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

Techniques d’enquête

Combinaison de données provenant d’enquêtes et de sources connexes

par Dexter Cahoy et Joseph Sedransk

Date de diffusion : le 30 juin 2023

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, juin 2023
Vol. 49, nº 1, pp. 75-96
Statistique Canada, nº 12-001-X au catalogue

Combinaison de données provenant d'enquêtes et de sources connexes

Dexter Cahoy et Joseph Sedransk[1]

Résumé

Pour accroître la précision des inférences et réduire les coûts, la combinaison de données provenant de plusieurs sources comme les enquêtes-échantillon et les données administratives suscite beaucoup d'intérêt. Une méthodologie appropriée est requise afin de produire des inférences satisfaisantes, puisque les populations cibles et les méthodes d'acquisition de données peuvent être assez différentes. Pour améliorer les inférences, nous utilisons une méthodologie qui a une structure plus générale que celles de la pratique actuelle. Nous commençons par le cas où l'analyste ne dispose que de statistiques sommaires provenant de chacune des sources. Dans la méthode principale, la combinaison incertaine, on suppose que l'analyste peut considérer une source, l'enquête *r*, comme étant de loin le meilleur choix pour l'inférence. Cette méthode part des données de l'enquête *r* et ajoute les données provenant des sources tierces, pour former des grappes qui comprennent l'enquête *r*. Nous considérons également les mélanges selon le processus de Dirichlet, l'une des méthodes bayésiennes non paramétriques les plus populaires. Nous utilisons des expressions analytiques et les résultats d'études numériques pour montrer les propriétés de la méthodologie.

Mots-clés : Données administratives; méthodes bayésiennes; mise en grappe; mélange selon le processus de Dirichlet; combinaison de données; enquêtes par sondage.

1. Introduction

Les taux de réponse considérablement réduits et les budgets limités nécessitent qu'une importance accrue soit accordée à l'utilisation efficace de tous les renseignements dont dispose l'analyste d'enquête. Plus précisément, il serait possible d'améliorer les inférences en utilisant les résultats provenant de plusieurs enquêtes-échantillons et de sources connexes comme les dossiers administratifs. La méthodologie que nous utilisons pour combiner les renseignements est plus structurée que les méthodes actuellement utilisées dans les enquêtes par sondage et devrait donc produire de meilleures inférences. En commençant par les données provenant de l'enquête qui constituent le meilleur choix pour l'inférence, nous complétons celles-ci par d'autres données concordantes. Nous utilisons des expressions analytiques et les résultats d'études numériques pour montrer les propriétés de la méthodologie.

La présente recherche fait suite à une étude (Ha et Sedransk, 2019) portant sur la couverture par l'assurance maladie dans les comtés de Floride où les auteurs ont noté des estimations très différentes établies provenant de trois enquêtes. En conséquence, nous pourrions avoir des estimations d'une enquête probabiliste bien établie et de deux enquêtes non probabilistes. Dans les deux cas, on s'interroge sur la façon de produire de meilleures inférences.

Dans la foulée, nous désignons la série d'études par le mot « enquêtes », en reconnaissant qu'il peut s'agir d'enquêtes probabilistes, d'enquêtes non probabilistes, de dossiers administratifs et d'autres sources. Nous étudions le cas où l'analyste ne dispose que d'une estimation ponctuelle et de l'erreur-type qui lui est

1. Dexter Cahoy, University of Houston-Downtown; Joseph Sedransk, University of Maryland. Courriel : jxs123@case.edu.

associée pour chaque enquête. Il s'agit d'une situation courante, comme l'indique la section 7 du document d'examen de Lohr et Raghunathan (2017). Dans l'exemple motivant ces travaux, et dans de nombreux autres cas, aucune covariable ne peut être utilisée pour améliorer les inférences. Notre méthodologie s'étend aux cas où les objectifs et les modèles inférentiels sont plus complexes.

Dans le cas des estimations d'enquête, $\left\{\hat{Y}_i : i = 1, \ldots, L\right\}$, nous supposons généralement que $\hat{Y}_i$ sont indépendantes :

$$\hat{Y}_i \sim N\left(\mu_i, V_i\right) \tag{1.1}$$

où V_i sont supposées connues.

Une distribution *a priori* commune exprimant la similarité entre $\left\{\mu_1, \ldots, \mu_L\right\}$ est :

$$\mu_i \,|\, \nu, \delta^2 \sim N\left(\nu, \delta^2\right) \tag{1.2}$$

indépendamment pour chaque i. ν et δ sont des distributions *a priori* uniformes localement.

La valeur attendue *a posteriori* résultante de μ_i est une combinaison convexe de l'estimation, $\hat{Y}_i$, pour la présente enquête, et d'une moyenne pondérée de $\left\{\hat{Y}_1, \ldots, \hat{Y}_L\right\}$. La faiblesse de cette approche tient au fait que la distribution *a priori* dans (1.2) suppose un échantillonnage indépendant des variables μ_i à partir d'une distribution commune. La moyenne *a posteriori* est :

$$E\left(\mu_i \,|\, \hat{Y}_1, \ldots, \hat{Y}_L\right) = \lambda_i \hat{Y}_i + \left(1 - \lambda_i\right) \hat{Y}_w \tag{1.3}$$

où $\lambda_i = \delta^2 \,/\left(\delta^2 + V_i\right)$ et $\hat{Y}_w = \sum_{i=1}^{L} \lambda_i \hat{Y}_i \Big/ \sum_{i=1}^{L} \lambda_i$. Cela peut donner lieu à des inférences insatisfaisantes lorsque, par exemple, les valeurs de $\mu_1, \ldots, \mu_b$ sont chacune proche de celle de μ^*, tandis que les valeurs de $\mu_{b+1}, \ldots, \mu_L$ sont chacune proche de celle de μ^{**} et $\mu^* << \mu^{**}$. Dans le cas présent, l'estimation de μ_1 comprendrait, peut-être de manière inappropriée, une contribution importante de $\hat{Y}_{b+1}, \ldots, \hat{Y}_L$. La difficulté est que la distribution *a priori* n'est pas suffisamment souple. Nous utilisons des distributions *a priori* plus souples qui permettent de déterminer le niveau et la nature de la combinaison à partir des données de l'échantillon.

La spécification dans (1.1) et (1.2) est courante dans les méta-analyses et dans les situations où l'on souhaite faire des inférences ayant trait à de petites sous-populations et à de petites régions géographiques. Par exemple, le U.S. Census Bureau s'appuie sur ces modèles (ajustés au moyen de termes pour tenir compte des covariables) pour faire des inférences pour les taux de pauvreté au niveau des comtés américains : voir l'exemple 6.1.2 dans Rao et Molina (2015). Cependant, comme nous venons de le mentionner, il est possible que les hypothèses de (1.1) et (1.2) portant sur l'échangeabilité complète ne soient pas appropriées, en particulier lorsqu'il s'agit de combiner les renseignements tirés des enquêtes L.

La théorie de notre méthode principale, la combinaison incertaine, a été formulée par Malec et Sedransk (1992) et Evans et Sedransk (2001) et des travaux supplémentaires figurant dans Evans et Sedransk (1999) et Yan et Sedransk (2011). À notre connaissance, cette méthodologie n'a pas été utilisée dans une

application d'échantillonnage d'enquête : le rapport d'examen exhaustif de Lohr et Raghunathan (2017) ne fait pas mention d'une quelconque technique semblable à la nôtre. Nous modifions également (1.1) et (1.2) en utilisant le mélange selon le processus de Dirichlet, comme celui employé par Polettini (2017) pour l'inférence sur petits domaines.

Dans le présent article, nous décrivons la méthodologie pour la combinaison incertaine et le mélange selon le processus de Dirichlet et nous les utilisons pour analyser les données tirées de Ha et Sedransk (2019). Nous avons modifié ces données pour montrer les propriétés des méthodes. Une étude par simulations a fini par être réalisée pour évaluer les propriétés d'échantillonnage. De toute évidence, les résultats de cette évaluation s'appliqueraient de la même manière si les trois sources étaient, par exemple, une enquête probabiliste bien établie et deux enquêtes non probabilistes ou s'il s'agissait d'autres sources de données.

Nous supposons que les variances d'échantillon $\{V_1, \ldots, V_L\}$ sont précisées. Dans notre contexte, aucune des solutions de rechange décrites dans les ouvrages spécialisés pour faire des inférences pour les variances $\{V_1, \ldots, V_L\}$, qui sont toutes basées sur des inférences pour de petits échantillons, n'est pleinement satisfaisante. Dans la section 4, nous discutons de ce problème difficile de faire des inférences dans le cas des variances des échantillons.

La combinaison incertaine et le mélange selon le processus de Dirichlet ont une structure plus générale que celle de la spécification commune, (1.1) et (1.2). Cela devrait aboutir à de meilleures inférences. Comme nous l'avons vu à la section 2, le modèle de combinaison incertaine devient un prolongement en soi de (1.1) et de (1.2), c'est-à-dire que le modèle dans (1.1) et (1.2) est un cas particulier. De plus, les résultats de la combinaison incertaine comprennent les probabilités *a posteriori* associées à la possible mise en grappes des L enquêtes.

Il convient de souligner que nous n'abordons qu'un seul des nombreux aspects de la « combinaison de données d'enquête », qui sont bien résumés par Lohr et Raghunathan (2017). La section 7 de leur article, « Hierarchical models for combining data sources », donne des exemples supplémentaires où notre méthodologie peut être utile.

La méthodologie que nous suivons est décrite à la section 2 et les résultats de nos études numériques sont résumés à la section 3. Un bref résumé et une discussion figurent dans la section 4.

2. Méthodologie

Comme à la section 1, nous supposons qu'il existe L estimations d'enquête, $\hat{Y}_1, \ldots, \hat{Y}_L$, comportant :

$$\hat{Y}_i \overset{\text{ind}}{\sim} N(\mu_i, V_i) \tag{2.1}$$

où V_i sont supposées connues.

2.1 Combinaison incertaine

La méthode de combinaison incertaine est fondée sur les travaux de Malec et Sedransk (1992) et Evans et Sedransk (2001). Ceux-ci ont montré qu'on peut choisir une loi *a priori* $\mu = (\mu_1, \mu_2, \ldots, \mu_L)^t$ pour refléter les croyances selon lesquelles il existe des sous-ensembles de μ de façon à ce que les μ_i dans chaque sous-ensemble soient « similaires » et qu'il existe une incertitude au sujet de la composition de ces sous-ensembles de μ. Soit G le nombre total de partitions de l'ensemble $\mathcal{L} = \{1, \ldots, L\}$, g une partition particulière $(g = 1, \ldots, G)$, $d(g)$ le nombre de sous-ensembles de $\mathcal{L}$ dans la g^{e} partition $(1 \le d(g) \le L)$ et $S_k(g)$ l'ensemble d'étiquettes dans le sous-ensemble $k\big(k = 1, \ldots, d(g)\big)$. Par exemple, pour $L = 3$, il existe $G = 5$ partitions : $\{g=1\} \sim \{(123)\}, \{g=2\} \sim \{(13),(2)\}, \{g=3\} \sim \{(12),(3)\}, \{g=4\} \sim \{(23),(1)\}$, $\{g=5\} \sim \{(1),(2),(3)\}$. Alors, $S_1(g=2) = \{(13)\}$, $S_2(g=2) = \{(2)\}$, $d_1(g=2) = 2$ et $d_2(g=2) = 1$.

Pour choisir une loi *a priori* pour μ, nous devons d'abord conditionner sur g, Malec et Sedransk (1992) et Evans et Sedransk (2001) supposent que les sous-ensembles sont indépendants et qu'au sein des sous-ensembles $S_k(g)$, les μ_i sont indépendantes et comportent ce qui suit :

$$\mu_i \,|\, \nu_k(g) \sim N\big(\nu_k(g), \delta_k^2(g)\big), \qquad i \in S_k(g). \tag{2.2}$$

Par ailleurs, les $\nu_k(g)$ sont mutuellement indépendantes compte tenu de :

$$\nu_k(g) \,|\, \theta_k(g) \sim N\big(\theta_k(g), \gamma_k^2(g)\big) \tag{2.3}$$

où $\theta_k(g)$ et $\gamma_k^2(g)$ sont des hyperparamètres. La définition dans (2.3) est la première étape menant à l'obtention d'une distribution *a priori* de référence pour $\nu_k(g)$, c'est-à-dire, une distribution *a priori* qui est dominée par la fonction de vraisemblance. Cela laisse notamment supposer que $\gamma_k^2(g) \to \infty$, mais la tâche est beaucoup plus compliquée, comme nous le décrivons ci-dessous. Les $\delta_k^2(g)$ sont aussi des hyperparamètres, auxquels il faut attribuer une distribution *a priori*.

La définition officielle dans (2.3) est la première étape menant à l'obtention d'une distribution *a priori* de référence pour $\nu_k(g)$, c'est-à-dire une distribution *a priori* qui est dominée par la vraisemblance, décrite ci-dessous dans l'évaluation de $f\big(g, \Delta^2 \,|\, y\big)$. Au moyen du conditionnement sur les échantillons $\delta_k^2(g)$ et $\gamma_k^2(g)$ (mais en les supprimant dans notre notation), et en supposant que $\gamma_k^2(g) \to \infty$, nous obtenons les résultats attendus suivants pour les moments *a posteriori* pour la partition g. Comme nous l'expliquons ci-dessous, il importe de faire preuve d'une grande rigueur pour obtenir la distribution *a posteriori* de g.

En définissant $y = \big(\hat{Y}_1, \ldots, \hat{Y}_L\big)^t$, en désignant $\Delta^2 = \big\{\delta_k^2(g) : k = 1, \ldots, d(g); g = 1, \ldots, G\big\}$ et en écrivant $\hat{\mu}_i = \hat{Y}_i$:

$$E\big(\mu_i \,|\, y, g, \Delta^2\big) = \big\{\lambda_i(g)\big\}\, \hat{\mu}_i + \big\{1 - \lambda_i(g)\big\}\, \hat{\mu}_k(g), \qquad i \in S_k(g) \tag{2.4}$$

et

$$\operatorname{cov}\left(\mu_i, \mu_j \mid y, g, \Delta^2\right) = \begin{cases} \delta_k^2(g)\left\{1-\lambda_i(g)\right\} + \dfrac{\left\{1-\lambda_i(g)\right\}^2 \delta_k^2(g)}{\sum_{i\in S_k(g)} \lambda_i(g)}, & i=j; i, j \in S_k(g)) \\ \dfrac{\left\{1-\lambda_i(g)\right\}\left\{1-\lambda_j(g)\right\}\delta_k^2(g)}{\sum_{i\in S_k(g)} \lambda_i(g)}, & i\neq j; i, j \in S_k(g) \\ 0, & i\in S_{k_1}(g), j\in S_{k_2}(g), k_1\neq k_2, \end{cases} \tag{2.5}$$

où

$$\lambda_i(g) = \frac{\delta_k^2(g)}{\delta_k^2(g)+V_i}, \hat{\mu}_k(g) = \frac{\sum_{j\in S_k(g)} \lambda_j(g)\hat{\mu}_j}{\sum_{j\in S_k(g)} \lambda_j(g)}. \tag{2.6}$$

Il convient de souligner que $E\left(\mu_i \mid y, g, \Delta^2\right)$ prend la forme familière d'une moyenne pondérée de $\hat{\mu}_i$ et de $\hat{\mu}_k(g),$ mais, ici, $\hat{\mu}_k(g)$ se limite aux enquêtes dans $S_k(g).$

Supposons que le modèle de base dans (1.1) et (1.2) correspond ici à la partition de l'échantillon regroupé $\{g=1\},$ où toutes les L enquêtes constituent une seule grappe. Ainsi, pour $\{g=1\},$ les moments dans (2.4), (2.5) et (2.6) sont ceux qui seraient obtenus à partir de l'analyse au moyen de (1.1) et de (1.2). Une analyse reposant sur (1.1) et (1.2) est un cas particulier d'une analyse aux fins de spécification de l'échantillonnage incertain.

L'inférence au sujet de μ comprend l'incertitude au sujet de la valeur de $g,$ c'est-à-dire :

$$f\left(\mu \mid y\right) = \int f\left(\mu \mid y, g, \Delta^2\right) f\left(g, \Delta^2 \mid y\right) dg d\Delta^2 \tag{2.7}$$

où la notation est simplifiée en utilisant l'intégration plutôt que la somme pour $g.$ L'utilisation de la partition « la plus probable » g^* (c'est-à-dire, $p\left(g^* \mid y\right) \geq p\left(g \mid y\right): g=1,\ldots,G)$ à des fins d'inférence entraînerait une sous-estimation de la précision globale.

Pour évaluer (2.7), nous avons besoin de $f\left(g, \Delta^2 \mid y\right).$ Cependant, lors de l'évaluation de $f\left(g \mid \Delta^2, y\right)$ il faut se montrer prudent quant à la spécification de la vitesse à laquelle $\gamma_k^2(g) \to \infty$: un choix naturel mène à une expression pour $f\left(g \mid \Delta^2, y\right)$ qui n'est pas invariante en cas de changements apportés à l'échelle de $Y;$ voir la section 4 de Malec et Sedransk (1992). Malec et Sedransk (1992) ont proposé une solution en utilisant un argument bayésien empirique. Ici, nous utilisons une autre approche entièrement bayésienne, décrite dans la section 5 de Evans et Sedransk (2001). Cette approche est invariante en cas de changements apportés à l'échelle de Y et son application est indiquée quand on ne dispose que de peu de renseignements *a priori* sur l'échantillon $\nu_k(g).$ Supposons que $\nu(g) = \left(\nu_1(g), \ldots, \nu_{d(g)}(g)\right)^t$ et $K\left\{f_1\left(\nu(g)\right), f_2\left(\nu(g) \mid y\right)\right\}$ est l'information de Kullback-Leibler sur $\nu(g).$ Compte tenu de $f\left(g, \Delta^2\right) = f(g) f\left(\Delta^2\right)$ *a priori* et en supposant que $\gamma_k^2(g) \to \infty$ sous la contrainte que $K\left\{f_1\left(\nu(g)\right), f_2\left(\nu(g) \mid y\right)\right\} = \text{constant},$

$$f\left(g, \Delta^2 \mid y\right) \propto f\left(\Delta^2\right) f(g) \exp\left\{\frac{-d(g)}{2}\right\} \prod_{k=1}^{d(g)} \prod_{i\in S_k(g)} \left\{1-\lambda_i(g)\right\}^{1/2} \times \exp\left[-\frac{1}{2}\sum_{k=1}^{d(g)} \sum_{i\in S_k(g)} \left\{\frac{\lambda_i(g)}{\delta_k^2(g)}\right\} \left\{\hat{\mu}_i - \hat{\mu}_k(g)\right\}^2\right]. \tag{2.8}$$

La valeur de l'exposant :

$$Q\{d(g)\} = \sum_{k=1}^{d(g)} \sum_{i \in S_k(g)} \left\{ \frac{\lambda_i(g)}{\delta_k^2(g)} \right\} \{\hat{\mu}_i - \hat{\mu}_k(g)\}^2, \tag{2.9}$$

devrait diminuer à mesure que $d(g)$ augmente, par exemple, pour une nouvelle partition $\bigcup_{k=1}^{d(g)} S_k(g)$ obtenue en créant des sous-ensembles de l'ensemble existant $\{S_k(g)\}$. Puisque $f(g, \Delta^2 \mid y)$ augmente à mesure que $Q\{d(g)\}$ diminue, il est utile d'avoir le deuxième terme, $\exp\{-d(g)/2\}$, qui pénalise les partitions ayant de grandes valeurs $d(g)$.

Pour notre analyse, nous posons que $\delta_k^2(g) = \delta^2$ et que $\lambda_i(g) = \delta^2 / (\delta^2 + V_i)$. L'inférence au sujet de μ est calculée à l'aide de (2.7) et de (2.8) par :

$$\mu \mid y, g, \delta^2 \sim N\left(E(\mu \mid y, g, \delta^2), V(\mu \mid y, g, \delta^2)\right) \tag{2.10}$$

où les moments *a posteriori* de μ sont donnés dans (2.4) et (2.5).

Nous supposons que $f(g)$ est constante, c'est-à-dire que toutes les partitions ont une probabilité égale, *a priori*, de se produire et nous appliquons une distribution inverse *a priori* bêta à δ^2, c'est-à-dire :

$$f(\delta^2) \propto 1 / (1 + \delta^2)\sqrt{\delta^2}, \quad 0 < \delta^2 < \infty. \tag{2.11}$$

L'inférence au sujet de μ est calculée à l'aide de (2.7). Pour commencer, évaluons le côté droit de (2.8) pour :

$$\{g : g = 1, \ldots, G; \quad R \text{ points de la grille pour } \delta^2\}, \tag{2.12}$$

puis standardisons la valeur en divisant les termes individuels de la grille par leur somme. Cela donne une approximation pour $f(g, \delta^2 \mid y)$. Sélectionnons ensuite un échantillon aléatoire de taille B à partir des valeurs normalisées de RG de $f(g, \delta^2 \mid y)$. Pour chaque sélection, (g_*, δ_*^2), sélectionnons un échantillon μ de $f(\mu \mid y, g_*, \delta_*^2)$. Ici, nous avons généré $B = 5\,000$ valeurs de μ. Enfin, il convient de mentionner que les approximations des distributions *a posteriori* marginales, c'est-à-dire $f(g \mid y)$ et $f(\delta^2 \mid y)$, peuvent être obtenues directement à partir de l'approximation par grille de $f(g, \delta^2 \mid y)$.

Si nous partons du principe que l'enquête r est le meilleur choix pour l'inférence, nous pouvons considérer la distribution *a posteriori* correspondant à l'enquête r comme faisant l'objet de l'inférence. Dans une application actuelle courante, il y aura des données provenant d'une enquête probabiliste bien établie (le meilleur choix) et des données provenant d'autres sources, comme des enquêtes non probabilistes, des dossiers administratifs, etc. Dans d'autres contextes, il est probable qu'une préférence sera accordée à l'une de ces enquêtes.

Ensuite, en utilisant la valeur attendue *a posteriori* à titre d'illustration, nous obtenons :

$$E(\mu_r \mid y) = E_{g, \delta^2 \mid y} E(\mu_r \mid y, g, \delta^2) \tag{2.13}$$

où $E\left(\mu_r \mid y, g, \delta^2\right)$ est défini dans (2.4). Ainsi, l'inférence pour μ_r est une fonction de $\hat{\mu}_r$ ainsi que des données provenant des $L-1$ autres études, déterminée par la forme de (2.4) et, surtout, par la vraisemblance associée à l'ensemble de sous-ensembles $S_k\left(g\right)$ contenant l'étude r. Consultez Evans et Sedransk (2001) pour obtenir des renseignements supplémentaires et une application à une étude digne de mention sur l'effet de l'usage de l'aspirine chez les patients à la suite d'un infarctus du myocarde.

Le modèle donné par Chakraborty, Datta et Mandal (2014) présente une ressemblance superficielle avec celui de (2.1), de (2.2) et de (2.3). En établissant x_i comme le facteur scalaire de valeur 1, le modèle de (2.1) de Chakraborty et coll. (2014) est :

$$\hat{Y}_i = \mu_i + e_i, \quad i = 1, \ldots, L \tag{2.14}$$

où $\mu_i = \xi + \left(1-\delta_i\right)\nu_{1i} + \delta_i \nu_{2i} + e_i$ avec $e_i, \delta_i, \nu_{1i}, \nu_{2i}$ qui sont indépendantes, $p\left(\delta_i = 1 \mid p\right) = 1 - p$, $\nu_{1i} \sim N\left(0, A_1\right)$ et $\nu_{2i} \sim N\left(0, A_2\right)$. Enfin, $e_i \sim N\left(0, V_i\right)$ avec V_i connue. Ainsi, contrairement à la méthode de combinaison incertaine, il n'y a qu'une seule variable, ξ. Cela permet un traitement approprié des valeurs aberrantes, mais ne permet pas de tirer parti de la mise en grappe possible des μ_i. Cela s'observe également dans (2.4) de Chakraborty et coll. (2014) où :

$$E\left(\mu_i \mid \xi, A_1, A_2, p, y\right) = \hat{Y}_i - \kappa_i\left(\hat{Y}_i - \xi\right) \tag{2.15}$$

et κ_i est une fonction de $V_i, A_1, A_2, p\left(\delta_i = 0 \mid \xi, A_1, A_2, p, y\right)$ avec $y = \left(\hat{Y}_1, \ldots, \hat{Y}_L\right)^t$. Chakraborty et coll. (2014) montrent que si l'enquête i est une valeur aberrante, $E\left(\mu_i \mid \xi, A_1, A_2, p, y\right) \approx \hat{Y}_i$, comme il est souhaité. Supposons maintenant que les enquêtes $\left\{1, \ldots, b\right\}$ et $\left\{b+1, \ldots, L\right\}$ forment deux grappes distinctes comportant une très grande séparation entre elles. L'inférence pour μ_1, par exemple, n'utilisera pas, en général, les données de façon appropriée. Dans (2.15), il devrait y avoir deux valeurs de ξ, c'est-à-dire des valeurs qui correspondent aux deux sous-ensembles. De plus, l'utilisation appropriée des renseignements relatifs aux sous-ensembles constitue l'essence de la méthode de combinaison incertaine.

2.2 Mélange selon le processus de Dirichlet

Une solution de rechange à la méthode de combinaison incertaine consiste à utiliser le mélange selon le processus de Dirichlet, l'une des méthodes bayésiennes non paramétriques les plus populaires. Cette méthodologie est présentée en détail dans les sections 2.1 et 2.2 de Muller, Quintana, Jara et Hanson (2015). Pour nos analyses, nous avons utilisé la fonction R DPmeta du progiciel DPpackage : consultez Jara, Hanson, Quintana, Muller et Rosner (2011) pour obtenir de plus amples précisions. Le modèle de DPmeta est :

$$y_i \mid \theta_i \overset{\text{iid}}{\sim} f_{\theta_i} \tag{2.16}$$

et

$$\theta_i \mid H \overset{\text{iid}}{\sim} H \tag{2.17}$$

avec $H \sim \mathrm{DP}\left(M, H_0\right)$.

Dans (2.16) et (2.17), $y_i = \hat{Y}_i, \theta_i = \mu_i, f_{\theta_i}$ est la fonction de densité de probabilité d'une variable aléatoire $N(\mu_i, V_i)$ dont V_i est fixe et $H_0 = N(\eta, \tau^2)$.

Les hyperparamètres (indépendants) sont :

$$M \mid a_0, b_0 \sim \text{Gamma}(a_0, b_0)$$

$$\eta \mid \eta_b, S_b \sim N(\eta_b, S_b)$$

$$\tau^{-2} \mid \phi_1, \phi_2 \sim \text{Gamma}(\phi_1/2, \phi_2/2). \tag{2.18}$$

Polettini (2017) a proposé d'utiliser un modèle par mélange selon le processus de Dirichlet de cette nature pour l'inférence sur les paramètres de petit domaine. Comme dans la section 2.1 de notre article, Polettini (2017) indique la pertinence de l'extension du modèle typique des effets aléatoires (par exemple, le modèle bien connu de Fay-Herriot) qui suppose l'échangeabilité complète de l'ensemble des paramètres de petit domaine.

La méthode de combinaison incertaine exige seulement que l'on précise une distribution *a priori* pour g et δ^2. En revanche, DPmeta nécessite beaucoup plus de données d'entrée, c'est-à-dire, les valeurs pour $a_0, b_0, \eta_b, S_b, \phi_1$ et ϕ_2. Sans une quantité suffisante de renseignements *a priori*, nous ne pouvons pas formuler d'inférences adéquates au sujet de ces quantités en ayant seulement $L = 3$ enquêtes. Ainsi, nous avons omis la spécification $M \mid a_0, b_0 \sim \text{Gamma}(a_0, b_0)$ et avons fait une inférence pour un ensemble sélectionné de valeurs de M comme le propose Escobar (1994). De plus, nous avons remplacé η_b, S_b, ϕ_1 et ϕ_2 en introduisant l'estimation de leur probabilité *a posteriori* maximale.

3. Résultats

Ha et Sedransk (2019) ont effectué une inférence pour la proportion d'adultes sans assurance maladie dans chacun des 67 comtés de la Floride et ont comparé ces estimations avec celles de deux autres sources. Certaines des différences étaient frappantes, nous motivant à examiner la méthodologie pour faire des inférences appropriées dans de tels cas. Nous utilisons ces données et appliquons des modifications à ces données pour montrer les avantages de l'utilisation de la méthodologie décrite dans la section 2. L'une des sources est le Small Area Health Insurance Estimates (SAHIE, ci-après appelé « l'enquête 1 »). Le programme SAHIE repose sur des estimations ponctuelles tirées de l'American Community Survey (ACS) ainsi que des données administratives, comme les déclarations de revenus fédérales et les taux de participation au Medicaid et au Children's Health Insurance Program (CHIP). Il y a une modélisation détaillée au niveau du domaine. Les principaux modèles sont les modèles d'estimations de l'ACS des proportions dans les groupes de revenu et des proportions de personnes assurées. Il existe d'autres modèles comme ceux modélisant le nombre de personnes assurées par Medicaid ou CHIP, le taux de participation au Supplemental Nutrition Assistance Program et les exemptions fiscales accordées par l'Internal Revenue

Service. Pour obtenir un aperçu complet du programme, consultez le rapport technique de vingt-deux pages de Bauder, Luery et Szelepka (2018). Nous avons ajouté à l'annexe un résumé non technique.

Dans les analyses qui s'appuient à la fois sur l'enquête 2, désignée par HS, fondée sur Ha et Sedransk (2019), et l'enquête 3 désignée par CDC (Centers for Disease Control and Prevention), nous utilisons des modèles au niveau de l'unité selon les données de 2010 provenant du Behavioral Risk Factor Surveillance System (BRFSS) recueillies lors d'entrevues téléphoniques. Bien que les plans d'échantillonnage diffèrent quelque peu d'un État à l'autre, celui de la Floride était typique, c'est-à-dire qu'il s'agissait d'un plan d'échantillonnage stratifié non proportionnel. En Floride, l'ensemble de numéros de téléphone a été divisé en deux strates (de haute et de moyenne densités) échantillonnées séparément. De plus, il y avait une stratification par indicatifs régionaux, c'est-à-dire trois strates géographiques et une quatrième strate composée d'indicatifs régionaux qui, selon les estimations, comportait de grandes populations hispaniques. Pour obtenir des renseignements généraux supplémentaires, consultez http://www.cdc.gov/brfss/annual_data/annual_2010.htm. Pour obtenir des renseignements techniques, consultez Pierannunzi, Xu, Wallace, Garvin, Greenlund, Bartoli, Ford, Eke et Town (2016) et Ha et Sédransk (2019). Les deux études reposent essentiellement sur les mêmes covariables, mais la modélisation dans HS est plus détaillée. De plus, Pierannunzi et coll. (2016) ne fournit que des estimations ponctuelles, soulignant que les erreurs-types étaient en cours d'élaboration. Une autre complication tient au fait que l'analyse par les CDC est fréquentiste, tandis que celles par SAHIE et HS sont bayésiennes. Ainsi, nous avons un écart-type *a posteriori* (empirique) de Bayes pour SAHIE, aucun pour les CDC et une estimation des erreurs-types pour HS obtenue en prenant l'intervalle de crédibilité à 95 % pour la proportion d'un comté et en le divisant par 3,92. Bien que les limites susmentionnées ne nous permettent pas de tirer des conclusions fermes à partir de ces données, elles illustrent la méthode employée. De plus, les conditions dans lesquelles les erreurs-types sont manquantes ou peu fiables sont, au moins, assez courantes pour les échantillons non probabilistes et constituent un thème central de la présente étude.

La première série d'analyses est basée sur les données observées. Pour montrer d'autres propriétés de la méthodologie, une deuxième série d'analyses s'appuie sur des modifications de ces données. Enfin, pour montrer les propriétés d'échantillonnage, une étude par simulations a été réalisée. Mentionnons que chacune de nos analyses est fondée *uniquement* sur les données provenant d'un seul comté. Des études supplémentaires sont requises afin de tirer des inférences à l'aide de données provenant de toutes les sources et de tous les comtés. La discussion à ce sujet figure à la section 4.

3.1 Analyses fondées sur des données

Un résumé des résultats pour le comté de Dixie obtenus au moyen de la méthode de combinaison incertaine est présenté dans le tableau 3.1. Ces résultats sont typiques de la plupart des analyses fondées sur les comtés que nous avons effectuées. Dans l'ensemble du tableau, l'erreur-type représente celui de l'échantillon. Il y a trois sections, qui correspondent à des choix de l'erreur-type des CDC, considérés comme étant égaux à 0,5; 1,0; 2,0 fois l'erreur-type de HS. Pour chaque section, les en-têtes de colonne sont les suivantes : proportion observée, moyenne *a posteriori* de la proportion du comté, erreur-type estimée de

la proportion observée, écart-type *a posteriori* et bornes inférieure et supérieure de l'intervalle de crédibilité à 95 % pour la proportion du comté. Au bas de chaque section se trouvent les valeurs de $p(g \mid y)$ avec $p(g \mid y) \geq 0{,}001$ où $\{g=1\} \sim \{(123)\}$, $\{g=2\} \sim \{(13),(2)\}$, $\{g=3\} \sim \{(12),(3)\}$, $\{g=4\} \sim \{(2,3),(1)\}$, $\{g=5\} \sim \{(1),(2),(3)\}$, ainsi que des résumés correspondant à $\{g=1\}$, étiquetés « échantillon regroupé ».

Nous analysons d'abord ces données en utilisant la méthode de combinaison incertaine avant de les comparer avec celles de DPmeta.

Une manière courante de résumer un ensemble de proportions d'échantillon consiste à supposer que l'ensemble correspondant de proportions réelles provient d'une source commune, c'est-à-dire $p(g=1)=1$. Cependant, pour chacun des trois cas présentés au tableau 3.1, $p(g=1 \mid y) \leq 0{,}001$. Ainsi, il existe peu d'éléments pour appuyer la combinaison de toutes les données provenant des trois enquêtes. Pour étudier plus en détail l'effet de l'hypothèse d'une source commune, supposons que $g=1$. Alors, comme dans (1.1) et (1.2),

$$\hat{Y}_i \overset{\text{ind}}{\sim} N(\mu_i, V_i)$$

$$\mu_i \overset{\text{iid}}{\sim} N(\nu, \delta^2), \quad i=1,\ldots,L. \tag{3.1}$$

Avec une loi *a priori* uniforme localement pour ν et une distribution inverse *a priori* bêta pour δ^2 dans (2.11)

$$f(\nu \mid y) = \int f(\nu \mid \delta^2, y) f(\delta^2 \mid y)\, d\delta^2 \tag{3.2}$$

où la distribution *a posteriori* de ν sachant que δ^2 est normal et que $E(\nu \mid \delta^2, y) = \frac{\sum_{i=1}^{L} \hat{Y}_i / (V_i + \delta^2)}{\sum_{i=1}^{L} 1/(V_i + \delta^2)}$ et $\text{Var}(\nu \mid \delta^2, y) = \left(\sum_{i=1}^{L} (V_i + \delta^2)^{-1}\right)^{-1}$.

Pour la section 1 du tableau 3.1, $E(\nu \mid y) = 0{,}313$, $\text{SD}(\nu \mid y) = 0{,}017$ et l'intervalle de crédibilité à 95 % est (0,290; 0,340). Les inférences fondées sur la distribution *a posteriori* de ν ne sont pas conformes à l'idée selon laquelle l'une ou l'autre des trois enquêtes constitue la norme de référence. Par exemple, si l'enquête 1 est prise comme la norme de référence, la moyenne *a posteriori* de μ_1, 0,254, serait considérablement plus faible que la moyenne *a posteriori* de ν, 0,313. De plus, 0,254 ne se situe pas dans l'intervalle à 95 % pour ν, (0,290; 0,340). Les conclusions tirées des sections 2 et 3 sont essentiellement les mêmes. Enfin, il convient de rappeler que $p(g=1 \mid y) \leq 0{,}001$, ce qui indique qu'il existe peu d'éléments pour appuyer la combinaison de toutes les données.

Dans l'exemple qui suit, supposons, à titre d'illustration, que nous préférons la méthode de HS. Il peut alors y avoir des gains de précision considérables (mesurés par l'écart-type *a posteriori*) en utilisant la méthode de combinaison incertaine. Le gain de précision est mesuré en comparant l'écart-type *a posteriori* de la méthode de combinaison incertaine avec celui obtenu en n'utilisant que les données de l'enquête particulière, ici l'enquête de HS (enquête 2). Pour ce dernier et une loi *a priori* uniforme localement pour μ_2, la distribution *a posteriori* de μ_2 est normale, alors qu'elle comporte une moyenne *a posteriori* égale à la proportion observée et un écart-type *a posteriori* égal à l'erreur-type estimée. Si nous prenons l'erreur-type des $\text{CDC} = k$ (erreur-type de HS) pour $k = 0{,}5$, 1 et 2, les réductions de l'écart-type *a posteriori* pour

l'enquête de HS (enquête 2) correspondant à $k = 0{,}5,\ 1$ et 2 sont alors de 29 %, 18 % et 7 %. (Par exemple, pour la section 1 du tableau 3.1, c'est-à-dire $k = 0{,}5,$ la baisse en pourcentage de l'écart-type *a posteriori* pour HS est de $100\,(0{,}028 - 0{,}020) / 0{,}028\,\% = 29\,\%$). Notons que la valeur relativement faible des erreurs-types pour chacune des enquêtes signifie que la partition de « tous les singletons », c'est-à-dire $\{g = 5\},$ a une probabilité *a posteriori* relativement élevée (d'environ 0,38). Les réductions correspondantes de l'écart-type *a posteriori* (combinaison incertaine par rapport à aucune combinaison) pour les CDC (enquête 3) sont de 7 %, 18 % et 14 %.

Tableau 3.1
Proportions observées, erreurs-types et valeurs sommaires *a posteriori* du comté de Dixie, en Floride, établies au moyen de la combinaison incertaine

	Enquête	Proportion observée	Moyenne *a posteriori*	Erreur-type observée	Écart-type *a posteriori*	Intervalle de crédibilité à 95 %
Erreur-type des CDC = 0,5 × erreur-type de HS	1	0,254	0,254	0,014	0,014	(0,225; 0,283)
	2	0,361	0,360	0,028	0,020	(0,317; 0,403)
	3	0,359	0,359	0,014	0,013	(0,333; 0,385)
	échantillon regroupé		0,313		0,017	(0,290; 0,340)
	$P(g=3\mid y) = 0{,}002;\ P(g=4\mid y) = 0{,}621;\ P(g=5\mid y) = 0{,}377.$					
Erreur-type des CDC = erreur-type de HS	1	0,254	0,254	0,014	0,014	(0,225; 0,283)
	2	0,361	0,360	0,028	0,023	(0,313; 0,406)
	3	0,359	0,359	0,028	0,023	(0,312; 0,404)
	échantillon regroupé		0,290		0,011	(0,268; 0,312)
	$P(g=2\mid y) = 0{,}002;\ P(g=3\mid y) = 0{,}002;\ P(g=4\mid y) = 0{,}619;\ P(g=5\mid y) = 0{,}376.$					
Erreur-type des CDC = 2 × erreur-type de HS	1	0,254	0,254	0,014	0,014	(0,226; 0,284)
	2	0,361	0,360	0,028	0,026	(0,307; 0,412)
	3	0,359	0,349	0,056	0,048	(0,256; 0,303)
	échantillon regroupé		0,279		0,012	(0,255; 0,305)
	$P(g=1\mid y) = 0{,}001;\ P(g=2\mid y) = 0{,}107;\ P(g=3\mid y) = 0{,}002;\ P(g=4\mid y) = 0{,}554;\ P(g=5\mid y) = 0{,}336.$					

Note : CDC signifie Centers for Disease Control and Prevention; HS signifie Ha et Sedransk.

Comme l'indique la section 2, une spécification complète de DPmeta nécessite la spécification des valeurs de nombreux hyperparamètres, et nous n'avons aucun renseignement *a priori* pour faire des choix éclairés. Nous avons donc remplacé η_b, S_b, ϕ_1 et ϕ_2 en introduisant l'estimation de leur probabilité *a posteriori* maximale. Nous avons suivi les travaux d'Escobar (1994) en prenant en compte $M \in \{L^{-1}, L^0, L^1, L^2\} = \{1/3; 1; 3; 9\}$.

À partir de (2.10), comme le décrit l'étude de Muller et coll. (2015), la probabilité *a priori* des k grappes est une fonction de M. Soit $p_M = (p_{1M}, p_{2M}, p_{3M})$ où p_{kM} est la probabilité *a priori* des k grappes avec précision M. p_{kM} peut alors être calculée à l'aide de la probabilité associée à n'importe quelle partition, c'est-à-dire :

$$\frac{M^{k-1}\prod_{j=1}^{k}\Gamma(L_j)}{(M+1)(M+2)\cdots(M+L-1)} \tag{3.3}$$

où L_j correspond au nombre d'enquêtes dans la grappe j avec $\sum_{j=1}^{k} L_j = L.$ Alors $p_{1/3} = (18/28; 9/28; 1/28),\ p_1 = (2/6; 3/6; 1/6),\ p_3 = (2/20; 9/20; 9/20)$ et $p_9 = (2/110; 27/110; 81/110).$

Puisque les valeurs de $p_{1/3}$ et de p_9 sont trop extrêmes, nous avons insisté sur le fait que $M = 1$ et $M = 3$. Les résultats correspondant à $M = 1$ et $M = 3$ étant très près les uns des autres, seuls ces derniers sont présentés dans le tableau 3.2, qui a le même format que le tableau 3.1.

Si nous comparons la méthode de combinaison incertaine à DPmeta, nous observons qu'en général, il y a une étroite correspondance entre les résultats. Pour les moyennes *a posteriori*, elles sont similaires, sauf pour ce qui est de la section 3 où nous constatons un rétrécissement plus important des valeurs pour les enquêtes 2 et 3. Les résultats pour l'écart-type *a posteriori* sont également proches, à l'exception de la section 3 de l'enquête 2, où la valeur est plus élevée. Il n'y a que de petites différences dans les intervalles, sauf pour la section 3, où les intervalles DPmeta pour les enquêtes 2 et 3 sont plus grands.

Tableau 3.2
Proportions observées, erreurs-types et valeurs sommaires *a posteriori* du comté de Dixie, en Floride, établies au moyen de DPmeta

	Enquête	Proportions observées	Moyenne *a posteriori*	Erreur-type observée	Écart-type *a posteriori*	Intervalle de crédibilité à 95 %
Erreur-type des CDC = 0,5 × erreur-type de HS	1	0,254	0,254	0,014	0,014	(0,227; 0,282)
	2	0,361	0,359	0,028	0,013	(0,334; 0,384)
	3	0,359	0,360	0,014	0,012	(0,335; 0,384)
Erreur-type des CDC = erreur-type de HS	1	0,254	0,256	0,014	0,016	(0,227; 0,290)
	2	0,361	0,357	0,028	0,024	(0,291; 0,399)
	3	0,359	0,357	0,028	0,025	(0,290; 0,399)
Erreur-type des CDC = 2 × erreur-type de HS	1	0,254	0,264	0,014	0,018	(0,230; 0,287)
	2	0,361	0,332	0,028	0,044	(0,261; 0,406)
	3	0,359	0,321	0,056	0,048	(0,249; 0,402)

Note : CDC signifie Centers for Disease Control and Prevention; HS signifie Ha et Sedransk.

La faible valeur de l'erreur-type du SAHIE observée dans presque tous les comtés impose des contraintes en ce qui concerne la portée de notre évaluation. Ainsi, nous avons utilisé des ensembles de données qui ont été modifiés en fonction des données originales. Ici, comme auparavant, nous considérons l'erreur-type des CDC comme équivalant à 0,5, 1 et 2 fois l'erreur-type de HS, mais aussi l'erreur-type de SAHIE comme 2, 5 et 10 fois l'erreur-type de HS. Les données du tableau 3.3, présentées dans le même format que celles du tableau 3.1, montrent, pour la combinaison incertaine, les résultats pour le comté d'Orange et que l'erreur-type des CDC équivaut à 0,5 fois l'erreur-type de HS. Ces résultats sont typiques de ce que nous avons observé dans le cadre de nos analyses pour un nombre important de comtés de la Floride. Il faut toutefois se rappeler que chaque analyse est fondée uniquement sur les données du comté en particulier. Pour la section 1 du tableau 3.3, $E\left(\nu \middle| \left\{\hat{Y}_i : i = 1, \ldots, L\right\}\right) = 0{,}199$, E.-T.$\left(\nu \middle| \left\{\hat{Y}_i : i = 1, \ldots, L\right\}\right) = 0{,}008$ et l'intervalle de crédibilité à 95 % est (0,184; 0,215). Les inférences fondées sur la distribution *a posteriori* de ν ne sont pas appropriées si l'on utilise l'une ou l'autre des trois enquêtes comme la norme de référence. Par exemple, la moyenne *a posteriori* de μ_1, 0,278, se situe au-delà de l'intervalle de crédibilité à 95 % pour ν. Comme dans le tableau 3.1, $p\left(g = 1 \middle| y\right) \leq 0{,}001$, ce qui indique qu'il existe peu d'éléments pour appuyer la combinaison de toutes les données.

Les réductions en pourcentage de l'écart-type *a posteriori* de μ_1, c'est-à-dire pour SAHIE (enquête 1), sont de 11 %, 26 % et 44 %, correspondant aux trois sections du tableau 3.3. Au fur et à mesure que la valeur

de l'erreur-type de SAHIE augmentait, il y a, comme il fallait s'y attendre, un autre combinaison de la proportion observée par SAHIE avec la proportion observée par les CDC. Soulignons que la proportion observée par SAHIE est de 0,294, tandis que les moyennes *a posteriori* pour μ_1 diminuent, pour passer de 0,278 (section 1) à 0,240 (section 3). Une raison à cela peut être déterminée en comparant les distributions *a posteriori* de g, c'est-à-dire $\{g, p(g \mid y) : g = 1, \ldots, 5\}$ qui se trouvent au bas de chaque section. Par exemple, $p(g = 2 \mid y)$ passe de 0,006 à 0,339, tandis que $p(g = 5 \mid y)$ passe de 0,479 à 0,253. Selon ces résultats, la méthode de combinaison incertaine tient compte comme il convient de la plus grande variabilité associée aux estimations de SAHIE, c'est-à-dire qu'elle augmente la probabilité que les données des enquêtes 1 et 3 soient regroupées.

Si nous comparons les résultats de la méthode de combinaison incertaine présentés dans le tableau 3.3 avec ceux de DPmeta présentés dans le tableau 3.4, nous observons que les écarts sont plus marqués que ceux observés dans les tableaux 3.1 et 3.2. En ce qui a trait aux moyennes *a posteriori*, il convient de souligner que pour l'enquête 1, la moyenne *a posteriori* de DPmeta est quelque peu inférieure à celle de la combinaison incertaine. Cela rend compte de la plus grande combinaison des données de l'enquête 1 avec celles de l'enquête 3. Pour les enquêtes 2 et 3, les deux ensembles de moyennes *a posteriori* sont similaires. La différence la plus marquée a trait aux écarts-types *a posteriori* où, pour l'enquête 1 (sections 2 et 3), les valeurs de DPmeta sont bien inférieures à celles de la combinaison incertaine. Pour les sept autres cas, les deux ensembles d'écarts-types *a posteriori* sont similaires. De même, pour l'enquête 1, les intervalles de DPmeta sont beaucoup plus étroits que ceux de la combinaison incertaine, tandis que ceux des enquêtes 2 et 3 ne sont qu'un peu plus grands. D'après le tableau 3.4 et pour l'enquête 1, les baisses en pourcentage des écarts-types *a posteriori* (par rapport aux erreurs-types observées) sont de (42 %, 55 %, 75 %). Des augmentations ont toutefois été observées pour l'enquête 2 dans les sections 2 et 3.

Tableau 3.3
Proportions observées, écarts-types et valeurs sommaires *a posteriori* du comté d'Orange, en Floride, où l'erreur-type des CDC = 0,5 × erreur-type HS est établie au moyen de la combinaison incertaine

	Enquête	Proportion observée	Moyenne *a posteriori*	Erreur-type observée	Écart-type *a posteriori*	Intervalle de crédibilité à 95 %
Erreur-type du SAHIE = 2 × erreur-type de HS	1	0,294	0,278	0,036	0,032	(0,227; 0,352)
	2	0,257	0,261	0,018	0,017	(0,226; 0,294)
	3	0,179	0,179	0,009	0,009	(0,162; 0,197)
	échantillon regroupé		0,199		0,008	(0,184; 0,215)
	$P(g = 2 \mid y) = 0{,}006;\ P(g = 3 \mid y) = 0{,}514;\ P(g = 5 \mid y) = 0{,}479.$					
Erreur-type du SAHIE = 5 × erreur-type de HS	1	0,294	0,251	0,089	0,066	(0,162; 0,417)
	2	0,257	0,258	0,018	0,018	(0,223; 0,293)
	3	0,179	0,179	0,009	0,009	(0,162; 0,197)
	échantillon regroupé		0,195		0,008	(0,180; 0,211)
	$P(g = 2 \mid y) = 0{,}224;\ P(g = 3 \mid y) = 0{,}468;\ P(g = 5 \mid y) = 0{,}308.$					
Erreur-type du SAHIE = 10 × erreur-type de HS	1	0,294	0,240	0,179	0,101	(0,059; 0,520)
	2	0,257	0,257	0,018	0,018	(0,222; 0,292)
	3	0,179	0,179	0,009	0,009	(0,162; 0,197)
	échantillon regroupé		0,195		0,008	(0,179; 0,211)
	$P(g = 2 \mid y) = 0{,}339;\ P(g = 3 \mid y) = 0{,}408;\ P(g = 5 \mid y) = 0{,}253.$					

Note : SAHIE signifie Small Area Health Insurance Estimates; CDC signifie Centers for Disease Control and Prevention; HS signifie Ha et Sedransk.

Tableau 3.4
Proportions observées, erreurs-types et valeurs sommaires *a posteriori* du comté d'Orange, en Floride, établies au moyen de DPmeta

	Enquête	Proportions observées	Moyenne *a posteriori*	Erreur-type observée	Écart-type *a posteriori*	Intervalle de crédibilité à 95 %
Erreur-type du SAHIE = 2 × erreur-type de HS	1	0,294	0,262	0,036	0,021	(0,202; 0,297)
	2	0,257	0,263	0,018	0,018	(0,222; 0,296)
	3	0,179	0,180	0,009	0,009	(0,162; 0,199)
Erreur-type du SAHIE = 5 × erreur-type de HS	1	0,294	0,226	0,089	0,040	(0,168; 0,290)
	2	0,257	0,246	0,018	0,023	(0,186; 0,291)
	3	0,179	0,182	0,009	0,011	(0,163; 0,205)
Erreur-type du SAHIE = 10 × erreur-type de HS	1	0,294	0,217	0,179	0,044	(0,166; 0,288)
	2	0,257	0,243	0,018	0,031	(0,185; 0,290)
	3	0,179	0,183	0,009	0,011	(0,162; 0,205)

Note : SAHIE signifie Small Area Health Insurance Estimates; HS signifie Ha et Sedransk.

3.2 Résultats de l'étude par simulations

Pour évaluer des propriétés comme le biais et la couverture de l'intervalle de crédibilité, nous avons réalisé une étude par simulations basée sur plusieurs modifications des données du comté d'Orange. Plus précisément, nous générons $\{\hat{Y}_i : i = 1, 2, 3\}$ à partir de :

$$\hat{Y}_1 \sim N(\psi_1, V_1)$$

$$\hat{Y}_2 \sim N(\psi_1, V_2)$$

$$\hat{Y}_3 \sim N(\psi_2 + \Delta, V_2) \tag{3.4}$$

où ψ_1 et ψ_2 proviennent de la section 3 du tableau 3.3, ψ_1 est la moyenne des proportions observées des enquêtes 1 et 2 et ψ_2 est la proportion observée de l'enquête 3 (CDC). Par ailleurs, nous nous attendions à ce que V_1 soit beaucoup plus grande que V_2. Nous avons fait ces choix afin de représenter une situation fréquente où l'enquête 1 est un échantillon probabiliste ayant une variance d'échantillon relativement élevée, tandis que les enquêtes 2 et 3 sont des échantillons non probabilistes ayant des variances d'échantillon beaucoup plus faibles. Enfin, $\Delta \in \{0; 4(0{,}0193) = 0{,}0772; 8(0{,}0193) = 0{,}1544\}$.

Le tableau 3.5 donne les valeurs de ψ_1, ψ_2, V_1 et V_2 dans la note au bas du tableau. Le tableau comporte trois rangées, correspondant à $\Delta =$ 0; 0,0772; 0,1544. Dans chaque rangée figurent les valeurs médianes de 500 répliques de $\{p(g \mid y) : g = 1, \ldots, 5\}, \{E(\mu_i \mid y) : i = 1, 2, 3\}$ et $\{\text{E.-T.}(\mu_i \mid y) : i = 1, 2, 3\}$ ainsi que les couvertures estimées.

Tableau 3.5
Résultats de la simulation provenant de 500 répliques de (3.4)

Grandeur de l'entreprise	$P(g \mid \text{données})$					Couverture			Moyenne *a posteriori*			Écart-type *a posteriori*		
	g : 1	2	3	4	5	i : 1	2	3	i : 1	2	3	i : 1	2	3
$\Delta = 0$	0	0,148	0,462	0	0,332	0,973	0,958	0,960	0,262	0,275	0,179	0,050	0,006	0,006
$\Delta =$ 0,0772	0,032	0,292	0,303	0,033	0,249	0,984	0,941	0,939	0,269	0,275	0,257	0,039	0,006	0,006
$\Delta =$ 0,1544	0	0,280	0,401	0	0,300	0,971	0,958	0,952	0,292	0,275	0,333	0,044	0,006	0,006

$\psi_1 =$ 0,276; $\psi_2 =$ 0,179; $V_1 = 0{,}06^2$; $V_2 = 0{,}006^2$.

Les principaux résultats montrent que : a) les médianes des moyennes *a posteriori* sont proches des valeurs utilisées pour générer les données, c'est-à-dire ψ_1 et ψ_2; b) les couvertures sont proches de la valeur nominale de 95 %; c) nous observons une diminution notable de l'écart-type *a posteriori* pour l'enquête 1 (SAHIE), soit de 16,7 %, de 35,0 % et de 26,7 %, ce qui correspond à $\Delta = 0$; 0,0772 et à 0,1544. Il n'y a pas de réduction de l'écart-type *a posteriori* pour les enquêtes 2 et 3.

Pour $\Delta = 0{,}1544$, notons que $p(g = 2 \mid y) = 0{,}280$ et que $p(g = 4 \mid y) = 0$, c'est-à-dire que les données des enquêtes 1 (SAHIE) et 3 (CDC), $\{g = 2\}$ ont dû être regroupées en raison de l'erreur-type relativement importante pour l'enquête 1 (SAHIE). Cependant, nous n'avons pas regroupé les données des enquêtes 2 (HS) et 3 (CDC), $\{g = 4\}$, en raison des erreurs-types relativement petites pour les enquêtes 2 (HS) et 3 (CDC). Bien sûr, nous avons regroupé les données des enquêtes 1 et 2, $\{g = 3\}$, car elles ont la même moyenne, ψ_1.

4. Discussion et résumé

Compte tenu des ressources et des taux de réponse réduits, la combinaison de données provenant de plusieurs sources comme les enquêtes-échantillons et les données administratives suscite beaucoup d'intérêt. Actuellement, on s'intéresse particulièrement aux cas où les sources comprennent des enquêtes non probabilistes. Une méthodologie appropriée est requise afin de produire des inférences satisfaisantes puisque les populations cibles et les méthodes d'acquisition de données peuvent être assez différentes.

Il existe de nombreuses situations où il pourrait être avantageux de combiner ces données, comme le montre l'article de synthèse de Lohr et Raghunathan (2017). Ici, nous avons étudié le cas où l'analyste ne dispose que de statistiques sommaires provenant de chacune des sources, et où l'on peut considérer une source, r, comme la meilleure source pour l'inférence. Bien qu'il soit souvent avantageux d'utiliser les données provenant de sources connexes pour améliorer les inférences dans r, il est essentiel que les données combinées concordent avec celles dans r. La méthodologie décrite dans le présent article peut également être utilisée lorsque les données ne se limitent pas à des statistiques sommaires et lorsque les objectifs et les modèles inférentiels sont plus complexes. Comme le montre l'article, à défaut de tenir compte des biais dus à la combinaison de données « dissemblables », cela peut donner lieu à de mauvaises inférences. À l'aide d'expressions analytiques et d'exemples, nous avons montré que les méthodes de combinaison incertaine et par mélange selon le processus de Dirichlet permettent de faire des inférences appropriées. Cependant, nos analyses basées sur la méthode de combinaison incertaine sont entièrement bayésiennes tandis que celles de DPmeta sont bayésiennes empiriques, compte tenu de la nécessité de préciser les valeurs pour de nombreux hyperparamètres. De plus, la méthode de combinaison incertaine fournit des renseignements supplémentaires sous la forme de probabilités *a posteriori* pour les partitions g.

Ces méthodes peuvent être mises en œuvre. Pour DPmeta, il existe un progiciel R, DPpackage (Jara et coll., 2011), tandis qu'un autre progiciel R est en cours de développement pour la méthode de combinaison incertaine. Une fois terminé, il sera intégré au référentiel Comprehensive R Archive Network. Ce

progiciel comporte des fonctions qui permettent des analyses bayésiennes, semblables à celles décrites dans le présent article : a) par des estimations ponctuelles fournies par l'utilisateur et les variances connexes; ou b) par des données binomiales, des cas (y) et le nombre total (n). Le cas (b) fournit une analyse basée sur la transformation logit de la proportion d'échantillon. Nous avons utilisé la dernière méthode lorsqu'il y avait 11 enquêtes.

Il est très difficile de faire des inférences pour les variances des échantillons $V_1, \ldots, V_L$. Polettini (2017) présente une analyse approfondie des méthodes qui ont été proposées. Les solutions proposées par You et Chapman (2006), Sugasawa, Tamae et Kubokawa (2017) et Polettini (2017) sont particulièrement intéressantes. Cependant, ces solutions sont proposées dans le contexte de l'inférence sur petits domaines, et non lorsque l'objectif est de combiner des données d'enquêtes et de sources connexes.

Bien que la discussion ci-dessous s'inscrive dans le contexte de l'extension de la méthode par mélange selon le processus de Dirichlet (section 2.2), les idées présentent aussi un intérêt pour la méthode de combinaison incertaine (section 2.1). Polettini (2017) enrichit le modèle par mélange selon le processus de Dirichlet décrit à la section 2.2 en y ajoutant :

$$\delta_i S_i^2 \overset{\text{ind}}{\sim} V_i \chi^2_{\delta_i}, \quad i = 1, \ldots, L \tag{4.1}$$

et

$$V_i^{-1} \overset{\text{iid}}{\sim} \text{Gamma}\left(a_1, b_1\right) \tag{4.2}$$

où S_i^2 correspond à la variance d'échantillonnage et δ_i est une mesure des degrés de liberté.

Comme le souligne Polettini (2017), l'hypothèse selon laquelle les données suivent la distribution χ^2 dans (4.1) est discutable, d'autant plus lorsque la conception de l'enquête est complexe. On ne peut pas vérifier la *distribution d'échantillonnage* de S_i^2 au moyen d'un seul échantillon, comme l'a également souligné Polettini (2017) à la page 731. De plus, dans les enquêtes par sondage, la forme de S_i^2 est susceptible d'être une fonction complexe des valeurs de la variable d'intérêt Y et des poids d'enquête. Ainsi, il est peu probable que la distribution des valeurs observées de Y puisse être utilisée pour inférer une valeur approximative et raisonnable de la distribution de S_i^2.

L'hypothèse selon laquelle les paramètres de population sont constants dans (4.2) pose problème dans notre cas, c'est-à-dire lorsqu'il s'agit de combiner les données. Nous nous attendons à observer d'importantes différences entre les enquêtes, par exemple, pour un ensemble d'échantillons probabilistes et non probabilistes. You et Chapman (2006) généralisent cette approche en remplaçant (4.2) par :

$$V_i \overset{\text{ind}}{\sim} \text{Inverse Gamma}\left(a_i, b_i\right). \tag{4.3}$$

Il convient donc de dériver les valeurs pour $\left(a_i, b_i\right)$, ce qui peut être difficile lorsqu'il n'y a pas de renseignements préalables. De plus, Gelman (2006) montre que le fait de sélectionner des très petites valeurs pour a_i et b_i, un choix naturel (et celui fait par You et Chapman (2006)), peut donner lieu à de mauvaises

inférences. Sugasawa et coll. (2017) propose une approche autre que celle de You et Chapman (2006) en supposant :

$$V_i \overset{\text{ind}}{\sim} \text{Inverse Gamma}\left(a_i, b_i\gamma\right), \tag{4.4}$$

avec une loi *a priori* pour γ, mais cette approche nécessite, elle aussi, que l'on définisse les valeurs pour a_i et b_i. De toute évidence, la production de meilleures inférences pour les variances d'échantillon est un sujet important pour les recherches à venir.

Il y a eu un intérêt accru pour la formulation d'inférences pour de petites sous-populations, c'est-à-dire l'inférence provenant de petits échantillons, lorsqu'il existe plusieurs sources de données; voir, par exemple, Manzi, Spiegelhalter, Turner, Flowers et Thompson (2011) et Nandram, Berg et Barboza (2014). Bien que de plus amples recherches soient requises pour étendre la méthodologie de combinaison incertaine à ce cas, l'approche est claire. Soit j une petite région, comme un comté américain, et i une source de données où $j=1,\ldots,J$ et $i=1,\ldots,L.$ Selon la définition susmentionnée, g désigne une partition générique ayant un sous-ensemble générique $S_k(g)$ pour $k=1,\ldots,d(g).$ Définissons $\mathcal{G}=\{(ij): j=1,\ldots,J; i=1,\ldots,L\}.$ Alors pour un g fixe, $S_k(g)$ est un sous-ensemble de $\mathcal{G}$ avec $S_k(g)\bigcap S_m(g)=\varnothing$ pour $k\neq m$ et $\bigcup_{k=1}^{d(g)} S_k(g)=\mathcal{G}.$ Par exemple, soit $J=2$ et $L=3.$ Chaque partition serait alors un ensemble de sous-ensembles disjoints de $\mathcal{G}=\{(11),(12),(21),(22),(31),(32)\}$ dont l'union est $\mathcal{G}.$ Par analogie avec la discussion de la section 2, il y aurait une seule meilleure source pour chaque petit domaine, désignée par $(j,i(j))$ pour quelques valeurs i dans le petit domaine $j.$

Alors le modèle suivant, analogue à celui de la section 2, est :

$$\hat{Y}_{ij} \overset{\text{ind}}{\sim} N\left(\mu_{ij}, V_{ij}\right). \tag{4.5}$$

Par analogie avec (2.2)

$$\mu_{ij} \overset{\text{ind}}{\sim} N\left(\nu_k(g), \delta_k^2(g)\right), \qquad ij\in S_k(g). \tag{4.6}$$

Si les mêmes hypothèses (à propos des limites) sont formulées pour $\gamma_k^2(g)$ et $\delta_k^2(g)=\delta^2,$ les expressions pour l'inférence *a posteriori* sur μ_{ij} seront les mêmes que celles qui figurent à la section 2. Cependant, l'hypothèse selon laquelle les valeurs de δ^2 sont constantes peut ne pas être raisonnable. En raison du très grand nombre de partitions, le calcul sera difficile, d'autant plus qu'il est prévu que la valeur de nombreuses $p(g\,|\,y)$ sera très petite.

La prémisse de notre travail est qu'il faut inclure la possibilité que les paramètres associés aux différentes enquêtes ne soient pas interchangeables. (Cela peut s'avérer particulièrement important lorsqu'il s'agit d'un échantillon probabiliste et de plusieurs échantillons non probabilistes.) De même, il est naturel de généraliser afin que les paramètres associés aux petites zones soient réputés ne pas être interchangeables. Cependant, si nous pouvons supposer l'interchangeabilité entre les enquêtes et les petites zones, le modèle décrit à la section 2.1 de Kim, Park et Kim (2015) (qui a possiblement été modifié afin de pouvoir utiliser l'approche bayésienne) devrait être plus facile à mettre en œuvre.

Même si le très grand nombre de partitions de $\mathcal{G}$ peut constituer un obstacle à la mise en œuvre, il serait possible d'appliquer DPmeta lorsqu'il existe des données provenant d'un ensemble de petits domaines et de

plusieurs sources de données. Un des problèmes tient à la spécification dans DPmeta d'une distribution commune pour μ_{ij}, c'est-à-dire, pour de petits domaines et enquêtes, ce qui est peu susceptible d'être appropriée. Comme exemples de possibilités, mentionnons un modèle de type analyse de la variance (ANOVA) (section 4.4.2) ou un modèle emboîté (section 7.3.1) de Muller et coll. (2015), bien que le modèle ANOVA ne comporte pas de termes d'interaction et que le modèle que nous avons utilisé soit un modèle recoupé dans un plan d'échantillonnage.

Il serait utile à l'avenir que la recherche comporte une approche d'inférence pour les variances d'échantillon, comme cela est indiqué ci-dessus. De plus, lorsqu'il existe des données provenant de plusieurs enquêtes, nous avons besoin d'une méthode améliorée pour gérer l'extension à l'inférence sur petits domaines. Dans certains cas, il serait possible de simplifier le modèle pour μ_{ij}. Une méthode d'échantillonnage fondée sur une grille pour g et δ^2 est difficile à mettre en œuvre lorsque la valeur de G est extrêmement grande. Ainsi, l'utilisation d'une approche de Monte Carlo par chaîne de Markov standard, ayant possiblement une distribution *a priori* informative sur g, peut être un meilleur moyen de tirer des inférences. Voir, par exemple, Dahl, Day et Tsai (2017).

D'autres approches pourraient également être étudiées. Par exemple, Park, Kim et Stukel (2017) proposent une approche différente pour combiner les données provenant de deux enquêtes. Dans ce cas, il y a des covariables, X, qui sont observées dans chaque enquête, tandis que Y_1, la variable d'intérêt de l'étude, est observée uniquement dans l'enquête 1 et que Y_2 est observée uniquement dans l'enquête 2. L'inférence pour la moyenne de la population de Y_1 est souhaitée, compte tenu des données provenant de deux enquêtes. Les densités utilisées sont $f_1\left(Y_1 \mid X, \theta_1\right), f_2\left(Y_2 \mid X, Y_1, \theta_2\right)$ et, pour l'identifiabilité, on suppose que $f_2\left(Y_2 \mid X, Y_1\right) = f_2\left(Y_2 \mid Y_1\right)$. Pour une analyse bayésienne, il serait nécessaire d'avoir un prolongement de plus de deux enquêtes ainsi que la spécification des distributions *a priori* appropriées aux paramètres. Il semble que modéliser la distribution de Y_2 en utilisant la distribution de Y_1 ne soit pas simple.

Remerciements

Les auteurs remercient les examinateurs de leurs commentaires détaillés ayant donné lieu à un article plus ciblé et d'une plus grande portée. Ils apprécient également les subventions de recherche fournies par le Pittsburgh Supercomputing Center d'ACCESS.

Annexe

Programme Small Area Health Insurance Estimates

Dans le résumé suivant, nous paraphrasons les parties pertinentes du US Census Bureau (2021). Pour éviter de déformer le sens des propos des auteurs, nous avons conservé le texte à la première personne.

Le programme Small Area Health Insurance Estimates (SAHIE) produit des estimations basées sur un modèle de la couverture par l'assurance maladie pour les groupes démographiques au sein des comtés et

des États. Nous publions les estimations de comté selon le sexe, l'âge et le revenu. Les groupes de revenu sont définis par le « income-to-poverty ratio », soit le ratio du revenu familial au seuil de pauvreté fédéral approprié.

Pour l'estimation, SAHIE s'appuie sur des modèles qui combinent les données d'enquête de l'American Community Survey (ACS) avec les données provenant de dossiers administratifs et les données des recensements. Les modèles sont des modèles régionaux parce que nous utilisons des estimations d'enquête et des données administratives à certains niveaux d'agrégation, plutôt que des données d'enquêtes et de dossiers administratifs individuels. Notre approche de modélisation est semblable à celle des modèles courants élaborés pour l'estimation sur petits domaines, mais elle comporte des éléments de complexité supplémentaires.

Les estimations publiées sont fondées sur des agrégats de groupes démographiques modélisés. Pour les comtés, nous modélisons les estimations à un niveau de base défini par les groupes d'âge, de sexe et de revenu.

Nous utilisons les estimations du Population Estimates Program du U.S. Census Bureau pour la population dans des groupes définis pour le comté selon l'âge et le sexe. Nous traitons ces populations comme si elles étaient connues. Au sein de chacun de ces groupes, le nombre de personnes couvertes par l'assurance maladie dans l'une des catégories de revenu est indiqué par cette population multipliée par deux proportions inconnues à estimer : la proportion de personnes dans la catégorie de revenu et la proportion de personnes assurées dans cette catégorie de revenu. Les modèles comportent deux parties largement distinctes – une « partie revenu » et une « partie assurance » – qui correspondent à ces proportions. Nous utilisons des estimations de l'enquête des proportions dans les groupes de revenu et des proportions de personnes assurés au sein de ces groupes. Nous supposons que ces estimations d'enquête sont impartiales et qu'elles suivent des distributions connues. Nous supposons également des formes fonctionnelles des variances des estimations d'enquête qui mettent en jeu des paramètres qui sont estimés. Nous traitons les variables supplémentaires qui permettent de prédire l'une ou les deux des proportions inconnues de contribuables et de prestataires d'assurance de l'une des deux manières présentées ci-dessous.

Certaines de ces variables sont utilisées comme prédicteurs fixes dans un modèle de régression. Il y a une composante de régression dans les parties du modèle relatives au revenu et à l'assurance. Dans chaque cas, une transformation de la proportion est prédite par une combinaison linéaire de prédicteurs fixes. Certains de ces prédicteurs sont des variables catégoriques qui définissent les groupes démographiques que nous modélisons. D'autres sont des variables continues. Les prédicteurs fixes continus comprennent des variables liées à l'emploi, au niveau de scolarité et à la population démographique.

Nous utilisons également des prédicteurs continus aléatoires, qui comportent des données provenant de l'ACS sur 5 ans, de l'Internal Revenue Service, du Supplemental Nutrition Assistance Program et du Medicaid ou du Children's Health Insurance Program. Il ne s'agit pas de prédicteurs fixes dans le modèle. Au lieu de les considérer comme des prédicteurs fixes, nous les traitons comme des prédicteurs aléatoires,

de façon similaire aux estimations d'enquête, mais différente des estimateurs sans biais des chiffres. À la place, nous supposons que les valeurs estimées sont des fonctions linéaires du nombre de personnes dans un groupe de revenu ou du nombre de personnes assurées dans un groupe de revenu. Nous supposons généralement que leur distribution est normale et que les variances dépendent de paramètres inconnus.

Nous formulons le modèle dans un cadre bayésien et déclarons les moyennes *a posteriori* comme des estimations ponctuelles. Nous utilisons les moyennes et les variances *a posteriori* avec une approximation normale pour calculer des intervalles de confiance symétriques à 90 % et utilisons la demi-largeur de ces intervalles comme marges d'erreur.

Nous ajustons les estimations pour assurer leur conformité avec les totaux nationaux précisés.

Bibliographie

Bauder, M., Luery, D. et Szelepka, S. (2018). *Small Area Estimation of Health Insurance Coverage in 2010-2016*. Rapport technique, U.S. Census Bureau.

Chakraborty, A., Datta, G.S. et Mandal, A. (2016). A two-component normal mixture alternative to the Fay-Herriot model. *Statistics in Transition new series and Survey Methodology*, 17, 1, 67-90, https://doi.org/10.21307/stattrans-2016-006.

Dahl, D., Day, R. et Tsai, J. (2017). Random partition distribution indexed by pairwise information. *Journal of the American Statistical Association*, 112, 721-732.

Escobar, M. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89, 268-277.

Evans, R., et Sedransk, J. (1999). Methodology for pooling subpopulation regressions when there is uncertainty about which subpopulations are similar. *Statistica Sinica*, 9, 345-359.

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.

Ha, N., et Sedransk, J. (2019). Assessing health insurance coverage in Florida using the Behavioral Risk Factor Surveillance System. *Statistics in Medicine*, 38(13), 2332-2352, https://doi.org/10.1002/sim.8108.

Jara, A., Hanson, T., Quintana, F., Müller, P. et Rosner, G. (2011). DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software*, 40(5), 1-30, https://doi.org/10.18637/jss.v040.i05.

Kim, J.-K., Park, S. et Kim, S.-Y. (2015). Estimation sur petits domaines en combinant des données provenant de plusieurs sources. *Techniques d'enquête*, 41, 1, 21-37. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2015001/article/14150-fra.pdf.

Lohr, S., et Raghunathan, T. (2017). Combining survey data with other data sources. *Statistical Science*, 32(2), 293-312.

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.

Manzi, G., Spiegelhalter, D.J., Turner, R.M., Flowers, J. et Thompson, S.G. (2011). Modelling bias in combining small area prevalence estimates from multiple surveys. *Journal of the Royal Statistical Society, A*, 174(1), 31-50, https://doi.org/10.1111/j.1467-985X.2010.00648.x.

Muller, P., Quintana, F., Jara, A. et Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer.

Nandram, B., Berg, E. et Barboza, W. (2014). A hierarchical Bayesian model for forecasting state-level corn yield. *Environmental and Ecological Statistics*, 21, 507-530, https://doi.org/10.1007/s10651-013-0266-z.

Park, S., Kim, J. et Stukel, D. (2017). A measurement error model approach to survey data integration: Combining information from two surveys. *Metron*, 75, 345-357.

Pierannunzi, C., Xu, F., Wallace, R., Garvin, W., Greenlund, K., Bartoli, W., Ford, D., Eke, P. et Town, G. (2016). A methodological approach to small area estimation for the Behavioral Risk Factor Surveillance System. *Preventing Chronic Disease*, 14 juillet 2016, 13, http://dx.doi.org/10.5888/pcd13.150480.

Polettini, S. (2017).A generalised semiparametric Bayesian Fay-Herriot model for small area estimation shrinking both means and variances. *Bayesian Analysis*, 2017, 12, 729-752.

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

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

US Census Bureau (2021). SAHIE 2008 – 2015 demographic and income model methodology: Summary for counties and for states. https://www.Census.gov/Small Area Health Insurance Estimates (SAHIE) Program/Technical Documentation/Methodology/Demographic and Income Model Methodology (2008-2015).

Yan, G., et Sedransk, J. (2011). Improved inference for a linear mixed-effects model when the subpopulations are clustered. *Journal of Statistical Planning and Inference*, 141, 3489-3497.

You, Y., et Chapman, B. (2006). Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances d'échantillonnage. *Techniques d'enquête*, 32, 1, 97-103. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006001/article/9263-fra.pdf.