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

Techniques d’enquête

Amélioration de l’inférence sur petits domaines grâce à l’intégration de données fondée sur des distributions a priori globales-locales

par Dexter Cahoy et Joseph Sedransk

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. 489-516
Statistique Canada, nº 12-001-X au catalogue

Amélioration de l'inférence sur petits domaines grâce à l'intégration de données fondée sur des distributions *a priori* globales-locales

Dexter Cahoy et Joseph Sedransk[1]

Résumé

Nous présentons et appliquons une méthodologie pour améliorer l'inférence des paramètres pour petits domaines en utilisant des données tirées de plusieurs sources. Les présents travaux prolongent ceux de Cahoy et Sedransk (2023) qui ont montré la façon d'intégrer des statistiques sommaires provenant de sources multiples. Notre méthodologie pour effectuer des inférences sur la proportion de personnes dans les comtés de Floride qui ne sont pas couvertes par un régime d'assurance-maladie s'appuie sur des distributions *a priori* hiérarchiques globales-locales. Les résultats d'une vaste étude par simulation montrent que cette méthodologie, fondée sur de multiples sources de données, permettra de produire de meilleures inférences. Parmi les cinq variantes du modèle qui ont été évaluées, celles reposant sur les distributions *a priori* de type *horseshoe* pour l'ensemble des variances produisent de meilleurs résultats que celles reposant sur les distributions *a priori* de type *LASSO* pour les variances locales.

Mots-clés : Combinaison de données; distribution *a priori* de type *horseshoe*; distribution *a priori* de type *LASSO*; observations aberrantes; régime d'assurance-maladie; regroupement de données; rétrécissement.

1. Introduction et contexte

En raison des changements importants observés dans le domaine de l'échantillonnage, qu'ils soient attribuables à la forte baisse des taux de réponse ou aux compressions budgétaires, une attention accrue est portée à l'élaboration et à l'utilisation d'une nouvelle méthodologie pour améliorer les inférences. Une approche prometteuse consiste à combiner des données provenant d'autres sources à celles d'une simple enquête probabiliste. Cahoy et Sedransk (2023) ont présenté deux méthodes pour combiner les statistiques sommaires tirées de plusieurs enquêtes-échantillons et les ont appliquées aux données du Behavioral Risk Factor Surveillance System (BRFSS) et du programme Small Area Health Insurance Estimates (SAHIE) du Bureau du recensement des États-Unis. Ces méthodes sont plus structurées que celles couramment utilisées dans l'échantillonnage d'enquête, et Cahoy et Sedransk (2023) ont démontré leur valeur pour ce qui est du regroupement de ce genre de données. Ils ont également souligné la valeur potentielle de cette méthodologie pour combiner les données provenant d'enquêtes-échantillons traditionnelles ainsi que d'enquêtes non probabilistes et, peut-être, de données administratives. Or, ayant fait une inférence pour chaque comté (en Floride) en s'appuyant uniquement sur les données du comté en question, Cahoy et Sedransk (2023) ont peut-être été privés d'autres améliorations dans l'inférence qui auraient été possibles en combinant les données entre les comtés. Dans le présent article, nous élargissons ces travaux en proposant une méthodologie pour faire des inférences simultanées pour tous les comtés lorsqu'il existe plusieurs sources de données. Cependant, la méthode principale utilisée par Cahoy et Sedransk (2023), soit la combinaison

1. Dexter Cahoy, University of Houston System, Houston, Texas, 77002. Courriel : cahoyd@uhd.edu; Joseph Sedransk, University of Maryland, College Park, Maryland, 20742. Courriel : jxs123@cwru.edu.

incertaine (Evans et Sedransk, 2001; Malec et Sedransk, 1992), ne permet pas, pour le moment, de satisfaire aux exigences de calcul de l'inférence pour plusieurs petits domaines en même temps que plusieurs sources de données. Nous utilisons donc des modèles globaux-locaux (GL), une option prometteuse qui fait l'objet d'une abondante littérature. Le document de synthèse de Bhadra, Datta, Polson et Willard (2019) fournit un résumé détaillé. Deux articles faisant partie de la littérature sur l'échantillonnage d'enquête ont été publiés : Tang, Ghosh, Ha et Sedransk (2018) et Tang et Ghosh (2023). Les résultats de l'étude de Tang et coll. (2018) ont motivé notre choix d'utiliser des modèles GL. Puisque les deux articles traitent d'une question plus simple que la nôtre, c'est-à-dire l'utilisation d'une seule source de données, nous avons étendu le modèle de Tang et coll. (2018).

Une version simplifiée du modèle Fay-Herriot qu'utilisent Tang et coll. (2018) est

$$y_i = \theta_i + e_i, \quad \theta_i = \eta + u_i, \quad i = 1, \ldots, I, \tag{1.1}$$

où y_i est l'estimation directe pour le petit domaine i et η est une constante. Dans ce cas-ci, nous supposons que $e = (e_1, \ldots, e_I)'$ et $u = (u_1, \ldots, u_I)'$ sont indépendants. Les éléments de e sont indépendants selon $e_i \sim N(0, V_i)$. Tang et coll. (2018) supposent que V_i est connue pour éviter des problèmes liés à l'identifiabilité. La distribution *a priori* en rétrécissement GL pour la u_i est

$$u_i \mid \lambda_i^2, \tau^2 \sim N(0, \lambda_i^2 \tau^2). \tag{1.2}$$

Au lieu de (1.1) et de (1.2), Tang et coll. (2018) remplacent η par une régression linéaire. Ils supposent une distribution *a priori* uniforme localement pour le paramètre de régression et envisagent plusieurs distributions *a priori* pour les variances; voir le tableau 1. Comme on peut le voir dans (1.2), il y a deux niveaux pour les variances des distributions *a priori*. La variance globale, τ^2, permet un rétrécissement pour tous les effets aléatoires, tandis que les variances locales, λ_i^2, fournissent des ajustements pour les domaines individuels. Si les distributions *a priori* sont à queues lourdes, on peut faire une inférence efficace tant pour les petits effets aléatoires que pour les effets aléatoires importants. Tang et Ghosh (2023) étendent les travaux de Tang et coll. (2018) en remplaçant (1.2) par une distribution normale ayant une structure spatiale. Une autre solution de rechange à (1.2) consiste à comparer une distribution *a priori* de type *spike-and-slab* (Datta et Mandal, 2015) à la distribution *a priori* GL dans Tang et coll. (2018). Autrement dit, $u_i = \xi_i v_i$, où $\xi_1, \ldots, \xi_I$ sont des variables aléatoires de Bernoulli indépendantes et identiquement distribuées comprenant $P(\xi_i = 1) = \gamma$ et $v_i \sim \mathrm{N}(0, \zeta^2)$ de façon indépendante. Une étude récente de Smith et Griffin (2023) comporte plusieurs éléments qui se rapprochent de notre prolongement de l'étude de Tang et coll. (2018) visant à effectuer des inférences sur petits domaines quand il y a plusieurs sources de données. Voir la section 2 pour obtenir plus de précisions.

Le reste du présent document est structuré comme suit. À la section 2, nous décrivons nos modèles et fournissons une explication de notre estimateur ponctuel, la moyenne *a posteriori* pour chaque comté de Floride. À la section 3, notre première analyse repose sur les données du BRFSS et du programme SAHIE que Cahoy et Sedransk (2023) ont utilisées pour faire des inférences sur la proportion d'adultes sans

assurance-maladie dans les comtés de Floride. La section 3.1 décrit brièvement l'enquête du BRFSS et le programme SAHIE. Les intervalles de crédibilité pour les moyennes des comtés de Floride se trouvent à la section 3.2, tandis que les résultats d'une comparaison de plusieurs modèles reposant sur des données du BRFSS et du programme SAHIE figurent à la section 3.3. La section 3.4 porte sur les avantages de l'utilisation de données provenant de deux sources plutôt que d'une seule source. La section 4 présente les résultats d'une vaste étude par simulation fondée en partie sur l'ensemble de données observées. Elle traite aussi de l'analyse de la littérature portant sur l'utilisation des distributions *a priori* de types *horseshoe* et *LASSO* (bayésien). La section 5 comporte une analyse des variances d'échantillonnage, de l'utilisation d'un plus grand nombre de sources de données et de la comparaison et de la robustesse des modèles, puis elle se termine par un résumé.

2. Modèles et inférence

La considération des données provenant de deux sources pour chacun des comtés de Floride menait à l'extension naturelle suivante du modèle de Tang et coll. (2018). Représentons par Y_{ij} la valeur de Y pour la j^{e} source dans l'i^{e} comté. Pour une valeur V_{ij} fixe,

$$\begin{aligned} Y_{ij} \mid \theta_{ij}, V_{ij} &\overset{\text{ind}}{\sim} N(\theta_{ij}, V_{ij}):\ j=1,...,J;\ i=1,...,I, \\ \theta_{ij} \mid \mu_i, \lambda_{ij}^2, \lambda_i^2, \tau_1^2 &\overset{\text{ind}}{\sim} N(\mu_i, \lambda_{ij}^2 \lambda_i^2 \tau_1^2), \\ \mu_i \mid \eta, \lambda_i^2, \tau_2^2 &\overset{\text{ind}}{\sim} N(\eta, \lambda_i^2 \tau_2^2), \\ f(\eta) &= \text{constante.} \end{aligned} \tag{2.1}$$

Comme dans l'étude de Cahoy et Sedransk (2023), Y est une estimation de la proportion de la population n'ayant pas d'assurance-maladie. Il convient de prendre note que les données pouvant être tirées du BRFSS et du programme SAHIE sont seulement Y et l'estimation de l'erreur-type de Y. Dans ce cas-ci, l'inférence pour les moyennes des comtés, μ_i, est d'un intérêt premier.

En définissant $\lambda = (\{\lambda_{ij}^2:\ j=1,\ldots,J, i=1,\ldots,I;\ \lambda_i^2: i=1,\ldots,I\})'$ et $\tau = (\tau_1^2, \tau_2^2)'$, la densité *a priori* conjointe est :

$$f(\lambda, \tau) \propto \left[\prod_{i,j} f(\lambda_{ij}^2)\right] \left[\prod_i f(\lambda_i^2)\right] f(\tau_1^2)\, f(\tau_2^2).$$

Pour choisir les distributions *a priori* pour les variances locales et les deux variances globales, nous nous sommes appuyés sur une analyse dans la littérature. Tang et coll. (2018) font remarquer que Gelman (2006) et Polson et Scott (2012) préfèrent utiliser les distributions *a priori* de type *horseshoe* pour les paramètres d'échelle de niveau supérieur, et il semble que ce choix reçoit un appui supplémentaire considérable. Pour obtenir plus de renseignements, voir la section 4.3.

Dans un document récent préparé par Smith et Griffin (2023), on utilise un modèle quelque peu similaire dans un contexte différent pour estimer la demande pour un grand éventail de biens différenciés. Leur

modèle s'articule autour de la représentation graphique d'un arbre de classification de produits. Dans cet arbre, le niveau le plus bas représente la définition la plus détaillée des groupes de produits, tandis que le niveau le plus élevé représente les plus grandes catégories. En utilisant cet arbre comme cadre, les auteurs élaborent des modèles hiérarchiques qui réduisent directement les élasticités au niveau des produits vers les élasticités des prix de niveau plus élevé. La structure dans notre problème est assez différente, c'est-à-dire l'emboîtement des sources (j) dans les comtés (i). Une autre hypothèse est que les objectifs des deux documents sont différents. Dans le contexte de l'inférence sur petits domaines, le choix naturel consiste à prendre η pour obtenir une distribution uniforme localement. Smith et Griffin (2023), quant à eux, supposent une distribution normale standard au niveau le plus élevé. Une autre différence est que Smith et Griffin (2023) ne considèrent que la structure hiérarchique pour les variances locales, c'est-à-dire $\lambda_{ij}^2\lambda_i^2$ dans (2.1); il s'agit d'une limitation, puisque nous avons observé de meilleurs résultats pour notre problème en utilisant la λ_{ij}^2 non hiérarchique dans la variance conditionnelle de θ_{ij}. Cependant, nous avons appliqué les idées de Smith et Griffin (2023) pour concevoir notre étude par simulation et clarifier certaines caractéristiques de notre approche.

Soit υ, une variance générale, et $\kappa \sim \mathcal{C}^+(0,1)$, la distribution demi-Cauchy ayant une fonction de densité de probabilité (fdp) $f(y) = 2/\pi(1+y^2),\ y \geq 0$. Par conséquent, $\upsilon = \kappa^2$ a la distribution de type *horseshoe* ayant une fdp $f(\upsilon) \propto \left[(1+\upsilon^2)\sqrt{\upsilon^2}\right]^{-1}$ pour $\upsilon > 0$. La distribution de type *LASSO* a une fdp $f(\upsilon) \propto \exp(-\upsilon)$ pour $\upsilon \geq 0$.

Nos choix sont les suivants :

Dans le modèle M11a, $\lambda_{ij}^2, \lambda_i^2, \tau_1^2, \tau_2^2$ ont des distributions de type *horseshoe*.

Le modèle M11b est le même que le modèle M11a, sauf que λ_{ij}^2 et λ_i^2 ont des distributions de type *LASSO*.

Le modèle M1a est le même que le modèle M11a, sauf que $\theta_{ij} \sim \mathrm{N}(\mu_i, \lambda_{ij}^2\tau_1^2)$.

Le modèle M1b est le même que le modèle M1a, sauf que λ_{ij}^2 et λ_i^2 ont des distributions de type *LASSO*.

Dans le modèle M12, $\lambda_{ij}^2 = \lambda_i^2 = 1$. Il s'agit du modèle hiérarchique habituel.

Les premiers modèles que nous avons choisis étaient les modèles M1a et M1b avec $\theta_{ij} \sim \mathrm{N}(\mu_i, \lambda_{ij}^2\tau_1^2)$. Nous avons ajouté les modèles plus génériques M11a et M11b (ayant une variance $\lambda_{ij}^2\lambda_i^2\tau_1^2$) après avoir lu Smith et Griffin (2023).

Pour étudier la moyenne *a posteriori* de μ_i étant donné les données et les variances observées, supposons le modèle M11a et définissons $A_i = \lambda_i^2\tau_2^2, \xi_{ij}^2 = V_{ij} + \lambda_{ij}^2\lambda_i^2\tau_1^2, \eta_i^2 = \left(\sum_{j=1}^J \xi_{ij}^{-2}\right)^{-1}$ et $\phi_i = A_i/(A_i + \eta_i^2)$.

En définissant $\overline{y}_i = \sum_{j=1}^J (y_{ij}/\xi_{ij}^2) \Big/ \sum_{j=1}^J (1/\xi_{ij}^2)$, la moyenne *a posteriori* de μ_i est

$$E(\mu_i \mid y, \Omega) = \phi_i \overline{y}_i + (1-\phi_i)\,\overline{y}_w, \tag{2.2}$$

où $\Omega = (\lambda, \tau, \{V_{ij}: j = 1,\ldots, J; i = 1,\ldots, I\})$, $y = \{y_{ij}: j = 1,\ldots, J; i = 1,\ldots, I\}$ et $\bar{y}_w = \sum_{i=1}^{I}(\bar{y}_i / (A_i + \eta_i^2)) \Big/ \sum_{i=1}^{I}(1/(A_i + \eta_i^2))$.

Le poids *au sein* des comtés, $\xi_{ij}^{-2} \Big/ \sum_{j=1}^{J} \xi_{ij}^{-2}$, se réduit à $V_{ij}^{-2} \Big/ \sum_{j=1}^{J} V_{ij}^{-2}$ si $\lambda_{ij}^2 \lambda_i^2 \tau_1^2 = 0$. Si ce n'est pas le cas, les composantes des variances locales et globales tiendront compte des valeurs observées qui ne sont pas conformes au modèle hiérarchique standard M12. (Dans la suite du présent document, ces observations seront désignées comme des observations « aberrantes ».) L'augmentation de V_{ij}, en raison d'une source aberrante (λ_{ij}^2 de grande taille) ou d'un comté aberrant (λ_i^2 de grande taille), réduira le poids sur y_{ij}. Si la valeur de V_{ij} est petite par rapport à la valeur de $\lambda_{ij}^2 \lambda_i^2 \tau_1^2$, $\xi_{ij}^2 = \lambda_{ij}^2 \lambda_i^2 \tau_1^2$, ce qui signifie que $\bar{y}_i = \sum_{j=1}^{J}(y_{ij} / \lambda_{ij}^2) \Big/ \sum_{j=1}^{J}(1/\lambda_{ij}^2)$. A_i est une mesure de la variation *dans l'ensemble* des comtés, tandis que η_i^2 est une mesure de la variation *au sein* des comtés (somme de la variance d'échantillonnage et de la variance d'une source à l'autre au sein des comtés). Ainsi, $\phi_i = A_i \big/ (A_i + \eta_i^2)$ correspond au facteur de rétrécissement global standard. Enfin, $\bar{y}_w$ est une moyenne pondérée des $\bar{y}_i$, où les poids comprennent les ajustements basés sur les variances locales λ_i^2 et λ_{ij}^2.

3. Analyse des données des comtés de Floride

3.1 Contexte

L'une des sources de données est le programme SAHIE. (Les données sont accessibles à l'adresse https://www.census.gov/data/datasets/time-series/demo/sahie/estimates-acs.html [en anglais seulement].) Le programme SAHIE repose sur des estimations ponctuelles tirées de l'American Community Survey ainsi que des données administratives, comme celles liées aux déclarations de revenus fédérales et aux taux de participation au Medicaid et au Children's Health Insurance Program. Il y a une modélisation détaillée au niveau du domaine. La deuxième source est le BRFSS, dont les données ont été recueillies lors d'interviews téléphoniques. (Les données peuvent être téléchargées à l'adresse https://www.flhealthcharts.gov/ [en anglais seulement].) Il repose sur un plan d'échantillonnage stratifié non proportionnel. Pour obtenir plus de renseignements, voir les sections 3 et 7 de Cahoy et Sedransk (2023). Afin d'assurer une cohérence avec les données de Cahoy et Sedransk (2023), nous utilisons les données de 2010 pour chaque source. Les données disponibles du BRFSS et du programme SAHIE pour chaque comté de Floride sont une estimation de la proportion de la population non assurée (Y) et de l'erreur-type de Y.

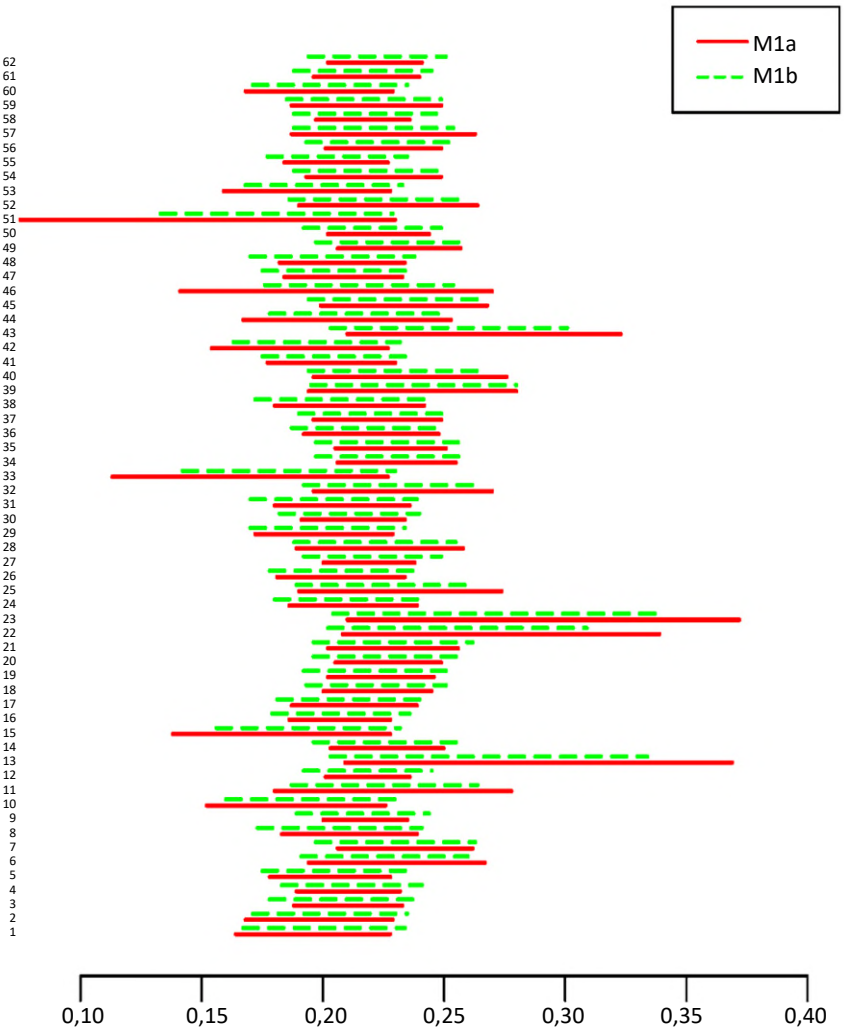
Puisque l'analyse des données des comtés de Floride est simple, nous la présentons en premier. Ainsi, elle fournit les renseignements sur lesquels s'appuie l'étude par simulation détaillée à la section 4.

3.2 Inférence

La figure 3.1 montre les intervalles de crédibilité à 95 % pour les comtés de Floride correspondant aux modèles M1a et M1b. Dans l'ensemble, les paires d'intervalles ont des centres similaires : les deux distributions, prises sur les I comtés, sont presque identiques (les premiers quartiles, les médianes et les troisièmes quartiles se situent à moins de 0,002). Même si la longueur médiane des intervalles est plus courte pour le modèle M1a (0,056) que pour le modèle M1b (0,064), il y a huit comtés où les intervalles du modèle M1a

sont beaucoup plus *grands*. Il s'agit de situations dans lesquelles les estimations ponctuelles du BRFSS et du programme SAHIE sont des valeurs aberrantes ou dans lesquelles il y a des écarts substantiels entre les deux estimations. Il s'agit d'une première indication que le modèle M1a (ayant des distributions *a priori* de type *horseshoe* pour l'ensemble des variances) produit des inférences plus appropriées lorsque des observations aberrantes sont présentes. Si l'on compare les modèles M11a et M11b, les conclusions sont similaires.

Figure 3.1 Intervalles de crédibilité à 95 % pour les 62 moyennes de comtés, selon les modèles M1a et M1b

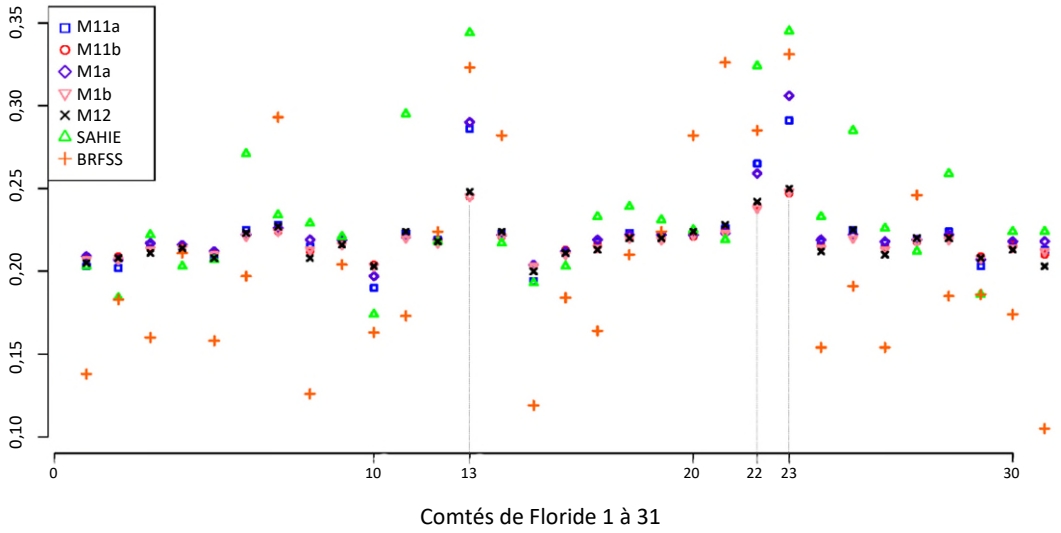


3.3 Comparaisons des modèles

Les figures 3.2 et 3.3 indiquent la moyenne *a posteriori*, μ_i, pour chaque comté de Floride dans les modèles M11a, M11b, M1a et M1b, le modèle standard M12 et les deux estimations ponctuelles. Premièrement, notons qu'il y a peu de variation entre les comtés dans le modèle M12. Pour la plupart des comtés, il y a accord dans les cinq modèles. Les valeurs de la moyenne *a posteriori* des modèles M1b et M11b (ayant

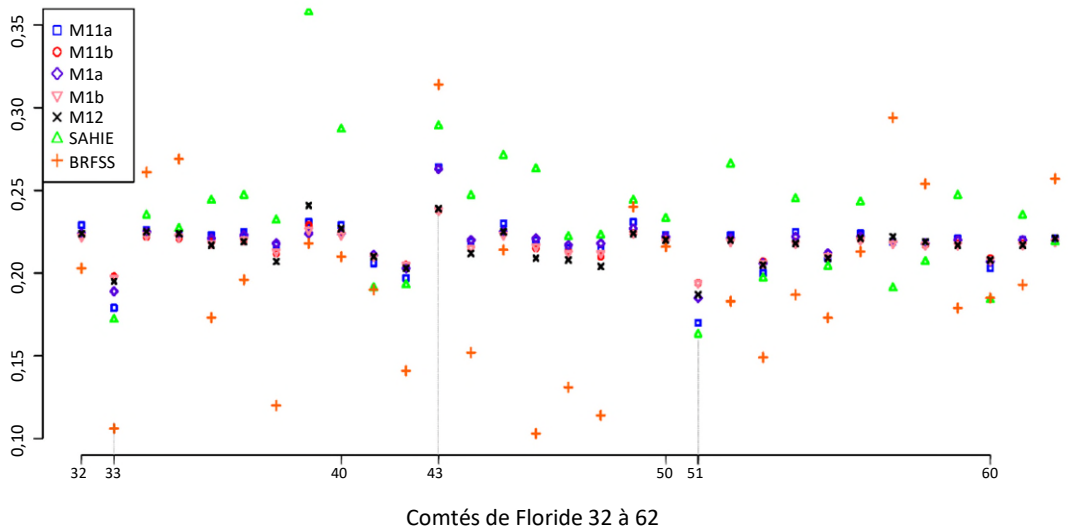
tous deux des distributions de type *LASSO*) sont beaucoup plus proches de celles du modèle M12 que de celles des modèles M1a et M11a (ayant tous deux des distributions de type *horseshoe*). On le voit bien dans le cas des comtés ayant des valeurs à la fois élevées et faibles des estimations ponctuelles du BRFSS et du programme SAHIE comme 10, 13, 22 et 23 (voir la figure 3.2). Les valeurs des modèles M11b et M1b sont similaires, tandis qu'en général, les valeurs du modèle M11a sont un peu plus extrêmes que celles du modèle M1a.

Figure 3.2 Moyennes *a posteriori*, estimations du BRFSS et du programme SAHIE pour les comtés de Floride 1 à 31



Note : BRFSS : Behavioral Risk Factor Surveillance System; SAHIE : Small Area Health Insurance Estimates.

Figure 3.3 Moyennes *a posteriori*, estimations du BRFSS et du programme SAHIE pour les comtés de Floride 32 à 62



Note : BRFSS : Behavioral Risk Factor Surveillance System; SAHIE : Small Area Health Insurance Estimates.

Pour la plupart des comtés, la différence entre les modèles M1a et M1b n'est pas importante. Généralement, les valeurs du modèle M1b sont plus proches de celles du modèle M1a que de celles du modèle standard M12. Il y a beaucoup plus de comtés pour lesquels les valeurs du modèle M12 sont plus proches de celles du programme SAHIE que de celles du BRFSS. Lorsque les valeurs du programme SAHIE sont faibles et que celles du BRFSS sont élevées ou vice versa, les valeurs des modèles M1a et M1b ont tendance à être semblables les unes aux autres et à se situer dans la fourchette définie dans le cadre du programme SAHIE et du BRFSS. La constatation la plus importante est que lorsque les valeurs du programme SAHIE et du BRFSS sont toutes élevées ou faibles, le modèle M1b rétrécit davantage que le modèle M1a. De même, le modèle M11b rétrécit davantage que le modèle M11a. Cela concorde avec ce qui est rapporté dans la littérature. Pour leur modèle en une étape, c'est-à-dire sans effets de source, les travaux de Tang et coll. (2018) font ressortir que les facteurs de rétrécissement dans les distributions *a priori* (de type *LASSO*) sont stochastiquement plus grands que ceux dans les distributions *a priori* de type *horseshoe* et causent donc un rétrécissement plus important. Smith et Griffin (2023) précisent que les queues d'une densité de mélange exponentielle sont plus légères que les queues polynomiales de la distribution demi-Cauchy, ce qui laisse croire que la méthode *LASSO* bayésienne peut avoir tendance à entraîner un rétrécissement excessif des grands coefficients de régression et un rétrécissement insuffisant des petits coefficients comparativement à la distribution de type *horseshoe*.

Comme le montre (2.2), l'ampleur de ϕ détermine la contribution des observations du comté i à la moyenne *a posteriori* de μ_i. Les figures A.1 à A.3 de l'annexe montrent les distributions *a posteriori* de $\{\phi_i\colon i=1,\ldots,I\}$ pour les modèles M1a, M1b, M11a, M11b et M12. Elles révèlent les écarts importants qui existent entre les modèles. Pour le modèle M12, l'ensemble des I distributions de ϕ varie peu, ce qui est nettement différent de celui des quatre modèles GL. En comparant le modèle M1a (ayant des distributions *a priori* de type *horseshoe* pour l'ensemble des variances) avec le modèle M1b (ayant des distributions *a priori* de type *LASSO* pour les variances locales et des distributions *a priori* de type *horseshoe* pour les variances globales), on peut constater des différences notables, notamment une variation beaucoup plus grande pour le modèle M1a tant pour l'ensemble des comtés qu'au sein des comtés. On constate également cette tendance pour les homologues des modèles M1a et M1b, les modèles M11a et M11b, qui se distinguent des modèles M1a et M1b par le fait qu'ils ont des variances locales $\lambda_{ij}^2\lambda_i^2$ plutôt que λ_{ij}^2.

3.4 Utilisation de deux sources plutôt qu'une

Nous utilisons l'ensemble de données des comtés de Floride pour illustrer les différences dans les inférences lorsqu'il n'existe qu'une seule source de données plutôt que deux sources de données. Le modèle à source unique est

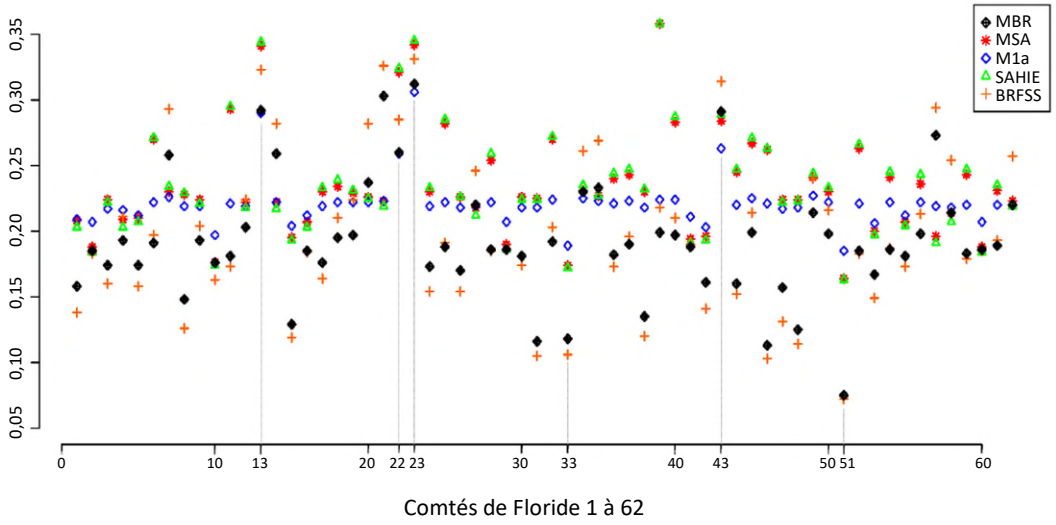
$$Y_i \mid \mu_i, V_i \overset{\text{ind}}{\sim} N(\mu_i, V_i),$$

$$\mu_i \mid \eta \overset{\text{ind}}{\sim} N(\eta, \lambda_i^2\tau^2),$$

$$f(\eta) = \text{ constante.}$$

La comparaison la plus significative est celle ayant trait à l'ensemble de données du BRFSS, puisque ce dernier comporte des erreurs-types observées beaucoup plus importantes; ainsi, il pourrait possiblement tirer parti de l'ajout des données du programme SAHIE lorsque cela convient. On peut considérer l'analyse présentée dans la présente section comme une illustration ou un choix réaliste en raison des hypothèses (apparemment) non vérifiées avancées dans le cadre du programme SAHIE. Pour chaque comté, la figure 3.4 donne les estimations ponctuelles et les moyennes *a posteriori* du BRFSS et du programme SAHIE correspondant au modèle M1a (de type *horseshoe*) et aux modèles de type *horseshoe* à source unique (MBR [modèle pour le BRFSS], MSA [modèle pour le SAHIE]) appliqués aux données du BRFSS et du programme SAHIE.

Figure 3.4 Estimations du MBR, du MSA et du M1a, et estimations observées du BRFSS et du programme SAHIE

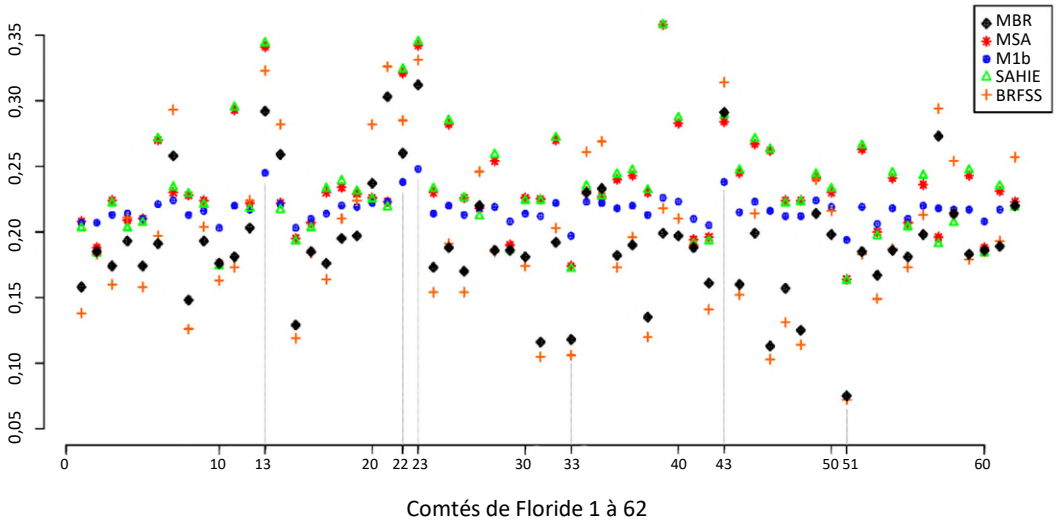


Note : BRFSS : Behavioral Risk Factor Surveillance System; MBR : Model BRFSS; MSA : Model SAHIE; SAHIE : Small Area Health Insurance Estimates.

En règle générale, la moyenne *a posteriori* du BRFSS (reposant sur le modèle de type *horseshoe* à source unique) permet d'ajuster l'estimation ponctuelle du BRFSS par l'augmentation des valeurs plus faibles et la diminution des valeurs plus élevées de cette dernière. La moyenne *a posteriori* du modèle M1a permet d'ajuster la moyenne *a posteriori* du BRFSS reposant sur le modèle à source unique par l'augmentation des valeurs plus faibles et la diminution des valeurs plus élevées de cette dernière. En reposant sur cet ensemble de données, la moyenne *a posteriori* du modèle M1a ne varie pas beaucoup d'un comté à l'autre. Ainsi, des ajustements ne doivent être apportés que pour un ensemble assez restreint de comtés. Néanmoins, il convient de noter que le modèle M1a possède d'excellentes propriétés, comme en témoignent les résultats de l'étude par simulation (section 4).

Les résultats pour le modèle M11a sont semblables à ceux pour le modèle M1a. Cela n'est pas surprenant, puisque les modèles M11a et M1a diffèrent seulement dans la variance de θ_{ij} (voir la section 2). À une exception importante près, les conclusions ci-dessus restent valables pour les modèles M11b et M1b, qui reposent tous deux sur des distributions *a priori* de type *LASSO*. Contrairement aux modèles M11a et M1a, on observe un rétrécissement important jusqu'à la valeur commune des comtés même pour certains comtés (par exemple les comtés 13, 22, 23, 33, 43 et 51) qui ont des valeurs quelque peu extrêmes de l'estimation ponctuelle du BRFSS; voir la figure 3.5, qui a le même format que la figure 3.4.

Figure 3.5 Estimations du MBR, du MSA et du M1b, et estimations observées du BRFSS et du programme SAHIE



Note : BRFSS : Behavioral Risk Factor Surveillance System; MBR : Model BRFSS; MSA : Model SAHIE; SAHIE : Small Area Health Insurance Estimates.

4. Étude par simulation

4.1 Introduction

Nous avons conçu notre étude par simulation d'après celles de Tang et coll. (2018) et de Smith et Griffin (2023). Dans la section qui suit, nous comparons le rendement des modèles M1a, M1b, M11a et M11b avec celui de la méthode de base M12. (Voir la section 2 pour obtenir les définitions de ces modèles.) Nous présentons les résultats obtenus en générant des données à l'aide des modèles suivants.

Représentons par Y_{ij} la valeur de Y pour la j^{e} source dans l'i^{e} comté. Pour une valeur V_{ij} fixe,

$$\begin{aligned} Y_{ij} &\overset{\text{ind}}{\sim} N(\theta_{ij}, V_{ij}): \; j = 1,\ldots,J; \; i = 1,\ldots,I, \\ \theta_{ij} &\overset{\text{iid}}{\sim} N(\mu_i, \gamma_1^2), \\ \mu_i &\overset{\text{iid}}{\sim} N(\eta, \gamma_2^2). \end{aligned} \tag{4.1}$$

Pour la γ^2, nous avons utilisé des représentations des mélanges et des valeurs aberrantes. Pour un mélange, $\gamma^2 = \delta_1\tau_{21}^2 + (1-\delta_1)\,\tau_{22}^2$, $\delta_1 \in \{0,1\}$ avec $P(\delta_1 = 1) = p_1$ et $\tau_{22}^2 = 0{,}05^2$. Pour une valeur aberrante, $\gamma^2 = \delta_2\tau_{11}^2$ avec $\delta_2 \in \{0,1\}$ et $P(\delta_2 = 1) = p_2$. Étant donné le grand nombre de quantités à préciser, nous avons simplifié le processus en omettant l'indexation par comté ou par source. Il existe des exceptions, par exemple pour les cas 5 et 6, décrits à la section 4.5.

Pour ancrer la présente étude par simulation, nous avons choisi des valeurs de certaines quantités en nous appuyant sur les données du BRFSS et du programme SAHIE : η est une constante de valeur 0,25 selon $I = 62$ comtés de Floride et $J = 2$ sources. De plus, les variances d'échantillonnage, V_{ij}, correspondent aux valeurs observées dans les enquêtes dont il est question. La section 5 présente un examen plus approfondi des variances d'échantillonnage et aborde le problème d'identifiabilité souligné par Tang et coll. (2018).

Dans le cas 1, les deux γ sont considérées comme des valeurs aberrantes, tandis que dans le cas 3, elles correspondent à des mélanges. Dans le cas 2, γ_1^2 est une valeur aberrante, tandis que γ_2^2 est un mélange. Dans le cas 4, γ_1^2 est un mélange, tandis que γ_2^2 est une valeur aberrante. Nous avons considéré plusieurs valeurs pour les probabilités (p_1, p_2) et les variances (τ^2) et avons présenté celles qui sont les plus informatives.

Pour chacun des quatre cas et chaque choix des probabilités et des variances, nous avons généré 100 ensembles de données. Pour chacun des ensembles de données, nous avons estimé μ_i à l'aide de sa moyenne *a posteriori*, $\hat{\mu}_i$, et avons évalué l'ajustement en employant quatre mesures courantes, à savoir le biais relatif absolu moyen (BRAM), le biais relatif quadratique moyen (BRQM), le biais absolu moyen (BAM) et la moyenne des écarts quadratiques (MEQ). Pour chaque ensemble de données, nous résumons les valeurs à l'aide du BRAM $= I^{-1}\sum_{i=1}^{I}\frac{(|\hat{\mu}_i - \mu_i|)}{\mu_i}$, du BRQM $= I^{-1}\sum_{i=1}^{I}\frac{(\hat{\mu}_i - \mu_i)^2}{\mu_i^2}$, du BAM $= I^{-1}\sum_{i=1}^{I}|\,\hat{\mu}_i - \mu_i\,|$ et de la MEQ $= I^{-1}\sum_{i=1}^{I}(\hat{\mu}_i - \mu_i)^2$, où I correspond au nombre de comtés de Floride. Pour chacune de ces mesures (BRAM, BRQM, BAM et MEQ), la statistique sommaire définitive correspond à la médiane des 100 ensembles de données. Par souci de simplicité, seuls les résultats pour le BRAM et le BRQM sont présentés.

4.2 Calcul *a posteriori*

Tous les calculs ont été réalisés en R à l'aide du Pittsburgh Supercomputing Center de l'Advanced Cyberinfrastructure Coordination Ecosystem : Services & Support (ACCESS). Nous avons généré 100 échantillons de données et avons calculé des estimations *a posteriori* en exécutant 18 000 simulations de la méthode de Monte Carlo par chaîne de Markov (MCMC) et en effectuant 3 000 rodages pour chaque échantillon. Pour tester la convergence, nous avons simulé cinq chaînes ayant une longueur de 7 000 (et 2 000 rodages) pour les distributions *a posteriori* des μ_i et avons obtenu des valeurs de la variante de fractionnement de R (*split-R variant*) du diagnostic de Gelman proches de 1,0 pour l'ensemble des 62 comtés. La section 11.4 de l'article de Gelman, Carlin, Stern, Dunson, Vehtari et Rubin (2013) traite des mesures de la convergence, y compris la variante de fractionnement de R. Les distributions conditionnelles complètes pour les modèles M11a et M11b sont présentées à l'annexe A. Les algorithmes de Gibbs correspondants pour les modèles M1a et M1b peuvent être obtenus en simplifiant les expressions données.

4.3 Évaluation globale

Nos comparaisons sont effectuées par rapport au modèle standard, le modèle M12, et sont présentées sous forme de ratios d'écart. Le modèle 12 est fourni à la section 2; il s'agit du cas particulier du modèle GL avec $\lambda_{ij}^2 = \lambda_i^2 = 1$. Pour les modèles Mx (x = 1a, 1b, 11a et 11b) et le BRAM, le ratio d'écart est désigné par BRAM(Mx)/BRAM(M12) avec de petites valeurs du ratio montrant des gains pour le modèle Mx par rapport au modèle M12. Nous résumons les distributions (sur l'ensemble des spécifications) du ratio d'écart pour les modèles M1a, M1b, M11a et M11b à l'aide des médianes.

Pour les cas 1 à 4, les médianes de BRAM(M1a)/BRAM(M12), de BRAM(M11a)/BRAM(M12), de BRAM(M1b)/BRAM(M12) et de BRAM(M11b)/BRAM(M12) sont de (0,589; 0,633; 0,802; 0,853), de (0,733; 0,760; 0,851; 0,874), de (0,780; 0,821; 0,870; 0,915) et de (0,786; 0,913; 0,835; 0,929). Ces valeurs révèlent des gains considérables pour l'ensemble des modèles GL, et le modèle M1a a les plus petites valeurs (soit le gain le plus important). Les résultats pour le BRQM sont similaires.

Des renseignements supplémentaires sont fournis dans les tableaux S1 à S4 de la documentation supplémentaire (https://arxiv.org/pdf/2412.07824 [en anglais seulement]) correspondant aux quatre cas. Dans chaque tableau, nous présentons, pour le BRAM et le BRQM ainsi que chacun des quatre ratios d'écart, la valeur minimale, le premier quartile, la médiane, la moyenne, le troisième quartile et la valeur maximale (s'appliquant à l'ensemble des spécifications).

Le tableau 4.1 donne un résumé global. Pour chaque ratio d'écart, cas et modèle Mx, la valeur de la cellule correspond au nombre de spécifications pour lesquelles le modèle Mx a la plus petite valeur parmi l'ensemble des modèles. De toute évidence, la préférence est donnée au modèle M1a. De plus, dans le cas du modèle M1a, la valeur de BRAM(M1a)/BRAM(M12) est plus petite que celle de BRAM(M1b)/BRAM(M12) pour 99 des 122 paramètres. Dans l'autre modèle de type *horseshoe*, soit le modèle M11a, la valeur de BRAM(M11a)/BRAM(M12) est plus petite que celle de BRAM(M11b)/BRAM(M12) pour 95 paramètres. Dans les résultats pour le BRQM, on note une préférence plus marquée pour ces modèles de type *horseshoe*. Voir le tableau S21 pour obtenir une ventilation selon le cas.

Ces résultats renforcent davantage la crédibilité de la littérature montrant une préférence pour la distribution *a priori* de type *horseshoe*. Par exemple, Smith et Griffin (2023) précisent que les queues d'une densité de mélange exponentielle sont plus légères que les queues polynomiales de la distribution demi-Cauchy, ce qui laisse croire que la méthode *LASSO* bayésienne peut avoir tendance à entraîner un rétrécissement excessif des grands coefficients de régression et un rétrécissement insuffisant des petits coefficients comparativement à la distribution de type *horseshoe*. Makalic et Schmidt (2016) précisent que le choix d'une distribution *a priori* demi-Cauchy pour les hyperparamètres GL entraîne un rétrécissement important des petits coefficients (c'est-à-dire le bruit) et quasi aucun rétrécissement des coefficients suffisamment grands (c'est-à-dire le signal). Cela contraste avec la méthode *LASSO* bayésienne bien connue… et les hiérarchies bayésiennes ridge où l'effet de rétrécissement est uniforme pour l'ensemble des coefficients.

Pour le modèle plus simple de (1.1) et de (1.2), Bhadra et coll. (2019) montrent la raison pour laquelle la distribution *a priori* de type *horseshoe* est efficace. Cet argument s'applique également pour l'inférence à propos des effets de source dans notre modèle de (2.1), puisque les deux premières expressions de (2.1) sont essentiellement les mêmes que dans le modèle de (1.1) et de (1.2) [l'indice i est supprimé]. Pour notre modèle préféré, soit le modèle M1a, nous supposons que $\lambda_i^2 = 1$. Alors

$$E(\theta_{ij} \mid y_{ij}) = \left[1 - E(\kappa_{ij} \mid y_{ij})\right] y_{ij} + E(\kappa_{ij} \mid y_{ij})\, \mu_i, \tag{4.2}$$

où le poids de rétrécissement, κ, est $\kappa_{ij} = V_{ij} \big/ (V_{ij} + \lambda_{ij}^2 \tau_1^2)$. Bhadra et coll. (2019) expriment la vraisemblance marginale de y_{ij} en fonction de κ_{ij}, soulignant que la densité *a posteriori* du poids de rétrécissement permet de déterminer les signaux et le bruit en laissant $\kappa_{ij} \to 0$ et $\kappa_{ij} \to 1$. La vraisemblance marginale étant nulle lorsque le poids de rétrécissement est nul, elle n'aide pas à déterminer les signaux. Cependant, comme le montrent Bhadra et coll. (2019), l'utilisation de la distribution *a priori* de type *horseshoe* permet à la distribution *a posteriori* de s'approcher à la fois de zéro et de un. Cela fournit une indication de la raison pour laquelle le modèle M1a (ayant une distribution *a priori* de type *horseshoe* au niveau de la source) présente un meilleur rendement que le modèle M1b (ayant une distribution *a priori* de type *LASSO* au niveau de la source).

Tableau 4.1
Pour chaque cas : nombre de spécifications où un modèle présente le meilleur rendement

		BRAM		
Cas	M11a	M11b	M1a	M1b
1	0	1	29	0
2	8	0	28	0
3	0	1	16	3
4	0	0	17	19
		BRQM		
Cas	M11a	M11b	M1a	M1b
1	5	0	15	0
2	6	0	30	0
3	9	0	8	3
4	2	1	21	12

Note : BRAM : biais relatif absolu moyen; BRQM : biais relatif quadratique moyen.

Les tableaux 4.2 à 4.5 donnent des renseignements plus détaillés sur le modèle M1a. Chaque tableau concerne un cas en particulier. Par exemple, dans le tableau 4.2, les deux γ^2 sont des valeurs aberrantes. Chaque rangée correspond à une spécification des probabilités et des variances, τ^2. Le premier ensemble de colonnes permet de déterminer les probabilités et les τ^2. Par exemple, pour le cas 1, les composantes de γ_1^2 dans (4.1) correspondent au modèle présentant des valeurs aberrantes et les composantes de γ_2^2 dans (4.1) correspondent au modèle présentant des valeurs aberrantes. Les deux dernières colonnes affichent les ratios d'écart pour le BRAM et le BRQM correspondant au modèle M1a, c'est-à-dire BRAM(M1a)/BRAM(M12) et BRQM(M1a)/BRQM(M12). Les tableaux S5 à S16 de la documentation supplémentaire fournissent les mêmes renseignements relativement aux autres modèles. Ces tableaux peuvent être utilisés pour étudier les effets d'un changement des valeurs des probabilités et des variances; voir la section 4.4.

4.4 Changement des probabilités et des variances

Nous étudions ensuite les effets d'un changement des probabilités et des valeurs des variances (τ^2). Il s'agit d'une nouvelle évaluation, puisque ces quantités sont fixes dans les études de Tang et coll. (2018) et de Smith et Griffin (2023). Motivés par les résultats présentés à la section 3.3 et les ouvrages mentionnés à la section 2, nous résumons uniquement nos constatations sur les propriétés du modèle M1a. Les valeurs des probabilités et des variances ont été choisies initialement en fonction des résultats présentés à la section 3. D'autres valeurs ont été ajoutées pour essayer de rendre les conclusions plus précises. Les résultats pour un sous-ensemble des probabilités et des variances que nous avons sélectionnées et prises en compte figurent aux tableaux 4.2 à 4.5. En général, les gains pour le modèle M1a augmentent à mesure que les probabilités et les variances augmentent. Il existe quelques exceptions qui ne peuvent être expliquées par un examen plus approfondi.

Tableau 4.2
Spécifications de l'étude par simulation et BRAM(M1a)/BRAM(M12), BRQM(M1a)/BRQM(M12) pour le cas 1

	p_2 pour μ_i	p_2 pour θ_{ij}	τ_{11} pour μ_i	τ_{11} pour θ_{ij}	M1a/M12(BRAM)	M1a/M12(BRQM)
1	0,100	0,100	0,025	0,025	0,803	0,670
2			0,050	0,050	0,563	0,466
3			0,100	0,100	0,345	0,226
4			0,200	0,200	0,244	0,058
5			0,050	0,100	0,487	0,327
6			0,100	0,050	0,425	0,282
7	0,200	0,200	0,025	0,025	0,828	0,832
8			0,050	0,050	0,635	0,601
9			0,100	0,100	0,429	0,293
10			0,200	0,200	0,298	0,123
11			0,050	0,100	0,636	0,544
12			0,100	0,050	0,562	0,486
13	0,400	0,400	0,025	0,025	0,896	0,874
14			0,050	0,050	0,772	0,784
15			0,100	0,100	0,636	0,500
16			0,200	0,200	0,612	0,481
17			0,050	0,100	0,780	0,727
18			0,100	0,050	0,728	0,711
19	0,100	0,200	0,025	0,025	0,923	0,888
20			0,050	0,050	0,595	0,442
21			0,100	0,100	0,405	0,328
22			0,200	0,200	0,255	0,096
23			0,050	0,100	0,609	0,579
24			0,100	0,050	0,456	0,370
25	0,200	0,100	0,025	0,025	0,796	0,773
26			0,050	0,050	0,600	0,504
27			0,100	0,100	0,415	0,240
28			0,200	0,200	0,309	0,093
29			0,050	0,100	0,544	0,369
30			0,100	0,050	0,583	0,433

Note : BRAM : biais relatif absolu moyen; BRQM : biais relatif quadratique moyen.

Examinons le cas 1 (celui des deux modèles présentant des valeurs aberrantes, tableau 4.2) et le BRAM. Les gains observés dans le modèle M1a (par rapport à ceux observés dans le modèle M12) augmentent, comme prévu, à mesure que les deux $\tau_{11}, \tau_{11}(\mu_i)$ pour μ_i et $\tau_{11}(\theta_{ij})$ pour θ_{ij}, augmentent. Les gains

observés dans le modèle M1a sont généralement plus importants pour $\left(\tau_{11}(\mu_i), \tau_{11}(\theta_{ij})\right) =$ (0,10; 0,05) que pour (0,05; 0,10). Autrement dit, les gains sont plus importants lorsque la variance de μ_i est plus grande que celle de θ_{ij} dans les deux modèles présentant des valeurs aberrantes. Les gains observés dans le modèle M1a diminuent à mesure que les deux p_2, $p_2(\theta_{ij})$ et $p_2(\theta_{ij})$, augmentent pour passer de 0,1 à 0,2, puis à 0,4. Cela s'explique du fait que la p_2 plus élevée dans le modèle présentant des valeurs aberrantes le rapproche du modèle M12. Aucune tendance n'a été observée dans les gains associés à $\left(p_2(\mu_i), p_2(\theta_{ij})\right) =$ (0,1; 0,2) par rapport à ceux associés à $\left(p_2(\mu_i), p_2(\theta_{ij})\right) =$ (0,2; 0,1). Les tendances observées en ce qui concerne le BRQM sont comparables à celles observées en ce qui concerne le BRAM.

Tableau 4.3
Spécifications de l'étude par simulation et BRAM(M1a)/BRAM(M12), BRQM(M1a)/BRQM(M12) pour le cas 2

	p_1 pour μ_i	p_2 pour θ_{ij}	τ_{21} pour μ_i	τ_{11} pour θ_{ij}	M1a/M12(BRAM)	M1a/M12(BRQM)
1	0,100	0,100	0,100	0,050	0,942	0,892
2				0,100	0,725	0,520
3				0,200	0,502	0,255
4			0,200	0,050	0,894	0,816
5				0,100	0,723	0,707
6				0,200	0,551	0,375
7		0,200	0,100	0,050	0,936	0,887
8				0,100	0,709	0,473
9				0,200	0,487	0,166
10			0,200	0,050	0,887	0,823
11				0,100	0,739	0,569
12				0,200	0,498	0,217
13		0,400	0,100	0,050	0,951	0,873
14				0,100	0,742	0,481
15				0,200	0,525	0,259
16			0,200	0,050	0,923	0,862
17				0,100	0,754	0,617
18				0,200	0,556	0,330
19	0,200	0,100	0,100	0,050	0,959	0,940
20				0,100	0,713	0,579
21				0,200	0,483	0,252
22			0,200	0,050	0,924	0,826
23				0,100	0,702	0,570
24				0,200	0,519	0,405
25		0,200	0,100	0,050	0,943	0,892
26				0,100	0,764	0,616
27				0,200	0,465	0,203
28			0,200	0,050	0,886	0,723
29				0,100	0,749	0,515
30				0,200	0,492	0,245
31		0,400	0,100	0,050	0,917	0,725
32				0,100	0,775	0,603
33				0,200	0,521	0,260
34			0,200	0,050	0,919	0,793
35				0,100	0,727	0,668
36				0,200	0,572	0,411

Note : BRAM : biais relatif absolu moyen; BRQM : biais relatif quadratique moyen.

Ensuite, examinons le cas 2 (μ_i est un « mélange »; θ_{ij} est une « valeur aberrante », tableau 4.3) et le BRAM. Les gains observés dans le modèle M1a (par rapport au modèle M12) augmentent, comme prévu, à mesure que $\tau_{21}(=\tau_{11})$ augmente. Les gains observés dans le modèle M1a sont plus importants pour $(\tau_{21}, \tau_{11}) =$ (0,1; 0,2) que pour (0,2; 0,1). Autrement dit, les gains sont plus importants lorsque la variance de θ_{ij} est plus grande que celle de μ_i. Les gains pour le modèle M1a augmentent à mesure que $p_1(= p_2)$ augmente pour passer de 0,1 à 0,2, alors que seules de légères différences ont été observées pour $(p_1, p_2) =$ (0,1; 0,2) par rapport à (0,2; 0,1). En ce qui concerne les spécifications comportant des variances changeantes, les tendances observées pour le BRQM sont comparables à celles observées pour le BRAM. Pour les spécifications comportant des probabilités changeantes, les tendances ne sont pas uniformes.

Dans le cas 3, μ_i et θ_{ij} sont des « mélanges » (tableau 4.4). Pour le BRAM, les gains augmentent à mesure que τ_{21} pour μ_i augmente pour passer de 0,1 à 0,2, puis à 0,4, tandis que les variations des gains sont généralement faibles à mesure que τ_{21} pour θ_{ij} augmente pour passer de 0,1 à 0,2, puis à 0,4. Seules de faibles variations ont été observées à mesure que p_1 pour μ_i, $p_1(\mu_i)$, augmente pour passer de 0,1 à 0,2, et aucune tendance n'a été décelée à mesure que p_1 pour θ_{ij}, $p_1(\theta_{ij})$, augmente. Les tendances observées pour le BRQM sont comparables à celles observées pour le BRAM.

Le cas 4 comporte un modèle présentant des valeurs aberrantes pour μ_i et un modèle de « mélange » pour θ_{ij} (tableau 4.5). Pour le BRAM, les gains augmentent à mesure que τ_{11} augmente pour passer de 0,05 à 0,10, puis à 0,20, et à mesure que τ_{21} augmente pour passer de 0,1 à 0,2, puis à 0,4. Pour presque l'ensemble des spécifications, les gains augmentent à mesure que p_1 et p_2 augmentent pour passer de 0,1 à 0,2. Les tendances observées pour le BRQM sont comparables à celles observées pour le BRAM.

Tableau 4.4
Spécifications de l'étude par simulation et BRAM(M1a)/BRAM(M12), BRQM(M1a)/BRQM(M12) pour le cas 3

	p_1 pour μ_i	p_1 pour θ_{ij}	τ_{21} pour μ_i	τ_{21} pour θ_{ij}	M1a/M12(BRAM)	M1a/M12(BRQM)
1	0,100	0,100	0,100	0,100	1,050	1,102
2			0,200	0,200	0,844	0,452
3			0,400	0,400	0,655	0,305
4			0,200	0,400	0,680	0,214
5			0,400	0,200	0,828	0,562
6	0,200	0,200	0,100	0,100	1,001	0,960
7			0,200	0,200	0,766	0,470
8			0,400	0,400	0,636	0,427
9			0,200	0,400	0,619	0,275
10			0,400	0,200	0,783	0,763
11	0,100	0,200	0,100	0,100	1,025	0,997
12			0,200	0,200	0,786	0,527
13			0,400	0,400	0,619	0,353
14			0,200	0,400	0,671	0,429
15			0,400	0,200	0,778	0,612
16	0,200	0,100	0,100	0,100	1,006	1,005
17			0,200	0,200	0,852	0,563
18			0,400	0,400	0,609	0,301
19			0,200	0,400	0,596	0,296
20			0,400	0,200	0,904	0,893

Note : BRAM : biais relatif absolu moyen; BRQM : biais relatif quadratique moyen.

Tableau 4.5
Spécifications de l'étude par simulation et BRAM(M1a)/BRAM(M12), BRQM(M1a)/BRQM(M12) pour le cas 4

	p_2 pour μ_i	τ_{11} pour μ_i	p_1 pour θ_{ij}	τ_{21} pour θ_{ij}	M1a/M12(BRAM)	M1a/M12(BRQM)
1	0,100	0,050	0,100	0,100	1,353	1,545
2				0,200	1,175	1,344
3				0,400	0,815	0,777
4			0,200	0,100	1,270	1,402
5				0,200	1,122	1,254
6				0,400	0,775	0,807
7		0,100	0,100	0,100	1,059	0,725
8				0,200	0,865	0,584
9				0,400	0,749	0,504
10			0,200	0,100	1,025	0,693
11				0,200	0,804	0,684
12				0,400	0,760	0,817
13		0,200	0,100	0,100	0,687	0,444
14				0,200	0,651	0,298
15				0,400	0,548	0,223
16			0,200	0,100	0,673	0,429
17				0,200	0,644	0,294
18				0,400	0,582	0,424
19	0,200	0,050	0,100	0,100	1,183	1,288
20				0,200	1,088	1,002
21				0,400	0,875	0,679
22			0,200	0,100	1,216	1,076
23				0,200	1,009	1,007
24				0,400	0,747	0,822
25		0,100	0,100	0,100	0,832	0,586
26				0,200	0,797	0,472
27				0,400	0,710	0,426
28			0,200	0,100	0,903	0,637
29				0,200	0,821	0,594
30				0,400	0,733	0,610
31		0,200	0,100	0,100	0,695	0,559
32				0,200	0,645	0,385
33				0,400	0,502	0,160
34			0,200	0,100	0,679	0,450
35				0,200	0,616	0,429
36				0,400	0,590	0,374

Note : BRAM : biais relatif absolu moyen; BRQM : biais relatif quadratique moyen.

4.5 Deux sources plutôt qu'une source

En présence de données provenant de deux sources, on peut faire des inférences en utilisant les données provenant d'une seule source. Nous avons d'abord généré les données à l'aide des cas 1 à 4 de sorte à permettre l'analyse. Pour chaque cas et chaque spécification, nous avons calculé le BRAM et le BRQM à l'aide des modèles M1a, M1b, M11a, M11b et M12. Nous avons également utilisé les modèles MSA et MBR, des modèles à source unique (section 3.4) fondés uniquement sur les données du BRFSS et du programme SAHIE. Comme à la section 3.2, nous résumons les résultats pour chaque cas en utilisant des ratios d'écart. Dans ce cas-ci, il s'agit de A(M12)/A(M1a), de A(MSA)/A(M1a) et de A(MBR)/A(M1a), où A désigne soit le BRAM soit le BRQM. Les tableaux 4.6 à 4.9 résument les résultats pour les cas 1 à 4. Pour chaque ratio, nous présentons la valeur minimale, le premier quartile, la médiane, la moyenne, le troisième quartile et la valeur maximale dans l'ensemble des spécifications du cas. À l'exception de quelques valeurs minimales, tous les ratios d'écart sont supérieurs à 1, ce qui signifie que l'utilisation des modèles M12,

MSA et MBR engendre des pertes comparativement à l'utilisation du modèle M1a. Pour le BRAM, les valeurs médianes du ratio BRAM(MBR)/BRAM(M1a) sont très grandes, c'est-à-dire (2,530; 2,255; 1,782 et 2,703) pour les cas 1 à 4. Cela démontre l'avantage que présente l'utilisation d'un modèle GL, comme le modèle M1a dans le cas présent. Les valeurs médianes pour le ratio BRQM(MBR)/BRQM(M1a), (4,593; 5,319; 3,636; 6,021), sont encore plus élevées et fournissent des données probantes supplémentaires appuyant l'utilisation du modèle M1a.

De même, les quatre distributions du ratio A(MSA)/A(M1a) ont des valeurs élevées, quoique en général, elles sont moins élevées que celles du ratio A(MBR)/A(M1a). Par exemple, pour le BRAM, les médianes des ratios {BRAM(MSA)/BRAM(M1a), BRAM(MBR)/BRAM(M1a)} sont de {(1,875; 2,530), (1,255; 2,255), (1,724; 1,782), (2,879; 2,703)}.

Nous avons ensuite peaufiné les spécifications dans les cas 1 à 4 en supposant des distributions différentes pour les θ. Cela est plus réaliste du fait qu'en général, les deux sources devraient avoir des caractéristiques différentes. Nous avons également simplifié les spécifications en supposant qu'il n'y avait pas de valeurs aberrantes au niveau des comtés. Tout au long du processus, nous générons des données en utilisant le BRFSS comme source 1 et le programme SAHIE comme source 2. Pour le cas 5, le modèle utilisé pour générer les données est

$$y_{ij} \overset{\text{ind}}{\sim} N(\theta_{ij}, V_{ij})$$

$$\theta_{i1} \overset{\text{ind}}{\sim} N(\mu_i, \tau_1^2)$$

$$\theta_{i2} \overset{\text{ind}}{\sim} N(\mu_i, \delta\tau_2^2),$$

où $\delta \in \{0,1\}$ et $P(\delta = 1) = p$. Enfin, $\mu_i \overset{\text{iid}}{\sim} N(0{,}25,\ \tau^2)$.

Tableau 4.6
Statistiques sommaires pour le BRAM et le BRQM pour le cas 1 : une source plutôt que deux

	BRAM		
	M12/M1a	MSA/M1a	MBR/M1a
Min.	1,141	1,089	1,458
1er quartile	1,494	1,516	2,019
Médiane	1,722	1,875	2,530
Moyenne	1,996	2,044	2,659
3e quartile	2,219	2,316	3,242
Max.	4,737	3,950	4,969
	BRQM		
	M12/M1a	MSA/M1a	MBR/M1a
Min.	1,176	1,319	1,826
1er quartile	1,786	2,153	2,790
Médiane	2,235	3,934	4,593
Moyenne	3,329	5,741	6,526
3e quartile	3,553	7,445	8,436
Max.	18,581	32,331	30,281

Note : BRAM : biais relatif absolu moyen; BRFSS : Behavioral Risk Factor Surveillance System; BRQM : biais relatif quadratique moyen; MBR : Model BRFSS; MSA : Model SAHIE; SAHIE : Small Area Health Insurance Estimates.

Tableau 4.7
Statistiques sommaires pour le BRAM et le BRQM pour le cas 2 : une source plutôt que deux

	BRAM		
	M12/M1a	MSA/M1a	MBR/M1a
Min.	1,012	0,965	1,871
1er quartile	1,123	1,107	2,044
Médiane	1,345	1,255	2,255
Moyenne	1,482	1,364	2,235
3e quartile	1,947	1,620	2,419
Max.	2,113	2,054	2,608
	BRQM		
	M12/M1a	MSA/M1a	MBR/M1a
1er quartile	1,247	1,639	3,568
Médiane	1,650	2,520	4,716
Moyenne	2,279	3,546	5,319
3e quartile	3,147	5,855	6,963
Max.	5,246	8,317	10,078

Note : BRAM : biais relatif absolu moyen; BRFSS : Behavioral Risk Factor Surveillance System; BRQM : biais relatif quadratique moyen; MBR : Model BRFSS; MSA : Model SAHIE; SAHIE : Small Area Health Insurance Estimates.

Tableau 4.8
Statistiques sommaires pour le BRAM et le BRQM pour le cas 3 : une source plutôt que deux

	BRAM		
	M12/M1a	MSA/M1a	MBR/M1a
Min.	0,943	1,370	1,445
1er quartile	1,155	1,551	1,594
Médiane	1,254	1,724	1,782
Moyenne	1,304	1,813	1,851
3e quartile	1,510	1,964	1,985
Max.	1,733	2,643	2,653
	BRQM		
	M12/M1a	MSA/M1a	MBR/M1a
Min.	0,935	1,787	2,139
1er quartile	1,081	2,198	2,395
Médiane	1,881	3,509	3,636
Moyenne	2,096	3,846	4,127
3e quartile	2,321	4,307	4,556
Max.	4,867	8,459	12,665

Note : BRAM : biais relatif absolu moyen; BRFSS : Behavioral Risk Factor Surveillance System; BRQM : biais relatif quadratique moyen; MBR : Model BRFSS; MSA : Model SAHIE; SAHIE : Small Area Health Insurance Estimates.

Tableau 4.9
Statistiques sommaires pour le BRAM et le BRQM pour le cas 4 : une source plutôt que deux

	BRAM		
	M12/M1a	MSA/M1a	MBR/M1a
Min.	0,726	1,681	1,590
1er quartile	1,019	2,400	2,317
Médiane	1,260	2,879	2,703
Moyenne	1,269	3,150	2,904
3e quartile	1,382	3,564	3,316
Max.	2,116	6,413	5,740
	BRQM		
	M12/M1a	MSA/M1a	MBR/M1a
Min.	0,530	1,326	1,154
1er quartile	1,002	3,203	3,393
Médiane	1,407	6,507	6,021
Moyenne	1,817	11,151	10,253
3e quartile	2,040	11,854	12,164
Max.	5,935	56,195	46,653

Note : BRAM : biais relatif absolu moyen; BRFSS : Behavioral Risk Factor Surveillance System; BRQM : biais relatif quadratique moyen; MBR : Model BRFSS; MSA : Model SAHIE; SAHIE : Small Area Health Insurance Estimates.

Nous avons utilisé 18 spécifications comprenant toutes les combinaisons de $\tau =$ 0,05, de $\tau_1 =$ (0,005; 0,010), de $p =$ (0,10; 0,20; 0,40) et de $\tau_2 =$ (0,05; 0,10; 0,20); voir le tableau 4.10. Les résultats résumés dans le tableau 4.11 ne montrent presque aucune spécification pour laquelle l'utilisation des données du BRFSS comme source unique est préférable à l'utilisation du modèle M1a. Pour le BRAM et le BRQM, les valeurs médianes de A(MBR)/A(M1a) (pour l'ensemble des 18 spécifications) sont assez élevées, à savoir 1,663 et 2,070. (Il convient de prendre note qu'une autre étude, reposant sur des valeurs plus élevées de τ, a permis de produire des résultats très similaires.) L'analyse des 18 ratios A(MBR)/A(M1a) individuels nous permet de constater que les valeurs des ratios diminuent à mesure que p et τ_2 augmentent, comme prévu. La valeur de τ_1 n'a qu'un effet minime. Voir les tableaux 4.10, S17 et S18.

Tableau 4.10
Spécifications de l'étude par simulation pour le cas 5 : $\mu_i \sim N(0{,}25, \tau^2), \theta_{i1} \sim N(\mu_i, \tau_1^2), \theta_{i2} \sim N(\mu_i, \delta\tau_2^2)$, $\delta \sim$ Ber(p)

	τ	τ_1	p	τ_2
1	0,05	0,005	0,1	0,05
2				0,1
3				0,2
4			0,2	0,05
5				0,1
6				0,2
7			0,4	0,05
8				0,1
9				0,2
10		0,01	0,1	0,05
11				0,1
12				0,2
13			0,2	0,05
14				0,1
15				0,2
16			0,4	0,05
17				0,1
18				0,2

Tableau 4.11
Statistiques sommaires pour le BRAM et le BRQM pour le cas 5

	BRAM		
	M12/M1a	MSA/M1a	MBR/M1a
Min.	1,000	1,080	1,088
1er quartile	1,065	1,346	1,443
Médiane	1,162	1,485	1,663
Moyenne	1,285	1,786	1,694
3e quartile	1,326	2,002	2,025
Max.	2,030	3,435	2,342
	BRQM		
	M12/M1a	MSA/M1a	MBR/M1a
Min.	1,000	1,468	0,962
1er quartile	1,133	2,233	1,712
Médiane	1,348	4,369	2,070
Moyenne	1,681	7,033	2,377
3e quartile	1,851	11,494	3,112
Max.	3,899	20,498	4,465

Note : BRAM : biais relatif absolu moyen; BRFSS : Behavioral Risk Factor Surveillance System; BRQM : biais relatif quadratique moyen; MBR : Model BRFSS; MSA : Model SAHIE; SAHIE : Small Area Health Insurance Estimates.

Pour le cas 6, le modèle utilisé pour générer les données est le même que celui du cas 5 sauf que $p = 1$. Puisque le cas 6 ne suppose aucune aberration, un modèle GL n'est pas nécessaire. Ainsi, il n'est pas surprenant qu'il existe des spécifications dans lesquelles l'utilisation du modèle MBR est préférable. Par exemple, les médianes de A(MBR)/A(M1a) pour les 12 spécifications sont de 0,854 et de 0,731 pour le BRAM et le BRQM (voir le tableau 4.13). Les 12 spécifications sont présentées dans le tableau 4.12. Comme le montrent les tableaux S19 et S20, la valeur de τ_1 a peu d'effet. Cependant, à mesure que τ_2 augmente, A(MBR)/A(M1a) diminue tandis que A(MSA)/A(M1a) augmente. Ainsi, pour des valeurs élevées de τ_2, il est plus avantageux d'utiliser le modèle MBR plutôt que le modèle GL, M1a. Mais comme on l'a vu dans le cas 5, il semble peu probable que l'utilisation d'une seule source de données soit préférable quand : a) il existe des observations aberrantes significatives et b) la source unique est celle ayant un faible biais et une grande variance d'échantillonnage.

Tableau 4.12
Spécifications de l'étude par simulation pour le cas 6 : $\mu_i \sim N(0{,}25,\ \tau^2), \theta_{ij} \sim N(\mu_i, \tau_j^2)$

	τ	τ_1	τ_2
1	0,05	0,005	0,01
2			0,02
3			0,05
4			0,1
5			0,2
6			0,4
7		0,01	0,01
8			0,02
9			0,05
10			0,1
11			0,2
12			0,4

Tableau 4.13
Statistiques sommaires pour le BRAM et le BRQM pour le cas 6

	BRAM		
	M12/M1a	MSA/M1a	MBR/M1a
Min.	0,954	1,065	0,546
1er quartile	0,967	1,208	0,634
Médiane	1,000	2,293	0,854
Moyenne	1,027	3,283	1,097
3e quartile	1,099	4,596	1,495
Max.	1,170	8,230	2,168
	BRQM		
	M12/M1a	MSA/M1a	MBR/M1a
Min.	0,882	1,083	0,293
1er quartile	0,911	1,452	0,380
Médiane	0,975	5,067	0,731
Moyenne	1,030	15,203	1,547
3e quartile	1,182	18,322	2,420
Max.	1,289	60,459	4,745

Note : BRAM : biais relatif absolu moyen; BRFSS : Behavioral Risk Factor Surveillance System; BRQM : biais relatif quadratique moyen; MBR : Model BRFSS; MSA : Model SAHIE; SAHIE : Small Area Health Insurance Estimates.

5. Discussion et résumé

Dans le contexte de l'estimation sur petits domaines à partir d'une seule enquête, Datta et Lahiri (1995) ont élaboré une méthode hiérarchique bayésienne reposant sur une classe générale de mélange d'échelles de distributions normales (voir le modèle 1 à la page 314). Pour certains choix de distribution *a priori*, leur méthodologie hiérarchique bayésienne assure la robustesse par rapport aux valeurs aberrantes (voir leur théorème 5). Dans le contexte d'une enquête unique, la distribution *a priori* de type *horseshoe* est comprise dans la classe de distributions *a priori* proposée par Datta et Lahiri (1995), ce qui offre une protection contre les valeurs aberrantes au sens de leur théorème 5. Une extension au cas comptant plusieurs enquêtes, comme celui du présent document, serait utile.

Un examinateur a posé des questions au sujet des comparaisons formelles des modèles. Pour le modèle de (1.1) et de (1.2) [ayant η comme régression linéaire], Tang et coll. (2018) ont envisagé de choisir un modèle unique au moyen du critère d'information de déviance (DIC) et ont constaté dans leur étude par simulation que le modèle sélectionné produit des mesures d'écart proches des plus petites parmi tous les modèles candidats et un meilleur taux de couverture que le modèle Datta-Mandal. Cependant, le DIC a été la cible de beaucoup de critiques, et il faudrait être prudent lors de son utilisation pour des modèles comme ceux de (1.1) et de (1.2) ainsi que pour les modèles plus complexes que nous avons utilisés. Plummer (2008) a mené une enquête approfondie sur le DIC et a détaillé plusieurs problèmes importants. Carlin et Louis (2009) ont cerné des problèmes supplémentaires concernant le DIC et ont également décrit (à la section 4.6.2) une autre approche générale, à savoir la sélection de modèles prédictifs.

Il est intéressant d'étudier l'effet d'avoir un plus grand nombre de sources d'information. Nous avons réalisé une étude préliminaire dans laquelle étaient utilisés les quatre cas constituant les éléments de base de notre étude par simulation. À titre d'illustration, considérons les 30 spécifications du cas 1 présentées dans le tableau 4.2. Nous avons généré des données comme à la section 3 pour $J = 2$ et $J = 4$ sources. Les V_{ij} ont été sélectionnées à partir des $\hat{V}_{ij}$ pour faire partie d'un échantillon bootstrap. Comme on pouvait s'y attendre, il existe des gains substantiels mesurés à l'aide des ratios $(\mathrm{A}(\mathrm{Mx}(J = 4))/\mathrm{A}(\mathrm{Mx}(J = 2))$, où A désigne le BRAM ou le BRQM. Mais ces gains varient considérablement selon les probabilités et les variances choisies. Par exemple, pour le modèle M1a, la valeur la plus faible et la valeur la plus grande du ratio sont de 0,142 et de 0,767, respectivement. Considérons le modèle M1a et le BRAM : à mesure que les deux τ_{11} augmentent, les gains liés à l'utilisation des $J = 4$ sources diminuent. À mesure que les deux p_2 augmentent pour passer de 0,10 à 0,20, puis à 0,40, le ratio diminue, puis augmente par la suite. D'autres études sont nécessaires pour obtenir des résultats plus concluants.

Faire des inférences sur des variances d'échantillonnage, les V_{ij}, est un problème difficile. La méthode la plus courante consiste à supposer un modèle pour les $\hat{V}_{ij}$, les estimations des V_{ij}. Cependant, on ne peut pas vérifier le modèle au moyen d'un seul échantillon où $\{\hat{V}_{ij}: j = 1,\ldots, J;\ i = 1,\ldots, I\}$. La méthode proposée dans la littérature est passée en revue par Cahoy et Sedransk (2023). Dans le contexte du modèle GL, mais

en ayant un seul ensemble de données, Tang et coll. (2018) supposent que V_i est fixe pour éviter les problèmes liés à l'identifiabilité. Notre situation est plus complexe, car il s'agit de petits domaines et de plusieurs sources de données.

Dans le cadre de futures recherches, nous prévoyons modéliser la $\{\hat{V}_{ij}\colon\ j=1,\ldots,J;\ i=1,\ldots,I\}$ en utilisant les ensembles de données des années précédentes. Une extension supplémentaire consiste à inclure des covariables dans nos modèles.

L'objectif du présent document est d'exposer une méthodologie pour améliorer l'inférence des paramètres pour petits domaines au moyen de données provenant de plusieurs sources. Bien qu'il existe une abondante littérature sur les modèles GL, il semble qu'il n'y ait qu'un seul autre article, celui de Smith et Griffin (2023), dans lequel on a recours à un modèle en plusieurs étapes. De plus, l'application de l'étude de Smith et Griffin (2023) est complètement différente de la nôtre. Par conséquent, il était nécessaire de compléter notre analyse des données des comtés de Floride au moyen d'une vaste étude par simulation pour montrer les propriétés de cette méthodologie.

Nous introduisons un ensemble de modèles GL et expliquons la façon dont ces modèles combinent l'information provenant de l'ensemble des comtés et des sources. L'analyse des données du BRFSS et du programme SAHIE au moyen de ces modèles nous donne l'occasion de montrer la façon dont ces modèles traitent les observations aberrantes. Cela comprend également une comparaison des inférences fondées sur toutes les données avec des inférences fondées sur une seule source de données. Dans la vaste étude par simulation, les données sont sélectionnées à partir du modèle présentant des valeurs aberrantes et du modèle de mélange. Ces résultats montrent que la méthode proposée peut être utilisée pour améliorer l'inférence des paramètres pour petits domaines au moyen de plusieurs sources de données.

Remerciements

Les auteurs remercient l'ACCESS pour le soutien informatique, ainsi que plusieurs évaluateurs qui ont posé des questions intéressantes utiles pour de futures recherches.

Annexe

A. Les ϕ_i

Figure A.1 Les ϕ_i pour les modèles M1a (haut) et M1b (bas)

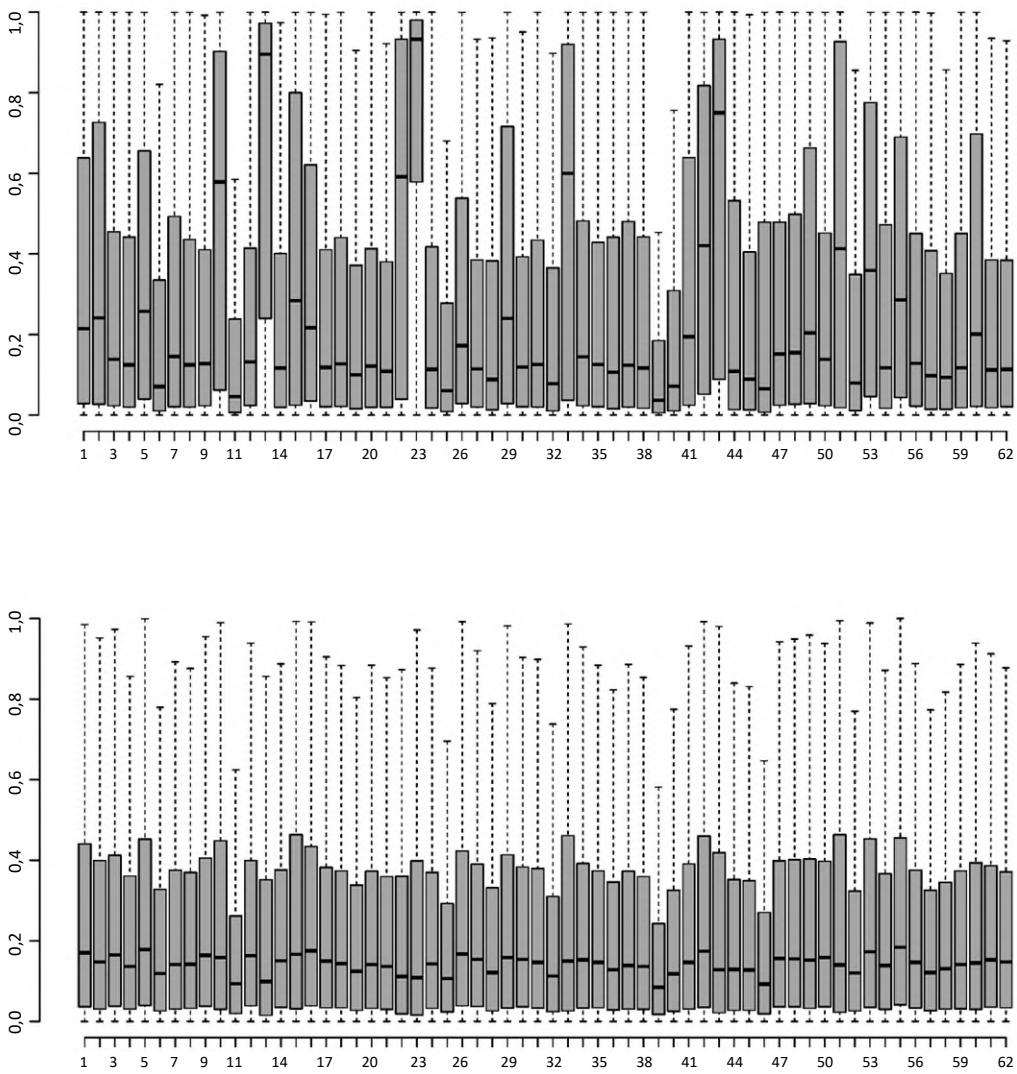


Figure A.2 Les ϕ_i pour les modèles M11a (haut) et M11b (bas)

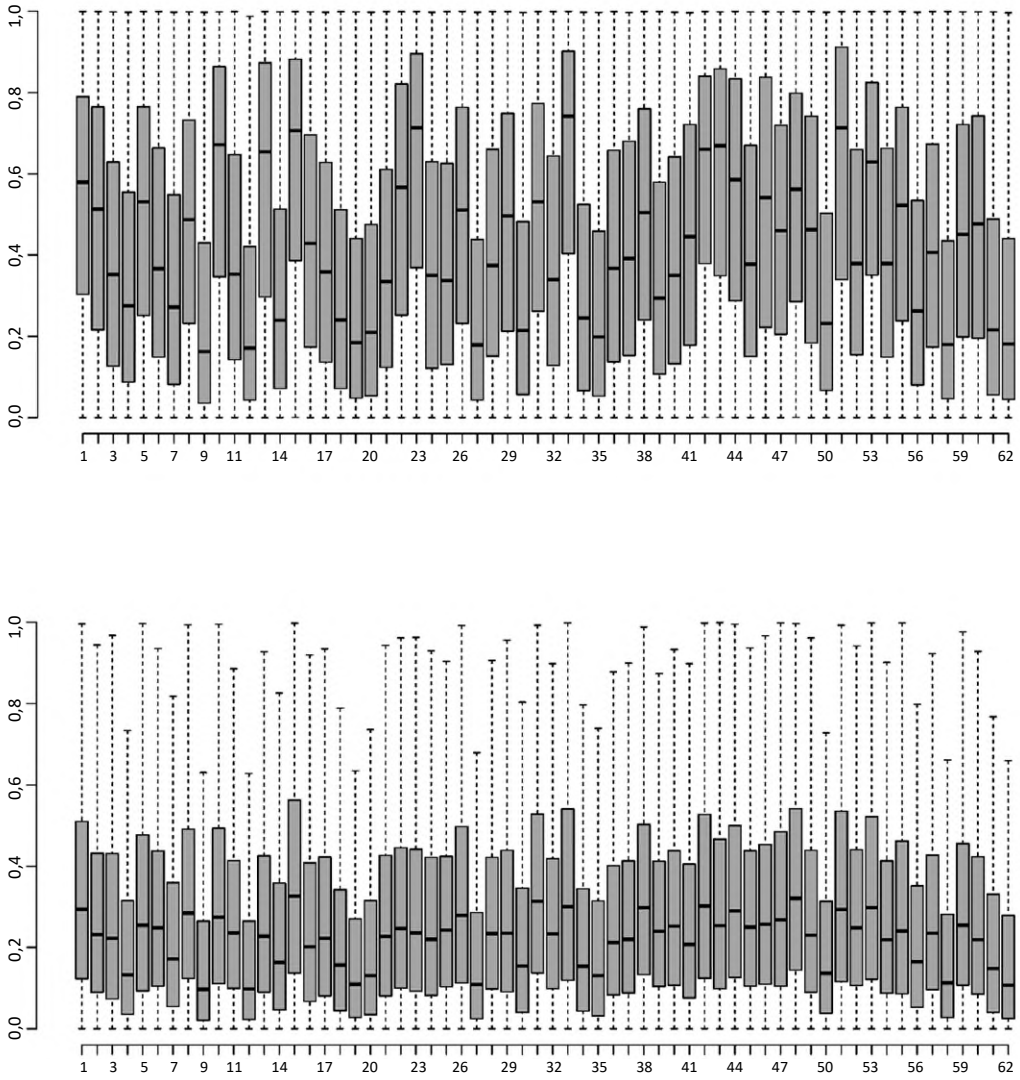
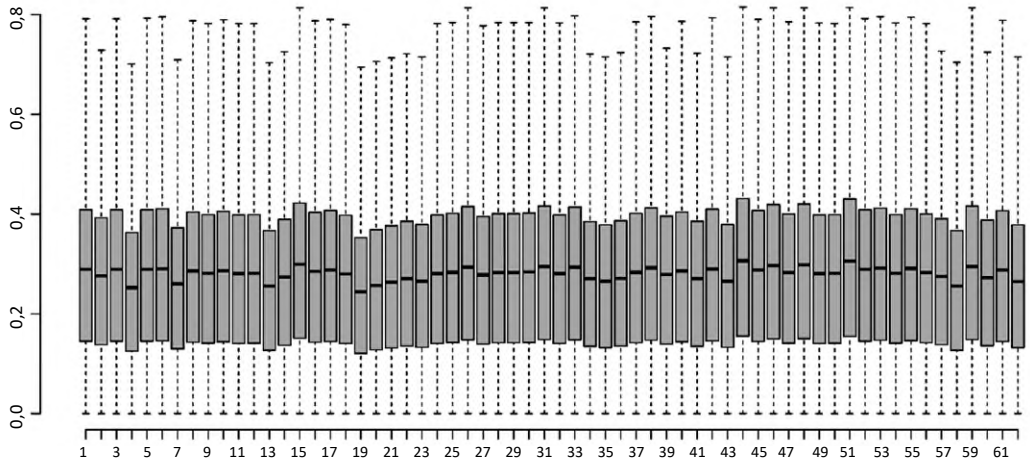


Figure A.3 Les ϕ_i pour le modèle M12



B. Distributions conditionnelles complètes pour les modèles M11a et M11b

Nous présentons ci-dessous les distributions conditionnelles complètes pour les effets d'interaction, θ_{ij}, les valeurs attendues pour petits domaines, μ_i, et les variances locales et globales.

Effets d'interaction : $\theta_{ij} \mid \text{sinon} \sim N(\bar{\theta}_{ij}, V_{\theta_{ij}})$ où

$$\bar{\theta}_{ij} = \frac{Y_{ij}/V_{ij} + \mu_i/\lambda_{ij}^2\lambda_i^2\tau_1^2}{1/V_{ij} + 1/\lambda_{ij}^2\lambda_i^2\tau_1^2} \text{ et } V_{\theta_{ij}} = \frac{1}{1/V_{ij} + 1/\lambda_{ij}^2\lambda_i^2\tau_1^2}.$$

Moyennes des valeurs attendues pour petits domaines : $\mu_i \mid \text{sinon} \sim N(\bar{\mu}_i, V_{\mu_i})$ où

$$\bar{\mu}_i, = \frac{\bar{\theta}_i c_i + \eta d_i}{c_i + d_i}, \qquad V_{\mu_i} = \frac{1}{c_i + d_i},$$

$a_{ij} = \lambda_{ij}^2\lambda_i^2\tau_1^2, c_i = \sum_j (1/a_{ij}), d_i = 1/(\lambda_i^2\tau_2^2)$ et $\bar{\theta}_i = \frac{\sum_j (\theta_{ij}/a_{ij})}{\sum_j (1/a_{ij})}$.

Variances locales : Pour faire suite aux travaux de Makalic et Schmidt (2016), nous utilisons la représentation de mélanges d'échelles gamma inverse (IG [forme, taux]) suivante : Si $\kappa \mid \xi \sim \text{IG}(1/2, 1/\xi)$ et $\xi \sim \text{IG}(1/2, 1)$, alors $\kappa \sim \mathcal{C}^+(0,1)$ et $\nu = \kappa^2$ (voir la section 2) a la distribution de type *horseshoe*. Il convient de prendre note que la densité IG (forme, taux) est proportionnelle à $x^{(\text{-1-forme})}\ e^{(\text{-taux}/x)}$.

En utilisant la distribution *a priori* de type *horseshoe* pour les variances locales,

$$\lambda_{ij}^2 \mid \text{sinon} \sim \text{IG}\left(1, \frac{(\theta_{ij} - \mu_i)^2}{2\lambda_i^2\tau_1^2} + \frac{1}{\xi}\right); \qquad \xi \mid \lambda_{ij}^2 \sim \text{IG}\left(1, 1 + \frac{1}{\lambda_{ij}^2}\right).$$

$$\lambda_i^2 \mid \text{sinon} \sim \text{IG}\left(\frac{(J+4)}{2} - 1, \sum_j \frac{(\theta_{ij} - \mu_i)^2}{2\lambda_{ij}^2\tau_1^2} + \frac{(\mu_i - \eta)^2}{2\tau_2^2} + \frac{1}{\xi}\right); \ \xi \mid \lambda_i^2 \sim \text{IG}\left(1, 1 + \frac{1}{\lambda_i^2}\right).$$

En utilisant la distribution *a priori* de type *LASSO* $(\propto \exp(-\kappa^2))$ pour les variances locales et la distribution GIG ou gaussienne inverse généralisée (forme, $\chi, \psi) \propto x^{(\text{forme-1})} e^{-\frac{1}{2}(\chi/x + \psi x)}$,

$$\lambda_{ij}^2 \mid \text{ sinon} \sim \text{GIG}\left(1/2, \frac{(\theta_{ij} - \mu_i)^2}{\lambda_i^2\tau_1^2}, 2\right)$$

et

$$\lambda_i^2 \mid \text{ sinon} \sim \text{GIG}\left(\frac{(-J+1)}{2}, \sum_j \frac{(\theta_{ij} - \mu_i)^2}{\lambda_{ij}^2\tau_1^2} + \frac{(\mu_i - \eta)^2}{\tau_2^2}, 2\right).$$

Variances globales : En utilisant la distribution *a priori* de type *horseshoe* pour les variances globales, nous avons la représentation de mélanges d'échelles suivante :

$$\tau_1^2 \mid \text{sinon} \sim \text{IG}\left(\frac{(IJ+3)}{2} - 1, \frac{(\theta_{ij} - \mu_i)^2}{2\lambda_{ij}^2\lambda_i^2} + \frac{1}{\xi}\right); \qquad \xi \mid \tau_1^2 \sim \text{IG}\left(1, 1 + \frac{1}{\tau_1^2}\right).$$

$$\tau_2^2 \mid \text{sinon} \sim \text{IG}\left(\frac{(I+3)}{2} - 1, \sum_j \frac{(\theta_{ij} - \mu_i)^2}{2\lambda_i^2} + \frac{1}{\xi}\right); \qquad \xi \mid \tau_2^2 \sim \text{IG}\left(1, 1 + \frac{1}{\tau_2^2}\right).$$

Bibliographie

Bhadra, A., Datta, J., Polson, N. et Willard, B. (2019). Lasso meets horseshoe: A survey. *Statistical Science*, 34(3), 405-427. https://www.jstor.org/stable/26874188.

Cahoy, D. et Sedransk, J. (2023). Combinaison de données provenant d'enquêtes et de sources connexes. *Techniques d'enquête*, 49(1), 75-96. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2023001/article/00003-fra.pdf.

Carlin, B. et Louis, T. (2009). *Bayesian Methods for Data Analysis* (3rd ed.). Chapman and Hall/CRC.

Datta, G. et Lahiri, P. (1995). Robust hierarchical Bayes estimation of small area characteristics in the presence of covariates and outliers. *Journal of Multivariate Analysis,* 54(2), 310-328.

Datta, G. et Mandal, A. (2015). Small area estimation with uncertain random effects. *Journal of the American Statistical Association*, 110, 1735-1744.

Evans, R. et Sedransk, J. (2001). Combining data from experiments that may be similar. *Biometrika*, 88(3), 643-656.

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

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. et Rubin, D. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018.

Makalic, E. et Schmidt, D. (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1), 179-182, Jan. 2016, DOI: 10.1109/LSP.2015.2503725.

Malec, D. et Sedransk, J. (1992). Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain. *Biometrika*, 79(3), 593-601.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparisons. *Biostatistics*, 9(3), 523-539.

Polson, N. et Scott, J. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7, 887-902.

Smith, A. et Griffin, J. (2023). Shrinkage priors for high-dimensional demand estimation. *Quant Mark Econ*, 21, 95-146, DOI: https://doi.org/10.1007/s11129-022-09260-7.

Tang, X. et Ghosh, M. (2023). Global-local priors for spatial small area estimates. *Calcutta Statistical Association Bulletin*, 75(2), 141-154.

Tang, X., Ghosh, M., Ha, N. et Sedransk, J. (2018). Modeling random effects using global-local shrinkage priors in small area estimation. *Journal of the American Statistical Association*, 113(524), 1476-1489, DOI: 10.1080/01621459.2017.1419135.