\psi}_i$ à $\tilde{\theta}_i^{\text{BLUP}}$, défini à l'aide de l'équation (3.4), donne la meilleure estimation linéaire sans biais empirique (EBLUP), $\hat{\theta}_i^{\text{EBLUP}}$, qui est obtenue comme suit :

$$\hat{\theta}_i^{\text{EBLUP}} = \begin{cases} \hat{\gamma}_i \hat{\theta}_i + (1-\hat{\gamma}_i)\,\mathbf{z}_i^T \hat{\boldsymbol{\beta}} & \text{pour } i \in A \\ \mathbf{z}_i^T \hat{\boldsymbol{\beta}} & \text{pour } i \in \bar{A} \end{cases}$$

où $\hat{\gamma}_i = (b_i^2\hat{\sigma}_v^2)/(\ddot{\psi}_i + b_i^2\hat{\sigma}_v^2)$, $\hat{\boldsymbol{\beta}} = \left(\sum_{i\in A}\mathbf{z}_i\mathbf{z}_i^T/(\ddot{\psi}_i + b_i^2\hat{\sigma}_v^2)\right)^{-1}\sum_{i\in A}\mathbf{z}_i\hat{\theta}_i^{\mathrm{DIR}}/(\ddot{\psi}_i + b_i^2\hat{\sigma}_v^2)$, et $\ddot{\psi}_i$ est choisi selon la procédure employée. Pour les procédures REML, FH et ADM, les $\ddot{\psi}_i$ sont les valeurs lissées des valeurs estimées de $\hat{\psi}_i$ obtenues à partir de l'échantillon ou d'une estimation fournie par l'utilisateur. Pour la procédure WF, nous disons que $\ddot{\psi}_i = \hat{\psi}_i$. Si la variance du modèle estimée $b_i^2\hat{\sigma}_v^2$ est relativement petite comparativement à $\ddot{\psi}_i$, alors $\hat{\gamma}_i$ est petit et plus de poids est attaché à l'estimateur synthétique $\mathbf{z}_i^T\hat{\boldsymbol{\beta}}$. De même, plus de poids est attaché à l'estimateur direct, $\hat{\theta}_i$, si la variance sous le plan $\ddot{\psi}_i$ est relativement petite.

Les détails des calculs requis se trouvent dans les spécifications de la méthodologie applicable au système de production décrites dans Estevao et coll. (2015).

## 3.1 Estimation de la variance sous le plan lisse

La variance sous le plan, $\psi_i$, pourrait être utilisée comme estimateur de la variance sous le plan lisse $\tilde{\psi}_i = E_m(\psi_i)$ si elle était connue. Dans la plupart des cas, elle est inconnue. Pour contourner ce problème, on suppose qu'un estimateur de variance sous le plan sans biais $\hat{\psi}_i$ de $\psi_i$ est disponible, c'est-à-dire que $E_p(\hat{\psi}_i) = \psi_i$. Selon cette hypothèse, nous avons ceci :

$$E_{mp}(\hat{\psi}_i) = E_m(\psi_i) = \tilde{\psi}_i.$$

Un estimateur simple sans biais de la variance sous le plan lisse $\tilde{\psi}_i$ est $\hat{\psi}_i$. Toutefois, $\hat{\psi}_i$ peut être assez instable lorsque la taille de l'échantillon dans le domaine $i$ est petite. On obtient un estimateur plus efficace en modélisant $\hat{\psi}_i$ avec $\mathbf{z}_i$. Dick (1995) et Rivest et Belmonte (2000) ont envisagé des modèles de lissage obtenus comme ceci :

$$\log(\hat{\psi}_i) = \mathbf{x}_i^T\boldsymbol{\alpha} + \varepsilon_i,$$

où $\mathbf{x}_i$ est un vecteur des variables explicatives qui sont des fonctions de $\mathbf{z}_i$, $\boldsymbol{\alpha}$ est un vecteur inconnu des paramètres modèles qu'il faut estimer, et $\varepsilon_i$ est une erreur aléatoire où $E_{mp}(\varepsilon_i) = 0$ et la variance constante $\sigma_\varepsilon^2 = V_{mp}(\varepsilon_i)$. Nous supposons également que les erreurs $\varepsilon_i$ sont identiquement distribuées conditionnellement à $\mathbf{z}_i$, $i = 1, \ldots, m$. À partir de ce modèle, nous constatons que :

$$\tilde{\psi}_i = E_{mp}(\hat{\psi}_i) = \exp(\mathbf{x}_i^T\boldsymbol{\alpha})\Delta,$$

où $\Delta = E_{mp}(\exp(\varepsilon_i))$. Dick (1995) a estimé $\tilde{\psi}_i$ en omettant le facteur $\Delta$. Rivest et Belmonte (2000) ont estimé $\Delta$ en supposant que les erreurs $\varepsilon_i$ étaient normalement distribuées. Cependant, nous avons observé de manière empirique que cet estimateur de $\Delta$ était sensible aux écarts par rapport à l'hypothèse de normalité. On évite cette hypothèse en utilisant une méthode des moments (voir Beaumont et Bocci, 2016). Cela donne l'estimateur sans biais de $\Delta$ :

$$\hat{\Delta}(\boldsymbol{\alpha}) = \frac{\sum_{i=1}^m \hat{\psi}_i}{\sum_{i=1}^m \exp(\mathbf{x}_i^T\boldsymbol{\alpha})}.$$

Il faut un estimateur $\hat{\boldsymbol{\alpha}}$ du vecteur des paramètres inconnus du modèle $\boldsymbol{\alpha}$ pour estimer $\tilde{\psi}_i$. On l'obtient à l'aide de la méthode des moindres carrés ordinaires comme suit :

$$\hat{\boldsymbol{\alpha}} = \left( \sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i=1}^{m} \mathbf{x}_i \log (\hat{\psi}_i).$$

On obtient alors l'estimateur $\hat{\tilde{\psi}}_i$ de $\tilde{\psi}_i$ comme suit :

$$\hat{\tilde{\psi}}_i = \exp (\mathbf{x}_i^T \hat{\boldsymbol{\alpha}}) \hat{\Delta} (\hat{\boldsymbol{\alpha}}).$$

Une belle propriété de $\hat{\tilde{\psi}}_i$ vient du fait que la moyenne de l'estimateur de la variance sous le plan lisse, $\hat{\tilde{\psi}}_i$, est égale à la moyenne de l'estimateur de variance directe, $\hat{\psi}_i$; c'est-à-dire :

$$\frac{\sum_{i=1}^{m} \hat{\tilde{\psi}}_i}{m} = \frac{\sum_{i=1}^{m} \hat{\psi}_i}{m}.$$

Cela garantit que $\hat{\tilde{\psi}}_i$ ne surestime ou ne sous-estime pas systématiquement $\tilde{\psi}_i = E_{mp} (\hat{\psi}_i)$.

## 3.2 Réconciliation

Si le paramètre d'intérêt $\theta_i$ est un total $(\theta_i = Y_i)$, l'utilisateur peut souhaiter que la somme des estimations sur petits domaines, $\hat{\theta} = \sum_{i \in A \cup \bar{A}} \hat{\theta}_i^{\text{EBLUP}}$, corresponde aux totaux estimés $\hat{Y} = \sum_{i \in A} \hat{Y}_i$ au niveau de l'échantillon général $s$; c'est-à-dire $\hat{\theta} = \hat{Y}$. Dans le cas d'une moyenne, $\theta_i = \overline{Y}_i$, cette condition de réconciliation devient $\sum_{i \in A \cup \bar{A}} N_i \hat{\theta}_i^{\text{EBLUP}} = \sum_{i \in A} N_i \hat{\theta}_i$, où $\hat{\theta}_i = \hat{\overline{Y}}_i$.

Dans le système de production, deux méthodes permettent d'assurer la réconciliation des estimations sur petits domaines au niveau du domaine. La première repose sur un rajustement de la différence et la seconde, sur un vecteur enrichi. Elles sont valides pour toute méthode servant à calculer $\hat{\theta}_i^{\text{EBLUP}}$ ou pour déterminer si l'estimation de la variance $\ddot{\psi}_i$ a été lissée. La réconciliation fondée sur un rajustement de la différence est une adaptation de la réconciliation présentée dans Battese et coll. (1988). La réconciliation fondée sur un vecteur enrichi est attribué à Wang, Fuller et Qu (2008).

*Rajustement de la différence* : Pour cette méthode, l'estimateur $\hat{\theta}_i^{\text{EBLUP}}$ est ajusté uniquement pour les domaines dont la taille d'échantillon réalisé $n_i \geq 1$, $i \in A$ et les estimations synthétiques $\mathbf{z}_i^T \hat{\boldsymbol{\beta}}$ pour $i \in \bar{A}$ restent telles quelles. L'estimateur ainsi réconcilié est obtenu à l'aide de $\hat{\theta}_i^{\text{EBLUP}, b}$ et est défini comme suit :

$$\hat{\theta}_i^{\text{EBLUP}, b} = \begin{cases} \hat{\theta}_i^{\text{EBLUP}} + \alpha_i \left( \hat{\theta}^* - \sum_{d \in A} \omega_d \hat{\theta}_d^{\text{EBLUP}} \right) & \text{pour } i \in A \\ \mathbf{z}_i^T \hat{\boldsymbol{\beta}} & \text{pour } i \in \bar{A} \end{cases}$$

où $\alpha_i = \left\{ \sum_{i \in U_A} \omega_i^2 (\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^2) \right\}^{-1} \omega_i (\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^2)$ pour $i \in A$, $\omega_i = 1$, si la réconciliation est pour un total, et $\omega_i = N_i / N$, si la réconciliation est pour la moyenne. L'estimateur $\hat{\theta}^*$ est une valeur fournie par l'utilisateur qui représente le total ou la moyenne des $y$- valeurs de population $U$. La réconciliation garantit que $\sum_{i \in A \cup \bar{A}} \omega_i \hat{\theta}_i^{\text{EBLUP}, b} = \hat{\theta}^*$.

*Vecteur enrichi* : Le vecteur $\mathbf{z}_i^T$ est enrichi à l'aide de $\omega_i \ddot{\psi}_i$, pour former $\mathbf{z}_i^{*T} = (\mathbf{z}_i^T, \omega_i \ddot{\psi}_i)$ où $\omega_i$ et $\ddot{\psi}_i$ sont définis comme précédemment. L'équation du modèle linéaire mixte généralisé (MLMG) enrichi qui en découle est obtenue comme ceci :

$$\hat{\theta}_i = \mathbf{z}_i^{*T} \boldsymbol{\beta}^* + b_i v_i^* + e_i \tag{3.5}$$

où $E_m(v_i^*) = 0$ et $V_m(v_i^*) = \sigma_v^{*2}$. Les estimations de $\boldsymbol{\beta}^*$ et $\sigma_v^{*2}$ sont une fois de plus résolues de manière récursive pour les quatre procédures EBLUP que nous désignons comme $\hat{\theta}_i^{\text{EBLUP*}}$.

L'estimateur réconcilié $\hat{\theta}_i^{\text{EBLUP*, }b}$ qui en découle est obtenu comme ceci :

$$\hat{\theta}_i^{\text{EBLUP*, }b} = \begin{cases} \hat{\gamma}_i^* \hat{\theta}_i^{\text{EBLUP*}} + (1 - \hat{\gamma}_i^*)\, \mathbf{z}_i^{*T} \hat{\boldsymbol{\beta}}^* & \text{pour } i \in A \\ \mathbf{z}_i^{*T}\, \hat{\boldsymbol{\beta}}^* & \text{pour } i \in \bar{A} \end{cases}$$

où $\hat{\gamma}_i^* = (b_i^2 \hat{\sigma}_v^{*2}) / (\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^{*2})$, et $\hat{\boldsymbol{\beta}}^* = \left( \sum_{i \in A} \mathbf{z}_i^* \mathbf{z}_i^{*T} / (\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^{*2}) \right)^{-1} \sum_{i \in A} \mathbf{z}_i^* \hat{\theta}_i / (\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^{*2})$.

Toutes les composantes de $\hat{\theta}_i^{\text{EBLUP*, }b}$ sont calculées à l'aide du modèle enrichi obtenu sous (3.5). On peut démontrer que $\sum_{i \in A \cup \bar{A}} \omega_i \hat{\theta}_i^{\text{EBLUP*, }b} = \sum_{i \in A} \omega_i \hat{\theta}_i$, et, par conséquent, la réconciliation tient.

Les méthodes du rajustement de la différence et du vecteur enrichi sont deux façons de satisfaire la réconciliation. Wang et coll. (2008) ont suggéré d'autres procédures à employer. Plus précisément, ils ont adapté l'estimateur à réconciliation automatique que You et Rao (2002) avaient conçu dans le contexte du modèle au niveau de l'unité pour le modèle au niveau du domaine. You, Rao et Hidiroglou (2013) ont obtenu un estimateur de l'erreur de prédiction quadratique moyenne et de son biais à l'aide d'un modèle mal spécifié.

## 3.3 Estimation de l'erreur quadratique moyenne

La fiabilité des estimateurs EBLUP est obtenue au moyen de $\text{EQM}\left(\hat{\theta}_i^{\text{EBLUP}}\right) = E\left(\hat{\theta}_i^{\text{EBLUP}} - \theta_i\right)^2$. Cette espérance prise par rapport au modèle (3.3) pour l'estimateur non réconcilié, et par rapport au modèle (3.5) pour l'estimateur réconcilié.

Les erreurs quadratiques moyennes (EQM) estimées des estimateurs au niveau du domaine sont présentées au tableau 3.1. La forme particulière des termes $g$ et les variances estimées se trouvent dans Rao et Molina (2015) ou dans Estevao et coll. (2015). Pour les estimateurs réconciliés, l'EQM estimée de l'approche de rajustement de la différence a recours à des formules de l'EQM non réconciliées. Dans le cas d'une approche vectorielle enrichie, l'EQM est fondée sur l'enrichissement du vecteur $\mathbf{z}_i^T$ avec $\omega_i \ddot{\psi}_i$.

**Tableau 3.1**
**Estimations de l'EQM pour les estimateurs au niveau du domaine**

| Estimateur | EQM |
|---|---|
| Fay-Herriot | $\text{eqm}\left(\hat{\theta}_i^{\text{FH}}\right) = \begin{cases} g_{0i} + g_{1i} + g_{2i} + 2g_{3i} & \text{pour } i \in A \\ \mathbf{z}_i^T \text{var}\left(\hat{\boldsymbol{\beta}}\right) \mathbf{z}_i + b_i^2 \hat{\sigma}_v^2 & \text{pour } i \in \bar{A} \end{cases}$ |
| ADM | $\text{eqm}\left(\hat{\theta}_i^{\text{ADM}}\right) = \begin{cases} g_{0i} + g_{1i} + g_{2i} + 2g_{3i} & \text{pour } i \in A \\ \mathbf{z}_i^T \text{var}\left(\hat{\boldsymbol{\beta}}\right) \mathbf{z}_i + b_i^2 \hat{\sigma}_v^2 & \text{pour } i \in \bar{A} \end{cases}$ |
| REML | $\text{eqm}\left(\hat{\theta}_i^{\text{REML}}\right) = \begin{cases} g_{1i} + g_{2i} + 2g_{3i} & \text{pour } i \in A \\ \mathbf{z}_i^T \text{var}\left(\hat{\boldsymbol{\beta}}\right) \mathbf{z}_i + b_i^2 \hat{\sigma}_v^2 & \text{pour } i \in \bar{A} \end{cases}$ |
| WF | $\text{eqm}\left(\hat{\theta}_i^{\text{WF}}\right) = \begin{cases} g_{1i} + g_{2i} + 2g_{3i} + g_{4i} & \text{pour } i \in A \\ \mathbf{z}_i^T \text{var}\left(\hat{\boldsymbol{\beta}}\right) \mathbf{z}_i + b_i^2 \hat{\sigma}_v^2 & \text{pour } i \in \bar{A} \end{cases}$ |

Les divers termes $g$ dans le tableau 3.1 peuvent être interprétés comme suit. Le $g_{0i}$ est un terme de correction du biais pour les méthodes FH et ADM. Le terme $g_{1i}$ obtenu à l'aide de $g_{1i} = \hat{\gamma}_i \ddot{\psi}_i$, représente la majeure partie de l'EQM si le nombre de domaines est élevé. Le terme $g_{2i}$ représente l'estimation de $\boldsymbol{\beta}$, et $2g_{3i}$, l'estimation de $\sigma_v^2$. Le terme $g_{4i}$ dans la procédure WF indique que la valeur estimée de $\psi_i$, $\hat{\psi}_i$, a été utilisée. La variance estimée de $\hat{\boldsymbol{\beta}}$, obtenue par $\text{var}\left(\hat{\boldsymbol{\beta}}\right) = \left(\sum_{i \in A} \frac{\mathbf{z}_i \mathbf{z}_i^T}{\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^2}\right)^{-1}$ dépend de la procédure particulière employée pour estimer $\sigma_v^2$.

# 4 Modèle au niveau de l'unité

Le modèle original au niveau de l'unité a été proposé par Battese et coll. (1988). Ces derniers ont supposé le modèle à erreurs emboîtées suivant :

$$y_{ij} = \mathbf{z}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij} \quad \text{pour} \quad i = 1, \ldots, m \quad \text{et} \quad j \in U_i \tag{4.1}$$

où $v_i \overset{\text{ind}}{\sim} (0, \sigma_v^2)$ sont les effets aléatoires et sont indépendants des erreurs aléatoires, $e_{ij}$, avec $e_{ij} \overset{\text{ind}}{\sim} (0, \sigma_e^2)$. Le système de production comprend une légère modification de la structure des erreurs aléatoires, c'est-à-dire que $e_{ij} \overset{\text{ind}}{\sim} \left(0, \sigma_e^2 / a_{ij}\right)$, où $a_{ij} > 0$ sont des constantes positives qui rendent compte de l'hétéroscédasticité.

Le système de production calcule les estimations sur petits domaines pour les moyennes $\left(\overline{Y}_{ic} = \sum_{j \in U_i} c_{ij} y_{ij} \Big/ \sum_{j \in U_i} c_{ij}\right)$ et les totaux $\left(Y_{ic} = \sum_{j \in U_i} c_{ij} \overline{Y}_i\right)$. Les valeurs $c_{ij}$ sont des constantes positives fixes connues pour toutes les unités de la population. Il a fallu ajouter $c_{ij}$ afin que certaines enquêtes-entreprises menées à Statistique Canada puissent utiliser le système (voir Rubin-Bleuer, Jang et Godbout, 2016). Les données auxiliaires disponibles sont des totaux $\mathbf{Z}_{ic} = \sum_{j \in U_i} c_{ij} \mathbf{z}_{ij}$, ou des moyennes $\overline{\mathbf{Z}}_{ic} = \sum_{j \in U_i} c_{ij} \mathbf{z}_{ij} \Big/ \sum_{j \in U_i} c_{ij}$.

Dans ce qui suit, nous fournissons les estimateurs de la moyenne de population $\overline{Y}_{ic}$, disons $\hat{\theta}_i^{\text{EPD}}$, où $i = 1, \ldots, M$. Les estimations des totaux correspondants $Y_{ic}$, sont obtenues en multipliant $\hat{\theta}_i^{\text{EPD}}$ par $\sum_{j=1}^{N_i} c_{ij}$.

La moyenne de l'échantillon pondéré déterminé par le plan d'échantillonnage des $y$ et des $\mathbf{z}$ est respectivement :

$$\overline{y}_{iwc} = \left(\sum_{j \in s_i} w_{ij} c_{ij}\right)^{-1} \sum_{j \in s_i} w_{ij} c_{ij} y_{ij}$$

et

$$\overline{\mathbf{z}}_{iwc} = \left(\sum_{j \in s_i} w_{ij} c_{ij}\right)^{-1} \sum_{j \in s_i} w_{ij} c_{ij} \mathbf{z}_{ij}.$$

Les moyennes pondérées fondées sur le modèle sont les suivantes :

$$\overline{y}_{ia} = \left(\sum_{j \in s_i} a_{ij}\right)^{-1} \left(\sum_{j \in s_i} a_{ij} y_{ij}\right)$$

et

$$\overline{\mathbf{z}}_{ia} = \left(\sum_{j \in s_i} a_{ij}\right)^{-1} \left(\sum_{j \in s_i} a_{ij} \mathbf{z}_{ij}\right).$$

Battese et coll. (1988) n'ont pas inclus les poids déterminés par le plan d'échantillonnage dans leur procédure, forçant ainsi l'uniformité du plan, à moins qu'il ne soit autopondéré. Nous appelons cet estimateur EBLUP $\left(\hat{\theta}_i^{\text{EBLUP}}\right)$. Cependant, EBLUP est l'estimateur le plus efficace selon le modèle (4.1), avec une structure d'erreur $e_{ij} \overset{\text{ind}}{\sim} \left(0, \sigma_e^2 / a_{ij}\right)$, et c'est la raison pour laquelle il est inclus dans le système de production.

Kott (1989), Prasad et Rao (1999), et You et Rao (2002) ont proposé d'utiliser des estimateurs fondés sur un modèle et convergents par rapport au plan pour les moyennes des domaines en incluant le poids d'enquête. La procedure de You et Rao (2002) a été modifiée de façon appropriée pour représenter les résidus hétéroscédastiques $c_{ij}$. Le pseudo-estimateur EBLUP qui en découle, désigné sous le nom de PEBLUP $\left(\hat{\theta}_i^{\text{PEBLUP}}\right)$, a été inclus dans le système de production parce qu'il est convergent par rapport au plan.

L'estimateur EBLUP est défini comme suit :

$$\hat{\theta}_i^{\text{EBLUP}} = \begin{cases} \hat{\gamma}_{ia}\overline{y}_{ia} + \left(\overline{\mathbf{Z}}_{ic} - \hat{\gamma}_{ia}\overline{\mathbf{z}}_{ia}\right)^T \hat{\boldsymbol{\beta}}^{\text{EBLUP}} & \text{si } i \in A \\ \overline{\mathbf{Z}}_{ic}^T \hat{\boldsymbol{\beta}}^{\text{EBLUP}} & \text{si } i \in \overline{A} \end{cases}$$

où $\hat{\gamma}_{ia} = \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \Big/ \sum_{j\in s_i} a_{ij}\right)^{-1} \hat{\sigma}_v^2$. Les termes $\overline{y}_{ia}$ et $\overline{\mathbf{z}}_{ia}$, sont les moyennes pondérées fondées sur le modèle déjà définies, respectivement pour $y$ et $\mathbf{z}$. Le vecteur de régression $\boldsymbol{\beta}$ est estimé comme ceci :

$$\hat{\boldsymbol{\beta}}^{\text{EBLUP}} = \left(\sum_{i=1}^{m}\sum_{j\in s_i} a_{ij}c_{ij}\left(\mathbf{z}_{ij} - \hat{\gamma}_{iac}\overline{\mathbf{z}}_{iac}\right)\mathbf{z}_{ij}^T\right)^{-1} \sum_{i=1}^{m}\sum_{j\in s_i} a_{ij}c_{ij}\left(\mathbf{z}_{ij} - \hat{\gamma}_{iac}\overline{\mathbf{z}}_{iac}\right) y_{ij}.$$

L'estimateur PEBLUP, $\hat{\theta}_i^{\text{PEBLUP}}$, est obtenu comme ceci :

$$\hat{\theta}_i^{\text{PEBLUP}} = \begin{cases} \hat{\gamma}_{iwc}\overline{y}_{iwc} + \left(\overline{\mathbf{Z}}_{ic} - \hat{\gamma}_{iwc}\overline{\mathbf{z}}_{iwc}\right)^T \hat{\boldsymbol{\beta}}^{\text{PEBLUP}} & \text{si } i \in A \\ \overline{\mathbf{Z}}_{ic}^T \hat{\boldsymbol{\beta}}^{\text{PEBLUP}} & \text{si } i \in \overline{A} \end{cases}$$

où $\hat{\gamma}_{iwc} = \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2\delta_{iwc}^2\right)^{-1}\left(\hat{\sigma}_v^2\right)$, et $\delta_{iwc}^2 = \left(\sum_{j\in s_i} w_{ij}c_{ij}\right)^{-2}\left(\sum_{j\in s_i}\left(w_{ij}c_{ij}\right)^2 \Big/ a_{ij}\right)$. Les termes $\overline{y}_{iwc}$ et $\overline{\mathbf{z}}_{iwc}$, sont les moyennes pondérées fondées sur le modèle déjà définies, respectivement pour $y$ et $\mathbf{z}$. Le vecteur de régression $\boldsymbol{\beta}$ est estimé comme ceci :

$$\hat{\boldsymbol{\beta}}^{\text{PEBLUP}} = \left(\sum_{i=1}^{m}\sum_{j\in s_i} w_{ij}a_{ij}\left(\mathbf{z}_{ij} - \hat{\gamma}_{iwa}\overline{\mathbf{z}}_{iwa}\right)\mathbf{z}_{ij}^T\right)^{-1} \sum_{i=1}^{m}\sum_{j\in s_i} w_{ij}a_{ij}\left(\mathbf{z}_{ij} - \hat{\gamma}_{iwa}\overline{\mathbf{z}}_{iwa}\right) y_{ij}$$

où $\overline{\mathbf{z}}_{iwa} = \left(\sum_{j\in s_i} w_{ij}a_{ij}\right)^{-1}\sum_{j\in s_i} w_{ij}a_{ij}\mathbf{z}_{ij}$, $\hat{\gamma}_{iwa} = \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2\delta_{iwa}^2\right)^{-1}\hat{\sigma}_v^2$ et où $\delta_{iwa}^2$ est calculé comme $\delta_{iwa}^2 = \left(\sum_{j\in s_i} w_{ij}a_{ij}\right)^{-2}\left(\sum_{j\in s_i}\left(w_{ij}a_{ij}\right)^2 \Big/ a_{ij}\right)$.

Les composantes de la variance, $\sigma_e^2$ et $\sigma_v^2$, sont estimées au moyen de l'ajustement des constantes (non pondérées par les poids d'enquête), comme l'ont indiqué Battese et coll. (1988) ou Rao (2003). Les estimateurs de $\sigma_e^2$ qui en découlent sont toujours supérieurs ou égaux à zéro, mais l'estimateur de $\sigma_v^2$ peut être négatif. Si $\hat{\sigma}_v^2 < 0$, il s'établit à zéro, ce qui implique qu'il n'y a pas d'effets sur le domaine. On obtient les EQM connexes en élargissant les méthodes de You et Rao (2002) et de Stukel et Rao (1997).

À noter que, si l'échantillon $s$ est sélectionné dans l'univers $U$, la fraction de l'échantillon réalisé, $f_i = n_i / N_i$, pourrait être non négligeable. Pour estimer une moyenne de population, $\overline{Y}_i$, Rao et Molina (2015) ont pris en compte des fractions d'échantillonnage non négligeables en les exprimant comme suit :

$$\overline{Y}_i = f_i \overline{y}_{is} + (1 - f_i)\, \overline{y}_{i\overline{s}}$$

où $\overline{y}_{is}$ est la moyenne de l'échantillon du $i^{\text{e}}$ domaine échantillonné et $\overline{y}_{i\overline{s}}$ est la moyenne de l'échantillon des unités non échantillonnées dans ce domaine. Ils ont établi des prédictions pour $\overline{y}_{i\overline{s}}$ à l'aide du modèle au niveau de l'unité obtenu par l'équation (4.1). Leurs expressions correspondent au cas où $c_{ij} = 1$. Cet estimateur a été élargi par Rubin-Bleuer (2014) afin d'inclure les estimateurs EBLUP et PEBLUP au cas où $c_{ij}$ serait arbitraire. Des détails précis qui tiennent également compte de l'estimation de l'EQM se trouvent dans Estevao et coll. (2015).

## 4.1 Réconciliation

Le système de production actuel ne renferme pas de procédure qui permette de réconcilier les estimations obtenues à l'aide du modèle au niveau de l'unité. Toutefois, on peut modifier la méthode du rajustement de la différence de façon appropriée pour remédier à la situation. Les estimateurs EBLUP et PEBLUP prennent la forme suivante :

$$\hat{\theta}_i^{\text{EPD}} = \begin{cases} \hat{\gamma}_i^* \overline{y}_i^* + (\overline{\mathbf{Z}}_{ic} - \hat{\gamma}_i^* \overline{\mathbf{z}}_i^*)^T \hat{\boldsymbol{\beta}}^{\text{EPD}} & \text{si } i \in A \\ \overline{\mathbf{Z}}_{ic}^T \hat{\boldsymbol{\beta}}^{\text{EPD}} & \text{si } i \in \overline{A} \end{cases}$$

où $\hat{\gamma}_i^*$, $\overline{y}_i^*$, $\overline{\mathbf{z}}_i^*$, et $\hat{\boldsymbol{\beta}}^{\text{EPD}}$ correspondent aux termes définis précédemment : $\hat{\gamma}_i^*$ est égal à $\hat{\gamma}_{ia}$ pour EBLUP, et à $\hat{\gamma}_{iwc}$ pour PEBLUP; $\overline{y}_i^*$ est égal à $\overline{y}_{ia}$ pour EBLUP, et à $\overline{y}_{iwc}$ pour PEBLUP; $\overline{\mathbf{z}}_i^*$ est égal à $\overline{\mathbf{z}}_{ia}$ pour EBLUP, et à $\overline{\mathbf{z}}_{iwc}$ pour PEBLUP; et, $\hat{\boldsymbol{\beta}}^{\text{EPD}}$ est égal à $\hat{\boldsymbol{\beta}}^{\text{EBLUP}}$ pour EBLUP, et à $\hat{\boldsymbol{\beta}}^{\text{PEBLUP}}$ pour PEBLUP.

Supposons qu'il faut réconcilier $\hat{\theta}_i^{\text{EPD}}$ comme $\hat{\theta}^*$. L'estimateur réconcilié correspondant est :

$$\hat{\theta}_i^{\text{EPD, }b} = \begin{cases} \hat{\theta}_i^{\text{EPD}} + \alpha_i \left( \theta^* - \sum_{d \in A} \omega_d \hat{\theta}_d^{\text{EPD}} \right) & \text{si } i \in A \\ \overline{\mathbf{Z}}_{ic}^T \hat{\boldsymbol{\beta}}^{\text{EPD}} & \text{si } i \in \overline{A} \end{cases}$$

où $\alpha_i = \left( \sum_{d \in A} \omega_d^2 \tau_d \right)^{-1} (\omega_i \tau_i)$. Le terme $\omega_i$ est défini comme suit : $\omega_i = 1$ si la réconciliation est pour un total et $\omega_i = N_i / N$ si la réconciliation est pour la moyenne. Les options possibles pour les $\tau_i$ sont $\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \delta_{ia}^2$, $\delta_{ia}^2 = \left( \sum_{j=1}^{n_i} a_{ij} \right)^{-1}$, pour EBLUP, et $\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \delta_{iwc}^2$ pour PEBLUP.

## 4.2 Estimation de l'erreur quadratique moyenne

Les estimations de l'erreur quadratique moyenne des estimateurs au niveau de l'unité sont fondées sur l'estimation de leur erreur quadratique moyenne, selon le modèle (4.1) et la structure d'erreur $e_{ij} \overset{\text{ind}}{\sim} \left(0, \sigma_e^2 / a_{ij}\right)$. Le tableau 4.1 présente ces EQM estimées.

**Tableau 4.1**
**Estimations des EQM pour les estimateurs au niveau de l'unité**

| Estimateur | EQM |
|---|---|
| EBLUP | $\text{eqm}\left(\hat{\theta}_i^{\text{EBLUP}}\right) = \begin{cases} g_{1ia} + g_{2ia} + 2g_{3ia} & \text{pour } i \in A \\ \overline{\mathbf{Z}}_i^T \text{ var}\left(\hat{\boldsymbol{\beta}}^{\text{EBLUP}}\right)\overline{\mathbf{Z}}_i + \hat{\sigma}_v^2 & \text{pour } i \in \overline{A} \end{cases}$ |
| PEBLUP | $\text{eqm}\left(\hat{\theta}_i^{\text{PEBLUP}}\right) = \begin{cases} g_{1iw} + g_{2iw} + 2g_{3iw} & \text{pour } i \in A \\ \overline{\mathbf{Z}}_i^T \text{ var}\left(\hat{\boldsymbol{\beta}}^{\text{PEBLUP}}\right)\overline{\mathbf{Z}}_i + \hat{\sigma}_v^2 & \text{pour } i \in \overline{A} \end{cases}$ |

Les divers termes $g$ au tableau 4.1 peuvent être interprétés de la même façon que ceux associés aux EQM au niveau du domaine. Les $g_{1i}$ désignés par $g_{1ia}$ pour EBLUP, et $g_{1iw}$ pour PEBLUP représentent la majeure partie des EQM si le nombre de domaines est élevé. Les $g_{2i}$ représentent l'estimation de $\boldsymbol{\beta}$, et les $2g_{3i}$, l'estimation de $\sigma_v^2$ et $\sigma_e^2$.

Les variances estimées de $\hat{\boldsymbol{\beta}}^{\text{EBLUP}}$ et $\hat{\boldsymbol{\beta}}^{\text{PEBLUP}}$ sont représentées respectivement par :

$$\text{var}\left(\hat{\boldsymbol{\beta}}^{\text{EBLUP}}\right) = \hat{\sigma}_e^2 \left( \sum_{i \in A} \sum_{j \in s_i} a_{ij} \left( \mathbf{z}_{ij} - \hat{\gamma}_{ia} \overline{\mathbf{x}}_{ia} \right) \mathbf{z}_{ij}^T \right)^{-1}$$

et

$$\text{var}\left(\hat{\boldsymbol{\beta}}^{\text{PEBLUP}}\right) = \hat{\sigma}_e^2 \left( \sum_{i \in A} \sum_{j \in s_i} \mathbf{z}_{ij}^* \mathbf{z}_{ij}^{*T} \right)^{-1} \left( \sum_{i \in A} \sum_{j \in s_i} \mathbf{z}_{ij}^* \mathbf{z}_{ij}^{*T} \Big/ a_{ij} \right) \left( \sum_{i \in A} \sum_{j \in s_i} \mathbf{z}_{ij}^* \mathbf{z}_{ij}^{*T} \right)^{-1}$$

où $\mathbf{z}_{ij}^* = w_{ij} a_{ij} \left( \mathbf{z}_{ij} - \hat{\gamma}_{iwa} \overline{\mathbf{z}}_{iwa} \right)$.

La forme précise des termes $g$ et des variances estimées se trouve dans Estevao et coll. (2015).

# 5 Méthode hiérarchique bayésienne (HB)

Le modèle de base de Fay-Herriot au niveau du domaine comprend un modèle d'échantillonnage linéaire pour les estimations directes découlant de l'enquête et un modèle de lien linéaire pour les paramètres d'intérêt. Ces modèles sont *appariés* parce que $\theta_i$ apparaît comme une fonction linéaire dans les modèles d'échantillonnage et de lien. Dans certains cas, ces équations ne sont pas appariées, par exemple, lorsqu'une fonction, $h(\theta_i)$, est modélisée comme une fonction linéaire de variables explicatives au lieu de $\theta_i$. La paire du *modèle d'échantillonnage* et du *modèle de lien* est :

$$\hat{\theta}_i = \theta_i + e_i \tag{5.1}$$

et

$$h(\theta_i) = \mathbf{z}_i^T \boldsymbol{\beta} + b_i v_i \tag{5.2}$$

où $e_i \overset{\text{ind}}{\sim} N(0, \psi_i)$ et $v_i \overset{\text{ind}}{\sim} N(0, \sigma_v^2)$.

La paire de modèles obtenue sous (5.1) et (5.2) porte le nom de modèle *non apparié*. Les modèles de lien non linéaire sont souvent nécessaires en pratique, car ils fournissent un modèle mieux adapté aux données. Par exemple, si le paramètre d'intérêt est une probabilité ou un taux dont la plage s'étend de 0 à 1, il se peut qu'un modèle de lien linéaire avec des effets aléatoires normaux ne convienne pas. Dans un tel

cas, un modèle de lien pourrait être un modèle logistique ou log-linéaire. Ce genre de modèle a été utilisé de manière à ajuster les chiffres à des niveaux détaillés pour le Recensement du Canada de 2011. Dick (1995) et You, Rao et Dick (2004) décrivent bien ce qu'il faut faire pour effectuer un tel ajustement.

Le système de production comprend les choix suivants de $h(\theta_i)$ :

$$h(\theta_i) = \begin{cases} \theta_i & : \text{modèle apparié de Fay-Herriot (FH)} \\ \log(\theta_i) & : \text{modèle log-linéaire non apparié} \\ \log(\theta_i/(\theta_i + C_i)) & : \text{modèle log du sous-dénombrement du recensement non apparié.} \end{cases} \tag{5.3}$$

L'inclusion de $h(\theta_i) = \theta_i$ correspond aux modèles appariés représentés par les équations (3.1) et (3.2). L'un des avantages de choisir la méthode hiérarchique bayésienne vient du fait que le $\sigma_v^2$ estimé ne peut pas être négatif. La fonction $\log(\theta_i)$, où $\theta_i$ est égal à la moyenne de population $\overline{Y}_i$, a été utilisée dans Fay et Herriot (1979). Leur contexte consistait à estimer le revenu par habitant (PCI) pour les petites localités des États-Unis dont la population est inférieure à 1 000 habitants. La fonction $h(\theta_i)$, $\log(\theta_i / (\theta_i + C_i))$, a été incluse pour appuyer la méthodologie d'estimation du sous-dénombrement net dans les recensements canadiens. Dans ce modèle, $\theta_i$ représente le nombre de personnes non dénombrées dans le recensement et $C_i$, les chiffres connus du recensement. Par conséquent, $\theta_i / (\theta_i + C_i)$ est la proportion des personnes sous-dénombrées par le recensement.

On présume que les variances d'échantillonnage, $\psi_i$, sont connues pour tous les modèles de lien représentés par (5.2). On suppose que les variances sont estimées pour les deux premières fonctions (modèle apparié de Fay-Herriot et modèle log-linéaire non apparié) obtenues sous (5.3). Si l'on présume que les variances d'échantillonnage, $\psi_i$, sont connues, alors les paramètres inconnus dans le modèle d'échantillonnage (5.1) et le modèle de lien (5.2) peuvent être présentés comme suit dans le cadre hiérarchique bayésien (HB) :

$$\left[\hat{\theta}_i \mid \theta_i\right] \sim N(\theta_i, \psi_i), \quad i = 1, \ldots, m$$

et

$$[h(\theta_i) \mid \beta, \sigma_v^2] \sim N(\mathbf{z}_i^T \boldsymbol{\beta}, b_i^2 \sigma_v^2).$$

Si les variances d'échantillonnage sont inconnues, elles sont estimées en ajoutant :

$$[d_i \hat{\psi}_i \mid \psi_i] \sim \psi_i \chi_{d_i}^2$$

où $\chi_{d_i}^2$ suit une distribution chi-carrée comportant $d_i = (n_i - 1)$ degrés de liberté.

On suppose que les paramètres du modèle $\boldsymbol{\beta}$, $\sigma_v^2$ et $\psi_i$ (quand il est inconnu) respectent des distributions a priori. Les distributions employées dans le système de production pour $\boldsymbol{\beta}$ et $\sigma_v^2$ sont des distributions a priori horizontales $\pi(\boldsymbol{\beta}) \propto 1$, et $\pi(\sigma_v^2) \propto (\sigma_v^2)^{-1/2}$. Si $\psi_i$ est estimé, la distribution a priori $\pi(\psi_i) \propto (\psi_i)^{-1/2}$ est ajoutée au modèle bayésien. Ces distributions a priori sont multipliées par les fonctions de densité des distributions associées aux modèles d'échantillonnage et de lien. Cela donne une fonction de vraisemblance conjointe en fonction des paramètres du modèle. Cette fonction permet d'obtenir une distribution conditionnelle (postérieure) complète pour chacun des paramètres inconnus. Pour certains d'entre eux, la distribution ainsi obtenue a une forme retraçable ou connue. Pour d'autres, la distribution obtenue est le produit de certaines fonctions de densité de forme inconnue. Toutes les méthodes HB font

intervenir l'estimation des paramètres du modèle par l'échantillonnage répété de leurs distributions conditionnelles complètes respectives.

Les méthodes de Monte Carlo par chaîne de Markov (MCMC) permettent d'obtenir des estimations à partir de la distribution conditionnelle complète de chaque paramètre. L'échantillonnage de Gibbs est utilisé à plusieurs reprises pour prélever un échantillon dans des distributions conditionnelles complètes. La méthode d'échantillonnage de Gibbs (Gelfand et Smith, 1990), combinée avec l'algorithme Metropolis-Hastings (Chib et Greenberg, 1995), permet de calculer les moyennes postérieures et les variances postérieures; voir les détails dans Estevao et coll. (2015). Les divers estimateurs de $\theta_i$ découlant de (5.3) sont représentés par $\hat{\theta}_i^{\mathrm{HB}}$.

## 5.1 Estimateur HB réconcilié

La réconciliation des estimateurs suit la *méthode du rajustement de la différence* décrite dans la section 3.2, c'est-à-dire que les estimateurs réconciliés $\hat{\theta}_i^{\mathrm{HB}}$ sont calculés comme ceci :

$$\hat{\theta}_i^{\mathrm{HB},\,b} = \begin{cases} \hat{\theta}_i^{\mathrm{HB}} + \alpha_i \left(\hat{\theta}^* - \sum_{d\in A} \omega_d \hat{\theta}_d^{\mathrm{HB}}\right) & \text{pour } i \in A \\ \mathbf{z}_i^T \hat{\boldsymbol{\beta}} & \text{pour } i \in \bar{A} \end{cases}$$

où $\alpha_i = \left(\sum_{i\in A} \omega_i^2 \left(\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^{2\mathrm{HB}}\right)\right)^{-1} \omega_i \left(\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^{2\mathrm{HB}}\right)$ pour $i \in A$, et $\hat{\theta}^*$ est la valeur de référence. Les termes $\omega_i$ sont définis comme suit : $\omega_i = 1$ si la réconciliation est pour un total, et $\omega_i = N_i / N$ si la réconciliation est pour la moyenne. Les $\ddot{\psi}_i$ sont connus ou inconnus. Le $\hat{\theta}^*$ peut être une valeur fournie par l'utilisateur qui représente le total ou la moyenne des valeurs $y$ de la population $U$. La réconciliation garantit que $\sum_{i\in A\cup\bar{A}} \omega_i \hat{\theta}_i^{\mathrm{HB},\,b} = \hat{\theta}^*$.

# 6 Application aux données de l'Enquête sur la population active (EPA)

L'EPA de Statistique Canada est une enquête mensuelle sous un plan d'échantillonnage stratifié à deux degrés, conçue pour produire des estimations fiables du taux de chômage pour les 55 régions économiques de l'assurance-emploi (REAE) au Canada. Le taux de chômage dans une région donnée $i$ est défini comme le ratio :

$$\theta_i = \frac{\sum_{j\in U_i} y_{1j}}{\sum_{j\in U_i} y_{2j}},$$

où $y_{1j}$ est une variable binaire indiquant si la personne $j$ est au chômage $\left(y_{1j} = 1\right)$ ou non $\left(y_{1j} = 0\right)$, et $y_{2j}$ est une variable binaire indiquant si la personne $j$ fait partie de la population active $\left(y_{2j} = 1\right)$ ou non $\left(y_{2j} = 0\right)$. L'estimateur direct de $\theta_i$ est l'estimateur composite du calage décrit dans Fuller et Rao (2001). Voir aussi Singh, Kennedy et Wu (2001) et Gambino, Kennedy et Singh (2001). Il peut être représenté sous la forme pondérée suivante :

$$\hat{\theta}_i = \frac{\sum_{j\in s_i} w_j y_{1j}}{\sum_{j\in s_i} w_j y_{2j}},$$

où $w_j$ est un poids composite de calage pour la personne $j$.

Tel que susmentionné, l'estimateur composite de calage est fiable pour l'estimation du taux de chômage dans les 55 REAE. Il est également intéressant d'obtenir des estimations fiables pour 149 régions (villes) au Canada. Mentionnons les 34 régions métropolitaines de recensement (RMR) et les 115 zones de recensement (ZR). Les RMR sont les plus grandes villes du point de vue de la taille de leur population et, généralement, elles représentent aussi une taille d'échantillon importante. Certaines ZR ont un échantillon de très petite taille, parfois même de 0. Pour ces ZR et d'autres plus larges, la taille de l'échantillon n'est pas suffisamment grande pour donner des estimations directes assez fiables du taux de chômage mensuel. Notre objectif était de déterminer si le modèle de Fay-Herriot pouvait permettre d'obtenir des estimations mensuelles suffisamment fiables pour être publiées.

Nous avons construit une variable auxiliaire $z_{1i}$ pour le domaine $i$, $z_{1i} = N_i^{\text{EIB}} / N_i^{15+}$, où $N_i^{\text{EIB}}$ est le nombre de bénéficiaires de l'assurance-emploi dans le domaine $i$ et $N_i^{15+}$ est le nombre de personnes âgées de 15 ans et plus dans le domaine $i$. Le numérateur provient d'une source administrative, tandis que le dénominateur est une projection du recensement calculée par Statistique Canada. Nous avons utilisé le vecteur $\mathbf{z}_i = (1, z_{1i})^T$, ainsi que $b_i = 1, i = 1, \ldots, m,$ pour obtenir les EPD. Nous avons eu recours aux données de mai 2016 lors de cette étude afin de comparer les estimations directes et EPD aux estimations du Recensement de 2016.

Puisque certains des 149 domaines d'intérêt avaient un échantillon de taille très réduite dans l'EPA, nous ne nous en sommes pas servi dans les modèles de Fay-Herriot et de lissage. Nous avons exclu des modèles les domaines où le nombre de personnes échantillonnées dans la population active était inférieur à 10. Il y avait 9 domaines, notamment six où aucune personne n'avait été échantillonnée dans la population active. De plus, pour neuf domaines, l'estimation directe du taux de chômage, $\hat{\theta}_i$, et l'estimation de sa variance directe, $\hat{\psi}_i$, étaient toutes deux égales à 0. Comme ces estimations directes n'étaient pas jugées suffisamment fiables, leurs domaines connexes ont été exclus des modèles. Nous n'avons donc utilisé que 131 domaines dans les modèles. Pour ces domaines, les estimations sur petits domaines étaient des estimations EBLUP, les 18 autres étant des estimations synthétiques.

L'estimateur $\hat{\psi}_i$ de la variance directe $\psi_i$ a été obtenu à l'aide de la méthode bootstrap de Rao-Wu. Nous avons ensuite obtenu les estimations de la variance sous le plan lisse en utilisant $\mathbf{x}_i = (1, \log(z_{1i}), \log(1 - z_{1i}), \log(N_i^{15+}))^T$. Un graphique des résidus du modèle de lissage, $\log(\hat{\psi}_i) - \mathbf{x}_i'\hat{\boldsymbol{\alpha}}$, en fonction des valeurs prédites, $\mathbf{x}_i'\hat{\boldsymbol{\alpha}}$, n'a révélé aucune erreur évidente de spécification du modèle. La figure 6.1 montre un graphique des estimations des variances directes, $\hat{\psi}_i$, en fonction des estimations de la variance lisse, $\tilde{\hat{\psi}}_i$. La ligne rouge est la ligne d'identité. Si le modèle de lissage est approprié, pour toute valeur de $\tilde{\hat{\psi}}_i$, la moyenne des estimations de la variance directe pour les domaines situés autour du domaine $i$ devrait être à peu près égale à $\tilde{\hat{\psi}}_i$. Autrement dit, la ligne rouge devrait passer partout à peu près au milieu des points. En faisant une inspection rapide de la figure 6.1, nous constatons que la ligne rouge se rapproche du milieu des points, même si elle se situe probablement légèrement au-dessus du milieu parce que certaines valeurs de $\hat{\psi}_i$ sont extrêmes, ce qui peut entraîner une légère surestimation de la vraie variance lisse $\tilde{\psi}_i = E_{mp}(\hat{\psi}_i)$. Une légère surestimation ne constitue pas un grave

problème. Il faut toutefois éviter une sous-estimation de $\tilde{\psi}_i$, car elle donne habituellement une sous-estimation de l'EQM. Cela donnerait à l'utilisateur une fausse impression de précision.

Dans l'ensemble, nous sommes satisfaits de nos estimations de la variance lisse. Toutefois, pour les domaines dont la taille de l'échantillon est grande, nous avons défini $\hat{\tilde{\psi}}_i = \hat{\psi}_i$ comme notre estimation de $\tilde{\psi}_i$. Nous avons supposé que les estimations de la variance directe étaient suffisamment stables lorsque la taille de l'échantillon était grande. Nous avons défini que $\hat{\tilde{\psi}}_i = \hat{\psi}_i$ quand le nombre de personnes échantillonnées dans la population active était supérieur à 400. Ce remplacement a été fait dans 35 domaines. Cette stratégie a été adoptée pour éviter d'éventuels petits biais de modèle dans $\hat{\tilde{\psi}}_i$ pour les plus grands domaines, qui pourraient donner lieu à des estimations EBLUP bien différentes des estimations directes. Ce n'est pas une propriété souhaitable pour les domaines dont la taille de l'échantillon est importante.

Les estimations de la variance lisse ont ensuite été utilisées pour obtenir des estimations sur petits domaines pour les 149 domaines d'intérêt. La figure 6.2 présente un graphique des estimations directes et sur petits domaines en fonction de la taille de l'échantillon (nombre de personnes échantillonnées dans la population active). Les estimations sur petits domaines sont bien moins volatiles que les estimations directes, en particulier pour les domaines dont la taille de l'échantillon est la plus petite. Pour les plus grands domaines, comme prévu, les deux estimations sont semblables.

Nous avons d'abord évalué la qualité du modèle sous-jacent de Fay-Herriot avant d'examiner les estimations de l'EQM. La figure 6.3 montre le graphique des estimations directes, $\hat{\theta}_i$, en fonction des valeurs prédites, $\mathbf{z}_i^T\hat{\boldsymbol{\beta}}$. La ligne rouge est la ligne d'identité et la ligne bleue est une courbe spline de lissage non paramétrique. Si l'hypothèse de linéarité est satisfaite, la ligne bleue devrait se rapprocher de la ligne rouge et cette dernière devrait passer partout à peu près au milieu des points. La figure 6.3 n'indique aucunement que l'hypothèse de la linéarité du modèle de Fay-Herriot est contestable.

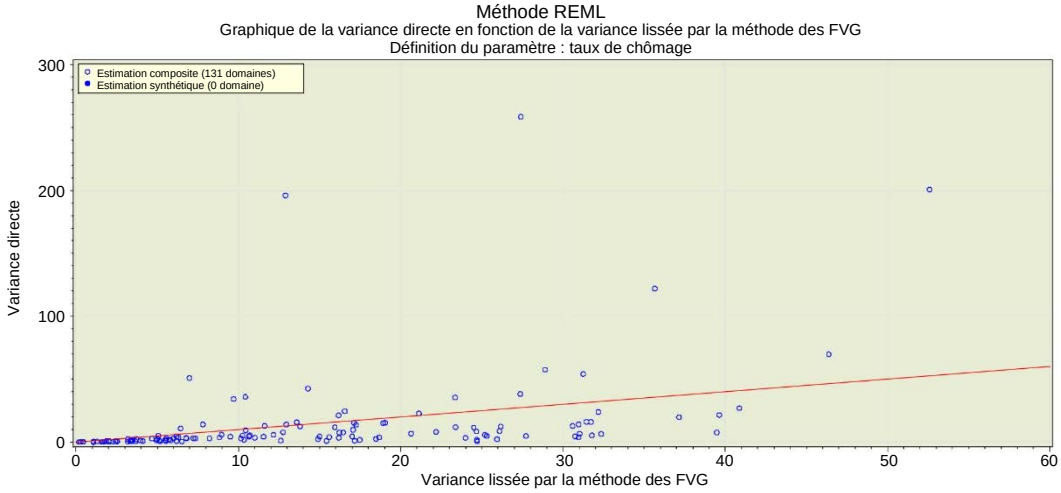


**Figure 6.1 Graphique des estimations de la variance directe, $\hat{\psi}_i$, en fonction des estimations de la variance lisse, $\hat{\tilde{\psi}}_i$.**

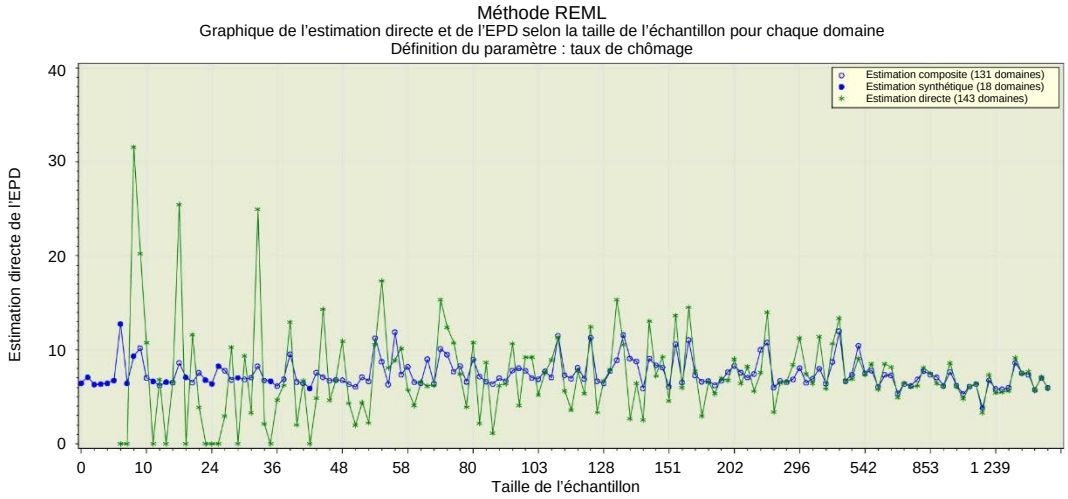


**Figure 6.2 Graphique des estimations sur petits domaines et des estimations directes en fonction de la taille de l'échantillon.**

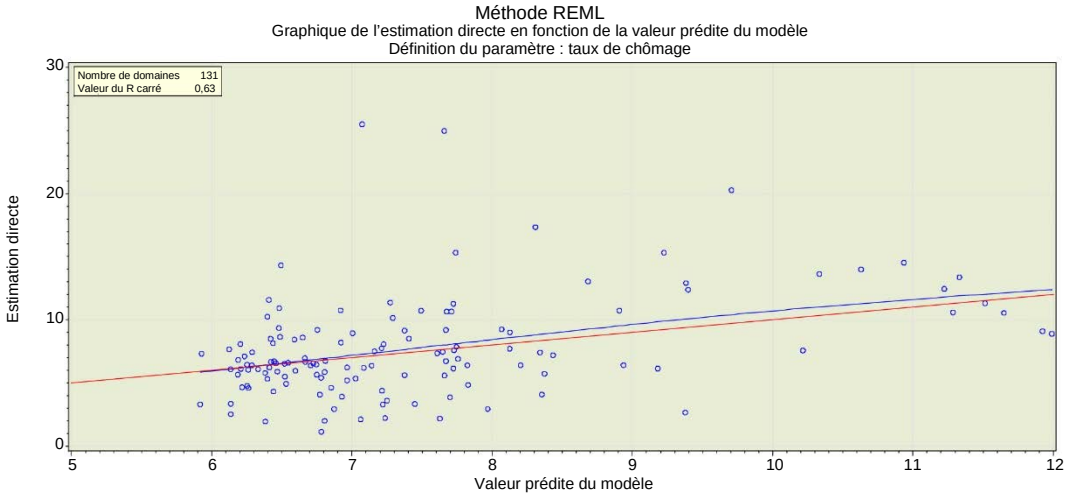


**Figure 6.3 Graphique des estimations directes en fonction des valeurs prédites du modèle.**

Il est également intéressant de calculer une mesure indiquant la vigueur de $\mathbf{z}_i$ pour la prédiction de $\theta_i$. Pour ce faire, nous avons élaboré un coefficient de détermination, ou une valeur $R^2$, associée au modèle de lien $\theta_i = \mathbf{z}_i^T\boldsymbol{\beta} + b_i v_i$. À noter que le coefficient de détermination associé au modèle combiné, $\hat{\theta}_i = \mathbf{z}_i^T\boldsymbol{\beta} + b_i v_i + e_i$, ne revêt aucun intérêt puisque l'objectif n'est pas de prédire $\hat{\theta}_i$, mais bien de prédire $\theta_i$. Nous obtenons notre coefficient de détermination comme ceci :

$$R^2 = 1 - \frac{\hat{\sigma}_v^2}{\dfrac{(m-q)}{(m-1)}\hat{\sigma}_v^2 + S^2\left(\hat{\boldsymbol{\beta}}\right)},$$

où $q$ est la dimension de $\mathbf{z}_i$ et $S^2\left(\hat{\boldsymbol{\beta}}\right)$ est la variance échantillonale de $\mathbf{z}_i^T\hat{\boldsymbol{\beta}}/b_i$ (voir à l'équation (A.6) la définition exacte de la fonction $S^2(\cdot)$). Les détails de la dérivation du coefficient de détermination susmentionné sont précisés en annexe. La figure 6.3 indique que la valeur $R^2$ est de 0,63. Ainsi, le modèle de lien n'est ni faible ni extrêmement fort, mais, espérons-le, suffisamment solide pour réaliser des gains d'efficacité par rapport à l'estimateur direct. Le système donne également des estimations des paramètres du modèle de Fay-Herriot ainsi que leurs erreurs-types. À partir de ces résultats, nous avons découvert que les estimations des paramètres de l'ordonnée à l'origine et de la pente du modèle de Fay-Herriot étaient significativement différentes de 0 au moyen d'un test de Wald au niveau de signification de 0,05.

La figure 6.4 montre un graphique des résidus normalisés, $\left(\hat{\theta}_i - \mathbf{z}_i^T\hat{\boldsymbol{\beta}}\right)/\sqrt{b_i^2\hat{\sigma}_v^2 + \hat{\psi}_i}$, en fonction des valeurs prédites normalisées, $\mathbf{z}_i^T\hat{\boldsymbol{\beta}}/\sqrt{b_i^2\hat{\sigma}_v^2 + \hat{\psi}_i}$. La ligne rouge est une ligne horizontale à zéro et la ligne bleue est une courbe spline de lissage non paramétrique. Tout comme la figure 6.3, la ligne bleue devrait être près de la ligne rouge sous la linéarité et cette dernière devrait passer partout à peu près au milieu des points. Une fois encore, la figure 6.4 n'indique aucune défaillance évidente de l'hypothèse de la linéarité sous-jacente au modèle de Fay-Herriot.

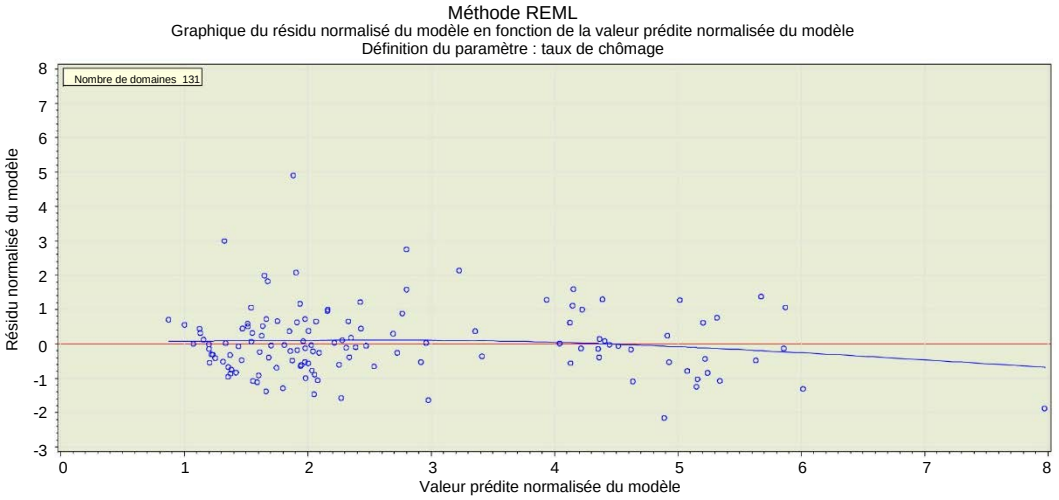


**Figure 6.4 Graphique des résidus normalisés en fonction des valeurs prédites normalisées.**

La figure 6.5 montre un graphique des résidus normalisés au carré en fonction des valeurs prédites normalisées. La ligne rouge est une ligne horizontale à un et la ligne bleue est à nouveau une courbe spline de lissage non paramétrique. Ce graphique sert à vérifier l'hypothèse de l'homoscédasticité; c'est-à-dire l'hypothèse voulant que la variance du modèle $\sigma_v^2$ soit constante. Selon l'homoscédasticité, la ligne bleue devrait être près de la ligne rouge partout. Le graphique ne révèle aucune présence évidente d'hétéroscédasticité.

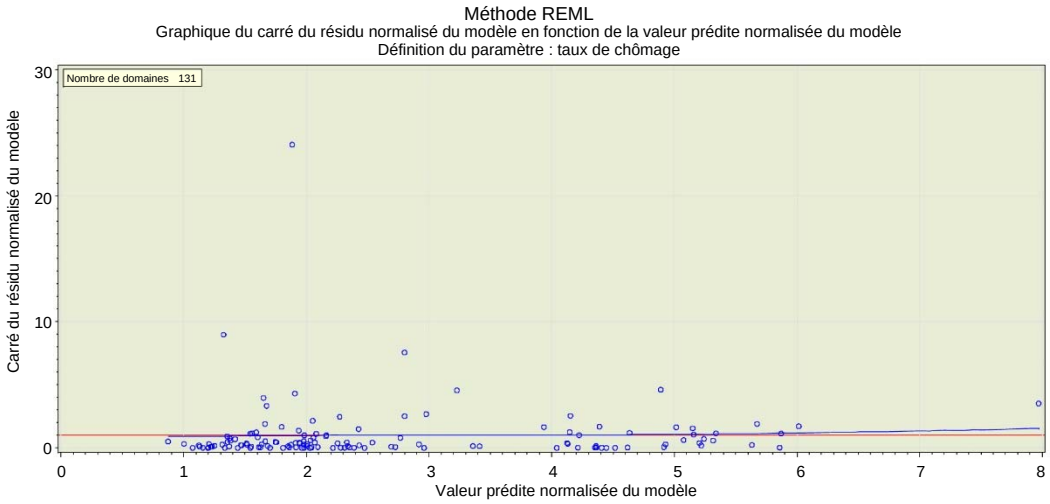


**Figure 6.5 Graphique des résidus normalisés au carré en fonction des valeurs prédites normalisées.**

La figure 6.6 montre un graphique QQ des quantiles des résidus normalisés en fonction des quantiles normaux. Elle sert à vérifier l'hypothèse de la normalité des erreurs $b_i v_i$ et $e_i$. Le graphique indique bien un léger écart par rapport à la normalité, mais Rao et Molina (2015, page 138) ont fait valoir que les estimations EBLUP et leurs estimations de l'EQM correspondantes étaient généralement robustes face aux écarts par rapport à la normalité.

Le système calcule également les distances de Cook pour indiquer les domaines qui pourraient avoir une influence importante sur l'estimation de $\hat{\boldsymbol{\beta}}$. La distance de Cook pour le domaine $i$ est obtenue comme ceci :

$$D_i = \frac{1}{q}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(-i)}\right)^T \sum_{j=1}^{m} \frac{\mathbf{z}_j \mathbf{z}_j^T}{b_j^2 \hat{\sigma}_v^2 + \hat{\tilde{\psi}}_j} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(-i)}\right),$$

où $\hat{\boldsymbol{\beta}}^{(-i)}$ est l'estimation de $\boldsymbol{\beta}$ après la suppression du domaine $i$. Un graphique des influences $D_i$ est présenté à la figure 6.7. Un domaine semble avoir une influence relativement importante par rapport aux autres $(D_i = 1{,}2851)$. Ce domaine a la plus grande valeur prédite normalisée et la deuxième valeur prédite la plus élevée. Son résidu normalisé est de -1,88, ce qui n'est pas extrême, même s'il n'est pas très faible non plus. La taille de l'échantillon est importante (le nombre de personnes échantillonnées dans la population active est de près de 500) et l'estimation de sa variance lisse, $\hat{\tilde{\psi}}_i$, est relativement faible par rapport aux autres domaines. Toutes ces raisons expliquent pourquoi nous avons jugé que ce domaine exerçait une influence. Dans cette application, nous avons décidé de garder ce domaine dans le modèle, car son influence n'était pas assez importante pour faire une grande différence dans les EPD et les estimations de l'EQM correspondantes.

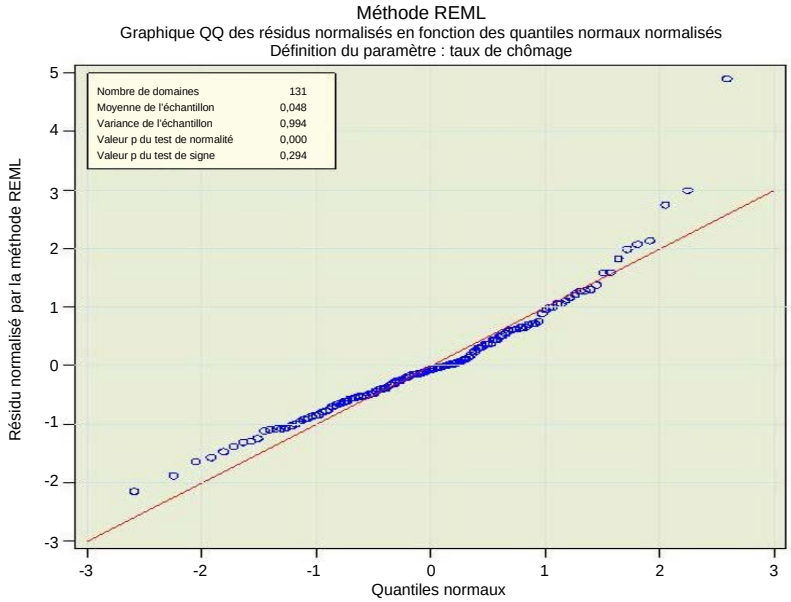


**Figure 6.6 Graphique QQ des quantiles résiduels normalisés en fonction des quantiles normaux normalisés.**

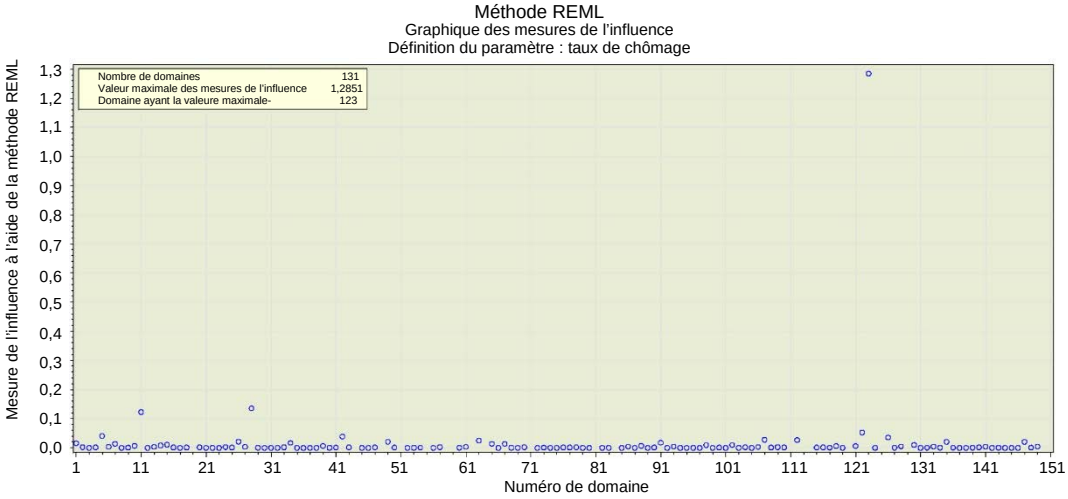


**Figure 6.7 Graphique des distances de Cook.**

Puisque le modèle de Fay-Herriot et le modèle de lissage étaient tous deux raisonnables, nous avons calculé des estimations de l'EQM pour évaluer la présence et l'ampleur des gains d'efficacité découlant du modèle de Fay-Herriot. La figure 6.8 montre le coefficient de variation (CV) direct estimé, défini comme $\sqrt{\hat{\psi}_i}\,/\hat{\theta}_i$, et la racine carrée de l'erreur quadratique moyenne relative (REQMR) de l'EPD estimée, définie comme $\sqrt{\hat{\phi}_i}\,/\hat{\theta}_i^{\mathrm{EPD}}$, où $\hat{\phi}_i$ est une estimation de l'EQM, $E_{mp}\left(\hat{\theta}_i^{\mathrm{EPD}} - \theta_i\right)^2$, et $\hat{\theta}_i^{\mathrm{EPD}}$ est l'estimation sur petits domaines (estimation EBLUP ou synthétique) de $\theta_i$. La taille de l'échantillon (nombre de personnes

échantillonnées dans la population active) est présentée sur l'axe horizontal. Les CV directs estimés sont en général beaucoup plus grands que les REQMR de l'EQM estimée, en particulier pour les domaines dont la taille de l'échantillon est la plus petite. Les REQMR de l'EQM estimée ne dépassent jamais 20 %, alors que le CV direct estimé est supérieur à 300 % pour un domaine. Les REQMR de l'EQM estimée sont également très stables en fonction de la taille de l'échantillon, contrairement au comportement erratique des CV directs estimés. Pour les domaines ayant les plus grandes tailles d'échantillon, les deux estimations sont très semblables, comme prévu. Cela indique que les méthodes d'EPD peuvent entraîner une augmentation importante de la précision par rapport aux méthodes d'estimation directe, en particulier pour les plus petits domaines.

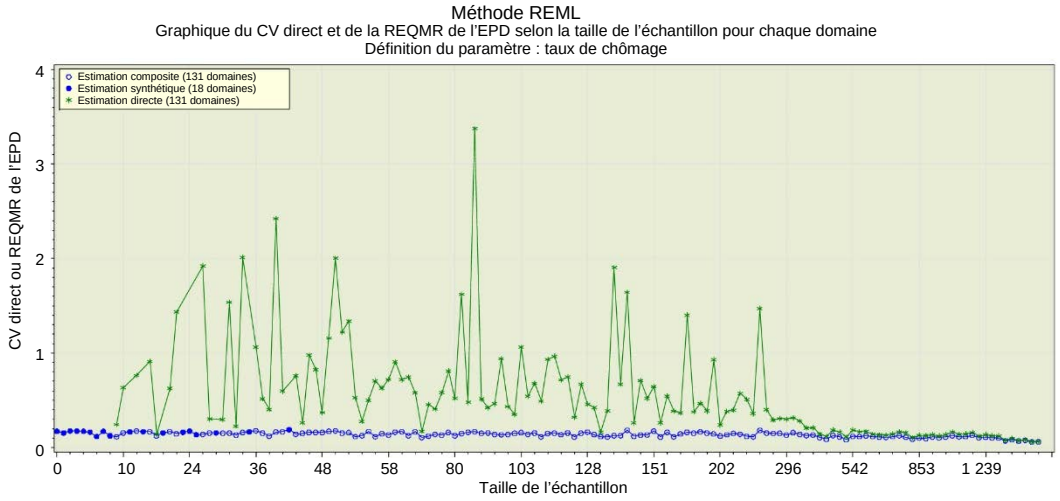


**Figure 6.8 Graphique des CV directs estimés et des REQMR de l'EPD en fonction de la taille de l'échantillon.**

Pour le mois de mai 2016, nous avons eu le luxe d'avoir une source très fiable pour l'estimation des taux de chômage : le questionnaire détaillé du recensement de 2016 distribué à peu près au quart des ménages canadiens. La taille de l'échantillon du recensement est bien plus grande que celle de l'EPA dans tous les domaines d'intérêt. Par conséquent, nous nous sommes servi des estimations directes du Recensement de 2016, représentées par $\hat{\theta}_i^{\text{Recensement}}$, comme standard de référence pour évaluer l'exactitude des estimations directes de l'EPA et des EPD. Nous avons calculé les différences relatives absolues (DRA) entre les estimations directes de l'EPA et les estimations du recensement, $\left|\hat{\theta}_i - \hat{\theta}_i^{\text{Recensement}}\right| / \hat{\theta}_i^{\text{Recensement}}$, ainsi que des DRA entre les EPD et les estimations du recensement, $\left|\hat{\theta}_i^{\text{EPD}} - \hat{\theta}_i^{\text{Recensement}}\right| / \hat{\theta}_i^{\text{Recensement}}$. Nous avons ensuite calculé la moyenne de ces DRA dans cinq sous-groupes homogènes différents de par la taille de l'échantillon. Le tableau 6.1 présente un résumé des résultats.

**Tableau 6.1**
**DRA moyenne des EPD et des estimations directes de l'EPA exprimée en pourcentage**

| Taille de l'échantillon | DRA moyenne entre les estimations directes de l'EPA et les estimations du recensement | DRA moyenne entre les EPD et les estimations du recensement | DRA moyenne entre les estimations de HB et les estimations du recensement |
|---|---|---|---|
| Les 28 domaines les plus petits | 70,4 % | 17,7 % | 18,3 % |
| Les 28 plus petits domaines suivants | 38,7 % | 18,9 % | 19,0 % |
| Les 28 plus petits domaines suivants | 26,2 % | 13,8 % | 14,1 % |
| Les 28 plus petits domaines suivants | 20,9 % | 12,7 % | 13,0 % |
| Les 28 domaines les plus grands | 13,2 % | 10,2 % | 10,3 % |
| Générale | 33,9 % | 14,7 % | 14,9 % |

Nota : Parmi les 149 domaines d'intérêt de l'EPA, neuf ont été exclus de ce tableau : six pour lesquels le nombre de personnes échantillonnées dans la population active de l'EPA était de 0 et trois qui ne figuraient plus sur la liste des RMR/ZR après le Recensement de 2016.

Comme prévu, la DRA entre les estimations directes de l'EPA et du recensement diminue à mesure que la taille de l'échantillon augmente. Cela peut donner à penser que les différences conceptuelles entre ces deux enquêtes et les erreurs non dues à l'échantillonnage sont raisonnablement faibles par rapport à l'erreur d'échantillonnage, surtout dans les plus petits domaines où l'erreur d'échantillonnage peut être le principal facteur à l'origine de la DRA. Les EPD sont beaucoup plus près des estimations du recensement que les estimations directes de l'EPA, particulièrement pour les plus petits domaines où des améliorations s'imposent, ce qui confirme que nos modèles sous-jacents sont raisonnables dans cette application.

À des fins de comparaison, nous avons également calculé les estimations HB, $\hat{\theta}_i^{\mathrm{HB}}$, à partir du modèle apparié de Fay-Herriot avec les distributions a priori non informatives de $\boldsymbol{\beta}$, $\sigma_v^2$ et $\tilde{\psi}_i$ fournies à la section 5. Nous avons ensuite calculé les DRA entre les estimations HB et les estimations du recensement, $\left|\hat{\theta}_i^{\mathrm{HB}} - \hat{\theta}_i^{\mathrm{Recensement}}\right| \Big/ \hat{\theta}_i^{\mathrm{Recensement}}$. Les résultats sont présentés dans la dernière colonne du tableau 6.1. Les DRA moyennes des estimations HB se rapprochent de celles des estimations EBLUP.

# 7 Conclusion

Les utilisateurs des données des ONS demandent souvent une plus grande précision pour répondre à leurs besoins de planification et de recherche sur les politiques. Les ONS ne peuvent plus se contenter d'augmenter la taille des échantillons de leurs enquêtes pour obtenir des estimations fiables au niveau de détail demandé. Les raisons à cela comprennent les coûts élevés de cette démarche, les préoccupations relatives au fardeau de réponse, ainsi que la difficulté d'obtenir des réponses des unités échantillonnées. Comme solution de rechange, de nombreux ONS envisagent d'utiliser des techniques d'estimation sur petits domaines qui permettent de répondre à la demande de données plus détaillées. C'est dans cet esprit que Statistique Canada a commencé à mettre au point un système de production des EPD au début des années 2000 et qu'il dispose maintenant d'un tel système pour ses programmes statistiques. Le système de production traite des modèles au niveau du domaine et de l'unité, qui offrent de nombreuses options, comme les différentes méthodes permettant d'estimer les composantes de la variance, les divers modèles de lien et

les méthodes d'estimation EBLUP et HB. Il sert actuellement à produire des estimations expérimentales pour plusieurs programmes statistiques de Statistique Canada et les premières estimations sur petits domaines devraient être diffusées en 2019.

Comme nous l'avons mentionné dans l'introduction, le seul logiciel existant en 2006 qui pouvait donner des estimations sur petits domaines et leurs erreurs quadratiques moyennes associées a été commandité par le projet EURAREA (2004). L'actuel système de production mis au point à Statistique Canada est rédigé en SAS, sa méthodologie suit de près Rao (2003) et comprend certaines percées récentes. À l'heure actuelle, il remplit les exigences en matière d'estimation sur petits domaines à Statistique Canada. Toutefois, à mesure que l'utilisation d'estimations sur petits domaines deviendra plus courante à Statistique Canada, il faudra ajouter des fonctionnalités au système pour répondre à cette demande. L'ouvrage récent de Rao et Molina (2015) donne une idée de l'ampleur de l'évolution des estimations sur petits domaines au cours des dernières années. Il faudrait énormément de temps pour intégrer tous ces progrès dans le système de production, cela coûterait cher et ne s'appliquerait peut-être pas directement aux besoins de Statistique Canada. Il faut donc envisager d'autres options que la programmation de ces nouvelles fonctionnalités dans l'actuel système de production en SAS. Une option consisterait à étudier de quelle manière des progiciels conçus ailleurs, comme ceux rédigés en *R*, pourraient y être intégrés. Parmi les progiciels rédigés en R, mentionnons *sae* (Molina et Marhuenda, 2015), *mme* (Lopez-Vizcaino, Lombardia et Morales, 2014), *saery* (Esteban, Morales et Perez, 2014) et *sae2* (Fay et Diallo, 2015). Ces progiciels comprennent des procédures sur petits domaines qui ne figurent pas dans le système actuel, comme des modèles linéaires mixtes multinomiaux, des modèles au niveau du domaine avec des effets temporels et des modèles au niveau du domaine en série chronologique soutenant des applications univariées et multivariées. Puisque le système de production actuel en SAS répond aux besoins de Statistique Canada à ce stade-ci, il n'y a pas de plans concrets pour y ajouter des fonctionnalités.

# Remerciements

Nous tenons à remercier les examinateurs pour leurs commentaires et suggestions qui ont permis d'améliorer le présent document.

# Annexe

## Justification du coefficient de détermination

Afin de déterminer un coefficient de détermination associé au modèle de lien, $\theta_i = \mathbf{z}_i^T\boldsymbol{\beta} + b_i v_i$, nous l'avons d'abord récrit comme suit :

$$\tilde{\theta}_i = \tilde{\mathbf{z}}_i^T\boldsymbol{\beta} + v_i,$$

où $\tilde{\theta}_i = \theta_i / b_i$ et $\tilde{\mathbf{z}}_i = \mathbf{z}_i / b_i$. Nous supposons qu'une ordonnée à l'origine est implicitement ou explicitement comprise dans $\tilde{\mathbf{z}}_i$; c'est-à-dire qu'il existe un vecteur $\boldsymbol{\lambda}$ de telle sorte que $\boldsymbol{\lambda}^T \tilde{\mathbf{z}}_i = 1$. En d'autres termes, nous supposons qu'il existe un vecteur $\boldsymbol{\lambda}$ de telle sorte que $b_i = \boldsymbol{\lambda}^T \mathbf{z}_i$. Si $\tilde{\theta}_i$, $i = 1, \ldots, m$, étaient connus, nous pourrions estimer le vecteur inconnu des paramètres du modèle $\boldsymbol{\beta}$ à l'aide de l'estimateur par les moindres carrés :

$$\hat{\boldsymbol{\beta}}_* = \left( \sum_{i=1}^m \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right)^{-1} \sum_{i=1}^m \tilde{\mathbf{z}}_i \tilde{\theta}_i$$

et la variance inconnue du modèle $\sigma_v^2$ à l'aide de l'estimateur sans biais :

$$\hat{\sigma}_{v*}^2 = \frac{\sum_{i=1}^m \left( \tilde{\theta}_i - \tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_* \right)^2}{m - q}.$$

Le coefficient de détermination ajusté bien connu est :

$$R^2_{\text{idéal}} = 1 - \frac{\hat{\sigma}_{v*}^2}{(m-1)^{-1} \sum_{i=1}^m \left( \tilde{\theta}_i - \bar{\tilde{\theta}} \right)^2}, \tag{A.1}$$

où $\bar{\tilde{\theta}} = m^{-1} \sum_{i=1}^m \tilde{\theta}_i$. Il s'agit d'un coefficient de détermination idéal parce qu'il ne peut pas être calculé (étant donné que $\tilde{\theta}_i$ est inconnu), mais c'est la cible que nous aimerions estimer. Le simple fait de remplacer $\theta_i$ par $\hat{\theta}_i$ ne règle pas le problème puisque $\hat{\theta}_i$ représente le modèle combiné et non simplement le modèle de lien. Le coefficient de détermination ainsi obtenu serait habituellement trop petit. Pour obtenir une meilleure estimation de $R^2_{\text{idéal}}$, nous devons d'abord décomposer $\sum_{i=1}^m \left( \tilde{\theta}_i - \bar{\tilde{\theta}} \right)^2$ comme ceci :

$$\sum_{i=1}^m \left( \tilde{\theta}_i - \bar{\tilde{\theta}} \right)^2 = \sum_{i=1}^m \left( \tilde{\theta}_i - \tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_* \right)^2 + \sum_{i=1}^m \left( \tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_* - \bar{\tilde{\theta}} \right)^2 + 2 \sum_{i=1}^m \left( \tilde{\theta}_i - \tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_* \right) \left( \tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_* - \bar{\tilde{\theta}} \right). \tag{A.2}$$

En supposant qu'une ordonnée à l'origine est implicitement ou explicitement comprise dans $\tilde{\mathbf{z}}_i$ et à partir de l'expression de $\hat{\boldsymbol{\beta}}_*$, nous obtenons ceci :

$$\sum_{i=1}^m \left( \tilde{\theta}_i - \tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_* \right) \tilde{\mathbf{z}}_i = \mathbf{0} \tag{A.3}$$

et

$$\sum_{i=1}^m \left( \tilde{\theta}_i - \tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_* \right) = 0. \tag{A.4}$$

En utilisant (A.4), nous pouvons récrire $\bar{\tilde{\theta}}$ sous la forme de $\bar{\tilde{\theta}} = \bar{\tilde{\mathbf{z}}}^T \hat{\boldsymbol{\beta}}_*$, où $\bar{\tilde{\mathbf{z}}} = m^{-1} \sum_{i=1}^m \tilde{\mathbf{z}}_i$. Par conséquent, le terme du produit croisé sous (A.2) s'estompe et l'équation (A.2) est ramenée à celle-ci :

$$\begin{aligned} \sum_{i=1}^m \left( \tilde{\theta}_i - \bar{\tilde{\theta}} \right)^2 &= \sum_{i=1}^m \left( \tilde{\theta}_i - \tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_* \right)^2 + \sum_{i=1}^m \left( \tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_* - \bar{\tilde{\mathbf{z}}}^T \hat{\boldsymbol{\beta}}_* \right)^2 \\ &= (m-q)\, \hat{\sigma}_{v*}^2 + (m-1)\, S^2 \left( \hat{\boldsymbol{\beta}}_* \right), \end{aligned} \tag{A.5}$$

où

$$S^2 \left( \hat{\boldsymbol{\beta}}_* \right) = \frac{\sum_{i=1}^m \left( \tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_* - \bar{\tilde{\mathbf{z}}}^T \hat{\boldsymbol{\beta}}_* \right)^2}{m-1}. \tag{A.6}$$

Selon (A.5), il s'ensuit que nous pouvons récrire le coefficient de détermination idéal (A.1) comme ceci :

$$R^2_{\text{idéal}} = 1 - \frac{\hat{\sigma}^2_{v*}}{\frac{(m-q)}{(m-1)}\hat{\sigma}^2_{v*} + S^2\left(\hat{\boldsymbol{\beta}}_*\right)} \equiv f(\hat{\boldsymbol{\beta}}_*, \hat{\sigma}^2_{v*}). \quad \text{(A.7)}$$

Les seules quantités inconnues dans (A.7) sont $\hat{\boldsymbol{\beta}}_*$ et $\hat{\sigma}^2_{v*}$. Nous pouvons donc obtenir un coefficient de détermination calculable en remplaçant $\hat{\boldsymbol{\beta}}_*$ et $\hat{\sigma}^2_{v*}$ dans (A.7) par $\hat{\boldsymbol{\beta}}$ et $\hat{\sigma}^2_v$, les estimateurs convergents de $\boldsymbol{\beta}$ et $\sigma^2_v$ mis en œuvre dans le système de l'EPD et décrits dans la section 3. Le coefficient de détermination ainsi obtenu peut s'exprimer comme $R^2 = f\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2_v\right)$, avec la fonction $f(\cdot,\cdot)$ définie dans (A.7), et il est un estimateur convergent du coefficient de détermination idéal $R^2_{\text{idéal}}$.

# Bibliographie


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

Beaumont, J.-F., et Bocci, C. (2016). Small area estimation in the Labour Force Survey. Document présenté au Comité consultatif des méthodes statistiques, mai 2016, Statistique Canada.

Brackstone, G.J. (1987). Small area data: Policy issues and technical challenges. Dans *Small Area Statistics*, (Éds., R. Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh), New York: John Wiley & Sons, Inc., 3-20.

Chib, S., et Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *American Statistician*, 49, 327-335.

Dick, P. (1995). Modélisation du sous-dénombrement net dans le recensement du Canada de 1991. *Techniques d'enquête*, 21, 1, 51-61. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1995001/article/14411-fra.pdf.

Drew, D., Singh, M.P. et Choudhry, G.H. (1982). Évaluation des techniques d'estimation pour les petites régions dans l'Enquête sur la population active au Canada. *Techniques d'enquête*, 8, 1, 19-52. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1982001/article/14328-fra.pdf.

Esteban, M.D., Morales, D. et Perez, A. (2014). saery: Small Area Estimation for Rao and Yu Model. URL http://CRAN.R-project.org/package=saery. R package version 1.0.

Estevao, V., Hidiroglou, M.A. et You, Y. (2015). *Area Level Model, Unit Level, and Hierarchical Bayes Methodology Specifications*. Document interne, Statistique Canada.

EURAREA (2004). *Enhancing Small Area Estimation Techniques to meet European Needs*. https://cordis.europa.eu/project/rcn/58374_en.html.

Fay, R.E., et Diallo, M. (2015). sae2: Small Area Estimation: Time-Series Models. URL https://CRAN.R-project.org/package=sae2. R package version 0.1-1.

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

Fuller, W.A., et Rao, J.N.K. (2001). Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada. *Techniques d'enquête*, 27, 1, 49-56. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2001001/article/5853-fra.pdf.

Gambino, J., Kennedy, B. et Singh, M.P. (2001). Estimation composite par regression pour l'Enquête sur la population active du Canada : Évaluation et application. *Techniques d'enquête*, 27, 1, 69-79. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2001001/article/5855-fra.pdf.

Gelfand, A.E., et Smith, A.F.M. (1990). Sample-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 972-985.

Ghangurde, P.D., et Singh, M.P. (1977). Synthetic estimation in periodic household surveys. *Techniques d'enquête*, 3, 2, 152-181.

Gonzalez, M.E., et Hoza, C. (1978). Small-area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73, 7-15.

Kott, P. (1989). Estimaton robuste pour petits domaines à l'aide du modèle des effets aléatoires. *Techniques d'enquête*, 15, 1, 3-13. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/1989001/article/14581-fra.pdf.

Li, H., et Lahiri, P. (2010). Adjusted maximum method in the small area estimation problem. *Journal of Multivariate Analysis*, 101, 882-892.

Lopez-Vizcaino, E., Lombardia, M.J. et Morales, D. (2014). mme: Multinomial Mixed Effects Models, 2014. URL http://CRAN.R-project.org/package=mme. R package version 0.1-5.

Molina, I., et Marhuenda, Y. (2015). sae: An R package for small area estimation. *The R Journal*, 7, 1, 81-98.

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, 163-171.

Prasad, N.G.N., et Rao, J.N.K. (1999). Estimation régionale robuste au moyen d'un modèle simple à effets aléatoires. *Techniques d'enquête*, 25, 1, 73-79. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/1999001/article/4713-fra.pdf.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

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

Rivest, L.-P., et Belmonte, E. (2000). Une erreur quadratique moyenne conditionnelle des estimateurs régionaux. *Techniques d'enquête*, 26, 1, 79-90. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2000001/article/5179-fra.pdf.

Rubin-Bleuer, S. (2014). *Specifications for EBLUP and Pseudo-EBLUP Estimators with Nonnegligible Sampling Fractions*. Document interne, Statistique Canada.

Rubin-Bleuer, S., Jang, L. et Godbout, S. (2016). The Pseudo-EBLUP estimator for a weighted average with an application to the Canadian Survey of Employment, Payrolls and Hours. *Journal of Survey Statistics and Methodology*, 4, 417-435.

Singh, M.P., et Tessier, R. (1976). Some estimators for domain totals. *Journal of the American Statistical Association*, 71, 322-325.

Singh, A.C., Kennedy, B. et Wu, S. (2001). Estimation composite par régression pour l'Enquête sur la population active du Canada avec plan de sondage à renouvellement de panel. *Techniques d'enquête*, 27, 1, 35-48. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2001001/article/5852-fra.pdf.

Stukel, D., et Rao, J.N.K. (1997). Small-area estimation under two-fold nested error regression model. *Journal of Statistical Planning and Inference*, 78, 131-147.

Wang, J., et Fuller, W.A. (2003). The mean square error of small area estimators constructed with estimated area variances. *Journal of American Statistical Association*, 98, 716-723.

Wang, J., Fuller, W.A. et Qu, Y. (2008). Estimation pour petits domaines sous une contrainte. *Techniques d'enquête*, 34, 1, 33-40. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2008001/article/10619-fra.pdf.

You, Y., et Rao, J.N.K. (2002). A pseudo empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, 30, 431-439.

You, Y., Rao, J.N.K. et Dick, P. (2004). Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, 6, 631-640.

You, Y., Rao, J.N.K. et Hidiroglou, M. (2013). De la performance des estimateurs sur petits domaines autocalés sous le modèle au niveau du domaine de Fay-Herriot. *Techniques d'enquête*, 39, 1, 243-255. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2013001/article/11830-fra.pdf.