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

Techniques d’enquête

Estimation hiérarchique bayésienne approximative pour petits domaines à l’aide de la famille exponentielle naturelle avec fonction de variance quadratique et poststratification

par Soumojit Das et Partha Lahiri

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. 433-460
Statistique Canada, nº 12-001-X au catalogue

Estimation hiérarchique bayésienne approximative pour petits domaines à l'aide de la famille exponentielle naturelle avec fonction de variance quadratique et poststratification

Soumojit Das et Partha Lahiri[1]

Résumé

On propose une approche hiérarchique bayésienne approximative pour laquelle on utilise la famille exponentielle naturelle avec fonction de variance quadratique en combinant de l'information tirée de multiples sources, afin d'améliorer les estimations de moyennes de population finie pour de petits domaines dans le cadre d'enquêtes traditionnelles. Contrairement aux autres approches bayésiennes relatives à l'échantillonnage d'une population finie, on ne suppose pas de modèle pour toutes les unités de la population finie et l'on n'a pas besoin de procéder au couplage d'unités échantillonnées à la base de population finie. On suppose un modèle uniquement pour les unités d'une population finie pour lesquelles on observe la variable dépendante, car, dans le cas de ces unités, le modèle supposé peut être vérifié à l'aide des outils statistiques existants. On ne suggère pas de modèle élaboré selon les moyennes réelles des unités non observées. On suppose plutôt que les moyennes de population des cellules ayant la même combinaison de niveaux de facteur sont identiques pour tous les petits domaines et que la moyenne de population d'une cellule est identique à la moyenne des unités observées dans cette cellule. On met en application la méthodologie que l'on propose pour une enquête réelle couplant des renseignements tirés de multiples sources de données disparates. On fournit également des moyens pratiques de sélectionner un modèle pouvant s'appliquer à un ensemble élargi de modèles dans le même contexte, mais pour un éventail diversifié de problèmes scientifiques.

Mots-clés : Données administratives; échantillonnage d'une population finie; échantillonnage informatif; enquêtes multiples; enquêtes non probabilistes; estimation synthétique; modélisation multiniveaux.

1. Introduction

Ericson (1969) a jeté les bases d'une approche bayésienne subjective pour l'échantillonnage d'une population finie. Dans le cadre de cette approche, on peut considérer la matrice complète des valeurs des caractéristiques d'intérêt pour toutes les unités de la population finie comme la matrice des paramètres de la population finie. En pratique, on évalue une fonction (par exemple la moyenne d'une population finie ou la proportion d'une caractéristique d'intérêt) ou un vecteur de fonctions (par exemple des moyennes de population finie ou des proportions de plusieurs caractéristiques) de cette matrice de paramètres de la population finie à des fins d'inférence. En appliquant une distribution *a priori* subjective à la matrice de paramètres de la population finie, on peut déduire des inférences en ce qui concerne le ou les paramètres d'intérêt de la population finie à partir de la distribution prédictive *a posteriori* des unités non observées de la population finie compte tenu de l'échantillon observé.

Reposant sur cette idée de base d'Ericson (1969), de nombreux articles ont été écrits sur l'estimation des paramètres d'une population finie. La technique élaborée peut servir à résoudre des problèmes associés à l'estimation pour petits domaines, aux enquêtes répétées et à d'autres applications importantes. Les articles peuvent être classés de façon générale dans la catégorie de l'approche bayésienne empirique (par exemple

1. Soumojit Das est chercheur au niveau postdoctoral à l'Institute for Social Research – University of Michigan, Ann Arbor. Courriel : soumojit@umd.edu; Partha Lahiri est professeur et directeur du Joint Program in Survey Methodology, en plus d'être professeur au Department of Mathematics de l'University of Maryland, College Park.

Ghosh et Meeden, 1986; Ghosh et Lahiri, 1987; Ghosh, Lahiri et Tiwari, 1989; Nandram et Sedransk, 1993; Arora, Lahiri et Mukherjee, 1997, et Jiang et Nguyen, 2012) et de l'approche hiérarchique bayésienne (par exemple Datta et Ghosh, 1991; Ghosh et Lahiri, 1992; Datta et Ghosh, 1995; Ghosh et Meeden, 1997; Malec, Sedransk, Moriarity et LeClere, 1997; Little, 2004; Chen, Elliott et Little, 2012; Ghosh, 2009; Liu et Lahiri, 2017; Nandram, Chen, Fu et Manandhar, 2018, et Ha et Sedransk, 2019). Dans le cadre d'une approche bayésienne empirique, les hyperparamètres sont estimés à l'aide d'une méthode classique (par exemple le maximum de vraisemblance). En revanche, dans le cadre de l'approche hiérarchique bayésienne, les distributions *a priori*, qui sont habituellement non informatives ou peu informatives, sont appliquées aux hyperparamètres.

En raison de l'accessibilité accrue des données administratives et des mégadonnées, ainsi que des avancées technologiques, les chercheurs ont désormais de nouvelles occasions de résoudre un vaste éventail de problèmes impossibles à résoudre en utilisant une seule source de données. Cependant, il arrive souvent que ces bases de données ne soient pas structurées et qu'elles soient accessibles dans un format disparate, ce qui rend les inférences à l'aide de couplages de données plutôt difficiles. Il existe donc un besoin croissant d'élaborer des outils statistiques novateurs de couplage de données afin de coupler ces multiples ensembles de données complexes. Il peut être efficace et fiable d'utiliser une seule enquête primaire pour répondre à des questions scientifiques sur l'ensemble de la population ou de vastes régions géographiques ou sous-populations. Cependant, le fait de l'utiliser pour des domaines plus petits ou de petites régions peut souvent mener à une estimation peu fiable et à des mesures non réalistes de l'incertitude. À l'aide d'un modèle statistique approprié servant à coupler des renseignements tirés de multiples sources de données, on peut souvent obtenir des estimations fiables pour les petits domaines. On trouve un examen efficace des différentes approches en matière d'estimation pour petits domaines dans, entre autres, Jiang (2007), Datta (2009), Pfeffermann (2013), Rao et Molina (2015) et Ghosh (2020).

Scott (1977), Pfeffermann et Sverchkov (1999), Bonnéry, Breidt et Coquet (2012) et d'autres ont abordé le concept de l'échantillonnage informatif dans le cadre duquel la distribution de l'échantillon pourrait différer grandement de celle supposée pour la population finie, même après avoir conditionné sur les variables auxiliaires connexes. Dans la majorité des articles cités dans les paragraphes précédents, on suppose un échantillonnage non informatif. On suppose donc que la distribution de l'échantillon est identique à la distribution supposée de la population finie. Cela pourrait être une hypothèse solide dans de nombreuses applications. L'approche en matière d'échantillonnage informatif dans Pfeffermann et Sverchkov (1999) et Pfeffermann et Sverchkov (2007) représente une solution possible. Cependant, cette approche exige une modélisation supplémentaire des poids de sondage. Rubin (1983) a suggéré de modéliser la variable dépendante d'intérêt de la population finie en fonction des probabilités d'inclusion (ou, de manière équivalente, des poids de base) pour faire des inférences au sujet des paramètres d'intérêt de la population finie. Cependant, en pratique, les probabilités d'inclusion pour toutes les unités d'une population finie ou les renseignements détaillés sur le plan de sondage peuvent ne pas être accessibles. En outre, les données d'enquête peuvent renfermer des renseignements qui portent uniquement sur les poids définitifs des répondants qui intègrent des ajustements pour la non-réponse et la calibration. Verret, Rao et Hidiroglou (2015) ont suggéré une autre approche dans le cadre de laquelle les probabilités d'inclusion sont utilisées pour augmenter le modèle

d'échantillon. Contrairement à Rubin (1983), leur approche nécessite des poids uniquement pour l'échantillon lorsque l'on estime les moyennes de domaine. Pfeffermann et Sverchkov (2007) et Verret et coll. (2015) ont envisagé la méthode du meilleur prédicteur linéaire sans biais empirique pour les moyennes de petits domaines.

Dans le présent article, on suggère une approche hiérarchique bayésienne approximative pour estimer les moyennes de population finie pour de petits domaines. La technique que l'on suggère combine des renseignements tirés de multiples sources de données disparates, comme des enquêtes probabilistes, des enquêtes non probabilistes, des données administratives, des enregistrements de recensement, des mégadonnées des médias sociaux ou d'autres sources de renseignements pertinents accessibles. On suggère de réduire le degré de valeur informative de l'échantillonnage en intégrant d'importantes variables auxiliaires qui ont potentiellement une incidence sur la variable dépendante d'intérêt. Suivant la méthode de Verret et coll. (2015), on a également ajouté les poids de sondage à notre modèle. Notre approche diffère intrinsèquement de l'approche bayésienne existante pour les petits domaines dans le cadre de l'échantillonnage de populations finies, qui suppose généralement un modèle hiérarchique pour toutes les unités de la population finie. On suppose un tel modèle aussi élaboré uniquement pour les unités de la population finie pour lesquelles la variable dépendante a été observée. La raison est de nature intuitive; dans le cas de ces unités, on peut vérifier le modèle à l'aide d'outils statistiques existants.

Pour formuler des hypothèses de modélisation raisonnables, on propose de créer plusieurs cellules, souvent appelées cellules de poststratification, pour chacun des petits domaines en utilisant des facteurs susceptibles d'influencer notre variable dépendante d'intérêt. On choisit judicieusement le nombre de cellules, afin que l'on puisse obtenir des tailles de population de cellule à partir de sources fiables (par exemple des données de projections démographiques). On s'attend à ce que cette stratégie procure un certain degré d'homogénéité dans une cellule donnée, ainsi que parmi les cellules de différents petits domaines qui sont établis avec la même combinaison de niveau de facteur. Cependant, contrairement à l'approche de régression multiniveau et de poststratification de Gelman et de Little (1997), notre construction de cellules n'atteindra pas l'homogénéité totale, que ce soit en ce qui concerne la variable dépendante d'intérêt ou les poids de sondage. Cependant, on souligne que, pour obtenir une homogénéité totale au chapitre de la variable dépendante et des poids, il faudra probablement utiliser de nombreuses cellules pour lesquelles des données fiables sur la taille de la population ne sont peut-être pas accessibles à partir des données de projections démographiques. On pourrait devoir compter sur des données d'enquête instables pour obtenir les chiffres de population des petites cellules.

Contrairement aux approches de modélisation habituelles dans des scénarios semblables (comme l'approche de régression multiniveau et de poststratification), on ne suppose pas de modèle échangeable dans les cellules ou de modèle élaboré pour les unités individuelles non observées, car le modèle supposé ne peut pas être vérifié lorsque l'on ne dispose pas de données sur la variable dépendante. S'inspirant plutôt des méthodes synthétiques pour petits domaines, on suppose que les moyennes de population de la variable à l'étude dans les cellules ayant la même combinaison de niveaux de facteur sont identiques pour tous les petits domaines, et que la moyenne de population d'une cellule est identique à la moyenne des valeurs

véritables des unités observées dans cette cellule. Puisque la taille de l'échantillon dans un domaine donné est petite, bon nombre de ces cellules ne sont pas représentées dans l'échantillon. Si l'on peut trouver un échantillon pour cette cellule à partir d'autres domaines, le modèle supposé est utilisé pour produire une estimation synthétique hiérarchique bayésienne de la moyenne de la population finie de la variable à l'étude dans cette cellule pour le domaine donné. On suppose que les observations dans une cellule donnée et dans un domaine donné sont semblables à celles pour la cellule d'autres domaines. Si la cellule n'est pas représentée dans tous les domaines, la moyenne de la cellule peut être prédite à partir du modèle de population supposé. Cependant, pour simplifier la méthodologie sans connaître la valeur informative de l'échantillonnage et pour induire un degré de stabilité accru par rapport aux erreurs de spécification du modèle, on peut tout simplement ignorer cette cellule lorsque sa contribution est négligeable, comme c'est le cas dans notre exemple illustratif. À la base, on cherche à rendre la méthodologie proposée résistante à tout non-respect éventuel du modèle supposé.

Pour mettre en application la méthodologie suggérée, on utilise un exemple de la vie réelle, en ayant recours à une enquête probabiliste sur la COVID-19 représentant l'ensemble de la population adulte américaine, à une enquête non probabiliste représentant uniquement les utilisateurs adultes actifs de Facebook aux États-Unis, aux estimations des chiffres de la population adulte du Bureau du recensement des États-Unis à des niveaux détaillés ainsi qu'à des données d'un site Web indépendant de déclaration de données sur la COVID-19, afin d'estimer les taux de réticence à l'égard de la vaccination (proportions) dans les états américains et le district de Columbia (petits domaines). À l'aide de cet exemple et de cette application, on démontrera les problèmes associés aux estimations régulières fondées sur le plan de sondage lorsqu'on les utilise pour de petits domaines. On illustrera également la façon dont notre méthodologie peut servir à obtenir des estimations plus stables (surtout dans le cas des domaines dont la taille de l'échantillon est petite ou nulle) ainsi que des mesures fiables de l'incertitude.

Voici la répartition du reste de l'article. Dans la section 2, on décrit de manière détaillée notre méthodologie. Dans la section 3, on décrit une application de notre méthode à une véritable enquête. Dans la section 4, on montre les résultats obtenus dans le cadre de cette application de la vie réelle. Enfin, dans la section 5, on résume notre article et on le conclut en fournissant des orientations de recherche futures pour notre méthodologie.

2. Méthodologie

Pour chaque petit domaine $i\,(i=1,\ldots,m),$ on subdivise la population finie en G cellules à l'aide de facteurs accessibles dans les données d'enquête qui peuvent potentiellement influer sur la variable dépendante d'intérêt. On suppose que les tailles de la population pour ces cellules sont connues à partir des données de recensement. Faisons de N_i et de N_{ig} les tailles connues de la population du domaine i et de la g^{e} cellule dans le i^{e} domaine, respectivement, $i=1,\ldots,m;\ g=1,\ldots,G.$ Naturellement, $\sum_g N_{ig}=N_i$ et $\sum_i N_i=N,$ où N est la taille de la population finie.

Disons que y_{igk} représente notre variable dépendante d'intérêt pour la k^{e} unité appartenant à la g^{e} cellule qui relève du i^{e} petit domaine $(i=1,\ldots,m;\ g=1,\ldots,G, k=1,\ldots,N_{ig})$. On suppose que l'on a accès à différents types de variables auxiliaires connues pour modéliser notre variable dépendante. Supposons que $x_{g(1)}$ est un vecteur de variables indicatrices correspondant à la cellule g, que $x_{i(2)}$ est un vecteur de variables auxiliaires connues propres au domaine i et que $x_{igk(3)}$ est un vecteur de variables auxiliaires au niveau de la personne $(i=1,\ldots,m;\ g=1,\ldots,G, k=1,\ldots,N_{ig})$.

Supposons que l'on a un échantillon de taille n tiré de la population finie. Supposons que G_i est le nombre de cellules représentées dans l'échantillon du domaine i $(i=1,\ldots,m)$, c'est-à-dire avec des échantillons de taille positive dans le domaine i. Faisons de n_i et de n_{ig} les tailles de l'échantillon pour le domaine i et pour la cellule g du domaine i, respectivement, afin que $\sum_{g=1}^{G_i} n_{ig} = n_i$ et que $\sum_{i=1}^{m} n_i = n$. Il est possible d'avoir $n_i = 0$ pour un certain domaine i ou $n_{ig} = 0$ pour une certaine cellule g dans un domaine i; $i=1,\ldots,m;\ g=1,\ldots,G$. Disons que w_{igk} représente le poids de sondage pour le k^{e} répondant échantillonné appartenant à la g^{e} cellule qui réside dans le i^{e} petit domaine $(i=1,\ldots,m;\ g=1,\ldots,G_i,\ k=1,\ldots,n_{ig})$.

2.1 Paramètre d'intérêt

On souhaite estimer les moyennes de population finie pour tous les domaines, c'est-à-dire

$$\bar{Y}_i = N_i^{-1} \sum_{g=1}^{G} \sum_{k=1}^{N_{ig}} y_{igk} = \sum_{g=1}^{G} \frac{N_{ig}}{N_i} \bar{Y}_{ig}, \tag{2.1}$$

où $N_i = \sum_g N_{ig}$, la taille de la population pour le i^{e} domaine $(i=1,\ldots,m)$.

2.2 Modèles hiérarchiques suggérés pour les répondants échantillonnés et la population finie

Notre approche exige une spécification d'un modèle explicite pour la variable dépendante y pour les N unités de la population finie. Compte tenu de θ_{igk}, on suppose que les y_{igk} sont indépendantes avec la moyenne θ_{igk} $i=1,\ldots,m;\ g=1,\ldots,G, k=1,\ldots,N_{ig}$.

On supposera un modèle élaboré pour tous les répondants, car on peut évaluer un tel modèle en suivant des étapes minutieuses de construction de modèle lors de l'analyse de données réelles. Dans le cas de $i=1,\ldots,m;\ g=1,\ldots,G_i, k=1,\ldots,n_{ig}$, on suggère le modèle suivant pour les répondants échantillonnés :

$$\begin{aligned}
&\text{Niveau 1 : } y_{igk} \mid \theta_{igk} \overset{\text{ind}}{\sim} f_1[\theta_{igk}, V(\theta_{igk})],\\
&\text{Niveau 2 : } \phi(\theta_{igk}) = x'_{g(1)}\alpha + x'_{i(2)}\beta + x'_{igk}\xi_c + \lambda h(w_{igk}) + v_i + u_{ig},\\
&\text{Niveau 3 : } v_i \text{ et } u_{ig} \text{ sont indépendants, et } v_i \overset{\text{iid}}{\sim} f_2[0,\sigma_v^2], u_{ig} \overset{\text{iid}}{\sim} f_3[0,\sigma_u^2],
\end{aligned} \tag{2.2}$$

où $f_j[a,b]$ $(j=1,2,3)$ représente une masse connue ou une fonction de densité ayant une moyenne a et une variance b. Par exemple, on peut choisir f_1 à partir de $\text{FEN}-\text{FVQ}[\mu, V(\mu)]$, la famille exponentielle naturelle ayant une moyenne μ et une fonction de variance quadratique $V(\mu) = v_0 + v_1\mu + v_2\mu^2$, où v_0, v_1 et

v_2 sont des constantes. La famille comprend la distribution de Gauss : $N(\mu,\sigma^2)\,(v_0=\sigma^2, v_1=v_2=0)$, la distribution gamma : $G(r,\lambda)\,(\mu=r\lambda, v_0=v_1=0, v_2=\frac{1}{r})$, la distribution de Poisson : $P(\lambda)\,(\mu=\lambda, v_0=v_2=0, v_1=1)$, la distribution binomiale : $\mathrm{B}(r,\theta)\,(\mu=r\theta, v_0=0, v_1=1, v_2=-\frac{1}{r})$, la distribution binomiale négative : $\mathrm{BN}(r,\theta)\,(\mu=r\frac{\theta}{1-\theta}, v_0=0, v_1=1, v_2=\frac{1}{r})$, etc. Pour obtenir les propriétés théoriques de la FEN-FVQ, voir Morris (1982). Dans le cas de f_2 et de f_3, on peut choisir une densité normale.

Comme c'est le cas pour le niveau 2, on peut choisir $\phi()$ comme fonction de lien connue pertinente (par exemple si f_1 est une distribution binomiale, on peut choisir $\phi()$ comme lien logit, et si f_1 est une distribution normale, $\phi()$ peut être le lien identité, etc.) et ξ_c comme vecteur de coefficients inconnus fixes (le suffixe c dans ξ_c permet la dépendance des coefficients de régression pour certaines combinaisons des facteurs qui forment les cellules démographiques; on déterminera c à l'étape de sélection du modèle). $h(\cdot)$ est une fonction connue des poids de sondage qui sera déterminée à l'étape de sélection du modèle et λ est un coefficient inconnu fixe. Verret et coll. (2015) ont envisagé de multiples choix pour $h(\cdot)$. Si l'analyse des données indique $\lambda=0$, on peut supposer un échantillonnage non informatif et réaliser l'analyse sans ce terme supplémentaire. Au niveau 2, on a ajouté les poids de sondage comme variable auxiliaire supplémentaire, afin de réduire la portée de la valeur informative. En ce qui concerne le niveau 3, on peut utiliser f_2 et f_3 comme densités normales pour les effets aléatoires propres au domaine v_i et les effets aléatoires propres au domaine et aux groupes u_{ig}, respectivement. Ces erreurs aléatoires de moyenne nulle permettront de s'occuper de toute variabilité restante qui n'a pas été prise en compte par les effets principaux fixes et de leurs interactions intégrées au modèle au niveau 2.

Abordons maintenant nos hypothèses de modélisation pour les $N-n$ unités non observées restantes de la population finie. À cette fin, pour les petits domaines $i=1,\ldots,m$, définissons :

$\mathcal{G}_{1i}=\{g$ de sorte que $n_{ig}>0, g=1,\ldots,G\}$: l'ensemble des cellules représentées dans l'échantillon pour le domaine i. $\mathcal{G}_{2i}=\{g$ de sorte que $n_{ig}=0, n_{jg}>0,$ pour un certain $j\neq i, g=1,\ldots,G\}$: l'ensemble des cellules non représentées dans l'échantillon du domaine i, mais représentées par au moins un autre domaine $j\neq i$. Et enfin, $\mathcal{G}_{3i}=\{g$ de sorte que $n_{ig}=0, \forall i, g=1,\ldots,G\}$: l'ensemble des cellules non représentées dans l'échantillon pour tout domaine. Définissons la moyenne de population pour l'ensemble de cellules $\mathcal{G}_{1i}$ comme suit :

$$\bar{\Theta}_{1i}=\frac{\sum_{g\in\mathcal{G}_{1i}}\sum_{k=1}^{N_{ig}}\theta_{igk}}{\sum_{g\in\mathcal{G}_{1i}}N_{ig}}=\sum_{g\in\mathcal{G}_{1i}}b_{ig;1}\bar{\Theta}_{ig},\ \text{où}\ b_{ig;1}=\frac{N_{ig}}{\sum_{g\in\mathcal{G}_{1i}}N_{ig}}\ \text{et}\ \bar{\Theta}_{ig}=\frac{\sum_{k=1}^{N_{ig}}\theta_{igk}}{N_{ig}}.$$

Dans le même ordre d'idées, on définit

$$\bar{\Theta}_{2i}=\frac{\sum_{g\in\mathcal{G}_{2i}}\sum_{k=1}^{N_{ig}}\theta_{igk}}{\sum_{g\in\mathcal{G}_{2i}}N_{ig}}=\sum_{g\in\mathcal{G}_{2i}}b_{ig;2}\bar{\Theta}_{ig},\ \text{où}\ b_{ig;2}=\frac{N_{ig}}{\sum_{g\in\mathcal{G}_{2i}}N_{ig}}\ \text{et}\ \bar{\Theta}_{ig}=\frac{\sum_{k=1}^{N_{ig}}\theta_{igk}}{N_{ig}}$$

et

$$\bar{\Theta}_{3i}=\frac{\sum_{g\in\mathcal{G}_{3i}}\sum_{k=1}^{N_{ig}}\theta_{igk}}{\sum_{g\in\mathcal{G}_{3i}}N_{ig}}=\sum_{g\in\mathcal{G}_{3i}}b_{ig;3}\bar{\Theta}_{ig},\ \text{où}\ b_{ig;3}=\frac{N_{ig}}{\sum_{g\in\mathcal{G}_{3i}}N_{ig}}\ \text{et}\ \bar{\Theta}_{ig}=\frac{\sum_{k=1}^{N_{ig}}\theta_{igk}}{N_{ig}}$$

comme les moyennes de la population pour l'ensemble de cellules $\mathcal{G}_{2i}$ et l'ensemble de cellules $\mathcal{G}_{3i}$, respectivement. Dans ce cas-ci, on formule une hypothèse synthétique selon laquelle :

$$\bar{\Theta}_{2i} = \sum_{g\in\mathcal{G}_{2i}} b_{ig;2}\bar{\Theta}_{ig;\text{syn}}, \text{ où } \bar{\Theta}_{ig;\text{syn}} = \frac{\sum_{j:\, n_{jg}>0,\, j\neq i}\sum_{k=1}^{n_{jg}}\theta_{jgk}}{\sum_{j:\, n_{jg}>0,\, j\neq i} N_{jg}},$$

c'est-à-dire que $\bar{\Theta}_{2i}$ est identique à la moyenne globale des cellules dans $\mathcal{G}_{2i}$ que l'on peut obtenir à partir d'au moins un domaine (autre que le domaine i). On ne formule aucune hypothèse en ce qui concerne le reste des unités de la population finie. Puisque, de manière générale, les tailles de population N_{ig} sont grandes, en s'appuyant sur la loi des grands nombres comme dans Jiang et Lahiri (2006), on peut approximer les moyennes de population finie $\bar{Y}_i$ dans l'équation (2.1) à l'aide des fonctions suivantes des paramètres du modèle de population finie supposé.

$$\begin{aligned}\bar{Y}_i \approx \bar{\Theta}_i = \sum_{g=1}^{G}\frac{N_{ig}}{N_i}\bar{\Theta}_{ig} &= a_{1i}\bar{\Theta}_{1i} + a_{2i}\bar{\Theta}_{2i} + (1-a_{1i}-a_{2i})\bar{\Theta}_{3i}, \text{(disons)} \\ &\approx a_{1i}\bar{\Theta}_{1i} + a_{2i}\bar{\Theta}_{2i},\end{aligned} \tag{2.3}$$

où $a_{ki} = N_i^{-1}\sum_{g\in\mathcal{G}_{ki}} N_{ig}$, la proportion démographique des unités appartenant à l'ensemble de cellules $\mathcal{G}_{ki}$; $k = 1,2,3$. L'approximation à l'équation (2.3) ne sera raisonnable que si $(1-a_{1i}-a_{2i}) \approx 0$.

2.3 Mise en œuvre bayésienne du modèle hiérarchique

Comme cela est fait couramment dans le cadre de la majorité des mises en œuvre de la méthode hiérarchique bayésienne, les paramètres du modèle seront estimés à l'aide de la méthode d'échantillonnage de Monte Carlo par chaîne de Markov (MCMC) à partir de distributions *a posteriori*. Supposons des distributions *a priori* faiblement informatives (Gelman, Jakulin, Pittau et Su, 2008) pour les hyper-paramètres du modèle : $\alpha, \beta, \xi_c, \lambda, \sigma_v$ et σ_v. Lors de chaque étape $(r = 1,\ldots,R)$ de la répétition MCMC, on produit ce qui suit pour le domaine i :

$$\bar{\theta}_i^{(r)} = a_{1i}\bar{\theta}_{1i}^{(r)} + a_{2i}\bar{\theta}_{2i}^{(r)}, r = 1,\ldots,R,$$

où

$$\bar{\theta}_{1i}^{(r)} = \sum_{g\in\mathcal{G}_{1i}} b_{ig;1}\bar{\theta}_{ig}^{(r)}, \text{ avec } \bar{\theta}_{ig}^{(r)} = \frac{\sum_{k=1}^{n_{ig}} w_{igk}\theta_{igk}^{(r)}}{\sum_{k=1}^{n_{ig}} w_{igk}},$$

$$\bar{\theta}_{2i}^{(r)} = \sum_{g\in\mathcal{G}_{2i}} b_{ig;2}\,\bar{\theta}_{ig;\text{syn}}^{(r)}, \text{ avec } \bar{\theta}_{ig;\text{syn}}^{(r)} = \frac{\sum_{j:\, n_{jg}>0,\, j\neq i}\sum_{k=1}^{n_{jg}} w_{jgk}\,\theta_{jgk}^{(r)}}{\sum_{j:\, n_{jg}>0,\, j\neq i}\sum_{k=1}^{n_{jg}} w_{jgk}},$$

$$\theta_{igk}^{(r)} = \phi^{-1}\left(x'_{g(1)}\alpha^{(r)} + x'_{i(2)}\beta^{(r)} + x'_{igk}\xi_c^{(r)} + \lambda^{(r)}h(w_{igk}) + v_i^{(r)} + u_{ig}^{(r)}\right).$$

On approxime $\bar{\Theta}_{1i} = \sum_{g\in\mathcal{G}_{1i}} b_{ig;1}\bar{\Theta}_{ig}$ par $\bar{\theta}_{1i} = \sum_{g\in\mathcal{G}_{1i}} b_{ig;1}\bar{\theta}_{ig}$, avec $\bar{\theta}_{ig} = \frac{\sum_{k=1}^{n_{ig}} w_{igk}\theta_{igk}}{\sum_{k=1}^{n_{ig}} w_{igk}}$. Jiang et Lahiri (2006) et Lahiri et Suntornchost (2020) ont utilisé des approximations semblables pour leur approche du meilleur

prédicteur empirique et leur approche hiérarchique bayésienne, respectivement. Une approximation du genre serait raisonnable si le nombre total de cellules était élevé, car on s'attendrait à l'homogénéité des réponses et à des poids correspondants dans les cellules. L'hétérogénéité résiduelle devrait être atténuée par la correction du biais décrite dans la sous-section suivante.

Dans le même ordre d'idées, il faut approximer $\bar{\Theta}_{2i} = \sum_{g \in \mathcal{G}_{2i}} b_{ig;2} \bar{\Theta}_{ig;\text{syn}}$ par $\sum_{g \in \mathcal{G}_{2i}} b_{ig;2} \bar{\theta}_{ig;\text{syn}} = \bar{\theta}_{2i}$, disons, où :

$$\bar{\theta}_{ig;\text{syn}} = \frac{\sum_{j: n_{jg} > 0,\, j \neq i} \sum_{k=1}^{n_{jg}} w_{jgk} \theta_{jgk}}{\sum_{j: n_{jg} > 0,\, j \neq i} \sum_{k=1}^{n_{jg}} w_{jgk}}.$$

L'ensemble de R répétitions MCMC pour le i^{e} domaine, c'est-à-dire $\{\bar{\theta}_i^{(r)}, r = 1, \ldots, R\}$, sera utilisé pour les inférences sur $\bar{\theta}_i$.

2.4 Correction du biais pour l'estimateur hiérarchique bayésien

L'estimateur hiérarchique bayésien présenté dans la section 2.3 intègre des poids de sondage pour tenir compte de l'échantillonnage informatif potentiel. Cependant, dans le cas des cellules dans $\mathcal{G}_{2i}$ (cellules qui ne sont pas représentées dans le domaine i, mais qui le sont dans d'autres domaines), les estimations synthétiques peuvent faire l'objet d'un biais en raison des différences dans la structure de pondération parmi les domaines. Pour tenir compte de cette situation, on met en application une procédure de correction du biais.

Faisons de $\bar{\theta}_{ig}$ notre estimateur hiérarchique bayésien de $\bar{\Theta}_{ig}$ du modèle dans l'équation (2.2) qui comprend des poids de sondage. On l'appelle « modèle pondéré ». Dans le même ordre d'idées, supposons que $\bar{\theta}_{ig}^*$ est l'estimateur d'un modèle qui exclut le terme de poids de sondage $\lambda h(w_{igk})$ (de manière semblable, le « modèle non pondéré »), ce qui est semblable aux approches standard de régression multiniveau et de poststratification (Gelman et Little, 1997).

Dans le cas des cellules dans $\mathcal{G}_{2i}$, notre estimateur ajusté en fonction du biais est :

$$\bar{\theta}_{ig;bc} = \bar{\theta}_{ig}^* + \widehat{\text{Biais}}_{ig;\text{syn}}, \tag{2.4}$$

où le terme de correction du biais est estimé comme suit :

$$\widehat{\text{Biais}}_{ig;\text{syn}} = \frac{\sum_{j: n_{jg} > 0,\, j \neq i} \sum_{k=1}^{n_{jg}} w_{jgk} (\hat{\theta}_{jgk} - \hat{\theta}_{jgk}^*)}{\sum_{j: n_{jg} > 0,\, j \neq i} \sum_{k=1}^{n_{jg}} w_{jgk}}. \tag{2.5}$$

Dans ce cas-ci, $\hat{\theta}_{jgk}$ et $\hat{\theta}_{jgk}^*$ sont les valeurs prédites pour l'unité k dans la cellule g du domaine j des modèles pondéré et non pondéré, respectivement.

Voici l'intuition sur laquelle repose cette correction : on calcule tout d'abord les estimations à l'aide du modèle non pondéré efficace sur le plan computationnel (semblable à l'approche de régression multiniveau et de poststratification). On corrige ensuite ces estimations en ajoutant la différence systématique estimée entre les modèles pondéré et non pondéré, calculée à l'aide de données d'autres domaines dans lesquels la

cellule g est observée. Cette approche nous permet de tirer parti de l'efficacité computationnelle des méthodes de type régression multiniveau et poststratification tout en corrigeant tout biais potentiel introduit par le fait d'avoir ignoré les poids de sondage.

Dans le cadre de l'exécution, lors de chaque répétition MCMC r, l'estimation corrigée en fonction du biais pour le domaine i devient :

$$\bar{\theta}_{i;bc}^{(r)} = a_{1i}\bar{\theta}_{1i}^{(r)} + a_{2i}\bar{\theta}_{2i;bc}^{(r)}, \tag{2.6}$$

où $\bar{\theta}_{2i;bc}^{(r)} = \sum_{g\in\mathcal{G}_{2i}} b_{ig;2}\bar{\theta}_{ig;bc}^{(r)}$.

En ce qui concerne la précision de nos estimations de l'incertitude, il est important de souligner que notre procédure de correction du biais ne tient pas uniquement compte de l'estimation ponctuelle, mais contribue aussi à obtenir des mesures plus fiables de l'incertitude. En tenant compte de la différence entre les modèles pondéré et non pondéré, on atténue le risque de sous-estimer les véritables erreurs. Bien qu'une étude par simulation exhaustive dépasse la portée du présent document, le cadre hiérarchique bayésien intègre naturellement plusieurs sources d'incertitude, notamment l'incertitude liée aux effets aléatoires, aux effets fixes et aux hyperparamètres, ce qui permet d'obtenir des intervalles de crédibilité qui reflètent raisonnablement l'incertitude totale du processus d'estimation. Les données probantes empiriques fournies dans la section 4.3, qui montrent une grande homogénéité des poids au sein des cellules, renforcent la fiabilité de nos mesures de l'incertitude en réduisant une source importante de biais potentiel dans les estimateurs de type Hájek.

2.5 Forme analytique spéciale de notre estimation bayésienne suggérée

On examine maintenant le cas spécial suivant de notre modèle général pour petits domaines afin de comprendre notre estimateur.

$$\begin{aligned}
&\text{Niveau 1 : } y_{igk} \mid \theta_{igk} \overset{\text{ind}}{\sim} N(\theta_{igk}, \sigma_y^2),\\
&\text{Niveau 2 : } \theta_{igk} \mid \Lambda_{igk}, v_i \overset{\text{ind}}{\sim} N(\Lambda_{igk} + v_i, \sigma_u^2),\\
&\text{Niveau 3 : } v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2),
\end{aligned}$$

où $\Lambda_{igk} = x'_{g(1)}\alpha + x'_{i(2)}\beta + x'_{igk(3)}\xi_c + \lambda h(w_{igk})$. On peut exprimer à nouveau le niveau 2 du même modèle comme suit : $\theta_{igk} = \Lambda_{igk} + v_i + u_{ig}$, où

$$u_{ig} \overset{\text{iid}}{\sim} N(0, \sigma_u^2).$$

Selon cette considération et dans le but de simplifier l'illustration, en ayant une hypothèse supplémentaire selon laquelle on connaît α, β, ξ_c, λ, σ_u^2 et σ_v^2, on peut obtenir une forme analytique de l'estimation bayésienne approximative de petits domaines se fondant sur l'approximation $a_{1i}\bar{\Theta}_{1i} + a_{2i}\bar{\Theta}_{2i}$ figurant dans l'équation (2.3). Naturellement, ce sera le « meilleur prédicteur » des moyennes de petits domaines de Rao et Molina (2015). Comme on l'a déjà mentionné dans les sous-sections 2.2 et 2.3, et en utilisant les mêmes

notations et terminologies, l'estimation approximative bayésienne de la moyenne de petits domaines pour un domaine donné i sera fournie par : $E(a_{1i}\bar{\theta}_{1i} + a_{2i}\bar{\theta}_{2i} \mid \text{données})$. Selon la normalité de chaque niveau du modèle illustratif, on obtient :

$$
\begin{aligned}
E(a_{1i}\bar{\theta}_{1i} + a_{2i}\bar{\theta}_{2i} \mid \text{données}) &= a_{1i}E(\bar{\theta}_{1i} \mid \text{données}) + a_{2i}E(\bar{\theta}_{2i} \mid \text{données}) \\
&= a_{1i}\sum_{g\in\mathcal{G}_{1i}} b_{ig;1} \times \frac{1}{\sum_k w_{igk}}\left[\sum_k w_{igk}\left\{\frac{\sigma_u^2 y_{igk} + \sigma_y^2 \Lambda_{igk}}{\sigma_u^2 + \sigma_y^2} + \right.\right. \\
&\quad \left.\left. \frac{n_i\sigma_y^2\sigma_v^2(\bar{y}_{i..} - \bar{\Lambda}_{i..})}{(\sigma_u^2 + \sigma_y^2)(n_i\sigma_v^2 + \sigma_u^2 + \sigma_y^2)}\right\}\right] \\
&\quad + a_{2i}\sum_{g\in\mathcal{G}_{2i}} b_{ig;2} \times \frac{1}{\sum_{j\neq i}\sum_k w_{jgk}}\left[\sum_{j\neq i}\sum_k w_{jgk}\left\{\frac{\sigma_u^2 y_{jgk} + \sigma_y^2 \Lambda_{jgk}}{\sigma_u^2 + \sigma_y^2} + \right.\right. \\
&\quad \left.\left. \frac{n_j\sigma_y^2\sigma_v^2(\bar{y}_{j..} - \bar{\Lambda}_{j..})}{(\sigma_u^2 + \sigma_y^2)(n_j\sigma_v^2 + \sigma_u^2 + \sigma_y^2)}\right\}\right]
\end{aligned}
$$

où $\bar{y}_{i..} = \frac{1}{n_i}\sum_{g,k} y_{igk}$. $\bar{\Lambda}_{i..}$, $\bar{y}_{j..}$ et $\bar{\Lambda}_{j..}$ sont définies de manière semblable.

À partir de cette illustration simple, selon les hypothèses habituelles formulées pour estimer le meilleur prédicteur, on peut obtenir une explication sous la forme de notre estimation bayésienne. On observe la façon dont les y_{igk} accessibles à partir des groupes de l'échantillon du domaine i intègrent l'estimateur par l'intermédiaire du premier terme de la somme. Les groupes qui ne sont pas représentés dans l'échantillon du même domaine sont intégrés à l'estimateur par l'intermédiaire du deuxième terme de la somme, en ce qui concerne les échantillons accessibles pour ces groupes provenant d'autres domaines.

3. Application dans la vie réelle

Pour expliquer clairement la méthodologie que l'on suggère, on examine un sujet ayant une importance pour la recherche contemporaine. On cherche à fournir des estimations raisonnablement stables, au niveau des états (nos petits domaines d'étude), de la proportion de personnes qui éprouvent de la « réticence » à se faire vacciner contre la COVID-19. Dans la présente section, on fournit tout d'abord une vue d'ensemble des sources de données disparates envisagées pour l'application.

3.1 Données primaires : Understanding America Study

La Understanding America Study (UAS), réalisée par l'Université de la Californie du Sud, réunit environ 9 000 répondants représentant l'ensemble de la population adulte aux États-Unis. Servant de panel Internet, les participants prennent part à des enquêtes à l'aide d'appareils dotés d'une connexion Internet au moment qui leur convient. Les ménages, au sens large, comprennent les personnes qui résident avec le participant

principal. Pour extraire des données utiles sur les attitudes, les comportements et les aspects sociétaux, le Center for Economic and Social Research de l'Université de la Californie du Sud a lancé l'enquête de suivi s'intitulant « Understanding Coronavirus in America » le 10 mars 2020. Reposant sur le panel de l'étude UAS, cette enquête permanente, qui reçoit le soutien de la Bill and Melinda Gates Foundation et du National Institute on Aging, porte sur différentes dimensions, y compris la santé mentale, les comportements en matière de soins de santé ainsi que la crise économique pendant la pandémie de COVID-19.

Au départ, l'enquête réunissait 8 547 participants admissibles du panel de l'étude UAS parmi les 9 603 répondants qui ont indiqué vouloir y participer. Notre étude met l'accent sur les données recueillies pendant la vague 16 (du 14 octobre au 11 novembre 2020), pour laquelle l'enquête a été transmise à 7 832 participants à l'étude UAS. Parmi eux, 6 081 ont répondu à l'enquête et sont donc nos répondants. Le taux de réponse a été d'environ 78 %. Les participants avaient 14 jours pour répondre à l'enquête. Ils ont reçu un dédommagement de 15 $ lorsqu'ils transmettaient le questionnaire rempli de l'enquête.

3.1.1 Estimations directes de la Understanding America Study

Dans le cadre de l'analyse sur la « réticence à l'égard de la vaccination », on a mis l'accent sur les réponses à la question d'enquête suivante : « Dans quelle mesure êtes-vous susceptible de vous faire vacciner contre le coronavirus lorsqu'un vaccin sera offert au public? ». Les répondants avaient cinq choix de réponse : très peu susceptible, assez peu susceptible, plutôt susceptible, très susceptible et incertain. On a transformé les réponses en variable binaire, attribuant une valeur de 1 si le répondant a fourni une réponse autre que « très susceptible » et une valeur de 0 dans le cas contraire. À l'aide de cette variable binaire et des poids de sondage correspondants, on a calculé les estimations directes fondées sur l'enquête pour les proportions de personnes qui *ne* sont *pas* « très susceptibles » de se faire vacciner dans les 50 états et le district de Columbia. Tandis que l'estimation nationale respectait des normes de grande qualité (marge d'erreur inférieure à 3 %), les estimations pour les régions géographiques plus petites (50 états et le district de Columbia) étaient moins fiables.

Le tableau 4.7 présente des estimations d'enquête directes des proportions de la réticence à l'égard de la vaccination pour les 50 états et le district de Columbia, accompagnées des erreurs-types estimées. On a notamment constaté une estimation peu réaliste de 1 et une erreur-type associée estimée de 0 pour le Delaware. Cette anomalie s'est produite, car aucun des répondants du Delaware a répondu « très susceptible » à la question d'enquête. Cependant, cela ne signifie pas pour autant que la réticence au Delaware est généralisée. L'estimation pour l'Alaska est de 0,21 et son erreur-type associée fondée sur le plan est élevée, soit 0,19, ce qui dépasse l'intervalle habituel de 0,02. On observe des tendances semblables pour le Vermont, le Rhode Island, le district de Columbia et d'autres états.

3.2 Descriptions de données supplémentaires

Dans les sous-sections subséquentes, on décrit les différentes sources de données supplémentaires examinées dans le contexte de l'application de notre méthodologie.

3.2.1 Enquête sur les symptômes de la COVID-19

L'enquête sur les symptômes de la COVID-19, une initiative de collaboration entre des spécialistes de la santé publique et Facebook, a été menée aux États-Unis par le Carnegie Mellon Delphi Research Center et à l'échelle internationale par le Joint Program in Survey Methodology de l'Université du Maryland. Cette enquête non probabiliste, représentative des utilisateurs adultes de Facebook, a permis de recueillir des données quotidiennes. À la recherche de covariables fiables au niveau de l'état, on a intégré des variables statiques de l'enquête, en mettant l'accent sur les réponses à deux questions qui portaient sur : 1) les problèmes de santé préexistants; on utilisait un code 1 si les répondants ont déclaré faire du diabète ou de l'asthme ou avoir une maladie pulmonaire chronique, et un code 0 dans tous les autres cas; 2) les vaccins contre la grippe reçus au cours des 12 mois précédents; on utilisait un code 1 si la réponse était « oui » et un code 0 dans tous les autres cas. Après avoir combiné les réponses à l'enquête obtenues du 5 août au 9 septembre 2020, on a calculé les estimations pondérées à l'échelle des états pour les personnes ayant des maladies précises et celles qui avaient été vaccinées contre la grippe. Ces estimations, obtenues sous forme de moyennes pondérées par les poids d'enquête des variables codées de façon binaire, servent de covariables continues dans notre modèle.

3.2.2 Le Population Estimates Program (PEP) du Bureau du recensement des États-Unis

Le Population Estimates Program (PEP) du Bureau du recensement des États-Unis fournit des estimations démographiques annuelles pour différents niveaux géographiques, notamment les états, les comtés, et les petites et grandes villes, ainsi que pour des composantes démographiques, comme les naissances, les décès, la migration et les logements. À l'aide des riches données du PEP, on produit des chiffres de population au niveau des états et au niveau des cellules, en mettant l'accent sur les cellules démographiques suivantes : `race × ethnicité × genre × catégorie d'âge`. Les données démographiques prises en compte comprennent quatre niveaux de `race` (Blanc, Noir, Asiatique et autres), deux niveaux d'`ethnicité` (Hispanique et non-Hispanique), deux niveaux de `genre` (homme et femme) et sept niveaux de `catégorie d'âge` (18 à 24 ans, 25 à 34 ans, 35 à 44 ans, 45 à 54 ans, 55 à 64 ans, 65 à 74 ans et 75 ans et plus). Les données utilisées sont associées aux estimations démographiques annuelles des résidents d'état pour cinq groupes raciaux (cinq groupes raciaux seuls ou combinés) selon l'âge, le sexe et l'origine hispanique, et ce, du 1er avril 2010 au 1er juillet 2019. Le prétraitement comprenait le regroupement et l'agrégation des données pour obtenir les chiffres pour les cellules démographiques précisées (cellules poststratifiées).

3.2.3 Le projet COVID Tracking

Dans le cadre de notre modèle, on a utilisé les données du site https://covidtracking.com, qui fournit des données sur la COVID-19 pouvant être téléchargées librement pour les 50 états et le district de Columbia. On a créé deux covariables continues propres aux états : le « taux de dépistage », soit le nombre total de tests réalisés dont les résultats ont été confirmés divisé par le nombre total de tests réalisés dans l'état (on

reconnaît des valeurs potentielles supérieures à un en raison du fait qu'une personne peut avoir fait de multiples tests) et le « taux de positivité », obtenu en divisant le nombre total de résultats positifs au test de dépistage par le nombre total de tests dont les résultats ont été confirmés.

4. Analyse et évaluation des données

On commence la présente section en soulignant que, compte tenu des données accessibles, on a réalisé un examen exhaustif de plusieurs variables. Dans le cadre d'une série d'essais et d'erreurs, on en est arrivé à un ensemble précis de covariables jugé pertinent pour notre analyse. On a également pris en compte des considérations supplémentaires par rapport à la sélection des modèles pour l'étape de leur sélection, et toutes étaient fondées sur une évaluation minutieuse du problème et des ensembles de données dont il est question. On recommande aux spécialistes des sciences appliquées de faire preuve de jugement et de tenir compte de certains aspects, comme la nature du problème et l'expertise en la matière, lorsqu'ils suivent des étapes semblables dans le cadre de leur analyse. Comme on l'a mentionné précédemment dans la sous-section 3.2.2, on désigne la « RACE » par a, qui a quatre niveaux (Blanc, Noir, Asiatique et autres), l'« ETHNICITÉ » par b, qui a deux niveaux, (Hispanique et non-Hispanique) et le « GENRE » par c, qui a deux niveaux (homme et femme). Dans le cas de la variable « ÂGE », qui est représentée par d, on a divisé l'ensemble de la population adulte en sept cellules, c'est-à-dire 18 à 24 ans, 25 à 34 ans, 35 à 44 ans, 45 à 54 ans, 55 à 64 ans, 65 à 74 ans et 75 ans et plus.

On tient compte des ordonnées à l'origine à effet fixe de la « RACE » × l'« ETHNICITÉ ». En tout, on a donc pris en compte huit ordonnées à l'origine à effet fixe pour tous les modèles. Au tableau 4.1, on trouve une liste des modèles concurrents.

En ce qui concerne les covariables au niveau de l'état $x_{i(2)}$ (voir la section 2 à titre de référence pour cette notation), on utilise les données d'enquête de Facebook (voir la sous-section 3.2.1) ainsi que les données du projet COVID Tracking (voir la sous-section 3.2.3). On utilise les données d'enquête de Facebook sur les symptômes de la COVID pour calculer deux variables sommaires au niveau de l'état, c'est-à-dire les proportions pondérées par les poids d'enquête des personnes qui ont l'une des trois maladies préexistantes ou les taux de comorbidité, et les proportions pondérées par les poids d'enquête des personnes qui se sont fait vacciner contre la grippe au cours des 12 mois précédents ou les taux de vaccination contre la grippe (comme on l'a indiqué à la sous-section 3.2.1). On a recours à des transformations logit de ces deux taux comme covariables dans nos modèles décrits dans le tableau 4.1. Les autres covariables au niveau de l'état utilisées provenaient du projet COVID Tracking, comme les taux de dépistage et les taux de positivité, qui sont définis dans la sous-section 3.2.3. Enfin, on a également utilisé les pourcentages au niveau de l'état des votes pour le Parti républicain au cours de l'élection présidentielle des États-Unis en 2020 comme covariable au niveau de l'état, c'est-à-dire « pourcentage de vote pour les républicains » dans notre modèle (aussi appelée « % rép. » dans le tableau 4.6).

Comme première covariable au niveau des répondants $x_{igk(3)}$, on a utilisé les sept catégories d'âge des répondants et on les a converties en chiffres $1,\ldots,7$ pour les utiliser comme covariable à échelle continue dans tous nos modèles. Cette approche est courante dans la littérature. Par exemple, voir Ha et Sedransk (2019), qui utilisent les intervalles d'âge comme variables continues. Même si leur méthode diffère en partie de la nôtre, ils ont jugé qu'il est avantageux de traiter les intervalles d'âge d'une manière semblable. En fonction de nos études empiriques, on a observé des tendances comparables à leurs constatations. Notre analyse empirique a montré que le fait de traiter les groupes d'âge propres à chaque genre comme des variables continues révélait une relation linéaire avec les taux de réticence, ce qui renforce le pouvoir explicatif de notre modèle linéaire (transformé par la fonction logit). Notre méthodologie d'estimation demeure valide même si les catégories d'âge sont traitées différemment. On utilise des pentes propres au genre, soit une pour les hommes et une pour les femmes, pour la covariable de l'âge dans notre modèle. Ce choix repose sur nos études empiriques. Dans le cas de la deuxième covariable au niveau des répondants, on utilise les poids de sondage associés à chaque répondant des données de l'étude UAS.

Tableau 4.1
Liste des modèles concurrents

Modèle	Ordonnée à l'origine de la race × l'ethnicité	Pente de l'âge selon le genre	Taux de comorbidité (Facebook)	Taux de vaccination contre la grippe (Facebook)	Taux de dépistage (Covid Tracking)	Taux de positivité (Covid Tracking)	Pourcentage de vote pour les républicains	Poids de sondage
M1	√	√	√	√	√	√	√	√
M2	√	√	√	X	√	X	√	√
M3	√	√	√	X	X	X	√	√
M4	√	√	√	X	X	X	√	X

4.1 Ajustement du modèle

Il ne faut pas oublier que l'approximation faite pour le paramètre d'intérêt $\overline{Y}_i$ de la population (dans ce cas-ci, 50 états et le district de Columbia) dans l'équation (2.3) ne serait valide que si la quantité $(1-a_{i1}-a_{i2})\approx 0$. Le tableau 4.2 résume la quantité $(1-a_{i1}-a_{i2})$ pour les 50 états et le district de Columbia. On peut justifier raisonnablement notre approximation, puisque la valeur maximum (0,0120) et le troisième quartile (0,0031) sont négligeables.

Tableau 4.2
Statistiques sommaires de $(1-a_{i1}-a_{i2})$ pour les 50 états et le district de Columbia

	Valeur minimum	1er quartile	Médiane	Moyenne	3e quartile	Valeur maximum
$1-a_{i1}-a_{i2}$	0,0004	0,0010	0,0016	0,0026	0,0031	0,0120

Abordons maintenant l'ajustement du modèle. On examine le cas spécial suivant du modèle général suggéré à l'équation (2.2) :

$$\begin{aligned}
&\text{Niveau 1 : } y_{igk} \mid \theta_{igk} \overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{igk}),\\
&\text{Niveau 2 : } \text{logit}(\theta_{igk}) = x'_{g(1)}\alpha + x'_{i(2)}\beta + x'_{igk(3)}\xi_c + \lambda h(w_{igk}) + v_i + u_{ig},\\
&\text{Niveau 3 : } v_i \overset{\text{iid}}{\sim} N(0,\sigma_v^2), u_{ig} \overset{\text{iid}}{\sim} N(0,\sigma_u^2).
\end{aligned}$$

Pour ajuster les modèles au tableau 4.1, on a utilisé le langage de programmation probabiliste **Stan** de Stan Development Team (2024) avec le logiciel statistique **R** de R Core Team (2020). Le paquet **R** `rstan` de Stan Development Team (2020) fournit l'interface **R** pour **Stan**. Pour toutes les ordonnées à l'origine (une pour chaque combinaison Race × Ethnicité) et les pentes, on a utilisé des distributions *a priori* indépendantes et identiquement distribuées $N(0, 5^2)$. Dans le cas des composantes de variance – σ_v et σ_u, on a utilisé la distribution *a priori* de $\text{Cauchy}^+(0, 5)$ tronquée à gauche à 0. Il faut souligner que l'on utilise souvent le nom « distributions *a priori* peu informatives » dans la littérature (voir Gelman, Carlin, Stern, Dunson, Vehtari et Rubin, 2013) pour désigner de telles distributions *a priori* pour les hyperparamètres du modèle. Ces choix relatifs à la distribution *a priori* ont permis de s'assurer que la densité *a posteriori* sera appropriée. On a exécuté 4 000 répétitions pour chaque modèle. Les 2 000 premières répétitions ont été éliminées, car il s'agissait d'échantillons de rodage. Ainsi, toutes les analyses *a posteriori* étaient fondées sur les 2 000 tirages MCMC après le rodage.

4.2 Sélection du modèle

Après avoir ajusté les quatre modèles, on a utilisé un critère de sélection de modèle pour retenir le « meilleur » modèle fonctionnel. Le paquet **R** `loo` (*leave-one-out* en anglais ou exclusion d'une unité) de Vehtari, Gabry, Magnusson, Yao, Bürkner, Paananen et Gelman (2020) propose une façon rapide et pratique de diagnostiquer et de comparer l'ajustement des modèles. On a utilisé cette méthode dans le cadre du présent article. Le logiciel calcule une validation croisée par exclusion d'une unité approximative à l'aide de l'échantillonnage préférentiel lissé de Pareto, une procédure relativement récente pour régulariser les poids d'échantillonnage préférentiel. On renvoie à Vehtari, Gelman et Gabry (2017) pour obtenir un examen détaillé sur l'exclusion d'une unité dans Stan. Comme conséquence indirecte des calculs, l'exclusion d'une unité peut fournir des erreurs-types approximatives pour les erreurs prédictives estimées et pour la comparaison des erreurs prédictives entre deux modèles. Lors de la comparaison de deux modèles ajustés, le paquet peut estimer la différence en ce qui concerne leur exactitude prédictive attendue en fonction de la différence de leur $\hat{\text{elpd}}_{\text{loo}}$, l'estimation bayésienne par exclusion d'une unité de la densité prédictive ponctuelle logarithmique attendue (elpd pour *expected log pointwise predictive density* en anglais). Dans le cas d'un échantillon d'une taille suffisamment grande (utilisé pour ajuster les données), cette différence suit, de manière approximative, une distribution normale standard. Cela facilitera la comparaison de tous les modèles utilisés dans le cadre du présent article. Le tableau 4.3, qui suit, fournit les résultats de la comparaison de la validation croisée par exclusion d'une unité pour les quatre modèles.

Tableau 4.3
Comparaison de la validation croisée par exclusion d'une unité pour les quatre modèles

	Différence relative à la densité prédictive ponctuelle logarithmique attendue	Différence relative à l'erreur-type
M3	0	0
M2	-0,958	0,804
M1	-1,672	0,836
M4	-64,216	18,477

Note : On a jugé que le modèle 3 est le meilleur modèle.

Dans le tableau 4.3, on peut voir que le critère d'exclusion d'une unité retient le modèle 3 comme meilleur modèle. Il est meilleur que le modèle 2, comme le montre la valeur significative de la différence relative à la densité prédictive ponctuelle logarithmique attendue, soit ~ 68 %. Il faut souligner que les poids de sondage sont intégrés de manière linéaire au modèle 3. Suivant la méthode de Verret et coll. (2015), on souhaitait déterminer le rendement du modèle 3 lors de l'intégration non linéaire des poids de sondage dans le modèle, en ne changeant rien d'autre, c'est-à-dire $\lambda \ln(w_{igk})$ et λw_{igk}^{-1} au lieu de λw_{igk}. Naturellement, pour vérifier le rendement de ces deux nouveaux modèles concurrents, on a à nouveau eu recours au critère d'exclusion d'une unité. Dans le tableau 4.4, on observe que les poids de sondage ayant subi une transformation logarithmique ou l'inverse des poids de sondage ne permettent pas d'améliorer le modèle 3 existant. C'est la raison pour laquelle on conserve le modèle 3 comme meilleur modèle fonctionnel et l'on poursuit les analyses suivantes avec le modèle 3.

Tableau 4.4
Comparaison de la validation croisée par exclusion d'une unité des différentes fonctions de pondération du modèle 3

	Différence relative à la densité prédictive ponctuelle logarithmique attendue	Différence relative à l'erreur-type
M3	0	0
M3 $(\ln(w_{igk}))$	-0.329	1.874
M3 (w_{igk}^{-1})	-3.644	2.576

Si l'on regarde les analyses *a posteriori* du modèle 3, mises en évidence dans le tableau 4.6, on constate que la variable de la comorbidité de l'enquête de Facebook semble non significative, celle-ci se situant à 70 %. C'est la raison pour laquelle on a décidé d'exécuter une autre comparaison de modèles, en utilisant le modèle 3 retenu et le modèle 3 sans cette variable précise. L'analyse relative à l'exclusion d'une unité décrite dans le tableau 4.5 indique que le fait d'enlever la comorbidité du modèle ne mène pas à des différences significatives. On a donc décidé de laisser cette variable dans notre modèle.

Tableau 4.5
Comparaison de la validation croisée par exclusion d'une unité du modèle 3 et du modèle 3 sans le taux de comorbidité de la COVID-19 Trends and Impact Survey de Facebook et du Delphi Group de l'Université Carnegie Mellon

	Différence relative à la densité prédictive ponctuelle logarithmique attendue	Différence relative à l'erreur-type
M3	0	0
M3 (sans la comorbidité de Facebook)	-0,740	0,747

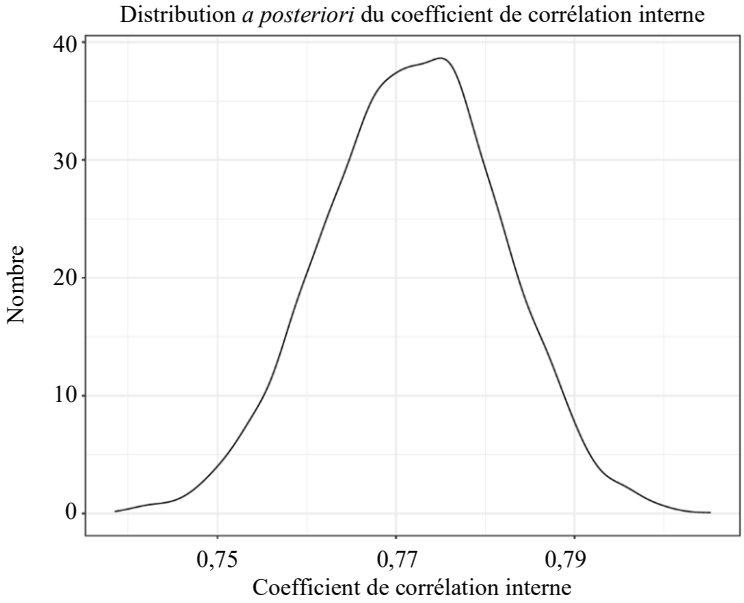
4.3 Justification empirique de l'approche de correction du biais

Pour justifier de manière empirique l'approche de correction du biais que l'on a décrite dans la section 2.4, on a réalisé une analyse portant sur la structure des poids de sondage parmi les cellules de poststratification. On a examiné la corrélation interne (ρ) pour les poids de sondage à l'aide d'un modèle hiérarchique bayésien. Dans le cas des poids de sondage (représentés par w_{igk}), on a précisé un modèle hiérarchique :

$$\begin{aligned} w_{igk} &\sim N(\mu_{igk}, \sigma_e^2) \\ \mu_{igk} &= \beta_0 + v_{ig} \\ v_{ig} &\sim N(0, \sigma_v^2) \end{aligned}$$

avec des distributions *a priori* peu informatives : $\beta_0 \sim N(0, 5^2)$, $\sigma_v \sim \text{Cauchy}^+(0, 5)$ et $\sigma_e \sim \text{Cauchy}^+(0, 5)$. Le coefficient de corrélation interne a été calculé comme suit : $\rho_w = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2)$. La figure 4.1 montre la distribution *a posteriori* du coefficient de corrélation interne pour les poids de sondage.

Figure 4.1 Distribution *a posteriori* du coefficient de corrélation interne pour les poids de sondage pour l'ensemble des cellules de poststratification



La distribution *a posteriori* de ρ était concentrée près de 0,77, et $P(\rho > 0{,}75) = 0{,}985$. Cette valeur de coefficient de corrélation interne élevée confirme une grande homogénéité des poids dans les cellules, ce qui fournit des données empiriques qui appuient notre approche méthodologique. Cela démontre que notre structure de poststratification tient compte avec efficacité de la structure des poids de sondage; les poids sont très homogènes dans les cellules (coefficient de corrélation interne $\approx$ 0,77). Cette homogénéité élevée dans les cellules appuie l'approximation dans nos estimateurs de type Hájek. Il est important de souligner que, même si l'on a réalisé l'analyse empirique pour les poids de sondage observés, il est impossible de réaliser une analyse semblable pour les valeurs θ_{igk} non observées, puisqu'il s'agit de quantités latentes qui ne peuvent pas être mesurées ou vérifiées directement. On remédie à la question de l'hétérogénéité restante à l'aide de la procédure de correction du biais décrite dans la section 2.4, qui fournit une protection supplémentaire, particulièrement pour les cellules de $\mathcal{G}_{2i}$ qui ne sont pas représentées directement dans leurs domaines respectifs.

4.4 Constatations tirées du modèle sélectionné

Dans la dernière sous-section, on a opté pour un modèle qui ajuste le mieux nos données à l'aide du critère d'exclusion d'une unité, en utilisant **Stan**, `rstan` et `loo`. On présente maintenant des constatations importantes tirées du « meilleur » modèle fonctionnel retenu, soit le modèle 3.

Le tableau 4.6 fournit les statistiques sommaires *a posteriori* pour le meilleur modèle retenu, soit le modèle 3. La colonne « moyenne » donne les moyennes *a posteriori*, la colonne « écart-type » fournit les écarts-types *a posteriori* (les erreurs-types estimées), les quatre colonnes suivantes fournissent les quatre centiles, la colonne N_{eff} fournit la taille effective des échantillons, c'est-à-dire le nombre d'échantillons indépendants utilisés pour remplacer le total N des tirages MCMC dépendants qui ont le même pouvoir d'estimation que les N échantillons autocorrélés, et, finalement, la colonne $\hat{R}$ fournit une mesure de l'équilibre des chaînes; une valeur près de 1 (mais $<$ 1,1) signifie que les chaînes se sont bien mélangées et que les échantillons *a posteriori* peuvent être utilisés avec certitude dans le cadre des analyses *a posteriori* (Stan Development Team, 2023; Gelman et coll., 2013). À titre de règle par défaut, Gelman et coll. (2013) recommandent d'exécuter la simulation jusqu'à ce que N_{eff} soit au moins $5m$, où m équivaut à deux fois le nombre de chaînes ou de séquences indépendantes. Pour nos calculs, on a exécuté deux chaînes indépendantes de 2 000 répétitions chacune. Dans notre cas, on a $5m = 5 \times 2 \times 2 = 20$. Comme on peut le voir dans le tableau 4.6, l'ensemble des $N_{\text{eff}} > 20$ et $\hat{R} <$ 1,1. On peut donc utiliser ces analyses *a posteriori* en étant plutôt certain que les chaînes ont atteint la stationnarité, en plus d'avoir été mélangées de manière appropriée. On utilise les estimations *a posteriori* pour calculer notre paramètre d'intérêt $\overline{Y}_i$ au niveau de l'état pour les 50 états et le district de Columbia. On obtient ainsi les estimations relatives à la réticence à l'égard de la vaccination au niveau de l'état. Dans le tableau 4.7, on peut voir ces chiffres ainsi que les erreurs-types estimées fondées sur le modèle correspondant. Pour faciliter la comparaison, les estimations directes fondées sur l'enquête sont également fournies dans le même tableau. On souligne que notre méthodologie hiérarchique bayésienne synthétique a permis de lisser considérablement les estimations directes fondées sur le plan. Il s'agit d'un phénomène bien connu dans la littérature; voir Rao et Molina (2015).

Ensuite, on compare nos estimations basées sur le modèle hiérarchique bayésien au niveau de l'état aux estimations pondérées par les poids de l'étude UAS au niveau de l'état. Dans les figures 4.2 et 4.3, on peut observer que, même si les estimations de l'étude UAS pour les états ayant un échantillon de taille adéquate sont, de manière générale, fiables, les estimations pour les états ayant un échantillon de petite taille (comme l'Alaska, le Delaware, le Rhode Island et le Vermont) montrent des valeurs insensées et des erreurs-types fondées sur le plan extrêmement élevées. En revanche, notre méthode hiérarchique bayésienne produit des estimations stables même dans le cas des états ayant une représentation minimale ou nulle sur le plan de l'échantillon. La figure 4.2 affiche les résultats sans la procédure de correction du biais décrite dans la section 2.3, alors que la figure 4.3 donne les résultats après correction du biais. Les deux tracés fournissent aussi les intervalles de crédibilité de 95 % correspondants pour les estimations hiérarchiques bayésiennes. Même si les estimations ponctuelles restent semblables avec et sans correction du biais, les intervalles de confiance pour les estimations dont le biais a été corrigé reflètent mieux l'incertitude supplémentaire inhérente au processus d'estimation pour les domaines pour lesquels les renseignements directs issus de l'échantillon sont limités.

Tableau 4.6
Statistiques sommaires *a posteriori* pour les paramètres du modèle 3, soit le meilleur modèle choisi en fonction du critère d'exclusion d'une unité

	Moyenne	Écart-type	10 %	15 %	85 %	90 %	N_{eff}	$\hat{R}$
α_1	-2,457	0,888	-3,608	-3,361	-1,566	-1,349	199,651	1,002
α_2	-1,232	0,870	-2,345	-2,132	-0,357	-0,141	201,311	1,003
α_3	-0,916	0,878	-2,078	-1,843	-0,034	0,159	187,865	1,001
α_4	-1,422	0,872	-2,540	-2,347	-0,545	-0,316	210,060	1,003
α_5	-1,391	0,885	-2,522	-2,326	-0,500	-0,294	198,465	1,002
α_6	-1,997	0,896	-3,157	-2,932	-1,096	-0,905	221,261	1,002
α_7	-1,121	0,979	-2,385	-2,140	-0,113	0,116	245,541	1,002
α_8	-1,574	1,153	-3,044	-2,762	-0,368	-0,154	302,163	1,002
Comorbidité	-0,573	0,682	-1,439	-1,273	0,109	0,280	202,383	1,002
Poids de sondage	-0,112	0,040	-0,161	-0,153	-0,070	-0,060	1 040,340	1,000
ξ_{homme}	0,209	0,019	0,184	0,189	0,229	0,233	1 122,518	1,001
ξ_{femme}	0,118	0,020	0,092	0,096	0,139	0,144	952,271	1,001
% rép.	-0,833	0,575	-1,571	-1,427	-0,248	-0,103	385,690	1,001
σ_v	0,121	0,072	0,026	0,035	0,196	0,215	31,222	1,039
σ_u	0,162	0,048	0,108	0,116	0,212	0,228	29,895	1,059

Note : « % rép. » = pourcentage de vote pour les républicains.

Dans la figure 4.4, on peut voir les améliorations qu'apportent nos estimations hiérarchiques bayésiennes par rapport aux estimations directes de l'étude UAS en ce qui concerne les erreurs-types estimées des estimations du degré de réticence. Les améliorations les plus importantes ont été observées pour les états dont la taille de l'échantillon était plus petite en ce qui concerne les données de l'étude UAS. On peut tout de même observer des estimations d'erreurs-types plus petites même pour les états dont l'échantillon est plus grand.

Le tableau 4.7 présente les estimations de la réticence à se faire vacciner au niveau de l'état à l'aide de trois approches différentes : l'approche hiérarchique bayésienne sans correction de biais, l'approche hiérarchique bayésienne avec correction de biais (comme elle est décrite dans la section 2.3) et l'approche des

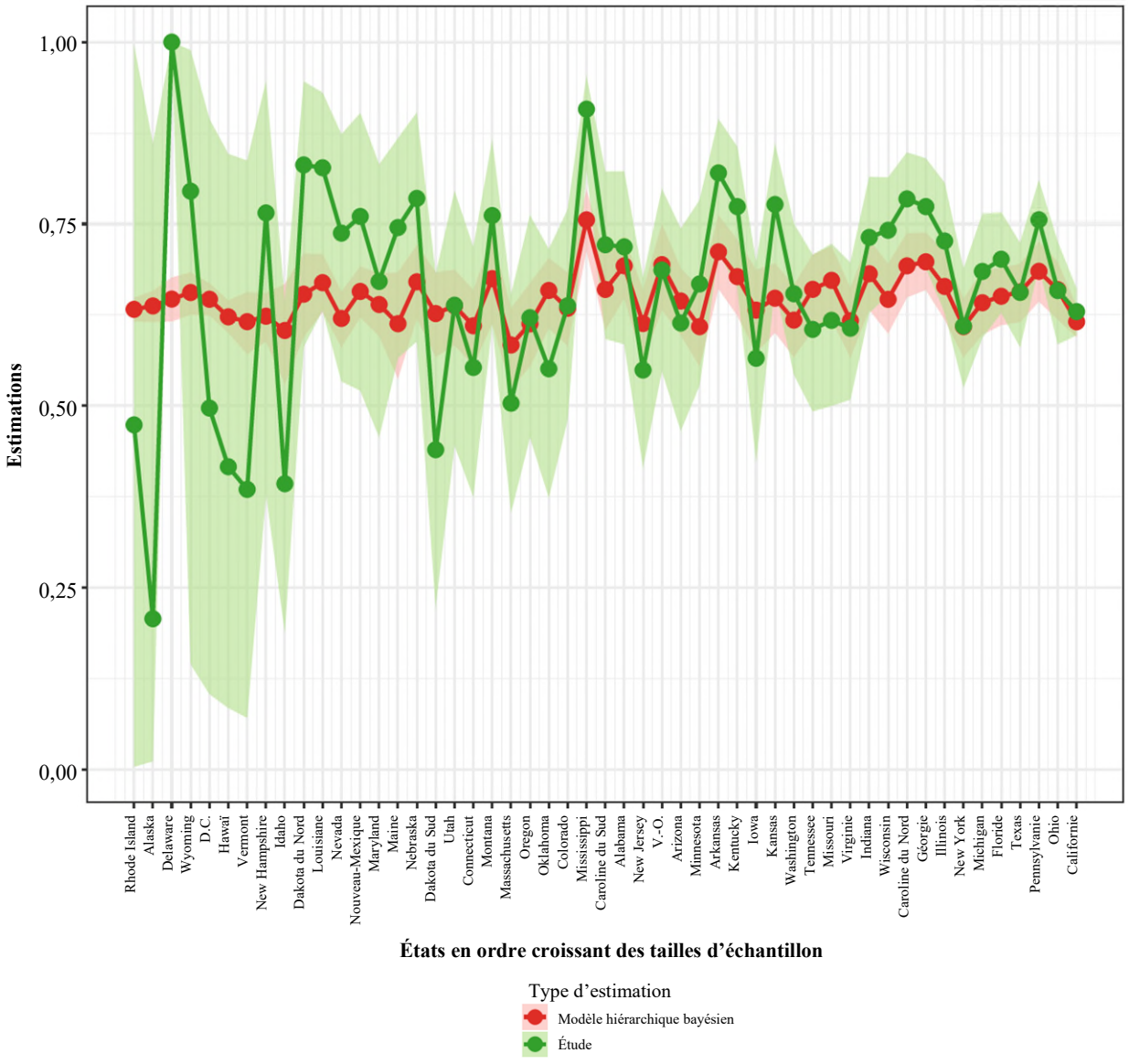
estimations directes fondées sur le plan de l'étude UAS. Dans le cas de chaque estimation, l'erreur-type correspondante est fournie entre parenthèses. La comparaison a révélé plusieurs tendances importantes. Tout d'abord, la correction de biais a des répercussions considérables sur les états dont la taille de l'échantillon est petite, comme le Delaware ($n = 5,\ 0{,}751 \rightarrow 0{,}632$), le district de Columbia ($n = 7,\ 0{,}701 \rightarrow 0{,}527$) et Hawaï ($n = 7,\ 0{,}573 \rightarrow 0{,}695$). Ensuite, les états ayant un échantillon minime affichent des augmentations plus élevées des erreurs-types après correction du biais, ce qui montre de manière appropriée l'incertitude supplémentaire inhérente à l'estimation synthétique (voir également la figure 4.3). Enfin, les deux approches hiérarchiques bayésiennes fournissent des estimations beaucoup plus stables que les estimations directes d'enquête pour les états ayant un échantillon de petite taille, comme le montrent les erreurs-types extrêmement élevées des estimations directes pour certains états, comme le Rhode Island (erreurtype = 0,321), le Vermont (erreur-type = 0,204) et Hawaï (erreur-type = 0,203). Le tableau souligne des améliorations particulièrement importantes pour l'Alaska, pour lequel l'estimation directe de 0,207 avec une erreur-type de 0,187 est remplacée par une estimation plus plausible après correction du biais de 0,663 avec une erreur-type de 0,060, et pour le Delaware, pour lequel l'estimation directe invraisemblable de 1,000 avec une erreur-type de 0,000 est remplacée par 0,632 avec une erreur-type de 0,051.

Tableau 4.7
Estimations hiérarchiques bayésiennes du degré de réticence (avec et sans correction de biais) et estimations directes fondées sur le plan de la Understanding America Study ainsi que leur erreur-type estimée pour les 50 états et le district de Columbia

			Estimations relatives à la réticence (ET)						Estimations relatives à la réticence (ET)		
Nº	État	n	Sans corr. de biais	Avec corr. de biais	Estimations directes	Nº	État	n	Sans corr. de biais	Avec corr. de biais	Estimations directes
1	Alabama	73	0,721 (0,021)	0,705 (0,028)	0,718 (0,060)	26	Missouri	117	0,686 (0,027)	0,679 (0,027)	0,617 (0,057)
2	Alaska	5	0,611 (0,022)	0,663 (0,060)	0,207 (0,187)	27	Montana	52	0,689 (0,031)	0,691 (0,032)	0,762 (0,063)
3	Arizona	82	0,668 (0,024)	0,670 (0,028)	0,614 (0,072)	28	Nebraska	37	0,643 (0,032)	0,689 (0,035)	0,785 (0,078)
4	Arkansas	93	0,752 (0,026)	0,728 (0,027)	0,820 (0,046)	29	Nevada	28	0,615 (0,023)	0,657 (0,036)	0,737 (0,085)
5	Californie	1 879	0,601 (0,009)	0,615 (0,011)	0,629 (0,016)	30	New Hampshire	9	0,576 (0,033)	0,616 (0,066)	0,765 (0,131)
6	Colorado	60	0,631 (0,027)	0,651 (0,033)	0,638 (0,074)	31	New Jersey	77	0,583 (0,026)	0,622 (0,030)	0,549 (0,067)
7	Connecticut	41	0,566 (0,030)	0,609 (0,033)	0,553 (0,088)	32	Nouveau-Mexique	33	0,669 (0,024)	0,678 (0,042)	0,760 (0,096)
8	Delaware	5	0,751 (0,026)	0,632 (0,051)	1,000 (0,000)	33	New York	178	0,628 (0,018)	0,616 (0,026)	0,610 (0,042)
9	District de Columbia	7	0,701 (0,016)	0,527 (0,062)	0,497 (0,219)	34	Caroline du Nord	157	0,719 (0,021)	0,708 (0,024)	0,784 (0,037)
10	Floride	241	0,662 (0,019)	0,659 (0,021)	0,701 (0,035)	35	Dakota du Nord	23	0,669 (0,042)	0,681 (0,036)	0,831 (0,086)
11	Géorgie	164	0,735 (0,019)	0,714 (0,022)	0,774 (0,038)	36	Ohio	281	0,689 (0,017)	0,665 (0,021)	0,658 (0,036)
12	Hawaï	7	0,573 (0,019)	0,695 (0,042)	0,416 (0,203)	37	Oklahoma	58	0,626 (0,028)	0,680 (0,034)	0,551 (0,089)
13	Idaho	23	0,506 (0,045)	0,626 (0,041)	0,393 (0,118)	38	Oregon	56	0,551 (0,034)	0,633 (0,034)	0,621 (0,079)
14	Illinois	176	0,674 (0,022)	0,679 (0,026)	0,726 (0,045)	39	Pennsylvanie	255	0,733 (0,015)	0,709 (0,023)	0,756 (0,030)
15	Indiana	152	0,687 (0,024)	0,687 (0,025)	0,731 (0,048)	40	Rhode Island	3	0,594 (0,019)	0,617 (0,075)	0,474 (0,321)
16	Iowa	94	0,602 (0,030)	0,643 (0,029)	0,565 (0,071)	41	Caroline du Sud	71	0,701 (0,023)	0,657 (0,028)	0,721 (0,058)
17	Kansas	107	0,676 (0,022)	0,676 (0,028)	0,777 (0,050)	42	Dakota du Sud	39	0,543 (0,036)	0,655 (0,035)	0,440 (0,122)
18	Kentucky	93	0,680 (0,026)	0,696 (0,029)	0,774 (0,049)	43	Tennessee	114	0,622 (0,026)	0,671 (0,028)	0,605 (0,055)
19	Louisiane	25	0,721 (0,028)	0,668 (0,034)	0,827 (0,071)	44	Texas	245	0,661 (0,015)	0,666 (0,022)	0,656 (0,037)
20	Maine	36	0,591 (0,044)	0,626 (0,038)	0,745 (0,076)	45	Utah	40	0,630 (0,034)	0,669 (0,037)	0,638 (0,090)
21	Maryland	34	0,670 (0,024)	0,631 (0,032)	0,671 (0,096)	46	Vermont	7	0,558 (0,037)	0,578 (0,066)	0,385 (0,204)
22	Massachusetts	56	0,531 (0,029)	0,579 (0,036)	0,504 (0,077)	47	Virginie	151	0,602 (0,022)	0,627 (0,026)	0,607 (0,048)
23	Michigan	192	0,701 (0,020)	0,649 (0,024)	0,685 (0,043)	48	Washington	110	0,613 (0,021)	0,647 (0,027)	0,654 (0,053)
24	Minnesota	83	0,674 (0,027)	0,623 (0,028)	0,667 (0,065)	49	Virginie-Occidentale	81	0,671 (0,026)	0,716 (0,032)	0,686 (0,064)
25	Mississippi	60	0,769 (0,024)	0,768 (0,024)	0,908 (0,031)	50	Wisconsin	155	0,644 (0,023)	0,666 (0,026)	0,741 (0,041)
						51	Wyoming	5	0,659 (0,028)	0,670 (0,070)	0,795 (0,184)

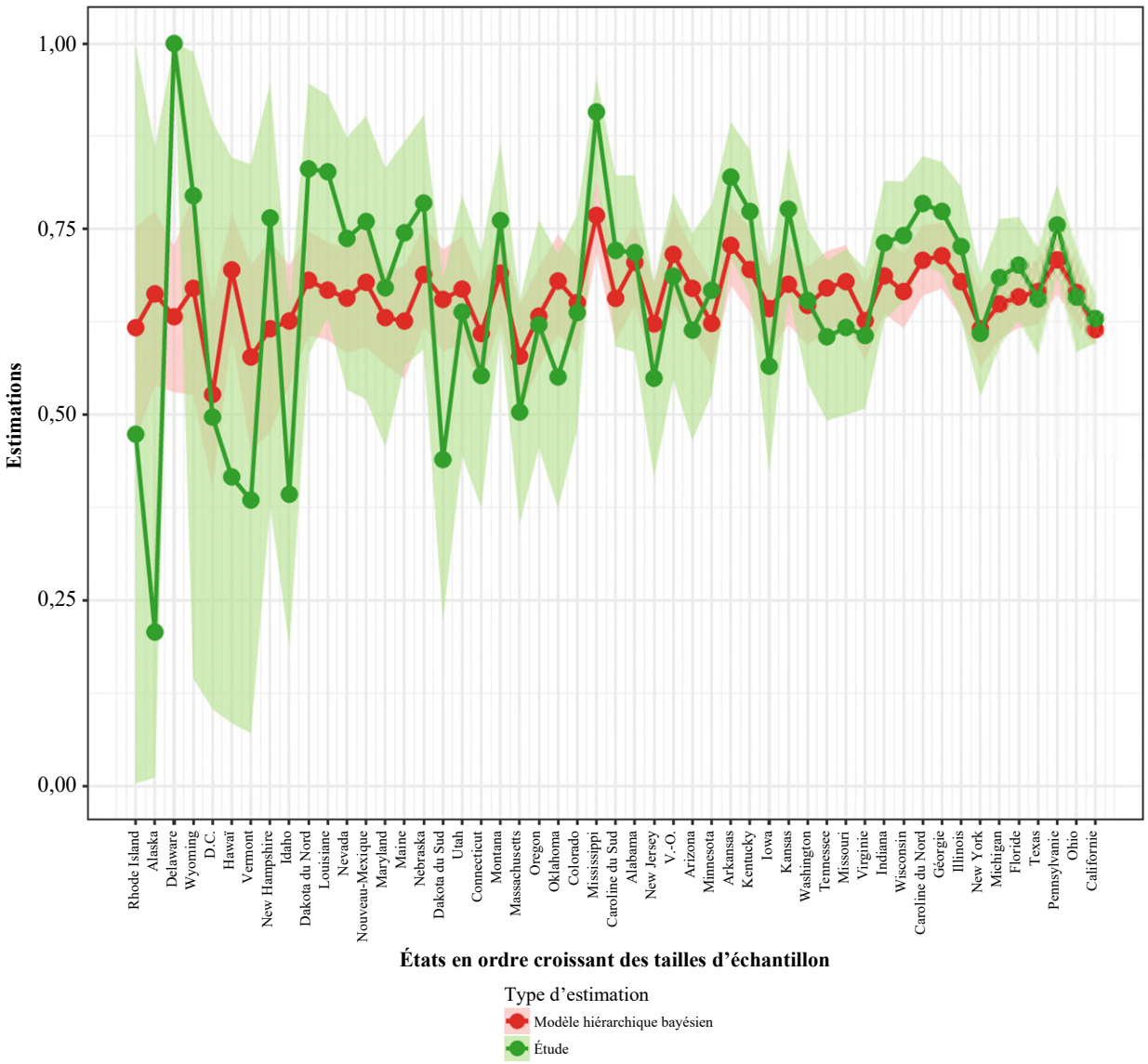
Notes : - n = taille de l'échantillon de la Understanding America Study; ET = erreur-type (estimée).
- On montre, entre parenthèses, les erreurs-types de toutes les estimations.

Figure 4.2 Comparaison du modèle hiérarchique bayésien par rapport aux estimations d'enquête avec des bandes de confiance



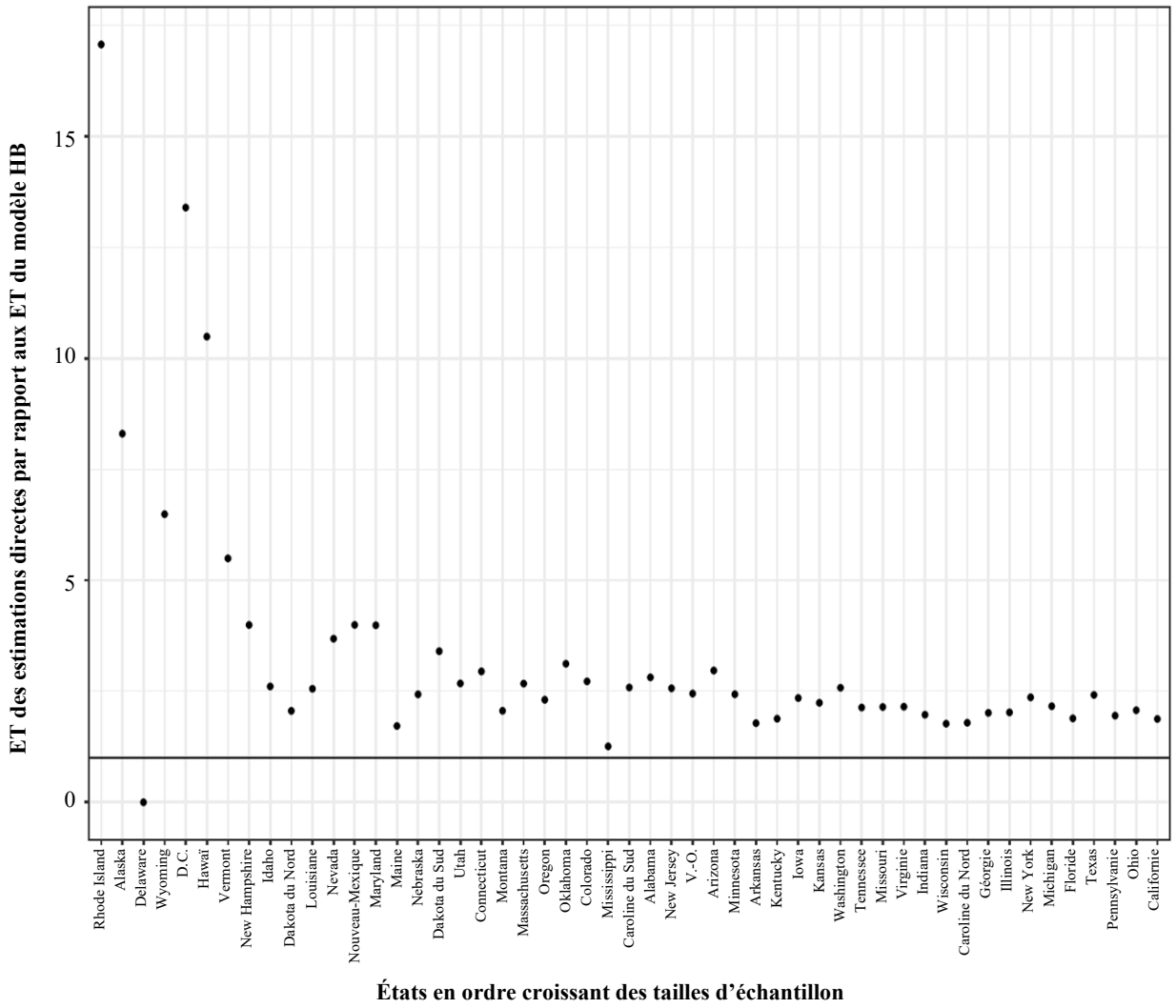
Note : Sur l'axe des X, on voit les états classés par ordre croissant des tailles d'échantillon de la Understanding America Study (UAS). Sur l'axe des Y, on observe les estimations ponctuelles de la proportion de personnes qui sont « réticentes » à se faire vacciner de l'étude UAS (en vert) et les estimations de notre modèle hiérarchique bayésien, le modèle 3 (en rouge), *sans* correction de biais, comme on l'a suggéré dans la section 2.3. Les intervalles de crédibilité de 95 % pour le modèle hiérarchique bayésien et les intervalles de confiance de 95 % pour les estimations directes sont illustrés en bandes.

Figure 4.3 Comparaison du modèle hiérarchique bayésien par rapport aux estimations d'enquête avec des bandes de confiance



Note : Sur l'axe des X, on voit les états classés par ordre croissant des tailles d'échantillon de la Understanding America Study (UAS). Sur l'axe des Y, on observe les estimations ponctuelles de la proportion de personnes qui sont « réticentes » à se faire vacciner de l'étude UAS (en vert) et les estimations de notre modèle hiérarchique bayésien, le modèle 3 (en rouge), *avec* correction de biais, comme on l'a suggéré dans la section 2.3. Les intervalles de crédibilité de 95 % pour le modèle hiérarchique bayésien et les intervalles de confiance de 95 % pour les estimations directes sont illustrés en bandes.

Figure 4.4



Notes: -Sur l'axe des X, on voit les états classés par ordre croissant des tailles d'échantillon de la Understanding America Study (UAS). Sur l'axe des Y, on voit des rapports d'erreurs-types des estimations directes fondées sur l'étude UAS par rapport à nos estimations du modèle hiérarchique bayésien. La ligne horizontale foncée coupe l'axe des Y à 1.
-ET = erreurs-types; HB = hiérarchique bayésien.

5. Conclusion

Pour conclure, dans le présent article, on présente une nouvelle procédure d'estimation hiérarchique bayésienne pour les moyennes de population finie sur petits domaines en ayant recours à une technique sophistiquée de couplage de données qui intègre une enquête primaire (probabiliste) à différentes sources de données, y compris des données d'une enquête non probabiliste réalisée sur les médias sociaux. Dans notre modèle élaboré, qui tient uniquement compte des répondants figurant dans l'échantillon, on ajoute des poids de l'enquête primaire et des facteurs susceptibles d'influer sur la variable dépendante d'intérêt, afin

de réduire la valeur informative de l'échantillon. Si l'analyse des données appliquée indique que les poids sont significatifs dans le modèle, cela voudrait dire que l'échantillonnage est informatif. Il ne faut plus supposer la validité du même modèle pour les répondants et la population. Contrairement à de nombreux travaux existants, on reconnaît cette étape essentielle et l'on en tient compte, en supposant un modèle simple pour la population finie, sans égard à la valeur informative de l'échantillonnage. L'estimation hiérarchique synthétique obtenue a été conçue pour réduire la vulnérabilité aux erreurs de spécification du modèle, en comptant moins sur les modèles supposés par rapport aux approches pleinement fondées sur un modèle. Cela est rendu possible sur le plan conceptuel grâce à notre approche consistant à ne supposer qu'un modèle élaboré pour les unités observées, tout en utilisant une approche synthétique pour les unités non observées. Bien que l'évaluation empirique exhaustive de cet avantage théorique nécessite des études par simulation approfondies qui dépassent la portée du présent article, notre plan méthodologique offre intrinsèquement une certaine protection contre les erreurs de spécification du modèle en raison de sa nature hybride. En utilisant les mêmes notations que celles utilisées dans la section 2, on démontre cette dépendance réduite en ignorant l'ensemble de cellules $\mathcal{G}_{3i}$ au lieu d'utiliser le modèle supposé pour estimer la moyenne de population correspondante Θ_{3i}. S'inspirant de travaux précurseurs, notre méthodologie est polyvalente et peut s'appliquer à divers défis liés à l'estimation de la moyenne d'une population finie. Cependant, il faut souligner que notre approche pourrait ne pas produire d'estimations pour petits domaines dans des scénarios précis exigeant la validation d'une condition cruciale avant de procéder à l'analyse des données. Encore une fois, en utilisant les mêmes notations, dans le cas de données d'enquête probabiliste, il est essentiel de respecter la condition $(1-a_{1i}-a_{2i}) \approx 0.$ Il faut vérifier cette condition avant de procéder à l'analyse des données. Enfin, on illustre l'application de notre méthodologie en produisant des estimations de la proportion (moyenne de petits domaines d'une variable dépendante binaire) d'une enquête sur la COVID-19 couplant statistiquement des renseignements tirés de différentes sources de données. Des extensions futures du présent article pourraient permettre d'évaluer des procédures d'estimation semblables pour modéliser conjointement les variables réponses catégoriques et continues. On évalue actuellement cette orientation au chapitre de la recherche.

Remerciements

La présente étude a été soutenue en partie par la subvention SES-1758808 de la National Science Foundation des États-Unis qui a été octroyée au deuxième auteur. Le projet décrit dans l'article repose sur des données d'une ou de plusieurs enquêtes menées dans le cadre de la Understanding America Study (UAS), qui est réalisée par le Center for Economic and Social Research de l'Université de Californie du Sud. Le contenu de l'article relève uniquement de la responsabilité des auteurs et ne représente pas nécessairement les opinions officielles de l'Université de Californie du Sud ou des responsables de l'étude UAS. La collecte de données de suivi sur la COVID-19 de l'étude UAS a été soutenue en partie par la Bill and Melinda Gates Foundation et par la subvention U01AG054580 de la National Institute on Aging. La

présente étude repose sur des données d'enquête du Delphi Group de l'Université Carnegie Mellon obtenues dans le cadre d'un échange de données. Pour en savoir davantage sur l'enquête, veuillez consulter le site Web suivant : https://cmu-delphi.github.io/delphi-epidata/symptom-survey/.

Bibliographie

Arora, V., Lahiri, P. et Mukherjee, K. (1997). Empirical bayes estimation of finite population means from complex surveys. *Journal of the American Statistical Association*, 92(440), 1555-1562.

Bonnéry, D., Breidt, F.J. et Coquet, F. (2012). Uniform convergence of the empirical cumulative distribution function under informative selection from a finite population. *Bernoulli*, 18(4), 1361-1385.

Chen, Q., Elliott, M.R. et Little, R.J.A. (2012). Inférence bayésienne pour les quantiles de population finie sous échantillonnage avec probabilités inégales. *Techniques d'enquête*, 38(2), 221-233. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2012002/article/11758-fra.pdf.

Datta, G.S. (2009). Model-based approach to small area estimation. *Handbook of Statistics*, 29, 251-288.

Datta, G.S. et Ghosh, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. *The Annals of Statistics*, 19(4), 1748-1770.

Datta, G.S. et Ghosh, M. (1995). Hierarchical bayes estimators of the error variance in one-way anova models. *Journal of Statistical Planning and Inference*, 45(3), 399-411.

Ericson, W.A. (1969). Subjective bayesian models in sampling finite populations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(2), 195-224.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. et Rubin, D.B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A., Jakulin, A., Pittau, M.G. et Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4), 1360-1383.

Gelman, A. et Little, T.C. (1997). Stratification a posteriori en un grand nombre de catégories par régression logistique hiérarchique. *Techniques d'enquête*, 23(2), 135-145. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1997002/article/3616-fra.pdf.

Ghosh, M. (2009). Bayesian developments in survey sampling. *Handbook of Statistics*, 29, 153-187. Elsevier.

Ghosh, M. (2020). Small area estimation: Its evolution in five decades. *Statistics in Transition New Series, Special Issue on Statistical Data Integration*, 1-67.

Ghosh, M. et Lahiri, P. (1987). Robust empirical bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 82(400), 1153-1162.

Ghosh, M. et Lahiri, P. (1992). A hierarchical bayes approach to small area estimation with auxiliary information. *Bayesian Analysis in Statistics and Econometrics*, 107-125. Springer.

Ghosh, M., Lahiri, P. et Tiwari, R.C. (1989). Nonparametric empirical bayes estimation of the distribution function and the mean. *Communications in Statistics-Theory and Methods*, 18(1), 121-146.

Ghosh, M. et Meeden, G. (1986). Empirical bayes estimation in finite population sampling. *Journal of the American Statistical Association*, 81(396), 1058-1062.

Ghosh, M. et Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall.

Ha, N.S. et Sedransk, J. (2019). Assessing health insurance coverage in florida using the behavioral risk factor surveillance system. *Statistics in Medicine*, 38(13), 2332-2352.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer.

Jiang, J. et Lahiri, P. (2006). Estimation of finite population domain means - a model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.

Jiang, J. et Nguyen, T. (2012). Small area estimation via heteroscedastic nested-error regression. *Canadian Journal of Statistics*, 40, 588-663.

Lahiri, P. et Suntornchost, J. (2020). A general bayesian approach to meet different inferential goals in poverty research for small areas. *Statistics in Transition, Special Issue*, 245-261.

Little, R.J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99(466), 546-556.

Liu, B. et Lahiri, P. (2017). Adaptive hierarchical bayes estimation of small area proportions. *Calcutta Statistical Association Bulletin*, 69(2), 150-164.

Malec, D., Sedransk, J., Moriarity, C.L. et LeClere, F.B. (1997). Small area inference for binary variables in the national health interview survey. *Journal of the American Statistical Association*, 92(439), 815-826.

Morris, C.N. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics*, 65-80.

Nandram, B., Chen, L., Fu, S. et Manandhar, B. (2018). Bayesian logistic regression for small areas with numerous households. *arXiv preprint arXiv:1806.00446*.

Nandram, B. et Sedransk, J. (1993). Empirical bayes estimation for the finite population mean on the current occasion. *Journal of the American Statistical Association*, 88(423), 994-1000.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28, 40-68.

Pfeffermann, D. et Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā: The Indian Journal of Statistics, Series B*, 166-186.

Pfeffermann, D. et Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102(480), 1427-1439.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienne, Autriche.

Rao, J.N.K. et Molina, I. (2015). *Small Area Estimation, 2nd Edition*. Wiley.

Rubin, D.B. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys: Comment. *Journal of the American Statistical Association*, 78(384), 803-805.

Scott, A.J. (1977). Some comments on the problem of randomisation in surveys. *Sankhyā C*, 39, 1-9.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2.

Stan Development Team (2023). *Stan Modeling Language Users Guide and Reference Manual*.

Stan Development Team (2024). *Stan: A Probabilistic Programming Language*. Version 2.36.0.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T. et Gelman, A. (2020). loo: Efficient leave-one-out cross-validation and waic for bayesian models. R package version 2.4.1.

Vehtari, A., Gelman, A. et Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5), 1413-1432.

Verret, F., Rao, J.N.K. et Hidiroglou, M.A. (2015). Estimation sur petits domaines fondée sur un modèle sous échantillonnage informatif. *Techniques d'enquête*, 41(2), 353-368. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2015002/article/14248-fra.pdf.