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

Techniques d’enquête

# Rendement des estimateurs hiérarchiques bayésiens sur petits domaines reposant sur des distributions a priori non informatives et informatives et application à l’Enquête sur la population active du Canada

par Yong You et Keven Bosa

Date de diffusion : le 23 décembre 2025


Statistique Canada
Statistics Canada
Canada

## Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

**Courriel à** infostats@statcan.gc.ca

**Téléphone** entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

## Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

## Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Techniques d'enquête, décembre 2025
Vol. 51, nº 2, pp. 461-488
Statistique Canada, nº 12-001-X au catalogue



# Rendement des estimateurs hiérarchiques bayésiens sur petits domaines reposant sur des distributions *a priori* non informatives et informatives et application à l'Enquête sur la population active du Canada

**Yong You et Keven Bosa**[1]

Résumé

Dans le présent article, nous étudions le rendement des estimateurs hiérarchiques bayésiens (HB) sur petits domaines reposant sur des distributions *a priori* non informatives et informatives. Nous appliquons les modèles bayésiens de You et Chapman (2006) et de You (2021) aux données de l'Enquête sur la population active (EPA) du Canada et évaluons l'incidence des distributions *a priori* sur les estimateurs HB. Une comparaison des modèles bayésiens et une étude par simulation sont également réalisées. Nos résultats indiquent qu'une distribution *a priori* informative exacte peut produire de très bons résultats et que les distributions *a priori* non informatives peuvent aussi avoir un très bon rendement. Les distributions *a priori* informatives inexactes peuvent produire de mauvais résultats en matière de biais important et de coefficient de variation (CV) élevé. Les distributions *a priori* non informatives sont recommandées dans la pratique pour l'estimation HB sur petits domaines, à moins que des distributions *a priori* informatives correctement précisées soient disponibles. Les distributions *a priori* informatives sont particulièrement utiles lorsque le nombre de petits domaines est relativement faible.

**Mots-clés :** Biais; distribution prédictive *a posteriori*; échantillonnage de Gibbs; modèle de Fay-Herriot; OPC; sélection de modèle.

## 1. Introduction

Dans la plupart des applications, les enquêtes par sondage sont habituellement conçues pour fournir des estimations directes fiables pour de grands domaines au moyen de données d'échantillon propres à un domaine. Toutefois, ces estimations directes ne permettent pas, bien souvent, de produire des estimations fiables pour les petits domaines en raison des petites tailles d'échantillon. Les estimations d'enquête directes pour les petits domaines présentent souvent de grandes erreurs types inadaptées. Pour accroître la précision et la fiabilité, il est nécessaire de tirer parti de renseignements ou « d'emprunter » de l'information aux domaines apparentés et donc d'augmenter la taille efficace de l'échantillon en vue de produire des estimations indirectes pour les petits domaines. Les méthodes fondées sur un modèle explicite, qui s'appuient sur des données supplémentaires, telles que des données de recensement et des données administratives associées aux petits domaines, ont été largement utilisées dans la pratique pour obtenir des estimations fiables fondées sur un modèle. Pour une vue d'ensemble et une évaluation des modèles pour l'estimation sur petits domaines, voir Rao et Molina (2015). Dans le présent article, nous étudions le rendement des estimateurs hiérarchiques bayésiens (HB) sur petits domaines au moyen de modèles au niveau du domaine.

L'avantage d'utiliser des modèles au niveau du domaine est que ces modèles tiennent compte du plan de sondage par l'utilisation d'estimations d'enquête directes et d'estimations de la variance fondées sur le plan

1. Yong You et Keven Bosa, Division des méthodes de la statistique économique, Statistique Canada, Ottawa, Ontario, Canada, K1A 0T6. Courriels : yong.you@statcan.gc.ca et keven.bosa@statcan.gc.ca.

de sondage connexes. Dans la pratique, on peut utiliser les modèles au niveau du domaine chaque fois que des estimations d'enquête directes et que des variables auxiliaires au niveau du domaine sont disponibles. Parmi les modèles au niveau du domaine, celui de Fay-Herriot (Fay et Herriot, 1979) est un modèle largement utilisé dans l'estimation sur petits domaines. Le modèle de Fay-Herriot comporte deux composantes : un modèle d'échantillonnage pour les estimations d'enquête directes et un modèle de liaison pour les paramètres de petits domaines d'intérêt. Le modèle d'échantillonnage suppose que, compte tenu de la taille d'échantillon de domaine spécifique $n_i > 1,$ pour le $i^{\text{e}}$ domaine, il existe un estimateur d'enquête direct $y_i$, habituellement sans biais par rapport au plan de sondage, pour le paramètre de petits domaines $\theta_i,$ de sorte que

$$y_i = \theta_i + e_i,\ i = 1, \ldots, m, \tag{1.1}$$

où $e_i$ est l'erreur d'échantillonnage liée à l'estimateur direct $y_i$ et où $m$ est le nombre de petits domaines. Dans la pratique, il est courant de supposer que les variables $e_i$ sont des variables aléatoires normales indépendantes de moyenne $E(e_i) = 0$ et de variance d'échantillonnage $\text{Var}(e_i) = \sigma_i^2$. Le modèle de liaison suppose que le paramètre de petits domaines d'intérêt $\theta_i$ est lié à la variable auxiliaire au niveau du domaine $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ par un modèle de régression linéaire qui est

$$\theta_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i, i = 1, \ldots, m, \tag{1.2}$$

où $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ est un vecteur $p \times 1$ de coefficients de régression et où les $v_i$ sont des effets aléatoires propres au domaine que nous supposons indépendants et identiquement distribués (i.i.d.) avec $E(v_i) = 0$ et $\text{Var}(v_i) = \sigma_v^2$. L'hypothèse de normalité est aussi généralement incluse, même s'il est plus difficile de justifier l'hypothèse. La variance $\sigma_v^2$ est appelée la variance du modèle, et elle est inconnue et doit être estimée à partir des données. Les effets aléatoires au niveau du domaine $v_i$ rendent compte de l'hétérogénéité non structurée entre les domaines que n'explique pas la variance d'échantillonnage $\sigma_i^2$. La combinaison des modèles (1.1) et (1.2) produit un modèle linéaire mixte au niveau du domaine donné par

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i + e_i, i = 1, \ldots, m. \tag{1.3}$$

Le modèle (1.3) comprend des erreurs aléatoires fondées sur le plan $e_i$ et des effets aléatoires fondés sur le modèle $v_i$. Aux fins du modèle de Fay-Herriot, la variance d'échantillonnage $\sigma_i^2$ est présumée connue. Dans la pratique, nous devons estimer la variance d'échantillonnage $\sigma_i^2$. Dans certaines applications, l'estimation directe de la variance d'échantillonnage est utilisée directement dans le modèle de Fay-Herriot, mais cette estimation est généralement instable pour les petits domaines. Une approche par lissage ou par modélisation est souvent utilisée pour obtenir une estimation plus stable de la variance d'échantillonnage. Dans le présent article, nous envisageons une approche de modélisation pour $\sigma_i^2$ grâce à l'utilisation de l'estimateur direct de la variance d'échantillonnage dans un cadre hiérarchique bayésien (HB) comme suit.

Supposons que $s_i^2$ désigne l'estimateur direct de la variance d'échantillonnage $\sigma_i^2$. Nous considérons un modèle couramment utilisé pour $s_i^2$ comme étant $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2,$ où $d_i = n_i - 1$ et $n_i$ est la taille de l'échantillon pour le $i^{\text{e}}$ domaine. Par exemple, supposons que nous ayons $n_i$ observations provenant du

petit domaine $i$ et que ces observations ont une distribution identique $N(\mu_i, \sigma^2)$. Supposons que $y_i$ soit la moyenne d'échantillon des $n_i$ observations. Ensuite, $y_i \sim N(\mu_i, \sigma_i^2)$ avec $\sigma_i^2 = \sigma^2 / n_i$. Nous pouvons alors obtenir un estimateur de $\sigma_i^2$ sous la forme $s_i^2 = s^2 / n_i$, où $s^2$ est la variance d'échantillon des $n_i$ observations. En outre, $y_i$ et $s_i^2$ sont indépendants et $(n_i - 1)\, s_i^2 \sim \sigma_i^2 \chi_{n_i-1}^2$. You et Chapman (2006) ont retenu une approche HB et ont combiné le modèle de variance d'échantillonnage $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$ au modèle de petits domaines (1.3) pour construire un modèle intégré. Le modèle ainsi obtenu tire sa puissance des estimations sur petits domaines et des estimations de la variance d'échantillonnage simultanément. L'approche intégrée de modélisation HB avec $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$ a été largement utilisée dans la pratique, notamment par You (2008a, 2021), Dass, Maiti, Ren et Sinha (2012), Sugasawa, Tamae et Kubokawa (2017), Ghosh, Myung et Moura (2018), et Hidiroglou, Beaumont et Yung (2019), entre autres. L'un des avantages de l'approche HB est qu'elle est simple et que les inférences pour les paramètres $\theta_i$ reposent sur des échantillons *a posteriori*. L'approche HB tient compte automatiquement des incertitudes associées aux paramètres inconnus, mais elle exige la spécification de distributions *a priori* pour les paramètres de modèle $\boldsymbol{\beta}$ et $\sigma_v^2$ dans le modèle de Fay-Herriot.

Dans le présent article, nous étudions des estimateurs HB sur petits domaines à l'aide de distributions *a priori* non informatives et informatives pour $\boldsymbol{\beta}$ et $\sigma_v^2$ au moyen d'une analyse sur données réelles et d'une étude par simulation. Dans l'analyse bayésienne, la spécification *a priori* est importante et assez difficile. De façon générale, la distribution *a priori* des paramètres du modèle pourrait être informative, faiblement informative ou non informative pour les modèles HB. Dans cet article, nous évaluons l'incidence du choix des distributions *a priori* sur les estimations sur petits domaines, en particulier pour la spécification de la distribution *a priori* du paramètre de régression $\boldsymbol{\beta}$ dans le modèle. Le choix d'une distribution *a priori* peut comprendre certaines étapes ou certains critères, par exemple, l'utilisation du contexte et de l'objet de l'analyse des données, la connaissance des données d'enquête sur de petits domaines et des critères comme la cohérence, c'est-à-dire que la distribution *a priori* doit être conforme à la logique et à la structure du modèle HB et de l'espace des paramètres. La distribution *a priori* doit être conjuguée, c'est-à-dire qu'elle doit être mathématiquement compatible avec la fonction de vraisemblance du modèle; elle doit être compatible avec les données de sorte qu'elle ne contredise pas ni ne domine les faits issus des données d'échantillonnage. La distribution *a priori* doit également témoigner des connaissances et des croyances *a priori* à l'égard du paramètre, en fonction des sources, des données de recensement et des données administratives pertinentes. Dans la pratique, des renseignements et des connaissances variés peuvent être utilisés pour définir des distributions *a priori* informatives, en particulier pour la spécification de moyenne *a priori* du paramètre de régression. Par exemple, Sakshaug, Wiśniowski, Ruiz et Blom (2019) et Wiśniowski, Sakshaug, Ruiz et Blom (2020) ont démontré que des distributions *a priori* informatives fondées sur des données non probabilistes peuvent entraîner des écarts dans les variances et les erreurs quadratiques moyennes pour les coefficients du modèle linéaire. En supposant un plan de sondage probabiliste ignorable, Sakshaug et coll. (2019) et Wiśniowski et coll. (2020) ont proposé et évalué au moyen d'études par simulation une approche bayésienne pour l'intégration des données des deux échantillons aux fins de l'estimation des paramètres d'un modèle, en précisant des distributions *a priori* informatives pour les paramètres du

modèle : les données de l'échantillon non probabiliste servent à déterminer la moyenne *a priori*, mais les données des deux échantillons sont utilisées pour déterminer la variance *a priori*, c'est-à-dire que le biais estimé de l'estimateur de l'échantillon non probabiliste sert à faire augmenter la variance *a priori*. En général, une distribution *a priori* pour les paramètres d'un modèle est davantage précisée dans l'optique de l'application et elle dépend dans de nombreux cas de l'application précise et des données réelles disponibles. Pour des exemples et des conseils, nous citons certaines références comme Consonni, Fouskakis, Liseo et Ntzoufras (2018), Grzenda (2016), Lemoine (2019), Mikkola, Martin, Chandramouli, Hartmann, Abril Pla, Thomas, Pesonen, Corander, Vehtari, Kaski, Bürkner et Klami (2024), et Zondervan-Zwijnenburg, Peeters, Depaoli et Van de Schoot (2017).

Pour la spécification des distributions *a priori* pour les composantes de la variance dans le modèle HB sur petits domaines, une distribution *a priori* gamma inverse (GI) est une distribution *a priori* adéquate et conjuguée de manière conditionnelle pour les composantes de la variance. Une distribution *a priori* GI est largement utilisée dans la documentation bayésienne (par exemple Gelman, Carlin, Stern et Rubin, 2004) et dans les logiciels d'inférence bayésienne (par exemple WinBUGS, Lunn, Thomas, Best et Spiegelhalter, 2000). You (2023) a démontré l'utilisation efficace d'une distribution *a priori* GI dans l'estimation HB sur petits domaines et son application à l'EPA. Par ailleurs, des distributions *a priori* uniformes peuvent être utilisées pour les variances du modèle. Une distribution *a priori* uniforme est généralement utilisée comme distribution *a priori* non informative dans la documentation (par exemple Gelman, 2006).

Dans le présent article, nous étudions les estimations HB sur petits domaines en fonction de différentes spécifications *a priori*, en particulier pour la distribution *a priori* du paramètre de régression dans le modèle HB sur petits domaines. L'article est structuré de la manière suivante. À la section 2, nous présentons les modèles HB sur petits domaines et les procédures correspondantes de Gibbs pour l'inférence fondée sur l'échantillonnage. À la section 3, nous comparons les estimations HB en fonction de différentes spécifications *a priori* par l'application d'une analyse de données réelles. À la section 4, nous présentons la sélection du modèle bayésien et l'analyse de la vérification du modèle. À la section 5, nous menons une étude par simulation pour évaluer les estimateurs HB d'après différentes spécifications *a priori*. Et à la section 6, nous comparons et étudions les estimations HB pour de petits domaines groupés au moyen de distributions *a priori* informatives et uniformes. Enfin, à la section 7, nous présentons la conclusion.

## 2. Spécification du modèle hiérarchique bayésien et inférence

Dans la présente section, nous présentons l'approche HB à l'aide de la méthode d'échantillonnage de Gibbs pour le modèle de Fay-Herriot lorsque les variances d'échantillonnage sont estimées. En particulier, nous considérons deux modèles HB pour l'estimation sur petits domaines au moyen de distributions *a priori* non informatives et informatives, à savoir les modèles proposés par You et Chapman (2006) et You (2021). Les modèles HB et les spécifications *a priori* sont les suivants.

Modèle HB 1 : Modèle de You-Chapman (You et Chapman, 2006), désigné par MYC :

- $y_i \mid \theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2),\ i = 1, \ldots, m;$
- $d_i s_i^2 \mid \sigma_i^2 \sim \sigma_i^2 \chi_{d_i}^2,\ d_i = n_i - 1,\ i = 1, \ldots, m;$
- $\theta_i \mid \boldsymbol{\beta}, \sigma_v^2 \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma_v^2),\ i = 1, \ldots, m;$
- Distribution *a priori* pour le paramètre de régression $\boldsymbol{\beta}$ :
  (1) Une distribution *a priori* uniforme pour $\boldsymbol{\beta}$ : $\pi(\boldsymbol{\beta}) \propto 1$;
  (2) Une distribution *a priori* informative normale pour $\boldsymbol{\beta}$ : $\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Phi}_0)$, où $\boldsymbol{\beta}_0$ est connu et $\boldsymbol{\Phi}_0$ est une matrice de covariance dont la variance est prédéterminée;
- Distribution *a priori* pour les composantes de la variance $\sigma_v^2$ et $\sigma_i^2$ : les distributions *a priori* gamma inverses sont $\pi(\sigma_v^2) \sim \text{GI}(a_v, b_v)$, $\pi(\sigma_i^2) \sim \text{GI}(a_i, b_i)$, où $a_v$, $b_v$ et $a_i$, $b_i$ sont des constantes prédéterminées. En particulier, $a_i, b_i$ $(0 \le i \le m)$ sont des constantes choisies très petites, disons 0,00001, pour traduire la vague connaissance de $\sigma_i^2$, comme dans You et Chapman (2006).

Nous nous intéressons principalement à l'estimation du paramètre de petits domaines $\theta_i$. Nous utilisons l'approche d'échantillonnage de Gibbs pour obtenir l'estimation *a posteriori* de $\theta_i$. Supposons que $\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \left(\sum_{i=1}^m \mathbf{x}_i \theta_i\right)$, et $\mathbf{V}_\beta = \sigma_v^2 \left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i'\right)^{-1}$, $\boldsymbol{\theta}' = (\theta_1, \ldots, \theta_m)$, alors les distributions conditionnelles complètes pour la procédure d'échantillonnage de Gibbs correspondant au modèle HB 1 sont données par :

- $\left[\theta_i \mid y_i, \boldsymbol{\beta}, \sigma_v^2\right] \sim \text{N}\left(\gamma_i y_i + (1 - \gamma_i)\, \mathbf{x}_i'\boldsymbol{\beta}, \gamma_i \sigma_i^2\right)$, où $\gamma_i = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_i^2}$, pour $i = 1, \ldots, m$;
- Nous avons les distributions conditionnelles suivantes correspondant à la spécification *a priori* pour $\boldsymbol{\beta}$ :
  (1) $\left[\boldsymbol{\beta} \mid y_i, \boldsymbol{\theta}, \sigma_v^2\right] \sim N\left(\hat{\boldsymbol{\beta}}, \mathbf{V}_\beta\right)$, si $\pi(\boldsymbol{\beta}) \propto 1$;
  (2) $\left[\boldsymbol{\beta} \mid y_i, \boldsymbol{\theta}, \sigma_v^2\right] \sim N\left(\tilde{\boldsymbol{\beta}}, \boldsymbol{\Psi}_{\boldsymbol{\beta}}\right)$, si $\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Phi}_0)$, où $\tilde{\boldsymbol{\beta}} = \boldsymbol{\Psi}_{\boldsymbol{\beta}}\left(\boldsymbol{\Phi}_0^{-1}\boldsymbol{\beta}_0 + \mathbf{V}_{\boldsymbol{\beta}}^{-1}\hat{\boldsymbol{\beta}}\right)$ et $\boldsymbol{\Psi}_\beta = \left(\boldsymbol{\Phi}_0^{-1} + \mathbf{V}_{\boldsymbol{\beta}}^{-1}\right)^{-1}$;
- $\left[\sigma_v^2 \mid y_i, \boldsymbol{\theta}, \boldsymbol{\beta}\right] \sim \text{GI}\left(a_v + \frac{m}{2}, b_v + \frac{1}{2}\sum_{i=1}^m \left(\theta_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2\right)$;
- $\left[\sigma_i^2 \mid y_i, \boldsymbol{\theta}, s_i^2\right] \sim \text{GI}\left(a_i + \frac{n_i}{2}, b_i + \frac{1}{2}\left((y_i - \theta_i)^2 + (n_i - 1)\, s_i^2\right)\right)$.

Le second modèle HB est semblable au modèle MYC, mais comporte un modèle log-linéaire pour la variance d'échantillonnage (You, 2021). Nous présentons maintenant le modèle comprenant les spécifications *a priori* comme suit :

Modèle HB 2 : Modèle log-linéaire de You (2021), désigné par MLLY :

- $y_i \mid \theta_i, \sigma_i^2 \sim N\left(\theta_i, \sigma_i^2\right), i = 1, \ldots, m;$
- $d_i s_i^2 \mid \sigma_i^2 \sim \sigma_i^2 \chi_{d_i}^2,\ d_i = n_i - 1, i = 1, \ldots, m;$
- $\theta_i \mid \boldsymbol{\beta}, \sigma_v^2 \sim N\left(\mathbf{x}_i'\boldsymbol{\beta}, \sigma_v^2\right), i = 1, \ldots, m;$

- $\log(\sigma_i^2) \mid \boldsymbol{\delta}, \tau^2 \sim N\left(z_i'\delta, \tau^2\right)$, $\boldsymbol{\delta}' = (\delta_1, \delta_2)$, $\mathbf{z}_i' = (1, \log(n_i))$, $i = 1, \ldots, m$;
- Distribution *a priori* pour le paramètre de régression $\boldsymbol{\beta}$ :
  (1) Une distribution *a priori* uniforme pour $\boldsymbol{\beta}$ : $\pi(\boldsymbol{\beta}) \propto 1$;
  (2) Une distribution *a priori* informative normale pour $\boldsymbol{\beta}$ : $\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Phi}_0)$, où $\boldsymbol{\beta}_0$ est connu et $\boldsymbol{\Phi}_0$ est une matrice de covariance dont la variance est prédéterminée;
- Distribution *a priori* pour la variance du modèle $\sigma_v^2$ : $\pi(\sigma_v^2) \sim \mathrm{GI}(a_v, b_v)$, où $a_v, b_v$ sont des constantes prédéterminées;
- Spécification *a priori* pour les hyperparamètres $\boldsymbol{\delta}$ et $\tau^2$ : $\pi(\boldsymbol{\delta}) \propto 1$, $\pi(\tau^2) \sim \mathrm{GI}(a_\tau, b_\tau)$, et $a_\tau$, $b_\tau$ sont fixés à 0,00001.

Les distributions conditionnelles complètes et la procédure d'échantillonnage pour le MLLY sont données comme suit :

- $\left[\theta_i \mid y_i, \boldsymbol{\beta}, \sigma_i^2, \sigma_v^2\right] \sim N\left(\gamma_i y_i + (1-\gamma_i)\mathbf{x}_i'\boldsymbol{\beta}, \gamma_i\sigma_i^2\right)$, où $\gamma_i = \frac{\sigma_v^2}{\sigma_v^2+\sigma_i^2}$, $i = 1, \ldots, m$;
- Pour $\boldsymbol{\beta}$, les distributions conditionnelles correspondant aux distributions *a priori* pour $\boldsymbol{\beta}$ sont présentées ci-dessous :
  (1) $\left[\boldsymbol{\beta} \mid y_i, \boldsymbol{\theta}, \sigma_v^2\right] \sim N\left(\hat{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}}\right)$, si $\pi(\boldsymbol{\beta}) \propto 1$;
  (2) $\left[\boldsymbol{\beta} \mid y_i, \boldsymbol{\theta}, \sigma_v^2\right] \sim N\left(\tilde{\boldsymbol{\beta}}, \boldsymbol{\Psi}_{\boldsymbol{\beta}}\right)$, si $\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Phi}_0)$, où $\tilde{\boldsymbol{\beta}} = \boldsymbol{\Psi}_{\boldsymbol{\beta}}\left(\boldsymbol{\Phi}_0^{-1}\boldsymbol{\beta}_0 + \mathbf{V}_{\boldsymbol{\beta}}^{-1}\hat{\boldsymbol{\beta}}\right)$ et $\boldsymbol{\Psi}_{\boldsymbol{\beta}} = \left(\boldsymbol{\Phi}_0^{-1} + \mathbf{V}_{\beta}^{-1}\right)^{-1}$;
- $\left[\sigma_v^2 \mid y_i, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_i^2\right] \sim \mathrm{GI}\left(a_v + \frac{m}{2}, b_v + \frac{1}{2}\sum_{i=1}^m (\theta_i - \mathbf{x}_i'\boldsymbol{\beta})^2\right)$;
- $\left[\sigma_i^2 \mid y_i, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \delta, \tau^2\right] \propto f\left(\sigma_i^2\right) \cdot h\left(\sigma_i^2\right)$, où $f\left(\sigma_i^2\right)$ et $h\left(\sigma_i^2\right)$ sont $f\left(\sigma_i^2\right) \sim \mathrm{GI}\left(\frac{d_i+1}{2}, \frac{(y_i-\theta_i)^2+d_i s_i^2}{2}\right)$ et $h\left(\sigma_i^2\right) = \exp\left(-\frac{(\log(\sigma_i^2) - z_i'\delta)^2}{2\tau^2}\right)$;
- $\left[\delta \mid y_i, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_i^2, \sigma_v^2, \tau^2\right] \sim N_2\left(\left(\sum_{i=1}^m \mathbf{z}_i\mathbf{z}_i'\right)^{-1}\left(\sum_{i=1}^m \mathbf{z}_i \log(\sigma_i^2)\right), \tau^2\left(\sum_{i=1}^m \mathbf{z}_i\mathbf{z}_i'\right)^{-1}\right)$;
- $\left[\tau^2 \mid y_i, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_i^2, \sigma_v^2, \boldsymbol{\delta}\right] \sim \mathrm{GI}\left(a_\tau + \frac{m}{2}, b_\tau + \frac{1}{2}\sum_{i=1}^m \left(\log(\sigma_i^2) - \mathbf{z}_i'\boldsymbol{\delta}\right)^2\right)$.

Comme dans You (2021), nous passons par l'étape de rejet de l'algorithme de Metropolis-Hastings (MH) (Chip et Greenberg, 1995) pour actualiser $\sigma_i^2$ dans la procédure d'échantillonnage de Gibbs. À l'itération $k$ dans la procédure d'échantillonnage de Gibbs :

(1) Tirer $\sigma_i^{2*}$ de $\mathrm{GI}\left(\frac{d_i+1}{2}, \frac{(y_i-\theta_i)^2+d_i s_i^2}{2}\right)$;

(2) calculer la probabilité d'acceptation $\alpha\left(\sigma_i^{2*}, \sigma_i^{2(k)}\right) = \min\left\{h\left(\sigma_i^{2*}\right) / h\left(\sigma_i^{2(k)}\right), 1\right\}$;

(3) tirer $u$ de la distribution uniforme (0,1), si $u < \alpha\left(\sigma_i^{2*}, \sigma_i^{2(k)}\right)$, la valeur candidate $\sigma_i^{2*}$ est acceptée, $\sigma_i^{2(k+1)} = \sigma_i^{2*}$; dans le cas contraire $\sigma_i^{2*}$ est rejeté et donne $\sigma_i^{2(k+1)} = \sigma_i^{2(k)}$.

L'étape de rejet de l'algorithme de Metropolis-Hastings est assez efficace dans la procédure d'échantillonnage de Gibbs et il n'y a pas de problème de calcul pour actualiser $\sigma_i^2$. Nous surveillons la procédure d'échantillonnage de Gibbs en vérifiant la convergence et le taux de rejet de l'algorithme de MH. Le taux de rejet moyen est assez faible dans notre application comme démontré à la section 3. Parker, Holan et Janicki (2023) ont utilisé une approche très semblable qui modélise les variances sur l'échelle logarithmique au moyen d'une distribution log-gamma multivariée (Bradley, Holan et Wikle, 2018, 2020) pour contourner la nécessité de suivre les étapes d'échantillonnage de l'algorithme de MH pour toutes les composantes de la variance $\sigma_i^2$. Parker, Holan et Janicki (2023) ont terminé leur modèle en établissant des paramètres qui témoignent de distributions *a priori* relativement vagues. Il pourrait être utile et intéressant d'étudier de quelle façon des distributions *a priori* informatives et non informatives peuvent être construites au moyen d'une distribution log-gamma multivariée. D'autres travaux étroitement liés à la modélisation de la variance d'échantillonnage comprennent par exemple ceux de Maiti, Ren et Sinha (2014), et de Sugasawa, Tamae et Kubokawa (2017).

Nous nous intéressons à l'estimation des paramètres de petits domaines $\theta_i$ et au coefficient de variation (CV) correspondant d'après différentes distributions *a priori* pour le paramètre de régression $\boldsymbol{\beta}$ et le modèle de variance $\sigma_v^2$. À la section suivante, nous comparons les estimations HB au moyen de l'analyse des données sur le taux de chômage de l'EPA du Canada.

# 3. Application à l'Enquête sur la population active

Dans la présente section, nous appliquons les deux modèles HB présentés à la section 2 à l'estimation du taux de chômage d'après les données de l'EPA pour évaluer les spécifications *a priori*. L'EPA du Canada permet de produire des estimations mensuelles du taux de chômage à l'échelle nationale et provinciale partout au Canada. L'EPA diffuse aussi des estimations du taux de chômage pour des régions infraprovinciales, comme les régions métropolitaines de recensement (RMR) et les agglomérations de recensement (AR) à l'échelle du Canada. Les détails de la méthodologie de l'EPA sont présentés dans le document Méthodologie de l'Enquête sur la population active du Canada (2017). Les variances estimées pour l'EPA sont calculées au moyen de la procédure bootstrap de Rao-Wu. Pour certaines régions infraprovinciales, les estimations directes ne sont pas fiables, car les tailles d'échantillon dans certaines régions sont très petites. La procédure d'estimation sur petits domaines, comme celle appliquée aux données de l'EPA, permet généralement d'estimer les taux de chômage pour des régions infraprovinciales locales comme les RMR/AR au moyen de modèles sur petits domaines. Ces modèles sur petits domaines et ces méthodes sont examinés, par exemple, dans Hidiroglou, Beaumont et Yung (2019), Lesage, Beaumont et Bocci (2021), You, Rao et Gambino (2003), et You (2008a, 2021). Nous appliquons les modèles HB décrits à la section 2 aux données mensuelles de 2011 des estimations du taux de chômage à l'échelle des RMR/AR afin de démontrer

l'utilisation de distributions *a priori* non informatives et informatives pour les estimations HB sur petits domaines. Au total, 65 RMR/AR (régions) sont prises en compte dans notre étude. La taille de l'échantillon pour les 65 RMR/AR varie de 58 à 3 773. La taille de l'échantillon est inférieure à 120 pour 9 régions et est supérieure à 1 000 pour 17 régions, y compris de grandes villes comme Toronto, Montréal et Vancouver.

Nous utilisons les données mensuelles sur le taux de chômage de l'EPA de 2011 dans notre étude d'application, car il y a eu un recensement en mai 2011. Pour le mois de mai, nous avons une estimation du taux de chômage provenant des données du recensement, qui sera considérée comme étant notre norme par excellence. Les estimations sur petits domaines de mai 2011 pour de nombreux scénarios HB seront comparées aux estimations du recensement de la présente section.

## 3.1 Spécification *a priori*

Nous avons d'abord utilisé les données mensuelles sur le taux de chômage de l'EPA de 2011 en fonction de 65 RMR/AR à l'échelle du Canada pour obtenir des estimations de $\boldsymbol{\beta}$ et le modèle de variance $\sigma_v^2$. Nous appliquons le modèle 1, le modèle de You-Chapman (2006) comprenant les distributions *a priori* non informatives, aux données mensuelles de 2011 sur le taux de chômage de janvier à décembre 2011, en omettant le mois de mai 2011. Nous prenons ensuite la moyenne des estimations *a posteriori* et les estimations de la variance *a posteriori* des paramètres du modèle en utilisant ces 11 mois d'estimations HB pour construire des estimations relativement stables comme distribution *a posteriori* pour $\boldsymbol{\beta}$ et $\sigma_v^2$.

Les distributions *a posteriori* sont précisées par la distribution normale et la distribution gamma inverse (GI) comme suit :

- $\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Phi}_0)$, où $\boldsymbol{\beta}_0 = (0{,}0507; 0{,}4255)'$, et
$$\boldsymbol{\Phi}_0 = \begin{pmatrix} 2{,}070245e-05 & -0{,}0002 \\ -0{,}0002 & 0{,}00405 \end{pmatrix};$$
- $\pi(\sigma_v^2) \sim$ GI(17,42; 0,00272), alors la moyenne *a priori* de $\sigma_v^2$ est 0,0001662.

La spécification *a priori* ainsi obtenue est présentée aux fins d'illustration seulement, de façon à avoir une distribution *a priori* pour $\boldsymbol{\beta}$ aussi stable que possible et se rapprochant de la vraie valeur de $\boldsymbol{\beta}$ dans le modèle reposant sur les données de 2011. La distribution *a priori* $\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Phi}_0)$ est définie comme la distribution *a priori* informative par défaut ou la distribution *a priori* exacte dans notre étude sur l'application à l'EPA. Dans la pratique, il n'est pas facile de savoir de quelle façon construire et définir des distributions *a priori* exactes pour les paramètres du modèle.

Pour comparer les distributions *a priori* non informatives et informatives, nous examinons les quatre spécifications *a priori* et les combinaisons suivantes :

- Cas de distribution *a priori* 1 : Distribution *a priori* non informative pour $\boldsymbol{\beta}: \pi(\boldsymbol{\beta}) \propto 1$, et distribution *a priori* vague pour $\sigma_v^2: \pi(\sigma_v^2) \sim$ GI(0,0001; 0,0001) en choisissant $a_v = b_v =$ 0,0001.

- Cas de distribution *a priori* 2 : Distribution *a priori* non informative pour $\boldsymbol{\beta}: \pi(\boldsymbol{\beta}) \propto 1$, et distribution *a priori* informative pour $\sigma_v^2: \pi(\sigma_v^2) \sim$ GI(17,42; 0,00272).
- Cas de distribution *a priori* 3 : Distribution *a priori* informative pour $\boldsymbol{\beta}: \pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Phi}_0)$, et distribution *a priori* vague pour $\sigma_v^2: \pi(\sigma_v^2) \sim$ GI(0,0001; 0,0001) en choisissant $a_v = b_v =$ 0,0001.
- Cas de distribution *a priori* 4 : Distribution *a priori* informative pour $\boldsymbol{\beta}: \pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Phi}_0)$, et distribution *a priori* informative pour $\sigma_v^2: \pi(\sigma_v^2) \sim$ GI(17,42; 0,00272).

En résumé, les distributions *a priori* pour le cas 1 et le cas 2 sont non informatives pour $\boldsymbol{\beta}$, tandis que les distributions *a priori* pour le cas 3 et le cas 4 sont informatives pour $\boldsymbol{\beta}$. Il importe de souligner que, pour les composantes de la variance, les distributions *a priori* uniformes $\pi(\sigma_v^2) \propto 1$ et $\pi(\tau^2) \propto 1$ peuvent également être utilisées pour les variances du modèle $\sigma_v^2$ et $\tau^2$ dans les modèles MYC et MLLY. Cependant, il n'y a aucun avantage à utiliser une distribution *a priori* uniforme pour les composantes de la variance dans les modèles MYC et MLLY par rapport à l'utilisation d'une distribution *a priori* GI pour ce qui est des estimations HB finales. Les modèles HB MYC et MLLY reposant sur la distribution *a priori* GI peuvent être très efficaces et permettent d'obtenir un meilleur ajustement du modèle que les modèles qui reposent sur une distribution *a priori* uniforme pour la composante de la variance, comme démontré dans You (2023). Par conséquent, dans notre étude, nous utilisons la distribution *a priori* GI vague GI(0,0001; 0,0001) et la distribution *a priori* GI informative pour $\sigma_v^2$.

## 3.2 Application aux données de l'Enquête sur la population active

Nous avons appliqué le modèle HB 1 (MYC) et le modèle 2 (MLLY) aux estimations directes du taux de chômage de l'EPA pour mai 2011 pour 65 RMR/AR à l'échelle du Canada avec les quatre types de spécifications *a priori* pour $\boldsymbol{\beta}$ et $\sigma_v^2$. Une proportion mensuelle de prestataires locaux de l'assurance-emploi est utilisée comme variable auxiliaire $x_i$ dans les modèles HB. On peut notamment consulter une description plus détaillée de la spécification de modèle fondée sur les données de l'EPA dans Hidiroglou, Beaumont et Yung (2019), You (2021), et You et Hidiroglou (2023).

Afin d'exécuter l'échantillonnage de Gibbs pour le MYC et le MLLY, nous avons une durée de « rodage » de $B =$ 2 000 et une taille d'échantillon de Gibbs de $G =$ 50 000. Pour le MLLY, pour différentes combinaisons de spécifications *a priori*, le taux de rejet moyen de l'algorithme de Metropolis-Hastings pour les 65 régions se situe autour de 8,12 %. Pour le modèle HB MLLY, nous prenons chaque 5$^{e}$ itération après la période de « rodage » et nous pouvons ainsi réduire l'autocorrélation pouvant résulter de l'algorithme d'acceptation-rejet de Metropolis-Hastings pour $\sigma_i^2$ dans l'exécution de l'échantillonnage de Gibbs. Par conséquent, pour le MLLY, nous avons en réalité une taille d'échantillon de Gibbs de 10 000 pour calculer les estimations *a posteriori*. La convergence de l'échantillonnage de Gibbs est surveillée pour le paramètre de petits domaines $\theta_i$, le paramètre de régression $\boldsymbol{\beta}$ et la composante de la variance $\sigma_v^2$ dans le modèle au moyen du facteur potentiel de réduction d'échelle (Gelman et Rubin, 1992; Gelman, Carlin,

Stern et Rubin, 2004). Nous avons calculé les facteurs de réduction pour tous les paramètres surveillés dans le modèle dans l'échantillonnage de Gibbs. Les valeurs de ces facteurs sont toutes très proches de 1 (moins de 1,05), ce qui donne à penser que la convergence désirée pour ces paramètres est obtenue par l'échantillonneur de Gibbs.

Nous comparons les estimations HB avec la valeur du recensement pour mai 2011 et calculons l'erreur relative absolue (ERA) et le coefficient de variation (CV). Les estimations HB sur petits domaines sont comparées au moyen de l'ERA des estimations directes et HB pour ce qui est des estimations du recensement pour chaque RMR/AR comme suit :

$$\text{ERA}_i = \left| \frac{\theta_i^{\text{Recensement}} - \theta_i^{\text{EST}}}{\theta_i^{\text{Recensement}}} \right|,$$

où $\theta_i^{\text{EST}}$ est l'estimation directe ou HB et $\theta_i^{\text{Recensement}}$ est la valeur correspondante du recensement pour le taux de chômage de l'EPA. Nous obtenons ensuite l'ERA moyenne pour toutes les régions. Le tableau 3.1 présente les résultats de l'ERA moyenne et du CV moyen selon le MYC (modèle 1), et le tableau 3.2 présente les résultats d'après le MLLY (modèle 2).

Dans le tableau 3.1, pour le MYC, les estimations HB ont amélioré les estimations directes en réduisant l'ERA et le CV. En utilisant une distribution *a priori* non informative pour $\beta$, l'ERA est de 0,1348 dans le cas d'une distribution *a priori* vague pour $\sigma_v^2$ et de 0,1337 dans le cas d'une distribution *a priori* GI informative pour $\sigma_v^2$. De plus, le CV correspondant est de 0,1192 dans le cas d'une distribution *a priori* vague pour $\sigma_v^2$ et de 0,1221 dans le cas d'une distribution *a priori* GI informative pour $\sigma_v^2$. Par conséquent, l'utilisation d'une distribution *a priori* GI informative pour $\sigma_v^2$ produit une ERA légèrement moins élevée, mais produit également un CV plus élevé. Lorsqu'une distribution *a priori* informative normale pour $\boldsymbol{\beta}$ est utilisée, l'ERA est de 0,1287 et le CV est égal à 0,1164 dans le cas d'une distribution *a priori* vague pour $\sigma_v^2$, et l'ERA est de 0,1286 et le CV est égal à 0,1195 dans le cas d'une distribution *a priori* GI informative pour $\sigma_v^2$. Ainsi, il y a une légère amélioration de l'ERA lorsque la distribution *a priori* informative normale pour $\boldsymbol{\beta}$ est utilisée. Cependant, l'utilisation de la distribution *a priori* GI informative pour $\sigma_v^2$ produit un CV plus élevé.

**Tableau 3.1**
**Comparaison de l'ERA et du CV pour le mois de mai 2011 d'après les données de l'EPA au moyen du modèle MYC**

| | EPA directes | HB pour le cas de distribution *a priori* 1 | HB pour le cas de distribution *a priori* 2 | HB pour le cas de distribution *a priori* 3 | HB pour le cas de distribution *a priori* 4 |
|---|---|---|---|---|---|
| ERA moyenne | 0,1739 | 0,1348 | 0,1337 | 0,1287 | 0,1286 |
| CV moyen | 0,2117 | 0,1192 | 0,1221 | 0,1164 | 0,1195 |

**Note :** CV = coefficient de variation; EPA = Enquête sur la population active du Canada; ERA = erreur relative absolue; HB = hiérarchique bayésien; MYC = modèle de You-Chapman.

Dans le tableau 3.2, le MLLY permet d'obtenir de meilleurs résultats que le MYC en termes d'ERA et de CV moins élevés dans tous les cas de distribution *a priori*. Là encore, l'utilisation de la distribution *a priori* informative normale pour $\boldsymbol{\beta}$ produit une ERA moins élevée pour les estimations HB sur petits domaines. L'utilisation de la distribution *a priori* GI informative produit un CV légèrement plus élevé, à l'instar des résultats présentés dans le tableau 3.1. Dans les tableaux 3.1 et 3.2, le meilleur résultat est obtenu dans le cas du MLLY qui repose sur une distribution *a priori* informative normale pour $\boldsymbol{\beta}$ et sur une distribution *a priori* vague pour $\sigma_v^2$, avec une ERA égale à 0,1278 et une valeur de CV de 0,1161.

**Tableau 3.2**
**Comparaison de l'ERA et du CV pour le mois de mai 2011 d'après les données de l'EPA au moyen du modèle MLLY**

| | EPA directes | HB pour le cas de distribution *a priori* 1 | HB pour le cas de distribution *a priori* 2 | HB pour le cas de distribution *a priori* 3 | HB pour le cas de distribution *a priori* 4 |
|---|---|---|---|---|---|
| ERA moyenne | 0,1739 | 0,1329 | 0,1325 | 0,1278 | 0,1277 |
| CV moyen | 0,2117 | 0,1185 | 0,1217 | 0,1161 | 0,1192 |

**Note :** CV = coefficient de variation; EPA = Enquête sur la population active du Canada; ERA = erreur relative absolue; HB = hiérarchique bayésien; MLLY = modèle log-linéaire de You.

Ensuite, nous utiliserons le MLLY comme modèle à l'étude et nous ajusterons les valeurs pour la distribution *a priori* informative normale afin d'évaluer l'incidence correspondante sur les estimations HB sur petits domaines. Nous utiliserons la distribution *a priori* vague GI(0,0001; 0,0001) pour $\sigma_v^2$, car la distribution *a priori* GI informative GI(17,42; 0,00272) produit toujours des CV plus élevés pour les estimateurs HB comparativement à la distribution *a priori* vague GI(0,0001; 0,0001). Nous utilisons la distribution *a priori* informative suivante pour $\boldsymbol{\beta}: \pi(\boldsymbol{\beta}) \sim N(a\boldsymbol{\beta}_0, b\boldsymbol{\Phi}_0)$, où $a$ et $b$ sont des constantes établies à $a$ = 0,5; 1; 2; 3; 4 et $b$ = 0,25; 1; 4; 9; 16; 25. Lorsque $a = b = 1$, nous avons la distribution *a priori* informative par défaut $\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Phi}_0)$ pour $\boldsymbol{\beta}$. Lorsque $a$ et $b$ prennent d'autres valeurs, nous obtenons d'autres distributions *a priori* incorrectes. Nous évaluerons le rendement des estimations HB sur petits domaines dans le cas des combinaisons de distribution *a priori* par défaut et des différentes distributions *a priori* incorrectes. En particulier, nous élargissons la variance *a priori* pour $\boldsymbol{\beta}$ en attribuant des valeurs plus élevées pour $b$, de sorte que nous puissions évaluer le rendement de l'estimateur HB pour différentes variances *a priori*. Lorsque la variance *a priori* est suffisamment élevée, les estimations HB dans le cas de la distribution *a priori* informative devraient se rapprocher des estimations HB dans le cas de la distribution *a priori* uniforme pour $\boldsymbol{\beta}$. Le tableau 3.3 présente la comparaison de l'ERA pour différentes distributions *a priori* au moyen de différentes valeurs de $a$ et de $b$.

**Tableau 3.3**
**Comparaison de l'ERA moyenne et (CV) dans le cas de différentes spécifications *a priori***

| | $b = 0{,}25$ | $b = 1$ | $b = 4$ | $b = 9$ | $b = 16$ | $b = 25$ (distribution *a priori* uniforme) |
|---|---|---|---|---|---|---|
| $a = 0{,}5$ | 0,1718 (0,2087) | 0,1701 (0,1665) | 0,1364 (0,1186) | 0,1340 (0,1185) | 0,1333 (0,1184) | 0,1328 (0,1183) |
| $a = 1$ | 0,1267 (0,1146) | 0,1278 (0,1161) | 0,1321 (0,1181) | 0,1326 (0,1182) | 0,1327 (0,1182) | 0,1326 (0,1182) |
| $a = 2$ | 0,1827 (0,1882) | 0,1805 (0,1837) | 0,1226 (0,1189) | 0,1312 (0,1183) | 0,1323 (0,1183) | 0,1325 (0,1182) |
| $a = 3$ | 0,1835 (0,2014) | 0,1815 (0,2051) | 0,1279 (0,1224) | 0,1286 (0,1185) | 0,1318 (0,1184) | 0,1322 (0,1183) |
| $a = 4$ | 0,1815 (0,2052) | 0,1814 (0,2055) | 0,1281 (0,1311) | 0,1292 (0,1183) | 0,1317 (0,1184) | 0,1322 (0,1183) |

**Note :** CV = coefficient de variation; ERA = erreur relative absolue.

Au départ, lorsque $a = b = 1$, nous avons la distribution *a priori* informative par défaut $\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Phi}_0)$ pour $\boldsymbol{\beta}$. L'ERA de l'estimateur HB est de 0,1278 et le CV moyen est de 0,1161. Dans le cas de la distribution *a priori* non informative pour $\boldsymbol{\beta}$, l'ERA est de 0,1326 et le CV est de 0,1183, ce qui donne à penser que l'estimateur HB a un rendement légèrement supérieur lorsque la distribution *a priori* informative par défaut est utilisée. Si nous gardons $b = 1$ (variance inchangée) et nous ajustons les valeurs pour $a$ à 0,5 (moyenne réduite) ou à 2, 3 et 4 (moyenne augmentée), l'ERA et le CV correspondant deviennent élevés. Par exemple, pour $a = 2$, l'ERA est égale à 0,1805 et le CV est de 0,1837. Par conséquent, si nous avons des valeurs erronées pour la moyenne *a priori*, l'estimateur HB aura une ERA élevée et un CV élevé, et ces distributions *a priori* erronées entraînent des résultats HB pires que la distribution *a priori* non informative. Si nous augmentons les valeurs pour $b$ de 1 à 9, 16 et 25, les valeurs de l'ERA et du CV se rapprochent toutes des résultats obtenus dans le cas de la distribution *a priori* non informative, comme prévu. Ainsi, si nous établissons une valeur de la variance très élevée pour la variance *a priori*, l'incidence de la valeur différente de la moyenne *a priori* est alors négligeable. Comme prévu, la distribution *a priori* avec une variance très grande est à peu près équivalente à la distribution *a priori* uniforme non informative pour $\boldsymbol{\beta}$. Les résultats présentés dans le tableau 3.3 indiquent que la spécification adéquate de la moyenne *a priori* est très importante. Une spécification erronée de la moyenne *a priori* pourrait entraîner un biais et un CV importants. Si nous ne pouvons pas avoir une valeur adéquate de la moyenne pour la distribution *a priori*, une distribution *a priori* uniforme pour $\boldsymbol{\beta}$ doit être utilisée.

# 4. Sélection du modèle bayésien et vérification du modèle

Dans la présente section, nous effectuons une sélection du modèle bayésien et une vérification du modèle pour l'application à l'EPA présentée à la section 3 afin d'évaluer le MLLY avec différentes distributions *a priori*. Nous utilisons d'abord ce qu'on appelle l'ordonnée prédictive conditionnelle (OPC) pour évaluer et sélectionner le modèle le mieux ajusté selon différentes spécifications *a priori* présentées dans le tableau 3.3. Les OPC sont les valeurs de vraisemblance observées en fonction de la densité prédictive sur validation

croisée $f(y_i \mid y_{\text{obs}(i)})$ pour chaque point de données $y_i$, où $y_{\text{obs}(i)}$ désigne toutes les données observées à l'exception de $y_i$. Nous calculons les valeurs de l'OPC pour chaque point de données observé $y_{i,\text{obs}}$, c'est-à-dire $\text{OPC}_i = f(y_{i,\text{obs}} \mid y_{\text{obs}(i)})$, et une augmentation de la valeur de l'$\text{OPC}_i$ indique un meilleur ajustement du modèle. Pour obtenir une formule plus détaillée du calcul de l'OPC, voir par exemple Chen, Shao et Ibrahim (2000), You et Rao (2000), et Rao et Molina (2015, page 346). Pour découvrir d'autres applications de l'OPC, voir par exemple Gilks, Richardson et Spiegelhalter (1996), Molina, Nandram et Rao (2014), et You (2021). Pour le choix du modèle bayésien entre le modèle A et le modèle B, nous pouvons calculer l'OPC pour tous les points de données dans le cas du modèle A et du modèle B et obtenir le ratio de l'OPC pour chaque point de données. Si ce ratio est supérieur à 1, alors $y_{i,\text{obs}}$ appuie le modèle A. Nous avons utilisé le modèle MLLY avec la distribution *a priori* informative exacte $(a = b = 1)$ pour $\boldsymbol{\beta}$ comme modèle par défaut (modèle A) et avons calculé les ratios de l'OPC pour le modèle avec d'autres spécifications *a priori* (valeurs $a$ et $b$ différentes présentées dans le tableau 3.3). Nous comptons ensuite les ratios de l'OPC qui sont supérieurs à 1. Nous avons 65 RMR/AR dans notre analyse de données. Le nombre de ratios qui contient plus de 33 points indique que le modèle HB MLLY ayant la distribution *a priori* exacte par défaut offre un meilleur ajustement du modèle.

Le tableau 4.1 présente le nombre de ratios de l'OPC supérieurs à 1 pour les 65 RMR/AR pour le MLLY selon différentes spécifications *a priori*. Les nombres plus élevés des ratios de l'OPC (33 et plus) indiqueront que la spécification *a priori* exacte offre un meilleur ajustement du modèle que les mauvaises spécifications *a priori*. Les résultats présentés dans le tableau 4.1 démontrent de toute évidence que le MLLY ayant la spécification *a priori* informative par défaut $(a = b = 1)$ pour $\boldsymbol{\beta}$ offre un bien meilleur ajustement du modèle que le modèle comprenant les spécifications *a priori* incorrectes, et particulièrement meilleur que le modèle ayant les spécifications de moyenne *a priori* erronées. Par exemple, dans la première colonne ($b$ = 0,25), pour le MLLY ayant les spécifications de moyenne *a priori* erronées ($a$ = 0,5; 2; 3 et 4), plus de 50 des 65 points de données appuient le MLLY selon une spécification de la moyenne *a priori* exacte. Dans le cas du modèle présentant une spécification de la moyenne *a priori* exacte ($a$ = 1), le modèle présentant une variance *a priori* plus faible ($b$ = 0,25) offre un meilleur ajustement du modèle, comme prévu, le nombre de ratios de l'OPC correspondant étant de 28, soit le seul nombre inférieur à 33 présenté dans le tableau 4.1. Nous pouvons également observer dans le tableau 4.1 que, pour le modèle ayant des spécifications de moyenne *a priori* erronées, lorsque la variance *a priori* est faible, davantage de points de données appuient le modèle selon une spécification *a priori* exacte, car le nombre de ratios supérieurs à 1 présentés dans les colonnes 1 à 3 varie généralement de 40 à 50. Lorsque la variance *a priori* est plus importante, par exemple, lorsque $b$ = 25, le nombre de ratios de l'OPC supérieurs à 1 se situe autour de 37 ou 38, comme présenté dans la dernière colonne du tableau 4.1, ce qui sous-entend que l'incidence d'une moyenne *a priori* erronée devient beaucoup moins importante lorsque la variance *a priori* est très importante. Le résultat de la sélection du modèle dans le tableau 4.1 démontre que les données appuient les modèles présentant une spécification de la moyenne *a priori* exacte, et le résultat de la sélection du modèle est tout à fait conforme aux résultats présentés dans le tableau 3.3.

**Tableau 4.1**
**Résumé du nombre de ratios de l'OPC supérieurs à 1**

| | $b = 0{,}25$ | $b = 1$ | $b = 4$ | $b = 9$ | $b = 16$ | $b = 25$ (distribution *a priori* uniforme) |
|---|---|---|---|---|---|---|
| $a = 0{,}5$ | 51 | 48 | 38 | 37 | 41 | 38 |
| $a = 1$ | 28 | - | 35 | 37 | 36 | 38 |
| $a = 2$ | 52 | 51 | 42 | 36 | 35 | 37 |
| $a = 3$ | 52 | 52 | 50 | 42 | 38 | 37 |
| $a = 4$ | 51 | 50 | 53 | 51 | 41 | 37 |

**Note :** OPC = ordonnée prédictive conditionnelle.

Ensuite, nous comparons les différents modèles de spécifications *a priori* au moyen du critère d'information de déviance (CID) bayésien proposé par Spiegelhalter, Best, Carlin et van der Linde (2002). Le CID repose sur la déviance du modèle $D(\theta)$, qui est égale à moins deux fois la log-vraisemblance du modèle. Le CID est habituellement calculé sous la forme $\text{CID} = D(\hat{\theta}) + 2\ P_D$, où $D(\hat{\theta})$ est la déviance du modèle, évaluée à la moyenne *a posteriori* des paramètres du modèle, qui résume la qualité de l'ajustement du modèle, et $P_D$ est le nombre effectif de paramètres, qui traduit la complexité du modèle. $P_D$ est défini comme étant $P_D = \bar{D}(\theta) - D(\hat{\theta})$, et $\bar{D}(\theta)$ est la moyenne *a posteriori* de la déviance du modèle. Ainsi, le CID est défini comme la somme de l'adéquation du modèle et de la complexité du modèle. De petites valeurs du CID indiquent un meilleur ajustement du modèle. Le calcul du CID est relativement simple à condition que la déviance $D(\theta)$ soit disponible sous une forme fermée, et $P_D$ peut être calculé après l'exécution de l'échantillonnage de Gibbs en prenant la moyenne d'échantillon des valeurs simulées de $D(\theta)$ dont on soustrait l'estimation de la déviance $D(\hat{\theta})$ obtenue par insertion de données. Le tableau 4.2 présente les valeurs du CID pour le modèle HB MLLY selon différentes combinaisons de spécifications *a priori*.

**Tableau 4.2**
**Comparaison des valeurs du CID pour différentes spécifications *a priori***

| | $b = 0{,}25$ | $b = 1$ | $b = 4$ | $b = 9$ | $b = 16$ | $b = 25$ (distribution *a priori* uniforme) |
|---|---|---|---|---|---|---|
| $a = 0{,}5$ | 280,13 | 271,15 | 257,72 | 256,49 | 256,22 | 255,87 |
| $a = 1$ | 254,26 | 254,52 | 255,37 | 255,45 | 255,69 | 255,83 |
| $a = 2$ | 282,72 | 281,49 | 262,56 | 256,73 | 255,63 | 255,79 |
| $a = 3$ | 285,67 | 285,91 | 285,33 | 264,01 | 257,39 | 256,21 |
| $a = 4$ | 286,22 | 286,11 | 286,06 | 286,31 | 261,48 | 257,19 |

**Note :** CID = critère d'information de déviance.

Il est évident que le modèle HB avec la moyenne *a priori* informative par défaut présente les valeurs du CID les plus faibles, ce qui indique clairement que le modèle MLLY présentant la spécification *a priori* exacte offre le meilleur ajustement du modèle. Les modèles dont les spécifications *a priori* sont exactes $a = 1$ et $b = 0{,}25$ ou $b = 1$ ont un meilleur rendement comme démontré avec les valeurs du CID les plus faibles. Les modèles présentant des spécifications erronées de moyenne *a priori* ($a = 0{,}5$; 2; 3; 4) ont des valeurs du CID importantes, ce qui indique un ajustement du modèle relativement faible, surtout lorsque la

variance *a priori* n'est pas grande ($b = 0{,}25; 1; 4$). Lorsque la variance *a priori* est importante, par exemple, lorsque $b = 16$ ou 25, les valeurs du CID deviennent plus petites et se rapprochent de la valeur du CID du modèle dont la spécification *a priori* est exacte, comme prévu. Les résultats présentés dans le tableau 4.2 démontrent que la spécification exacte de la moyenne *a priori* offre le meilleur ajustement du modèle et que, lorsque la variance *a priori* est très grande, l'ajustement du modèle se rapproche du modèle présentant la distribution *a priori* exacte ou uniforme. La comparaison des valeurs du CID est conforme aux résultats de la comparaison des valeurs de l'OPC présentés dans le tableau 4.1 et aux résultats HB présentés dans le tableau 3.3.

Enfin, nous évaluons l'ajustement global du modèle MLLY présentant la spécification *a priori* exacte au moyen de la valeur p prédictive *a posteriori* selon la distribution prédictive *a posteriori*. La valeur p prédictive *a posteriori* est généralement utile si nous considérons le modèle actuel comme un point final plausible auquel des modifications doivent être apportées que si un manque considérable d'ajustement est constaté, comme souligné dans Sinharay et Stern (2003). Supposons que $y_{\text{rép}}$ désigne l'observation répétée pour le modèle. La distribution prédictive *a posteriori* de $y_{\text{rép}}$ compte tenu des données observées $y_{\text{obs}}$ est définie comme étant

$$f(y_{\text{rép}} \mid y_{\text{obs}}) = \int f(y_{\text{rép}} \mid \theta) f(\theta \mid y_{\text{obs}}) d\theta.$$

Selon cette approche, nous pouvons définir une statistique de test $T\left(y, \theta\right)$ qui dépend des données $y$ et possiblement du paramètre $\theta$ et comparer la valeur observée $T(y_{\text{obs}}, \theta \mid y_{\text{obs}})$ à la distribution prédictive *a posteriori* de $T(y_{\text{rép}}, \theta \mid y_{\text{obs}})$, tout écart significatif indiquant une défaillance du modèle. Un manque d'ajustement aux données par rapport à la distribution prédictive *a posteriori* peut être mesuré au moyen d'une valeur p de la statistique de test (Meng, 1994; Gelman, Meng et Stern, 1996). La valeur p prédictive *a posteriori* est définie comme étant

$$p = P(T\left(y_{\text{rép}}, \theta\right) \geq T\left(y_{\text{obs}}, \theta\right) \mid y_{\text{obs}}).$$

Il s'agit d'une extension naturelle de la valeur p habituelle dans un contexte bayésien. Si un modèle est ajusté aux données observées, les deux valeurs de la mesure de divergence sont semblables. Autrement dit, si le modèle est correctement ajusté aux données observées, $T(y_{\text{obs}}, \theta \mid y_{\text{obs}})$ devrait alors se situer près de la partie centrale de l'histogramme des valeurs $T(y_{\text{rép}}, \theta \mid y_{\text{obs}})$ si $y_{\text{rép}}$ est généré de façon répétée à partir de la distribution prédictive *a posteriori*. Par conséquent, la valeur p prédictive *a posteriori* devrait s'approcher de 0,5 si le modèle est bien ajusté aux données. Les valeurs p extrêmes (proches de 0 ou de 1) indiquent un mauvais ajustement. Nous pouvons estimer la valeur p prédictive *a posteriori* de la façon suivante. Supposons que $\theta^*$ représente un tirage à partir de la distribution *a posteriori* $f(\theta \mid y_{\text{obs}})$, et supposons que $y^*_{\text{rép}}$ représente un tirage à partir de $f(y_{\text{rép}} \mid \theta^*)$. Alors, marginalement, $y^*_{\text{rép}}$ est un échantillon provenant de la distribution prédictive *a posteriori* $f(y_{\text{rép}} \mid y_{\text{obs}})$. Il est relativement facile de calculer la valeur p en utilisant les valeurs simulées de $\theta^*$ provenant de l'échantillonneur de Gibbs. Pour chaque valeur simulée $\theta^*$, nous pouvons simuler $y^*_{\text{rép}}$ à l'aide du modèle et calculer $T\left(y^*_{\text{rép}}, \theta^*\right)$ et $T\left(y_{\text{obs}}, \theta^*\right)$. La valeur p est alors estimée par la proportion de fois que $T\left(y^*_{\text{rép}}, \theta^*\right)$ dépasse $T\left(y_{\text{obs}}, \theta^*\right)$.

Pour faire la vérification du modèle prédictif *a posteriori*, nous devons spécifier une statistique de test $T(y, \theta)$. Dans notre application à l'EPA, nous considérons deux statistiques de test. La première est $T_1 = \sum_{i=1}^{m}(y_i - \theta_i)^2 \Big/ \text{var}(y_i \mid \theta)$, une statistique de test générale de la qualité de l'ajustement qui ressemble à la mesure classique $\chi^2$ de la qualité de l'ajustement. La statistique de test $T_1$ est proposée dans Gelman, Carlin, Stern et Rubin (2004) et est largement utilisée dans la pratique comme mesure omnibus de l'ajustement du modèle. Suivant You (2008b) et You et Zhou (2011), nous considérons aussi une autre statistique de test, $T_2 = \left| \max(y_i) - \text{moyenne}(\theta_i) \right| - \left| \min(y_i) - \text{moyenne}(\theta_i) \right|$, qui est particulièrement utile pour vérifier les modèles HB au niveau du domaine. Pour le MLLY présentant la spécification *a priori* exacte (lorsque $a = b = 1$) comme présenté dans le tableau 3.3, nous obtenons une valeur p moyenne estimée de 0,428 en utilisant la statistique de test $T_1$ et une valeur p de 0,477 en utilisant la statistique de test $T_2$. Les deux valeurs permettent d'appuyer l'ajustement du modèle. Ainsi, rien n'indique un manque d'ajustement global du modèle dans le cas du modèle MLLY.

# 5. Étude par simulation

Dans la présente section, nous menons une étude par simulation pour évaluer le rendement de la spécification *a priori* pour le paramètre de régression $\boldsymbol{\beta}$. Selon You et Hidiroglou (2023), nous générons les données au moyen du modèle de Fay-Herriot avec les variances d'échantillonnage estimées selon les données de l'EPA étudiées à la section 3. Nous considérons m = 65 régions dans l'étude par simulation. Supposons que $\theta_i$ soit le véritable paramètre d'intérêt simulé. Les valeurs de $\theta_i$ sont générées comme étant $\theta_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i = \beta_0 + x_i\beta_1 + v_i$, où $x_i$ est le taux de prestataires au mois de mai 2011 selon l'EPA, $\beta_0$ = 0,0507 et $\beta_1$ = 0,42550, et $v_i$ est généré à partir de $N\left(0, \sigma_v^2\right)$, et $\sigma_v^2$ = 0,0001662. Nous obtenons les valeurs de $\beta_0, \beta_1$ et de $\sigma_v^2$ à partir de la spécification *a priori* de la section 3. L'estimation directe $\hat{\theta}_i$ est générée comme étant $\hat{\theta}_i = \theta_i + e_i$, où $e_i = N(0, \sigma_i^2)$. La variance d'échantillonnage $\sigma_i^2$ est la variance d'échantillonnage lissée obtenue à partir des estimations directes de la variance d'échantillonnage au moyen de la variance d'échantillonnage moyenne lissée de You et Hidiroglou (2023). La variance d'échantillonnage lissée $\sigma_i^2$ est considérée comme la véritable variance d'échantillonnage dans la simulation. L'estimation directe de la variance d'échantillonnage $s_i^2$ est ensuite générée comme étant $s_i^2 = (d_i)^{-1}\sigma_i^2\chi_{d_i}^2$, où $d = n_i - 1$, comme dans Rivest et Vandal (2002), Wang et Fuller (2003), et You (2021). Nous effectuons 5 000 exécutions de la simulation. Pour chaque exécution, nous utilisons la procédure d'échantillonnage de Gibbs pour estimer les paramètres du modèle, et la procédure d'échantillonnage de Gibbs est exécutée avec une période de rodage de 1 000 et 2 000 itérations de plus pour chaque exécution de la simulation.

Nous appliquons le modèle 2 MLLY aux données simulées au moyen de distributions *a priori* non informatives et informatives pour $\boldsymbol{\beta}$. La distribution *a priori* pour $\sigma_v^2$ est la distribution *a priori* vague GI(0,0001; 0,0001). Pour comparer l'estimateur HB dans le cas de différentes distributions *a priori* pour l'ensemble des régions, nous considérons le biais relatif absolu (BRA) moyen pour l'estimateur HB $\hat{\theta}_i$ de la moyenne de petits domaines simulée $\theta_i$ définie comme étant $\overline{\text{BRA}} = \left(\sum_{i=1}^{m} \text{BRA}_i\right) \Big/ m$, où

$$\text{BRA}_i = \left| \frac{1}{R} \sum_{r=1}^{R} \frac{(\hat{\theta}_i^{(r)} - \theta_i^{(r)})}{\theta_i^{(r)}} \right|,$$

et $\hat{\theta}_i^{(r)}$ est l'estimation HB et $\theta_i^{(r)}$ est la moyenne générée selon le $r^{\text{e}}$ échantillon simulé, $R = 5\,000$, $m = 65$.

Nous calculons aussi la racine de l'erreur quadratique moyenne relative (REQMR) pour l'estimateur HB dans le cas de différentes distributions *a priori*. Nous prenons ensuite la moyenne de la REQMR pour les régions. La REQMR moyenne de la simulation est calculée comme étant $\overline{\text{REQMR}} = \left(\sum_{i=1}^{m} \text{REQMR}_i\right) \Big/ m$, où $\text{REQMR}_i = \left(\sum_{r=1}^{R} \text{REQMR}_i^{(r)}\right) \Big/ R$, et

$$\text{REQMR}_i^{(r)} = \frac{\sqrt{(\hat{\theta}_i^{(r)} - \theta_i^{(r)})^2}}{\theta_i^{(r)}}.$$

Pour l'estimateur HB, nous pouvons également comparer le CV estimé moyen comme étant $\overline{\text{CV}} = \left(\sum_{i=1}^{m} \text{CV}_i\right) \Big/ m$, où $\text{CV}_i = \left(\sum_{r=1}^{R} \text{CV}_i^{(r)}\right) \Big/ R$ et

$$\text{CV}_i^{(r)} = \sqrt{\text{Var}(\hat{\theta}_i^{(r)})} \Big/ \hat{\theta}_i^{(r)},$$

où $\text{var}(\hat{\theta}_i^{(r)})$ est la variance *a posteriori* estimée de l'estimateur HB $\hat{\theta}_i^{(r)}$ pour la $i^{\text{e}}$ région.

Dans l'étude par simulation, les distributions *a priori* pour $\boldsymbol{\beta}$ sont définies comme étant une distribution *a priori* uniforme non informative $\pi(\boldsymbol{\beta}) \propto 1$ et une distribution *a priori* informative $\pi(\boldsymbol{\beta}) \sim N(a\boldsymbol{\beta}_0, b\boldsymbol{\Phi}_0)$, où $a =$ 0,5, 1, 2, 3, 4 et $b =$ 1 et 16. Lorsque $a = 1$, la moyenne *a priori* est correctement précisée. Lorsque $a$ est 0,5, 2, 3 ou 4, la moyenne *a priori* est incorrectement précisée. Lorsque $b =$ 16, la variance *a priori* est très importante et la distribution *a priori* doit être à peu près équivalente à une distribution *a priori* uniforme non informative pour $\boldsymbol{\beta}$. Le tableau 5.1 présente une comparaison du BRA, de la REQMR et du CV des estimateurs HB selon différentes spécifications de moyenne *a priori* (valeur $a$ différente) avec la même spécification de variance *a priori* $b = 1$ (variance *a priori* régulière). Le BRA, la REQMR et le CV de l'estimateur direct sont également compris dans le tableau à des fins d'analyse comparative. Nous avons également inclus le taux de couverture moyen $(\overline{\text{TC}})$ dans les 65 petits domaines pour les intervalles de confiance bayésiens à 95 % des estimateurs bayésiens à des fins de comparaison.

**Tableau 5.1**
**Comparaison du BRA, de la REQMR, du CV et du TC moyens dans l'étude par simulation selon différentes moyennes *a priori* et la même variance *a priori* $b\boldsymbol{\Phi}_0$ ($b = 1$)**

| | **Estimation directe** | **Distribution *a priori* uniforme** | **Moyenne *a priori* exacte $a = 1$** | **Moyenne *a priori* incorrecte $a =$ 0,5** | **Moyenne *a priori* incorrecte $a = 2$** | **Moyenne *a priori* incorrecte $a = 3$** | **Moyenne *a priori* incorrecte $a = 4$** |
|---|---|---|---|---|---|---|---|
| $\overline{\text{BRA}}$ moyen | 0,0067 | 0,0189 | 0,0190 | 0,0781 | 0,0653 | 0,0318 | 0,0213 |
| $\overline{\text{REQMR}}$ moyenne | 0,1822 | 0,1135 | 0,1115 | 0,1426 | 0,1778 | 0,1819 | 0,1826 |
| $\overline{\text{CV}}$ moyen | 0,3035 | 0,1365 | 0,1338 | 0,1961 | 0,2172 | 0,2484 | 0,2613 |
| $\overline{\text{TC}}$ moyen | - | 0,9445 | 0,9468 | 0,9307 | 0,9462 | 0,9471 | 0,9478 |

**Note :** BRA = biais relatif absolu; CV = coefficient de variation; REQMR = racine de l'erreur quadratique moyenne relative; TC = taux de couverture.

Lorsque $\boldsymbol{\beta}$ a une distribution *a priori* uniforme, le BRA moyen est de 0,0189, la REQMR moyenne est de 0,1135 et le CV moyen est de 0,1365 pour l'estimateur HB, comme présenté dans le tableau 5.1. Lorsque $\boldsymbol{\beta}$ a la distribution *a priori* par défaut $\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Phi}_0)$ avec $a = 1$ et $b = 1$, le BRA moyen est de 0,0190, la REQMR moyenne est de 0,1115 et le CV moyen est de 0,1338. Ainsi, dans le cas de la spécification *a priori* exacte, l'estimateur HB a un rendement légèrement supérieur avec une REQMR et un CV légèrement moins élevés. Toutefois, le BRA moyen est à peu près le même dans le cas de la distribution *a priori* uniforme et de la distribution *a priori* informative exacte, comme le démontrent les colonnes 3 et 4 dans le tableau 5.1. L'estimateur direct présente une REQMR et un CV importants comme le démontre la colonne 2 dans le tableau 5.1. Les taux de couverture moyens pour l'intervalle de confiance bayésien à 95 % des estimateurs HB selon différentes distributions *a priori* sont à peu près les mêmes et tournent autour de 95 %, à l'exception de la moyenne *a priori* $a =$ 0,5, où le taux de couverture moyen est légèrement inférieur et s'établit à 93 %.

Si la distribution *a priori* a une spécification de la moyenne *a priori* incorrecte $\pi(\boldsymbol{\beta}) \sim N(a\boldsymbol{\beta}_0, \boldsymbol{\Phi}_0)$ avec $a =$ 0,5 ou $a = 2$, 3 ou 4, l'estimateur HB aura alors un BRA plus élevé ou une REQMR et un CV plus élevés comme démontré dans le tableau 5.1. Par exemple, lorsque $a =$ 0,5, la moyenne *a priori* est de 0,5 $\beta_0$, le BRA moyen est alors de 0,0781, la REQMR moyenne est de 0,1426 et le CV moyen est de 0,1961. Ainsi, les valeurs du BRA, de la REQMR et du CV sont de toute évidence plus élevées que les valeurs correspondantes dans le cas de la distribution *a priori* uniforme ou de la distribution *a priori* exacte. Il est intéressant de souligner que, lorsque la moyenne *a priori* de $\boldsymbol{\beta}$ devient plus importante (par exemple $a = 3$ ou 4), l'estimation de la variance du modèle $\sigma_v^2$ devient alors considérablement plus grande. Par exemple, l'estimation *a posteriori* de $\sigma_v^2$ est de 0,00018 lorsque $a =$ 1, de 0,0215 lorsque $a =$ 3 et de 0,0502 lorsque $a = 4$. L'estimateur HB $\hat{\theta}_i$ accorde une bien plus grande pondération à l'estimation directe $y_i$ et une moins grande pondération à l'estimation de la régression $\mathbf{x}_i'\boldsymbol{\beta}$ lorsque la variance du modèle est plus grande. Ainsi, le rendement de l'estimateur HB ressemble davantage à celui de l'estimateur direct dans cette situation, comme le démontrent les deux dernières colonnes dans le tableau 5.1, présentant un BRA plus faible, mais une REQMR et un CV plus élevés.

Les résultats présentés dans le tableau 5.1 indiquent que l'estimateur HB qui repose sur une distribution *a priori* uniforme ou une distribution *a priori* exacte permet d'améliorer l'estimateur direct et d'obtenir de meilleurs résultats en réduisant la REQMR et le CV moyens. Le rendement de la distribution *a priori* uniforme et de la distribution *a priori* exacte est très semblable dans notre étude par simulation. Si la moyenne *a priori* n'est pas correctement précisée, l'estimateur HB peut présenter un biais et un CV plus importants ou le rendement de l'estimateur HB peut être identique à celui de l'estimateur d'enquête direct comme le démontre le tableau 5.1. Par conséquent, si nous ne pouvons pas déterminer la distribution *a priori* exacte, la meilleure solution consiste à utiliser la distribution *a priori* uniforme pour les paramètres de régression dans les modèles HB sur petits domaines de You et Chapman (2006) et de You (2021).

À titre comparatif, nous calculons aussi le BRA, la REQMR, le CV et le TC moyens des estimations HB lorsque la variance *a priori* est très importante en établissant la valeur pour $b$ aussi élevée que $b =$ 16. Le

tableau 5.2 présente une comparaison du BRA, de la REQMR et du CV pour l'étude par simulation selon différentes spécifications de moyenne *a priori* (valeur $a$ différente) avec la même spécification de variance *a priori* $b = 16$ (variance *a priori* très importante).

**Tableau 5.2**
**Comparaison du BRA, de la REQMR, du CV et du TC dans l'étude par simulation selon différentes moyennes *a priori* et la même variance *a priori* $b\Phi_0$ ($b = 16$)**

| | Estimation directe | Distribution *a priori* uniforme | Moyenne *a priori* exacte $a = 1$ | Moyenne *a priori* incorrecte $a = 0{,}5$ | Moyenne *a priori* incorrecte $a = 2$ | Moyenne *a priori* incorrecte $a = 3$ | Moyenne *a priori* incorrecte $a = 4$ |
|---|---|---|---|---|---|---|---|
| $\overline{\text{BRA}}$ moyen | 0,0067 | 0,0189 | 0,0188 | 0,0185 | 0,0195 | 0,0205 | 0,0217 |
| $\overline{\text{REQMR}}$ moyenne | 0,1822 | 0,1135 | 0,1135 | 0,1134 | 0,1136 | 0,1137 | 0,1137 |
| $\overline{\text{CV}}$ moyen | 0,3035 | 0,1365 | 0,1366 | 0,1364 | 0,1363 | 0,1361 | 0,1359 |
| $\overline{\text{TC}}$ moyen | - | 0,9445 | 0,9448 | 0,9445 | 0,9438 | 0,9441 | 0,9443 |

**Note :** BRA = biais relatif absolu; CV = coefficient de variation; REQMR = racine de l'erreur quadratique moyenne relative; TC = taux de couverture.

Comme prévu, lorsque la variance *a priori* est très grande, le BRA, la REQMR, le CV et le TC moyens des estimateurs HB sont tous respectivement semblables dans le cas de différentes spécifications de moyenne et de la distribution *a priori* uniforme. Les résultats de la simulation présentés dans le tableau 5.2 démontrent que le fait d'utiliser une valeur très grande pour la variance *a priori* équivaut à la distribution *a priori* uniforme. Comme le démontre le tableau 5.2, toutes les estimations HB ont amélioré les estimations directes en réduisant considérablement la REQMR et le CV. Les taux de couverture moyens pour l'intervalle bayésien sont presque tous les mêmes et présentent une valeur se situant autour de 94,45 %. Les résultats de la simulation confirment également les résultats présentés dans le tableau 3.3, selon lesquels une distribution *a priori* avec une spécification de la variance très grande ($b = 16$ ou 25) dans l'application à l'EPA est à peu près équivalente à la distribution *a priori* uniforme non informative pour $\boldsymbol{\beta}$.

# 6. Estimation pour de petits domaines groupés

Aux sections 3 et 4, nous avons étudié le rendement des estimateurs HB sur petits domaines au moyen de distributions *a priori* informatives par l'entremise de l'application à l'EPA et dans une étude par simulation. Dans la présente section, nous considérons différents groupes de régions pour l'application à l'EPA et l'étude par simulation. Pour les 65 RMR/AR à l'échelle du Canada, nous divisons maintenant ces régions en cinq groupes comme suit : groupe de régions 1 : régions des provinces de l'Est, 1 à 14, comprenant 14 RMR/AR; groupe de régions 2 : régions du Québec, 15 à 23, comprenant 9 RMR/AR; groupe de régions 3 : régions de l'Ontario, 24 à 46, comprenant 23 RMR/AR; groupe de régions 4 : régions des provinces du centre, 47 à 58, comprenant 12 RMR/AR; groupe de régions 5 : régions de la Colombie-Britannique, 59 à 65, comprenant 7 RMR/AR.

Le but d'avoir recours à des régions groupées est d'étudier le rendement de différentes distributions *a priori* lorsque le nombre de petits domaines est relativement petit. Nous appliquons maintenant le modèle HB MLLY à ces cinq groupes séparément au moyen d'une distribution *a priori* informative et d'une distribution *a priori* uniforme pour $\boldsymbol{\beta}$, c'est-à-dire la distribution *a priori* informative $\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Phi}_0)$ et la distribution *a priori* uniforme $\pi(\boldsymbol{\beta}) \propto 1$. Nous obtenons les estimations HB et nous comparons ces estimations aux valeurs du recensement dans chaque groupe de régions. Le tableau 6.1 présente l'ERA et le CV moyens de chaque groupe de régions pour l'estimateur HB en fonction des distributions *a priori* informative et uniforme.

**Tableau 6.1**
**Comparaison de l'ERA et du CV pour les différents groupes de régions dans le cas de la modélisation de chaque groupe pris séparément**

| | Distribution *a priori* informative $\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Phi}_0)$ | | Distribution *a priori* uniforme $\pi(\boldsymbol{\beta}) \propto 1$ | |
|---|---|---|---|---|
| | **ERA** | **CV** | **ERA** | **CV** |
| Groupe de régions 1 (14 régions) | 0,1898 | 0,1142 | 0,2526 | 0,1274 |
| Groupe de régions 2 (9 régions) | 0,1133 | 0,0572 | 0,1279 | 0,0739 |
| Groupe de régions 3 (23 régions) | 0,1126 | 0,1107 | 0,0893 | 0,0989 |
| Groupe de régions 4 (12 régions) | 0,0915 | 0,0443 | 0,0937 | 0,0772 |
| Groupe de régions 5 (7 régions) | 0,0841 | 0,1219 | 0,0813 | 0,1688 |
| Ensemble des régions (65 régions) | 0,1226 | 0,0933 | 0,1297 | 0,1050 |

**Note :** CV = coefficient de variation; ERA = erreur relative absolue.

Pour le groupe de régions 1, l'ERA dans le cas d'une distribution *a priori* informative est de 0,1898, ce qui est beaucoup moins élevé que l'ERA de 0,2526 dans le cas d'une distribution *a priori* uniforme. Le CV est de 0,1142 dans le cas d'une distribution *a priori* informative, ce qui est également moins élevé que le CV de 0,1274 dans le cas d'une distribution *a priori* uniforme. Ainsi, il est évident que l'utilisation de la distribution *a priori* informative adéquate dans le modèle MLLY permet d'obtenir de bien meilleures estimations HB pour les 14 régions du groupe 1. Pour le groupe 2, qui comprend 9 régions au Québec, l'ERA et le CV sont eux aussi moins élevés dans le cas de la distribution *a priori* informative. Cependant, pour le groupe 3, qui comprend 23 régions en Ontario, l'ERA et le CV sont légèrement plus élevés dans le cas de la distribution *a priori* informative que les valeurs correspondantes dans le cas de la distribution *a priori* uniforme. Pour le groupe 4, l'ERA et le CV sont moins élevés dans le cas de la distribution *a priori* informative. Pour le groupe 5, l'ERA est légèrement plus élevée dans le cas de la distribution *a priori* informative, mais le CV est beaucoup moins élevé. Pour l'ensemble des 65 régions, l'ERA est de 0,1226 dans le cas de la distribution *a priori* informative et de 0,1297 dans le cas de la distribution *a priori* uniforme. Le CV est de 0,0933 dans le cas de la distribution *a priori* informative et de 0,1050 dans le cas de la distribution *a priori* uniforme. En résumé, nous démontrons que l'utilisation de la distribution *a priori* informative pour la modélisation à l'intérieur des groupes est utile et permet d'obtenir de meilleurs résultats en termes d'ERA et de CV moins élevés pour la plupart des groupes de régions à l'exception du groupe 3 de l'Ontario. Lorsque le nombre de régions est petit, l'utilisation de la distribution *a priori* informative exacte dans le

modèle HB peut donner de meilleurs résultats, car les distributions *a priori* auront plus de poids dans l'estimateur HB. Par conséquent, lorsque le nombre de régions de modélisation est petit, la distribution *a priori* informative pourrait être plus utile.

Nous évaluons maintenant les modèles au moyen de distributions *a priori* informatives et uniformes pour les groupes de régions selon le ratio de l'OPC et les valeurs du CID, comme à la section 4. Nous calculons la valeur du ratio de l'OPC du MLLY qui repose sur une distribution *a priori* informative par rapport au MLLY qui repose sur une distribution *a priori* uniforme. Le tableau 6.2 présente le nombre de ratios de l'OPC supérieurs à 1 et les valeurs du CID pour le MLLY qui repose sur des distributions *a priori* informatives et uniformes.

Le tableau 6.2 démontre que, pour les groupes de régions comprenant un nombre moins élevé de régions, le ratio de l'OPC et la valeur du CID appuient tous deux le modèle qui repose sur une distribution *a priori* informative. Par exemple, pour le groupe de régions 1 (14 régions), le ratio de l'OPC démontre que 9 des 14 régions appuient le modèle qui repose sur une distribution *a priori* informative, et la valeur du CID est aussi moins élevée (65,59) lorsque la distribution *a priori* informative est utilisée. Pour le groupe de régions 3 (Ontario), l'OPC et le CID appuient tous deux le modèle qui repose sur une distribution *a priori* uniforme, ce qui est conforme au résultat présenté dans le tableau 6.1. Pour l'Ontario (23 régions), la distribution *a priori* uniforme est suffisante pour obtenir un meilleur ajustement du modèle en termes d'OPC et de CID, et d'ERA et de CV moins élevés comme présenté dans le tableau 6.1. Pour l'ensemble des régions (65 régions), l'OPC et le CID appuient légèrement le modèle ayant une distribution *a priori* informative, 38 des 65 régions appuyant le modèle ayant la distribution *a priori* informative. Les résultats des tableaux 6.1 et 6.2 indiquent que la distribution *a priori* uniforme peut être suffisante pour obtenir les mêmes résultats que pour la distribution *a priori* informative exacte lorsque le nombre de régions est suffisamment grand. Toutefois, lorsque le nombre de régions est petit (groupes de régions 1, 2, 4, 5), l'utilisation de la distribution *a priori* informative exacte permet d'obtenir de meilleurs résultats et un meilleur ajustement du modèle.

**Tableau 6.2**
**Nombre de ratios de l'OPC supérieurs à 1 et valeurs du CID pour les groupes de régions**

| | Nombre de ratios de l'OPC supérieurs à 1 | CID Distribution *a priori* informative $\pi(\beta) \sim N(\beta_0, \Phi_0)$ | CID Distribution *a priori* uniforme $\pi(\beta) \propto 1$ |
|---|---|---|---|
| Groupe de régions 1 (14 régions) | 9 | 65,59 | 66,72 |
| Groupe de régions 2 (9 régions) | 7 | 25,41 | 27,84 |
| Groupe de régions 3 (23 régions) | 10 | 88,12 | 87,83 |
| Groupe de régions 4 (12 régions) | 8 | 30,61 | 32,47 |
| Groupe de régions 5 (7 régions) | 4 | 34,76 | 35,45 |
| Ensemble des régions (65 régions) | 38 | 254,52 | 255,95 |

**Note :** CID = critère d'information de déviance; OPC = ordonnée prédictive conditionnelle.

Pour une comparaison plus poussée et une évaluation de la modélisation à l'intérieur de chaque groupe, nous utilisons maintenant les résultats du tableau 3.2 selon la modélisation de l'ensemble des 65 régions et

nous les comparons aux valeurs du recensement à l'intérieur de chaque groupe de régions, comme pour la comparaison présentée dans le tableau 6.1. Si le nombre de régions de modélisation est grand, les données de l'échantillon peuvent fournir une estimation et des renseignements plus fiables et, par conséquent, nous nous attendons à ce que l'utilisation de la distribution *a priori* uniforme ou d'une distribution *a priori* informative ne fasse pas une grande différence. Le tableau 6.3 présente les résultats de la comparaison pour chaque groupe de régions dans le cas de la modélisation de l'ensemble des régions. Comme prévu, le tableau 6.3 démontre que les valeurs moyennes de l'ERA et du CV dans le cas de la distribution *a priori* informative et de la distribution *a priori* uniforme sont semblables et que la différence est petite pour l'ensemble des cinq groupes de régions.

**Tableau 6.3**
**Comparaison de l'ERA et du CV pour les différents groupes de régions dans le cas de la modélisation de l'ensemble des régions**

| | Distribution *a priori* informative $\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Phi}_0)$ | | Distribution *a priori* uniforme $\pi(\boldsymbol{\beta}) \propto 1$ | |
|---|---|---|---|---|
| | **ERA** | **CV** | **ERA** | **CV** |
| Groupe de régions 1 (14 régions) | 0,1663 | 0,1104 | 0,1853 | 0,1226 |
| Groupe de régions 2 (9 régions) | 0,1166 | 0,1154 | 0,1206 | 0,1145 |
| Groupe de régions 3 (23 régions) | 0,1327 | 0,1153 | 0,1316 | 0,1153 |
| Groupe de régions 4 (12 régions) | 0,0995 | 0,1228 | 0,1041 | 0,1224 |
| Groupe de régions 5 (7 régions) | 0,0983 | 0,1188 | 0,0983 | 0,1198 |
| Ensemble des régions (65 régions) | 0,1278 | 0,1161 | 0,1329 | 0,1185 |

**Note :** CV = coefficient de variation; ERA = erreur relative absolue.

À partir des résultats présentés dans les tableaux 6.1, 6.2 et 6.3, il est évident que, lorsque le nombre de petits domaines est relativement grand, l'utilisation de la distribution *a priori* uniforme pour le paramètre de régression peut tout de même permettre de produire des estimations HB efficaces pour l'estimation sur petits domaines. Cependant, lorsque le nombre de petits domaines est petit, l'utilisation de la distribution *a priori* informative exacte peut permettre de produire de meilleures estimations HB présentant un biais et un CV moins importants.

Enfin, nous avons également réalisé une étude par simulation comme à la section 5 à l'intérieur des cinq régions groupées. Le tableau 6.4 présente les résultats de la simulation comparant le BRA, la REQMR et le CV moyens des estimateurs HB à l'intérieur de chaque région groupée au moyen d'une distribution *a priori* informative $\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Phi}_0)$ et d'une distribution *a priori* uniforme non informative $\pi(\boldsymbol{\beta}) \propto 1$ pour $\boldsymbol{\beta}$. Le tableau 6.4 démontre que le BRA moyen dans le cas de la distribution *a priori* informative est légèrement plus élevé que le BRA moyen dans le cas de la distribution *a priori* uniforme, mais les valeurs du $\overline{\text{BRA}}$ sont toutes deux très petites. La REQMR et le CV moyens dans le cas de la distribution *a priori* informative sont moins élevés que les valeurs correspondantes dans le cas de la distribution *a priori* uniforme pour tous les groupes de régions. Par exemple, pour le groupe de régions 1, le $\overline{\text{BRA}}$ est de 0,0108 dans le cas de la distribution *a priori* informative et de 0,0104 dans le cas de la distribution *a priori* uniforme, tandis que le $\overline{\text{CV}}$ est de 0,1288 dans le cas de la distribution *a priori* informative et de 0,1427 dans le cas de la distribution

*a priori* uniforme. Les résultats de la simulation indiquent que l'utilisation de la distribution *a priori* informative exacte peut permettre d'améliorer l'estimation HB en réduisant la REQMR et le CV dans l'estimation sur petits domaines groupés.

**Tableau 6.4**
**Comparaison du BRA, de la REQMR et du CV pour différents groupes de régions dans l'étude par simulation**

| | Distribution *a priori* informative $\pi(\beta) \sim N(\beta_0, \Phi_0)$ | | | Distribution *a priori* uniforme $\pi(\beta) \propto 1$ | | |
|---|---|---|---|---|---|---|
| | $\overline{\text{BRA}}$ | $\overline{\text{REQMR}}$ | $\overline{\text{CV}}$ | $\overline{\text{BRA}}$ | $\overline{\text{REQMR}}$ | $\overline{\text{CV}}$ |
| Groupe de régions 1 (14 régions) | 0,0108 | 0,0967 | 0,1288 | 0,0104 | 0,1091 | 0,1427 |
| Groupe de régions 2 (9 régions) | 0,0181 | 0,1199 | 0,1587 | 0,0163 | 0,1303 | 0,1751 |
| Groupe de régions 3 (23 régions) | 0,0183 | 0,1172 | 0,1447 | 0,0152 | 0,1224 | 0,1537 |
| Groupe de régions 4 (12 régions) | 0,0185 | 0,1202 | 0,1563 | 0,0146 | 0,1323 | 0,1732 |
| Groupe de régions 5 (7 régions) | 0,0203 | 0,1169 | 0,1594 | 0,0191 | 0,1335 | 0,1839 |

**Note :** BRA = biais relatif absolu; CV = coefficient de variation; REQMR = racine de l'erreur quadratique moyenne relative.

# 7. Conclusions

Dans le présent article, nous avons étudié le rendement des estimateurs hiérarchiques bayésiens sur petits domaines qui reposent sur des distributions *a priori* uniformes informatives et non informatives pour les paramètres de régression dans le modèle de liaison des modèles HB sur petits domaines au niveau du domaine. Notre analyse de données réelles et notre étude par simulation indiquent que la distribution *a priori* informative par défaut et la distribution *a priori* uniforme pour les paramètres de régression peuvent toutes deux avoir un très bon rendement lorsque le nombre de petits domaines est relativement grand. La distribution *a priori* uniforme $\pi(\boldsymbol{\beta}) \propto 1$ peut être suffisante en général pour les modèles HB sur petits domaines lorsque le nombre de petits domaines est grand. Cette conclusion est également soutenue par les résultats de la sélection du modèle bayésien et la vérification du modèle bayésien dans l'analyse des données. La spécification erronée de moyennes *a priori* informatives peut produire un biais et un CV élevés. Lorsque le nombre de petits domaines est relativement petit, l'utilisation d'une distribution *a priori* informative exacte dans la modélisation HB peut être utile pour obtenir de meilleurs résultats en termes de biais et de CV moins élevés. Par exemple, dans l'application de l'estimation du sous-dénombrement du recensement étudiée dans You et Rao (2002) et dans You, Rao et Dick (2004), le nombre de petits domaines (provinces) est très petit, seulement 10 estimations de sous-dénombrement directes sont disponibles à l'échelle provinciale partout au Canada, auquel cas les distributions *a priori* informatives exactes peuvent être utiles dans les modèles HB pour obtenir de meilleures estimations en fonction d'un modèle et un meilleur ajustement du modèle. Évidemment, dans la pratique, il n'est pas facile de savoir de quelle façon construire une distribution *a priori* exacte et adéquate. Nous avons déjà donné des directives générales quant au choix des distributions *a priori* en introduction. Dans le cadre de travaux futurs sur l'application à l'EPA, nous prévoyons examiner et étudier certaines possibilités de spécification *a priori*, par exemple, en empruntant des renseignements aux données mensuelles précédentes au moyen d'une approche itérative pour les données de l'EPA, c'est-à-dire, en utilisant les estimations *a posteriori* du mois précédent comme renseignements antérieurs pour le mois

actuel ou le prochain mois, et évaluer les estimations HB correspondantes par l'analyse de données réelles et l'étude par simulation.

L'utilisation de données historiques pour construire des distributions *a priori* pour les paramètres du modèle peut être utile et peut fournir des renseignements directs sur les paramètres du modèle. Une autre façon d'utiliser des données historiques consiste à emprunter des renseignements au fil du temps à l'aide d'un modèle de séries chronologiques. Par exemple, un terme autorégressif ou de marche aléatoire peut être ajouté au modèle de moyenne de petits domaines (par exemple Rao et Yu, 1994; Datta, Lahiri, Maiti, et Lu, 1999). You, Rao et Gambino (2003) et You (2008a) ont utilisé des modèles transversaux et de séries chronologiques pour les données de l'EPA afin d'améliorer les estimations directes de l'EPA. Il serait intéressant d'examiner et de comparer la spécification *a priori* et le modèle de séries chronologiques pour l'utilisation de données historiques dans les travaux futurs.

L'étalonnage est une autre méthode permettant d'utiliser des données externes ou antérieures afin d'améliorer les estimations fondées sur un modèle au niveau agrégé. Pour des méthodes d'étalonnage bayésiennes, Zhang et Bryant (2020) présentent un bon aperçu de différentes méthodes d'étalonnage bayésiennes dans la documentation et proposent une approche exclusivement bayésienne et flexible à l'égard de l'étalonnage permettant d'utiliser un large éventail de modèles et de valeurs repères. Ils révisent la vraisemblance en la multipliant par une distribution de probabilités qui mesure la concordance avec les valeurs repères et décrivent des méthodes de Monte Carlo par chaîne de Markov pour générer des échantillons à partir de distributions *a posteriori* étalonnées. Plus récemment, Okonek et Wakefield (2024) font une critique des méthodes de réconciliation bayésiennes dans le contexte de l'estimation sur petits domaines et ils améliorent la méthode de Zhang et Bryant (2020) en proposant une approche novatrice dans laquelle des échantillons *a posteriori* de caractéristiques de petits domaines à partir d'un modèle non étalonné peuvent être combinés à un échantillonneur de rejet ou à des algorithmes de Metropolis-Hastings pour produire des distributions *a posteriori* étalonnées de manière efficace du point de vue du calcul. Pour l'application à l'EPA dans notre étude, il serait intéressant de comparer et d'évaluer les estimations HB étalonnées avec les estimations HB en fonction de différentes spécifications *a priori*.

## Remerciements

Nous aimerions remercier trois examinateurs, le rédacteur en chef adjoint et le rédacteur en chef pour leurs commentaires constructifs et leurs suggestions pour nous permettre d'améliorer considérablement notre article et le rendre plus intéressant à lire et plus pertinent pour les utilisateurs, dans la pratique.

## Bibliographie

Bradley, J.R., Holan, S.H. et Wikle, C.K. (2018). Computationally efficient multivariate spatio-temporal models for high-dimensional count-valued data (with Discussion). *Bayesian Analysis*, 13, 253-310.

Bradley, J.R., Holan, S.H. et Wikle, C.K. (2020). Bayesian hierarchical models with conjugate full-conditional distributions for dependent data from the natural exponential family. *Journal of the American Statistical Association*, 115, 2037-2052.

Chen, M.-H., Shao, Q.-M. et Ibrahim, J.G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer.

Chip, S. et Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49, 327-335.

Consonni, G., Fouskakis, D., Liseo, B. et Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, 13, 627-679.

Dass, S.C., Maiti, T., Ren, H. et Sinha, S. (2012). Estimation des intervalles de confiance des paramètres de petit domaine avec rétrécissement des moyennes et des variances. *Techniques d'enquête*, 38(2), 187-203. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2012002/article/11756-fra.pdf.

Datta, G.S., Lahiri, P., Maiti, T. et Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.

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, 268-277.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

Gelman, A. et Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.

Gelman, A., Meng, X.L. et Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, 6, 733-807.

Gelman, A., Carlin, J.B., Stern, H.S. et Rubin, D.B. (2004). *Bayesian Data Analysis*. 2nd Edition, Chapman & Hall/CRC.

Ghosh, M., Myung, J. et Moura, F.A.S. (2018). Estimation bayésienne robuste sur petits domaines. *Techniques d'enquête*, 44(1), 109-124. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2018001/article/54959-fra.pdf.

Gilks, W.R., Richardson, S. et Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC.

Grzenda, W. (2016). Informative versus non-informative prior distributions and their impact on the accuracy of Bayesian inference. *Statistics in Transition New Series*, 17, 763-780.

Hidiroglou, M.A., Beaumont, J.-F. et Yung, W. (2019). Élaboration d'un système d'estimation sur petits domaines à Statistique Canada. *Techniques d'enquête*, 45(1), 107-133. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2019001/article/00009-fra.pdf.

Lemoine, N.P. (2019). Moving beyond noninformative priors: Why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, 128: 912-928. DOI: 10.1111/oik.05985.

Lesage, E., Beaumont, J.-F. et Bocci, C. (2021). Deux diagnostics locaux pour évaluer l'efficacité du meilleur prédicteur empirique issu du modèle de Fay-Herriot. *Techniques d'enquête*, 47(2), 303-323. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2021002/article/00001-fra.pdf.

Lunn, D.J., Thomas, A., Best, N. et Spiegelhalter, D.J. (2000). WinBUGS – A Bayesian modeling framework: Concepts, structure and extensibility. *Statistics and Computing*, 10, 325-337.

Maiti, T., Ren, H. et Sinha, S. (2014). Prediction error of small area predictors shrinking both means and variances. *Scandinavian Journal of Statistics*, 41, 775-790.

Meng, X.L. (1994). Posterior predictive *p* value. *The Annals of Statistics*, 22, 1142-1160.

Méthodologie de l'Enquête sur la population active du Canada (2017). Statistique Canada : 71-526-X, ISBN 978-0-660-24068-8. Disponible à l'adresse https://www150.statcan.gc.ca/n1/pub/71-526-x/71-526-x2017001-fra.htm.

Mikkola, P., Martin, O.A., Chandramouli, S., Hartmann, M., Abril Pla, O., Thomas, O., Pesonen, H., Corander, J., Vehtari, A., Kaski, S., Bürkner, P.C. et Klami, A. (2024). Prior knowledge elicitation: The past, present, and future. *Bayesian Analysis*, 19, 1129-1161.

Molina, I., Nandram, B. et Rao, J.N.K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *Annals of Applied Statistics*, 8, 852-885.

Okonek, T. et Wakefield, J. (2024). A computationally efficient approach to fully Bayesian benchmarking. *Journal of Official Statistics*, 40, 283-316.

Parker, P.A., Holan, S.H. et Janicki, R. (2023). Conjugate modeling approaches for small area estimation with heteroscedastic structure. *Journal of Survey Statistics and Methodology*, smad002, https://doi.org/10.1093/jssam/smad002.

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

Rao, J.N.K. et Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *The Canadian Journal of Statistics*, 22, 511-528.

Rivest, L.P. et Vandal, N. (2002). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, July 10-13, 2002, Ottawa, Canada.

Sakshaug, J.W., Wiśniowski, A., Ruiz, D.A.P. et Blom, A.G. (2019). Supplementing small probability samples with nonprobability samples: A Bayesian approach. *Journal of Official Statistics*, 35, 653-681.

Sinharay, S. et Stern, H.S. (2003). Posterior predictive modelchecking in hierarchical models. *Journal of Statistical Planning and Inference*, 111, 209-221.

Spiegelhalter, D.J., Best, N., Carlin, B.P. et van der Linde, A. (2002). Bayesain measures of model complexity and fit. *Journal of Royal Statistical Society*, B, 64, 583-639.

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

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

Wiśniowski, A., Sakshaug, J.W., Ruiz, D.A.P. et Blom, A.G. (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, 8, 120-147.

You, Y. (2008a). Une approche intégrée de modélisation de l'estimation du taux de chômage pour les régions infraprovinciales au Canada. *Techniques d'enquête*, 34(1), 21-31. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2008001/article/10614-fra.pdf.

You, Y. (2008b). Small area estimation using area level models with model checking and applications. *Proceedings of Survey Methods Section*, *Statistical Society of Canada*.

You, Y. (2021). Estimation sur petits domaines à l'aide du modèle au niveau de domaine de Fay-Herriot avec lissage et modélisation de variance d'échantillonnage. *Techniques d'enquête*, 47(2), 389-399. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2021002/article/00007-fra.pdf.

You, Y. (2023). An empirical study of hierarchical Bayes small area estimators using different priors for model variances. *Statistics in Transition New Series*, 24, 139-148.

You, Y. et Chapman, B. (2006). Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances d'échantillonnage. *Techniques d'enquête*, 32(1), 107-114. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006001/article/9263-fra.pdf.

You, Y. et Hidiroglou, M. (2023). Application of sampling variance smoothing methods for small area proportion estimation. *Journal of Official Statistics*, 39, 571-590.

You, Y. et Rao, J.N.K. (2000). Estimation bayésienne hiérarchique des moyennes pour petites régions à l'aide de modèles à plusieurs niveaux. *Techniques d'enquête*, 26(2), 197-206. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2000002/article/5537-fra.pdf.

You, Y. et Rao, J.N.K. (2002) Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 30, 3-15.

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 Gambino, J. (2003). Estimation du taux de chômage fondée sur un modèle pour l'Enquête sur la population active du Canada : Une approche bayésienne hiérarchique. *Techniques d'enquête*, 29(1), 27-36. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2003001/article/6602-fra.pdf.

You, Y. et Zhou, Q.M. (2011). Estimation sur petits domaines hiérarchique bayésienne sous un modèle spatial avec application à des données d'enquête sur la santé. *Techniques d'enquête*, 37(1), 31-44. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2011001/article/11445-fra.pdf.

Zhang, J.L. et Bryant, J. (2020). Fully Bayesian benchmarking of small area estimation models. *Journal of Official Statistics*, 36, 197-223. DOI: https://doi.org/10.2478/jos-2020-0010.

Zondervan-Zwijnenburg, M., Peeters, M., Depaoli, S. et Van de Schoot, R. (2017). Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Research in Human Development*, 14, 305-320. DOI: 10.1080/15427609.2017.1370966.