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

Techniques d’enquête

Ajustement de pondération hiérarchique bayésienne et inférence d’enquête

par Yajuan Si, Rob Trangucci, Jonah Sol Gabry et Andrew Gelman

Date de diffusion : le 15 décembre 2020

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

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

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

Programme des services de dépôt

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Techniques d'enquête, décembre 2020
Vol. 46, nº2, pp. 193-228
Statistique Canada, nº 12-001-X au catalogue

Ajustement de pondération hiérarchique bayésienne et inférence d'enquête

Yajuan Si, Rob Trangucci, Jonah Sol Gabry et Andrew Gelman[1]

Résumé

Nous combinons pondération et prédiction bayésienne dans une approche unifiée pour l'inférence d'enquête. Les principes généraux de l'analyse bayésienne impliquent que les modèles pour les résultats d'enquête devraient être conditionnés par toutes les variables influant sur les probabilités d'inclusion. Nous intégrons toutes les variables servant à l'ajustement de pondération dans un cadre de régression multiniveau et de poststratification pour obtenir un sous-produit générant des poids basés sur un modèle après lissage. Nous améliorons l'estimation sur petits domaines en traitant les divers problèmes complexes que posent les applications dans la vie réelle pour obtenir une inférence robuste à des niveaux plus fins pour les sous-domaines d'intérêt. Nous examinons les interactions profondes et introduisons des distributions a priori structurées pour le lissage et la stabilisation des estimations. Le calcul se fait par Stan et avec le paquet `rstanarm` du code source libre R, disponible pour utilisation publique. Nous évaluons les propriétés selon le plan de la procédure bayésienne. Nous recourons à des études en simulation pour illustrer comment la prédiction basée sur un modèle et l'inférence pondérée peuvent donner de meilleurs résultats que la pondération classique. Nous appliquons la méthode à la *New York Longitudinal Study of Wellbeing* (LSW). La nouvelle approche produit des poids lissés et rend plus efficace une inférence robuste de population finie, plus particulièrement pour des sous-ensembles de la population.

Mots-clés : Pondération; prédiction; régression multiniveau et poststratification; distribution a priori structurée.

1 Introduction

1.1 Contexte

En recherche dans le domaine des enquêtes, les méthodes par plan et les méthodes par modèle sont depuis longtemps mises en contraste (Little, 2004). Les premières tiennent automatiquement compte du plan de sondage et les secondes peuvent donner lieu à une inférence robuste dans une estimation de petit échantillon. Rao (2011) présente son appréciation des méthodes fréquentistes et bayésiennes dans la pratique de l'échantillonnage d'enquête. Dans les méthodes classiques par plan, nous recourons à la pondération pour aligner l'échantillon sur la population. On trouvera dans Chen, Elliott, Haziza, Yang, Ghosh, Little, Sedransk et Thompson (2017) un examen de divers estimateurs pondérés d'une moyenne de population. Il reste que, en temps normal, la pondération classique d'enquête fait appel à de nombreux choix définis par l'utilisateur, de sorte que le processus de pondération peut se révéler difficile à codifier dans les enquêtes de la vie réelle (Gelman, 2007). La méthode bayésienne pour l'inférence de population finie (Ghosh et Meeden, 1997) permet d'intégrer l'information antérieure s'il y a lieu, mais des erreurs de spécification de modèle sont possibles.

Dans cet article, nous combinons la prédiction bayésienne et la pondération dans une approche unifiée pour l'inférence d'enquête en appliquant des modèles extensibles et robustes de régression bayésienne pour prendre en compte des caractéristiques complexes de plan de sondage dans un cadre de régression multiniveau et de poststratification (RMP, Gelman et Little (1997); Park, Gelman et Bafumi (2005); Ghitza

1. Yajuan Si, Survey Research Center, Institute for Social Research, Université du Michigan à Ann Arbor. Courriel : yajuan@umich.edu; Rob Trangucci, département de statistique, Université du Michigan; Jonah Sol Gabry, département de statistique, Université Columbia; Andrew Gelman, départements de statistique et de sciences politiques, Université Columbia.

et Gelman (2013); Si, Pillai et Gelman (2015)). La méthode RMP est adaptée à des mécanismes complexes d'échantillonnage et de réponse et vient améliorer l'estimation sur petits domaines (Fay et Herriot, 1979; Rao et Molina, 2015). Nous traitons avec différentes questions complexes liées aux applications réelles et à des niveaux bien plus fins d'inférence de sous-domaine. Notre méthode permet une inférence efficace et valide de population finie, plus particulièrement pour des sous-groupes; elle construit des poids basés sur un modèle après lissage.

Le présent article apporte une double contribution : 1) en innovant en méthodologie bayésienne, nous développons un nouveau cadre structuré a priori pour le traitement des termes d'interaction d'ordre supérieur; 2) pour améliorer la recherche et l'exécution d'enquête, nous combinons la prédiction bayésienne et la pondération en un mode unique d'une approche unifiée pour l'inférence d'enquête en tenant compte des caractéristiques du plan de sondage dans la modélisation bayésienne. Nous généralisons la méthode de régression multiniveau et de poststratification pour une inférence de population finie et construisons une pondération stable et calée par modèle pour résoudre les problèmes de la pondération classique. Nous employons le paquet `rstanarm` en R pour que les méthodes proposées entrent dans le domaine public et nous préconisons l'adoption de méthodes par modèle dans la recherche et l'exécution d'enquête. Aspect plus important encore, nous innovons de manière à employer le cadre RMP en ajustement de pondération d'enquête et en intégration de données, par exemple, en vue d'une inférence avec des enquêtes non probabilistes. Les méthodes que nous proposons donnent un important outil d'ordre pratique pour la conception et la pondération d'échantillons d'enquête (Valliant, Dever et Kreuter, 2018).

1.2 Cadre

Pour une population finie de N unités, nous désignons la variable d'intérêt par $y = (y_1, \ldots, y_N)$ et la variable indicatrice d'inclusion par $I = (I_1, \ldots, I_N)$, où $I_i = 1$ si l'unité i est incluse dans l'échantillon et $I_i = 0$ dans les autres cas. Ici, inclusion signifie sélection et réponse. Le cadre général d'inférence considère la codistribution de I et y. Dans une inférence par plan, nous regardons la distribution de I et traitons y comme fixe. Dans un échantillonnage probabiliste, l'inférence par modèle peut être fondée sur la seule distribution de y, car les variables qui influent sur les mécanismes d'inclusion sont comprises dans le modèle (Royall, 1968), c'est-à-dire dans le mécanisme d'inclusion ignorable lorsque la distribution de I étant donné y est indépendante de la distribution de y (Rubin, 1976, 1983).

Pour tenir compte des facteurs influant sur l'inclusion, la pondération classique par plan ajustée pour les probabilités inégales d'échantillonnage, avec ajustements subséquents à la pondération pour tenir compte des problèmes de couverture et de la non-réponse lors de la collecte ou du nettoyage des données. La pondération classique est donc le produit de facteurs d'ajustement multiples : probabilité inverse de sélection, score inverse de propension de réponse, poststratification (aussi appelée calage ou étalonnage; Holt et Smith (1979)). Chacun de ces ajustements peut être approché en cas d'estimation de la probabilité de sélection, de la probabilité de réponse ou des totaux de population à partir des données. Par-delà les questions d'approximation et même si le modèle d'inclusion est parfaitement connu, des valeurs extrêmes de

pondération causeront une haute variabilité et donc des problèmes d'inférence, surtout lorsque les poids sont en faible corrélation avec la variable de résultat d'enquête (Rao, 1966a, b; Hájek, 1971; Särndal, Swensson et Wretman, 1992). Quand le mode de pondération fait intervenir une poststratification ou une correction de non-réponse – où les poids sont eux-mêmes des variables aléatoires –, l'estimation de la variance ne sera pas la même que dans les cas où la pondération par plan est simplement fixe. Il ne va pas de soi de dégager par l'analyse un estimateur de variance avec un ajustement de pondération à plusieurs degrés ou un plan complexe de sondage.

Dans la pratique, la construction d'une pondération d'enquête exige des décisions quelque peu arbitraires en ce qui concerne la sélection des variables et des interactions, le regroupement de cellules de pondération et l'élagage des poids. On ne voit pas toujours clairement s'il convient d'intégrer de l'information auxiliaire et comment le faire (Groves et Couper, 1995). Dans l'examen de la question du lissage et de l'élagage dans les études consacrées à la pondération d'enquête (voir, par exemple, Potter, 1988, 1990; Elliott et Little, 2000; Elliott, 2007; Xia et Elliott, 2016), on s'est attaché à l'estimation d'un total ou d'une moyenne de population finie et on s'est moins intéressé aux estimations de sous-domaine. Beaumont (2008) propose d'opérer la régression des poids sur les variables d'enquête et de se servir des valeurs prédites comme poids lissés. Cette orientation est inspirante, mais tangentielle à l'objectif d'une inférence aux bonnes propriétés recherchées pour la variable d'intérêt de l'enquête par opposition à la pondération. En empruntant de l'information aux résultats d'enquête, on peut accroître l'efficacité et faire appel à un cadre général.

Gelman (2007) recommande des modèles de régression avec comme covariables des variables influant sur la sélection et la réponse, dont des variables de stratification, des grappes et de l'information auxiliaire. Toutes ces approches peuvent être sensibles à la spécification a priori pour une estimation stable. C'est là la contrepartie par modèle aux décisions nécessaires au lissage ou à l'élagage d'une pondération classique d'enquête. On a proposé des techniques souples de prédiction comme les fonctions spline, les modèles de régression à pénalité et les modèles à arborescence aux fins de l'estimation d'enquête assistée par modèle (Särndal et coll., 1992; Wu et Sitter, 2001; Breidt et Opsomer, 2017; McConville et Toth, 2018).

Les modes d'ajustement de pondération par modèle ou assistés par modèle pour une estimation de total de population finie ont été mis en comparaison par Henry et Valliant (2012). Les méthodes de pondération par modèle dans une perspective de superpopulation (Valliant, Dorfman et Royall, 2000) recourent aux prédictions de modèles de régression pour obtenir une pondération de cas où les prédictions reposent sur des modèles de régression linéaire hiérarchique avec diverses corrections de biais (Chambers, Dorfman et Wehrly, 1993; Firth et Bennett, 1998). À partir de l'estimation de total de population finie, les méthodes assistées par modèle tirent une pondération de cas principalement d'un calage sur variables d'étalonnage (Kott, 2009) par l'estimateur de régression généralisé (GREG, Deville et Särndal (1992)). Il reste que la pondération de cas tirée de prédictions de régression peut être hautement variable et même négative et risque de dégrader certaines estimations de domaine. Les méthodes par modèle jouent un rôle primordial dans l'estimation sur de petits domaines, mais s'exposent aux erreurs de spécification et doivent être développées

plus avant lorsque les domaines sont nombreux et que le mécanisme d'inclusion n'est pas simplement aléatoire.

Pour se garder des erreurs de spécification de modèle, Little (1983) recommande de modéliser les différences de répartition des résultats entre des classes définies par des probabilités différentielles d'inclusion. Si et coll. (2015) construisent des cellules de poststratification fondées sur les valeurs uniques de probabilité d'inclusion et édifient des modèles hiérarchiques pour lisser les estimations des cellules, comme le préconise Little (1991, 1993).

Nous proposons d'employer des modèles hiérarchiques bayésiens tenant compte du plan de sondage pour produire une pondération pouvant servir à une inférence fondée sur le plan de sondage. L'inférence est bien calée et valide et offre de bonnes propriétés fréquentistes (Little, 2011). Dans le cas des grands échantillons, elle sera à mettre en parallèle avec l'inférence par plan. Dans le cas des petits, le lissage par modèle hiérarchique stabilisera l'estimation de domaine et produira un ajustement de pondération robuste.

Nous utilisons les variables intrinsèques servant à la construction d'une pondération par plan, à la correction de non-réponse et au calage en supposant qu'elles sont discontinues et nous construisons des cellules de poststratification en tableau croisé. Nous tirons les poids d'une régression des résultats d'enquête par rapport aux variables de pondération étant donné la poststratification. L'inclusion de la variable de résultat dans la pondération et la poststratification permet d'éviter les erreurs de spécification de modèle et est de nature à accroître l'efficacité (Fuller, 2009). Les estimations de modèle multiniveau rétrécissent les estimations des cellules à la dimension des prédictions venant du modèle de régression. Le cadre RMP tient compte des caractéristiques de plan de sondage dans le paradigme bayésien et est alors bien armé pour traiter les caractéristiques complexes de plan d'échantillonnage. Notre proposition se distingue de la pondération par modèle dont parlent les études spécialisées en employant la structure des cellules de poststratification et en améliorant le tout par lissage, ce qui permet d'éviter les valeurs négatives de pondération.

Si et coll. (2015) intègrent la pondération au cadre RMP et accroissent la souplesse et l'efficacité par rapport à un traitement en pseudo vraisemblance (Pfeffermann, 1993). Nous allons plus loin ici en prenant pour point de départ les variables servant à la pondération et en construisant une pondération par modèle comme sous-produit de la régression multiniveau et de la poststratification. Nous concevons une nouvelle spécification antérieure en régularisation pour pouvoir traiter ce qui peut être une abondance de cellules de poststratification. La construction d'une distribution a priori permet de choisir des variables et de conserver la structure hiérarchique pour les effets principaux et les termes d'interaction d'ordre supérieur de variables catégoriques. En d'autres termes, si une variable n'est pas prédictive, les interactions d'ordre supérieur avec cette variable ne devraient pas l'être non plus, ce qui facilitera l'interprétation du modèle. McConville et Toth (2018) emploient des méthodes à arborescence pour sélectionner automatiquement des poststrates fondées sur des variables auxiliaires en corrélation possible avec les résultats d'enquête. La distribution a priori structurée que nous proposons joue un même rôle avec l'algorithme récursif de partition en facilitant la sélection de poststrates, mais en améliorant en même temps l'efficacité par un regroupement partiel. Nous utilisons les poids lissés et des estimations plus stables que l'estimateur de régression à

arborescence; le cadre bayésien propage toutes les sources d'incertitude là où McConville et Toth (2019) écartent la variance pour l'arborescence et se servent de l'erreur quadratique moyenne pour approcher cette variance.

Pour le calcul, nous avons employé le paquet `rstanarm` en R (Goodrich et Gabry, 2017). Une inférence pleinement bayésienne est réalisable par Stan (Stan Development Team, 2018, 2017), plateforme qui utilise un échantillonnage de Monte Carlo hamiltonien avec segments de chemin adaptés (Hoffman et Gelman, 2014). Stan favorise des approches robustes par modèle en allégeant la charge de calcul que comportent la construction et l'essai de nouveaux modèles. Le paquet `rstanarm` permet une modélisation hiérarchique bayésienne et une inférence pondérée efficaces. Les codes appartiennent au domaine public et sont reproductibles. Notre logiciel perfectionné de calcul procure une plateforme accessible et est de nature à favoriser l'adoption d'un cadre unifié d'inférence d'enquête.

À la section 2, nous posons le problème motivant de la pondération servant aux enquêtes permanentes en sciences sociales. Nous examinons la méthode en détail à la section 3. Nous décrivons à la section 4 l'évaluation statistique d'une inférence à prédiction et à pondération par modèle et démontrons les gains d'efficacité à en attendre par rapport à la pondération classique. À la section 5, nous appliquons notre proposition à une enquête de la vie réelle. À la section 6 enfin, nous résumons les éléments d'amélioration et analysons ce qui devrait suivre.

2 Motivation de l'application

La recherche méthodologique que nous avons faite a pour motivation la pratique opérationnelle de pondération des enquêtes permanentes. Notre but immédiat est de construire une pondération pour la *New York City* (NYC) *Longitudinal Study of Wellbeing* (LSW; Si et Gelman (2014); Wimer, Garfinkel, Gelblum, Lasala, Phillips, Si, Teitler et Waldfogel (2014)), enquête organisée par le *Population Research Center* de l'Université Columbia et qui vise à évaluer l'insuffisance du revenu, les difficultés matérielles et le bien-être de l'enfance et de la famille chez les citadins.

Pour nous, la LSW illustre les questions pratiques de pondération et l'amélioration que nous proposons en sachant que des problèmes semblables se posent dans d'autres enquêtes. La LSW comporte un échantillon téléphonique à composition aléatoire et un échantillon en personne et en adaptation au répondant qui est formé des bénéficiaires des services philanthropiques de l'organisme Robin Hood ainsi que de leurs connaissances. Nous nous intéressons ici à une enquête téléphonique en guise d'illustration. Dans la LSW, on interroge au téléphone 2 002 résidents adultes de la ville de New York, 500 par téléphone cellulaire et 1 502 par téléphone ordinaire. La moitié des unités échantillonnées du volet « téléphone ordinaire » viennent de secteurs à faible revenu délimités par les codes zip. Les échantillons de base observés font l'objet d'un suivi aux trois mois. Nous alignons ces échantillons sur les enregistrements de l'*American Community Survey* (ACS) de 2011 pour la ville de New York. Les écarts tiennent principalement à un suréchantillonnage des quartiers à faible revenu et à la non-réponse.

Le mode de pondération de base (Si et Gelman, 2014) tient compte des probabilités inégales de sélection, du biais de couverture et de la non-réponse. La pondération classique s'obtient par les probabilités inverses d'inclusion estimées et par la méthode itérative du quotient (Deville, Särndal et Sautory, 1993). Toutefois, les praticiens doivent faire des choix arbitraires ou subjectifs au moment d'opérer leur sélection de facteurs de pondération avec leurs valeurs. S'il s'agit, par exemple, de construire une pondération pour des gens d'âge adulte, nous devons majorer le poids des répondants des ménages de grande taille, car un seul adulte figurera dans l'échantillon par ménage échantillonné. Gelman et Little (1998) recommandent de prendre la racine carrée du rapport entre taille de ménage et taille de famille pour cet ajustement de pondération, parce qu'une pondération par taille de ménage (ACS Weighting Method, 2014, par exemple) tend à produire une surcorrection dans les enquêtes téléphoniques. Dans la pratique, la méthode itérative du quotient tient compte des facteurs sociodémographiques sans adaptation précise aux particularités des enquêtes.

Les organisateurs d'enquêtes s'intéressent aux aspects de la qualité de vie des citadins et regardent, par exemple, la proportion d'enfants en butte à la pauvreté et aux difficultés matérielles. Il importe donc de dégager des estimations exactes pour des sous-populations. Nous souhaitons concevoir une procédure objective et laisser les données d'enquête recueillies déterminer le mode de pondération. Le principe fondamental est de prendre en compte dans la pondération toutes les variables susceptibles d'influer sur la sélection et la réponse. Idéalement, nous nous attendrions à ce que les variables servant à la pondération tiennent compte de la disponibilité de téléphones (nombre de téléphones cellulaires ou ordinaires et durée d'interruption du service téléphonique), de la structure des familles et des ménages, des caractéristiques sociodémographiques et peut-être de leurs termes d'interaction d'ordre supérieur. Les enregistrements de l'ACS nous renseignent seulement sur la taille de la famille, l'âge, l'origine ethnique, le sexe, l'éducation et l'écart de pauvreté (mesure de la pauvreté des familles). Les organisateurs d'enquêtes recommandent d'inclure ce qui suit dans la pondération pour combler l'écart de distribution avec la population : objet de l'analyse de fond, variables du nombre de personnes âgées et d'enfants dans la famille et de la taille de l'unité familiale, interactions de ces facteurs avec les écarts de pauvreté.

Pour construire une pondération classique, nous choisissons les facteurs de proportion qui, dans la méthode itérative du quotient, peuvent influer sur la sélection et la réponse comme le sexe, l'âge, l'éducation, l'origine ethnique, l'écart de pauvreté, le nombre d'enfants, de personnes âgées et de personnes en âge de travailler dans la famille, les interactions binaires entre l'écart de pauvreté, d'une part, et l'âge et les nombres respectifs de membres, d'enfants et de personnes âgées dans la famille, d'autre part. Nous prenons les distributions marginales de l'enquête ACS et procédons à un ajustement par la méthode itérative du quotient. Les poids ainsi produits doivent être élagués à cause d'un certain nombre de valeurs extrêmes.

Il est néanmoins possible que l'ajustement de pondération à caractère subjectif comporte un certain nombre de variables ou d'interactions non essentiellement prédictives ou ne tienne pas compte de tous les facteurs d'un intérêt fondamental par la suite. Dans l'ajustement par la méthode itérative du quotient, nous posons que ces facteurs sont indépendants. Il s'ensuivra un biais d'inférence de domaine en tableau croisé si la structure de corrélation n'est pas la même dans l'échantillon et dans la population. Idéalement, il devrait y

avoir appariement par la codistribution de ces variables liées de pondération. Il reste que des cellules petites ou vides pour les interactions profondes créeront des poids extrêmement importants qui obligeront à regrouper des cellules.

Les problèmes que pose la pondération classique pour l'enquête de base LSW sont le reflet des problèmes que pose le plus souvent la pratique de la pondération dans l'exécution des enquêtes réelles, lesquelles sont souvent compliquées à cause de plans de sondage complexes, d'une structure longitudinale ou de mécanismes de réponse à plusieurs degrés. Les décisions spéciales qui doivent fréquemment être prises avec les régimes classiques de pondération peuvent avoir pour résultat que la pondération variera selon les praticiens pour une même enquête. Pour prévenir toute subjectivité, il importe de proposer une procédure de pondération par modèle permettant de laisser aux données le choix des facteurs de pondération. Nous aimerions intégrer les variables de pondération dans le modèle des résultats d'enquête pour des gains d'efficacité, modéliser leurs termes d'interaction d'ordre supérieur dans une distribution a priori régularisée et produire des poids se prêtant à un même traitement que dans la pondération classique. Un grand nombre de variables de pondération et d'interactions profondes apporteront de petites cellules de pondération en tableau croisé. Les petites cellules demandent un ajustement statistique de lissage et de stabilisation.

La régression multiniveau avec poststratification a eu du succès dans l'estimation de domaine à des niveaux bien plus fins. En empruntant la puissance d'un cadre hiérarchique de modélisation avec une distribution a priori informative, nous devrions pouvoir estimer après lissage des cellules éparses. Une poststratification par l'information du recensement permettra un appariement de l'estimation entre l'échantillon et la population. La combinaison d'une régression et d'une poststratification rappelle le concept de poststratification endogène (Breidt, 2008; Dahlke, Breidt, Opsomer et Keilegom, 2013). Nous décrirons en détail le cadre de régression multiniveau et de poststratification.

3 Méthode

3.1 Régression multiniveau et poststratification

Dans une configuration de base, nous nous proposons d'estimer la distribution de population du résultat d'enquête y. Avec un mode de pondération transparent, nous pouvons directement inclure la variable auxiliaire X dans la modélisation de régression pour y. Ici, X est un vecteur q-dimensionnel de variables influant sur le plan de sondage, la non-réponse et la couverture. Conditionnée par X, la distribution de la variable indicatrice d'inclusion I est ignorable.

La sélection de variables auxiliaires et la disponibilité de leurs codistributions dans la population sont la clé du succès pour la méthode RMP et pour toutes les autres méthodes s'il s'agit de tenir compte du biais de sélection d'échantillon et de non-réponse et d'obtenir des inférences valides de population. Nous recommandons d'inclure toutes les variables susceptibles d'influer sur l'inclusion dans l'échantillon, qu'il

soit question de l'information de plan de sondage, des paradonnées ou des caractéristiques sociodémographiques. Un avantage avec la méthode RMP réside dans la possibilité de choisir les variables et de stabiliser les poids par opposition à une pondération classique entachée de bruit.

Un autre problème d'ordre pratique est que la distribution de population des variables de calage peut être inconnue. Nous obtenons la codistribution de l'ACS comme variable de contrôle de population dans notre étude d'application. Wang, Rothschild, Goel et Gelman (2015), Zhang, Holt, Yun, Lu, Greenlund et Croft (2015) et Yougov (2017) se sont respectivement servis de données agrégées de sortie, de données de secteurs de recensement et des données de la Current Population Survey pour obtenir directement l'information nécessaire à l'ajustement de poststratification. Dans la pratique, nous recommandons de tirer l'information de population directement du recensement ou de grandes études aux erreurs minimales ou encore de l'estimer à l'aide de l'information disponible d'études apparentées. La distribution de population de certaines variables auxiliaires pourrait ne pas être disponible dans la base des données du recensement comme le nombre d'appareils téléphoniques, et nous pouvons estimer cette variable à partir d'autres enquêtes comme échantillons de référence. Reilly, Gelman et Katz (2001) ont appliqué des modèles pour prédire l'information inconnue de population en poststratification. Là où des distributions marginales sont disponibles, Little et Wu (1991) proposent une modélisation équivalente pour la méthode itérative du quotient et Si et Zhou (2020), une estimation bayésienne par cette même méthode du quotient pour l'estimation de taille des cellules de population. Nous proposons de pousser cette démarche et de mettre au point un cadre d'intégration pour l'incertitude de l'estimation de l'information de contrôle inconnue à la section 6. La disponibilité d'une information de contrôle de population d'une grande qualité et d'une puissance prédictive influe directement sur la validité de l'inférence par modèle ou par plan.

Dans le cadre RMP, les variables auxiliaires X sont discontinues; nous construisons en tableau croisé les cellules de poststratification j avec la taille de cellule de population N_j et la taille de cellule d'échantillon n_j, pour $j = 1, \ldots, J$, où J est le nombre total de cellules de poststratification (Little, 1991, 1993; Gelman et Little, 1997; Gelman et Carlin, 2001). La taille totale de population est alors $N = \sum_{j=1}^{J} N_j$ et la taille d'échantillon, $n = \sum_{j=1}^{J} n_j$.

L'inférence poststratifiée diffère de l'inférence par plan d'échantillonnage stratifié du fait que les n_j sont maintenant des fonctions aléatoires de la distribution d'échantillonnage I. Dans l'échantillonnage répété de I, la probabilité est non nulle que $n_j = 0$ pour certains j. On résout habituellement ce problème avec un conditionnement par les n_j observés dans l'échantillon réalisé, mais l'inférence d'échantillon n'est pas exempte d'un biais de plan de sondage conditionné par les n_j. Le cadre de régression multiniveau et de poststratification pose qu'un modèle pour les n_j tiendra compte déjà des caractéristiques du plan de sondage.

Dans la poststratification, nous supposons implicitement que les unités sont équiprobables dans les diverses cellules. Si θ est l'estimande de population comme la moyenne d'ensemble ou de domaine et qu'il peut s'exprimer comme somme pondérée sur tout sous-ensemble ou domaine D des poststrates,

$$\theta = \frac{\sum_{j \in D} N_j \theta_j}{\sum_{j \in D} N_j}, \tag{3.1}$$

où θ_j est l'estimande correspondant de la cellule j. L'estimateur poststratifié proposé sera de la forme générale

$$\tilde{\theta}^{\mathrm{PS}} = \frac{\sum_{j \in D} N_j \tilde{\theta}_j}{\sum_{j \in D} N_j}, \tag{3.2}$$

où $\tilde{\theta}_j$ est l'estimation correspondante dans la cellule j. Diverses méthodes de modélisation sont possibles pour établir les estimations des cellules comme les modèles bayésiens non paramétriques souples et les algorithmes d'apprentissage machine (Rasmussen et Williams (2006); Hastie, Tibshirani et Friedman (2009)). Nous emploierons ici un modèle de régression hiérarchique en guise d'illustration.

Dans la pratique, un poids d'enquête s'attache à chaque unité, bien que les poids ne soient pas des attributs individuels des unités. Il est naturel de produire une pondération au niveau des unités en se fondant sur tout le plan de sondage et d'utiliser les moyennes pondérées de cette forme comme $\tilde{\theta} = \sum_{i=1}^{n} w_i y_i \Big/ \sum_{i=1}^{n} w_i$. Notre but ici est d'obtenir un jeu équivalent de poids w_i au niveau des unités à l'aide d'une procédure par modèle d'estimation de $\tilde{\theta}^{\mathrm{PS}}$ pour lier pondération et poststratification. Les modèles de régression peuvent donc servir à obtenir des comptes de poststratification $\tilde{\theta}_j$ pour l'information de population; la pondération par modèle est alors recalculée par l'expression en (3.2).

Dans les modèles classiques de régression, une poststratification intégrale demeure un cas d'espèce où les estimations se calculent séparément pour les diverses cellules sans effet collectif et, par conséquent, sans regroupement de cellules. Si nous nous intéressons, par exemple, à la moyenne de population, la moyenne des cellules sera alors notre estimation. En règle générale, les modèles classiques de régression s'exécutent sur les caractéristiques des cellules sans un ajustement extrême de modèle pour chaque cellule. Si plus d'interactions entre les caractéristiques sont prévues, la pondération résultante devient plus variable. Par ailleurs, un regroupement intégral fait fi de l'hétérogénéité entre cellules. Des modèles hiérarchiques de régression se trouveront à lisser les estimations des variables dans un regroupement partiel.

Gelman (2007) prend le modèle normal échangeable pour illustrer et montre que l'estimation $\tilde{\theta}^{\mathrm{PS}}$ en poststratification de la moyenne de population peut s'exprimer comme moyenne pondérée entre la moyenne des cellules et la moyenne d'ensemble, ce qui donne les poids unitaires, tout comme en moyenne pondérée entre des poids intégralement lissés, $w_j = 1$, et des poids en poststratification intégrale, $w_j = (N_j/N)/(n_j/n)$. La poststratification hiérarchique équivaut en gros à un rétrécissement des poids par un rétrécissement des estimations des paramètres. Le degré de rétrécissement tend vers zéro à mesure que croît la taille d'échantillon, d'où l'implication que les estimations du modèle seront proches des valeurs réelles par le plan de sondage. Il faudra cependant développer l'approche pour le traitement d'un grand nombre de cellules et d'interactions profondes, ainsi qu'évaluer rigoureusement le rendement d'une pondération par modèle.

Dans notre application de l'étude LSW, les variables servant à la pondération sont notamment l'âge (cinq catégories), l'origine ethnique ou la race (cinq catégories), l'éducation (quatre catégories), le sexe (deux catégories), la mesure de la pauvreté (cinq catégories), la taille de la famille (quatre catégories), le nombre de personnes âgées (trois catégories) et d'enfants (quatre catégories) dans la famille, ce qui donne $J = 5 \times 5 \times 4 \times 2 \times 5 \times 3 \times 4 \times 4 = 48\,000$ poststrates. Les cellules de poststratification sont en majeure partie vides ou éparses en raison de la taille limitée d'échantillon (2 002 unités). Les tailles des cellules d'échantillon sont en déséquilibre. Souvent, les cellules sont arbitrairement regroupées ou combinées (Little, 1993) sans que la chose se justifie en théorie. Les procédures récentes de lissage par modèle des poids entre les cellules ne pouvaient traiter ces cellules éparses (Elliott et Little, 2000). Xia et Elliott (2016) ont introduit une distribution a priori laplacienne pour le lissage des poids sur un nombre modeste de poststrates en fonction des probabilités d'inclusion, mais en écartant les variables servant à la pondération et leur structure hiérarchique. Avec le cadre RMP, nous tenons compte de la structure hiérarchique des variables pour lisser et regrouper les estimations sur les tailles de cellules éparses et déséquilibrées en employant un ensemble nouveau de distributions a priori.

3.2 Distribution a priori structurée

Nous présentons ici une distribution a priori structurée pour améliorer la régression multiniveau et la poststratification dans le cas des structures de cellules éparses et déséquilibrées, ce qui donne une pondération stable d'enquête par modèle qui tient compte de l'information du plan de sondage. Supposons que la distribution de population de X est connue et que nous pouvons tirer les N_j de données externes pour décrire une codistribution des variables servant à la pondération. L'extension à des N_j inconnus est examinée à la section 6. Dans la pratique, les cellules de poststratification J peuvent être nombreuses; leur nombre peut même l'emporter très largement sur la taille de l'échantillon n. Les variables de pondération pourraient influer sur l'inclusion à cause d'une relation complexe ou d'un mécanisme différentiel de réponse. Les interactions profondes sont essentielles dans une structure de relation complexe, mais nous ne pouvons les inclure toutes et devons choisir ce qui sera prédictif comme interactions et effets principaux.

Posons que la réponse d'enquête recueillie est continue, y_i pour $i = 1, \ldots, n$. Nous nous intéressons ici à l'estimation $\bar{Y}$ de la moyenne de population. Nous employons $\left(X^{1\mathrm{T}}, \ldots, X^{J\mathrm{T}}\right)^{\mathrm{T}}$ pour représenter la matrice de prédiction $J \times q$ dans la population avec le cadre de poststratification. Pour illustrer, posons une distribution normale,

$$y_i \sim N\left(\theta_{j[i]},\ \sigma_y^2\right), \tag{3.3}$$

où $j[i]$ désigne la cellule j à laquelle appartient l'unité i. Nous pouvons aussi penser à des variances inégales et nous laisserons l'échelle de cellule σ_y varier entre les cellules en indexation par σ_j. Pour la spécification de distribution a priori de θ_j, un choix possible est $\theta_j = X^j\beta$ où β reçoit une certaine distribution a priori. Dans l'exemple de régression hiérarchique de Gelman (2007), une distribution normale à plusieurs variables est considérée, $y_i \sim N\left(X_i\beta, \Sigma_y\right)$ et $\beta \sim N\left(0, \Sigma_\beta\right)$, où les covariables

tiennent compte de tous les effets principaux et de quelques interactions binaires choisies dans X et où la matrice des covariances Σ_β est diagonale et à échelles différentes. Ce modèle peut être mal spécifié et certains poids produits pourraient être négatifs.

Comme X^j consiste en indicateurs à niveaux différents des q variables auxiliaires discontinues, nous pouvons exprimer la moyenne de cellule de population θ_j par

$$\theta_j = \alpha_0 + \sum_{k \in S^{(1)}} \alpha_{j,k}^{(1)} + \sum_{k \in S^{(2)}} \alpha_{j,k}^{(2)} + \ldots + \sum_{k \in S^{(q)}} \alpha_{j,k}^{(q)}, \tag{3.4}$$

où $S^{(l)}$ est l'ensemble des termes d'interaction l-multiples possibles et où $\alpha_{j,k}^{(l)}$ représente le k^{e} des l-multiples termes d'interaction de l'ensemble $S^{(l)}$ pour la cellule j. Ainsi, les $\alpha_{j,k}^{(1)}$ avec $k \in S^{(1)}$ correspondent aux effets principaux et les $\alpha_{j,k}^{(2)}$ avec $k \in S^{(2)}$ aux termes d'interaction binaire pour la cellule j. Cette décomposition rend compte de toutes les interactions possibles des q variables. Avec une structure de cellules éparses, une sélection de variables est nécessaire. Dans la pratique, nous recommandons d'inclure initialement les covariables et les termes d'interaction d'importance fondamentale et d'intérêt scientifique dans le modèle (3.4) et de procéder à une sélection bayésienne de variables dans ce que nous proposons comme distribution a priori structurée.

Nous induisons des distributions a priori structurées pour pouvoir traiter les interactions profondes et tenir compte de leur structure hiérarchique; les termes d'interaction d'ordre supérieur seront exclus si un des effets principaux correspondants n'est pas choisi. Plus augmentent les effets principaux, plus augmentent aussi les effets des termes d'interaction en question. Idéalement, il devrait y avoir plus de rétrécissement sur les interactions d'ordre supérieur que sur les effets principaux et la distribution a priori devrait traduire la structure d'imbrication. Ce défi est le problème qui se pose en inférence bayésienne pour les paramètres de variance au niveau du groupe dans une structure d'analyse de variance (Gelman, 2005, 2006). Volfovsky et Hoff (2014) présentent une classe de distributions a priori hiérarchiques pour des ensembles d'interactions en adaptation à une similarité possible entre niveaux adjacents, là où la matrice des covariances pour les interactions d'ordre supérieur est posée comme le produit de Kronecker des matrices des covariances des effets principaux après rajustement des grandeurs relatives. Nous développons notre proposition en induisant plus de structure entre les paramètres de variance, plus de rétrécissement et un effet de lissage de manière à pouvoir traiter un nombre extrêmement grand de cellules aux tailles déséquilibrées par rapport à la structure généralement équilibrée de Volfovsky et Hoff (2014), ce qui permet un calcul d'un meilleur rendement.

Nous prenons comme point de départ des distributions a priori indépendantes sur les paramètres de régression α :

$$\alpha_{j,k}^{(l)} \sim N\left(0,\ \left(\lambda_k^{(l)}\sigma\right)^2\right),$$

où $\lambda_k^{(l)}$ représente l'échelle locale et où σ est l'échelle d'erreur globale pour $k \in S^{(l)}$ et $l = 1, \ldots, q$. L'échelle d'erreur est la même pour les effets principaux et les interactions d'ordre supérieur, alors que les

échelles locales sont différentes. L'effet de rétrécissement est induit par la spécification des échelles locales. Nous posons que l'échelle locale des interactions d'ordre supérieur est le produit des échelles locales des effets principaux correspondants après rajustement des grandeurs relatives :

$$\lambda_k^{(l)} = \delta^{(l)} \prod_{l_0 \in M^{(k)}} \lambda_{l_0}^{(l)},$$

où $\delta^{(l)}$ est le rajustement de grandeur relative et $M^{(k)}$ l'ensemble d'effets principaux correspondants qui construisent la k^{e} interaction l-multiple dans l'ensemble $S^{(l)}$. Ainsi, l'échelle locale de l'interaction trinaire de l'âge, du sexe et de l'éducation chez les hommes d'âge moyen ayant fait des études collégiales sera le produit des échelles locales des effets principaux sur l'âge, le sexe et l'éducation, c'est-à-dire le produit des paramètres d'échelle locale respectivement pour les hommes d'âge moyen ayant fréquenté le collège.

Nous utilisons les hyperdistributions a priori suivantes pour les paramètres d'échelle :

$$\begin{aligned}
&\text{échelle d'erreur : } \sigma \sim \text{Cauchy}_+(0,1) \\
&\text{échelle locale des effets principaux : } \lambda_k^{(1)} \sim N_+(0,1) \\
&\text{échelle locale des interactions d'ordre supérieur : } \lambda_k^{(l)} = \delta^{(l)} \prod_{l_0 \in M^{(k)}} \lambda_{l_0}^{(1)} \qquad (3.5) \\
&\text{grandeur relative des interactions d'ordre supérieur : } \delta^{(l)} \sim N_+(0,1), \text{ pour } l = 2, \ldots, q.
\end{aligned}$$

Dans ce cas, Cauchy_+ et N_+ désignent la partie positive des distributions de Cauchy et normale respectivement. Gelman (2006) propose une distribution a priori semi-Cauchy pour le paramètre d'échelle des modèles hiérarchiques, une propriété intéressante étant que l'on peut obtenir des valeurs d'échelle arbitrairement proches de 0 avec de lourdes queues aux valeurs importantes lorsqu'elles sont confirmées par les données. Avec $\lambda_k^{(l)}$ proche de 0, les échantillons a posteriori de $\alpha_{j,k}^{(l)}$ sont en rétrécissement vers 0. Le paramètre d'échelle des termes d'interaction d'ordre supérieur sera nul si un des paramètres d'échelle liés des effets principaux est également nul. L'effet global de régularisation est déterminé par l'échelle d'erreur et les paramètres d'échelle multiplicative des effets principaux correspondants. Nous attribuons une distribution a priori non informative au terme à l'origine et des distributions a priori faiblement informatives aux deux paramètres d'échelle d'erreur globale (σ_y, σ), où $\sigma_y \sim \text{Cauchy}_+(0, 5)$.

La distribution a priori en rétrécissement global-local peut stabiliser la modélisation d'effets aléatoires dans l'estimation sur petits domaines (Tang, Ghosh, Ha et Sedransk, 2018). La spécification de distribution a priori que nous proposons se caractérise par ce rétrécissement global-local et une sélection de groupe pour tous les indicateurs de niveau possibles de la même variable dans ce qui ressemblera à un « lasso de groupe » (Yuan et Lin, 2006). Nous parvenons à sélectionner des variables selon la même spécification avec la distribution a priori horseshoe (Carvalho, Polson et Scott, 2010) et nous améliorons le tout par une sélection

de groupe et des échelles multiplicatives pour les interactions d'ordre supérieur, d'où des gains en matière de cellules éparses. Nous prenons des distributions a priori semi-Cauchy faiblement informatives pour les échelles d'erreur et des distributions a priori semi-normales informatives pour les paramètres d'échelle locale, ce qui améliore l'estimation de rétrécissement des paramètres et l'efficacité du calcul. Si l'estimation a posteriori du paramètre d'échelle est proche de 0, indice que la variable n'est pas prédictive, on peut en post traitement exclure cette variable de la construction des cellules de poststratification pour une réduction de dimension. Cette classe de distributions a priori permet une sélection de variables de forte dimension. Elle permet en outre de conserver la structure hiérarchique entre effets principaux et interactions.

Piironen et Vehtari (2016) recommandent le choix a priori d'hyperparamètres de rétrécissement global en fonction de ce qu'on croit antérieurement être le nombre de coefficients non nuls dans le modèle. Une configuration hiérarchique avec des variables corrélées exige plus d'investigation. Nous prenons par défaut le choix de la distribution de Cauchy$_+$ (0, 1) et soumettons à une vaste analyse de sensibilité la spécification des hyperparamètres où les résultats ne changent pas.

3.3 Pondération par modèle

Nous pouvons reformuler (3.4) et (3.5) comme modèle normal échangeable :

$$\theta_j \sim N(\alpha_0, \sigma_\theta^2), \quad \sigma_\theta^2 = \sum_{l=1}^{q} \sum_{k \in S^{(l)}} \left(\lambda_k^{(l)} \sigma\right)^2. \tag{3.6}$$

Conditionnée par les paramètres de variance, la moyenne a posteriori est une fonction linéaire des données dans le modèle normal avec distribution normale a priori. Ainsi, nous pouvons déterminer des *poids équivalents* w_i^* de manière à pouvoir reformuler l'estimation lissée $\sum_{j=1}^{J} N_j \big/ N\tilde{\theta}_j$ comme moyenne pondérée classique, $\sum_{i=1}^{n} w_i^* y_i \big/ \sum_{i=1}^{n} w_i^*$. En combinant les estimations de moyenne a posteriori pour θ_j et l'estimation par modèle donnée par l'expression en (3.2), Gelman (2007) calcule les poids unitaires équivalents pouvant être utilisés sur le mode classique dans la cellule j :

$$w_j \approx \frac{n_j \big/ \sigma_y^2}{n_j \big/ \sigma_y^2 + 1 \big/ \sigma_\theta^2} \cdot \frac{N_j \big/ N}{n_j \big/ n} + \frac{1 \big/ \sigma_\theta^2}{n_j \big/ \sigma_y^2 + 1 \big/ \sigma_\theta^2} \cdot 1, \tag{3.7}$$

où le poids par modèle est une moyenne pondérée entre une poststratification intégrale sans regroupement (poids de $(N_j \big/ N) \big/ (n_j \big/ n)$) et un regroupement intégral (équipondération unitaire). Le facteur de regroupement ou de rétrécissement est $1 \big/ \left(1 + n_j \sigma_\theta^2 \big/ \sigma_y^2\right)$ et dépend du groupe et des variances individuelles, tout comme de la taille d'échantillon dans la cellule. Les poids par modèle sont des variables aléatoires et une inférence pleinement bayésienne propagera la variabilité correspondante. Nous recueillons les valeurs de moyenne a posteriori et les traitons comme des poids pouvant aussi être utilisés comme des poids classiques.

3.4 Calcul

La procédure d'inférence par prédiction bayésienne hiérarchique et pondération est reproductible et extensible. Nous appliquons les distributions a priori structurées que nous proposons dans le paquet `rstanarm` du code source libre R (Goodrich et Gabry, 2017). Le code de ce calcul est disponible en ligne (Si, Trangucci et Gabry, 2020) au public. Nous présentons comme exemple le code de l'application en données réelles à l'annexe A pour faire la démonstration d'une interface de calcul conviviale et efficace que les praticiens des enquêtes peuvent utiliser et adapter en toute simplicité. L'inférence pleinement bayésienne s'effectue à l'aide de Stan, logiciel convivial de code source libre qui contribue à l'application généralisée de la modélisation bayésienne. Les praticiens des enquêtes résistent aux approches par modélisation principalement à cause de la charge de calcul. Il reste que les méthodes par modèle sont prêtes à relever les nouveaux défis des mégadonnées d'enquête (structures déséquilibrées de cellules, combinaisons d'enquêtes, analyses de flux de données, etc.). Le développement de Stan peut venir bonifier la généralisation de l'approche par modélisation et apporter la plateforme de calcul nécessaire à un cadre unifié d'inférence d'enquête.

Dans notre application, les échantillons de Monte Carlo à chaîne de Markov se dosent bien et les chaînes convergent rapidement. La rapidité du calcul élargit l'utilité des modes d'inférence d'enquête par modèle. La spécification de distribution a priori que nous proposons rend plus stable la pondération lissée avec un regroupement partiel. Nous comparons la pondération par modèle à la pondération classique aux sections 4 et 5 pour démontrer ce qu'est le calage pour les propriétés de plan de sondage (Little, 2011). De plus, nous illustrons l'amélioration proposée de l'estimation de domaine dans des structures de cellules d'échantillon déséquilibrées et éparses.

4 Études en simulation

Nous évaluons la procédure bayésienne par les propriétés de plan de sondage et en démontrons la validité. Nous regardons deux grands scénarios de simulation, soit une structure légèrement déséquilibrée avec un nombre modéré de cellules de poststratification et une structure très déséquilibrée avec un grand nombre de cellules de poststratification. Nous évaluons la validité statistique de l'estimation pondérée par modèle de population finie et de l'inférence de domaine pour démontrer le gain de capacité de solution des problèmes de la pondération classique. Pour illustrer la capacité de sélection de variables et de maintien de la hiérarchie avec les gains d'efficacité qui s'ensuivent, nous comparons l'estimation a posteriori à celle d'une distribution a priori indépendante, mais sans la contrainte d'échelle multiplicative, ce qui ressemblera à la distribution a priori horseshoe sous spécification de groupe comme distribution a priori indépendante dans notre exposé : $\lambda_k^{(l)} \sim N\left(0,\ \left(\sigma_k^{(l)}\right)^2\right)$.

Nous examinons les prédictions par modèle avec distribution a priori structurée (Str-P) et distribution a priori indépendante (Ind-P). Pour une inférence pondérée, nous évaluons l'estimation après avoir appliqué la pondération par modèle avec distribution a priori tant structurée qu'indépendante, la pondération avec

méthode itérative du quotient (Rake-W), la pondération classique de poststratification (PS-W) et la pondération à probabilités inverses de sélection (IP-W). Nous présentons les outils graphiques de diagnostic permettant de comparer les poids et l'inférence pondérée.

Nous nous reportons à l'ACS de 2011 auprès des résidents adultes de la ville de New York comme « population » et en tirons au hasard des échantillons par le modèle préspécifié de sélection sans non-réponse. Nous prenons les covariables de l'ACS et simulons la variable de résultat afin d'obtenir la distribution réelle comme étalon. Les détails des spécifications de modèle sont donnés à l'annexe B pour les scénarios qui suivent. Nous appliquons la méthode itérative du quotient en équilibrant les distributions marginales des variables de calage dans le modèle de sélection et produisons la pondération par cette même méthode du quotient. Nous obtenons les poids classiques de poststratification N_j / n_j en appariant les indices des cellules de l'échantillon et des cellules de la population. Le modèle de sélection peut nous donner des poids en probabilité inverse de sélection par appariement des indices des unités de l'échantillon. Nous produisons aussi une pondération par modèle avec des distributions a priori indépendantes pour les effets principaux et les termes d'interaction d'ordre supérieur des variables de l'ACS. Nous normalisons les poids produits à la moyenne 1 pour faciliter la comparaison.

4.1 Structure légèrement déséquilibrée

Nous traitons d'abord une structure légèrement déséquilibrée où le nombre de cellules de poststratification et la taille des cellules de l'échantillon sont modérés. Nous procédons par échantillonnage répété pour examiner les propriétés fréquentistes des prédictions par modèle et des inférences pondérées. Comme il n'y a guère d'effet de rétrécissement sur les interactions d'ordre supérieur, la prédiction et la pondération par modèle avec des distributions a priori structurées sont d'un même rendement qu'avec des distributions a priori indépendantes, tout en dépassant le rendement de la pondération classique.

Posons que trois variables sont incluses dans les modèles de sélection et de résultat, à savoir l'âge, l'origine ethnique et l'éducation. Nous employons les trois variables de l'ACS en discontinu avec *age* (18–34, 35–44, 45–54, 55–64, 65+), *eth* (Blancs non hispanophones, Noirs non hispanophones, Asiatiques, Hispanophones et autres) et *edu* (moins que l'école secondaire, école secondaire, années de collège, baccalauréat ou études supérieures). Le nombre de cellules de poststratification est de $5 \times 5 \times 4 = 100$. Nous posons que le résultat dépend des interactions profondes avec tous les effets principaux et les termes d'interaction binaire et trinaire entre les trois variables. Nous posons aussi que la variable indicatrice de sélection dépend des trois effets principaux. Les valeurs spécifiques des coefficients figurent aux tableaux B.2 et B.3 à l'annexe B. Nous fixons les valeurs de manière à traduire l'étroitesse de la corrélation entre les covariables et les variables dépendantes. Et les effets ne sont pas nécessairement les mêmes entre les niveaux adjacents des facteurs, ce qui diffère des scénarios dans Volfovsky et Hoff (2014). Nous fixons à 1 l'échelle d'erreur du modèle de résultat, la valeur réelle étant toujours intégralement tirée de l'estimation a posteriori. Le modèle de génération de données diffère du modèle d'estimation, mais ce dernier est assez souple pour comprendre le premier, puisque la structure de dépendance est recouvrée par l'estimation. Notre proposition est robuste à l'égard des erreurs de spécification de modèle.

Nous répétons l'échantillonnage 500 fois. Les tailles d'échantillon varient de 2 141 à 2 393 et la médiane est de 2 288. Nous relevons des cellules vides dans l'échantillon avec des probabilités de sélection (variant de 0,001 à 0,269) étalées sur le processus d'échantillonnage répété. Le nombre de cellules occupées dans l'échantillon est de 80 à 93 pour une médiane de 87. Une structure légèrement déséquilibrée des cellules est fréquente dans les enquêtes pratiques dont le plan de sondage est simple et propre. Les quantités de population d'intérêt sont la moyenne générale, la moyenne de domaine sur les 13 (= 5 + 4 + 4) niveaux marginaux des trois variables et la moyenne de domaine pour les jeunes non blancs (un exemple d'interaction entre l'âge et l'origine ethnique/race). Nous examinons la valeur absolue du biais d'estimation, la racine de l'erreur quadratique moyenne (REQM), l'erreur-type (ET) approchée par la valeur moyenne des écarts-types (e.-t. moyen) et le taux nominal de couverture des intervalles de confiance à 95 %.

Les résultats à la figure 4.1 indiquent que les prédictions par modèle se caractérisent entre toutes les méthodes par le REQM et l'ET les plus bas avec des taux raisonnables de couverture et un biais comparable. Toutes les variables influant sur le mécanisme de résultat et de sélection sont comprises dans le modèle suivant le principe bayésien d'un mécanisme d'échantillonnage ignorable. Le modèle prédira toutes les estimations des cellules, y compris des cellules vides de l'échantillon, dans ce qui représente une utilisation intégrale de l'information de population et de la structure des cellules de poststratification. L'inférence pondérée est conditionnelle aux unités observées dans les cellules occupées et est donc moins efficace que les prédictions par modèle. Généralement parlant, l'inférence sur pondération par modèle présente des valeurs REQM et ET moindres, mais des taux de couverture plus raisonnables que dans la pondération classique. L'ajustement par la méthode itérative du quotient ne vaut pas pour l'estimation de domaine avec un biais et une REQM importants et une piètre couverture, bien que le mécanisme de sélection dépende seulement des effets principaux. L'inférence sur pondération à probabilités inverses de sélection tend à présenter des ET élevées mais des taux de couverture faibles, plus particulièrement pour l'estimation de domaine. L'inférence sur pondération de poststratification est proche de l'estimation pondérée par modèle, car les tailles de domaine sont modestement grandes. L'effet de rétrécissement des cellules vers des poids nuls est ténu (de 0 à 0,19 pour une moyenne de 0,05) lorsque le plan de sondage est légèrement déséquilibré. Le nombre de cas ayant moins que les études secondaires est petit (80 environ) d'où un grand biais d'estimation et une ET importante pour les inférences pondérées, mais non dans les prédictions par modèle. Ces prédictions stabilisent l'estimation sur petits domaines par lissage, comme on peut le voir au tableau 4.1 qui présente la comparaison numérique pour l'inférence de sous-domaine.

La prédiction par modèle fonctionne bien et de la même manière avec une distribution a priori structurée et une distribution a priori indépendante. C'était à prévoir à cause du léger effet de rétrécissement. La structure des cellules est peu déséquilibrée et les modèles de résultat et de sélection dépendent de l'ensemble des effets principaux et des termes d'interaction d'ordre supérieur. Toutefois, la distribution a priori structurée permet une inférence plus efficace que la distribution a priori indépendante avec une ET moindre. Cette amélioration est manifeste lorsque le plan de sondage est très déséquilibré, comme on peut le voir dans la simulation à la section 4.2.

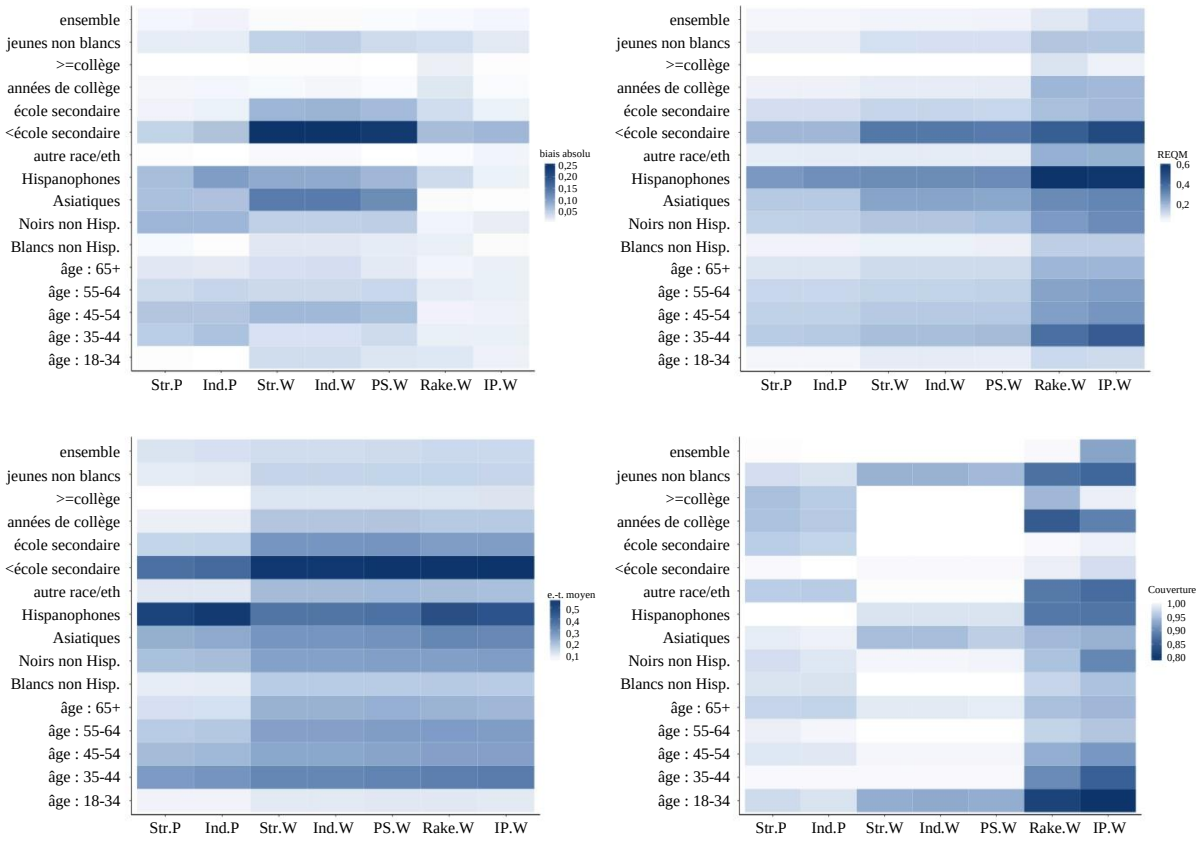

Figure 4.1 Comparaison des rendements de la prédiction et de la pondération pour la validité de l'inférence de population finie avec un plan de sondage légèrement déséquilibré. L'axe des y désigne les différents groupes pour l'estimation de moyenne. L'axe des x comprend deux méthodes de prédiction par modèle (Str-P et Ind-P), deux méthodes de pondération par modèle (Str-W, Ind-W) et trois méthodes de pondération classique (PS-W, Rake-W et IP-W). Str-P : prédiction par modèle avec distribution a priori structurée; Ind-P : prédiction par modèle avec distribution a priori indépendante; Str-W : pondération par modèle avec distribution a priori structurée; Ind-W : pondération par modèle avec distribution a priori indépendante; Rake-W : pondération avec méthode itérative du quotient; PS-W : pondération de poststratification; IP-W : pondération à probabilités inverses de sélection. Les graphiques indiquent que les prédictions par modèle sont d'un meilleur rendement que les pondérations avec les plus petites valeurs REQM et ET, des taux raisonnables de couverture et un biais comparable entre toutes les méthodes. L'inférence pondérée par modèle présente des valeurs REQM et ET moindres mais des taux de couverture plus raisonnables que dans la pondération classique.

Tableau 4.1
Comparaison des rendements de la prédiction et de la pondération pour la moyenne de sous-groupe des jeunes non blancs avec un plan de sondage légèrement déséquilibré

	Str-P	Ind-P	Str-W	Ind-W	PS-W	Rake-W	IP-W
Biais absolu	0,02	0,02	0,05	0,05	0,04	0,03	0,02
REQM	0,07	0,07	0,11	0,11	0,10	0,17	0,17
ET moyenne	0,08	0,08	0,13	0,13	0,13	0,13	0,13
Couverture	0,97	0,98	0,94	0,94	0,94	0,88	0,86

De plus, nous avons examiné neuf cas avec différents modèles de résultat d'enquête et de sélection d'échantillon et selon divers prédicteurs comme au tableau B.1 à l'annexe B. Les valeurs spécifiques des coefficients figurent dans les tableaux B.2 et B.3. La conclusion qui revient est que la prédiction et la pondération par modèle permettent une inférence plus efficace et précise que la pondération classique, plus particulièrement dans le cas de l'estimation de domaine.

4.2 Structure très déséquilibrée

La complexité des plans de sondage et des mécanismes de réponse tend à créer des structures de données très déséquilibrées où la plupart des cellules de poststratification sont éparses ou vides. La distribution a priori structurée que nous proposons produit un fort effet de régularisation qui stabilise la prédiction par modèle et améliore l'efficacité de l'estimation, surtout de l'estimation de domaine; son rendement se révèle supérieur à celui de la distribution a priori indépendante. L'inférence a posteriori portant sur les paramètres d'échelle peut éclairer la sélection des variables et ainsi améliorer l'interprétation du modèle. Si les effets principaux ne sont pas prédictifs, les interactions liées d'ordre supérieur ne le seront pas non plus. Il reste que l'inférence a posteriori avec des distributions a priori indépendantes déforme la structure hiérarchique entre les effets principaux et les interactions d'ordre supérieur et n'éclaire guère la sélection des variables. Les inférences par pondération classique sont d'une haute variabilité dans un scénario de cellules éparses.

Comme on le fait dans la LSW, nous puisons dans les données de l'ACS de 2011 pour la ville de New York huit variables qui influent sur l'inclusion dans l'échantillon : *age* (18–34, 35–44, 45–54, 55–64, 65+), *eth* (Blancs non hispanophones, Noirs non hispanophones, Asiatiques, Hispanophones, autres), *edu* (moins que l'école secondaire, école secondaire, années de collège, baccalauréat ou études supérieures), *sexe* (masculin, féminin), *pov* (ménage à un revenu ou mesure de la pauvreté, écart de pauvreté de moins de 50 %, de 50 à 100 %, de 100 à 200 %, de 200 à 300 % et de plus de 300 %), *cld* (0, 1, 2, 3+ jeunes enfants dans la famille), *eld* (0, 1, 2+ personnes âgées dans la famille) et *fam* (1, 2, 3, 4+ membres dans la famille). Le nombre de cellules uniques occupées par les éléments de cette classification est de 8 874 et le nombre de cellules de poststratification construites en tableau croisé intégral s'élève à 48 000.

Dans la simulation décrite aux tableaux B.4 et B.5, la probabilité de sélection dépend des effets principaux de toutes les variables et le résultat, des effets principaux de cinq variables. Les probabilités de sélection des cellules sont en grappe, c'est-à-dire que certaines cellules présenteront la même probabilité de sélection. L'échelle d'erreur est fixée à 1 dans le modèle de résultat. Les probabilités de sélection vont de 0 à 0,90 pour une moyenne de 0,12. Nous sélectionnons 6 374 unités. Bien que les tailles d'échantillon soient importantes, la simulation crée une structure très déséquilibrée. La majorité des cellules sont vides et 1 096 cellules sélectionnées sur 1 925 comportent une seule unité. À partir d'un modèle d'estimation avec cellules éparses, nous posons que le modèle en (3.4) des estimations des cellules comprend les effets principaux des huit variables, huit interactions binaires et deux interactions trinaires. Ces termes peuvent être des facteurs importants en pondération du point de vue de l'organisateur d'enquête. Notre proposition peut éclairer la sélection des variables et donc faciliter la réduction de dimension.

Lorsque seuls les effets principaux sont prédictifs, les valeurs de médiane a posteriori avec la distribution a priori structurée pour les échelles de *cld*, *eld* et *fam* sont petites (0,002, 0,003 et 0,000); pour les échelles de toutes les interactions d'ordre supérieur, elles sont proches de 0 (et d'une grandeur inférieure ou à peu près égale à 0,0001). La moyenne a posteriori de l'échelle d'erreur est de 0,99 avec une ET de 0,008, ce qui est proche de la valeur réelle 1. Cela s'accorde avec ce que laisse prévoir le plan de sondage en simulation. Avec la distribution a priori indépendante cependant, la structure hiérarchique entre les effets principaux et les termes d'interaction d'ordre supérieur est écartée. Les échantillons a posteriori des paramètres d'échelle des interactions d'ordre supérieur peuvent l'emporter sur ceux des effets principaux. On ne sait au juste quelle est leur puissance prédictive, et il est alors difficile de choisir des termes d'interaction. Les échantillons a posteriori des paramètres de variance avec les distributions a priori indépendantes tendent à une haute variabilité avec de lourdes queues. Ainsi, les variances des effets principaux de l'âge et du sexe présentent des valeurs échantillonnées extrêmement grandes (14 496 et 390 000) et les distributions sont asymétriques. Dans le cas des variables à peu de niveaux comme la variable du sexe, l'estimation de variance au niveau du groupe est sensible à la distribution a priori; la distribution a priori indépendante ne peut bien se régulariser dans ce cas. La distribution a priori structurée donne un meilleur résultat, car nous supposons que les distributions a priori ont un certain paramètre en commun et exploitent plus d'information aux fins de l'estimation, d'où la capacité de stabiliser l'estimation de variance. La distribution a priori structurée rend l'inférence plus stable que la distribution a priori indépendante et se trouve en plus à pouvoir faciliter la sélection des variables.

Ce que nous proposons comme distribution a priori structurée nous amènerait à exclure les effets principaux et les interactions d'ordre supérieur non prédictifs dans le modèle de régression des estimations des cellules, soit par un post traitement à 0 des échantillons a posteriori des échelles et des coefficients correspondants, soit par un nouvel ajustement du modèle modifié. Dans le plan de sondage en simulation, trois variables influent sur les probabilités de sélection, mais ne sont pas liées au résultat. L'inclusion de ces variables dans le modèle de régression accroîtra la variabilité de l'inférence. La structure des cellules de poststratification tient compte des huit variables, de sorte que l'hypothèse d'échantillonnage ignorable soit respectée. Une autre modification pourrait être l'exclusion des trois variables de la poststratification, ce qui pourrait rendre vulnérable l'hypothèse d'ignorabilité, mais en offrant des gains d'efficacité. C'est là un arbitrage entre efficacité et robustesse qui doit se faire en fonction de l'intérêt de fond de l'étude. Nous devons examiner plus avant la sélection de variables de résultat d'enquête dans le processus de pondération, ce que nous ferons à la section 6. Nous avons comparé cette inférence à l'inférence après exclusion des termes non prédictifs et nous avons obtenu des résultats semblables pour l'estimation de population finie et de domaine, car les estimations des paramètres sont proches de 0 pour les termes non prédictifs. Ici, nous présentons les résultats en gardant ces variables dans la construction des cellules de poststratification et dans le modèle de régression.

D'abord, nous comparons les poids produits en pondération par modèle et en pondération classique. Nous observons les échantillons a posteriori des poids produits et présentons la moyenne a posteriori comme base

de pondération par modèle. La pondération par modèle est d'une variabilité et d'une étendue inférieures à celles de la pondération classique, comme on peut le voir à la figure 4.2. La procédure itérative d'ajustement proportionnel ne converge pas après les 10 itérations par défaut et leur nombre doit augmenter. Nous examinons la distribution du résultat après prise en compte des poids et la comparons à la distribution de population et d'échantillon dans la partie droite de la figure 4.2. La distribution d'échantillon diffère de la distribution de population par une sous-estimation des valeurs de résultat. La distribution pondérée s'infléchit en direction de la population réelle. Les distributions de résultat après pondération se ressemblent entre la pondération par modèle et la pondération classique. La pondération par modèle crée une distribution lisse des résultats, ce qui est raisonnable puisque nous nous attendons à ce que cette pondération soit d'un même rendement que la pondération classique dans l'estimation ponctuelle, mais en améliorant l'efficacité par une moindre variabilité. L'effet de rétrécissement avec la distribution a priori structurée est marqué et varie de 0,86 à 1,00 pour une moyenne de 0,90. La structure très déséquilibrée des cellules nécessite un fort effet de lissage entre cellules. La pondération par modèle avec les distributions a priori structurée et indépendante est d'une même répartition avec les poids de poststratification, d'où l'omission des deux derniers ensembles de poids à la figure 4.2.

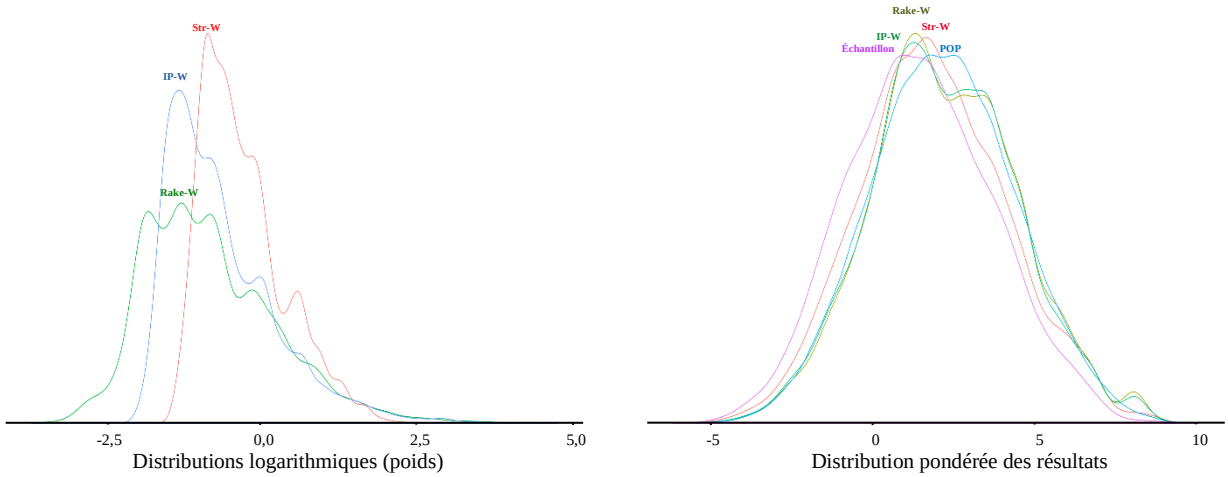

Figure 4.2 Comparaison des poids produits après une transformation logarithmique et des distributions pondérées de résultat avec un plan de sondage très déséquilibré. Str-W : pondération par modèle avec distribution a priori structurée; Rake-W : pondération avec méthode itérative du quotient; IP-W : pondération avec probabilités inverses de sélection. Échantillon : distribution d'échantillon des résultats; POP : distribution de population des résultats. La pondération par modèle est plus stable et donne une distribution plus lisse des résultats après pondération que la pondération avec méthode itérative du quotient et par probabilités inverses de sélection.

Nous examinons l'inférence de la moyenne d'ensemble et de la moyenne de domaine sur les niveaux marginaux et pour les jeunes adultes non blancs. Les conclusions sont les mêmes qu'à la section 4.1. La prédiction par modèle donne un meilleur résultat que l'inférence pondérée et le biais et l'erreur-type sont les plus bas. Cet avantage peut s'expliquer du fait que le modèle exploite l'information de population pour prédire les cellules vides en régularisation. L'inférence sur pondération par modèle présente des ET

inférieures à celles de l'inférence par pondération classique. Lors même que les probabilités de sélection dépendraient seulement des effets principaux, la méthode itérative du quotient laisse un petit biais, mais fonctionne mal avec une ET élevée.

Dans un plan de sondage très déséquilibré, l'inférence pondérée par modèle avec distribution a priori structurée est plus efficace qu'avec une distribution a priori indépendante ou dans une pondération de poststratification. Nous mettons en comparaison les ET des estimations de moyenne marginale des huit variables pour les trois méthodes de pondération et nous mettons en graphique les ratios relatifs dans la partie gauche de la figure 4.3. L'inférence pondérée par modèle présente des ET inférieures à celles de l'estimation par pondération de poststratification. C'est la pondération avec distribution a priori structurée qui comporte les ET les plus basses. Comme les tailles d'échantillon et de domaine sont importantes et que le modèle de génération des données est à cellules éparses, l'inférence sur pondération par modèle représente une légère amélioration sans plus sur l'inférence par pondération de poststratification en raison d'un léger effet de lissage.

La prédiction et l'inférence par modèle sont plus efficaces avec une distribution a priori structurée qu'avec une distribution a priori indépendante. Les ET sont moins élevées avec la première qu'avec la seconde dans la partie droite de la figure 4.3. Pour démontrer le gain d'efficacité, nous considérons les erreurs-types des estimations de population des cellules. En général, l'inférence bayésienne est d'une moindre variabilité avec une distribution structurée qu'avec une distribution indépendante, plus particulièrement dans les scénarios à cellules éparses.

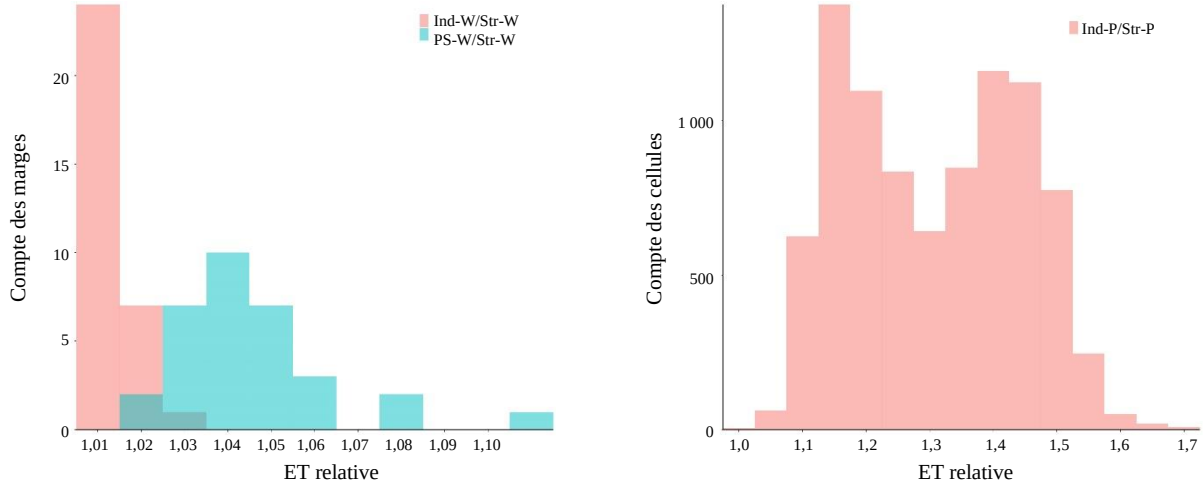

Figure 4.3 Comparaison d'efficacité dans les rendements de prédiction et de pondération pour l'inférence de population finie et de domaine avec un plan de sondage très déséquilibré. Le graphique de gauche porte sur l'estimation de la moyenne sur les marges définies par les huit variables. Le graphique de droite présente l'estimation de population de la moyenne des cellules. La prédiction et la pondération par modèle donnent des erreurs-types moindres avec la distribution a priori structurée qu'avec la distribution a priori indépendante. La pondération par modèle se caractérise par des erreurs-types inférieures à celles de la pondération de poststratification.

Nous posons différents modèles de résultat et de sélection avec diverses covariables dans les scénarios résumés au tableau B.4 et parvenons aux mêmes conclusions en matière d'évaluation.

5 Application à l'étude longitudinale du bien-être

Dans le contexte décrit à la section 2, nous appliquons l'inférence par prédiction et pondération à la *Longitudinal Study of Wellbeing* (LSW) de la ville de New York. Nous apparions la LSW et la population d'âge adulte par l'ACS. Nous désirons procéder à une inférence de population finie et de domaine et produire des poids se prêtant à une utilisation en analyse générale. Le résultat d'intérêt est la cote autodéclarée de satisfaction dans la vie sur une échelle de 1 à 10. Nous modélisons le résultat en distribution normale, ce qui n'est pas tout à fait juste si nous considérons que les réponses sont discontinues, mais ce qui devrait convenir en pratique pour l'estimation des moyennes. Nous prenons d'abord les huit mêmes variables pour construire les cellules de poststratification et employons le même modèle d'estimation qu'à la section 4.2 avec une distribution a priori structurée. L'inférence a posteriori indique que les variables *sex, cldx, eldx* et *psx* ne sont pas prédictives, pas plus que les interactions liées d'ordre supérieur. Les estimations d'échelle de ces termes ont des valeurs de médiane a posteriori proches de 0 et plusieurs valeurs importantes sont à longue queue. Les échantillons a posteriori d'échelle de plusieurs interactions d'ordre supérieur parmi les quatre variables restantes se concentrent autour de 0, ce qui montre que les valeurs ne sont pas prédictives. Un autre facteur de complexité est que, pour les cellules de l'échantillon de la LSW, les cellules correspondantes de population ne sont pas disponibles dans les données de l'ACS, ce qui pourrait s'expliquer par le fait que la base de sondage n'est pas l'enquête ACS. L'information de population est inconnue pour les cellules en question et des hypothèses invérifiables doivent être posées. L'ajustement du modèle s'améliore après la sélection des variables lorsque nous vérifions les erreurs de prédiction dans les estimations des cellules.

Nous employons, par conséquent, quatre variables après sélection, soit *age, eth, edu* et *pov*, ce qui permet de construire 500 cellules de poststratification. Les 2 002 unités de la LSW se répartissent entre 359 cellules. La cellule la plus grande de l'échantillon compte 86 unités et 92 cellules comportent une seule unité. Les covariables du modèle en (3.4) pour les estimations des cellules sont les effets principaux des quatre variables, cinq interactions binaires (*age * eth*, *age * edu*, *eth * edu*, *age * inc* et *eth * inc*) et deux interactions trinaires (*age * eth * edu* et *age * eth * inc*). Nous effectuons une inférence pleinement bayésienne avec les distributions a priori structurées. Nous désirons estimer la cote moyenne de satisfaction dans la vie pour l'ensemble et certains sous-groupes de la population adulte de la ville de New York et nous construisons une pondération à des fins d'analyse générale à l'aide de la LSW.

La médiane a posteriori de l'échelle des unités à l'intérieur des cellules σ_y est de 1,93 pour un intervalle de crédibilité à 95 % [1,87, 1,99]. La médiane a posteriori de l'échelle de groupe σ_θ est de 0,79 pour un intervalle de crédibilité à 95 % [0,65, 1,02]. Cela donne des effets de rétrécissement modérément importants qui varient de 0,11 à 0,90 pour une moyenne de 0,30 entre cellules. L'effet modéré de rétrécissement est

logique en cas d'inclusion des quatre variables et des interactions jusqu'aux interactions trinaires. Nous présentons les valeurs de moyenne a posteriori des poids par modèle dans la partie gauche de la figure 5.1. Nous pouvons produire la pondération avec la méthode itérative du quotient après ajustement en fonction des distributions marginales des quatre variables et la pondération de poststratification avec les données de l'ACS. Nous obtenons l'information de population après application des poids des personnes de l'ACS.

À comparer à la pondération classique, la pondération par modèle est d'une moindre variabilité avec 0,32 comme écart-type et 3,87 comme rapport maximum-minimum; ces valeurs sont bien inférieures à celles de la pondération avec méthode itérative du quotient et de la pondération de poststratification, comme le montre le tableau 5.1. La partie droite de la figure 5.1 présente la distribution après pondération des cotes de satisfaction dans la vie. Les distributions pondérée basée sur un modèle et pondérée classique se ressemblent comme on pouvait s'y attendre et s'accordent avec les résultats à la section 4.2. Le processus de pondération tient compte de la distribution d'échantillon en majorant la pondération des cotes hautes et en minorant celle des cotes basses. La LSW suréchantillonne les résidents pauvres qui sont généralement insatisfaits dans la vie et l'ajustement de pondération comble l'écart.

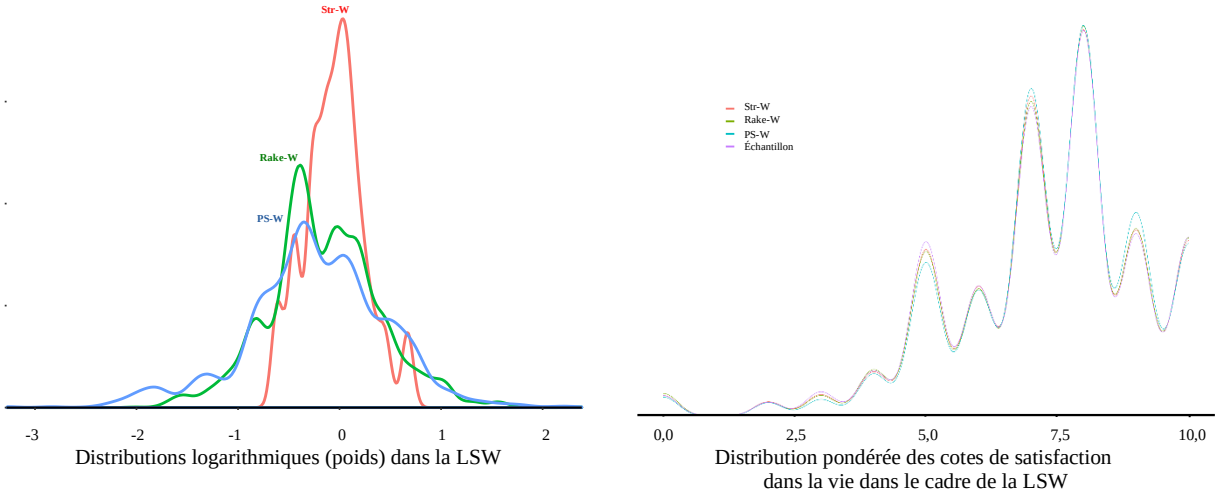

Figure 5.1 Comparaison des poids produits après transformation logarithmique et des distributions pondérées des cotes de satisfaction dans la vie dans le cadre de la LSW. Str-W : pondération par modèle avec distribution a priori structurée; Rake-W : pondération avec méthode itérative du quotient; IP-W : pondération à probabilités inverses de sélection; échantillon : distribution d'échantillon des résultats. Les distributions pondérées se ressemblent entre pondération par modèle et pondération classique, mais la première de ces pondérations est plus stable que la seconde.

Le tableau 5.1 et la figure 5.2 présentent l'inférence de population finie et de domaine. La cote moyenne de satisfaction dans la vie des adultes dans la ville de New York s'établit à 7,24 et l'erreur-type est de 0,05 selon les prédictions du modèle structurel. L'estimation est la même pour la pondération par modèle que

pour la pondération avec méthode itérative du quotient, mais elle est moindre que dans le cas de la pondération de poststratification, mais sans que la différence soit significative. Ainsi, le modèle structurel prédit la cote moyenne de satisfaction dans la vie chez les Blancs d'âge moyen ayant fait des études collégiales et dont le revenu est plus de trois fois le seuil de pauvreté avec une valeur de 7,40 pour une erreur-type de 0,10. Ce sont des valeurs supérieures à celles des inférences pondérées. Cependant, les cotes prédites pour les personnes âgées au revenu relativement bas (7,37, ET de 0,15) et pour les New-Yorkais noirs à faible revenu (7,01, ET de 0,18) sont inférieures à celles des inférences pondérées. La différence pourrait s'expliquer par la non-représentativité de la LSW et par les interactions profondes incluses par le modèle. Le sous-groupe des Blancs d'âge moyen ayant fréquenté le collège peut être sous-observé dans la LSW (cellules de poststratification vides). Il y aurait en revanche surobservation des Noirs pauvres et âgés. La pondération des échantillons observés ne permet ni inférence ni extrapolation sur les gens absents de l'enquête. Les différences ne sont pas significatives, mais les inférences conditionnées par les échantillons prélevés ne sauraient restituer la vérité, surtout dans le cas des estimations de cellules vides. La figure 5.2 indique que la prédiction par modèle livre, par rapport aux inférences pondérées, des cotes supérieures pour les jeunes adultes hautement instruits et hispanophones de la ville de New York, mais inférieures pour les citadins correspondants dont l'écart de pauvreté est de moins de 50 %.

Tableau 5.1
Comparaison du rendement de la prédiction et de la pondération dans l'estimation de diverses moyennes de domaine pour la satisfaction dans la vie dans le cadre de la LSW. Str-P : prédiction par modèle avec distribution a priori structurée; Str-W : pondération par modèle avec distribution a priori structurée; Rake-W : pondération avec méthode itérative du quotient; PS-W : pondération de poststratification

	Str-P	Str-W	Rake-W	PS-W
Écart-type des poids/moyenne des poids		0,32	0,66	0,80
Rapport maximum/minimum des poids		3,87	81,28	274,65
Moyenne générale pour les adultes de la ville de New York (n = 2 002)				
Estimation	7,24	7,23	7,24	7,30
ET	0,05	0,05	0,05	0,06
Moyenne pour les Blancs d'âge moyen ayant fréquenté le collège et dont l'écart de pauvreté est de > 300 % (n = 222)				
Estimation	7,40	7,34	7,34	7,34
ET	0,10	0,11	0,11	0,11
Moyenne pour les personnes âgées dont l'écart de pauvreté est de < 200 % (n = 154)				
Estimation	7,37	7,52	7,49	7,53
ET	0,15	0,18	0,19	0,22
Moyenne pour les Noirs dont l'écart de pauvreté est de < 50 % (n = 57)				
Estimation	7,01	7,16	7,30	7,16
ET	0,18	0,26	0,28	0,29

Les ET sont semblables pour l'estimation de moyenne générale entre les prédictions et diverses inférences pondérées en raison de la grande taille de l'échantillon. Dans le cas de l'estimation de domaine, la prédiction et la pondération par modèle sont plus efficaces que la pondération avec méthode itérative du quotient et la pondération de poststratification; la prédiction par modèle présente l'erreur-type la plus basse. L'estimation de moyenne de domaine pour les cotes de satisfaction dans la vie sur les niveaux marginaux des quatre variables (voir la figure 5.2) témoigne encore plus des gains d'efficacité qu'offrent la prédiction et

la pondération par modèle. Celles-ci améliorent particulièrement l'estimation sur petits domaines et accroissent l'efficacité.

Les praticiens des enquêtes comparent souvent la distribution pondérée des caractéristiques sociodémographiques à la distribution de population pour vérifier l'effet de la pondération. Les diagnostics de pondération doivent être mieux étudiés et gérés, mais nous y faisons appel et mettons en comparaison la pondération par modèle et la pondération classique. Nous calculons les distances euclidiennes entre les distributions pondérées et la distribution de population pour les effets principaux et les interactions d'ordre supérieur en ce qui concerne les quatre variables de la LSW (voir le tableau B.6 à l'annexe B). Les distributions pondérées sont généralement proches des distributions réelles. La méthode itérative du quotient permet un ajustement en fonction des distributions marginales des variables de calage, mais elle déforme les codistributions là où la structure de dépendance est uniquement déterminée par l'échantillon hors de tout calage. La pondération de poststratification tient compte de la codistribution, mais les cellules vides dans l'échantillon empêchent un parfait appariement. La structure déséquilibrée des cellules rend l'inférence instable. La pondération par modèle lisse les poids de poststratification et réussit mieux que la pondération avec méthode itérative du quotient à apparier les distributions des termes d'interaction trinaire et quatrinaire. Souvent, les praticiens s'appuient sur les distributions marginales pour évaluer le rendement en pondération, favorisant ainsi la méthode itérative du quotient. Il reste que cette méthode donne lieu à des inférences hautement variables et potentiellement biaisées (voir la section 4), même là où l'ajustement qu'elle permet est correct. Un pas en avant dans ce domaine sera d'intégrer des contraintes en modifiant la pondération par modèle dans un désir d'appariement des distributions marginales.

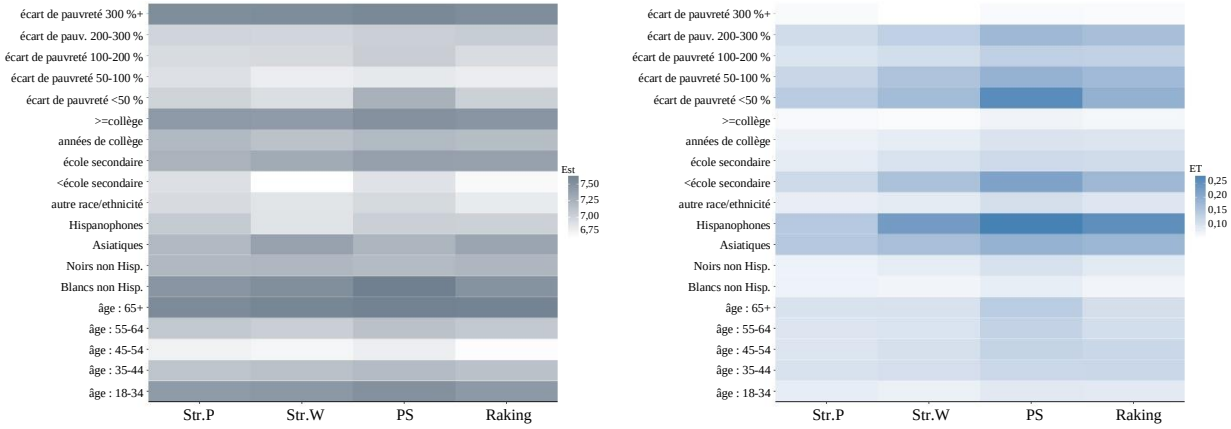

Figure 5.2 Comparaison du rendement de la prédiction et de la pondération dans l'estimation des cotes de satisfaction dans la vie sur les distributions marginales de quatre variables de la LSW. Str-P : prédiction par modèle avec distribution a priori structurée; Str-W : pondération par modèle avec distribution a priori structurée; Rake-W : pondération avec méthode itérative du quotient; PS-W : pondération de poststratification. La prédiction et la pondération par modèle donnent des estimations différentes pour plusieurs sous-ensembles et sont généralement plus efficaces que la pondération classique.

6 Analyse

Nous combinons la prédiction bayésienne et la pondération en une approche unifiée pour l'inférence d'enquête. La régression multiniveau avec distributions a priori structurées et la poststratification appliquée à l'inférence de population permettent une estimation efficace si elles tiennent compte des caractéristiques du plan de sondage. Le calcul se fait au moyen de Stan et est rendu disponible dans le paquet `rstanarm` en R pour le public. L'évolution des logiciels favorise les approches par modèle dans la recherche et l'exécution d'enquête. Nous construisons une pondération par modèle stable et calée pour résoudre les problèmes de la pondération classique. Dans le présent article, nous développons le cadre de prédiction et de pondération par modèle et apportons une première contribution à l'évaluation des propriétés statistiques de la pondération par modèle. Nous mettons en comparaison cette pondération et la pondération classique. Les poids par modèle font l'objet d'un lissage sur les cellules de poststratification et améliorent l'estimation sur petits domaines.

La distribution a priori structurée exploite la structure hiérarchique entre les effets principaux et les termes d'interaction d'ordre supérieur pour introduire des contraintes multiplicatives sur les paramètres d'échelle correspondants. Elle vient informer la sélection des variables. Il est possible d'améliorer le modèle après le post traitement des inférences a posteriori. Le modèle structurel bayésien rend l'inférence plus stable qu'avec les distributions a priori indépendantes. L'hypothèse posée d'une hiérarchie pourrait ne pas être valide dans certains cas spéciaux comme pour le problème « ou exclusif » où deux variables n'ont pas d'effets principaux, mais une parfaite interaction. Nous n'avons néanmoins aucune preuve solide dans les études d'application à opposer à la notion de vraisemblance d'un traitement hiérarchique. Qui plus est, un cadre unifié de prédiction et de pondération est bien armé pour traiter les cas de plans de sondage complexes et de mégadonnées dans les enquêtes, qu'il s'agisse des données de flux ou des combinaisons d'enquêtes.

Le cadre général de la régression multiniveau et de la postratification (RMP) se prête à l'adoption de stratégies souples de modélisation. Dans cet article, nous illustrons la chose par un modèle de régression avec l'ensemble des variables d'intérêt et les interactions d'ordre supérieur et intégrons des distributions a priori structurées à des fins de régularisation. D'autres méthodes (modèles non paramétriques, outils d'apprentissage automatique, etc.) peuvent s'appliquer dans ce même cadre et s'avèrent robustes à l'égard des erreurs de spécification de modèle. Si et coll. (2015) se servent de modèles de régression à processus gaussien pour emprunter de l'information à la structure de cellules de poststratification selon les distances entre les poids à probabilités inverses d'inclusion. D'autres possibilités consisteraient à appliquer de telles approches souples au lissage des poids et au calcul d'une pondération par modèle.

Les vastes possibilités d'application qui s'offrent ne vont pas sans un certain nombre de défis à étudier. La pondération par modèle est dépendante des résultats, d'où un gain d'efficacité mais une perte possible de robustesse. Les organisateurs d'enquêtes préfèrent un ensemble de poids pouvant servir à une analyse générale sans une sensibilité à la sélection des résultats. Nous pouvons comparer différentes pondérations construites selon plusieurs résultats importants et procéder à une analyse de sensibilité. Là où la pondération par modèle donne des conclusions d'inférence différentes, nous recommandons de choisir la pondération qui

donne les résultats les plus raisonnables en appliquant la science, tout comme de s'aligner sur l'inférence de population.

Les distributions marginales pondérées des variables de calage diffèrent quelque peu des inférences de population (voir la section 5), lesquelles ne respectent pas la norme habituelle de diagnostic de pondération des organisateurs d'enquêtes. La pondération par modèle tend à apparier la distribution conjointe à celle dans la population, mais le lissage des poids peut introduire un biais. Il est possible d'inclure des contraintes optionnelles au modèle pour rencontrer les distributions marginales.

Un autre défi d'ordre pratique est que la distribution de population des variables de calage peut être inconnue. Dans ce cas, les tailles N_j des cellules de poststratification de population ne seront pas connues. Il nous faut un modèle supplémentaire pour l'estimation de cette information dans l'échantillon. Il nous faut l'intégrer au cadre RMP pour propager toutes les sources d'incertitude en complément, comme dans le cadre de Si et Zhou (2020), par l'incorporation dans le modèle des distributions marginales connues. Les inférences par prédiction et pondération par modèle doivent encore être développées de manière à pouvoir traiter les résultats discrets, l'inférence sur coefficients de régression et les plans de sondage non probabilistes ou informatifs (Kim et Skinner, 2013). Il serait bon de faire le lien entre ces idées sur l'inférence d'enquête et les études biostatistiques et économétriques sur les scores inverses de propension et la pondération à double robustesse (Kang et Schafer, 2007).

Remerciements

Nous remercions la *National Science Foundation*, les *National Institutes of Health*, l'*Office of Naval Research*, l'*Institute of Education Sciences* et la *Sloan Foundation* pour l'aide apportée sous forme de subventions.

Annexes

A. Exemple de code

Nous présentons ici le code de l'application décrite dans les données. Nous avons formulé une fonction `model_based_cell_weights` pour calculer la pondération par modèle dans le cas d'un modèle `rstanarm` ajusté.

```
model_based_cell_weights <- function(object, cell_table) {
 stopifnot(
    is.data.frame(cell_table),
    colnames(cell_table) == c("N", "n")
 )
 draws <- as.matrix(object)
 Sigma <- draws[, grep("^Sigma\\[", colnames(draws)), drop = FALSE]
```

```
 sigma_theta_sq <- rowSums(Sigma)
 sigma_y_sq <- draws[, "sigma"]^2
 Ns <- cell_table[["N"]] # population cell counts
 ns <- cell_table[["n"]] # sample cell counts
 J <- nrow(cell_table)
 N <- sum(Ns)
 n <- sum(ns)
 # implementing equation 7 in the paper (although i did some algebra first to
 # simplify the expression a bit)
 Nsy2 <- N * sigma_y_sq
 ww <- matrix(NA, nrow = nrow(draws), ncol = J)
 for (j in 1:J) {
    ww[, j] <-
       (Nsy2 + n * Ns[j] * sigma_theta_sq) / (Nsy2 + N * ns[j] * sigma_theta_sq)
 }
 return(ww)
}
# prepare population data: acs_ad has age, eth, edu and inc
acs_ad %>%
 mutate(
    cell_id = paste0(age, eth, edu, inc)
 ) -> acs_ad
acs_design <- svydesign(id = ~1, weights = ~perwt, data = acs_ad)
agg_pop <-
 svytable( ~ age + eth + edu + inc, acs_design) %>%
 as.data.frame() %>%
 rename(N = Freq) %>%
 mutate(
    cell_id = paste0(age, eth, edu, inc)
 ) %>%
 filter(cell_id %in% acs_ad$cell_id)
# prepare data to pass to rstanarm
# SURVEYdata has 4 variables used for weighting: age, eth, edu and inc; and outcome Y
dat_rstanarm <-
 SURVEYdata %>%
 mutate(
    cell_id = paste0(age, eth, edu, inc)
 )%>%
 group_by(age, eth, edu, inc) %>%
 summarise(
    sd_cell = sd(Y),
    n = n(),
    Y = mean(Y),
    cell_id = first(cell_id)
```

```
 ) %>%
 mutate(sd_cell = if_else(is.na(sd_cell), 0, sd_cell)) %>%
 left_join(agg_pop[, c("cell_id", "N")], by = "cell_id")
# Stan fitting under structured prior in rstanarm
fit <-
 stan_glmer(
    formula =
       Y ~ 1 + (1 | age) + (1 | eth) + (1 | edu) + (1 | inc) +
       (1 | age:eth) + (1 | age:edu) + (1 | age:inc) +
       (1 | eth:edu) + (1 | eth:inc) +
       (1 | age:eth:edu) + (1 | age:eth:inc),
    data = dat_rstanarm, iter = 1000, chains = 4, cores = 4,
    prior_covariance =
       rstanarm::mrp_structured(
          cell_size = dat_rstanarm$n,
          cell_sd = dat_rstanarm$sd_cell,
          group_level_scale = 1,
          group_level_df = 1
       ),
    seed = 123,
    prior_aux = cauchy(0, 5),
    prior_intercept = normal(0, 100, autoscale = FALSE),
    adapt_delta = 0.99
 )
# model-based weighting
cell_table <- fit$data[,c("N","n")]
weights <- model_based_cell_weights(fit, cell_table)
weights <- data.frame(w_unit = colMeans(weights),
                      cell_id = fit$data[["cell_id"]],
                      Y = fit$data[["Y"]],
                      n = fit$data[["n"]]) %>%
         mutate(
             w = w_unit / sum(n / sum(n) * w_unit), # model-based weights
             Y_w = Y * w
         )
with(weights, sum(n * Y_w / sum(n)))# mean estimate
```

B. Plans des simulations

Nous présentons ici les plans des simulations, les valeurs des coefficients et une comparaison portant sur les distributions pondérées des caractéristiques sociodémographiques en complément d'information aux sections 4 et 5.

Tableau B.1
Covariables dans les modèles de résultat et de sélection pour un plan de sondage légèrement déséquilibré

	Cas 1		Cas 2		Cas 3		Cas 4		Cas 5		Cas 6		Cas 7	
	O	S	O	S	O	S	O	S	O	S	O	S	O	S
age	√	√	√	√	√	√	√	√	√	√	√	√	√	√
eth	√	√	√	√	√	√	√	√	√			√		√
edu	√	√	√	√	√		√	√	√		√	√	√	
age*eth	√			√	√	√				√				√
age*edu	√			√	√						√		√	
eth*edu	√			√	√									
age*eth*edu	√			√	√									

Tableau B.2
Coefficients hypothétiques de régression dans le modèle de résultat pour la simulation avec un plan de sondage légèrement déséquilibré

	Ensemble	Effets principaux	Deux variables
age	(0,5; 1,375; 2,25; 3,125; 4)	(0,5; 1,375; 2,25; 3,125; 4)	(0,5; 1,375; 2,25; 3,125; 4)
eth	(-2; -1; 0; 1; 2)	(2; -1; 0; 1; 2)	$\vec{0}$
edu	(3; 2; 1; 0)	(3; 2; 1; 0)	(3; 2; 1; 0)
age*eth	(4; 2; 1; 1; 3; 3; 2; 1; 1; 1; 2; 3; 2; 2; 1; 4; 4; 3; 2; 3; 2; 4; 1; 4; 1)	$\vec{0}$	$\vec{0}$
age*edu	(-2; -1; 2; 2; 1; -2; 2; 1; 0; -2; 1; -2; -1; 2; 1; -1; -1; 2; 0; 2)	$\vec{0}$	(2; 0; -2; -2; 1; 1; -1; -2; -2; -1; -1; 1; 0; -1; -1; 2; 2; 1; -1; 0)
eth*edu	(1; -2; 0; -3; -1; 0; -1; -2; 0; -1; -3; -3; 0; -1; -1; 0; 0; -1; 0; -1)	$\vec{0}$	$\vec{0}$
age*eth*edu	(-1; -0,5; 0,5; -1; -1; -0,5; -1; 0; -1; 0; -1; 0; 1; 1; 0,5; 1; 1; -1; -1; 0; -1; -0,5; -0,5; -1; 1; -1; -0,5; -1; 1; 0; 0,5; 0,5; 1; 0,5; 1; 1; 1; 0,5; 1; 0; 0; -0,5; 0; 1; -1; -1; 0; -1; -1; -1; -0,5; -0,5; 0 1; -1; 0; 0; -0,5; 1; -0,5; 0,5; -1; 1; 0; 1; 0; -1; 0; -0,5; 1; -0,5; -1; -0,5; 0; 0,5; -0,5; 1; 0,5; -0,5; 0,5; 0; 1; 0; 1; 0,5; 0,5; 0,5; 0; 0; -0,5; 1; -1; 0; 1; 1; 1; 1; -0,5; -1; -1)	$\vec{0}$	$\vec{0}$

Tableau B.3
Coefficients hypothétiques de régression dans le modèle de sélection pour la simulation avec un plan de sondage légèrement déséquilibré

	Ensemble	Effets principaux	Deux variables
Valeur à l'origine	-2	-2	-2
age	(-2; -1,75; -1,5; -1,25; -1)	(0; 0,5; 1; 1,5; 2)	(-2; -1,5; -1; -0,5; 0)
eth	(-1; -0,25; 0,5; 1,25; 2)	(-2; -1,5; -1; -0,5; 0)	(-1; -0,5; 0; 0,5; 1)
edu	(0; 0,67; 1,33; 2)	(0; 1; 2; 3)	$\vec{0}$
age × eth	(1; 1; -1; 1; -1; 1; -1; 0; 0; -1; 0; 0; -1; 1; 0; 0; -1; 1; 1; -1; -1; 0; 1; -1; 1)	$\vec{0}$	(-1; 1; 1; 1; -1; -1; -1; 0; -1; -1; -1; -1; 1; -1; -1; 0; 1; 1; -1; 1; -1; -1; 1; 0; 0)
age × edu	(0; 1; -1; -1; 0; 1; 1; 0; 1; 0; 1; -1; -1; 1; 1; -1; 0; -1; 1; 1)	$\vec{0}$	$\vec{0}$
eth × edu	(-1; -1; 0; -1; -1; 1; 1; 1; 1; 0; -1; 0; -1; 0; -1; 1; 0; -1; -1; -1)	$\vec{0}$	$\vec{0}$
age × eth × edu	(0,8; -0,4; 0,6; -0,2; 0,8; 0,2; 0,4; 0,8; 0,4; -0,6; -0,8; -0,4; -0,8; -0,4; 0,4; -1; 0,6; -0,8; -0,6; 0,6; -0,2; 0,2; 0,6; -0,6; 0; 0; -1; -0,2; 0,6; 0,8; -0,4; 0,2; -0,8; 0,4; 0,6; -0,6; 0,8; 0; 0,2; -1; 1; 0,4; 0; 0,8; -0,2; 0; 0; 0,6; -0,8; -0,8; -0,2; 0,4; -1; -0,8; 1; -0,2; 0; 0,8; 0,6; 0,8; -0,2; -0,2; -0,8; 1; 0,8; 0,8; -0,4; -0,8; 0,4; -0,4; 1; -0,6; -1; -0,6; -0,2; 1; 1; -0,2; 1; 0,6; 0,4; 0,8; 0,2; -0,2; -0,6; 0; 0,8; -0,4; 0,4; 0,4; 0,6; -1; -0,8; -0,8; 1; 1; 0,4; 0,6; 0,4; 0,8)	$\vec{0}$	$\vec{0}$

Tableau B.4
Covariables dans les modèles de résultat (O) et de sélection (S) pour un plan de sondage très déséquilibré

	Cas 1		Cas 2		Cas 3		Cas 4	
	O	S	O	S	O	S	O	S
age	√	√	√	√	√	√	√	√
eth	√	√	√	√	√	√	√	√
edu	√	√	√	√	√	√	√	√
sex	√	√	√	√	√	√	√	√
pov	√	√	√	√	√	√	√	√
cld		√		√		√	√	√
eld	√	√		√	√	√	√	√
fam	√	√		√	√	√	√	√
age*eth	√	√			√			√
age*edu	√	√			√			√
eth*edu	√	√			√			√
eth*pov	√	√			√			√
age*pov	√	√			√			√
pov*fam	√	√			√			√
pov*eld	√	√			√			√
pov*cld		√						√
age*eth*edu	√	√			√			√
age*eth*pov	√	√			√			√

Tableau B.5
Coefficients hypothétiques de régression dans les modèles de résultat (O) et de sélection (S) pour un plan de sondage très déséquilibré

	O	S
age	(2; 0; -2; -2; 1)	(0; 0,75; 1,5; 2,25; 3)
eth	(1; -1; -2; -2; -1)	(-1; -0,5; 0; 0,5; 1)
edu	(-1; 1; 0; -1)	(0; 0,67; 1,33; 2)
sex	(-1; 2)	(-1; 0)
pov	(2; 1; -1; 0; -1)	(0; 1; 2; 3; 4)
cld	$\vec{0}$	(-1; -0,33; 0,33; 1)
eld	$\vec{0}$	(-2; -1; 0)
fam	$\vec{0}$	(-1; -0,67; -0,33; 0)

Tableau B.6
Distances euclidiennes entre les distributions pondérées et la distribution de population. Str-W : pondération par modèle avec distribution a priori structurée; Rake-W : pondération avec méthode itérative du quotient; PS-W : pondération de poststratification

	Str-W	PS-W	Rake-W
age	0,04	0,02	0,00
eth	0,08	0,06	0,00
edu	0,08	0,03	0,00
inc	0,02	0,02	0,00
age * eth	0,05	0,03	0,05
age * edu	0,05	0,02	0,05
age * inc	0,03	0,01	0,03
eth * edu	0,06	0,04	0,05
eth * inc	0,04	0,04	0,03
edu * inc	0,06	0,03	0,04
age * eth * edu	0,03	0,02	0,05
age * eth * inc	0,03	0,02	0,04
age * edu * inc	0,03	0,01	0,04
eth * edu * inc	0,04	0,02	0,04
age * eth * edu * inc	0,02	0,01	0,04

Bibliographie

ACS Weighting Method (2014). *American Community Survey Design and Methodology*, Chapitre 11 : Weighting and Estimation. United States Census Bureau.

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, 95, 539-553.

Breidt, F.J. (2008). Endogenous post-stratification in surveys: Classifying with a sample-fitted model. *Annals of Statistics*, 36, 403-427.

Breidt, F., et Opsomer, J. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32, 190-205.

Carvalho, C.M., Polson, N.G. et Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97, 465-480.

Chambers, R.L., Dorfman, A.H. et Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88, 260-269.

Chen, Q., Elliott, M.R., Haziza, D., Yang, Y., Ghosh, M., Little, R., Sedransk, J. et Thompson, M. (2017). Approaches to improving survey-weighted estimates. *Statistical Science,* 32(2), 227-248.

Dahlke, M., Breidt, F., Opsomer, J. et Keilegom, I.V. (2013). Nonparametric endogenous poststratification in surveys. *Statistica Sinica*, 23, 189-211.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Deville, J.-C., Särndal, C.-E. et Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88(423), 1013-1020.

Elliott, M.R. (2007). Bayesian weight trimming for generalized linear regression models. *Journal of Official Statistics*, 33(1), 23-34.

Elliott, M.R., et Little, R.J. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics,* 16(3), 191-209.

Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366a), 269-277.

Firth, D., et Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society, Series B*, 60, 3-21.

Fuller, W. (2009). *Sampling Statistics*. Hoboken: John Wiley & Sons, Inc.

Gelman, A. (2005). Analysis of variance: Why it is more important than ever (avec discusion). *Annals of Statistics*, 33(1), 1-53.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 3, 515-533.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2), 153-164.

Gelman, A., et Carlin, J.B. (2001). Poststratification and weighting adjustments. Dans *Survey Nonresponse*, (Éds., R. Groves, D. Dillman, J. Eltinge et R. Little).

Gelman, A., et Little, T.C. (1997). Stratifcation 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.

Gelman, A., et Little, T.C. (1998). Improving on probability weighting for household size. *Public Opinion Quarterly*, 62, 398-404.

Ghitza, Y., et Gelman, A. (2013). Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, 57(3), 762-776.

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

Goodrich, B., et Gabry, J.S. (2017). `rstanarm`: Bayesian applied regression modeling via Stan. https://cran.r-project.org/web/packages/rstanarm/.

Groves, R., et Couper, M. (1995). Theoretical motivation for post-survey nonresponse adjustment in household surveys. *Journal of Offcial Statistics*, 11, 93-106.

Hájek, J. (1971). Comment on "An essay on the logical foundations of survey sampling" by D. Basu. Dans *The Foundations of Survey Sampling*, (Éds., V.P. Godambe et D.A. Sprott), 236. Holt, Rinehart and Winston.

Hastie, T., Tibshirani, R. et Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition. Springer.

Henry, K., et Valliant, R. (2012). Comparing alternative weight adjustment methods. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Hoffman, M.D., et Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15, 1351-1381.

Holt, D., et Smith, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical Society, Series A*, 142(1), 33-46.

Kang, J.D.Y., et Schafer, J.L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4), 523-539.

Kim, J.K., et Skinner, C.J. (2013). Weighting in survey analysis under informative sampling. *Biometrika*, 100, 385-398.

Kott, P. (2009). Calibration weighting: Combining probability samples and linear prediction models. Dans *Handbook of Statistics, Sample Surveys: Design, Methods and Application*, (Éds., D. Pfeffermann et C.R. Rao), Volume 29B. Elsevier.

Little, R. (1983). Comment on "An evaluation of model-dependent and probability-sampling inferences in sample surveys", par M.H. Hansen, W.G. Madow et B.J. Tepping. *Journal of the American Statistical Association*, 78, 797-799.

Little, R. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.

Little, R. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

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

Little, R. (2011). Calibrated Bayes, for statistics in general, and missing data in particular. *Statistical Science*, 26, 162-174.

Little, R., et Wu, M. (1991). Models for contingency tables with known margins when target and sampled populations differ. *Journal of the American Statistical Association*, 86, 87-95.

McConville, K.S., et Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46(2), 389-413.

Park, D.K., Gelman, A. et Bafumi, J. (2005). State-level opinions from national surveys: Poststratification using multilevel logistic regression. Dans *Public Opinion in State Politics*, (Éd., J.E. Cohen), Standord University Press.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Revue Internationale de Statistique*, 61(2), 317-337.

Piironen, J., et Vehtari, A. (2016). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. https://arxiv.org/abs/1610.05559.

Potter, F.A. (1988). Survey of procedures to control extreme sample weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 453-458.

Potter, F.A. (1990). A study of procedures to identify and trim extreme sampling weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 225-230.

Rao, J.N.K. (1966a). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā, Series A*, 28(1), 47-60.

Rao, J.N.K. (1966b). On the relative effciency of some estimators in PPS sampling for multiple characteristics. *Sankhyā, Series A*, 28(1), 61-70.

Rao, J.N.K. (2011). Impact of frequentist and bayesian methods on survey sampling practice: A selective appraisal. *Statistical Science*, 26(2), 240-256.

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

Rasmussen, C.E., et Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.

Reilly, C., Gelman, A. et Katz, J. (2001). Poststratication without population level information on the poststratifying variable, with application to political polling. *Journal of the American Statistical Association*, 96, 1-11.

Royall, R.M. (1968). An old approach to finite population sampling theory. *Journal of the American Statistical Association*, 63, 1269-1279.

Rubin, D.B. (1976). Inference and missing data (avec discussion). *Biometrika*, 63, 581-592.

Rubin, D.B. (1983). Comment on "An evaluation of model-dependent and probability-sampling inferences in sample surveys", par M.H. Hansen, W.G. Madow et B.J. Tepping. *Journal of the American tatistical Association*, 78, 803-805.

Särndal, C.-E., Swensson, B. et Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Si, Y., et Gelman, A. (2014). Survey weighting for New York Longitudinal Survey on Poverty Measure. Rapport technique, Columbia University.

Si, Y., et Zhou, P. (2020). Bayes-raking: Bayesian finite population inference with known margins. *Journal of Survey Statistics and Methodology*, smaa008.

Si, Y., Pillai, N.S. et Gelman, A. (2015). Nonparametric Bayesian weighted sampling inference. *Bayesian Analysis*, 10, 605-625.

Si, Y., Trangucci, R. et Gabry, J.S. (2020). Computation codes for manuscript "Bayesian hierarchical weighting adjustment and survey inference". https://github.com/yajuansisophie/weighting.

Stan Development Team (2017). Stan modeling language user's guide and reference manual. http://mc-stan.org.

Stan Development Team (2018). Stan: A C++ library for probability and sampling. http://mcstan.org.

Tang, X., Ghosh, M., Ha, N.S. et Sedransk, J. (2018). Modeling random effects using global-local shrinkage priors in small area estimation. *Journal of the American Statistical Association*, 0(0), 1-14.

Valliant, R., Dever, J.A. et Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples*, 2nd Edition. New York: Springer.

Valliant, R., Dorfman, A. et Royall, R. (2000). *Finite Population Sampling and Inference*. New York: John Wiley & Sons, Inc.

Volfovsky, A., et Hoff, P. (2014). Hierarchical array priors for ANOVA decompositions of cross-classified data. *Annals of Applied Statistics*, 8(1), 19-47.

Wang, W., Rothschild, D., Goel, S. et Gelman, A. (2015). Forecasting elections with nonrepresentative polls. *International Journal of Forecasting*, 31(3), 980-991.

Wimer, C., Garfinkel, I., Gelblum, M., Lasala, N., Phillips, S., Si, Y., Teitler, J. et Waldfogel, J. (2014). Poverty tracker − Monitoring poverty and well-being in NYC. Columbia Population Research Center and Robin Hood Foundation.

Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453), 185-193.

Xia, X., et Elliott, M.R. (2016). Weight smoothing for generalized linear models using a Laplace prior. *Journal of Official Statistics*, 32(2), 507-539.

Yougov (2017). Introducing the Yougov referendum model. https://yougov.co.uk.

Yuan, M., et Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68, 49-67.

Zhang, X., Holt, J.B., Yun, S., Lu, H., Greenlund, K.J. et Croft, J.B. (2015). Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. *American Journal of Epidemiology*, 182(2), 127-137.