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

Techniques d’enquête

Techniques d’enquête
50-2

Date de diffusion : le 20 décembre 2024

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Publication autorisée par le ministre responsable de Statistique Canada

© Sa Majesté le Roi du chef du Canada, représenté par le ministre de l'Industrie, 2024

L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada.

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête

Nº 12-001-XPB au catalogue

Une revue
éditée
par Statistique Canada

Décembre 2024

•

Volume 50

•

Numéro 2

Statistique Canada Statistics Canada

Canada

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans *The ISI Web of knowledge* (*Web of science*), *The Survey Statistician*, *Statistical Theory and Methods Abstracts* et *SRM Database of Social Research Methodology*, *Erasmus University*. On peut en trouver les références dans *Current Index to Statistics*, et *Journal Contents in Qualitative Methods*. La revue est également citée par *SCOPUS* sur les bases de données *Elsevier Bibliographic Databases*.

COMITÉ DE DIRECTION

Président E. Rancourt

Anciens présidents C. Julien (2013-2018)
J. Kovar (2009-2013)
D. Royce (2006-2009)
G.J. Brackstone (1986-2005)
R. Platek (1975-1986)

Membres J.-F. Beaumont
D. Haziza
W. Yung

COMITÉ DE RÉDACTION

Rédacteur en chef J.-F. Beaumont, *Statistique Canada*

Anciens rédacteurs en chef W. Yung (2016-2020)
M.A. Hidiroglou (2010-2015)
J. Kovar (2006-2009)
M.P. Singh (1975-2005)

Rédacteurs associés

- J.M. Brick, *Westat Inc.*
- P.J. Cantwell, *U.S. Census Bureau*
- G. Chauvet, *École nationale de la statistique et de l'analyse de l'information*
- S. Chen, *University of Oklahoma Health Sciences Center*
- J. Chipperfield, *Australian Bureau of Statistics*
- J.L. Eltinge, *U.S. Census Bureau*
- A. Erciulescu, *Westat Inc.*
- W.A. Fuller, *Iowa State University*
- D. Haziza, *University of Ottawa*
- M.A. Hidiroglou, *Statistique Canada*
- D. Judkins, *ABT Associates Inc Bethesda*
- J.K. Kim, *Iowa State University*
- P.S. Kott, *RTI International*
- P. Lahiri, *University of Maryland*
- É. Lesage, *L'Institut national de la statistique et des études économiques*
- A. Matei, *Université de Neuchâtel*
- K. McConville, *Reed College*
- I. Molina, *Universidad Complutense de Madrid*
- J. Opsomer, *Westat Inc*
- D. Pfeffermann, *University of Southampton*
- J.N.K. Rao, *Carleton University*
- L.-P. Rivest, *Université Laval*
- A. Ruiz-Gazen, *Toulouse School of Economics*
- F.J. Scheuren, *National Opinion Research Center*
- P.L.d.N. Silva, *Escola Nacional de Ciências Estatísticas*
- P. Smith, *University of Southampton*
- D. Steel, *University of Wollongong*
- M. Torabi, *University of Manitoba*
- D. Toth, *U.S. Bureau of Labor Statistics*
- J. van den Brakel, *Statistics Netherlands*
- C. Wu, *University of Waterloo*
- W. Yung, *Statistique Canada*
- L.-C. Zhang, *University of Southampton*

Rédacteurs adjoints C. Bocci, K. Bosa, C. Boulet, S. Matthews, C.O. Nambeu et Y. You, *Statistique Canada*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie généralement des articles innovateurs de recherche théorique ou appliquée, et parfois des articles de synthèse, qui fournissent des idées nouvelles sur les méthodes statistiques pertinentes pour les bureaux nationaux de statistique et d'autres organismes statistiques. Les sujets d'intérêt sont mentionnés sur le site Web de la revue (www.statcan.gc.ca/techniquesdenquete). Les auteurs peuvent soumettre leurs articles à la section régulière de la revue ou à la section des notes courtes pour les contributions contenant moins de 3 000 mots, incluant les tableaux, les figures et la bibliographie. Bien que le processus d'examen puisse être simplifié pour les notes courtes, tous les articles sont soumis à une évaluation par des pairs. Cependant, les auteurs demeurent responsables du contenu de leur article et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée en version électronique deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le soumettre en français ou en anglais via le **portail de *Techniques d'enquête* sur le site Web de ScholarOne Manuscripts** (https://mc04.manuscriptcentral.com/surveymeth). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue ou sur le site web (www.statcan.gc.ca/techniquesdenquete). Pour communiquer avec le rédacteur en chef, veuillez utiliser l'adresse suivante : (statcan.smj-rte.statcan@statcan.gc.ca).

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 50, numéro 2, décembre 2024

Table des matières

Article sollicité Waksberg

Richard Valliant
Plan d'échantillonnage à partir de modèles 167

Articles réguliers

Daniell Toth et Kelly S. McConville
Modèles de forêt aléatoire convergents par rapport au plan pour la collecte de données recueillies à partir d'un échantillon complexe 207

Glen Meeden et Muhammad Nouman Qureshi
Échantillonnage en grappes adaptatif, une approche quasi bayésienne 231

Caren Hasler
Inférence à l'aide de l'échantillonnage s'appuyant sur les probabilités de réponse estimées par calage 259

Nicholas T. Longford
Calage assoupli de poids d'enquête 287

Xueying Tang et Liangliang Zhang
Distribution *a priori* gamma hiérarchique pour la modélisation des effets aléatoires dans l'estimation sur petits domaines 313

Xiyue Liao, Mary C. Meyer et Xiaoming Xu
Estimation fondée sur un modèle des domaines petits et vides dans l'analyse des données d'enquête à l'aide de contraintes d'ordre 331

Teng Liu, F. Jay Breidt et Jean D. Opsomer
Une approche d'estimation sur petits domaines pour concilier les différences entre deux enquêtes sur l'effort de pêche récréative 351

Shirley Mathur, Yajuan Si et Jerome P. Reiter
Données entièrement synthétiques pour des enquêtes complexes 375

Abel Dasylva, Arthur Goussanou et Christian-Olivier Nambeu
Modèles d'erreur de couplage pour l'estimation par capture-recapture sans vérifications manuelles 407

Wenshan Yu, Michael R. Elliott et Trivellore E. Raghunathan
Examiner les effets du mode d'enquête dans les variances de l'intervieweur grâce à deux enquêtes multimodales représentatives 443

Shiya Wu, Harm-Jan Boonstra, Mirjam Moerbeek et Barry Schouten
Plan d'enquête adaptatif robuste pour les variations temporelles des propensions à répondre dans un contexte multimodal 473

Ashley Lockwood et Balgobin Nandram
Inférence prédictive bayésienne d'une moyenne de population finie sans préciser la relation entre la variable d'étude et les covariables 507

Jacek Wesołowski, Robert Wieczorkowski et Wojciech Wójciak
Algorithme récursif de Neyman pour la répartition optimale d'échantillons sous contraintes de boîtes sur les tailles d'échantillons dans les strates 531

Jorge González Chapela
Rythme quotidien de la qualité des données : résultats d'une enquête menée auprès des chômeurs au New Jersey 559

Communication brève

Phillip S. Kott et Burton Levine
Exploration d'une conjecture sur l'asymétrie : élargissement de la règle de Cochran à une proportion estimée à partir d'un échantillon complexe 591

Remerciements 601
Annonces 603
Autres revues 605

Techniques d'enquête, décembre 2024
Vol. 50, nº2, pp. 165-166
Statistique Canada, nº 12-001-X au catalogue

Série Waksberg d'articles sollicités

La revue *Techniques d'enquête* a mis sur pied en 2001 une série annuelle d'articles sollicités en l'honneur de Joseph Waksberg, en reconnaissance de ses contributions exceptionnelles à la statistique et méthodologie d'enquête. Chaque année, un éminent statisticien d'enquête est choisi par un comité de sélection composé de quatre personnes désignées par *Techniques d'enquête* et l'*American Statistical Association*. Le statisticien sélectionné est invité à rédiger un article pour *Techniques d'enquête* qui passe en revue l'évolution et l'état actuel d'un thème important du domaine de la statistique et méthodologie d'enquête. L'article reflète le mélange de théorie et de pratique caractéristique des travaux de Joseph Waksberg. Le lauréat du prix Waksberg est également invité à présenter la communication sollicitée Waksberg, habituellement au Symposium de Statistique Canada, et reçoit une prime en argent.

Veuillez consulter la section annonces à la fin de la revue pour des informations sur le processus de nomination et de sélection du prix Waksberg 2026.

Ce numéro de *Techniques d'enquête* commence par le 24e article de la série du prix Waksberg. Le comité de rédaction remercie les membres du comité de sélection, composé de Maria Giovanna Ranalli (présidente), Denise Silva, Jae-Kwang Kim et Kristen Olson d'avoir choisi Richard Valliant comme auteur de l'article du prix Waksberg de 2024.

Communication sollicitée pour le prix Waksberg 2024

Auteur : Richard Valliant

Richard Valliant est professeur-chercheur émérite à l'Université du Michigan et au Joint Program for Survey Methodology à l'Université du Maryland. Il est titulaire d'un doctorat en biostatistique de l'Université Johns Hopkins et d'une maîtrise en statistique de l'Université Cornell. Il possède plus de 45 ans d'expérience en matière d'échantillonnage, de théorie de l'estimation et d'informatique statistique. Dans le passé, il a été codirecteur à Westat et statisticien-mathématicien au Bureau of Labor Statistics (BLS). Il détient une vaste expérience de l'estimation d'enquête et de l'établissement de plan de sondage acquise dans le cadre d'une variété d'enquêtes menées auprès des établissements et des ménages, y compris l'Indice des prix à la consommation, l'Enquête sur l'état de la population et d'autres enquêtes menées par le BLS, le National Center for Education Statistics, la Consumer Product Safety Commission, le Department of Energy et le National Agricultural Statistical Service, entre autres. Il est membre de l'American Statistical Association et membre élu de l'Institut international de statistique. Il a été corédacteur du *Journal of the American Statistical Association* (section sur les théories et les méthodes, de 1989 à 1993, et section sur les applications et les études de cas, de 1996 à 1999), ainsi que du *Journal of Official Statistics* (de 2003 à 2010) et de la revue *Techniques d'enquêtes* (de 1996 à 2007).

Richard est coauteur des trois livres suivants : « Finite Population Sampling and Inference: A Prediction Approach » (2000), avec A. Dorfman et R. M. Royall; « Survey Weights: A Step-by-step Guide to Calculation » (2018), avec J.A. Dever; et « Practical Tools for Designing and Weighting Survey Samples » (2018, 2e édition), avec J.A. Dever et F. Kreuter. La première édition du livre « Practical Tools » a remporté le « Book Award » de l'American Association for Public Opinion Research en 2020. De plus, Richard est l'auteur des progiciels R PracTools et svydiags.

Gagnants du prix Waksberg et leurs articles sollicités depuis 2001

2025 Michael A. **Hidiroglou**, Manuscrit en préparation prévu pour le numéro de décembre 2025.

2024 Richard **Valliant**, « Plan d'échantillonnage à partir de modèles ». *Techniques d'enquête*, vol. 50, 2, 167-205.

2023 Raymond **Chambers**, « Le principe de l'information manquante – Un paradigme d'analyse de données désordonnées d'enquête par sondage ». *Techniques d'enquête*, vol. 49, 2, 237-278.

2022 Roderick **Little**, « Bayes, étayé par des idées fondées sur le plan, est le meilleur paradigme global pour l'inférence en enquête par échantillonnage ». *Techniques d'enquête*, vol. 48, 2, 279-306.

2021 Sharon **Lohr**, « Les enquêtes à bases de sondage multiples pour un monde fait de sources de données multiples ». *Techniques d'enquête*, vol. 47, 2, 247-285.

2020 Roger **Tourangeau**, « Science et gestion d'enquête ». *Techniques d'enquête*, vol. 47, 1, 3-32.

2019 Chris **Skinner.**

2018 Jean-Claude **Deville**, « De la pratique à la théorie : l'exemple du calage à poids bornés ». 10ème Colloque francophone sur les sondages, Université Lumière Lyon 2.

2017 Donald **Rubin**, « Le calage conditionnel et le sage statisticien ». *Techniques d'enquête*, vol. 45, 2, 199-210.

2016 Don **Dillman**, « Inciter les participants aux enquêtes à mode mixte à répondre sur le Web : les promesses et les défis ». *Techniques d'enquête*, vol. 43, 1, 3-34.

2015 Robert **Groves**, « Vers un cadre de qualité pour les mélanges de données conçues et de données organiques ». Recueil : *Symposium 2016, Croissance de l'information statistique : défis et bénéfices.*

2014 Constance **Citro**, « Des modes multiples pour les enquêtes à des sources de données multiples pour les estimations ». *Techniques d'enquête*, vol. 40, 2, 151-181.

2013 Ken **Brewer**, « Trois controverses dans l'histoire de l'échantillonnage ». *Techniques d'enquête*, vol. 39, 2, 275-289.

2012 Lars **Lyberg**, « La qualité des enquêtes ». *Techniques d'enquête*, vol. 38, 2, 115-142.

2011 Danny **Pfeffermann**, « Modélisation des données d'enquêtes complexes : Pourquoi les modéliser ? Pourquoi est-ce un problème ? Comment le résoudre ? ». *Techniques d'enquête*, vol. 37, 2, 123-146.

2010 Ivan **Fellegi**, « L'organisation de la méthodologie statistique et de la recherche méthodologique dans les bureaux nationaux de la statistique ». *Techniques d'enquête*, vol. 36, 2, 131-139.

2009 Graham **Kalton**, « Méthodes de suréchantillonnage des sous-populations rares dans les enquêtes sociales ». *Techniques d'enquête*, vol. 35, 2, 133-152.

2008 Mary **Thompson**, « Enquêtes internationales : motifs et méthodologies ». *Techniques d'enquête*, vol. 34, 2, 145-157.

2007 Carl-Erik **Särndal**, « La méthode de calage dans la théorie et la pratique des enquêtes ». *Techniques d'enquête*, vol. 33, 2, 113-135.

2006 Alastair **Scott**, « Études cas-témoins basées sur la population ». *Techniques d'enquête*, vol. 32, 2, 137-147.

2005 J.N.K. **Rao**, « Évaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage ». *Techniques d'enquête*, vol. 31, 2, 127-151.

2004 Norman **Bradburn**, « Comprendre le processus de question et réponse ». *Techniques d'enquête*, vol. 30, 1, 5-16.

2003 David **Holt**, « Enjeux méthodologiques de l'élaboration et de l'utilisation d'indicateurs statistiques pour des fins de comparaisons internationales ». *Techniques d'enquête*, vol. 29, 1, 5-19.

2002 Wayne **Fuller**, « Estimation par régression appliquée à l'échantillonnage ». *Techniques d'enquête*, vol. 28, 1, 5-25.

2001 Gad **Nathan**, « Méthodes de téléenquêtes applicables aux enquêtes-ménages – Revue et réflexions sur l'avenir ». *Techniques d'enquête*, vol. 27, 1, 7-34.

Techniques d'enquête, décembre 2024
Vol. 50, nº 2, pp. 167-205
Statistique Canada, nº 12-001-X au catalogue

Plan d'échantillonnage à partir de modèles

Richard Valliant[1]

Résumé

Joseph Waksberg était une figure de premier plan dans le domaine des statistiques d'enquête, principalement en raison de ses travaux appliqués portant sur la conception d'échantillons. Il a par ailleurs adopté une approche fondée sur le plan de sondage à l'égard des plans d'échantillonnage en mettant l'accent sur les utilisations de la randomisation dans le but de créer des estimateurs ayant de bonnes propriétés fondées sur le plan. Depuis son époque, des progrès ont été réalisés dans l'utilisation de modèles aux fins de construction de plans et dans les logiciels permettant la mise en œuvre de plans sophistiqués. Le présent article passe en revue les utilisations de modèles dans l'échantillonnage équilibré, aux échantillons définis par un seuil d'inclusion, à la stratification au moyen de modèles, à l'échantillonnage à plusieurs degrés et à la programmation mathématique pour déterminer la taille et la répartition des échantillons.

Mots-clés : Assisté par un modèle; échantillons définis par un seuil d'inclusion; échantillons équilibrés; fondé sur le plan; fondé sur un modèle; programmation mathématique; variances anticipées.

1. Introduction

Joseph Waksberg a eu une grande influence sur la pratique de l'échantillonnage et les statistiques officielles, et ce, de plusieurs manières. En effet, en poste au Bureau du recensement des États-Unis des années 1950 aux années 1970, il a dirigé les premières études sur l'erreur de mémoire, l'erreur de couverture, l'estimation sur petits domaines, l'échantillonnage de populations rares, les progrès réalisés dans l'échantillonnage des ménages, notamment l'utilisation de listes d'adresses, la rotation des zones d'échantillonnage, l'utilisation de données administratives et l'amélioration des méthodes d'échantillonnage par téléphone. L'entrevue de David Morganstein et de David Marker avec lui dans *Statistical Science* traite de bon nombre des domaines auxquels il a contribué (Morganstein et Marker, 2000).

Il a surtout travaillé sur des questions liées au plan de sondage, mais sa réflexion ne se limitait pas à des considérations mathématiques. Selon l'application, il adaptait les méthodes pour tenir compte des aspects pratiques. Au début des années 1960, il a étudié avec Neter le télescopage dans une enquête sur les dépenses de consommation (Neter et Waksberg, 1964). Les erreurs de réponse dans les enquêtes sur les dépenses étaient un problème connu (par exemple voir Cole et Utting, 1956; Ferber, 1955), mais peu d'études s'y intéressaient directement. Neter et Waksberg (1964) ont mené une expérience parrainée par le Bureau du recensement des États-Unis pour étudier la tendance qu'ont les gens à fournir des renseignements inexacts sur la période à laquelle les dépenses ont eu lieu. En particulier, on déclarait souvent que les dépenses importantes avaient eu lieu à une date plus proche du moment présent que leur date réelle, c'est-à-dire qu'elles avaient fait l'objet d'un *télescopage* en aval. Selon leurs constatations, ils ont été les premiers à proposer la technique appelée *rappel borné* comme solution possible. Lors de la deuxième interview ou d'une interview ultérieure dans le cadre d'une enquête continue, les dépenses qui avaient été déclarées lors de l'interview précédente sont indiquées au répondant et il lui est ensuite demandé les dépenses supplémentaires engagées depuis.

1. Richard Valliant, professeur émérite, Universities of Michigan & Maryland, États-Unis. Courriel : valliant@umich.edu.

Par ailleurs, Waksberg a étudié les données erronées utilisées aux fins de conception d'échantillons. Lorsqu'il est devenu statisticien en chef de la Current Population Survey (CPS) au début des années 1960, les méthodes d'échantillonnage probabiliste aréolaire étaient bien établies. Cependant, l'enquête s'est heurtée à de nouveaux problèmes causés par l'expansion de l'économie américaine. La migration des villes vers les banlieues battait son plein et les données du recensement de 1960 devenaient de plus en plus dépassées. Les cartes utilisées pour le travail sur le terrain étaient obsolètes, et il avait été constaté que certains segments de région comptant quelques maisons de ferme selon le recensement comptaient en fait de grands lotissements bâtis sur les terrains des maisons. La croissance rapide de quartiers a donné lieu à des mesures incorrectes de la taille fondées sur le dernier recensement, ce qui entraînait des charges de travail horriblement coûteuses si le plan d'échantillonnage initial était mis en œuvre. Cela a conduit le statisticien à instaurer l'utilisation d'échantillons de permis de construire pour repérer les nouvelles constructions à l'avance et éviter de tels segments d'échantillon « surprises ».

Il a également mené des recherches sur les erreurs de couverture, qui étaient des problèmes connus des recensements et des enquêtes. Dans les années 1960, alors qu'il dirigeait la CPS, l'échantillonnage au moyen de listes d'adresses a été introduit dans cette enquête et dans d'autres enquêtes du Bureau du recensement des États-Unis comme moyen de réduire le nombre de ménages omis par inadvertance par les personnes responsables du listage sur le terrain. La méthode de compilation de listes d'adresses a commencé par l'achat d'une liste à la Donnelley Corporation. Comme Waksberg l'a expliqué dans Morganstein et Marker (2000) : « le bureau de poste avait les adresses postales dans de petites boîtes. Des envois fictifs ont été préparés pour toutes les adresses de la liste de Donnelley. Les facteurs ont placé le courrier dans ces petites boîtes, ont vérifié s'il y avait des adresses manquantes et ont rempli une fiche pour chaque adresse manquante. » [traduction] Au moyen de cette méthode et de certaines procédures spéciales, comme la vérification des bâtiments convertis en immeubles d'appartements sans numéro d'appartement désigné, ils ont compilé une liste plus complète à utiliser aux fins de l'échantillonnage dans certaines zones. Ce genre d'inventivité était caractéristique de la façon dont Joseph Waksberg, Morris Hansen et ses collègues du recensement ont résolu des problèmes pratiques.

Figure 1.1 Joseph Waksberg vers 1998.

La méthode de composition aléatoire de Mitofsky-Waksberg (MW) [Waksberg, 1978] était une autre solution à un problème pratique. Au début des années 1970, l'échantillonnage aléatoire non restreint de numéros de téléphone aux États-Unis était extrêmement inefficace pour l'échantillonnage des ménages, car environ 80 % des numéros de téléphone à 10 chiffres étaient attribués à des entreprises, à des établissements institutionnels ou à des organismes gouvernementaux, ou n'étaient pas attribués. Dans la méthode de MW, on traitait les huit premiers chiffres de la liste triée de numéros de téléphone comme des grappes (appelées « 100 îlots »), on présélectionnait les grappes en téléphonant à un numéro sélectionné au hasard dans un échantillon de 100 îlots et l'on conservait une grappe seulement si le numéro composé était résidentiel. Dans une grappe retenue, on ajoutait deux chiffres supplémentaires au numéro de grappe à huit chiffres et on téléphonait au numéro pour obtenir la taille d'échantillon souhaitée. La méthode de MW n'exige pas la connaissance des probabilités de sélection au premier ou au second degré, mais elle produit un échantillon à probabilité égale de numéros de téléphone. Étant donné qu'un pourcentage élevé de 100 îlots n'avaient pas de numéros résidentiels, l'échantillonnage de MW était beaucoup plus rentable que l'échantillonnage aléatoire non restreint. Il s'agit d'un autre exemple de son approche d'échantillonnage très pratique : face à un problème précis, il a conçu une solution intelligente adaptée à l'application visée.

L'une de ses plus importantes contributions au domaine a consisté à former des dizaines de jeunes statisticiens. Comme toutes les personnes qui ont eu la chance de travailler avec lui peuvent en témoigner, il transmettait ses connaissances principalement par une formation sur le tas plutôt que par un enseignement structuré. Il était habile à ramener des problèmes techniques complexes à des explications intuitives et compréhensibles, ce qui était particulièrement utile pour les clients et les novices encore au stade d'apprentissage. Il insistait notamment sur l'importance pour un statisticien d'échantillonnage de réfléchir non seulement aux questions précises posées, mais aussi aux aspects plus généraux de ces questions, à savoir si les questions sont logiques et peuvent être résolues, ou si elles devraient être modifiées ou remplacées.

Son approche en ce qui concerne le plan de sondage était fondée sur la randomisation, et les propriétés d'échantillonnage répété étaient primordiales. L'utilisation explicite de modèles est progressivement devenue une partie intégrante des plans et des estimations des enquêtes au fil des ans, depuis l'époque où Waksberg et ses collègues du Bureau du recensement des États-Unis et de Westat étaient en activité. Certes, leur utilisation des données auxiliaires dans le plan de sondage a certainement des aspects d'un modèle, mais ils ont rarement, voire jamais, eu recours de façon officielle à des modèles dans leur travail. Bien entendu, les modèles ont joué un rôle central dans le domaine connexe des plans expérimentaux pendant de nombreuses années (par exemple voir Box, Hunter et Hunter, 2005; Wu et Hamada, 2021). Dans le présent article, on passe en revue certaines des utilisations les plus explicites de modèles pour guider la conception d'échantillons de population finie au cours des dernières décennies. À la section 2, on examine l'échantillonnage équilibré motivé par des modèles. L'échantillonnage défini par un seuil d'inclusion, dont il est question à la section 3, est parfois utilisé dans les enquêtes auprès des établissements quand de grandes unités représentent la majeure partie de la population totale. La formation de strates au moyen de modèles, couramment utilisée dans les enquêtes auprès des entreprises, est abordée à la section 4. L'échantillonnage

à plusieurs degrés au moyen de modèles visant à estimer les composantes de la variance est présenté à la section 5. La programmation mathématique de la section 6 est très utile pour trouver des répartitions efficaces dans les enquêtes polyvalentes. Enfin, la section 7 conclut par un résumé.

2. Échantillons équilibrés

Les spécialistes utilisent depuis longtemps l'échantillonnage systématique à partir de listes triées par variables auxiliaires disponibles dans une base de sondage comme moyen d'exercer un contrôle sur la répartition d'un échantillon sélectionné. Cette technique est particulièrement utile quand plusieurs variables auxiliaires (x) se trouvent dans une base de sondage, mais que la taille de l'échantillon est trop petite pour permettre de croiser tous les x pour former des strates distinctes. Par exemple, une base de sondage d'écoles pourrait être stratifiée par région géographique et triée au sein d'une strate selon le nombre d'inscriptions afin de contrôler la répartition de l'échantillon par région et taille d'école. Une base de sondage d'îlots urbains peut être numérotée de façon sinueuse pour que les îlots proches les uns des autres dans la numérotation en série soient également géographiquement proches (Hansen, Hurwitz et Madow, 1953a; Bureau du recensement des États-Unis, 2006). Une base de sondage d'hôpitaux pourrait être stratifiée en fonction du nombre de consultations aux services d'urgence et triée géographiquement au sein des strates (Schroeder et Ault, 2001). Les établissements commerciaux peuvent être stratifiés par région géographique et code d'industrie, puis triés par le nombre d'employés dans les strates. La sélection systématique à partir de chaque liste triée produira, en espérance, un échantillon ayant un type d'*équilibre* qui dépend du fait que l'échantillon est sélectionné selon une probabilité égale ou selon des probabilités proportionnelles à une mesure de taille. Le concept d'*échantillons équilibrés* a été mis en forme par plusieurs auteurs, comme cela est expliqué ci-dessous.

L'échantillonnage équilibré a été introduit dans les années 1970 comme une méthode de protection contre le biais de prédiction (Royall et Herson, 1973a, b) au moyen d'une méthode d'échantillonnage fondée sur un modèle. Pour la notation, supposons que s désigne l'ensemble des éléments d'échantillon; U, la population de N éléments; n, la taille de l'échantillon; y_i, une variable d'analyse pour l'élément i; et x_i, une variable auxiliaire connue pour chaque élément de la population. À titre d'exemple, prenons l'estimateur par le ratio $\hat{\bar{y}}_R = \bar{y}_s\left(\bar{x}_U/\bar{x}_s\right)$, où $\bar{y}_s = \sum_{i\in s} y_i/n$, $\bar{x}_s = \sum_{i\in s} x_i/n$ et $\bar{x}_U = \sum_{i\in U} x_i/N$. L'estimateur par le ratio est le meilleur prédicteur linéaire sans biais (BLU pour *best linear unbiased*) de la moyenne, $\bar{y}_U = \sum_{i\in U} y_i/N$, selon un modèle ayant une moyenne $E_M\left(y_i\right) = \beta x_i$ et une variance $V_M\left(y_i\right) = \sigma^2 x_i$. Cependant, si la moyenne du modèle est $E_M\left(y_i\right) = \alpha + \beta x_i$, $\hat{\bar{y}}_R$ a un biais de modèle (ou un biais de prédiction) défini comme étant $E_M\left(\hat{\bar{y}}_R - \bar{y}_U\right)$, qui est égal à $\alpha\left(\bar{x}_U/\bar{x}_s - 1\right)$. Par conséquent, si le modèle a une ordonnée à l'origine plutôt que de passer par l'origine, l'estimateur par le ratio a un biais de modèle. Ce biais est nul dans tout échantillon équilibré en ce sens que $\bar{x}_s = \bar{x}_U$. Ce résultat s'étend aux cas plus complexes où, par exemple, le bon modèle est polynomial plutôt que linéaire avec une ordonnée à l'origine (Valliant, Dorfman

et Royall, 2000, section 3.1). Selon un plan d'échantillonnage aléatoire simple (EAS), $E_\pi\left(\bar{x}_s - \bar{x}_U\right) = 0,$ où E_π est l'espérance pour ce qui est de l'échantillonnage répété, et $\bar{x}_s - \bar{x}_U \xrightarrow{p} 0$ dans les grands EAS, dans certaines conditions normalisées sur la façon dont la population et l'échantillon croissent quand n et N grandissent. S'il y a d'autres covariables, $\mathbf{z},$ qui doivent être dans le modèle pour $y,$ l'EAS, en moyenne, s'équilibre aussi sur leurs moyennes, même si les $\mathbf{z}$ peuvent être inconnus au moment de l'échantillonnage. Ces résultats s'étendent à d'autres plans d'échantillonnage probabiliste qui donnent des estimateurs sans biais par rapport au plan ou asymptotiquement sans biais par rapport au plan de N et de $\bar{x}_U$.

Une exigence clé dans les calculs fondés sur un modèle ci-dessus est que l'échantillon ne soit pas *informatif* selon la définition de Pfeffermann et Sverchkov (2009). Un échantillon est informatif si le modèle qui est ajusté dans un échantillon est différent de celui qui est ajusté à la population, même après la prise en compte des covariables. Formulé en symboles, le fait d'être informatif signifie que $f\left(y_i \mid x_i, i \in s\right) \neq f\left(y_i \mid x_i, i \in U\right),$ où $f(\cdot)$ est une densité. L'informativité peut être attribuable, par exemple, à l'échantillonnage lui-même ou à la réponse à l'échantillon qui dépend de y.

Le fait que la moyenne de l'échantillon de x vise effectivement sa moyenne de population constitue un argument pour expliquer la raison pour laquelle l'EAS ou d'autres plans d'échantillonnage probabiliste peuvent être des méthodes utiles de sélection de l'échantillon aux fins de protection contre des biais inconnus. Ces propriétés pourraient laisser entendre que l'utilisation d'un plan d'échantillonnage probabiliste élimine la nécessité de se préoccuper du biais de modèle. Toutefois, si l'estimateur ponctuel a un biais de modèle, le biais de modèle au carré peut ne pas diminuer assez rapidement pour devenir une partie négligeable de l'erreur quadratique moyenne du modèle, ce qui renforce la notion selon laquelle une modélisation appropriée est essentielle. Dans l'exemple ci-dessus, le carré du biais du modèle de l'estimateur par le ratio dans un plan d'EAS est $O_P\left(n^{-1}\right),$ où O_p est l'ordre par rapport au plan d'échantillonnage probabiliste, et il en est de même pour la variance du modèle. Royall et Cumberland (1985) ont montré que dans l'EAS, un pourcentage prédictible d'échantillons seront mal équilibrés sur toute variable auxiliaire, quelle que soit la taille des échantillons. Dans ces échantillons mal équilibrés, même les intervalles de confiance construits à l'aide d'un estimateur de la variance, comme l'estimateur jackknife de la variance, qui est robuste pour certains types d'erreurs de spécification de modèle, auront une mauvaise couverture. Par conséquent, pour éliminer les préoccupations concernant le biais de modèle, il faut un plan d'échantillonnage qui réduit l'ordre du biais au carré plus rapidement que l'EAS. C'est exactement ce qu'a fait Kott (1986) en montrant que l'échantillonnage systématique à probabilité égale à partir d'une liste ordonnée est un moyen d'atteindre l'équilibre plus rapidement que par EAS.

Ces résultats de biais de modèle et l'attrait cosmétique de l'utilisation d'échantillons « bien répartis » incitent à restreindre les échantillons à l'étape du plan en faveur de ceux qui sont en quelque sorte équilibrés. Royall (1992) a généralisé l'idée d'un échantillonnage équilibré à des modèles linéaires ayant la forme

$$E_M\left(\mathbf{Y}\right) = \mathbf{X}\boldsymbol{\beta}; \qquad V_M\left(\mathbf{Y}\right) = \mathbf{V}\sigma^2, \tag{2.1}$$

où $\mathbf{Y}$ est un vecteur N de variables d'analyse, $\mathbf{X}$ est une matrice $N \times p$ de covariables et $\mathbf{V} = \text{diag}\{v_i\}_{i \in U}$ est une matrice de covariance diagonale $N \times N$. Le modèle (2.1) est assez souple, puisque les covariables peuvent être des interactions ou des transformations de variables auxiliaires.

La variance de l'erreur ou la variance de prédiction d'un estimateur, $\hat{\theta}$, d'une quantité de population, θ_U, est $E_M\left(\hat{\theta} - \theta_U\right)^2$. Par souci de précision, nous considérons les estimateurs, $\hat{t}$ du total de la population, $t_U = \sum_U y_i$. Le meilleur prédicteur linéaire sans biais de t_U est

$$\hat{t}_{\text{BLU}} = \sum_s y_i + \sum_{U-s} \hat{y}_i = \sum_{i \in U} \hat{y}_i + \sum_{i \in s} \hat{e}_{Mi}, \tag{2.2}$$

où $\hat{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$, $\hat{e}_{Mi} = y_i - \hat{y}_i$, $\mathbf{x}_i^T$ est la i^{e} ligne de $\mathbf{X}$ et $\hat{\boldsymbol{\beta}} = \left(\sum_s \mathbf{x}_i \mathbf{x}_i^T / v_i\right)^{-1} \sum_s \mathbf{x}_i y_i / v_i$ est l'estimateur par les moindres carrés généralisés de $\boldsymbol{\beta}$.

Définissons $\mathbf{1}_N$ comme étant un vecteur de taille N dont tous les éléments sont des 1 et $\mathbf{1}_s$ comme étant un vecteur de taille n dont tous les éléments sont des 1. Ensuite, quand $\mathbf{V}\mathbf{1}_N$ et $\mathbf{V}^{1/2}\mathbf{1}_N$ sont tous deux dans l'espace colonne de $\mathbf{X}$, l'échantillon qui produit la variance de l'erreur minimale pour le meilleur prédicteur linéaire sans biais est un échantillon équilibré pondéré qui satisfait

$$\frac{1}{n} \mathbf{1}_s^T \mathbf{V}_s^{-1/2} \mathbf{X}_s = \frac{\mathbf{1}_N^T \mathbf{X}}{\mathbf{1}_N^T \mathbf{V}^{1/2} \mathbf{1}_{\mathbf{N}}} \text{ ou, de façon équivalente, } \frac{1}{n} \sum_s \frac{\mathbf{x}_i}{v_i^{1/2}} = \frac{\overline{\mathbf{x}}_U}{\overline{v}_U^{(1/2)}}, \tag{2.3}$$

où $\mathbf{V}_s$ est la matrice de covariance de dimension $n \times n$ pour les n unités de l'échantillon, $\mathbf{X}_s$ est la matrice des covariables de dimension $n \times p$ pour les unités de l'échantillon et $\overline{v}_U^{(1/2)} = N^{-1} \sum_{i \in U} v_i^{1/2}$. Déterminer un ensemble d'éléments satisfaisant (2.3) avant l'échantillonnage nécessite une base de sondage où les x_i et les v_i sont individuels. Si ces derniers dépendent d'une fonction quelconque des x, il est peut être possible de les calculer. Si tous les v_i sont égaux, alors (2.3) se réduit à un équilibre simple, $\overline{\mathbf{x}}_s = \overline{\mathbf{x}}_U$.

En ayant un échantillon équilibré pondéré, le meilleur prédicteur linéaire sans biais du total des y se réduit à

$$\hat{t}_{\text{BLU}} = \frac{N \overline{v}_U^{(1/2)}}{n} \sum_{i \in s} \frac{y_i}{v_i^{1/2}}. \tag{2.4}$$

Consultez Valliant et coll. (2000, théorème 4.2.1) pour en savoir plus. On remarque que (2.3) dépend de y seulement par l'entremise de la matrice de covariance $\mathbf{V}$, mais que la réduction à (2.4) exige que $v_i^{1/2}$ et v_i soient des combinaisons linéaires des colonnes de $\mathbf{X}$. Par conséquent, si différents y ont la même structure, l'équilibre pondéré sera optimal pour eux aussi. Tam (1995) a étendu l'idée d'équilibre à des populations en grappes dans lesquelles les éléments au sein des grappes ont une corrélation du modèle. Ses résultats semblent plus difficiles à mettre en œuvre en pratique parce que l'équilibre sur les variables auxiliaires doit tenir compte des corrélations intragrappes qui dépendront de y.

Le plan de sondage assisté par un modèle repose à la fois sur un modèle et la sélection d'un échantillon aléatoire pour les analyses. Un des principaux outils de cette méthode est une variance anticipée (VA) ayant la forme

$$\mathrm{VA}\left(\hat{t}\right) = E_M E_\pi\left[\left(\hat{t} - t_U\right)^2\right] - \left[E_M E_\pi\left(\hat{t} - t_U\right)\right]^2.$$

Quand $\hat{t}$ est π -sans biais, c'est-à-dire que $E_\pi(\hat{t} - t_U) = 0,$ la variance anticipée est réduite à $E_M V_\pi(\hat{t} - t_U)$. L'optimalité de l'échantillonnage équilibré pondéré est étroitement liée aux résultats antérieurs de la variance anticipée sur l'échantillonnage à probabilité inégale. La forme réduite du meilleur prédicteur linéaire sans biais dans (2.4) est égale à l'estimateur π quand l'échantillon est sélectionné à l'aide de probabilités proportionnelles à $v_i^{1/2}$. Godambe et Joshi (1965) et Isaki et Fuller (1982) ont présenté des circonstances où la variance anticipée d'un estimateur par la régression de la moyenne de population est minimisée quand les probabilités de sélection sont proportionnelles à la racine carrée d'une variance de modèle. Une hypothèse essentielle est que les erreurs de modèle ne sont pas corrélées.

Dans le contexte de l'échantillonnage probabiliste, Deville et Tillé (2004) et Fuller (2009) proposent des méthodes qui restreignent les échantillons aléatoires à ceux dont les moyennes d'échantillon pondérées des variables auxiliaires sont proches des moyennes correspondantes de la population, c'est-à-dire

$$N^{-1}\sum_s \frac{\mathbf{x}_i}{\pi_i} \doteq \overline{\mathbf{x}}_U, \tag{2.5}$$

où π_i est la probabilité de sélection d'un élément i dans un échantillon probabiliste. (Voir aussi Ardilly, Haziza, Lavallée et Tillé [2024].) Un estimateur ayant la forme du premier terme de (2.5) est généralement appelé estimateur π (Särndal, Swensson et Wretman, 1992). La méthode de Deville-Tillé sélectionne directement des échantillons probabilistes qui satisfont approximativement (2.5); la méthode de Fuller rejette les échantillons probabilistes où l'expression (2.5) n'est pas satisfaite selon une tolérance précise. Selon la terminologie de Cumberland et Royall (1981) et de Royall (1992), les échantillons qui satisfont (2.5) sont équilibrés sur π. Deville et Tillé (2004, 2005) abordent le calcul des poids et des estimateurs de la variance pour les échantillons probabilistes équilibrés en utilisant ce qu'ils appellent la méthode du « cube » qui a une interprétation fondée sur le plan. Nedyalkova et Tillé (2008) ont généralisé les résultats de Godambe-Joshi et de Fuller-Isaki pour montrer qu'une stratégie optimale assistée par un modèle (c'est-à-dire une stratégie qui réduit le plus possible la variance anticipée) pour l'estimateur π selon le modèle (2.1) consiste à sélectionner un échantillon équilibré sur π de taille fixe sur les x dans le modèle. On peut obtenir un échantillon de taille fixe en incluant $\mathbf{x}_i = \pi_i$ dans les conditions d'équilibre.

Le paquet de R `sampling` (Tillé et Matei, 2023) sélectionnera des échantillons équilibrés pondérés ou non pondérés qui satisfont soit (2.3), soit (2.5). Si un échantillon probabiliste est conçu de telle sorte que $\pi_i = nv_i^{1/2} / (N\overline{v}_U^{(1/2)}),$ le résultat sera optimal à la fois selon le modèle et selon le plan, au moins pour la variable clé y servant à attribuer les probabilités de sélection. On peut ensuite utiliser un estimateur de la variance comme l'estimateur jackknife, qui a de bonnes propriétés de plan et de modèle. Si $V_M\left(y_i\right) \neq v_i$ pour certains $y,$ alors un échantillon probabiliste sélectionné où $\pi_i = nv_i^{1/2} \Big/ \left(N\overline{v}_U^{(1/2)}\right)$ peut être quelque peu inefficace dans le modèle pour ces $y,$ mais permettra tout de même d'obtenir une estimation sans biais par rapport au plan ou convergente.

L'Institut national de la statistique et des études économiques (INSEE) a utilisé la méthode du cube pour sélectionner un échantillon maître d'unités primaires d'échantillonnage (UPE) [Costa, Guillo, Paliod, Merly-Alpa, Vincent, Chevalier et Deroyon, 2018] et l'a utilisée dans son recensement pour sélectionner des échantillons de municipalités à probabilité égale qui sont équilibrés sur un ensemble de variables démographiques (voir Deville et Tillé, 2004). L'application de l'INSEE comporte une caractéristique clé qui manque à de nombreuses applications : les municipalités ne peuvent pas être des non-répondants dans le recensement. Dans les cas où les unités peuvent être des non-répondants, il se peut que l'équilibre initial d'un échantillon soit perdu, ce qui, au mieux, est une nuisance et, au pire, entraîne des biais. En principe, le fait de substituer un non-répondant par un élément ayant la même valeur $\mathbf{x}_i$ préserve l'absence de biais et l'optimalité du modèle dans (2.1). Cependant, cela perturberait les propriétés du plan de sondage parce qu'une substitution est une imputation qui ajoute de la variance, voire un biais.

La restriction de la configuration géographique des unités d'échantillonnage du premier degré est souhaitée depuis longtemps lors de la conception d'échantillons aréolaires (Kish, 1965, section 12.8) et est liée à l'équilibrage. Le fait que les UPE de l'échantillon soient dispersées sur une carte de l'univers est particulièrement satisfaisant dans les échantillons probabilistes aréolaires. De plus, il peut y avoir plusieurs variables de stratification possibles, comme la densité de la population, le niveau de scolarité et la concentration des groupes ethniques, dont l'utilisation pourrait améliorer l'efficacité des estimateurs, mais qui ne peuvent pas être pleinement utilisées en raison de la taille limitée de l'échantillon, comme dans les cas d'échantillonnage systématique mentionnés ci-dessus. Goodman et Kish (1950) ont proposé une méthode à une UPE par strate appelée *sélection contrôlée* conçue pour ces types de restrictions. Cette méthode et d'autres, comme les carrés latins (Frankel et Stock, 1942), restreignent les configurations d'unités d'échantillonnage et attribuent une probabilité de sélection à chaque configuration acceptable (Hansen et coll., 1953a, section 11.4). Bien que l'on puisse obtenir la restriction de l'échantillon par ces méthodes, les estimateurs de la variance qui étaient utilisés à ce moment-là ne représentaient pas de gains de précision (le cas échéant) en raison de l'échantillonnage restreint.

Grafström, Lundström et Schelin (2012) et Grafström (2012) ont introduit d'autres méthodes qui contrôlent la dispersion d'un échantillon sur une population. Au lieu de reposer sur une cartographie, ces méthodes reposent sur la distance entre les unités pour créer de petites probabilités d'inclusion conjointes pour les unités avoisinantes, ce qui force les échantillons à être bien dispersés. Grafström et Tillé (2013) ont proposé une méthode qui est doublement équilibrée dans le sens qu'elle sélectionne des échantillons équilibrés sur plusieurs variables auxiliaires et aussi bien dispersés pour d'autres variables comme les coordonnées topographiques. Grafström et Tillé (2013) ont utilisé un modèle linéaire ayant la forme $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i$. On suppose que les erreurs de modèle ont la structure de covariance $\text{cov}_\xi(\varepsilon_i, \varepsilon_j) = \sigma_i\sigma_j\rho^{d_{ij}}$, où d_{ij} est une mesure de la distance entre les éléments i et j et $0 < \rho < 1$. Par conséquent, la corrélation diminue à mesure que les éléments s'éloignent. Dans l'échantillonnage aréolaire, un centroïde de latitude et de longitude est souvent associé à chaque unité du premier degré et peut servir à calculer la distance entre toute paire d'unités.

Toutes les méthodes d'équilibrage sont disponibles dans le paquet de R appelé `BalancedSampling` (Grafström, Lisic et Prentius, 2023). Étant donné que les méthodes d'équilibrage sont des échantillons probabilistes ayant des probabilités connues de sélection unique et conjointe, on peut utiliser des estimateurs de la variance normalisés fondés sur le plan de sondage. En raison de limites pratiques, il est souvent impossible de sélectionner des échantillons exactement équilibrés. Dans de tels cas ou quand des x supplémentaires se révèlent prédictifs des variables d'analyse, on peut utiliser soit des estimateurs par la régression généralisée, soit des estimateurs de moyennes et de totaux purement fondés sur un modèle, en plus des estimateurs de la variance de Valliant et coll. (2000), de Valliant (2002) ou de Särndal et coll. (1992).

Étant donné que l'analyse ci-dessus repose sur des modèles linéaires, on se demande naturellement si les résultats d'équilibrage s'étendent aux modèles non linéaires comme $E_M(y_i) = \mu(\mathbf{x}_i, \boldsymbol{\theta})$ selon $V_M(y_i) = v_i(\boldsymbol{\theta})$, où $\mu(\mathbf{x}_i, \boldsymbol{\theta})$ est une fonction d'un vecteur de covariables et d'un paramètre inconnu $\boldsymbol{\theta}$. La moyenne, $\mu(\mathbf{x}_i, \boldsymbol{\theta})$, peut être linéaire ou non linéaire dans $\mathbf{x}_i$, de sorte que les y quantitatifs ou catégoriques sont couverts. Un estimateur assisté par un modèle d'un total, conçu pour ce modèle non linéaire, est (Breidt et Opsomer, 2009)

$$\hat{t}_{\text{MA}} = \sum_U \hat{\mu}_i + \sum_s \frac{e_{\text{MA}i}}{\pi_i},$$

où $e_{\text{MA}i} = y_i - \hat{\mu}_i$. Bien que la forme soit semblable à $\hat{t}_{\text{BLU}}$ dans (2.2), aucun résultat d'équilibrage n'est disponible pour $\hat{t}_{\text{MA}}$ ou pour l'estimateur calé sur un modèle de Wu et Sitter (2001), qui est également fondé sur un modèle non linéaire.

3. Échantillons définis par un seuil d'inclusion

Dans les applications où un petit nombre d'unités représente une part démesurée des totaux de population, la méthode standard consiste à inclure les grandes unités dans l'échantillon avec certitude et à sélectionner un échantillon aléatoire à partir du reste de la population. Une méthode plus radicale consiste à sélectionner un échantillon défini par un seuil d'inclusion. Les échantillons définis par un seuil d'inclusion sont ceux dans lesquels seuls les éléments présentant une caractéristique précisée sont échantillonnés. Les seuils d'inclusion sont souvent quantitatifs, par exemple les revenus dans les établissements commerciaux ou les niveaux de production des raffineries de pétrole. Si l'on souhaite avoir des estimations pour une population totale, elles peuvent être justifiées si a) les unités de l'échantillon et les unités hors de l'échantillon suivent le même modèle de superpopulation ou si b) le seuil d'inclusion présente un certain caractère aléatoire de sorte que la propension à être inclus dans l'échantillon peut être modélisée selon des covariables au niveau des éléments. Les échantillons définis par un seuil d'inclusion peuvent aussi être considérés comme des cas particuliers de plans stratifiés comportant des strates à tirage nul, à tirage partiel et à tirage complet, comme le décrit la section 4.

Des échantillons de ce type sont même mentionnés dans Hansen et coll. (1953a, pages 486 à 490), qui font remarquer que ce plan peut être efficace dans les populations d'établissements où un petit nombre de grandes unités représentent un fort pourcentage des totaux estimés et où la collecte de données tirées de petites unités ne serait pas rentable. Il se peut que la restriction d'un échantillon d'une façon ou d'une autre soit inévitable si certains membres d'une population cible sont inaccessibles. À titre d'exemple, si la collecte des données doit être réalisée par téléphone, les ménages sans téléphone sont exclus; les personnes vivant en établissement (par exemple les personnes incarcérées ou vivant dans des établissements de soins infirmiers) peuvent être exclues des enquêtes auprès des ménages en raison de la difficulté à recueillir leurs données. Si des estimations sont nécessaires pour l'ensemble de la population, une exigence essentielle pour justifier le point a) ci-dessus est que les prédictions pour les unités hors de l'échantillon peuvent être effectuées à partir des unités de l'échantillon. Cela est possible quand le même modèle se vérifie pour les unités de l'échantillon et les unités hors de l'échantillon. L'inclusion dans l'échantillon défini par un seuil d'inclusion doit également pouvoir être ignorée, c'est-à-dire qu'elle ne peut pas dépendre directement des y à analyser.

Dans certaines applications, un échantillon défini par un seuil d'inclusion non aléatoire sera optimal selon le modèle aux fins d'estimation d'un total. Par exemple, considérons le modèle de ratio $E_M\left(y_i\right)=\beta x_i$ et $V_M\left(y_i\right)=\sigma^2 x_i,$ où les y sont indépendants. Supposons que l'objectif est d'estimer le total des y pour l'ensemble de la population. Les valeurs pour les unités hors de l'échantillon sont prédites comme étant $\hat{\beta}x_i,$ et la variance de l'erreur de l'estimateur par le ratio est $V_M\left(\hat{\bar{y}}_R-\bar{y}_U\right)=\left(1-n/N\right)\sigma^2\bar{x}_r\bar{x}_U/\left(n\bar{x}_s\right),$ où $\bar{x}_r=\left(N-n\right)^{-1}\sum_{i\in U-s}x_i.$ Dans ce cas, le plan de sondage optimal qui minimise la variance de l'erreur est non aléatoire et consiste à prendre les n unités ayant les plus grandes valeurs de x. Si y est une valeur de la période courante et x est une valeur du recensement pour la même variable à une période antérieure et que les conditions économiques n'ont pas changé radicalement depuis le recensement, le modèle de ratio peut être bien ajusté.

Dans les populations extrêmement asymétriques où les plus grandes unités représentent un pourcentage élevé de la population totale, la perspective de faire de mauvaises prédictions pour les unités plus petites est moins inquiétante. Toutefois, ce type d'échantillonnage défini par un seuil d'inclusion est risqué parce qu'il élimine la possibilité de tester le degré d'ajustement du modèle pour les unités plus petites. Si des estimations par domaine pour des unités de petite et de moyenne taille sont requises, il ne faut pas utiliser d'échantillon défini par un seuil d'inclusion en raison du risque que les domaines suivent un modèle différent de celui qui ajuste l'échantillon défini par un seuil d'inclusion. La classification erronée est une autre préoccupation. Si une grande unité qui devrait faire partie de l'échantillon défini par un seuil d'inclusion est classée de façon erronée comme faisant partie de la portion à tirage nul de la population, l'échantillon défini par un seuil d'inclusion peut exclure un facteur important pour le total de la population. La dégradation des modèles au fil du temps est également un sujet de préoccupation. En période de volatilité économique, l'ajustement d'un modèle peut convenir pendant un certain temps, mais échouer en cas de récession ou de ralentissement. En vue de s'en protéger partiellement, Benedetti, Bee et Espa (2010) ont

étendu l'idée de définition par un seuil d'inclusion en stratifiant la population en trois – la strate à tirage complet, celle à tirage partiel et celle à tirage nul – et ont élaboré un algorithme pour diviser une population en strates et leur attribuer l'échantillon.

De plus, la non-réponse est une autre source de préoccupation. Si une ou plusieurs unités extrêmement grandes ne coopèrent pas, il peut être difficile de tenir compte de cette non-réponse dans les échantillons définis par un seuil d'inclusion ou dans des échantillons plus classiques. Si une grande unité est unique, il se peut que la pondération des répondants ou l'imputation pour un non-répondant ne soit pas une solution viable. Par exemple, dans une enquête portant sur les cultures agricoles, si les exploitations agricoles appartenant à une grande entreprise agroalimentaire refusent de fournir des données, une grande partie de la production de maïs, de blé et d'autres cultures sera manquante. Dans de tels cas, les valeurs des répondants peuvent avoir une utilisation limitée comme sources d'imputation. La présence de grandes unités non répondantes qui ont été échantillonnées avec certitude peut nuire à l'objectif d'estimation pour toute la population, à moins que soit conçue une bonne méthode pour les imputer. Aux États-Unis, tous les cinq ans, le recensement de l'agriculture est mené pour dénombrer toutes les fermes et tous les ranchs. Les données du recensement peuvent être utiles aux fins d'imputation des données manquantes sur les cultures agricoles dans les enquêtes-échantillons menées hors des années de recensement, si toutes les grandes unités participent au recensement. Cependant, dans le cadre d'autres enquêtes menées auprès d'établissements commerciaux ou d'autres établissements, on n'a pas nécessairement ce luxe.

Yorgason, Bridgman, Cheng, Dorfman, Lent, Liu, Miranda et Rumburg (2011) ont examiné certaines applications de l'échantillonnage défini par un seuil d'inclusion réalisées par des organismes fédéraux américains. En particulier, l'Energy Information Administration (EIA) du département de l'Énergie des États-Unis mène des enquêtes mensuelles auprès des producteurs de pétrole brut et de gaz naturel au moyen d'échantillons définis par un seuil d'inclusion qui couvrent au moins 85 % de la production totale de pétrole et de gaz de chaque État (U.S. Energy Information Administration, 2018). La production des entreprises non échantillonnées est implicitement imputée au moyen d'un estimateur par le ratio. L'EIA effectue également une enquête mensuelle sur les services d'électricité au moyen d'échantillons définis par un seuil d'inclusion similaires de grands producteurs (Kirkendall, 1992; Knaub, 2008). L'évolution rapide de l'économie de l'énergie en 2008 illustre le risque posé par les échantillons définis par un seuil d'inclusion. Selon Yorgason et coll. (2011, page 3) : « Quand les prix du pétrole et du gaz naturel ont commencé à augmenter rapidement en 2008, les grands exploitants de puits de gaz naturel (compris dans l'échantillon) ont augmenté leurs taux de production plus rapidement que les petits exploitants (non compris dans l'échantillon). Outre l'incitatif à la production généré par la hausse des prix, les progrès technologiques ont permis à certaines grandes entreprises d'augmenter leurs taux d'extraction de gaz de schiste. Les taux réels de couverture de l'échantillon EIA-914 ont augmenté, et les taux de couverture estimés, fondés sur des données antérieures, n'ont pas traduit les changements assez rapidement. C'est la raison pour laquelle l'EIA a surestimé la production de gaz naturel dans certains États. » [traduction] Les analystes de l'industrie ont affirmé que les estimations surestimées de la production avaient artificiellement fait baisser les prix du gaz naturel sur le marché.

Haziza, Chauvet et Deville (2010) ont cité un exemple de données fiscales de Statistique Canada tiré de Fecteau et Jocelyn (2005). Les entreprises canadiennes non constituées en société peuvent déclarer leurs états financiers sur papier ou par voie électronique. En raison des coûts élevés de la conversion des données recueillies sur papier en format électronique, les déclarants sur papier sont délibérément exclus de la possibilité de sélection dans l'échantillon. En effet, les estimations de la population sont fondées sur un échantillon sélectionné à partir des déclarants par voie électronique seulement. Dans cet exemple portant sur des données fiscales, s'il est raisonnable de concevoir la situation comme une situation où il y a une probabilité de produire une déclaration sur papier ou par voie électronique, la propension à produire une déclaration par voie électronique peut être estimée à partir de covariables. (Il convient de mentionner que les covariables ne doivent pas comprendre la variable qui détermine si un élément se trouve ou non dans l'échantillon défini par un seuil d'inclusion.) Ensuite, la pondération par l'inverse de la propension à répondre au questionnaire de l'enquête des déclarants par voie électronique peut servir à l'estimation, comme cela se fait parfois dans les échantillons non probabilistes (par exemple voir Elliott et Valliant, 2017). S'il n'y a aucune probabilité qu'un déclarant par voie électronique produise sa déclaration sur papier, la pondération par l'inverse de la propension à répondre au questionnaire échouerait, mais il y a peu de chance que l'on sache cela dans une situation donnée. Cela reviendrait à avoir des non-répondants irréductibles qui n'ont aucune probabilité de participer à l'enquête.

4. Stratification au moyen de modèles

Supposons que la population est divisée en $h = 1, \ldots, H$ strates et que N_h éléments se trouvent dans la strate h. La population d'éléments dans la strate h est désignée par U_h. La proportion d'unités dans la strate h est $W_h = N_h / N$, et la moyenne de la population de y est $\bar{y}_U = \sum_{h=1}^{H} W_h \bar{y}_{Uh}$, où $\bar{y}_{Uh}$ est la moyenne de la population de y au sein de la strate h. Pour les analyses fondées sur le plan, supposons qu'un échantillon aléatoire simple est sélectionné sans remise dans chaque strate. La taille de l'échantillon dans la strate h est n_h, l'ensemble des unités d'échantillon dans h est s_h et la taille totale de l'échantillon est $n = \sum_{h=1}^{H} n_h$. On obtient alors un estimateur de la moyenne de la population $\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_{sh}$, où $\bar{y}_{sh}$ est la moyenne de l'échantillon dans la strate h. La variance par rapport au plan de $\bar{y}_{st}$ est $V_\pi\left(\bar{y}_{st}\right) = \sum_{h=1}^{H} W_h^2 \left(n_h^{-1} - N_h^{-1}\right) S_{yUh}^2$, où S_{yUh}^2 est la variance de la population de y dans la strate h. La variance relative par rapport au plan de $\bar{y}_{st}$ est définie comme étant $V_\pi\left(\bar{y}_{st}\right) / \bar{y}_U^2$; le coefficient de variation (CV) de $\bar{y}_{st}$ est la racine carrée de la variance relative. Les questions de base du plan portent sur la meilleure façon de former les strates et d'attribuer l'échantillon aux strates.

Les modèles sont plus utiles quand il faut former des strates dans des populations où un échantillonnage à un degré, comme l'échantillonnage aléatoire simple stratifié sans remise, est utilisé. Cela comprend les populations d'établissements commerciaux, d'écoles, d'hôpitaux ou d'autres établissements institutionnels. Dans certains échantillons, les strates peuvent être dictées par des objectifs de publication pour lesquels les modèles peuvent présenter une utilisation limitée. Par exemple, dans une enquête auprès des entreprises, il peut falloir des statistiques distinctes pour le commerce de détail, le commerce de gros, la fabrication et

d'autres secteurs. Toutefois, dans un même secteur, on peut utiliser un modèle pour former des sous-strates au moyen des méthodes décrites ci-dessous.

Quand une variable y est liée à une variable auxiliaire quantitative unique x, connue pour tous les éléments d'une population, un modèle peut servir à guider la formation de strates. Ce raisonnement, souvent utilisé dans les enquêtes auprès des entreprises ou des établissements institutionnels, est appelé stratification par taille. Une littérature abondante traite de la formation des strates; voir, par exemple, Rivest (2002) et sa liste de références. La manière habituelle consiste à trier la base de sondage par x, à diviser la population en strates et à déterminer la répartition optimale d'un échantillon aléatoire simple stratifié sans remise aux strates. Dans ce scénario, l'objectif est de trouver des limites de strate, $(b_{h-1}, b_h]$ pour $h = 1, \ldots, H-1$, qui mènent à la variance par rapport au plan de l'estimateur π d'une moyenne ou d'un total qui est minimisé ou à n qui est minimisé selon un CV cible.

4.1 Analyses assistées par un modèle

Lavallée et Hidiroglou (1988) ont présenté des algorithmes itératifs pour trouver les limites de strates qui réduiront le plus possible la taille totale de l'échantillon assujettie à une contrainte sur le CV de $\bar{y}_{st}$, en supposant que la répartition aux strates est une répartition de puissance (Bankier, 1988). Dans une répartition de puissance, la proportion de la taille totale de l'échantillon attribuée aux strates h est proportionnelle à $\left(W_h \bar{y}_{Uh}\right)^p$ pour $p \in (0,1]$.

Rivest (2002) a étendu l'algorithme de Lavallée et Hidiroglou (1988) quand i) $\log(y) = \alpha + \beta_{\log} x + \varepsilon$, où ε est normalement réparti selon une moyenne de 0 et une variance constante qui ne dépend pas de x, ou ii) $y = \beta_{\text{lin}} x + \varepsilon$, ε ayant une moyenne de 0 et la variance $\sigma^2_{\text{lin}} x^{\gamma}$ où γ est non négatif. Rivest a donné des résultats pour la répartition de puissance et la répartition de Neyman. Rivest attribue la strate H pour qu'elle soit une strate à tirage complet, une procédure souvent utilisée dans les enquêtes auprès des entreprises pour les grandes unités, de sorte que $n_H = N_H$. Pour $h < H$, la taille de l'échantillon dans la strate h peut s'écrire $\left(n - n_H\right) a_h$, où n est la taille totale de l'échantillon, $a_h = \left(W_h \bar{y}_{Uh}\right)^p \Big/ \sum_{h=1}^{H-1} \left(W_h \bar{y}_{Uh}\right)^p$ pour une répartition de puissance et $a_h = W_h S_{yUh} \Big/ \sum_{h=1}^{H-1} W_h S_{yUh}$ pour la répartition de Neyman. La résolution de $V_\pi\left(\bar{y}_{st}\right)$ pour n et l'utilisation de variances conditionnelles à l'appartenance à la strate donnent

$$n = NW_H + \frac{\sum_{h=1}^{H-1} W_h^2 V_M\left(y \mid b_h \geq x \geq b_{h-1}\right) \Big/ a_{hX}}{\bar{y}_U^2 c^2 + \sum_{h=1}^{H-1} W_h^2 V_M\left(y \mid b_h \geq x \geq b_{h-1}\right) \Big/ N}, \tag{4.1}$$

où V_M désigne la variance du modèle, $\bar{y}_U$ est la moyenne de la population de y, c^2 est la variance relative cible pour $\bar{y}_{st}$ et a_{hX} est écrit au moyen de la relation du modèle entre y et x. L'expression (4.1) donne des équations différentielles pour $\partial n / \partial b_h$ et $\partial n / \partial b_{h-1}$, qui sont résolues de façon itérative au moyen d'un algorithme de Sethi (1963). Une limite pratique des algorithmes itératifs de Lavallée-Hidiroglou (LH) et de Rivest est qu'ils peuvent converger vers des limites qui ne donnent pas le vrai minimum de n ou peuvent ne pas converger du tout pour certaines configurations de y (Slanta et Krenzke, 1994, 1996; Rivest, 2002).

Gunning et Horgan (2004) et Horgan (2006) ont présenté un autre algorithme pour trouver les limites de la strate, $(b_{h-1}, b_h]$, basé sur une seule mesure asymétrique de la taille x. Leur solution consistait à calculer les limites de la strate de la façon $b_h = b_0 r^h \quad (h = 1, \ldots, H)$, où $r = \left(b_H / b_0\right)^{1/H}$ et b_H et b_0 sont les valeurs maximale et minimale de x. Autrement dit, les limites suivent une progression géométrique. Si la répartition de x est uniforme dans chaque strate, cet ensemble de limites est approximativement égal aux CV de strate de x définis comme étant $S_{xh}/\overline{x}_h$, où S_{xh} est l'écart-type des unités de la base de sondage pour la strate h et $\overline{x}_h$ est la moyenne de la base de sondage pour les unités de la strate h. L'algorithme n'est pas motivé par un modèle, mais il faisait concurrence à l'algorithme de LH et à la méthode $\text{cum}(\sqrt{f})$ de Dalenius et Hodges (1959), et ses calculs sont plus faciles à mettre en œuvre.

Baillargeon et Rivest (2009) ont étendu la méthode de Rivest (2002) pour tenir compte des taux de non-réponse propres à la strate et permettre les strates à tirage nul, à tirage partiel et à tirage complet. L'exclusion de certaines unités de l'échantillon au moyen d'une strate à tirage nul peut être raisonnable quand certaines unités ont des valeurs y proches de zéro ou sont si petites comparativement aux grandes unités qu'elles contribuent peu à un total de population. (Un cas extrême serait celui de l'échantillonnage défini par un seuil d'inclusion présenté à la section 3.) Une strate à tirage nul peut être une façon de réduire la racine de l'erreur quadratique moyenne relative de $\overline{y}_{st}$, mais elle entraîne un biais. Les spécialistes répartissent l'échantillon aux strates au moyen d'une règle générale qui comprend une répartition proportionnelle, une répartition de Neyman et une répartition de puissance comme cas particuliers. Comme l'ont constaté Baillargeon et Rivest (2009), leur solution pour n exige une solution itérative pour laquelle ils utilisent un algorithme proposé par Kozak (2004). Quand la non-réponse est prise en compte dans la strate à tirage complet ou quand il y a une strate à tirage nul, une solution sans contrainte peut être négative. Par conséquent, on obtient une minimisation contrainte pour n au-delà des limites $\{b_h\}$ qui donnent une taille d'échantillon positive.

Le paquet de R appelé `stratification` (Baillargeon et Rivest, 2011) met en œuvre plusieurs méthodes de stratification, dont la méthode $\text{cum}\sqrt{f}$, la méthode géométrique et la méthode de LH. L'algorithme de LH qui trouve des limites de strate qui minimisent la taille totale de l'échantillon n tout en atteignant un CV cible peut être mis en œuvre au moyen des algorithmes de Sethi ou de Kozak. Au moyen de l'algorithme de Kozak, le paquet trouvera également des limites de strate qui minimisent le CV de $\overline{y}_{st}$ pour une taille d'échantillon fixe n plutôt que de minimiser n pour un CV prédéterminé.

4.2 Analyses purement fondées sur un modèle

Dorfman et Valliant (2000) ont étudié les propriétés fondées sur un modèle de la stratification par taille d'un point de vue purement fondé sur un modèle. Certains de leurs résultats sont synthétisés dans le présent article. Quand un modèle commun se vérifie pour l'ensemble de la population comme dans (2.1) et que $\mathbf{V}\mathbf{1}_N$ et $\mathbf{V}^{1/2}\mathbf{1}_N$ sont dans l'espace colonne de $\mathbf{X}$, le meilleur prédicteur linéaire sans biais avec un échantillon équilibré pondéré constituent la meilleure stratégie, comme l'explique la section 2. La stratification par taille est alors, au mieux, un mécanisme de sélection d'un échantillon équilibré pondéré.

Néanmoins, d'autres analyses fondées sur un modèle éclaireront la pertinence des différentes variations de la stratification par taille parfois utilisées en pratique.

Premièrement, envisageons une version stratifiée du modèle (2.1) dans laquelle les paramètres peuvent dépendre des strates :

$$E_M\left(\mathbf{Y}_h\right)=\mathbf{X}_h\boldsymbol{\beta}_h;\quad V_M\left(\mathbf{Y}_h\right)=\sigma_h^2\mathbf{V}_h;\quad h=1,\ldots,H, \tag{4.2}$$

où $\mathbf{Y}_h$ est $N_h\times 1$, $\mathbf{X}_h$ est $N_h\times p_h$, σ_h^2 est un scalaire positif, $\mathbf{V}_h=\text{diag}\left(v_{hi}\right)$ est $N_h\times N_h$ et $\boldsymbol{\beta}_h$ est un vecteur de paramètre $p_h\times 1$. Le meilleur prédicteur linéaire sans biais est alors la somme des meilleurs prédicteurs linéaires sans biais dans chaque strate.

Dans la strate h, définissons un échantillon équilibré pondéré qui satisfait

$$\frac{1}{n_h}\mathbf{1}_{sh}^T\mathbf{V}_{sh}^{-1/2}\mathbf{X}_{sh}\;=\;\frac{\mathbf{1}_{Nh}^T\mathbf{X}_h}{\mathbf{1}_{Nh}^T\mathbf{V}_h^{1/2}\mathbf{1}_{Nh}}, \tag{4.3}$$

où $\mathbf{1}_{sh}$ est un vecteur de taille n_h dont tous les éléments sont des 1, $\mathbf{1}_{Nh}$ est un vecteur de taille N_h dont tous les éléments sont des 1, $\mathbf{V}_{sh}$ est la matrice de covariance diagonale de dimension $n_h\times n_h$ pour les unités de l'échantillon et $\mathbf{X}_{sh}$ est la matrice de variables auxiliaires de dimensions $n_h\times p_h$ pour les unités de l'échantillon. Tout échantillon de strate satisfaisant (4.3) sera désigné par $B\left(\mathbf{X}_h:\mathbf{V}_h\right)$, et quand l'expression (4.3) est satisfaite dans chaque strate, l'échantillon entier est un échantillon équilibré pondéré stratifié.

Si $\mathbf{V}_h\mathbf{1}_{Nh}$ et $\mathbf{V}_h^{1/2}\mathbf{1}_{Nh}$ sont tous deux dans l'espace colonne de $\mathbf{X}_h$, alors le meilleur prédicteur linéaire sans biais atteint sa variance minimale quand chaque échantillon de strate est $B\left(\mathbf{X}_h:\mathbf{V}_h\right)$. Dans ce cas, le meilleur prédicteur linéaire sans biais se réduit à

$$\hat{t}_{\text{BLU}}\;=\;\sum_{h=1}^{H}N_h\overline{v}_h^{(1/2)}\frac{1}{n_h}\sum_{i\in s_h}\frac{y_{hi}}{v_{hi}^{(1/2)}} \tag{4.4}$$

et la variance de l'erreur est

$$V_M\left(\hat{t}_{\text{BLU}}-t_U\right)\;=\;\sum_h\left[\frac{1}{n_h}\left(N_h\overline{v}_h^{(1/2)}\right)^2-N_h\overline{v}_h\right]\sigma_h^2, \tag{4.5}$$

où $\overline{v}_h^{(1/2)}=N_h^{-1}\sum_{U_h}v_{hi}^{1/2}$ et $\overline{v}_h=N_h^{-1}\sum_{U_h}v_{hi}$.

Ainsi, dans un échantillon équilibré pondéré stratifié, l'estimateur optimal se réduit à la somme d'estimateurs de la moyenne des ratios. Comme dans le cas non stratifié, un échantillon équilibré pondéré est la meilleure sélection possible en ce sens qu'il permet de réduire la variance de l'erreur du meilleur prédicteur linéaire sans biais. L'estimateur du total, $\hat{t}_{\text{BLU}}$, est aussi l'estimateur π quand chaque échantillon au sein de la strate est sélectionné au moyen de probabilités proportionnelles à $v_{hi}^{1/2}$. Bien qu'un échantillon probabiliste sélectionné au moyen de probabilités proportionnelles à $v_{hi}^{1/2}$ soit équilibré en espérance, le résultat fondé sur un modèle n'exige pas que l'échantillon équilibré soit obtenu par échantillonnage

probabiliste. Toutefois, si l'on souhaite avoir un échantillon probabiliste, on peut utiliser les méthodes de Deville et Tillé (2004).

La répartition optimale limitée par le coût entre les strates peut être calculée au moyen de méthodes normalisées. Supposons que la fonction de coût est $C = C_0 + \sum_h c_h n_h$, où C est le coût total, C_0 est le coût qui ne varie pas selon la taille de l'échantillon et c_h est le coût par unité dans la strate h. Supposons que $\mathbf{V}_h \mathbf{1}_{Nh}$ et $\mathbf{V}_h^{1/2} \mathbf{1}_{Nh}$ sont dans l'espace colonne de $\mathbf{X}_h$ et qu'un échantillon équilibré pondéré, $B(\mathbf{X}_h : \mathbf{V}_h)$, est sélectionné dans chaque strate. Dans ce cas, la répartition qui minimise la variance de $\hat{t}_{\text{BLU}}$ assujettie à un coût total fixe est

$$\frac{n_h}{n} = \frac{N_h \overline{v}_h^{(1/2)} \sigma_h \Big/ \sqrt{c_h}}{\sum_{h'} N_{h'} \overline{v}_{h'}^{(1/2)} \sigma_{h'} \Big/ \sqrt{c_{h'}}}. \tag{4.6}$$

Si tous les coûts sont égaux, le meilleur prédicteur linéaire sans biais ayant la répartition optimale et équilibrée de l'échantillon est alors égal à

$$\hat{t}_{\text{BLU}} = \frac{1}{n} \left(\sum_h N_h \overline{v}_h^{(1/2)} \sigma_h \right) \sum_h \sum_{s_h} \frac{y_{hi}}{v_h^{1/2} \sigma_h}. \tag{4.7}$$

et sa variance d'erreur est

$$V_M \left(\hat{t}_{\text{BLU}} - t_U \right) = \frac{1}{n} \left(\sum_h N_h \overline{v}_h^{(1/2)} \sigma_h \right)^2 - \sum_h N_h \overline{v}_h \sigma_h^2. \tag{4.8}$$

Pour déterminer la façon de former des strates, prenons le cas d'un modèle unique qui ajuste toute la population, c'est-à-dire le cas particulier de (2.1) et de (4.2) défini par

$$E_M \left(\mathbf{Y}_h \right) = \mathbf{X}_h \boldsymbol{\beta}, V_M \left(\mathbf{Y}_h \right) = \mathbf{V}_h \sigma^2. \tag{4.9}$$

Supposons que nous choisissons un échantillon équilibré pondéré stratifié et que nous utilisons la répartition optimale donnée dans (4.6) pour le cas à coût égal. Au moyen de (4.7) ainsi que de $\sigma_h = \sigma$ et de $\overline{v}^{(1/2)} = N^{-1} \sum_h \sum_{U_h} v_{hi}^{1/2}$, le meilleur prédicteur linéaire sans biais ayant la répartition optimale est

$$\hat{t}_{\text{BLU}} = \frac{1}{n} \left(\sum_h N_h \overline{v}_h^{(1/2)} \right) \sum_h \sum_{s_h} \frac{y_{hi}}{v_{hi}^{1/2}} = \frac{1}{n} \left(N \overline{v}^{1/2} \right) \sum_h \sum_{s_h} \frac{y_{hi}}{v_{hi}^{1/2}},$$

qui est la forme du meilleur prédicteur linéaire sans biais dans (2.4) dans un échantillon équilibré pondéré pour un échantillon non stratifié. En d'autres termes, la stratification avec la répartition optimale d'un échantillon équilibré pondéré stratifié n'apporte dans ce cas-ci aucun avantage comparativement à la stratégie de sélection d'un échantillon non stratifié avec un équilibre pondéré global.

Un cas particulier important d'un modèle de population unique se présente quand il y a un seul x et que le modèle est polynomial :

$$E_M \left(y_{hi} \right) = \delta_0 \beta_0 + \delta_1 \beta_1 x_{hi} + \cdots + \delta_J \beta_J x_{hi}^J, V_M \left(y_{hi} \right) = \sigma^2 x_{hi}^{\gamma},$$

où $\delta_j = 1$ si le j^{e} terme d'ordre est dans le modèle et 0 sinon, et les β_j sont des paramètres de régression. Parmi les modèles de cette classe, on trouve celui de l'estimateur par le ratio : $E_M\left(y_{hi}\right) = \beta_1 x_{hi},\ V_M\left(y_{hi}\right) = \sigma^2 x_{hi}$. Au moyen de la spécification de la variance, $V_M\left(y_{hi}\right) = \sigma^2 x_{hi}^{\gamma},$ on obtient la répartition optimale limitée par le coût en effectuant une spécialisation de (4.6) :

$$\frac{n_h}{n} = \frac{N_h \bar{x}_h^{(\gamma/2)} \Big/ \sqrt{c_h}}{\sum_{h'} N_{h'} \bar{x}_{h'}^{(\gamma/2)} \Big/ \sqrt{c_{h'}}}.$$

où $\bar{x}_h^{(\gamma/2)} = N_h^{-1} \sum_{U_h} x_{hi}^{\gamma/2}$. La variance de l'erreur selon cette répartition est

$$V_M\left(\hat{t}_{\text{BLU}} - t_U\right) = \sigma^2 \sum_h \left[\frac{1}{n_h} \left(N_h \bar{x}_h^{(\gamma/2)} \right)^2 - N_h \bar{x}_h^{(\gamma)} \right]. \tag{4.10}$$

Il est plus pratique d'étudier le problème de la création de strates quand un nombre égal d'unités d'échantillonnage, $n_h \equiv n_0,$ est attribué à chaque strate. Dans ce cas,

$$V_M\left(\hat{t}_{\text{BLU}} - t_U\right) = \frac{\sigma^2}{n_0} \sum_h \left(N_h \bar{x}_h^{(\gamma/2)} \right)^2 - N \bar{x}^{(\gamma)} \sigma^2 \tag{4.11}$$

et $\bar{x}^{(\gamma)} = N^{-1} \sum_h \sum_{U_h} x_{hi}^{\gamma}$.

Supposons que $Z_h = N_h \bar{x}_h^{(\gamma/2)}$. La stratification optimale se produit quand le terme principal dans (4.11), $\sum_h Z_h^2 = \sum_h \left(N_h \bar{x}_h^{(\gamma/2)} \right)^2,$ est minimisé. Le fait d'additionner et de soustraire $\sigma^2 H \bar{Z}^2 / n_0,$ où $\bar{Z} = \sum_{h=1}^{H} Z_h / H,$ donne

$$V_M\left(\hat{t}_{\text{BLU}} - t_U\right) = \frac{\sigma^2}{n_0} S_Z^2 + \frac{\sigma^2}{n_0} \frac{\left(N \bar{x}^{(\gamma/2)} \right)^2}{H} - \sigma^2 N \bar{x}^{(\gamma)}, \tag{4.12}$$

où $S_Z^2 = \sum_h \left(Z_h - \bar{Z} \right)^2$. Le seul terme dans (4.12) qui dépend de la formation des strates est le premier, que l'on élimine en rendant tous les Z_h égaux. L'expression (4.12) devient alors $V_M\left(\hat{t}_{\text{BLU}} - t_U\right) = \frac{\sigma^2}{n} \left(N \bar{x}^{(\gamma/2)} \right)^2 - \sigma^2 N \bar{x}^{(\gamma)}$.

L'égalisation de $Z_h = N_h \bar{x}_h^{(\gamma/2)}$ mène à plusieurs règles de « taille d'agrégat égale » pour la formation de strates trouvées dans la littérature, par exemple Cochran (1977, page 172), Godfrey, Roshwalb et Wright (1984) et Hansen et coll. (1953a, page 382). Quand $\gamma = 0,$ les valeurs égales de $N_h \bar{x}_h^{\gamma/2}$ correspondent à des nombres égaux d'unités N_h dans chaque strate. Quand $\gamma = 1,$ nous avons une racine carrée de taille d'agrégat égale, et $\gamma = 2$ donne un agrégat égal x. Ainsi, les analyses fondées sur un modèle précisent dans quels cas les différentes méthodes de stratification par taille seront efficaces.

Les résultats des sous-sections 4.1 et 4.2 renvoient à un seul y et à un nombre prédéterminé de strates $H,$ mais des extensions sans ces restrictions ont été effectuées. Pour des raisons pratiques, on utilise habituellement un seul ensemble de strates, en sachant qu'il ne sera pas aussi efficace pour toutes les estimations. Par conséquent, il faut un ensemble de strates de compromis et une répartition qui fonctionnent relativement bien pour différentes estimations tout en respectant le budget, la charge de travail et d'autres contraintes du plan. La programmation mathématique, abordée à la section 6, est particulièrement utile à cet

égard. Benedetti, Espa et Lafratta (2008) et Ballin et Barcaroli (2013) ont abordé le problème du plan pour les enquêtes polyvalentes, c'est-à-dire celles où il y a plusieurs y, au moyen d'algorithmes des arbres de décision et génétiques. Leurs solutions définissent un ensemble optimal de strates à partir du croisement d'un ensemble de covariables catégoriques et d'une répartition à ces strates. Le nombre total de strates H est un sous-produit de leurs solutions. Les algorithmes de Ballin et Barcaroli (2013) sont mis en œuvre dans le paquet de R appelé `SamplingStrata` (Barcaroli, Ballin, Odendaal, Pagliuca, Willighagen et Zardetto, 2022).

5. Échantillonnage à plusieurs degrés et variances anticipées

Pendant des décennies, l'échantillonnage à plusieurs degrés a été un outil standard dans les enquêtes auprès des ménages qui exigent la collecte de données en personne. Une séquence emboîtée de régions géographiques est sélectionnée jusqu'à ce qu'à la dernière étape, les ménages ou les personnes faisant partie d'un ménage soient échantillonnés. L'échantillonnage à plusieurs degrés est courant aussi dans les enquêtes sur l'éducation, où les écoles et les élèves des écoles sont les degrés, et dans les enquêtes auprès des entreprises, où les établissements et les employés sont échantillonnés. Pour concevoir ces enquêtes, il faut des estimations des composantes de la variance. Les variances anticipées peuvent être utiles pour éviter le problème des estimations de la variance négatives, comme cela est décrit ci-dessous.

5.1 Plans à deux degrés

Prenons le cas d'un échantillon à deux degrés dans lequel les UPE sont sélectionnées avec des probabilités variables et avec remise (échantillonnage avec probabilité proportionnelle à la taille avec remise), tandis que les éléments du deuxième degré sont sélectionnés par échantillonnage aléatoire simple sans remise. Les plans avec remise ne sont peut-être pas souvent utilisés en pratique, mais ils comportent des formules de variance simples qui facilitent le calcul de la taille de l'échantillon. Soit y_k la valeur d'une variable d'analyse pour l'élément k, m le nombre d'UPE de l'échantillon, M le nombre d'UPE dans la population, s l'ensemble d'UPE de l'échantillon, N_i le nombre d'éléments dans la population pour l'UPE i, n_i le nombre d'éléments de l'échantillon dans l'UPE de l'échantillon i et s_i l'ensemble d'éléments de l'échantillon dans l'UPE i. L'estimateur pwr (probabilité avec remise) du total des y est

$$\hat{t}_{\text{pwr}} = \frac{1}{m}\sum_{i\in s}\frac{\hat{t}_i}{p_i},$$

où $\hat{t}_i = \frac{N_i}{n_i}\sum_{k\in s_i} y_k$ est le total estimé pour l'UPE i à partir d'un échantillon aléatoire simple et p_i est la probabilité de sélection à tirage unique de l'UPE i. La variance par rapport au plan de $\hat{t}_{\text{pwr}}$ tirée de Cochran (1977, pages 308 à 310) est :

$$V_\pi\left(\hat{t}_{\text{pwr}}\right) = \frac{1}{m}\sum_{i\in U} p_i\left(\frac{t_i}{p_i} - t_U\right)^2 + \sum_{i\in U}\frac{N_i^2}{mp_in_i}\left(1-\frac{n_i}{N_i}\right)S_{U2i}^2, \tag{5.1}$$

où t_U est le total de la population de y et $S_{U2i}^2 = \left(N_i - 1\right)^{-1} \sum_{k \in s_i} \left(y_k - \bar{y}_{Ui}\right)^2$, $\bar{y}_{Ui}$ étant la moyenne de la population de y dans l'UPE i.

Il est difficile de calculer la taille des échantillons quand les n_i sont autorisés à varier, mais pour maîtriser les charges de travail, les échantillons sont souvent conçus pour sélectionner le même nombre d'éléments dans chaque UPE. En supposant que $\bar{n}$ éléments sont sélectionnés dans chaque UPE et que la fraction de sondage au sein de l'UPE, $\bar{n}/N_i$, est négligeable, la variance relative par rapport au plan de $\hat{t}_{\text{pwr}}$, définie comme étant $V_\pi\left(\hat{t}_{\text{pwr}}\right)/t_U^2$, est approximativement (Valliant, Dever et Kreuter, 2018, section 9.2.4) :

$$\frac{V_\pi\left(\hat{t}_{\text{pwr}}\right)}{t_U^2} \doteq \frac{B^2}{m} + \frac{W^2}{m\bar{n}} = \frac{\tilde{V}}{m\bar{n}}\, k\left[1 + \delta\left(\bar{n} - 1\right)\right], \tag{5.2}$$

où

$$B^2 = S_{U1(\text{pwr})}^2 / t_U^2,\ S_{U1(\text{pwr})}^2 = \sum_{i \in U} p_i \left(\frac{t_i}{p_i} - t_U\right)^2,$$

t_i est le total de la population de y pour l'UPE i,

$$W^2 = t_U^{-2} \sum_{i \in U} N_i^2 \frac{S_{U2i}^2}{p_i},\ k = (B^2 + W^2)/\tilde{V}$$

et $\delta = B^2 / \left(B^2 + W^2\right)$ est une mesure d'homogénéité. La variance relative de l'unité (c'est-à-dire la population) est $\tilde{V} = S_U^2 / \bar{y}_U^2$, S_U^2 étant la variance de la population de y et $\bar{y}_U$, la moyenne de la population de y.

L'estimateur des composantes de la variance fondé sur le plan de B^2, que l'on trouve par exemple dans Särndal et coll. (1992), peut être négatif, selon la configuration des données. L'utilisation des variances anticipées permet d'écrire la variance relative de l'estimateur pwr en termes de composantes de la variance du modèle. On peut estimer les composantes du modèle au moyen d'algorithmes qui évitent les problèmes numériques que posent les formules d'analyse de la variance fondées sur le plan de base. La littérature semble donner peu d'exemples de l'utilisation des estimations des composantes de la variance fondées sur un modèle dans le plan d'enquête, bien que les spécialistes utilisent souvent cette technique. Chromy et Myers (2001), Hunter, Bowman et Chromy (2005), Judkins et Van de Kerckhove (2003), Valliant et Gentle (1997) et Waksberg, Sperry, Judkins et Smith (1993) en sont quelques exemples. Searle, Casella et McCulloch (1992) ont examiné les méthodes disponibles, notamment l'estimation sans biais quadratique de la variance minimale (MIVQUE0 pour *minimum variance quadratic unbiased estimation*), le maximum de vraisemblance et le maximum de vraisemblance restreint (REML pour *restricted maximum likelihood*). L'utilisation des variances anticipées clarifie aussi le rôle essentiel, souligné ci-dessous, que jouent les tailles des UPE et des unités secondaires d'échantillonnage (USE) dans la détermination des mesures de l'homogénéité.

Comme l'indique la section 2, quand l'estimateur est sans biais par rapport au plan ou approximativement sans biais, c'est-à-dire que $E_\pi(\hat{t}_{\text{pwr}}) \doteq t_U$, la variance anticipée est $\text{VA}\left(\hat{t}_{\text{pwr}}\right) = E_M\left[V_\pi\left(\hat{t}_{\text{pwr}} - t_U\right)\right]$. Par conséquent, on peut calculer l'espérance du modèle E_M d'une formule comme (5.2), ce qui donne des

composantes de la variance du modèle estimables au moyen d'un logiciel standard. Dans une population en grappes, le modèle le plus simple à envisager est un modèle ayant une moyenne commune, μ, et des effets aléatoires pour les grappes, α_i, et les éléments, ε_{ij} :

$$y_k = \mu + \alpha_i + \varepsilon_{ik},\ k \in U_i, \tag{5.3}$$

où $\alpha_i \sim (0, \sigma_\alpha^2)$, $\varepsilon_{ik} \sim (0, \sigma_\varepsilon^2)$, et les erreurs sont indépendantes. Ce modèle est bien trop simple, mais les résultats peuvent être étendus à un cas où $E_M(y_k) = \mathbf{x}_k^T\boldsymbol{\beta}$.

Le cas des grappes d'échantillonnage avec probabilité proportionnelle à la taille N_i revêt une importance pratique particulière, c'est-à-dire que $p_i = N_i/(M\bar{N})$, où $\bar{N} = \sum_{i\in U} N_i/M$ est le nombre moyen d'éléments par grappe. Dans ce cas particulier, après quelques calculs, les espérances du modèle de B^2 et de W^2 sont

$$\begin{aligned} E_M\left(B^2\right) &= \frac{1}{\mu^2}\left\{\sigma_\alpha^2\left[1 - \frac{1}{M^2}\left(2 - \frac{1}{\bar{N}}\right)\left(\nu_N^2 + 1\right)\right] + \frac{\sigma_\varepsilon^2}{\bar{N}}\right\} \\ E_M\left(W^2\right) &= \sigma_\varepsilon^2/\mu^2, \end{aligned}$$

$\nu_N^2 = \sum_{i\in U}\left(N_i - \bar{N}\right)^2 \Big/ \left[\left(M-1\right)\bar{N}^2\right]$ est la variance relative des tailles de grappe. Quand M est grand, la mesure anticipée de l'homogénéité est approximativement

$$E_M(\delta) \doteq \frac{\sigma_\alpha^2 + \sigma_\varepsilon^2/\bar{N}}{\sigma_\alpha^2 + \sigma_\varepsilon^2\left(1 + 1/\bar{N}\right)} \doteq \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}. \tag{5.4}$$

L'expression (5.4) est la corrélation dans le modèle (5.3) de deux éléments de la même grappe. Il convient de constater l'absence d'hypothèse selon laquelle toutes les UPE ont la même taille $\left(N_i \equiv \bar{N}\right)$ pour obtenir (5.4). Tant que M est grand, ν_N^2 a un effet limité sur B^2 et δ. Cela diffère du résultat obtenu quand les deux degrés sont sélectionnés par échantillonnage aléatoire simple, où la variation des tailles d'UPE joue un rôle important dans la détermination de δ même quand M est de grande taille (voir Valliant et coll., 2018, équation (9.43); Valliant, Dever et Kreuter, 2015).

5.2 Plans à trois degrés

Maze (2021) a étendu l'analyse ci-dessus à l'échantillonnage à trois degrés, où les USE et les éléments de troisième degré sont stratifiés. Considérons un plan à trois degrés dans lequel les degrés sont les UPE, les USE et les unités de logement (UL). L'utilisation d'UL comme unités de troisième degré n'est qu'à titre d'exemple. La formulation ci-dessous s'applique aussi à d'autres utilisations. Supposons que m UPE sont sélectionnées au moyen d'un échantillonnage avec probabilité proportionnelle à la taille avec remise, que les USE sont stratifiées au sein de chaque UPE et que les n_{ia} sont sélectionnés au moyen d'un échantillonnage avec probabilité proportionnelle à la taille avec remise au sein de l'UPE i, la strate d'USE a. Les unités de logement sont stratifiées au sein de chaque USE et les q_{iajb} sont sélectionnés par échantillonnage aléatoire simple sans remise à partir du total de la population de Q_{iajb} unités de logement au sein de l'UPE i, de l'USE j dans la strate d'USE $a\,(a = 1, \ldots, A)$ et de la sous-strate d'UL

$b (b = 1, \ldots, B)$. Supposons que les strates d'USE sont définies de la même façon dans chaque UPE et que les définitions des strates d'UL sont les mêmes dans chaque UPE et USE.

Les strates d'USE peuvent être définies en fonction du pourcentage de la population d'USE dans un domaine (par exemple les Hispaniques) important pour l'enquête. Les strates d'UL pourraient être définies par le groupe ethnique ou racial du chef de ménage. Par exemple, la Health and Retirement Study (HRS ou étude sur la santé et la retraite en français; https://hrs.isr.umich.edu/about) de l'Université du Michigan est une étude longitudinale par panel portant sur les personnes de 50 ans et plus, recevant l'appui de la National Institute on Aging et de la Social Security Administration des États-Unis. Ses UPE sont des comtés ou des groupes de comtés et ses USE, des îlots de recensement ou des groupes d'îlots. Les USE sont stratifiées selon la concentration d'Afro-Américains et d'Hispaniques. Périodiquement, pour la HRS, on recrute une nouvelle cohorte de personnes dont l'âge les rend admissibles. En 2016, la cohorte « de la fin du baby-boom » (années de naissance de 1960 à 1965) a été recrutée à l'aide d'UL stratifiées selon la race ou l'origine ethnique du chef de ménage, conformément au codage d'une liste commerciale de logements (Valliant, Hubbard, Lee et Chang, 2014). La disponibilité de listes commerciales aux États-Unis et de panels de ménages comme le panel AmeriSpeak du NORC à l'Université de Chicago

(voir https://amerispeak.norc.org/us/en/amerispeak/about-amerispeak/panel-design.html)

rend ces échantillons ciblés possibles pour les organisations non gouvernementales. Dans d'autres pays, les registres de la population permettent aux organismes gouvernementaux de mettre en œuvre des plans similaires.

L'estimateur pwr d'un total de population, t_U, de y est

$$\hat{t}_{\text{pwr}} = \frac{1}{m} \sum_{i \in s_1} \frac{1}{p_i} \sum_{a=1}^{A} \frac{1}{n_{ia}} \sum_{j \in s_{ia}} \frac{1}{p_{j|ia}} \sum_{b=1}^{B} \frac{Q_{iajb}}{q_{iajb}} \sum_{k \in s_{iajb}} y_k, \tag{5.5}$$

où p_i est la probabilité à 1 tirage de sélectionner l'UPE i; s_1 est l'ensemble d'UPE de l'échantillon; $p_{j|ia}$ est la probabilité à 1 tirage conditionnelle de sélectionner l'USE j dans l'UPE i, la strate d'USE a; s_{ia} est l'ensemble d'USE de l'échantillon dans l'UPE i, la strate d'USE a; s_{iajb} est l'ensemble d'UL de l'échantillon dans l'UPE i, l'USE j dans la strate d'USE a et la strate d'UL b.

Pour simplifier les calculs de la variance fondée sur le plan et de la taille de l'échantillon, une solution de rechange standard consiste à supposer que le même nombre d'USE, $\overline{n}_a$, est sélectionné dans chaque UPE et strate d'USE et que le même nombre d'UL, $\overline{\overline{q}}_{ab}$, est échantillonné dans chaque combinaison $iajb$. Ensuite, définissons U comme étant l'univers des UPE, U_{ia} comme étant l'univers des USE dans ia, U_{iajb} comme étant l'univers des UL dans $iajb$, $K_a = t_{Ua} / t_U$ comme étant la proportion du total de la population de y qui est dans la strate d'USE a et $K_{ab} = t_{Uab} / t_U$ comme étant la proportion dans la combinaison de strates ab. Après des calculs, la variance relative par rapport au plan de l'estimateur peut s'écrire comme la somme de trois termes, qui sont semblables à ceux de Hansen, Hurwitz et Madow (1953b, chapitre 9) :

$$\frac{V_\pi\left(\hat{t}_{\text{pwr}}\right)}{t_U^2} = \frac{B^2}{m} + \sum_{a=1}^{A} K_a^2 \frac{W_{2a}^2}{m\overline{n}_a} + \sum_{a=1}^{A} \sum_{b=1}^{B} K_{ab}^2 \frac{W_{3ab}^2}{m\overline{n}_a \, \overline{\overline{q}}_{ab}} \tag{5.6}$$

où $B^2 = \frac{S^2_{U1(\text{pwr})}}{t_U^2}$,

$$W^2_{2a} = \frac{1}{t^2_{U_a}} \sum_{i \in U} \frac{S^2_{U2(\text{pwr})ia}}{p_i},\ W^2_{3ab} = \frac{1}{t^2_{U_{ab}}} \sum_{i \in U} \frac{1}{p_i} \sum_{j \in U_{ia}} \frac{Q^2_{iajb} S^2_{U3iajb}}{p_{j|ia}}$$

selon

$$S^2_{U1(\text{pwr})} = \sum_{i \in U} p_i \left(\frac{t_{U_i}}{p_i} - t_U \right)^2,\ S^2_{U2(\text{pwr})ia} = \sum_{j \in U_{ia}} p_{j|ia} \left(\frac{t_{U_{iaj}}}{p_{j|ia}} - t_{U_{ia}} \right)^2$$

et

$$S^2_{U3iajb} = \frac{1}{Q_{iajb} - 1} \sum_{k \in U_{iajb}} \left(y_k - \overline{y}_{U_{iajb}} \right)^2.$$

Les totaux de population, t_{Ui} et t_{Uiaj}, sont pour les y dans l'UPE i et l'USE iaj; $\overline{y}_{U_{iajb}}$ est la moyenne par élément dans $iajb$. Cela suppose que la fraction de sondage au troisième degré est négligeable. Les renseignements du calcul se trouvent dans Maze (2021, section 2.3). Le CV de $\hat{t}_{\text{pwr}}$ est

$$\text{CV}_\pi \left(\hat{t}_{\text{pwr}} \right) = \sqrt{V_\pi \left(\hat{t}_{\text{pwr}} \right) / t_U^2}.$$

Pour une fonction non linéaire différentiable, $\hat{\theta}_{\text{pwr}}$, comme une moyenne de ratio, on peut faire une approximation linéaire de $\hat{\theta}_{\text{pwr}}$ et dériver une formule analogue à (5.6). Une des complications non abordées dans le présent article est que certaines UPE peuvent être des certitudes (c'est-à-dire qu'elles sont sélectionnées avec une probabilité 1). La variance relative dans (5.6) est ensuite divisée entre les certitudes et les non-certitudes, et aucune composante de variance de l'UPE n'est utilisée pour les certitudes. Les UPE de non-certitude sont aussi habituellement stratifiées selon la région géographique ou d'autres caractéristiques. L'extension de ces autres complications est simple.

Une limitation bien connue de (5.6) est que les estimations de B^2 et de W^2_{2a} supposent des soustractions, de sorte que les estimations puissent être négatives dans certains échantillons. Comme pour l'échantillonnage à deux degrés, les variances anticipées peuvent servir à contourner ce problème. Les routines existantes pour l'estimation des composantes de la variance du modèle peuvent ensuite servir à contraindre les estimations des paramètres à être positives.

Dans le cas à trois degrés, un modèle simple pour y_k a une moyenne commune, μ, et des effets aléatoires α_i pour les UPE, γ_{iaj} pour les USE et ε_{iajbk} pour les UL dans la sous-strate d'USE ou d'UL ab:

$$y_k = \mu + \alpha_i + \gamma_{iaj} + \varepsilon_{iajbk},$$

où

$$\alpha_i \sim \left(0, \sigma^2_\alpha\right),\ \gamma_{iaj} \sim \left(0, \sigma^2_\gamma\right),\ \varepsilon_{iajbk} \sim \left(0, \sigma^2_{\varepsilon_{ab}}\right)$$

et les erreurs sont indépendantes, de sorte que $V_M \left(y_k \right) = \sigma^2_\alpha + \sigma^2_\gamma + \sigma^2_{\varepsilon_{ab}}$ et $E_M \left(y_k \right) = \mu$ pour $k \in U_{iajb}$. Il est possible d'étendre les modèles où $E_M(y_k) = \mathbf{x}_k^T \boldsymbol{\beta}$.

6. Solutions de programmation mathématique pour les répartitions d'échantillons

La plupart des enquêtes nationales auprès des ménages, des établissements et des établissements institutionnels ont plusieurs objectifs. Souvent, on recherche des estimations distinctes selon les groupes démographiques ou les types d'entreprises. Il est possible de les mettre en œuvre comme tailles d'échantillon cibles pour les sous-groupes ou comme CV cibles pour les estimations. De plus, il y a habituellement des contraintes concernant le budget, l'attribution de la charge de travail aux responsables de la collecte de données, le nombre maximal de tentatives de communication avec une unité d'échantillonnage pour obtenir la participation, entre autres choses. Une façon de déterminer une répartition de l'échantillon aux strates, aux UPE, aux USE et à d'autres degrés de l'échantillonnage consiste à procéder par essais et erreurs. En essayant suffisamment de combinaisons, un concepteur peut trouver par itération une répartition qui répond à la plupart des objectifs. La programmation mathématique (PM) est un moyen plus officiel et plus exact de trouver une répartition, qui peut être appliqué à des calculs fondés sur un modèle ou sur le plan. La PM est une technique extrêmement utile pour trouver des répartitions dans des problèmes complexes où il est impossible d'obtenir une solution sous forme fermée directe.

Des méthodes permettant de trouver des solutions approximatives ont été élaborées dans le domaine de la recherche opérationnelle. Bien que la PM soit décrite dans la littérature aux fins de répartition de l'échantillon, elle semble sous-utilisée en pratique. Certaines des méthodes de formation de strates de la section 4 peuvent être considérées comme des algorithmes de PM. Bethel (1989) a proposé un algorithme non linéaire spécialisé pour certains problèmes de répartition stratifiée. Ballin et Barcaroli (2008) ont étendu l'algorithme de Bethel aux tâches visant à créer des strates pour un échantillonnage aléatoire simple stratifié et à trouver une répartition efficace. Hughes et Rao (1979) ont abordé la répartition optimale aux strates à contraintes multiples. Chromy (1987) a présenté un algorithme spécialisé pour trouver des tailles d'échantillon qui réduisent au minimum les coûts en présence de plusieurs contraintes. Valliant et Gentle (1997) ont décrit la répartition dans un échantillon à deux degrés avec des variances anticipées lissées utilisées pour les composantes et avec des contraintes sur la taille des échantillons et les CV d'un ensemble d'estimateurs. Choudhry, Rao et Hidiroglou (2012) ont utilisé la programmation non linéaire dans un échantillonnage aléatoire simple stratifié pour résoudre des problèmes de répartition avec des contraintes sur les CV des estimateurs de strate et de domaine pour les domaines qui recoupent plusieurs strates. Plus récemment, de Moura Brito, Silva, Semaan et Maculan (2015) ont examiné les répartitions d'échantillons aléatoires simples stratifiés au moyen de la programmation par nombres entiers. Valliant et coll. (2018, chapitre 5) ont présenté une introduction à la programmation mathématique ainsi que plusieurs exemples de l'utilisation de la PM aux fins de répartition d'échantillons. Une application particulièrement complexe consiste à répartir les points de vente et les articles pour la détermination des prix dans l'indice des prix à la consommation des États-Unis (Gomes et Johnson, 2016; Leaver et Solk, 2005). La présente section donne deux exemples, l'un simple et l'autre plus complexe, de l'utilisation de la PM pour déterminer les répartitions à plusieurs critères.

La résolution de problèmes de PM nécessite un logiciel spécialisé qui met en œuvre les algorithmes sophistiqués élaborés en recherche opérationnelle. Une des lacunes des premiers articles sur la question était qu'ils n'étaient pas accompagnés de logiciels accessibles au public, ce qui n'est plus le cas. Schwendinger et Borchers (2023) donnent une longue liste de paquets de R dotés de fonctions d'optimisation. Les paquets de R `alabama` (Varadhan, 2023) et `nloptr` (Ypma, Johnson, Borchers, Eddelbuettel, Ripley, Hornik, Chiquet, Adler, Dai, Stamm et Ooms, 2022), en particulier, résolvent les types de problèmes non linéaires nécessaires à la répartition d'échantillon. Les procédures `NLP` et `OPTMODEL` dans SAS[MD] (http://www.sas.com) et le module complémentaire solveur (http://www.solver.com) fourni avec Microsoft Excel[MD] sont également utiles, surtout ce-dernier en raison de son interface conviviale.

Valliant et coll. (2014) proposent une application assez simple de la PM pour trouver la taille d'un échantillon au moyen des données de la National Survey of Family Growth (NSFG ou Enquête nationale sur la croissance des familles) des États-Unis de 2011-2012. L'application est présentée schématiquement dans la présente section; vous trouverez plus de renseignements dans l'article. Au moyen d'un échantillon probabiliste national de ménages, la NSFG permet de recueillir des renseignements sur la vie familiale, le mariage et le divorce, les grossesses, l'infertilité, l'utilisation de la contraception, et la santé des hommes et des femmes de 15 à 44 ans (Groves, Mosher, Lepkowski et Kirgis, 2009). L'objectif de cet exemple est d'obtenir des tailles d'échantillon cibles pour un groupe d'âge à partir d'une liste commerciale imparfaite d'adresses pour l'échantillonnage des ménages. Aux États-Unis, les organismes d'enquête peuvent acheter des listes d'adresses auprès de fournisseurs privés. En Europe et dans d'autres pays dotés de registres de la population, les organismes gouvernementaux peuvent avoir accès à de meilleures listes contenant de très nombreuses données auxiliaires.

L'une des approches afin d'obtenir une taille d'échantillon cible pour un groupe d'âge particulier (ou un autre groupe démographique) est de sélectionner un échantillon à probabilités égales d'UL, dresser la liste de toutes les personnes dans les UL et retenir la totalité ou un sous-ensemble de personnes faisant partie du groupe visé. Sans renseignement préalable sur les UL, cette méthode pourrait être la seule solution. L'inconvénient est qu'il peut être nécessaire de présélectionner de nombreuses UL, surtout quand le domaine cible est petit. Une autre tactique, moins coûteuse, consiste à utiliser une liste d'adresses qui élimine une partie de la présélection en repérant les UL susceptibles d'être dans le groupe souhaité. Même quand la liste n'est pas tout à fait exacte, cela peut être plus efficace qu'un échantillonnage à probabilités égales.

Dans le questionnaire de présélection de la NSFG, la liste de toutes les personnes est recueillie auprès de chaque UL répondante en plus de données démographiques limitées. En particulier, l'âge de chaque personne est obtenu et sert de variable de définition de domaine dans cet exemple. Bien que la fourchette d'âge cible de la NSFG soit de 15 à 44 ans, l'âge de chaque personne dans l'UL est obtenu pendant la présélection. Les adresses des répondants de la NSFG de 2011-2012 ont été envoyées à un fournisseur de listes commerciales à des fins d'appariement. Les données de présélection de la NSFG ont été considérées comme exactes et pouvaient ensuite être comparées aux données démographiques de la liste commerciale pour que soit évaluée l'exactitude de la liste concernant la classification des personnes par âge et la conception des futures versions de l'enquête.

Dans cet exemple, nous voulons obtenir une taille d'échantillon cible dans le groupe d'âge des 65 ans et plus en stratifiant les UL au moyen des données sur l'âge tirées de la liste commerciale. Pour formuler le problème, définissons la notation suivante :

d = domaine d'âge cible (65 ans et plus);

h = strate d'échantillonnage fondée sur les renseignements de la liste commerciale des adresses individuelles, $h = 1, \ldots, 4$ comme l'indique le tableau 6.1;

$p_h(d)$ = proportion d'UL dans la strate d'échantillonnage h qui comptent au moins une personne dans le domaine des 65 ans et plus selon la NSFG;

$a_h(d)$ = nombre moyen de personnes par UL dans la strate d'échantillonnage h qui sont réellement dans le domaine d selon les données de la NSFG; cette moyenne est fondée sur toutes les UL, y compris celles qui n'ont aucune personne dans le domaine;

n_h = nombre d'UL de l'échantillon attribuées à la strate h;

$n(d) = \sum_h n_h a_h(d)$ = nombre attendu de personnes de l'échantillon admissibles parce qu'elles sont dans le domaine d.

Comme l'indiquent les descriptions des strates du tableau 6.1, il se peut que la liste commerciale n'ait pas d'enregistrement pour une adresse. Quand la liste contient un enregistrement, cela peut ou non indiquer que l'UL compte des personnes de 65 ans et plus. De fait, d'après Valliant et coll. (2014, tableau 1), 36,8 % des enregistrements ne contenaient aucun renseignement sur l'âge, et dans le groupe des 65 ans et plus, dans l'ensemble, la liste comprenait 74,6 % des personnes trouvées dans la NSFG. Dans la strate 2, « la liste a un enregistrement; 1 personne ou plusieurs dans le groupe d'âge », 67,1 % $\left(p_h(d)\right)$ des UL ont réellement une personne de 65 ans ou plus, la moyenne étant de 0,947 $\left(a_h(d)\right)$ personne de 65 ans ou plus par UL. Bien que la strate 2 ait, de loin, la plus forte incidence de personnes de 65 ans et plus, environ le tiers des UL dans cette strate n'a personne dans le groupe d'âge cible. Dans les trois autres strates, il y a de petits pourcentages d'UL comptant des personnes de 65 ans et plus, même si la liste ne l'indique pas. Par conséquent, une répartition efficace attribuera le plus grand nombre d'unités à la strate 2, mais les autres strates devraient être échantillonnées pour assurer une couverture complète du groupe d'âge.

Tableau 6.1
Strates fondées sur la présence de personnes de 65 ans et plus dans la liste commerciale. Les proportions et les moyennes sont estimées à partir des données de la NSFG.

Strate h	Description	Proportion d'UL comptant au moins une personne âgée de 65 ans ou plus, $p_h(d)$	Nombre moyen de personnes de 65 ans et plus par UL, $a_h(d)$
1	La liste comporte un enregistrement; 0 personne dans le groupe d'âge	0,062	0,071
2	La liste comporte un enregistrement; 1 personne ou plus dans le groupe d'âge	0,671	0,947
3	La liste comporte un enregistrement; aucun renseignement sur l'âge	0,122	0,159
4	Aucun enregistrement dans la liste	0,102	0,128
	Total	0,176	0,236

Pour obtenir une approximation des coûts, supposons que le coût de la présélection et de l'abandon d'une UL non admissible soit c_S et que le coût moyen de la présélection d'une UL et de l'interview de toutes les personnes dans une UL admissible soit de c_{S+I}. Le coût attendu d'une UL de l'échantillon sélectionnée aléatoirement dans la strate d'échantillonnage h, quand la présélection sert à trouver un membre du domaine d, est

$$c_h(d) = p_h(d)c_{S+I} + \left[1 - p_h(d)\right]c_S.$$

Définissons l'effet de plan attribuable à l'utilisation de poids inégaux (Kish, 1992) comme étant $\text{deff}_w = 1 + \sum_{i=1}^{n}\left(w_i - \bar{w}\right)^2 \Big/ \left(n\bar{w}^2\right)$, où n est la taille de l'échantillon, w_i le poids d'échantillonnage pour l'élément i et $\bar{w}$ est leur moyenne. La taille effective de l'échantillon est $n_{\text{eff}} = n / \text{deff}_w$ et la taille attendue de l'échantillon du domaine dans un échantillon à probabilités égales d'UL est $n_{\text{eq}}(d)$. Le problème d'optimisation s'énonce comme suit :

Objectif : Trouver $\{n_h\}_{h=1}^{4}$ pour réduire le plus possible le coût total de la présélection et des interviews, $C_d = \sum_{h=1}^{4} n_h c_h(d)$.

Selon les contraintes suivantes :

1) Taille minimale de l'échantillon de la strate des UL : $n_h \geq n_{\min}$;
2) Taille effective de l'échantillon de personnes : $n_{\text{eff}}(d) = n_{\text{eq}}(d)$;
3) Taille maximale de l'échantillon de la strate des UL : $n_h \leq N_h$, N_h étant le nombre d'UL dans la population de la strate h;
4) Effet de plan pour la pondération des personnes : $\text{deff}_w(d) \leq d_0$, une constante fixe.

La deuxième contrainte sert à faciliter la comparaison de la répartition reposant sur la PM avec une autre reposant sur un échantillon à probabilités égales d'UL. Si les deux n'avaient pas la même taille effective d'échantillon et que la taille de l'échantillon découlant de la programmation mathématique était beaucoup plus petite, la répartition reposant sur la PM aurait l'air indûment bonne comparativement à celle reposant sur un échantillonnage à probabilités égales. La troisième contrainte n'a pas d'effet sur la grande population des ménages aux États-Unis, mais pourrait être nécessaire dans certaines applications.

Les paramètres de contrainte ont été fixés à $n_{\text{eq}}(d) = 2\,000$, $n_{\min} = 250$ et $d_0 = 1{,}5$. Les coûts unitaires en heures-personnes étaient de $c_S = 3$ et de $c_{S+I} = 10$. Bien que la fonction objectif soit linéaire dans les tailles d'échantillon de sous-strate, l'effet de plan, $\text{deff}_w(d)$, a les tailles d'échantillon dans les dénominateurs des poids d'échantillonnage, ce qui en fait un problème de programmation non linéaire.

L'estimation nationale de la NSFG du nombre moyen de personnes de 65 ans et plus par UL était de 0,236. Le nombre approximatif d'UL à présélectionner dans un échantillon à probabilités égales pour trouver 2 000 personnes de 65 ans et plus serait de 8 475 ≐ 2 000/0,236. Par ailleurs, la solution de PM pour obtenir une taille effective d'échantillon de 2 000 personnes était de 4 746, soit 56 % de l'échantillon

à probabilités égales. Le coût attendu de la répartition par PM était inférieur de 19 % à celui de l'échantillon à probabilités égales. Valliant et coll. (2014) ont également présenté des résultats pour le groupe des 18 à 44 ans. Étant donné qu'il est beaucoup plus fréquent d'avoir entre 18 et 44 ans, la PM reposant sur la liste imparfaite d'UL était moins efficace que l'échantillonnage à probabilités égales pour ce groupe d'âge.

La programmation mathématique peut être appliquée à des situations beaucoup plus complexes que l'exemple de la NSFG, qui ne tient pas compte de la précision des estimations. Une enquête auprès des ménages comme la HRS, présentée à la section 5.2, en est un exemple motivant. Cette enquête a de nombreux objectifs, dont l'estimation de statistiques sur les sources de revenu, les actifs et l'état de santé des unités financières, qui sont semblables aux ménages et aux personnes. La HRS comporte des cibles de taille d'échantillon pour un ensemble de domaines désignés par $d = 1, \ldots, D$. La HRS s'appuie également sur des listes d'UL commerciales pour chaque UPE, qui classent une UL selon la race ou l'origine ethnique et l'âge du chef de ménage. Elles servent à attribuer les UL à la sous-strate b. Cependant, les listes n'étant pas toujours exactes, il faut donc tenir compte de ce problème dans la répartition d'échantillon. Définissons $p_{ab}(d)$ comme étant la proportion d'UL dans la strate d'USE ou la sous-strate d'UL ab qui sont bien reconnues par les données de la liste commerciale comme étant dans le domaine d.

Supposons qu'il y a des coûts par UPE de l'échantillon, par USE de l'échantillon dans la strate a et par UL de l'échantillon dans la sous-strate ab, désignés respectivement par C_1, C_{2a} et C_{3ab}. Au moyen de la formulation selon laquelle $\overline{n}_a$ USE sont sélectionnées dans la strate a dans chaque UPE et que $\overline{\overline{q}}_{ab}$ UL sont échantillonnées dans chaque sous-strate ab dans chaque UPE/USE, on obtient la fonction de coût simple

$$C = C_0 + C_1 m + \sum_{a=1}^{A} C_{2a} m \overline{n}_a + \sum_{a=1}^{A} \sum_{b=1}^{B} C_{3ab} m \overline{n}_a \overline{\overline{q}}_{ab}. \tag{6.1}$$

Soit $\hat{\theta}_\ell$, $\ell = 1, \ldots, L$, un ensemble d'estimateurs importants dans le plan de sondage. Le problème d'optimisation consiste à trouver $\{m, \overline{n}_a, \overline{\overline{q}}_{ab};\ a = 1, \ldots, A, b = 1, \ldots, B\}$ qui minimisent la somme pondérée des variances relatives (c'est-à-dire la fonction objectif),

$$\phi = \sum_{\ell=1}^{L} \omega_\ell \, \mathrm{CV}^2\left(\hat{\theta}_\ell\right),$$

où les $\hat{\theta}_\ell$ sont des estimations à calculer à partir de l'échantillon et ω_ℓ est un poids d'échantillonnage préférentiel pour l'estimation ℓ.

La taille du poids d'échantillonnage préférentiel, ω_ℓ, attribuée à chaque variable d'analyse ℓ comprise dans l'optimisation, dépend des objectifs de l'enquête. Dans certaines enquêtes, il est parfois possible de déterminer les variables qui sont les principaux résultats d'intérêt, ce qui leur donne plus de poids dans l'optimisation. Par exemple, les variables de la HRS, comme le revenu, les actifs et l'état de santé, pourraient avoir plus de poids dans la fonction objectif ci-dessus. Les CV sont calculés au moyen de (5.6) et sont utilisés plutôt que les variances parce que les CV sont sans unité. Cela permet d'inclure des estimations mesurées à différentes échelles – comme le revenu moyen, la valeur moyenne des propriétés et les

proportions de personnes en mauvaise santé ou qui font des dons à des organismes de bienfaisance – dans ϕ sans que certaines de ces estimations en dominent la valeur comme elles le feraient si les variances des estimateurs étaient utilisées.

Plusieurs contraintes peuvent être utilisées pour les CV individuels et la taille des échantillons à différents degrés. Celles qui suivent sont fondées sur une enquête auprès des ménages, mais elles pourraient être adaptées à d'autres types d'échantillons.

a) Taille maximale de l'échantillon d'unités primaires : $m \le m_{\max}$, un ensemble maximal établi par le concepteur du plan de sondage;

b) Taille minimale de l'échantillon d'unités primaires : $m \ge m_{\min}$, par exemple $m_{\min} = 2$ pour tenir compte de l'estimation de la variance;

c) Taille maximale de l'échantillon des strates d'USE : $\overline{n}_a \le \min\{N_{ia} \mid i = 1, \ldots, m\}$ pour tous les a, c'est-à-dire que le nombre d'USE de l'échantillon ne peut pas dépasser le chiffre de la population d'USE dans ia;

d) Taille minimale de l'échantillon de strates d'USE : $\overline{n}_a \ge \overline{n}_{a,\min}$ pour tous les a;

e) Taille maximale de l'échantillon de sous-strates d'UL : $\overline{\overline{q}}_{ab} \le \min\{Q_{iajb} \mid i = 1, \ldots, m;\ j = 1, \ldots, \overline{n}_a\}$ pour tous les ab, c'est-à-dire que $\overline{\overline{q}}_{ab}$ a pour borne supérieure la plus petite valeur de Q_{iajb} dans toutes les combinaisons d'UPE ou d'USE;

f) Taille minimale de l'échantillon de sous-strates : $\overline{\overline{q}}_{ab} \ge \overline{\overline{q}}_{ab,\min}$;

g) Taille minimale et maximale de l'échantillon d'UL par UPE : $\mathrm{UL}_{\min} \le q_{i\cdot\cdot} \equiv \sum_a \sum_b \overline{n}_a \overline{\overline{q}}_{ab} \le \mathrm{UL}_{\max}$, c'est-à-dire un nombre minimal et maximal d'UL échantillonnées par UPE; cela pourrait être établi en fonction des exigences en matière de charge de travail pour les responsables de la collecte de données;

h) Coûts fixes : Le coût total de la variable est inférieur au montant prévu au budget $C - C_0 \le C_{\text{budget}}$;

i) Tailles d'échantillon cibles pour les domaines analytiques $d = 1, \ldots, D$ tenant compte de l'inexactitude des enregistrements dans les données provenant de listes commerciales : le nombre attendu d'UL de l'échantillon jugées admissibles en raison de leur appartenance au domaine analytique d est $q(d) = \sum_{a=1}^{A} \sum_{b=1}^{B} m \overline{n}_a \overline{\overline{q}}_{ab} p_{ab}(d)$. Des contraintes peuvent être fixées pour ce nombre, par exemple $q(d) = q_0(d)$. Il est également possible de fixer des contraintes pour la proportion d'UL attribuées à chaque domaine sans contraindre leurs totaux. Par exemple, si l'on souhaite avoir des échantillons d'UL de taille environ identique dans chaque domaine, la contrainte pourrait être

$$\frac{q(d)}{\sum_{d=1}^{D} q(d)} = \frac{1}{D} \pm \epsilon$$

pour une certaine tolérance ϵ;

j) Effets maximaux du plan pour les poids dans chaque domaine : $\mathrm{deff}_w(d) \le \mathrm{deff}_{\max}$, où $\mathrm{deff}_w(d)$ est l'effet de plan (Kish, 1992) attribuable à la pondération différentielle dans le domaine d.

La dernière contrainte peut être utile ou non. Elle vise à éviter que les poids de base varient trop. Toutefois, le fait de contraindre $\text{deff}_w(d)$ dans chaque domaine peut représenter un conflit avec d'autres contraintes, comme la taille des échantillons cible pour les domaines. Une variation du problème ci-dessus consisterait à utiliser un échantillon sélectionné au préalable d'UPE et à optimiser la répartition de l'échantillon dans cet ensemble. Cela pourrait se faire dans le cadre d'une enquête continue qui repose sur le même échantillon d'UPE pendant de longues périodes. La configuration d'un programme mathématique en présence de contraintes conflictuelles est une question assez courante qui mène à un problème sans solution faisable. Un bon logiciel vous permettra de le savoir.

L'application dans Maze (2021) repose sur les données de la HRS et intègre 11 estimateurs différents dans la fonction objectif, ϕ. Les résultats sont détaillés et ne sont pas présentés dans le présent article, mais en utilisant les estimations des composantes de la variance anticipée, on pourrait trouver une répartition des UPE, des USE et des UL qui respecte toutes les contraintes de taille de l'échantillon dans les limites d'un budget réaliste donné.

La mise en pratique de la PM mentionnée ci-dessus nécessite plusieurs étapes :

1) Estimer les composantes de la variance, B^2, W_{2a}^2 et W_{3ab}^2, à utiliser dans la formule de la variance relative (5.6). Celles-ci différeront pour chaque y;
2) Estimer les proportions, K_a et K_{ab}, des totaux de population des y dans les strates a et ab;
3) Estimer les taux d'exactitude, $p_{ab}(d)$, de la liste utilisée;
4) Obtenir les chiffres de population, Q_{iajb}, des UL dans l'UPE i, l'USE j dans la strate d'USE a et la sous-strate d'UL b;
5) Obtenir les coûts unitaires, C_0, C_1, C_{2a} et C_{3ab} nécessaires pour la fonction de coût.

Les étapes ci-dessus seront facilitées si des ajouts antérieurs à une enquête ont été effectués et si leurs données sont disponibles aux fins d'analyse. Même si le problème de PM a été soigneusement formulé, une solution peut donner des résultats inhabituels ou déconcertants. Il est toujours judicieux d'examiner méticuleusement les données de sortie, pour reformuler le problème, au besoin.

Le concepteur de l'échantillon dispose généralement d'une certaine marge de manœuvre pour formuler un problème de PM aux fins de répartition de la taille de l'échantillon. Presque toujours, le budget est la contrainte la plus importante. La façon d'atteindre les objectifs d'estimation dans les limites d'un budget fixe vient au deuxième rang. Les objectifs peuvent être énoncés en termes de taille d'échantillon cible pour les sous-groupes analytiques ou de CV pour les estimations importantes. Les contraintes, autres que budgétaires, peuvent être déterminées par les charges de travail gérables par les responsables de la collecte des données ainsi que par la nécessité d'estimer les variances des estimateurs à partir des données recueillies. Cette dernière contrainte peut, par exemple, dicter qu'au moins deux unités au premier degré soient sélectionnées dans chaque strate d'UPE. Il y a invariablement de la souplesse dans la façon de formuler le problème de répartition, dont la solution fait partie de l'élaboration du plan de sondage. Les concepteurs de

plan doivent tenir compte des préoccupations énoncées ci-dessus d'une façon ou d'une autre. La programmation mathématique est une façon officielle de le faire et peut souvent permettre de trouver des solutions plus efficaces que les méthodes moins systématiques.

7. Sommaire

La littérature s'est beaucoup plus intéressée à l'utilisation de modèles pour l'estimation de population finie que leur utilisation dans les plans de sondage. Valliant (2024) a examiné bon nombre des solutions de rechange qui ont été étudiées, dont l'estimation fondée sur un modèle (par exemple la meilleure prédiction linéaire sans biais et la régression multiniveau avec poststratification) et l'estimation assistée par un modèle (par exemple l'estimateur par la régression généralisée, l'estimateur calé par un modèle et l'estimateur par la vraisemblance empirique). Cependant, les modèles peuvent aussi jouer un rôle important dans la conception d'échantillons efficaces. Dans la conception d'échantillons de population finie, les modèles offrent un moyen de prendre officiellement en compte les effets des données auxiliaires disponibles avant la sélection de l'échantillon. Souvent, la puissance de prédiction des variables auxiliaires est considérée de façon informelle dans le plan de sondage au moment de créer des strates ou de déterminer les probabilités de sélection, mais l'attrait explicite des modèles peut aider à créer des plans plus efficaces et à clarifier l'efficacité d'un échantillon pour les différentes variables analytiques qui doivent être recueillies dans un échantillon.

Même à l'ère des mégadonnées, où des quantités considérables de données peuvent être moissonnées sur le Web, le plan de sondage a encore sa place. Les coûts de vérification attribuable au nettoyage des données moissonnées sur le Web peuvent être exorbitants, étant donné qu'une enquête peut avoir ses propres définitions spécialisées pour des variables comme la situation d'emploi ou les prix ajustés en fonction de la qualité aux fins d'un indice des prix. Les données moissonnées sur le Web n'ont pas nécessairement la forme requise par une enquête, ce qui entraîne des activités de vérification importantes et coûteuses. Un sous-échantillon bien conçu à partir d'un grand ensemble de données peut réduire les exigences de vérification et fournir autant de renseignements à un coût moindre.

Les techniques envisagées dans le présent article sont l'échantillonnage équilibré, l'échantillonnage défini par un seuil d'inclusion, la création de strates et l'échantillonnage à plusieurs degrés. Dans les enquêtes polyvalentes, la programmation mathématique est un outil technique précieux qui peut officiellement tenir compte de plusieurs contraintes dont les concepteurs du plan de sondage doivent tenir compte. Le plan adaptatif, présenté dans Groves et Heeringa (2006), Schouten, Shlomo et Skinner (2011), Tourangeau, Brick, Lohr et Li (2017), Wagner et Raghunathan (2010) et dans de nombreux autres articles, est un domaine qui dépend fortement des modèles, mais qui n'est pas abordé dans le présent article. L'article de Tourangeau portant sur Waksberg (2021) constitue une excellente revue et critique de la littérature sur les méthodes adaptatives.

Joe Waksberg a su relever les défis liés au plan de sondage, en s'appuyant sur son expérience et une intuition fine. Depuis son époque, les progrès réalisés, reposant sur des modèles soigneusement choisis, ont

donné de nombreux outils sophistiqués pour concevoir des enquêtes visant un large éventail d'objectifs difficiles à atteindre. Cela est particulièrement vrai pour la programmation mathématique, désormais disponible dans plusieurs modules de R et dans plusieurs feuilles de calcul. Ces logiciels permettent de trouver des répartitions d'échantillons complexes à degré unique et à plusieurs degrés dans le respect de nombreuses contraintes pratiques.

Remerciements

L'auteur tient à remercier Jill Dever, Alan Dorfman, le rédacteur associé et les examinateurs de leurs commentaires utiles qui ont permis d'améliorer la couverture de l'article.

Bibliographie

Ardilly, P., Haziza, D., Lavallée, P. et Tillé, Y. (2024). Les contributions de Jean-Claude Deville à la théorie des sondages et à la statistique officielle. *Techniques d'enquête*, 49, 2, 279-321. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2023002/article/00017-fra.pdf.

Baillargeon, S., et Rivest, L. (2009). A general algorithm for univariate stratification. *Revue Internationale de Statistique*, 77(3), 331-344. DOI : https://doi.org/10.1111/j.1751-5823.2009.00093.x.

Baillargeon, S., et Rivest, L.-P. (2011). Élaboration de plans stratifiés en R à l'aide du programme *stratification*. *Techniques d'enquête*, 37, 1, 59-72. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2011001/article/11447-fra.pdf.

Ballin, M., et Barcaroli, G. (2008). Optimal stratification of sampling frames in a multivariate and multidomain sample design. Rapport technique, Insituto Nationale di Statistica, Rome Italie. https://www.istat.it/it/files//2018/07/10_2008.pdf.

Ballin, M., et Barcaroli, G. (2013). Détermination conjointe de la stratification et de la répartition optimales de l'échantillon en utilisant un algorithme génétique. *Techniques d'enquête*, 39, 2, 405-432. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2013002/article/11884-fra.pdf.

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42(3), 174-177. DOI: https://doi.org/10.1080/00031305.1988.10475556.

Barcaroli, G., Ballin, M., Odendaal, H., Pagliuca, D., Willighagen, E. et Zardetto, D. (2022). *SamplingStrata: Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys*, v.1.5-4. https://cran.r-project.org/package=SamplingStrata.

Benedetti, R., Bee, M. et Espa, G. (2010). A framework for cut-off sampling in business survey design. *Journal of Official Statistics*, 26(4), 651-671.

Benedetti, R., Espa, G. et Lafratta, G. (2008). Une approche arborescente de la formation de strates dans les enquêtes-entreprises polyvalentes. *Techniques d'enquête*, 34, 2, 217-226. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2008002/article/10760-fra.pdf.

Bethel, J. (1989). Répartition de l'échantillon dans les enquêtes à plusieurs variables. *Techniques d'enquête*, 15, 1, 49-60. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1989001/article/14578-fra.pdf.

Box, G.E.P., Hunter, W.G. et Hunter, J.S. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*. New York: John Wiley & Sons, Inc., 2nd edition. ISBN:978-0-471-71813-0.

Breidt, F.J., et Opsomer, J.D. (2009). Nonparametric and semiparametric estimation in complex surveys. Dans *Handbook of Statistics, Sample Surveys: Inference and Analysis*, (Éd., C.R. Rao), volume 29B, chapitre 27, 103-119. Amsterdam : Elsevier.

Bureau de recensement des États-Unis (2006). Current Population Survey: Design and Methodology. https://www2.census.gov/programs-surveys/cps/methodology/tp-66.pdf.

Choudhry, G.H., Rao, J.N.K. et Hidiroglou, M.A. (2012). À propos de la répartition de l'échantillon pour une estimation sur domaine efficace. *Techniques d'enquête*, 38, 1, 25-32. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2012001/article/11682-fra.pdf.

Chromy, J.R. (1987). Design optimization with multiple objectives. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 194-199.

Chromy, J.R., et Myers, L.E. (2001). Variance models applicable to the NHSDA. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Cochran, W. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc., 3rd edition.

Cole, D., et Utting, J.E.G. (1956). Estimating expenditure, saving and income from household budgets. *Journal of the Royal Statistical Society, Series A*, 119, 371-392.

Costa, L., Guillo, C., Paliod, N., Merly-Alpa, T., Vincent, L., Chevalier, M. et Deroyon, T. (2018). Le tirage coordonné du nouvel échantillon-maître nautile avec l'échantillon de l'enquête emploi en continu. *Journées de Méthodologie Statistique, Insee*.

Cumberland, W.G., et Royall, R.M. (1981). Prediction models and unequal probability sampling. *Journal of the Royal Statistical Society B*, 43, 353-367.

Dalenius, T., et Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54(285), 88-101.

de Moura Brito, J.A., Silva, P.L.N., Semaan, G.S., et Maculan, N. (2015). Application des formulations de la programmation en nombres entiers à la répartition optimale dans l'échantillonnage stratifié. *Techniques d'enquête*, 41, 2, 451-467. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2015002/article/14249-fra.pdf.

Deville, J.-C., et Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91(4), 893-912.

Deville, J.-C., et Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128(2), 411-425.

Dorfman, A.H., et Valliant, R. (2000). Stratification by size revisited. *Journal of Official Statistics*, 16, 139-154.

Elliott, M.R., et Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264. DOI : https://doi.org/10.1214/16-STS598.

Fecteau, S., et Jocelyn, W. (2005). Une application de l'échantillonnage équilibré : le plan de sondage des entreprises non incorporées. *Méthodes d'Enquêtes et Sondages*, (Éds., P. Lavallée et L.-P. Rivest), 405-411. Paris : Dunod.

Frankel, L.R., et Stock, S. (1942). On the sample survey of unemployment. *Journal of the American Statistical Association*, 37(443), 77-80. DOI: https://doi.org/10.1080/01621459.1942.10500615.

Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4), 933-944. DOI: https://doi.org/10.1093/biomet/asp042.

Godambe, V.P., et Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations – I. *The Annals of Mathematical Statistics*, 36, 6, 1707-1723.

Godfrey, J., Roshwalb, A. et Wright, R. (1984). Model-based stratification in inventory cost estimation. *Journal of Business and Economic Statistics*, 2(1), 1-9.

Gomes, H., et Johnson, W.H. (2016). Sample size optimization of the Consumer Price Index: An implementation using R. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 2137-2151.

Goodman, R., et Kish, L. (1950). Controlled selection – A technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.

Grafström, A. (2012). Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference*, 142, 139-147.

Grafström, A., Lisic, J. et Prentius, W. (2023). *BalancedSampling: Balanced and Spatially Balanced Sampling,* R package version 1.6.3. https://CRAN.R-project.org/package=BalancedSampling.

Grafström, A., Lundström, N.L.P. et Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2), 514-520. DOI: https://doi.org/10.1111/j.1541-0420.2011.01699.x.

Grafström, A., et Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 24, 120-131.

Groves, R.M., et Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A*, 169, 3, 439-457. DOI: http://dx.doi.org/10.1111/j.1467-985x.2006.00423.x.

Groves, R.M., Mosher, W.D., Lepkowski, J. et Kirgis, N.G. (2009). Planning and development of the continuous National Survey of Family Growth. Vital Health Statistics, Series 1, No. 48, National Center for Health Statistics. https://www.cdc.gov/nchs/data/series/sr_01/sr01_048.pdf.

Gunning, P., et Horgan, J.M. (2004). Un nouvel algorithme pour la construction de bornes de stratification dans les populations asymétriques. *Techniques d'enquête*, 30, 2, 177-185. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2004002/article/7749-fra.pdf.

Hansen, M.H., Hurwitz, W.N. et Madow, W.G. (1953a). *Sample Survey Methods and Theory, Volume I. Methods and Applications*. New York: John Wiley & Sons, Inc.

Hansen, M.H., Hurwitz, W.N. et Madow, W.G. (1953b). *Sample Survey Methods and Theory, Volume II. Theory*. New York: John Wiley & Sons, Inc.

Haziza, D., Chauvet, G. et Deville, J.-C. (2010). Sampling and estimation in the presence of cut-off sampling. *Australia & New Zealand Journal of Statistics*, 52(3), 303-319. DOI: https://doi.org/10.1111/j.1467-842X.2010.00584.x.

Horgan, J.M. (2006). Stratification of skewed populations: A review. *Revue Internationale de Statistique*, 74(1), 67-76. https://www.jstor.org/stable/25472691.

Hughes, E., et Rao, J.N.K. (1979). Problems of optimal allocation in sample surveys involving inequality constraints. *Communications in Statistics – Theory and Methods*, 8(15), 1551-1574.

Hunter, S.R., Bowman, K.R. et Chromy, J.R. (2005). Results of the variance component analysis of sample allocation by age in the National Survey on Drug Use and Health. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3132-3136.

Isaki, C.T., et Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377), 89-96. DOI: https://doi.org/10.2307/2287773.

Judkins, D., et Van de Kerckhove, W. (2003). RECS 2005 optimization. Préparé pour le U.S. Department of Energy, nº 16.3, Tâche 98-010, contrat nº : DE-AC01-96E123968. Rapport technique, Westat, Rockville MD.

Kirkendall, N.J. (1992). When is model-based sampling appropriate for EIA surveys? *Proceedings of the Section on Survey Methods Research*, 637-642. http://www.asasrms.org/Proceedings/index.html.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Kish, L. (1992). Weighting for unequal pi. *Journal of Official Statistics*, 8(2), 183-200.

Knaub, J.R. (2008). Cutoff vs. design-based sampling and inference for establishment surveys. *InterStat*. http://interstat.statjournals.net/YEAR/2008/abstracts/0806005.php.

Kott, P.S. (1986). Some asymptotic results for the systematic and stratified sampling of a finite population. *Biometrika*, 73(2), 485-491.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6(5), 797-806.

Lavallée, P., et Hidiroglou, M.A. (1988). Sur la stratification de populations asymétriques. *Techniques d'enquête*, 14, 1, 35-45. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1988001/article/14602-fra.pdf.

Leaver, S., et Solk, D.T. (2005). Handling program constraints in the sample design for the commodities and services component of the U.S. Consumer Price Index. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3290-3298.

Maze, A. (2021). *Using Commercial List Information in Screening Eligible Housing Units*. Thèse de doctorat, University of Maryland. DOI: https://doi.org/10.13016/xdzx-dto7.

Morganstein, D., et Marker, D. (2000). A conversation with Joseph Waksberg. *Statistical Science*, 15(3), 299-312.

Nedyalkova, D., et Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model. *Biometrika*, 95, 521-537.

Neter, J., et Waksberg, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, 59(305), 18-55. DOI: https://doi.org/10.1080/01621459.1964.10480699.

Pfeffermann, D., et Sverchkov, M. (2009). Inference under informative sampling. Dans *Handbook of Statistics, Sample Surveys: Design, Methods, and Applications*, (Éd., C.R. Rao), volume 29A, chapitre 39. Amsterdam: Elsevier.

Rivest, L.-P. (2002). Une généralisation de l'algorithme de Lavallée et Hidiroglou pour la stratification dans les enquêtes auprès des entreprises. *Techniques d'enquête*, 28, 2, 207-214. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2002002/article/6432-fra.pdf.

Royall, R.M. (1992). Robustesse et optimalité de plan dans des modèles de prédiction pour populations finies. *Techniques d'enquête*, 18, 2, 193-199. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1992002/article/14488-fra.pdf.

Royall, R.M., et Cumberland, W.G. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, 80(390), 355-359.

Royall, R.M., et Herson, J. (1973a). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68(344), 880-889.

Royall, R.M., et Herson, J. (1973b). Robust estimation in finite populations II. *Journal of the American Statistical Association*, 68(344), 890-893.

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

Schouten, B., Shlomo, N. et Skinner, C.J. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, 27, 1-24.

Schroeder, T., et Ault, K. (2001). The NEISS Sample (Design and Implementation) from 1979 to 1996. Rapport technique, U.S. Consumer Product Safety Commission, Washington DC. https://www.cpsc.gov/s3fs-public/2001d010-6b6.pdf.

Schwendinger, F., et Borchers, H.W. (2023). CRAN Task View: Optimization and Mathematical Programming. Rapport technique, The R Foundation, Vienne Autriche. https://CRAN.R-project.org/view=Optimization.

Searle, S., Casella, G. et McCulloch, C. (1992). *Variance Components*. New York: John Wiley & Sons, Inc.

Sethi, V.K. (1963). A note on the optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5, 20-33.

Slanta, J.G., et Krenzke, T.R. (1994). Applying the Lavallée and Hidiroglou method to obtain stratification boundaries for the Census Bureau's Annual Capital Expenditure Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 693-698.

Slanta, J.G., et Krenzke, T.R. (1996). Utilisation de la méthode de Lavallée et Hidiroglou pour le calcul des limites de stratification aux fins de l'enquête annuelle sur les dépenses en capital du Bureau of the Census. *Techniques d'enquête*, 22, 1, 65-75. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1996001/article/14384-fra.pdf.

Tam, S.M. (1995). Optimal and robust strategies for cluster sampling. *Journal of the American Statistical Association*, 90, 379-382.

Tillé, Y., et Matei, A. (2023). *sampling: Survey Sampling*, R package version 2.10. http://CRAN.R-project.org/package=sampling.

Tourangeau, R. (2021). Science et gestion d'enquête. *Techniques d'enquête*, 47, 1, 3-32. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2021001/article/00006-fra.pdf.

Tourangeau, R., Brick, J.M., Lohr, S. et Li, J. (2017). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society, Series A*, 180(1), 203-223. DOI: http://onlinelibrary.wiley.com/doi/10.1111/rssa.12186.

U.S. Energy Information Administration (2018). EIA-914 monthly crude oil and lease condensate, and natural gas production report methodology. Rapport technique, U.S. Department of Energy, Washington DC. https://www.eia.gov/petroleum/production/pdf/eia914methodology.pdf.

Valliant, R. (2002). Estimation de la variance de l'estimateur de régression généralisée. *Techniques d'enquête*, 28, 1, 109-122. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2002001/article/6424-fra.pdf.

Valliant, R. (2024). Hansen lecture 2022: The evolution of the use of models in survey sampling. *Journal of Survey Statistics and Methodology*, 12(2), 275-304. DOI: https://doi.org/10.1093/jssam/smad021.

Valliant, R., Dever, J.A. et Kreuter, F. (2015). Effects of cluster sizes on variance components in two-stage sampling. *Journal of Official Statistics*, 31(4), 763-782. DOI: http://dx.doi.org/10.1515/JOS-2015-0044.

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.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.

Valliant, R., et Gentle, J.E. (1997). An application of mathematical programming to sample allocation. *Computational Statistics & Data Analysis*, 25(3), 337-360. DOI: https://doi.org/10.1016/S0167-9473(97)00007-8.

Valliant, R., Hubbard, F., Lee, S. et Chang, C. (2014). Efficient use of commercial lists in U.S. household sampling. *Journal of Survey Statistics and Methodology*, 2(2), 182-209. DOI: https://doi.org/10.1093/jssam/smu006.

Varadhan, R. (2023). *alabama: Constrained Nonlinear Optimization*, R package version 2023.1.0. https://CRAN.R-project.org/package=alabama.

Wagner, J., et Raghunathan, T.E. (2010). A new stopping rule for surveys. *Statistics in Medicine*, 29(9), 1014-1024. DOI: https://doi.org/10.1002/sim.3834.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73(361), 40-46. DOI: https://www.tandfonline.com/doi/abs/10.1080/01621459.1978.10479995.

Waksberg, J., Sperry, S., Judkins, D. et Smith, V. (1993). The National Survey of Family Growth, Cycle IV, evaluation of linked design. *Vital and Health Statistics*, 2, (117), (PHS) 93-1391.

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. http://www.jstor.org/stable/2670358.

Wu, C.F.J., et Hamada, M. (2021). *Experiments: Planning, Analysis, and Optimization*, 3rd edition. New York: John Wiley & Sons, Inc. DOI: https://doi.org/10.1002/9781119470007.

Yorgason, D., Bridgman, B., Cheng, Y., Dorfman, A., Lent, J., Liu, Y.K., Miranda, J. et Rumburg, S. (2011). Cutoff sampling in federal surveys: An inter-agency review. Rapport technique, Bureau of Labor Statistics, Washington DC. https://www.bls.gov/osmr/research-papers/2011/pdf/st110050.pdf.

Ypma, J., Johnson, S.G., Borchers, H.W., Eddelbuettel, D., Ripley, B., Hornik, K., Chiquet, J., Adler, A., Dai, X., Stamm, A. et Ooms, J. (2022). *nloptr: R Interface to NLopt*, version 2.0.3. https://CRAN.R-project.org/package=nloptr.

Techniques d'enquête, décembre 2024
Vol. 50, nº 2, pp. 207-230
Statistique Canada, nº 12-001-X au catalogue

Modèles de forêt aléatoire convergents par rapport au plan pour la collecte de données recueillies à partir d'un échantillon complexe

Daniell Toth et Kelly S. McConville[1]

Résumé

Les modèles de forêt aléatoire, qui sont obtenus en calculant la moyenne des valeurs estimées d'un grand nombre de modèles arborescents, représentent un outil utile et souple pour modéliser les données de manière non paramétrique afin de fournir des valeurs hautement prédictives. Il existe un grand nombre d'applications potentielles pour ces types de modèles lorsqu'on traite des données d'enquête. Toutefois, puisque les données d'enquête sont habituellement recueillies à l'aide d'un plan de sondage informatif, il est nécessaire que l'algorithme utilisé pour créer les modèles de forêt aléatoire tienne compte de ce plan pendant l'estimation du modèle.

Les modèles arborescents utilisés dans la forêt sont généralement obtenus en estimant les modèles arborescents sur des échantillons bootstrap des données d'origine. Comme les modèles dépendent des données observées et que les données observées dans l'échantillon dépendent du plan de sondage informatif, la méthode d'estimation habituelle est susceptible de mener à un modèle de forêt aléatoire biaisé lorsque ce dernier est appliqué aux données d'enquête.

Dans le présent article, nous fournissons un algorithme et un ensemble de conditions produisant des modèles de forêt aléatoire convergents dans le cadre d'un plan de sondage informatif et comparons cette méthode avec la méthode habituelle de modélisation de type forêt aléatoire. Nous démontrons que le fait de ne pas tenir compte du plan de sondage peut donner lieu à des estimations de modèle comportant un biais.

Mots-clés : Apprentissage automatique; données d'enquête; modèles arborescents; non paramétrique; plan de sondage.

1. Introduction

Les algorithmes de partitionnement récursif ont d'abord été suggérés par Morgan et Sonquist (1963) comme méthode d'analyse de données d'enquête, du fait des relations complexes (notamment les effets d'interaction) entre les variables typiques de ces ensembles de données. Les variables recueillies à partir d'une enquête sont souvent hautement corrélées entre elles (voire colinéaires), sont fréquemment catégoriques et peuvent contenir de nombreuses valeurs manquantes. Du fait de ces complications, faire des inférences avec ces données sur la population cible en utilisant des modèles paramétriques traditionnels peut être difficile. Les modèles arborescents, estimés en appliquant un algorithme de partitionnement récursif à l'ensemble de données, permettent de traiter facilement ce type de données. Les variables utilisées dans ce modèle ainsi que tout effet d'interaction sont automatiquement sélectionnés et la structure binaire segmentée obtenue rend ces modèles faciles à interpréter et permet de relever les effets d'interaction complexes entre les variables de l'ensemble de données (Phipps et Toth, 2012; Earp, Toth, Phipps et Oslund, 2018).

Soit un ensemble de n observations $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ de la variable de réponse aléatoire Y et d variables prédictives aléatoires $\mathbf{X} = (X_1, \ldots, X_d)$, provenant d'un échantillon informatif, nous souhaitons estimer k

1. Daniell Toth, Office of Survey Methodology, U.S. Bureau of Labor Statistics. Courriel : toth.daniell@bls.gov; Kelly S. McConville, Department of Statistics, Harvard University.

nouvelles valeurs $\{y_i,\}_{i=n+1}^{n+k}$, étant donné les valeurs prédictives $\{\mathbf{x}_i\}_{i=n+1}^{n+k}$ pour les unités non échantillonnées de la population. En estimant la fonction moyenne $E[Y \mid \mathbf{X} = \mathbf{x}] = h(\mathbf{x})$ à partir des données observées, nous pouvons obtenir des prédictions de y_i avec le modèle $\tilde{y}_i = \tilde{h}(\mathbf{x})$.

Utiliser des données d'enquête pour estimer un modèle afin d'obtenir de bonnes prédictions, plutôt que des estimations d'un paramètre de population finie, est un sujet suscitant un intérêt croissant (Wieczorek, 2023). Par exemple, Hong et He (2010) utilisent des données d'études longitudinales pour ajuster un modèle pouvant servir à prédire l'état de mobilité fonctionnelle des aînés. Parallèlement, Kshirsagar, Wieczorek, Ramanathan et Wells (2017) ainsi que Krebs, Reeves et Baggett (2019) utilisent des modèles d'apprentissage automatique pour prédire les niveaux de pauvreté et la structure de la végétation de sous-bois, respectivement. De manière similaire, nous souhaitons estimer la fonction de régression à l'aide d'une approche d'apprentissage automatique reposant sur des données provenant d'un plan de sondage informatif. Comme Nalenz, Rodemann et Augustin (2024), nous proposons une approche de modélisation de forêts aléatoires à l'aide de données d'enquête.

Un modèle arborescent, $\tilde{h}(\mathbf{x})$, est un modèle non paramétrique obtenu à partir d'un algorithme partitionnant de façon récursive les données observées, puis estimant la statistique souhaitée séparément pour chaque boîte de partitionnement final (nœud final). L'algorithme de partitionnement récursif consiste à choisir une variable X_j parmi toutes les d variables disponibles données par le vecteur $\mathbf{X}$ et une valeur a où l'on peut fractionner l'ensemble d'observations en deux nœuds : les observations où $\mathbf{x}_j \le a$ et $\mathbf{x}_j > a$. Cette procédure est ensuite répétée pour chaque nœud jusqu'à ce qu'il ne reste plus assez d'observations à fractionner ou jusqu'à ce qu'un critère d'arrêt soit atteint (Hothorn, Hornik et Zeileis, 2006). Cet algorithme donne un ensemble de q boîtes, $Q^n = \{B_1, \ldots, B_q\}$, qui partitionnent entièrement le support de $\mathbf{X}$ et dépendent des valeurs des données observées.

Même si les modèles arborescents sont faciles à interpréter (ce qui fait en sorte qu'ils sont l'outil parfait pour répondre aux exigences de nombreuses applications d'inférence), ils ne sont pas des modèles très efficaces pour la production d'estimations ponctuelles. Ils sont particulièrement inefficaces pour les modèles présentant des effets linéaires (Loh, 2008). Un modèle de forêt aléatoire, contrairement au modèle arborescent facile à interpréter, estime la valeur attendue de la variable réponse conditionnellement aux variables prédictives, en calculant la moyenne des estimations d'un ensemble de modèles d'arbre de régression.

Soit un ensemble de M modèles d'arbre de régression, $\{\tilde{h}_j\}_{j=1}^{M}$, l'estimateur de forêt aléatoire de $h(\mathbf{x})$ est

$$\mathcal{F}_0(\mathbf{x}) = \frac{1}{M}\sum_{j=1}^{M} \tilde{h}_j(\mathbf{x}), \tag{1.1}$$

où chaque modèle arborescent, $\tilde{h}_j$, est ajusté à l'aide d'un sous-ensemble aléatoire des variables prédictives sur un échantillon bootstrap des données observées (Breiman, 2001). Même si ces modèles éliminent la caractéristique de la facilité d'interprétation des modèles arborescents, ils sont connus pour fournir des prédictions très exactes et demeurent tout de même applicables à un vaste éventail de types de données

(Breiman, 2001). Cela fournit un outil utile et souple pour modéliser avec exactitude la variable réponse d'un ensemble de données déterminé, ce qui pourrait avoir de nombreuses applications dans l'analyse de données provenant d'un échantillon informatif.

Buskirk (2018) ainsi que Bilton, Jones, Ganesh et Haslett (2017) présentent, par exemple, des applications d'arbres de régression et de forêts aléatoires sur des données recueillies à l'aide d'un plan de sondage complexe. Malheureusement, les algorithmes standard de forêt aléatoire sont conçus pour des données indépendantes et identiquement distribuées (i.i.d.) et bon nombre d'enquêtes s'appuient sur un plan de sondage complexe pour recueillir des observations, enfreignant l'hypothèse de données i.i.d. et, dans de nombreuses applications de modèles arborescents, les renseignements disponibles sur le plan de sondage sont souvent ignorés, ce qui donne probablement lieu à des estimations biaisées comme le démontrent les résultats de Toth et Eltinge (2011).

Dagdoug, Goga et Haziza (2021) ont étendu les travaux de McConville et Toth (2019) en utilisant un modèle arborescent plutôt qu'un arbre simple comme modèle secondaire dans un estimateur fondé sur un modèle pour estimer un total de population finie. Ils soulignent que, si les variables utilisées pour déterminer le plan de sondage sont disponibles, il est possible de réduire la majeure partie du biais en englobant ces variables dans le modèle. Ces variables sont extrêmement utiles pour estimer des paramètres de population dans le contexte d'une estimation assistée par un modèle, mais ces variables ne sont pas disponibles pour prédire les valeurs relatives à des unités hors de l'échantillon.

Il est alors souhaitable de disposer d'un algorithme qui permet une estimation convergente d'une fonction de régression de la population, estimée à l'aide de données provenant d'un plan de sondage informatif, pouvant être utilisé pour la prédiction. Par exemple, des travaux effectués au Bureau de la statistique du travail des États-Unis (BLS) attendent d'un modèle qu'il prédise le fardeau du répondant pour les ménages sélectionnés dans l'enquête sur les dépenses des consommateurs à partir de caractéristiques de ménage que l'on pense être associées au fardeau (Yang et Toth, 2022).

Dans le présent article, nous proposons un modèle de forêt aléatoire convergent par rapport au plan de sondage pour la fonction de régression qui est fondé sur une moyenne pondérée des nœuds finaux obtenus d'un ensemble d'arbres purement aléatoires comprenant des poids d'échantillonnage dans leur estimation. Ce processus évite de devoir produire des échantillons bootstrap raisonnables à partir d'un plan de sondage général. Les forêts construites à partir d'arbres complètement aléatoires ont été étudiées dans la littérature portant sur la méthode couramment appelée *algorithme de forêt aléatoire uniforme* (Biau, Devroye et Lugosi, 2008; Scornet, 2016; Arlot et Genuer, 2014), mais ces modèles ne sont généralement pas efficaces en pratique. Ils sont principalement utilisés pour étudier les propriétés théoriques et pour comprendre le comportement et les limites de méthodes d'ensemble. Dans l'algorithme de forêt aléatoire uniforme standard, les estimations des nœuds finaux sont de simples moyennes, alors que les facteurs de pondération des nœuds finaux de notre méthode permettent de bonnes propriétés prédictives.

À notre connaissance, notre algorithme est le premier à proposer d'utiliser des facteurs de pondération au nœud final, plutôt qu'au niveau de l'arbre. Parce que les arbres sont produits à l'aide de fractionnements

entièrement aléatoires de l'espace des variables d'explication, tout le travail réel provient de ces facteurs de pondération, fournissant un estimateur adaptatif et plus efficace. Nous montrons que ce modèle fournit des estimations convergentes par rapport au plan et qu'il est donc plus approprié pour une utilisation avec des données recueillies à l'aide d'un plan de sondage informatif.

À la section 2, nous présentons ce modèle arborescent ainsi que les hypothèses nécessaires à la convergence par rapport au plan. À la section 3, nous présentons la méthode permettant d'utiliser des modèles arborescents dans un modèle de forêt aléatoire nécessitant de pondérer chaque estimation de modèle arborescent ainsi que l'énoncé du principal résultat théorique de l'article, qui est que l'estimateur de forêt aléatoire proposé est un estimateur convergent par rapport au plan de la fonction de régression. L'annexe présente les lemmes et les démonstrations auxiliaires des résultats. La section 4 résume les études par simulation dans le cadre desquelles nous comparons le rendement de la méthode proposée à l'estimateur de forêt aléatoire standard sur des données provenant d'échantillons aléatoires simples (EAS) et d'échantillons de probabilité proportionnelle à la taille (PPT). Nous appliquons, en particulier, le modèle proposé de forêt aléatoire aux données de résultats de l'indice de réussite scolaire (API) provenant des résultats de tests standardisés d'élèves calculés pour tous les établissements scolaires de Californie comptant au moins 100 élèves ainsi qu'aux données des dépenses des consommateurs (CE) du Bureau de la statistique du travail des États-Unis (BLS). Ces résultats démontrent que ne pas tenir compte des poids d'échantillonnage lorsque nous utilisons des données provenant d'un plan de sondage informatif pourrait donner lieu à des estimateurs de forêts biaisés. À la section 4, nous effectuons aussi une étude par simulation sur données générées afin d'étudier la cohérence de la méthode proposée et du modèle de forêt aléatoire standard.

2. Modèles arborescents convergents par rapport au plan

Considérons une population finie de taille N, $\{(Y_i, \mathbf{X}_i, \mathbf{Z}_i)\}_{i=1}^{N}$, générée à partir d'un modèle de superpopulation ξ, où Y_i est la variable d'intérêt associée à l'unité i et $\mathbf{X}_i$ est un vecteur de dimension d de valeurs prédictives potentielles associées à l'unité i qui font partie des données diffusées à la disposition de l'analyste. Nous utilisons $\mathbf{Z}_i$ pour désigner un vecteur de dimension d^* de variables associées à l'unité i, connues du concepteur de l'enquête, mais non diffusées avec les données d'enquête aux fins d'analyse.

Un échantillon aléatoire $S \subset U = \{1,\ldots,N\}$ de taille n est tiré à l'aide d'un plan de sondage à probabilités d'inclusion $\pi_i = P(i \in S)$. Le plan de sondage, défini par les probabilités d'inclusion, peut dépendre de variables associées à l'unité, dont certaines sont connues de l'analyste et d'autres ne le sont pas, $\pi_i = P(i \in S \mid \mathbf{x}_i, \mathbf{z}_i)$. Si ces probabilités d'inclusion sont associées à Y, le plan peut avoir une incidence sur les estimations et l'inférence concernant la population qui découlent de l'utilisation des données-échantillon. De tels plans de sondage sont appelés des plans de sondage *informatifs*. Si le plan de sondage dépend uniquement de variables à la disposition de l'analyste de données, $\pi_i = P(i \in S \mid \mathbf{x}_i, \mathbf{z}_i) = P(i \in S \mid \mathbf{x}_i)$, il peut être possible d'obtenir des modèles convergents par rapport au plan en intégrant toutes ces variables utilisées pour définir les probabilités d'inclusion au processus de modélisation (Gelman, King et Liu, 1998; Little, 2004). Toutefois, dans la plupart des ensembles de données d'enquête à la disposition du public, bon

nombre des variables utilisées dans le plan d'enquête ne sont pas mises à la disposition de l'analyste de données. Les données sont plutôt diffusées avec un ensemble de facteurs de pondération d'enquête $\{w_i\}_{i\in S}$ destinés à l'usage de l'analyste de données pour tenir compte du plan d'enquête dans son analyse (Lavallée et Beaumont, 2015; Pfeffermann, 1993). En plus de représenter la probabilité de la sélection dans l'échantillon, ces facteurs de pondération comportent souvent des ajustements pour la non-réponse ou des totaux connus de l'information auxiliaire clé. Même si nos arguments dans le présent article pourraient être utilisés pour des facteurs de pondération d'enquête généraux, pour simplifier l'exposé, nous supposons que le facteur de pondération associé à l'unité i est l'inverse de la probabilité de sélection de cette unité, π_i^{-1}.

Afin d'étudier de grandes propriétés d'échantillon de l'estimateur dans ce contexte, il est nécessaire de tenir compte d'une séquence de populations de taille croissante et distribuées de manière indépendante et identique à partir d'une super-population ainsi qu'une séquence des plans de sondage connexes. Pour chaque paire de population-plan de sondage, une séquence correspondante d'échantillons est tirée, également de taille croissante, et chaque échantillon est sélectionné conformément au plan de sondage. Plus concrètement, supposons une séquence de populations finies $\{(Y_1, \mathbf{X}_1), \ldots, (Y_{N_\nu}, \mathbf{X}_{N_\nu})\}$, d'indice ν, de sorte que $U_1 \leq \ldots, \leq U_\nu$, pour des tailles $N_1 \leq \ldots \leq N_\nu$. Chaque population finie est générée par tirages i.i.d. à partir de la répartition de la super-population ξ. Les échantillons aléatoires, $S_1 \subset U_1, \ldots S_\nu \subset U_\nu$, sont tirés de chaque population finie conformément au plan de sondage correspondant, avec des tailles croissantes $n_1 \leq \ldots \leq n_\nu$. Le comportement de la séquence des estimations obtenues de ces échantillons est pris en considération.

Si un modèle arborescent, $\tilde{h}_\nu(\mathbf{x})$, est obtenu par partitionnement récursif des données-échantillon observées, puis par estimation de la moyenne de chaque boîte, le modèle arborescent obtenu est un estimateur de la fonction de moyenne conditionnelle $h(\mathbf{x}) = E_\xi[Y \mid \mathbf{x}]$. Toth et Eltinge (2011) fournissent un algorithme permettant d'estimer $h(\mathbf{x})$ ainsi qu'un ensemble de conditions pour lesquelles cet estimateur est un estimateur de $h(\mathbf{x})$ convergent par rapport à L^2. L'objectif du présent article est de proposer un modèle de forêt aléatoire construit à partir d'une moyenne pondérée de ces modèles arborescents convergents par rapport au plan. Dans le reste de cette section, nous passons en revue la notation et les résultats nécessaires pour établir un algorithme convergent par rapport au plan pour des modèles de forêt aléatoire, en présentant un exposé sur les conditions et le principal résultat du modèle arborescent donné dans Toth et Eltinge (2011).

Soit $Q^{n_\nu} = \{B_1^{n_\nu}, \ldots, B_q^{n_\nu}\}$, l'ensemble de boîtes de partitionnement obtenu en appliquant un algorithme de partitionnement récursif à l'échantillon observé S_ν. Pour faciliter l'analyse de la valeur prédite d'une observation avec des variables prédictives $\mathbf{x}$ pour un arbre donné, nous définissions maintenant certaines fonctions permettant de simplifier la notation. Soit $B^{n_\nu}(\mathbf{x})$, la boîte dans Q^{n_ν} contenant la valeur $\mathbf{x}$. Les fonctions $\#B^{n_\nu}(\mathbf{x}) = \sum_{i\in S} \mathbb{I}_{\{\mathbf{x}_i \in B^{n_\nu}(\mathbf{x})\}}$ et $\tilde{\#}B^{n_\nu}(\mathbf{x}) = \sum_{i\in S} \pi_i^{-1} \mathbb{I}_{\{x_i \in B^{n_\nu}(\mathbf{x})\}}$ fournissent respectivement le nombre d'unités d'échantillonnage observées et le nombre estimé d'unités de population dans la boîte $B^{n_\nu}(\mathbf{x})$. La moyenne estimée de la boîte contenant la valeur $\mathbf{x}$ est définie par

$$\tilde{\mu}(\mathbf{x}) = \left[\tilde{\#}B^{n_\nu}(\mathbf{x})\right]^{-1} \sum_{i\in S} \pi_i^{-1} y_i \mathbb{I}_{\{\mathbf{x}_i \in B^{n_\nu}(\mathbf{x})\}}. \tag{2.1}$$

Il convient de souligner qu'il s'agit de l'estimateur standard de Hájek de la moyenne des unités de population contenues dans la boîte de partitionnement $B^{n_v}(\mathbf{x})$ (Hájek, 1960).

Chaque boîte B^{n_v} d'une partition Q^{n_v} donnée compte deux vecteurs d'indice correspondants qui définissent les limites de la boîte dans toutes les dimensions d. Étant donné une boîte de partition, B^{n_v}, soient $\mathbf{a}(B^{n_v}) = (a_1(B^{n_v}), \ldots, a_d(B^{n_v}))$ et $\mathbf{b}(B^{n_v}) = (b_1(B^{n_v}), \ldots, b_d(B^{n_v}))$, où pour chaque $\mathbf{x} \in B^{n_v}$, $a_l(B^{n_v}) \leq x_l < b_l(B^{n_v})$, pour $l = 1 \ldots d$.

Pour une valeur donnée $\mathbf{x}$ corroborant $\mathbf{X}$, nous utilisons la notation $\tilde{F}_i^v(\cdot)$ pour désigner la fonction de distribution marginale empirique de x_l dépendant de la partition. Cela signifie que, pour une constante c et une valeur donnée $\mathbf{x}$, la fonction de distribution marginale empirique de x_l dépendant de la partition est

$$\begin{aligned} \tilde{F}_l^v(c \mid Q^{n_v}) &= \tilde{F}_l^v\left(c \mid B^{n_v}(\mathbf{x})\right) \\ &= \left(\tilde{\#}_{N_v}\left(B^{n_v}(\mathbf{x})\right)\right)^{-1} \sum_{i \in S_v} \pi_{vi}^{-1} \mathbb{I}_{\{x_{li} \leq c\}} \mathbb{I}_{\{\mathbf{x}_i \in B^{n_v}(\mathbf{x})\}}. \end{aligned} \tag{2.2}$$

La fonction de distribution marginale empirique conditionnelle continue de gauche $\tilde{F}_l^-$ est définie en remplaçant la fonction indicatrice $\mathbb{I}_{\{x_{li} \leq x\}}$ dans la définition ci-dessus par $\mathbb{I}_{\{x_{li} < x\}}$. Nous utiliserons également la fonction de probabilité empirique $\tilde{P}_{n_v}(\mathcal{A})$ d'un événement donné $\mathcal{A}$. La fonction de probabilité empirique est définie comme suit :

$$\tilde{P}_{n_v}(\mathcal{A}) = \tilde{N}_v^{-1} \sum_{i \in S_v} \pi_{vi}^{-1} \mathbb{I}_{\{\mathcal{A}\}}(\mathbf{x}_i), \tag{2.3}$$

où $\mathbb{I}_{\{\mathcal{A}\}}(\mathbf{x}_i) = 1$, si l'événement $\mathcal{A}$ est satisfait pour l'observation $\mathbf{x}_i$ et où $\tilde{N}_v = \sum_{i \in S_v} \pi_i^{-1}$.

Nous définissons ensuite la norme l de la partition Q^{n_v} par rapport à $\tilde{F}_l$ sous la forme

$$\| Q^{n_v} \|_l^{\tilde{F}} = \sum_{B^{n_v} \in Q^{n_v}} \left\{ \left[\tilde{F}_l\left(b_l(B^{n_v})\right) - \tilde{F}_l\left(a_l(B^{n_v})\right) \right] \tilde{P}(\mathbf{x} \in B^{n_v}) \right\} \tag{2.4}$$

puis, la norme l de la partition Q^{n_v} par rapport à $\tilde{F}^-$

$$\| Q^{n_v} \|_l^{\tilde{F}^-} = \sum_{B^{n_v} \in Q^{n_v}} \left\{ \left[\tilde{F}_l^-\left(b_l(B^{n_v})\right) - \tilde{F}_l^-\left(a_l(B^{n_v})\right) \right] \tilde{P}(\mathbf{x} \in B^{n_v}) \right\}. \tag{2.5}$$

Les conditions suivantes sur le modèle de super-population, le plan de sondage et la partition créée à partir de l'algorithme sont suffisantes pour montrer qu'un modèle arborescent de régression fondé sur les données-échantillon est un estimateur convergent L^2 de la moyenne conditionnelle réelle de la variable d'intérêt, Y. À la section 3, nous montrons que ces conditions (avec le renforcement d'une condition) sont suffisantes pour obtenir également des estimateurs arborescents convergents. Les démonstrations d'arbres de régression convergents nécessitent uniquement un moment fini de second ordre de la variable d'intérêt, mais nous avons besoin d'un moment fini de quatrième ordre pour démontrer la convergence de l'estimateur arborescent proposé.

Bon nombre des conditions nécessaires pour obtenir une convergence exigent de comprendre le taux de convergence des éléments. Avant de préciser les conditions relatives à la population, le plan de sondage et

l'algorithme, nous définissons d'abord deux fonctions scalaires qui serviront de taux de convergence. Soient $\gamma(x)$ et $k(x)$, les fonctions limitées au-dessus de 0 pour tous les $x > 0$ satisfaisant aux conditions suivantes :

1: $\gamma(x) \to \infty$

2: $x^{-1}k(x) \to 0$

3: $k(x)^{-1}\gamma(x)x^{1/2} \to 0,$

à mesure que $x \to \infty$. Ces contraintes exigent que les deux fonctions ne soient pas limitées, où $\gamma(x)$ augmente vers ∞ plus lentement que $\sqrt{x}$, alors que $k(x)$ augmente plus rapidement que $\sqrt{x}$, mais plus lentement que x. Il convient de mentionner qu'un nombre infini de paires de fonctions respectent ces trois contraintes. Ci-dessous, nous utilisons ces fonctions pour préciser les vitesses relatives auxquelles différents termes convergent par rapport aux tailles de la population N_v et de l'échantillon n_v. Nous utiliserons également la fraction d'échantillonnage définie par $f_v = n_v / N_v$.

Condition 1 : $\lim_{v\to\infty} N_v^{-1}\sum_{i=1}^{N_v} Y_i^4 < \infty$

Condition 2 : $\limsup_{v\to\infty}(N_v \min_{i\in U_v}\pi_{vi})^{-1} = O(n_v^{-1})$

Condition 3 : $\limsup_{v\to\infty}\max_{i,j\in U_v i\neq j}\left|\frac{\pi_{vij}}{\pi_{vi}\pi_{vj}} - 1\right| = O(N_v^{-1})$

Condition 4 : $f_v^{-1} = O(n_v^{1/2}\gamma(n_v)^{-1})$

Condition 5 : $E_p[\delta_{vi}\delta_{vj} \mid Q^{n_v}] = \pi_{vij} \quad \forall i, j \in U_v$

Condition 6 : $\tilde{P}\left(k(n_v)^{-1}\#_{n_v}(B^{n_v}(\mathbf{x})) \geq 1\right) \to_p 1$

Condition 7 : $\| Q^{n_v} \|_l^{\tilde{F}_{n_v}} \to_p 0$ et $\| Q^{n_v} \|_l^{\tilde{F}_{n_v}^-} \to_p 0,$ pour $l = 1, \ldots, d.$

où l'on suppose que toutes les conditions ci-dessus ont une probabilité ξ de 1.

La condition 1 est la seule portant directement sur la répartition du modèle de super-population. Les données n'ont pas à suivre de répartition prédéfinie; il est uniquement nécessaire que la variable dépendante, Y, présente un moment fini de quatrième ordre. Cela rend généralement ces modèles applicables à une vaste catégorie de problèmes. Nous utilisons cette condition au moment de quatrième ordre pour établir la convergence par rapport au plan de l'estimateur arborescent proposé. Toutefois, comme nous l'avons mentionné ci-dessus, cette condition pourrait être réduite pour ne nécessiter qu'un moment fini de deuxième ordre dans le cas d'estimateurs arborescents convergents par rapport au plan.

Les conditions 2 à 4 sont des conditions standard de plan de sondage nécessitant que chaque unité de la population puisse être sélectionnée selon une probabilité minimale, l'effet de rétrécissement de regroupement par rapport à la taille de la population et une légère exigence sur le taux d'échantillonnage (Isaki et Fuller, 1982; Breidt et Opsomer, 2000). La condition 4 est une faible limite applicable à la taille maximale que peuvent atteindre les populations finies par rapport à la taille de l'échantillon afin de permettre une taille d'échantillonnage arbitrairement petite.

La condition 5 est une condition provenant de Toth et Eltinge (2011), nécessitant que les probabilités de sélection soient indépendantes de la partition donnée. L'ensemble de partitionnement Q^{n_ν} est une fonction d'un algorithme appliqué aux données sélectionnées; par conséquent, cette condition sur l'algorithme et le plan de sondage limite l'influence que peut avoir toute unité sélectionnée sur la partition obtenue.

Les conditions 6 et 7 sont des conditions sur l'algorithme de partitionnement. La première exige que le nombre d'observations de chaque boîte de partitionnement augmente à un certain taux par rapport à la taille de l'échantillon, alors que la deuxième exige que les normes l des boîtes de partitionnement, définies par 5 et 6, rétrécissent pour se rapprocher de zéro à mesure que la taille de l'échantillon augmente.

Proposition 2.1 (Toth et Eltinge, 2011)**.** *Soient* $\{(Y_i, \mathbf{X}_i)\}_{i \in U_\nu}$, *une séquence de populations finies, désignées par* ν *et distribuées de manière indépendante et identique à partir du modèle de super-population* ξ *et* S_ν, *un échantillon aléatoire de* U_ν *sélectionné conformément au plan de sondage. Étant donné* Q^{n_ν}, *l'ensemble de boîtes de partitionnement créée à partir de l'algorithme appliqué aux données-échantillon,* S_ν, *définit*

$$\tilde{h}_{n_\nu}(\mathbf{x}) = (\tilde{\#}B^{n_\nu}(\mathbf{x}))^{-1} \sum_{i \in S_\nu} \pi_i^{-1} y_i \mathbb{I}_{\{\mathbf{x}_i \in B^{n_\nu}(\mathbf{x})\}}. \tag{2.6}$$

Si $\lim_{\nu \to \infty} N_\nu^{-1} \sum_{i=1}^{N_\nu} Y_i^2 < \infty$ *et les conditions 2 à 7 sont satisfaites avec une probabilité* ξ *de 1, alors*

$$\lim_{\nu \to \infty} E_{\xi p}\left[\left| \tilde{h}_{n_\nu}(\mathbf{x}) - h(\mathbf{x}) \right|^2 \right] = 0.$$

Il convient de mentionner que la partie droite de l'équation (2.6) est l'estimateur de Hájek de la moyenne de la boîte contenant $\mathbf{x}$ donné par (2.1). Puisque l'ensemble de boîtes partitionne les données, chaque observation correspond exactement à une boîte et la valeur de Y prédite par le modèle pour une observation avec des variables auxiliaires $\mathbf{X} = \mathbf{x}$ est simplement

$$\tilde{h}_{n_\nu}(\mathbf{x}) = \tilde{\mu}(\mathbf{x}). \tag{2.7}$$

Par conséquent, la proposition 2.1 indique qu'un modèle d'arbre de régression estimant la moyenne de Y dans chaque nœud final est un estimateur de la fonction $E_\xi[Y \mid \mathbf{x}]$ convergent par rapport à L^2; donc

$$\tilde{\mu}(\mathbf{x}) \to_{L^2} E_\xi[Y \mid \mathbf{x}]. \tag{2.8}$$

Nous nous appuierons sur ce résultat dans la section suivante pour montrer que l'estimateur proposé de forêt aléatoire est convergent ainsi que le corollaire suivant.

Corollaire 2.1. *Pour un arbre donné* j, *si les conditions 1 à 7 sont satisfaites, alors*

$$\left[\tilde{\#}B_j^{n_\nu}(\mathbf{x}) \right]^{-1} \sum_{i \in S} \pi_i^{-1} y_i^2 \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_\nu}(\mathbf{x})\}} \to_p E_\xi[Y^2 \mid \mathbf{x}].$$

Démonstration en annexe.

3. Modèles de forêt convergents par rapport au plan

Les estimations de modèle de forêt aléatoire sont obtenues en calculant la moyenne des estimations de M modèles arborescents. Cela nécessite l'utilisation d'une procédure permettant de produire plusieurs modèles arborescents *différents* à l'aide des mêmes données; obtenir la moyenne pour le même modèle n'améliore pas l'estimation. Les modèles arborescents utilisés dans la forêt sont généralement obtenus en estimant les modèles arborescents sur des échantillons bootstrap des données d'origine et en utilisant un sous-ensemble aléatoire des variables prédictives de chaque fractionnement (Breiman, 2001). Toutefois, il n'est pas toujours facile ni possible de produire l'échantillon bootstrap d'un ensemble de données pour un plan de sondage informatif général (Mashreghi, Haziza et Léger, 2016). L'approche typique qui consiste à ne pas tenir compte du plan de sondage pendant l'estimation mènera probablement à un modèle de forêt aléatoire biaisé lors de son application sur des données d'enquête.

En d'autres termes, pour un estimateur $\hat{m}_n$ de la fonction de régression m et un point $\mathbf{x}$,

$$\text{Biais}(\hat{m}_n(\mathbf{x})) := \mathbb{E}[\hat{m}_n(\mathbf{x})] - m(\mathbf{x}), \tag{3.1}$$

où l'espérance se rapporte à la répartition conjointe du modèle de super-population et au plan de sondage.

Nous proposons maintenant un modèle arborescent convergent par rapport au plan pour une famille de plans de sondage informatifs, tant que le plan de sondage, la répartition de la super-population et la procédure de partitionnement récursif satisfont aux conditions 1 à 7 de la section 2. Pour obtenir des modèles arborescents de régression différents à partir d'un échantillon donné, à chaque étape du partitionnement récursif, nous sélectionnons la variable entièrement au hasard à partir des d variables prédictives possibles et le point de fractionnement au hasard à partir du support observé de la variable sélectionnée. La figure 3.1 présente cet algorithme.

Figure 3.1 Algorithme de partitionnement récursif pour produire des modèles arborescents aléatoires.

Algorithme de partitionnement récursif	
1.	Soit $n_{\text{end}} = \max(5,\ \text{plancher}\ \{10^{-7} n\})$.
2.	Si l'ensemble de données contient au moins $2n_{\text{end}}$ observations, passer à l'étape suivante; sinon, arrêter.
3.	Parmi les variables auxiliaires $x_l,\ l = 1, \ldots, d_1,$ choisir au hasard une variable pour laquelle nous pouvons fractionner les données.
4.	Fractionner les données en deux ensembles S_L et S_R en sélectionnant au hasard une valeur de la variable sélectionnée x_l pour que chaque sous-ensemble de données contienne au moins n_{end} observations.
5.	Appliquer l'algorithme à partir de l'étape 2 pour chacun des deux sous-ensembles obtenus S_L et S_R.

Note : Il convient de souligner que n_{end} est défini de sorte que les nœuds finaux de chaque arbre satisfont aux conditions 6 et 7.

Il convient de mentionner que n_{end} est défini pour satisfaire à la condition 6, car il est linéaire dans n et donc domine $\sqrt{n}$, mais permet tout de même un nombre relativement réduit d'observations. Cela est important, car, en pratique, un petit nombre d'observations est efficace pour obtenir des estimations exactes.

3.1 Notation étendue pour les forêts

Puisque nous nous intéressons aux forêts, lesquelles nécessitent un ensemble d'arbres, nous étendons une partie de la notation et des fonctions utilisées à la section 2 pour faciliter cet analyse; par exemple, plutôt qu'un ensemble de boîtes de partitionnement, comme v, on suppose une boîte pour chacun des M arbres du modèle. Soit $Q_j^{n_v} = \{B_{j1}^{n_v}, \ldots, B_{jq_j}^{n_v}\}$, la partition pour le j^{e} arbre. Il convient de mentionner que le nombre de boîtes de partitionnement q, créées à l'aide de l'échantillon S_v, peut différer pour les divers arbres du modèle de forêt, et dépend donc de j. Soit $\mathcal{Q}^{n_v} = \{Q_j^{n_v}\}_{j=1}^{M}$, l'ensemble de toutes les partitions constituant le modèle de forêt.

La fonction $B_j^{n_v}(\mathbf{x})$ désigne la boîte dans l'arbre j qui contient la valeur $\mathbf{x}$, alors que $\#B_j^{n_v}(\mathbf{x})$ et $\tilde{\#}B_j^{n_v}(\mathbf{x})$ correspondent au nombre d'unités d'échantillonnage observées et au nombre estimé d'unités de population dans la boîte, respectivement. De la même manière, $\tilde{\mu}_j(\mathbf{x})$ désigne la moyenne estimée des observations dans la boîte contenant la valeur $\mathbf{x}$, définie par l'équation (2.1) pour le j^{e} arbre.

3.2 Facteurs de pondération pour calculer la moyenne des estimations

Notons que l'algorithme, et donc la structure de chaque arbre, dépend uniquement des valeurs observées de variables de modélisation $\{\mathbf{X}_i\}_{i=1}^{n}$, ce qui donne lieu à une estimation du plus proche voisin k de Y en fonction de la proximité d'un sous-échantillonnage aléatoire des variables de modélisation. Le modèle de forêt est alors une moyenne sur M estimations du plus proche voisin k. Toutefois, puisque, les arbres sont construits en fonction de fractionnements sélectionnés aléatoirement, l'homogénéité de Y varie probablement entre les boîtes, ce qui se traduit par des boîtes plus ou moins informatives. Par conséquent, même si la moyenne simple des estimations d'arbres aléatoires fournie par (1.1) donne lieu à un estimateur asymptotiquement sans biais, il sera plutôt inefficace.

Afin d'améliorer l'efficacité de l'estimateur de forêt, nous utilisons une moyenne pondérée, dans le but de fournir davantage de poids aux estimations générées par des modèles arborescents présentant une exactitude prédictive supérieure. Soit $\{\tilde{h}_j^{n_v}\}_{j=1}^{M}$, l'ensemble de modèles arborescents de régression. Un modèle de forêt pondéré prendrait la forme

$$\mathcal{F}_w(\mathbf{x}) = \sum_{j=1}^{M} \lambda_j(\mathbf{x})\, \tilde{h}_j^{n_v}(\mathbf{x}), \tag{3.2}$$

où $\lambda_j(\mathbf{x})$ est un facteur de pondération dépendant du nœud final j auquel $\mathbf{x}$ appartient.

Des méthodes reposant sur une moyenne pondérée d'estimations d'arbres pour produire une estimation de forêt ont été envisagées (Gajowniczek, Grzegorczyk, Ząbkowski et Bajaj, 2020; Shahhosseini et Hu, 2020; Winham, Freimuth et Biernacka, 2013), mais elles nécessitent d'avoir recours à un facteur de pondération fondé sur l'ajustement de chaque arbre uniquement. Lors de la mise à l'essai de plusieurs approches différentes, nous avons constaté qu'utiliser un facteur de pondération dépendant du nœud final produisait les meilleurs résultats. Toutefois, cette approche induit un biais dans les estimations qui nécessite un ajustement.

Pour la méthode que nous proposons, nous pondérons chaque estimation d'arbre à l'aide d'un facteur de pondération inversement proportionnel à l'estimation de un plus la variance du nœud final, $V_{B_j^{n_v}(\mathbf{x})} = \text{Var}_\xi(Y \mid \mathbf{X} \in B_j^{n_v}(\mathbf{x}))$, de manière similaire aux facteurs de pondération obtenus utilisés dans certaines méthodes adaptatives (Williams, Neilley, Koval et McDonald, 2016). Si la moyenne réelle, $\mu_j(\mathbf{x}) = E[Y \mid B_j^{n_v}(\mathbf{x})]$, des valeurs de Y pour les observations dans la boîte $B_j^{n_v}(\mathbf{x})$ est connue, un estimateur de $V_{B_j^{n_v}(\mathbf{x})}$ convergent par rapport au plan est obtenu par

$$\left[\tilde{\#}B_j^{n_v}(\mathbf{x})\right]^{-1} \sum_{i \in S} \pi_i^{-1} (y_i - \mu_j(\mathbf{x}))^2 \, \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_v}(\mathbf{x})\}}.$$

Toutefois, puisque la valeur réelle de $\mu_j(\mathbf{x})$ est inconnue, nous utilisons l'estimateur

$$\tilde{V}_{B_j^{n_v}(\mathbf{x})} = \left[\tilde{\#}B_j^{n_v}(\mathbf{x})\right]^{-1} \sum_{i \in S} \pi_i^{-1} (y_i - \tilde{\mu}_j(\mathbf{x}))^2 \, \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_v}(\mathbf{x})\}}, \tag{3.3}$$

où la valeur estimée $\tilde{\mu}_j(\mathbf{x})$ remplace la moyenne réelle.

Si $\mathbf{x} \in B_j^{n_v}(\mathbf{x})$, alors le facteur de pondération pour l'arbre j dans la forêt est défini par

$$\lambda_j(\mathbf{x}) = \frac{(\tilde{V}_{B_j^{n_v}(\mathbf{x})} + 1)^{-1}}{\sum_{j=1}^{M} (\tilde{V}_{B_j^{n_v}(\mathbf{x})} + 1)^{-1}}. \tag{3.4}$$

de sorte que les facteurs de pondération sont $\lambda_j(\mathbf{x}) \propto (\tilde{V}_{B_j^{n_v}(\mathbf{x})} + 1)^{-1}$ et donnent 1.

La valeur de y estimée par la forêt en fonction de $\mathbf{x}$ est

$$\sum_{j=1}^{M} \lambda_j(\mathbf{x}) \, \tilde{h}_j^{n_v}(\mathbf{x}) = \sum_{j=1}^{M} \lambda_j(\mathbf{x}) \, \tilde{\mu}_j(\mathbf{x}). \tag{3.5}$$

Il convient de mentionner une fois de plus l'équivalence entre l'estimation du j^{e} arbre et la moyenne estimée du nœud final contenant $\mathbf{x}$ pour l'arbre j. Pour un ensemble donné de données-échantillon, chaque $\lambda_j(\mathbf{x})$ et $\tilde{\mu}_j(\mathbf{x})$ sont des fonctions du processus de partitionnement aléatoire; on peut donc les considérer comme M observations indépendantes du vecteur aléatoire, $(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x}))$. Par conséquent, l'expression donnée dans (3.5) peut être considérée comme une somme des produits des composantes de ces M vecteurs aléatoires.

3.3 Estimation du biais à partir de facteurs de pondération

Utiliser cette moyenne pondérée améliore l'efficacité de l'estimateur, mais rend également l'estimateur potentiellement biaisé dans le cadre du plan de sondage. En particulier, si nous étudions l'espérance du modèle de forêt pondérée relativement à la probabilité de sélection et au caractère aléatoire de l'algorithme de partitionnement récursif (figure 3.1) désigné par E_{p^*}, nous obtenons alors

$$\begin{aligned} E_{p^*}\left[\sum_{j=1}^{M} \lambda_j(\mathbf{x}) \, \tilde{h}_j(\mathbf{x})\right] &= \sum_{j=1}^{M} E_{p^*}\left[\lambda_j(\mathbf{x})\right] E_{p^*}\left[\tilde{h}_j(\mathbf{x})\right] + \sum_{j=1}^{M} \text{cov}_{p^*}\left(\lambda_j(\mathbf{x}), \tilde{h}_j(\mathbf{x})\right) \\ &= \tilde{h}^* E_p\left[\sum_{j=1}^{M} \lambda_j(\mathbf{x})\right] + \sum_{j=1}^{M} \text{cov}_{p^*}\left(\lambda_j(\mathbf{x}), \tilde{\mu}_j(\mathbf{x})\right), \end{aligned}$$

où $E_{p^*}[\tilde{h}_j] = \tilde{h}^*$ pour chaque j.

Par conséquent,

$$E_{p^*}\left[\sum_{j=1}^{M} \lambda_j(\mathbf{x})\,\tilde{h}_j(\mathbf{x})\right] = E_\xi[Y \mid \mathbf{x}] + \sum_{j=1}^{M} \mathrm{cov}_{p^*}(\lambda_j(\mathbf{x}), \tilde{\mu}_j(\mathbf{x})),$$

puisque $\tilde{h}^*(\mathbf{x}) \to E_\xi[Y \mid \mathbf{X} = \mathbf{x}] = E_\xi[Y \mid \mathbf{x}]$ selon la proposition 2.1 et $\sum_{j=1}^{M} \lambda_j(\mathbf{x}) = 1$ conformément au plan de sondage.

Puisque chaque $\lambda_j(\mathbf{x})$ et chaque $\tilde{\mu}_j(\mathbf{x})$ sont des observations de variables aléatoires $\lambda(\mathbf{x})$ et $\tilde{\mu}(x)$, pour une valeur donnée $\mathbf{x}$, le terme de biais est

$$\sum_{j=1}^{M} \mathrm{cov}(\lambda_j(\mathbf{x}), \tilde{\mu}_j(\mathbf{x})) = M\mathrm{cov}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x})). \tag{3.6}$$

Afin de corriger ce biais pour un échantillon fixe, nous estimons $\mathrm{cov}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x}))$ à l'aide des M observations d'après

$$\widehat{\mathrm{cov}}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x})) \;=\; (M-1)^{-1} \sum_{j=1}^{M} (\lambda_j(\mathbf{x}) - \bar{\lambda})\,(\tilde{\mu}_j(\mathbf{x}) - \bar{\mu}), \tag{3.7}$$

où $\bar{\lambda} = M^{-1}\sum_{j=1}^{M} \lambda_j(\mathbf{x})$ et $\bar{\mu} = M^{-1}\sum_{j=1}^{M} \tilde{\mu}_j(\mathbf{x})$. Par conséquent, l'estimateur de forêt proposé pour la fonction $h(\mathbf{x})$ est

$$\mathcal{F}_{n_v}(\mathbf{x}) = \sum_{j=1}^{M} \lambda_j(\mathbf{x})\,\tilde{\mu}_j(\mathbf{x}) - M\widehat{\mathrm{cov}}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x})). \tag{3.8}$$

Le résultat suivant est le principal résultat théorique du présent article. Il indique que l'estimateur de forêt proposé, $\mathcal{F}_{n_v}(\mathbf{x})$, défini par l'équation (3.8) est asymptotiquement sans biais par rapport au plan convergeant en probabilité vers $h(\mathbf{x}) = E_\xi[Y \mid \mathbf{X} = \mathbf{x}]$. Ce résultat fournit une justification théorique de l'utilisation de cette méthode d'estimation par forêt aléatoire sur des données recueillies à partir d'un plan de sondage informatif.

Proposition 3.1. *Pour $M > 0$, fixe, si les conditions 1 à 7 sont satisfaites pour chaque arbre de la forêt, alors*

$$\mathcal{F}_{n_v}(\mathbf{x}) \to_p E_\xi[Y \mid \mathbf{x}],$$

pour tous les $\mathbf{x}$ lorsque $n_v \to \infty$.

4. Rendement relatif des estimateurs

Afin de comparer la méthode de forêt aléatoire proposée et la forêt aléatoire i.i.d. typique de Breiman (2001), nous mettons les deux méthodes à l'essai sur des échantillons répétés de deux ensembles de données accessibles au public en utilisant deux plans de sondage différents : l'échantillonnage aléatoire simple (EAS) et l'échantillonnage avec probabilité proportionnelle à la taille (PPT). Nous évaluons l'efficacité et le biais des prédictions de chaque méthode de manière empirique et les comparons à l'estimateur standard de Hájek. L'approche de type forêt aléatoire proposée fournie par l'algorithme de la figure 3.1 a été mise à l'essai en

utilisant l'algorithme disponible dans le paquet R *rpms* (partitionnement récursif pour la modélisation de données d'enquête) (Toth, 2024) et, pour la méthode de forêt aléatoire proposée par Breiman (2001), nous utilisons l'algorithme disponible dans le paquet R *randomForest* (Liaw et Wiener, 2002).

Pour les populations finies, nous utilisons les deux ensembles de données décrites ci-dessous. Dans chaque description d'ensemble de données, nous déterminons également la variable d'intérêt, les variables prédictives et la variable utilisée comme mesure de la taille pour le plan de sondage PPT.

API L'ensemble de données de l'indice de réussite scolaire californien (California Academic Performance Index) disponible dans le paquet *survey* (Lumley, 2020) contient des données sur 5 973 établissements scolaires, notamment le résultat scolaire moyen au test standardisé API ainsi que des données démographiques et administratives sur l'établissement et le quartier desservi. Nous traitons le résultat moyen au test de l'établissement scolaire pour l'année scolaire 2000 comme la variable d'intérêt avec cinq variables prédictives au niveau de l'établissement, notamment : 1) le niveau de scolarité moyen atteint par les parents des élèves de l'établissement; 2) le pourcentage d'élèves apprenant l'anglais; 3) le pourcentage d'élèves inscrits dans un programme de repas subventionné; 4) le pourcentage d'enseignants possédant des qualifications complètes; 5) l'admission ou non de l'établissement à un programme de récompenses. Pour comparer les méthodes sur un plan de sondage PPT informatif, nous procédons à un échantillonnage proportionnel selon la taille des inscriptions dans l'établissement.

CEx Un sous-ensemble de l'enquête sur les dépenses des consommateurs repose sur un fichier de données d'interview du Bureau de la statique du travail des États-Unis disponible dans le paquet *rpms*. Cet ensemble de données contient des renseignements de 2015 sur 45 308 ménages enregistrant des dépenses totales supérieures à 0 $. Nous considérons les dépenses totales des ménages pour le trimestre en cours comme variable d'intérêt avec cinq variables prédictives, notamment : 1) si le ménage vit dans un logement lui appartenant (avec ou sans prêt hypothécaire), loué ou faisant partie d'un logement étudiant; 2) la région où se trouve le ménage; 3) si le ménage vit dans une région urbaine ou non; 4) si un membre du ménage gagne actuellement un salaire; 5) l'âge de la personne désignée comme la principale personne du ménage gagnant un revenu. Pour comparer les méthodes à l'aide d'un plan de sondage PPT informatif, nous procédons à un échantillonnage proportionnel selon la taille (nombre de résidents) du ménage.

En traitant chaque ensemble de données comme une population finie, nous tirons $D = 500$ échantillons répétés de taille n au sein de la population, où $n = 600$ pour API et $n = 1\,000$ pour CEx. Pour chaque échantillon aléatoire, nous ajustons des modèles de forêt aléatoire de 500 arbres aux données-échantillon à l'aide des paramètres par défaut des deux algorithmes et de toutes les variables prédictives disponibles. Les paramètres par défaut nécessitent que chaque nœud final de chaque arbre contienne au moins 5 observations. À l'aide de ces modèles, nous prédisons les valeurs de la variable d'intérêt pour chaque unité de la population finie et les valeurs prédites sont comparées aux valeurs réelles.

En particulier, pour chaque échantillon s_l, $l = 1, \ldots, D$, nous trouvons le modèle estimé $\tilde{h}^{(l)}(\mathbf{x})$ et calculons l'erreur moyenne empirique

$$b_l = \frac{1}{N}\sum_{i=1}^{N}(\tilde{h}^{(l)}(\mathbf{x}_i) - y_i), \tag{4.1}$$

l'erreur relative moyenne empirique, $b_l / \bar{y}$, où $\bar{y} = N^{-1}\sum_{i=1}^{N} y_i$, et l'erreur quadratique moyenne empirique

$$m_l = \frac{1}{N}\sum_{i=1}^{N}(\tilde{h}^{(l)}(\mathbf{x}_i) - y_i)^2. \tag{4.2}$$

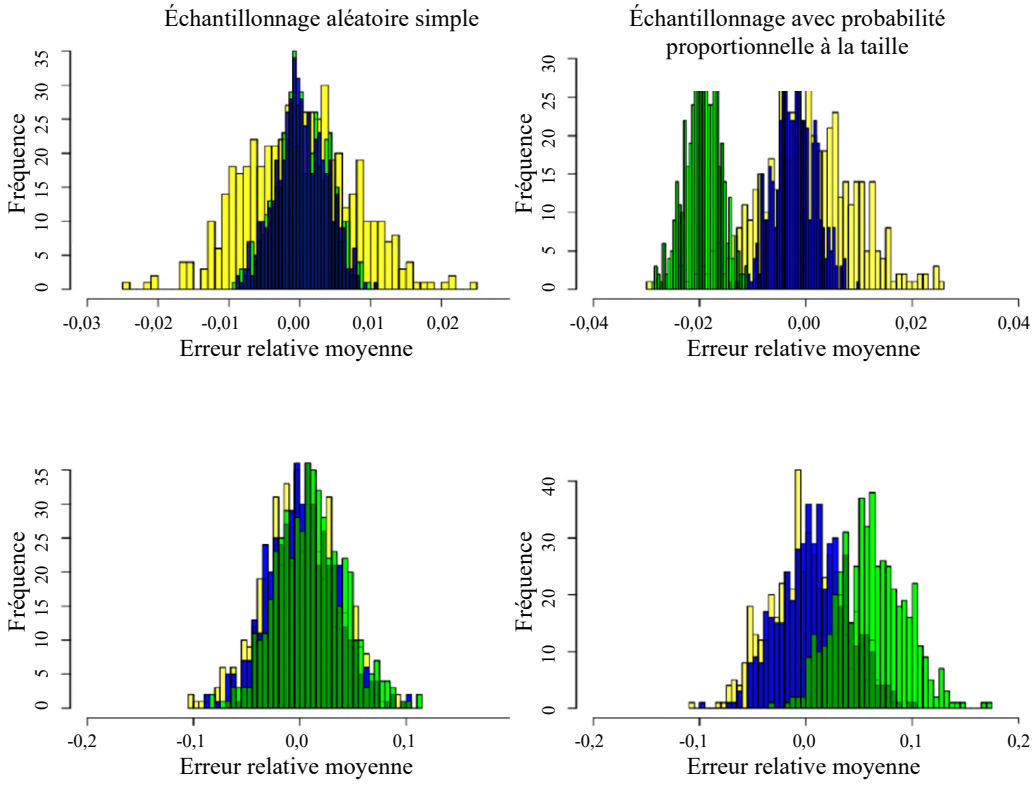
Il convient de souligner que l'erreur moyenne empirique est une estimation du biais moyen, défini par

$$\text{ABiais}(\hat{m}_n(\mathbf{x})) := \mathbb{E}_{\xi}[\mathbb{E}_{\xi p}[\hat{m}_n(\mathbf{x})] - m(\mathbf{x})], \tag{4.3}$$

où la première espérance se rapporte à la répartition conjointe de la population et de l'échantillon et la deuxième se rapporte à la répartition de la population.

Puisque l'un des plus gros risques de ne pas tenir compte d'un plan de sondage informatif lors de la modélisation d'un ensemble de données est l'introduction d'un biais dans le modèle, il est important d'évaluer le biais potentiel de chaque estimateur. La figure 4.1 présente les répartitions empiriques des erreurs moyennes relatives empiriques sur des échantillons répétés pour deux ensembles de données en utilisant deux plans de sondage pour les deux algorithmes de forêt ainsi que l'estimateur de Hájek.

Figure 4.1 Répartition des erreurs relatives moyennes des trois estimateurs sur 500 échantillons répétés provenant de deux ensembles de données.



Note : L'histogramme jaune représente la répartition de l'erreur moyenne relative pour l'estimateur de Hájek; le vert représente la forêt aléatoire non pondérée; le bleu représente la méthode de forêt pondérée. Les diagrammes supérieurs représentent les répartitions des erreurs relatives moyennes à l'aide de l'ensemble de données API et les deux diagrammes inférieurs, à l'aide de l'ensemble de données CEx.

Lorsqu'un plan de sondage EAS est utilisé, les deux répartitions de l'erreur relative présentées du côté gauche de la figure 4.1 montrent que les trois estimateurs produisent des estimations relativement sans biais, leurs erreurs étant centrées très proche de zéro. De plus, la répartition des erreurs des deux modèles de forêt présente une étendue de valeur plus limitée comparativement à l'estimateur de Hájek, ce qui montre que l'utilisation de ces modèles entraîne une efficacité supérieure. Ce gain d'efficacité est particulièrement visible dans les répartitions établies au moyen des données API.

Puisque l'algorithme de forêt aléatoire ne tient pas compte des poids de sondage, on pourrait s'attendre à davantage de biais des estimations des valeurs obtenues en utilisant ce modèle par rapport à celles obtenues au moyen de l'estimateur de Hájek ou le modèle de forêt utilisant l'algorithme et le paquet *rpms*, lorsque le plan de sondage est informatif. Les répartitions représentées des erreurs relatives moyennes sur des échantillons PPT répétés présentées du côté droit de la figure 4.1 confirment cela. Pour des échantillons PPT répétés, l'estimateur de Hájek semble toujours sans biais et la répartition des erreurs relatives est toujours plus étendue que celle des deux modèles de forêt. Même si la répartition des erreurs relatives du modèle RPMS (en bleu) semble centrée proche de zéro, les erreurs relatives du modèle de forêt aléatoire ne tenant pas compte des poids sont centrées près de -2 % pour l'ensemble de données API et autour de 6 % pour l'ensemble de données CEx. Cela laisse entendre que ne pas prendre en considération les facteurs de pondération entraîne un biais bien plus important que la méthode proposée du plan de sondage PPT.

Le tableau 4.1 présente les moyennes des erreurs moyennes relatives et de l'erreur quadratique moyenne $\bar{m} = D^{-1}\sum_{l=1}^{D} m_l$ pour les 500 échantillons aléatoires pour chacun des trois modèles, les deux ensembles de données et les deux plans de sondage. Les statistiques de l'erreur moyenne relative sont présentées en pourcentage, $(\bar{b} / \bar{y})\,100$ %, et celles de l'erreur quadratique moyenne sont fournies par rapport à celles de l'estimateur de Hájek, $\bar{m} / \bar{m}_H$, où $\bar{m}_H$ désigne la moyenne des erreurs quadratiques moyennes de l'estimateur de Hájek pour les 500 échantillons.

Tableau 4.1
Moyennes pour 500 échantillons aléatoires comparant l'erreur de prédiction à l'aide de l'estimateur de Hájek et les deux méthodes de forêt aléatoire sur deux ensembles de données et pour deux plans de sondage.

	Indice de réussite scolaire (API) $N = 5\,973$ $n = 600$				**Enquête sur les dépenses des consommateurs (CEx)** $N = 45\,308$ $n = 1\,000$			
	Erreur relative en %		**REQM**		**Erreur relative en %**		**REQM**	
Méthode	**EAS**	**PPT**	**EAS**	**PPT**	**EAS**	**PPT**	**EAS**	**PPT**
Hájek	-0,010	-0,007	1,000	1,000	-0,206	0,066	1,000	1,000
Forêt aléatoire	0,043	-1,945	0,209	0,218	0,806	6,111	0,862	0,865
RPMS	0,021	-0,210	0,204	0,204	-0,056	0,878	0,844	0,844

Note : L'erreur relative en pourcentage est l'erreur moyenne des valeurs estimées pour la population entière par rapport à la moyenne pour la population de la variable d'intérêt, multipliée par 100. La racine carrée de l'erreur quadratique moyenne (REQM) relative est la moyenne pour les 500 échantillons de l'erreur quadratique moyenne calculée des valeurs estimées pour la population complète, par rapport à celle de l'estimateur de Hájek.
API = Academic Performance Index; CEx = Consumer Expenditure Survey; PPT = Probabilité proportionnelle à la taille; RPMS = Recursive Partitioning for Modeling Survey Data; EAS = Échantillons aléatoires simples.

On peut constater dans les résultats du tableau 4.1 que l'erreur quadratique moyenne relative des deux estimateurs obtenue en utilisant les méthodes de modélisation de type forêt est inférieure à celle de

l'estimateur de Hájek pour les deux ensembles de données pour les plans de sondage EAS et PPT. Toutefois, la procédure de forêt aléatoire qui ne tient pas compte des poids de sondage produit des estimations biaisées dans le cadre du plan de sondage PPT pour les deux ensembles de données, alors que les deux procédures proposées de modélisation de type forêt aléatoire et l'estimateur de Hájek fournissent des estimations relativement sans biais.

4.1 Démonstration de la convergence

Jusqu'à présent, nous avons examiné l'efficacité et le biais de notre estimateur de forêt par rapport à l'algorithme i.i.d. habituel de forêt aléatoire et l'estimateur moyen standard de Hájek, afin de nous concentrer sur la différence entre les estimations du modèle et la réelle valeur de la variable d'intérêt y à l'aide de deux ensembles de données réels comme population finie. Un estimateur $\tilde{h}(\mathbf{x})$ est convergent si

$$E_{\xi p}[(\tilde{h}(\mathbf{x}) - E_{\xi}[Y \mid \mathbf{x}])^2] \to 0 \text{ lorsque } \nu \to \infty, \tag{4.4}$$

ce qui nécessite de connaître la réelle fonction moyenne $h(\mathbf{x}) = E_{\xi}[Y \mid \mathbf{x}]$ et laisser la taille de l'échantillon et de la population aller vers l'infini ∞. Lors de l'utilisation d'un ensemble de données réel comme population finie, nous ne connaissons pas $h(\mathbf{x})$. Par conséquent, pour étudier la convergence, nous utilisons des données simulées $(Y, \mathbf{X})_{i=1}^{N}$, dont les valeurs sont obtenues par tirages aléatoires à partir d'une répartition connue, et nous étudions le comportement des estimateurs pour une séquence de tailles d'échantillon.

Pour chaque observation générée aléatoirement i, nous générons un vecteur aléatoire

$$\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3}, V_{i1}, V_{i2}, V_{i3})$$

de 6 variables aléatoires indépendantes. Les variables X_{i1} et X_{i2} suivent une répartition uniforme $U(-10, 20)$ et $X_{i3} \sim U(-100, 200)$, alors que V_{i1} jusqu'à V_{i3} sont des variables aléatoires catégoriques ayant une probabilité égale parmi les catégories. V_{i1} et V_{i2} correspondent à une parmi 5 catégories et V_{i3} correspond à une parmi 14 catégories. Il s'agit des variables auxiliaires à la disposition de l'analyste pour chaque unité de la population et qui peuvent être utilisées dans le modèle pour la variable d'intérêt Y.

Puisque l'on sait que les forêts aléatoires sont des modèles non paramétriques très souples, plutôt que de mettre cette méthode à l'essai sur un ensemble de modèles paramétriques standard, nous présentons les résultats pour des données qui suivent le modèle moyen $y = \mu(x) + \epsilon$ où

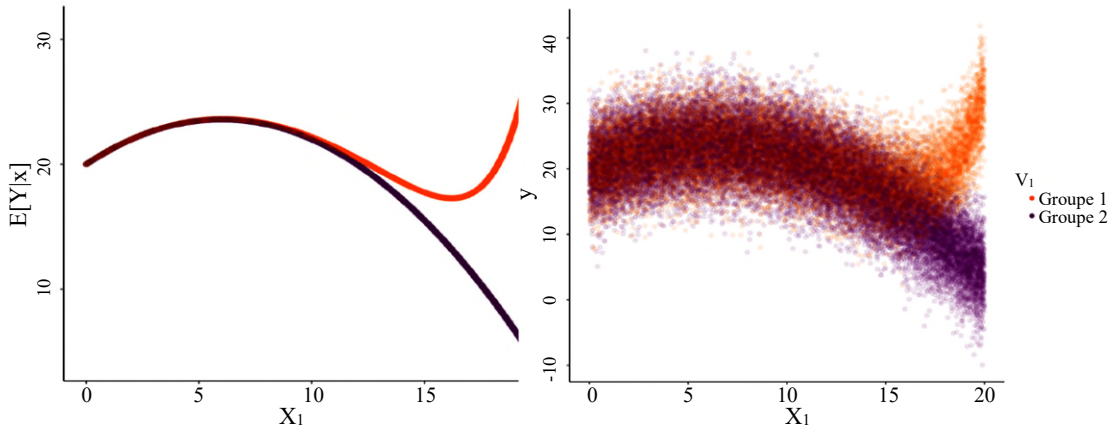
$$\mu(\mathbf{x}) = 0{,}2X_1(X_1 - 12) + 0{,}5\exp\{(X_1 - 15)\}\,\mathbb{I}_{\{V_1 \in \{A, B\}\}}.$$

Le côté gauche de la figure 4.2 présente la fonction moyenne, $\mu(\mathbf{x})$, et le côté droit présente un diagramme des valeurs pour la population générées aléatoirement à partir du modèle.

Nous générons également une variable Z, que nous utilisons comme variable de taille afin de tester les méthodes pour un plan de sondage PPT. Les valeurs de la variable de taille Z sont générées indépendamment à partir du modèle $Z = \frac{1}{2}\mu(\mathbf{x}) + 5\eta$, où η a une répartition khi carré à 5 degrés de liberté. La corrélation entre la variable de taille et Y est de 0,663; par conséquent, dans cet exemple, le plan de sondage PPT est informatif.

Pour cette simulation, nous générons des populations aléatoires finies ayant 1, 2, 4, 8, 16 et 32 milliers d'unités. Nous tirons ensuite 500 échantillons répétés de chacune de ces six populations finies. Pour chaque échantillon aléatoire, nous échantillonnons 5 % des unités, ce qui correspond à 50, 100, 200, 400, 800 et 1 600 observations, respectivement. Nous utilisons une fois de plus les données-échantillon pour estimer le modèle de forêt, nous utilisons ce modèle pour prédire les valeurs de Y pour les unités non échantillonnées selon les valeurs de $\mathbf{X}$ dans la population, puis nous utilisons ces valeurs pour estimer la moyenne pour la population.

Figure 4.2 Valeurs de $E[Y \mid \mathbf{x}]$ par rapport à la variable X_1.



Note : La couleur indique les valeurs des deux groupes d'observations en fonction de la valeur de la variable catégorique V_1. Le groupe 1 comprend des observations où $V_1 \in \{A, B\}$ et le groupe 2, toutes les autres observations.

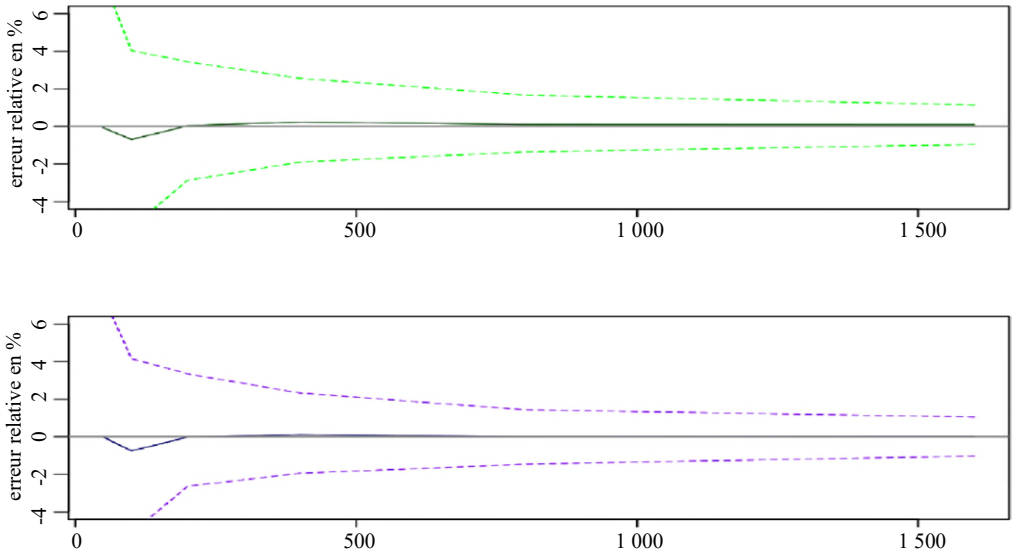
En tirant des échantillons aléatoires de taille croissante et en comparant notre estimateur à la fonction moyenne réelle, nous évaluons le comportement de l'exactitude et la variance de notre estimateur par rapport à la taille de l'échantillon; par exemple, pour un estimateur convergent, l'intervalle de confiance empirique des différences moyennes entre les valeurs estimées et les valeurs générées à partir de la fonction moyenne réelle devrait contenir zéro. De plus, à mesure que la taille d'échantillon augmente, la variance de la différence moyenne devrait diminuer pour s'approcher de zéro.

Nous pouvons observer cela pour le plan de sondage EAS à la figure 4.3 qui présente la moyenne pour les 250 échantillons des erreurs relatives moyennes des valeurs de population estimées à mesure que la taille de l'échantillon augmente pour passer de 50 à 1 600. La répartition des erreurs moyennes pour les 500 échantillons répétés est centrée autour de 0 pour toutes les tailles d'échantillons pour les deux méthodes de modélisation de type forêt. De plus, la variance des erreurs moyennes se rapproche de zéro à mesure que la taille de l'échantillon augmente à peu près au même rythme pour les deux méthodes de modélisation.

Comme nous pouvons nous y attendre, cela devient plus intéressant lorsque l'échantillon est tiré à l'aide d'un plan de sondage PPT et que la variable de taille est liée à la variable d'intérêt. La figure 4.4 montre que les moyennes des erreurs relatives en pourcentage pour les 250 échantillons répétés sont plus longues près de zéro pour les deux méthodes. Le zéro fait partie du milieu des 95 % des valeurs des erreurs relatives en pourcentage lorsque la taille de l'échantillon est inférieure à 800 pour les deux méthodes, du fait des importantes variances dans ces valeurs pour de petites tailles d'échantillon. Cet intervalle ne contient plus

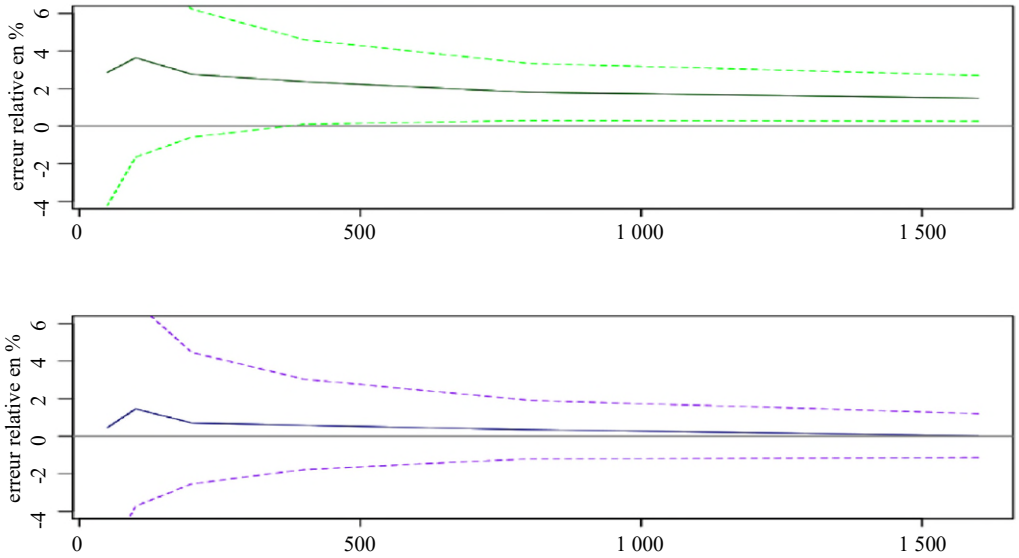
zéro pour l'algorithme de forêt aléatoire standard, car les écarts d'erreurs diminuent lorsque les tailles d'échantillon augmentent. Toutefois, la méthode de type forêt aléatoire proposée contient zéro pour toutes les tailles d'échantillon pour la méthode proposée.

Figure 4.3 Erreur relative en pourcentage selon la taille de l'échantillon pour les algorithmes de forêt aléatoire habituels (en haut) et convergent par rapport au plan (en bas) pour des échantillons aléatoires simples répétés.



Note : La ligne continue correspond à l'erreur relative moyenne en pourcentage pour tous les échantillons, alors que les lignes pointillées représentent les valeurs des 2,5 et 97,5 centiles.

Figure 4.4 Erreur relative en pourcentage selon la taille de l'échantillon pour les algorithmes de forêt aléatoire habituels (en haut) et convergent par rapport au plan (en bas) pour des échantillons par probabilité proportionnelle à la taille répétés.



Note : La ligne continue correspond à l'erreur relative moyenne en pourcentage pour tous les échantillons, alors que les lignes pointillées représentent les valeurs des 2,5 et 97,5 centiles.

Ces résultats de simulation confirment la principale conclusion de la présente étude; en d'autres termes, l'algorithme proposé satisfaisant à certaines conditions, dont on a démontré théoriquement qu'il est asymptotiquement sans biais et convergent par rapport au plan, a prouvé pour des échantillons répétés qu'il converge vers la moyenne réelle et est relativement sans biais.

5. Conclusions

Traditionnellement, des données d'enquête complexes sont recueillies pour estimer des quantités de population finie. Toutefois, l'avènement des méthodes par apprentissage automatique accroît l'intérêt porté à l'emploi de données d'enquête dans des problèmes prédictifs; par conséquent l'adaptation de méthodes d'apprentissage automatique pour gérer des données de probabilité inégale est désormais un domaine de recherche dynamique. Dans la présente étude, nous présentons un nouvel algorithme permettant d'estimer des modèles de forêt aléatoire. Cette méthode qui repose sur des arbres aléatoires indépendants et une procédure de pondération fondée sur la variabilité pondérée des valeurs de y est plus adaptée à des données d'enquête et à d'autres données recueillies à partir d'un plan de sondage informatif. Nous présentons un ensemble de conditions selon lesquelles nous montrons que cette méthode est convergente par rapport au plan pour l'espérance conditionnelle de la variable d'intérêt. L'absence de biais asymptotique et la convergence théoriques de cet algorithme sont démontrées par une simulation. Des études de simulation sont effectuées en utilisant des données réelles et générées; nous montrons qu'en pratique, la méthode proposée réduit considérablement le biais d'un algorithme de forêt aléatoire dans le cadre de plans de sondage informatifs. En revanche, les estimations de la méthode habituelle de forêt aléatoire, laquelle ne tient pas compte du plan de sondage, sont sans biais pour des échantillons informatifs répétés. Les estimations des deux méthodes ont présenté des erreurs quadratiques moyennes relativement similaires. Pour assurer l'indépendance des arbres individuels, notre algorithme construit des arbres réellement aléatoires dans le cadre desquels une variable aléatoire et une valeur seuil aléatoire sont sélectionnées pour chaque fractionnement.

L'approche de Nalenz et coll. (2024) d'ajustement d'un plan de sondage informatif est simple et intéressante, mais elle a vu le jour à l'étape de révision du présent article; nous ne l'avons donc pas comparée à notre méthode d'arbres aléatoires. Plutôt que d'éviter un échantillonnage bootstrap, Nalenz et coll. (2024) utilisent une méthode bootstrap de Hájek. Dans leur application, cela fonctionne très bien puisque les valeurs aberrantes se trouvent dans la partie suréchantillonnée de la population (unités ayant de faibles facteurs de pondération) et donc la pondération de ces unités est réduite par l'algorithme. Toutefois, dans un ensemble de données général, les valeurs aberrantes peuvent être également associées à des unités d'enquête à pondération élevée. De futurs travaux devraient comparer ces méthodes et, dans l'esprit de l'agrégation de modèles, envisager de combiner les deux approches.

Remerciements

Les auteurs souhaitent remercier un grand nombre de personnes ultérieurement.

Annexe

Démonstrations et résultats mineurs

Démonstration du corollaire 2.1. La condition 1 exige que la variable Y présente un moment fini de quatrième ordre, alors la variable aléatoire Y^2 présente un moment fini de deuxième ordre. Par conséquent, en choisissant Y^2 comme variable d'intérêt dans la proposition 2.1, nous obtenons un estimateur de $E_\xi[Y^2 \mid \mathbf{x}]$ convergent.

Lemme 6.1. *Pour un arbre donné j, si les conditions 1 à 7 sont satisfaites, alors*

$$\tilde{V}_{B_j^{n_v}(\mathbf{x})} \to_p \mathrm{Var}(Y \mid \mathbf{x}) < \infty, \text{ lorsque } \nu \to \infty.$$

Démonstration du lemme 6.1.

$$\begin{aligned}
\tilde{V}_{B_j^{n_v}(\mathbf{x})} &= \left(\tilde{\#}B_j^{n_v}(\mathbf{x})\right)^{-1} \sum_{i\in S} \pi_i^{-1}\left(y_i - \tilde{\mu}_j(\mathbf{x})\right)^2 \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_v}(\mathbf{x})\}} \\
&= \left(\tilde{\#}B_j^{n_v}(\mathbf{x})\right)^{-1} \sum_{i\in S} \pi_i^{-1}\left(y_i^2 - 2y_i\tilde{\mu}_j(\mathbf{x}) + \tilde{\mu}_j^2(\mathbf{x})\right) \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_v}(\mathbf{x})\}} \\
&= \left(\tilde{\#}B_j^{n_v}(\mathbf{x})\right)^{-1} \sum_{i\in S} \pi_i^{-1} y_i^2 \, \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_v}(\mathbf{x})\}} \\
&\quad - 2\tilde{\mu}_j(\mathbf{x})\left(\tilde{\#}B_j^{n_v}(\mathbf{x})\right)^{-1} \sum_{i\in S} \pi_i^{-1} y_i \, \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_v}(\mathbf{x})\}} + \tilde{\mu}_j^2(\mathbf{x}) \\
&= \underbrace{\left(\tilde{\#}B_j^{n_v}(\mathbf{x})\right)^{-1} \sum_{i\in S} \pi_i^{-1} y_i^2 \, \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_v}(\mathbf{x})\}}}_{I} \quad - \quad \underbrace{\tilde{\mu}_j^2(\mathbf{x})}_{II}
\end{aligned}$$

selon l'équation (2.1).

Selon la proposition 2.1, $\tilde{\mu}_j(\mathbf{x}) \to_p E_\xi[Y \mid \mathbf{x}]$, donc le terme $II \to_p E_\xi^2[Y \mid \mathbf{x}]$ pour chaque $j = 1 \ldots M$ et selon le lemme 2.1,

$$I = \left(\tilde{\#}B_j^{n_v}(\mathbf{x})\right)^{-1} \sum_{i\in S} \pi_i^{-1} y_i^2 \, \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_v}(\mathbf{x})\}} \to_p E_\xi[Y^2 \mid \mathbf{x}],$$

pour chaque $j = 1 \ldots M$. Par conséquent, $\tilde{V}_{B_j^{n_v}(\mathbf{x})} \to_p \mathrm{Var}(Y \mid \mathbf{x})$, et selon la condition 1, les deux quantités I et II sont finies avec une probabilité ξ de 1.

Lemme 6.2. *Si les conditions 1 à 7 sont satisfaites pour chaque arbre d'une forêt donnée de $M > 0$ arbres, alors*

$$\lambda_j(\mathbf{x}) \to_p \frac{1}{M}$$

pour tous les $\mathbf{x}$ *et tous les* $j = 1 \ldots M$, *lorsque* $\nu \to \infty$.

Démonstration du lemme 6.2. Puisque, pour chaque $j = 1, \ldots, M$, la variable aléatoire $\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} \geq 0$, selon l'équation (3.3), la fonction

$$\lambda_j(\mathbf{x}) = \frac{(\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} + 1)^{-1}}{\sum_{j=1}^{M} (\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} + 1)^{-1}}$$

est continue pour chaque j. De plus, selon le lemme 6.1, $\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} \to_p \mathrm{Var}(Y \mid \mathbf{x})$ pour chaque j. Par conséquent, selon le théorème de correspondance continue,

$$\lambda_j(\mathbf{x}) \to_p \frac{(\mathrm{Var}(Y \mid \mathbf{x})) + 1)^{-1}}{\sum_{j=1}^{M} (\mathrm{Var}(Y \mid \mathbf{x}) + 1)^{-1}} = \frac{1}{M}.$$

Ce lemme indique que du fait de la convergence de la variance vers $V[Y \mid \mathbf{x}]$, tous les facteurs de pondération des nœuds finaux convergent asymptotiquement vers $1/M$. Toutefois, il convient de mentionner que pour une valeur n finie, les facteurs de pondération diffèrent de façon substantielle selon le gain d'efficacité des fractionnements aléatoires donnant le nœud final. Il s'agit du caractère adaptatif ajouté à la procédure et dans le cadre duquel tous les travaux réel se déroulent.

Lemme 6.3. *Pour* $M > 0$ *fixe, si les conditions 1 à 7 sont satisfaites pour chaque arbre de la forêt de* M *arbres, alors*

$$M\widehat{\mathrm{cov}}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x})) \to_p 0,$$

pour tous les $\mathbf{x}$ *lorsque* $\nu \to \infty$.

Démonstration du lemme 6.3. Puisque $\lambda_j(\mathbf{x}) \to_p \frac{1}{M}$ selon le lemme 6.2, $\bar{\lambda} = M^{-1} \sum_{j=1}^{M} \lambda_j(\mathbf{x}) \to_p \frac{1}{M}$ selon le théorème de correspondance continue. De la même manière, $\bar{\mu} = M^{-1} \sum_{j=1}^{M} \tilde{\mu}_j(\mathbf{x}) \to_p E[Y \mid \mathbf{x}]$ puisque chaque $\tilde{\mu}_j(\mathbf{x}) \to_p E[Y \mid \mathbf{x}]$ selon la proposition 2.1. Par conséquent, chacun des termes de l'équation (3.7),

$$(\lambda_j(\mathbf{x}) - \bar{\lambda})(\tilde{\mu}_j(\mathbf{x}) - \bar{\mu}) \to_p 0.$$

Une fois de plus, appliquer le théorème de correspondance continue à

$$M\widehat{\mathrm{cov}}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x})) = \frac{1}{(1 - M^{-1})} \sum_{j=1}^{M} (\lambda_j(\mathbf{x}) - \bar{\lambda})(\tilde{\mu}_j(\mathbf{x}) - \bar{\mu}),$$

complète la démonstration du lemme 6.3.

Démonstration de la proposition 3.1. Puisque chaque $\lambda_j(\mathbf{x}) \to_p \frac{1}{M}$ selon le lemme 6.2, $\tilde{\mu}_j(\mathbf{x}) \to_p E_\xi[Y \mid \mathbf{x}]$ selon la proposition 2.1 et $M\widehat{\mathrm{cov}}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x})) \to_p 0$, selon le lemme 6.3, nous pouvons appliquer le théorème de correspondance continue pour obtenir

$$\mathcal{F}_{n_v}(\mathbf{x}) = \sum_{j=1}^{M} \lambda_j(\mathbf{x})\, \tilde{\mu}_j(\mathbf{x}) - M\widehat{\mathrm{cov}}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x})) \to_p \sum_{j=1}^{M} \frac{1}{M} E_\xi[Y \mid \mathbf{x}] + 0 = E_\xi[Y \mid \mathbf{x}].$$

Bibliographie

Arlot, S., et Genuer, R. (2014). Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*.

Biau, G., Devroye, L. et Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(9).

Bilton, P., Jones, G., Ganesh, S. et Haslett, S. (2017). Classification trees for poverty mapping. *Computational Statistics & Data Analysis*, 115, 53-66.

Breidt, F.J., et Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28, 1026-1053.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Buskirk, T.D. (2018). Surveying the forests and sampling the trees: An overview of classification and regression trees and random forests with applications in survey research. *Surv Pract*, 11, 2709.

Dagdoug, M., Goga, C. et Haziza, D. (2021). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 1-18.

Earp, M., Toth, D. Phipps, P. et Oslund, C. (2018). Assessing nonresponse in a longitudinal establishment survey using regression trees. *Journal of Official Statistics*, 34(2), 463-481.

Gajowniczek, K., Grzegorczyk, I., Ząbkowski, T. et Bajaj, C. (2020). Weighted random forests to improve arrhythmia classification. *Electronics*, 9(1), 99.

Gelman, A., King, G. et Liu, C. (1998). Not asked and not answered: Multiple imputation for multiple surveys. *Journal of the American Statistical Association*, 93(443), 846-857.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science*, 5, 361-74.

Hong, H.G., et He, X. (2010). Prediction of functional status for the elderly based on a new ordinal regression model. *Journal of the American Statistical Association*, 105(491), 930-941.

Hothorn, T., Hornik, K. et Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 651-674.

Isaki, C.T., et Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Krebs, M.A., Reeves, M.C. et Baggett, L.S. (2019). Predicting understory vegetation structure in selected western forests of the United States using fia inventory data. *Forest Ecology and Management*, 448, 509-527.

Kshirsagar, V., Wieczorek, J., Ramanathan, S. et Wells, R. (2017). Household poverty classification in data-scarce environments: A machine learning approach. *arXiv preprint arXiv:1711.06813*, 2017.

Lavallée, P., et Beaumont, J.-F. (2015). Why we should put some weight on weights. *Survey Methods: Insights from the Field (SMIF)*.

Liaw, A., et Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3), 18-22. URL: https//CRAN.R-project.org/doc/Rnews/.

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

Loh, W.-Y. (2008). Classification and regression tree methods. *Encyclopedia of Statistics in Quality and Reliability*, 1, 315-323.

Lumley, T. (2020). survey: analysis of complex survey samples, 2020. R package version 4.0.

Mashreghi, Z., Haziza, D. et Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10, 1-52.

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.

Morgan, J.N., et Sonquist, J.A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302), 415-434.

Nalenz, M., Rodemann, J. et Augustin, T. (2024). Learning de-biased regression trees and forests from complex samples. *Machine Learning*, 113(6), 3379-3398.

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

Phipps, P., et Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *The Annals of Applied Statistics*, 772-794.

Scornet, E. (2016). On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146, 72-83.

Shahhosseini, M., et Hu, G. (2020). Improved weighted random forest for classification problems. *International Online Conference on Intelligent Decision Science*, 42-56. Springer.

Toth, D. (2024). *rpms: Recursive Partitioning for Modeling Survey Data*. R package version 1.0.0.

Toth, D., et Eltinge, J. (2011). Building consistent regression trees from complex sample data. *Journal of the American Statistical Association*, 106, 1626-1636.

Wieczorek, J. (2023). Prédiction conforme fondée sur le plan. *Techniques d'enquête*, 49, 2, 477-512. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2023002/article/00007-fra.pdf.

Williams, J.K., Neilley, P.P., Koval, J.P. et McDonald, J. (2016). Adaptable regression method for ensemble consensus forecasting. *Thirtieth AAAI Conference on Artificial Intelligence*.

Winham, S.J., Freimuth, R.R. et Biernacka, J.M. (2013). A weighted random forests approach to improve predictive performance. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(6), 496-505.

Yang, D.K., et Toth, D.S. (2022). Analyzing the association of objective burden measures to perceived burden with regression trees. *Journal of Official Statistics*, 38(4), 1125-1144.

Techniques d'enquête, décembre 2024
Vol. 50, nº 2, pp. 231-258
Statistique Canada, nº 12-001-X au catalogue

Échantillonnage en grappes adaptatif, une approche quasi bayésienne

Glen Meeden et Muhammad Nouman Qureshi[1]

Résumé

Des plans d'échantillonnage en grappes adaptatif ont été proposés comme méthode d'échantillonnage de populations rares dont les unités tendent à apparaître en grappes. L'estimateur résultant n'est basé sur aucune hypothèse de modèle et il est sans biais par rapport au plan. Sa variance peut être plus petite que celle de l'estimateur classique qui ne tient pas compte du fait que l'on a affaire à une population rare. Dans le présent article, nous démontrerons que, lorsque l'échantillonnage en grappes adaptatif est approprié, son estimateur ne tient pas compte de tous les renseignements disponibles dans le plan. Nous présentons une approche quasi bayésienne qui intègre les informations qui sont présentement ignorées. Nous verrons que l'estimateur résultant constitue une amélioration significative par rapport aux méthodes actuelles.

Mots-clés : Échantillonnage de population finie; échantillonnage en grappes adaptatif; inférence bayésienne; information *a priori*.

1. Introduction

Observons le problème de l'estimation du nombre total d'une espèce végétale ou animale qui vit dans une région géographique donnée qui a été divisée en une collection de carrés de dimension égale. De plus, présumons que l'espèce d'intérêt est rare dans cette région de sorte que la majorité des carrés ne comportera pas d'espèce. Également, supposons que les quelques carrés qui contiennent l'espèce ont tendance à se regrouper dans quelques quartiers de carrés adjacents.

Pour analyser ce problème, Thompson (1990) a proposé la notion d'échantillonnage en grappes adaptatif (EGA). Un premier échantillon aléatoire simple des carrés est pris et le nombre de l'espèce dans chaque carré sélectionné est observé. Pour la majorité des carrés observés, le compte sera généralement nul. Cependant, chaque fois que le compte dans un carré est supérieur à zéro, les carrés adjacents, qui se trouvent à gauche, à droite, au-dessus et en dessous, sont ajoutés à l'échantillon. Lorsqu'un de ces carrés a un compte supérieur à zéro, alors tous ses carrés adjacents non observés sont également observés. Le processus se poursuit jusqu'à ce que nous obtenions un ensemble de carrés non vides contigus, entouré par des carrés vides. Un ensemble de carrés non vides contigus est appelé un réseau et les carrés vides qui l'entourent sont ses arêtes. Par définition, un carré vide est un réseau de dimension un. Pour ce plan d'échantillonnage en grappes adaptatif, l'estimateur habituel de la population totale en fonction des comptes dans tous les carrés observés aura un biais vers le haut. Thompson (1990) a développé un estimateur sans biais pour le total de la population, ainsi qu'un estimateur de sa variance. Des renseignements supplémentaires ainsi que des références peuvent être obtenus dans Thompson (2012). Bon nombre de chercheurs du domaine ont adopté diverses versions de l'échantillonnage en grappes adaptatif, et cette méthode est utilisée dans une variété de

1. Glen Meeden, professeur émérite, School of Statistics, University of Minnesota, Minneapolis, MN 55455. Courriel : gmeeden@umn.edu; Muhammad Nouman Qureshi, School of Statistics, University of Minnesota, Minneapolis, MN 55455. Courriel : qures089@umn.edu.

disciplines. Turk et Borkowski (2005) en présentent plusieurs exemples. Latpate, Kshirsagar, Gupta et Chandra (2021) abordent certaines modifications de l'approche par EGA classique.

Dans l'approche bayésienne à l'échantillonnage, les renseignements *a priori* sur la population d'intérêt sont incorporés dans une distribution *a priori*. Après l'observation des unités dans un échantillon, les inférences à propos de la population sont fondées sur la distribution *a posteriori* des unités non observées, en tenant compte des unités observées. De plus, cette distribution *a posteriori* ne dépend pas de la méthode de sélection des unités de l'échantillon. Cette approche a été décrite en détail par Basu (Ghosh, 1988). Trois approches bayésiennes d'échantillonnage en grappes adaptatif sont présentées dans Rapley et Welsh (2008), Pacifici, Reich, Dorazio et Conroy (2016), et Goncalves et Moura (2016). Dans ces approches, les auteurs construisent un modèle bayésien de populations possibles cohérent avec les hypothèses qui sous-tendent l'échantillonnage en grappes adaptatif. Nolau, Goncalves et Pereira (2022) proposent un modèle bayésien qui comprend des variables auxiliaires qui pourraient comporter des renseignements supplémentaires sur le compte dans un carré.

Pour un échantillon par EGA donné, notre objectif est de trouver un estimateur ponctuel et une borne supérieure du nombre total de l'espèce dans la population qui ont de bonnes propriétés fréquentistes. Puisque nous évaluons des espèces rares, nous présumerons que le rapport entre le nombre d'unités ayant un compte supérieur à zéro et le nombre total d'unités dans la population est petit. Nous scinderons le problème en deux parties. Premièrement, nous précisons une distribution *a priori* du nombre de carrés ou d'unités dans la population où le compte est supérieur à zéro, soit θ, un paramètre inconnu. Cette valeur *a priori* reflétera notre hypothèse du petit nombre d'unités. Compte tenu de l'exemple et de notre valeur *a priori*, nous avons un *a posteriori* pour θ. Ainsi, notre première étape consiste à simuler une possible valeur pour θ, soit $\hat{\theta}$. Puis, en fonction de $\hat{\theta}$, nous obtenons une estimation du total de tous les comptes supérieurs à zéro en utilisant une distribution qui s'appuie sur l'échangeabilité entre les comptes observés et non observés qui sont supérieurs à zéro. Cette distribution ne découle pas d'une distribution *a priori*, mais est plutôt précisée après l'observation de l'EGA. Cette procédure ne correspond pas à la procédure bayésienne classique puisque cette deuxième distribution « *a posteriori* » ne découle pas de quelque distribution *a priori* définie pour la population finie inconnue. Nous expliquons ainsi le terme « quasi bayésienne » de notre titre. Deux autres exemples récents dans lesquels les inférences reposent sur des pseudo distributions *a posteriori* sont présentés dans Si, Pliiai et Gelman (2015), et Savitsky et Toth (2014). Ce choix est sensé lorsqu'il y a des renseignements dans le plan d'échantillonnage qui ne peuvent pas être incorporés dans une distribution *a priori*. Toutefois, nous avons ensuite combiné ces deux distributions pour simuler des copies complètes de la population inconnue. Nous verrons que les estimateurs obtenus, par point et par intervalle, du total de la population ont de meilleures propriétés fréquentistes que les estimateurs par EGA classiques.

Dans la section 2, nous passons brièvement en revue l'approche par échantillonnage en grappes adaptatif et décrivons notre façon d'aborder le problème. Dans la section 3, nous expliquons notre approche en détail et présentons nos estimateurs. Nous avons développé notre estimateur en faisant des simulations sur un ensemble de six populations pour lesquelles l'EGA serait approprié. Dans la section 4, nous décrivons ces

six populations. Dans la section 5, nous présentons des simulations pour comparer notre approche à l'approche par EGA classique. Cette méthode est appliquée aux six populations de la section 4 et à six nouvelles populations qui n'ont pas été utilisées dans l'élaboration de notre méthode. Dans la section 6, nous examinons de possibles extensions qui s'appliqueraient si davantage de renseignements *a priori* étaient disponibles à propos de la population d'intérêt. Dans la section 7, nous présentons quelques observations finales.

2. Échantillonnage en grappes adaptatif

2.1 Les éléments fondamentaux

Nous commençons en présentant une notation. Nous présumons que la population inconnue correspond à une région rectangulaire composée de N_r par N_c carrés ou unités. Par conséquent, $N = N_r \times N_c$ est la taille de la population. Pour les entiers (i, j), où $1 \le i \le N_r$ et $1 \le j \le N_c$, supposons que $y_{i,j}$ indique le nombre de l'espèce dans le carré $(i^{\text{e}}, j^{\text{e}})$. Soulignons que $y_{i,j}$ est un entier non négatif. Supposons que Y désigne la matrice des valeurs $y_{i,j}$. Pour un carré donné, les voisins sont les carrés qui se trouvent juste au-dessus et en dessous et les carrés qui se trouvent juste à la droite et à la gauche de ce carré et des modifications sont évidentes pour les carrés à la limite de la population.

Dans l'échantillonnage en grappes adaptatif, pour chaque carré de l'échantillon aléatoire initial ayant une valeur y supérieure à zéro, tous ses voisins sont observés et si l'un d'entre eux a une valeur y supérieure à zéro, alors ses voisins sont également observés et ainsi de suite. Ce processus se poursuit jusqu'à ce que seules des valeurs nulles soient observées. Pour un carré donné, l'ensemble des carrés obtenus de cette façon, présentant des valeurs y supérieures à zéro, est appelé un **réseau**. Ainsi, un réseau est formé d'un ensemble de carrés ayant des valeurs non nulles et, si l'un d'eux apparaît dans l'échantillon, alors tous les autres y seront. L'ensemble des carrés ayant une valeur y de zéro qui ont été observés dans le processus sont appelés les arêtes d'un réseau. Comme nous l'avons souligné dans l'introduction, un réseau pour un carré dont la valeur y est zéro se limite à lui-même. Pour un carré (i, j), supposons que $\Psi_{i,j}$ correspond à tous les carrés de son réseau.

Supposons maintenant qu'un échantillon aléatoire simple initial sans remise de taille n_1 est pris en compte. Pour $k = 1, \ldots, n_1$, supposons que Ψ_{i_k,j_k} indique le réseau de carrés qui apparaît au k^{e} tirage de l'échantillon. Précisons que si deux carrés, qui appartiennent au même réseau, sont dans le premier échantillon, alors leurs réseaux sont identiques et ils seront tous deux inclus dans l'estimation du total de la population. Pour chaque k, nous supposons que m_k est le nombre de carrés dans Ψ_{i_k,j_k} et que $\bar{y}_k^*$ est la moyenne des comptes des unités qui apparaissent dans Ψ_{i_k,j_k}. Dans le cas de l'échantillonnage en grappes adaptatif, Thompson (1990) a produit un estimateur sans biais du total de la population et un estimateur sans biais de sa variance. Pour un échantillon par EGA, cette estimation, $\hat{T}_{ac}$ et sa variance estimée, $\hat{v}_{ac}$ sont données par

$$\hat{T}_{ac} = N\frac{\sum_{k=1}^{n_1}\bar{y}_k^*}{n_1} \text{ et } \hat{v}_{ac} = \frac{N(N-n_1)}{n_1}\frac{\sum_{i=k}^{n_1}(\bar{y}_k^* - \hat{T}_{ac}/N)^2}{n_1-1}. \quad (2.1)$$

Nous constatons que les estimateurs par EGA, point et intervalle, dépendent uniquement des moyennes du réseau. Par conséquent, la variabilité des comptes à l'intérieur d'un réseau donné ne joue aucun rôle. En effet, il est présumé que tous les comptes à l'intérieur d'un réseau donné sont les mêmes. De plus, le fait qu'il n'y avait que quelques carrés dont les comptes étaient supérieurs à zéro dans la population ne semble jouer aucun rôle explicite dans l'étape d'inférence après la sélection d'un échantillon par EGA.

Dryver et Chao (2007) proposent une autre façon d'analyser l'estimateur par EGA dans l'équation ci-dessus. Ils tiennent compte d'une deuxième version de la population, mais qui est liée. Pour constituer cette deuxième population, ils procèdent comme suit. Pour chaque réseau de la population, nous remplaçons chaque valeur y par la moyenne de toutes les valeurs y de ce réseau. Si un réseau contient un seul carré, alors sa valeur y est inchangée. Toutefois, si un réseau contient plus d'un carré, alors chacune de ses valeurs y est remplacée par la moyenne de toutes les valeurs y des carrés qui forment le réseau.

De toute évidence, cette population de rechange présente le même total que la population originale. De plus, les valeurs y_k observées dans la deuxième population, pour les unités de l'échantillon initial, sont identiques aux valeurs $\bar{y}_k^*$ de l'équation (2.1). Cette observation confirme clairement que l'estimateur par EGA est un estimateur sans biais, et qu'à l'étape d'inférence, l'estimateur par EGA ne tient pas compte du fait que l'échantillonnage se fait à partir d'une population rare.

Finalement, certains pourraient s'étonner que $\hat{T}_{ac}$ dépende de N et comprenne les réseaux ayant une moyenne de zéro. Toutefois, une situation similaire se produit lors de l'estimation du total d'un domaine, lorsque la taille du domaine est inconnue et que l'on dispose d'un échantillon aléatoire de la population entière. Voir, par exemple, la discussion à propos de l'estimation par domaine dans Cochran (1977).

2.2 Une nouvelle approche

L'approche par EGA s'appuie sur deux hypothèses de base : il n'y a que quelques carrés dont le compte est supérieur à zéro et ces carrés ont tendance à être groupés en grappes. Bien que le plan d'échantillonnage repose sur ces deux hypothèses, comme nous venons de le souligner à la fin de la section précédente, l'estimateur par EGA ne semble jamais utiliser l'information indiquant que la proportion des carrés qui ont un compte supérieur à zéro est petite à l'étape d'inférence. Il serait possible de mieux faire en incorporant ces renseignements au moment de construire un estimateur.

Supposons que D_b corresponde à l'ensemble de toutes les unités de la population dont le compte est supérieur à zéro, et supposons que θ corresponde au nombre d'unités dans D_b. Pour nous, θ est un paramètre inconnu et jouera un rôle important dans ce qui suit. Supposons que T_b corresponde au total de toutes les unités dans D_b. Évidemment, T_b correspond également au total de toutes les unités dans la population. Toutefois, nous introduisons cette notation pour insister sur le fait que notre approche est axée sur D_b. Nous scindons le problème en deux parties. Dans la première partie, nous obtenons une estimation

de θ en utilisant les renseignements contenus dans l'échantillon aléatoire initial de taille n_1. À partir de cette estimation et de tous les comptes dans les réseaux observés, nous trouvons ensuite une estimation de T_b.

D'après nos connaissances, la notion de rareté n'a jamais été explicitement définie dans la littérature. D'une certaine manière, elle serait l'analogue de quelques éléments du paradoxe du tas pour plusieurs éléments. Une version de ce paradoxe est « combien faut-il de cailloux pour faire un tas ? ». Si nous enlevons un seul caillou d'un tas, ce tas reste un tas, mais si nous répétons ce geste un nombre suffisant de fois, nous n'aurons plus de tas. De même, si une espèce est rare dans notre population, l'ajout d'un compte positif à un carré qui en contient zéro ne changerait pas notre perception de ce qui constitue la rareté. Cependant, si nous ajoutons un nombre suffisant de comptes positifs, alors l'espèce ne serait plus rare. Supposons que K_θ serait le plus grand entier inférieur ou égal à $N/10$. Nous commencerons par supposer que notre espace de paramètre de θ correspond à l'ensemble des entiers

$$\Theta = \{i : i = 0, i = 1, \ldots, i = K_\theta\}. \tag{2.2}$$

On pourrait soulever l'argument que ce choix est quelque peu arbitraire, mais il est cohérent avec la notion de rareté et vaut pour plusieurs exemples étudiés dans la littérature. Ultérieurement, nous verrons que nous pouvons assouplir cette hypothèse.

Nous ne faisons pas d'autres hypothèses à propos de la population, hormis que les carrés ou les unités ayant un compte supérieur à zéro tendent à se présenter en grappes et à former des réseaux. L'endroit où les réseaux sont situés dans la population n'a aucune incidence lorsque l'échantillonnage en grappes adaptatif est utilisé. Nous ne faisons aucune hypothèse explicite à propos de l'étendue des possibles comptes. Quiconque souhaite utiliser l'échantillonnage en grappes adaptatif et présume que le nombre de comptes positifs est rare, comme nous l'avons décrit dans le précédent paragraphe, pourrait utiliser notre approche.

Pour développer notre estimateur, nous avons choisi six populations possibles. Trois d'entre elles ont déjà été mentionnées dans la littérature et les trois autres ont été construites. Elles seront décrites en détail dans la section 4, où nous analyserons les résultats de notre simulation. Notre estimateur a été obtenu par essai et erreur, lors de simulations à partir de ces six populations établies.

Il faut préciser que la détermination des estimateurs ponctuels et d'intervalle pertinents pour T_b n'est pas un problème facile à résoudre lorsque la valeur de θ est petite. Dans pareil cas, un échantillon par EGA peut contenir un compte de zéro ou encore un ou deux comptes supérieurs à zéro. En outre, nous croyons qu'il est possible d'obtenir des estimateurs sensés seulement si nous disposons de renseignements supplémentaires *a priori*. Toutefois, dans le cas présent, nous présumons que nous ne disposons pas de ces renseignements. Dans l'esprit de notre approche objective, nous avons décidé d'ignorer de tels échantillons dans nos études de simulation.

Dans la prochaine section, nous présenterons nos estimateurs et nous expliquerons la logique sous-jacente et l'intuition qui nous y a menés.

3. Un nouvel estimateur

Il convient de rappeler que D_b est le sous-ensemble de Y comportant tous les carrés ayant une valeur y supérieure à zéro. Si T correspond au total de la population de Y, alors T est égal à T_b, soit le total des unités appartenant à D_b. Notre but est d'estimer T_b. Il faut rappeler que θ désigne la taille de D_b, soit le nombre de $y_{i,j}$ dans la population qui sont plus grands que zéro. Pour nous, θ est un paramètre inconnu et nous désignons son espace de paramètre Θ, l'ensemble des entiers qui sont supérieurs ou égaux à zéro et inférieurs ou égaux à K_θ.

Supposons que X est le nombre d'unités qui appartiennent à D_b dans le premier échantillon aléatoire simple sans remise de taille n_1. Alors, X a une distribution hypergéométrique qui dépend de N, de n_1 et du paramètre inconnu θ. Étant donné que $X = x$, nous pouvons utiliser la fonction de vraisemblance résultante pour l'estimation de θ. Supposons que y_b corresponde aux valeurs de toutes les unités de l'échantillon par EGA ayant des valeurs supérieures à zéro. Supposons que n_b corresponde au nombre d'unités dans y_b. Mentionnons que $n_b \geq x$. Supposons que $y_{b'}$ corresponde aux membres restants de D_b qui n'ont pas été observés dans l'échantillon par EGA. Soulignons que si $\theta = n_b$, alors $y_{b'}$ est l'ensemble vide.

D'un point de vue bayésien, le problème de l'estimation de T_b peut être scindé en deux étapes. Premièrement, nous simulons une possible valeur de θ à partir de sa distribution *a posteriori*. Puis, à partir de cette valeur, nous simulons de possibles comptes pour l'ensemble de valeurs possibles $\theta - n_b \geq 1$ des valeurs non observées constituant $y_{b'}$. En combinant cet ensemble de valeurs simulé aux valeurs observées y_b, nous avons obtenu une copie complète de D_b et de sa valeur correspondante de T_b. En répétant ces deux étapes plusieurs fois, nous pouvons utiliser les résultats obtenus, d'une façon bayésienne, pour déterminer une estimation ponctuelle et une borne supérieure du total des unités appartenant à D_b.

Nous présenterons maintenant notre distribution *a priori* de θ et nous expliquerons notre façon de simuler des copies complètes de D_b après l'observation d'un échantillon par EGA.

3.1 Estimation du nombre de comptes supérieurs à zéro

Étant donné que $X = x$, une estimation naturelle de θ correspond à l'estimation du maximum de vraisemblance. Toutefois, nous pouvons observer dans les simulations que la fonction de vraisemblance tend à attribuer un facteur de pondération trop important aux grandes valeurs de θ pour les plus petites valeurs de x que nous serions susceptibles d'observer dans l'EGA. Une autre méthode qui pourrait être adoptée serait l'approche bayésienne classique, où nous pourrions préciser une distribution *a priori* sur Θ. Une possible distribution *a priori* non informative pourrait être la distribution uniforme sur Θ, définie dans l'équation (2.2). Son espérance *a priori* est approximativement $N/20$, ce qui serait une bonne estimation par défaut de la valeur de θ. Toutefois, les simulations démontrent qu'elle présente la même faiblesse que l'estimation du maximum de vraisemblance. Les deux tendent à surestimer θ pour les échantillons par EGA.

Notre but était de déterminer une bonne distribution par défaut de θ qui conviendrait bien à une variété de populations où l'EGA serait utilisé. Au lieu de mettre notre répartition sur Θ, nous avons plutôt choisi

les distributions sur Θ/N, la proportion de la population de $y_{i,j}$ qui est supérieure à zéro. Une telle répartition peut alors être vue comme une distribution sur θ.

Rappelons que la loi bêta, selon les paramètres $\alpha > 0$ et $\beta > 0$, est établie sur l'intervalle d'unité et que sa fonction de densité est donnée par

$$f_{\alpha,\beta}(z) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} z^{\alpha-1}(1-z)^{\beta-1} \quad \text{pour} \quad 0 \le z \le 1. \tag{3.1}$$

Si, pour tout choix des paramètres α et β, nous prenons les valeurs de la fonction ci-dessus, pour tous les membres de θ dans Θ, et que nous les divisions par leur somme, l'ensemble des nombres obtenu correspond à une distribution de probabilités définie sur Θ. Après plusieurs essais sur nos six populations d'essai, nous choisissons, comme distribution *a priori* sur Θ, la densité bêta renormalisée où $\alpha = 1$ et $\beta = 90$.

Pour aider à comprendre ce choix, prenons le cas où $N = 400$. La moyenne *a priori* de notre choix est 3,92, alors que le choix $\alpha = \beta = 1$ a une moyenne *a priori* de 20. Si la valeur réelle de θ/N se situe autour de 0,02 ou 0,03, le dernier choix donnera une surestimation de la taille de θ. Notre choix aidera à diminuer le biais vers le haut de la fonction de vraisemblance pour l'échantillon aléatoire initial. Cet aspect sera abordé de manière plus détaillée dans la section 5, lorsque nous présentons les simulations.

Cela est lié à un problème similaire concernant l'estimateur par EGA $\hat{T}_{ac}$ (défini dans l'équation [2.1]) que nous avons vu dans nos simulations. Pour le plan par EGA, une importante propriété de $\hat{T}_{ac}$ est liée au fait qu'il s'agit d'un estimateur sans biais du total de la population. Si n_1, la taille de l'échantillon aléatoire original, est petit et que θ est petit, alors il y a une probabilité non triviale que nous observions des échantillons où chaque compte est zéro. Pour de tels échantillons, $\hat{T}_{ac}$ donne une estimation de zéro pour le total de la population. Il s'agira d'une sous-estimation, à moins que la population totale soit réellement zéro. Pour compenser cette sous-estimation, nous avons constaté que $\hat{T}_{ac}$ est surestimé lorsque l'échantillon contient « beaucoup » d'unités dont les comptes sont supérieurs à zéro. Le rapprochement à la vraie valeur de θ est obtenu lorsque le nombre d'unités de l'échantillon, dont les comptes sont supérieurs à zéro, est proche de $\theta/2$.

Nous devons souligner que notre estimateur de θ ne dépend pas du fait que nos unités de base sont des carrés et que la population est un rectangle. De plus, il ne repose sur aucune hypothèse à propos d'une forme ou l'autre des réseaux ou des valeurs des comptes à l'intérieur d'un réseau. Il dépend seulement du fait que nous avons un échantillon aléatoire obtenu à partir des unités de base de la population. Il devrait être valide chaque fois que les unités de base correspondent approximativement à la même taille et que la notion de voisin puisse être définie d'une manière sensée. La forme générale de la configuration de la population est sans importance. De plus, elle ne repose sur aucune hypothèse à propos des valeurs dans y_b, soit les comptes dans l'échantillon par EGA supérieurs à zéro.

3.2 Estimation du total de la population

Pour une valeur de θ, disons $\hat{\theta}$, obtenue à partir de notre distribution *a posteriori*, et n_b, le nombre de comptes observés dans y_b, nous devons définir une distribution qui simule des valeurs possibles pour les

unités $\hat{\theta} - n_b \geq 1$ restantes, dont les comptes doivent être supérieurs à zéro. Nous présumons qu'on en sait très peu sur la forme des réseaux et l'incidence de la taille d'un réseau sur ses valeurs y. Dans ce cas, une hypothèse simple est de présumer que y_b, soit tous les comptes supérieurs à zéro dans l'échantillon par EGA complet, correspond approximativement à un échantillon « représentatif » des valeurs dans D_b. Selon cette hypothèse, une façon sensée de simuler les membres non observés de D_b est d'utiliser l'échantillonnage de Pólya.

Plus précisément, plaçons n_b balles dans une urne, chaque balle représentant un des comptes appartenant à y_b. Donnons à chacune des balles le facteur de pondération $w > 0$. Prenons aléatoirement une balle dans l'urne. Remettons-la dans l'urne avec une autre balle à laquelle est attribué le compte de la balle sélectionnée. Nous donnons à cette nouvelle balle un facteur de pondération de un. Puis, une autre balle est sélectionnée dans l'urne, qui contient maintenant $n_b + 1$ balles, ayant une probabilité proportionnelle à leur facteur de pondération. La balle sélectionnée est remise dans l'urne avec une autre balle à laquelle est attribué un compte égal à celui de la balle sélectionnée. Un facteur de pondération de un est donné à cette nouvelle balle et l'urne contient désormais $n_b + 2$ balles. Ce processus se poursuit jusqu'à ce que l'urne contienne $\hat{\theta}$ balles. Selon cette distribution, la valeur attendue du total de tous les comptes dans l'urne qui fait l'objet d'une simulation ne dépend pas de w et correspond à $\hat{\theta}\bar{y}_b$, où $\bar{y}_b$ est la moyenne des comptes dans y_b. La variance, toutefois, dépend de w et diminue lorsque w augmente. Il est démontré dans Meeden (1999), équation (2.5), que la variance est donnée par

$$\mathrm{Var}(T_b \mid y_b, n_b, \hat{\theta}, w) = \hat{\theta}(\hat{\theta} - n_b) \frac{\mathrm{var}(y_b)}{n_b} \frac{n_b - 1}{1 + n_b w} \frac{\hat{\theta} + n_b w - n_b}{\hat{\theta}} \tag{3.2}$$

où $\mathrm{var}(y_b)$ correspond à la variance d'échantillon des comptes dans y_b.

Maintenant, nous devons préciser une valeur de w, le facteur de pondération de chaque balle dans l'urne au début de ce processus. Mentionnons que w apparaît seulement dans les deux dernières fractions de l'équation (3.2). Il est facile de vérifier que la valeur du produit de ces deux fractions est un, si nous prenons comme valeur de w

$$w^* = \frac{n_b(\hat{\theta} - n_b + 1) - 2\hat{\theta}}{n_b(\hat{\theta} - n_b + 1)}. \tag{3.3}$$

Il est facile de vérifier que $w^* > 0$ lorsque $n_b > 2$.

Selon ce choix de w^* et pour une valeur y_b établie et une valeur $\hat{\theta}$ établie, notre variance conditionnelle est donnée par

$$\mathrm{Var}(T_b \mid y_b, n_b, \hat{\theta}, w^*) = \hat{\theta}(\hat{\theta} - n_b) \frac{\mathrm{var}(y_b)}{n_b}. \tag{3.4}$$

Cette équation reflète notre hypothèse selon laquelle nous voyons les valeurs dans y_b comme échangeables et découlant de quelque chose reproduisant, de manière approximative, un échantillon aléatoire. Maintenant, si n_b est raisonnablement grand, disons autour de 20, et que y_b correspond approximativement à un

échantillon aléatoire, alors l'équation ci-dessus devrait donner une estimation raisonnablement bonne de la variance. Toutefois, pour des valeurs plus petites de n_b, disons moins de cinq, l'équation n'est pas aussi bonne. Dans ce qui suit, nous décrirons une façon de traiter ce problème.

D'abord, désignons $\hat{T}_{ab}$ comme indiquant notre quasi-estimateur bayésien ou approximatif de T_b, le nombre total d'espèces dans la population, selon notre processus de simulation en deux étapes. Supposons que $\overline{y}_b$ corresponde à la moyenne de toutes les unités dans y_b. Alors, notre estimation de T_b est

$$\hat{T}_{ab} = E(T_b) = E(E(T_b \mid \theta)) = E(\theta \overline{y}_b) = \hat{\theta}_{1,90} \overline{y}_b \tag{3.5}$$

où $\hat{\theta}_{1,90}$ est la moyenne de notre distribution *a posteriori* pour θ.

Pour déterminer la variance de $\hat{T}_{ab}$, nous utilisons la formule bien connue selon laquelle une variance peut être écrite comme la variance d'une espérance conditionnelle plus l'espérance d'une variance conditionnelle. Par conséquent, pour un échantillon par EGA établi, nous avons

$$\begin{aligned} \text{Var}(\hat{T}_{ab}) &= \text{Var}\Big(E(\hat{T}_{ab} \mid \theta)\Big) + E\Big(\text{Var}(\hat{T}_{ab} \mid \theta)\Big) \\ &= \text{Var}(\theta \overline{y}_b) + E\left(\theta(\theta - n_b)\frac{\text{var}(y_b)}{n_b}\right) \\ &= \overline{y}_b^2 V(\theta) + \Big(E(\theta^2) - n_b E(\theta)\Big)\frac{\text{var}(y_b)}{n_b}. \end{aligned} \tag{3.6}$$

Nous devons tout de même trouver une façon d'utiliser la variance de notre estimateur pour produire un bon estimateur de l'intervalle du nombre total d'espèces dans la population. Dans l'EGA, l'échantillon aléatoire initial comportera majoritairement des comptes de zéro. Rappelons que x correspond au nombre de comptes supérieurs à zéro dans l'échantillon initial de taille n_1 et que n_b correspond au nombre d'unités dans l'échantillon par EGA complet comportant des comptes supérieurs à zéro. Notons que x est inférieur ou égal à n_b. Une borne supérieure très simpliste correspondrait à notre estimation ponctuelle plus le produit de 1,96 et de la racine carrée de $\text{Var}(\hat{T}_{ab})$. Mais, comme nous l'avons déjà souligné, il n'est pas surprenant d'obtenir de piètres résultats parce que, dans l'EGA, les valeurs de x et de n_b peuvent être très petites.

Examinons un cas où une population contient un ou deux réseaux de taille un ou deux ayant des comptes beaucoup plus grands que les autres comptes de la population. Pour de telles populations, lorsque x est petit, ces réseaux ne seront probablement pas observés et notre estimation de variance sera trop petite. Pour nous aider à nous protéger contre cette possibilité, nous devons augmenter la borne supérieure simpliste évoquée ci-dessus, particulièrement lorsque n_b est petit. Pour ce faire, nous présumons que

$$\lambda = 10^{(2,5/n_b)}. \tag{3.7}$$

Mentionnons que λ diminuera alors que n_b augmente. Nous ne prétendons pas qu'il s'agit d'un choix optimal pour ajuster vers le haut notre borne supérieure. Nous avons simplement constaté que ce choix convient bien à nos populations d'essai.

Pour calculer les bornes supérieures de notre estimation, nous utiliserons

$$\hat{T}_{ab} + \lambda \times 1{,}96\sqrt{\mathrm{Var}(\hat{T}_{ab})} \tag{3.8}$$

comme notre limite de confiance supérieure approximative de 95 %. En raison de la petite taille des échantillons de l'EGA, il pourrait sembler étonnant de présumer que $\hat{T}_{ab}$ a, approximativement, une distribution normale. Toutefois, comme nous le verrons dans les simulations, ce choix semble fonctionner raisonnablement bien pour la majorité des cas où l'EGA est approprié en raison de notre choix de λ.

La deuxième partie de notre processus de simulation en deux étapes n'est clairement pas bayésienne puisque nos choix pour w^* et λ dépendent de n_b, qui provient des données observées. Ces valeurs ne découlent pas d'une certaine distribution *a priori*, même si le fait que l'EGA initial était un échantillon aléatoire est un renseignement important pour nous. Comme nous l'avons souligné dans l'introduction, deux autres exemples où des distributions « *a posteriori* » sont définies sans une répartition *a priori* et qui reposent, en partie, sur le plan d'échantillonnage se trouvent dans Si et coll. (2015), et Savitsky et Toth (2014). Cela semble logique lorsqu'il y a des renseignements *a priori* disponibles qui ne peuvent pas être incorporés dans une distribution *a priori*.

Une borne inférieure sensée pour T_b n'est que la somme de tous les comptes de l'échantillon par EGA qui sont supérieurs à zéro. Pour l'estimateur par EGA, nous utiliserons la même borne inférieure et, pour sa borne supérieure, nous utiliserons son estimation plus le produit de 1,96 et de sa variance estimée. De nouveau, nous présumons que l'estimateur par EGA a, approximativement, une distribution normale.

Une objection pourrait être soulevée relativement à l'effet que notre hypothèse voulant que y_b soit approximativement un échantillon « représentatif » des valeurs dans D_b est excessive. Mais, rappelons qu'à la fin de la section 2.1 nous avons vu que la seule information utilisée dans l'estimateur par EGA est la moyenne des comptes dans un réseau et qu'en prenant l'autre population de Dryver et Chao (2007), nous avons constaté que l'estimateur par EGA présume essentiellement l'échangeabilité entre les moyennes du réseau. Nous croyons que tous les comptes observés peuvent comporter certains renseignements supplémentaires et que notre hypothèse selon laquelle l'échantillon observé est approximativement un échantillon représentatif n'est pas plus solide que de présumer l'échangeabilité entre les moyennes du réseau.

Notre avons découvert notre estimateur en faisant des simulations sur un ensemble de six populations pour lesquelles l'EGA serait approprié. Trois d'entre elles ont déjà été citées dans la littérature et les trois autres ont été construites. Nous avons ensuite utilisé des études de simulation pour voir de quelle façon différents choix d'une distribution *a priori* et d'une variance estimée fonctionneraient. Ce que nous avons décrit ci-dessus correspond à ce que nous avons trouvé de mieux. Nous avons trouvé d'autres solutions qui semblaient fonctionner presque aussi bien, mais ces choix nous semblent les plus adaptés pour le moment. Dans la prochaine section, nous décrivons ces six populations.

4. Les populations

Pour les simulations, nous avons utilisé trois populations différentes dont il a déjà été question dans la littérature. Une de ces populations correspond au premier exemple analysé dans Thompson (1990). L'auteur

la présente comme un exemple type où l'EGA pourrait être utilisé. Il ne donne pas d'autres détails; il est donc probable que ce soit lui qui l'ait construite. La population comporte 400 unités, formant trois réseaux dont les tailles sont de 6, 11 et 4 et dont les moyennes sont 6,0, 9,73 et 11,75 respectivement. La moyenne de tous les comptes supérieurs à zéro est 9,4 et le compte le plus élevé, soit 39, apparaît dans le réseau de taille 11. Nous désignons cette population par « thmp ». Gattone, Mohamed et Di Battista (2016) décrivent deux échantillons de buffles africains et de bubales africains pris en 2010. Nous désignons ces populations par « afrbuf » et « afrhart », conformément à l'article de référence.

Nous avons également construit trois populations supplémentaires. Pour construire une population, nous étudions une grille de points sur la surface terrestre. Nous présumons que leur latitude et leur longitude sont espacées également, bien qu'il ne soit pas nécessaire que les différences successives dans les deux directions soient les mêmes. Selon la topographie de l'emplacement des points, leur altitude, mesurée en mètre, peut présenter un comportement de regroupement qui est l'hypothèse sous-jacente de l'EGA.

Pour déterminer l'altitude sur une grille, nous utilisons la fonction *elevation* dans R (R Core Team, 2023) au moyen du paquet rgbif (Chamberlain, Ram, Mcglinn et Barve, 2019). Pour obtenir l'ensemble final des « comptes », nous avons procédé selon trois étapes. Premièrement, nous avons arrondi chaque altitude de l'ensemble à sa valeur entière la plus proche. Deuxièmement, nous avons choisi une valeur $\epsilon > 0$, mais proche de zéro, et calculé le percentile $1-\epsilon$ de notre ensemble de valeurs entières, soit q_ϵ. Nous avons fixé à zéro chaque compte inférieur à q_ϵ. Finalement, nous avons soustrait un certain entier, inférieur à q_ϵ, de chaque compte supérieur à zéro. Le nombre de comptes obtenus supérieurs à zéro serait alors assez petit pour représenter les comptes d'une espèce rare. Par exemple, si nous fixons $\epsilon = 0{,}05$, l'ensemble obtenu des comptes aura 5 % de ses valeurs supérieures à zéro. Ainsi, dans la population obtenue, les comptes ou les valeurs $y_{i,j}$ sont soit zéro si leur altitude arrondie est inférieure à un certain niveau ou la différence entre leur altitude arrondie et une certaine constante. Selon la topographie d'une région, la grille des « comptes » obtenue peut présenter le comportement de regroupement qui fait de l'EGA un choix sensé. Cette méthode souple permet de repérer plusieurs populations réalistes qui peuvent être utilisées dans les études de simulation de l'EGA.

Ensuite, nous décrivons les trois grilles que nous avons prises pour construire les trois populations utilisées dans nos études de simulation. Pour la première, nous avons choisi une grille à Paris, en France; pour la deuxième, une grille à Niagara Falls, à la frontière entre le Canada et les États-Unis; pour la troisième, une grille près de Devil's Tower dans l'ouest des États-Unis. Pour les deux premières, notre grille a une dimension de 23 par 23, alors que la troisième fait 26 par 45. Nous désignons ces populations par « paris », « nfalls » et « devt ». Les renseignements sommaires des six populations sont présentés au tableau 4.1. Soulignons que la proportion des comptes supérieurs à zéro varie de 0,038 à 0,084. Des graphiques en trois dimensions des populations sont présentés dans les six premières illustrations sur les trois pages suivantes.

Tableau 4.1
Renseignements sommaires sur les populations utilisées dans les simulations.

Population	N	N_{ntw}	θ	θ/N	$y_{\max}$	T_b	T_b/θ
afrbuf	391	5	15	0,038	99	334	22,3
nfalls	529	5	20	0,038	59	368	18,4
thmp	400	3	21	0,053	39	190	9,4
devt	1 170	10	85	0,073	34	868	10,2
paris	529	6	43	0,081	63	1 112	25,9
afrhart	391	9	33	0,084	20	171	5,18

Notes : Il convient de rappeler que N correspond à la taille de la population, θ correspond au nombre d'unités supérieures à zéro et que T_b est leur somme. Nous désignons N_{ntw} comme le nombre de réseaux dans la population et $y_{\max}$ comme la valeur maximale de y dans la population.

Figure 4.1 Graphique en trois dimensions de la population afrbuf.

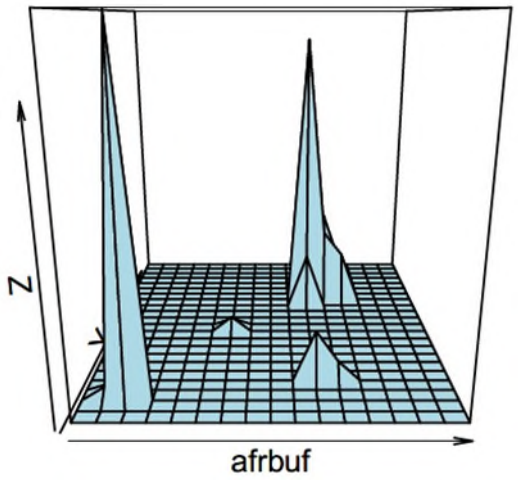


Figure 4.2 Graphique en trois dimensions de la population nfalls.

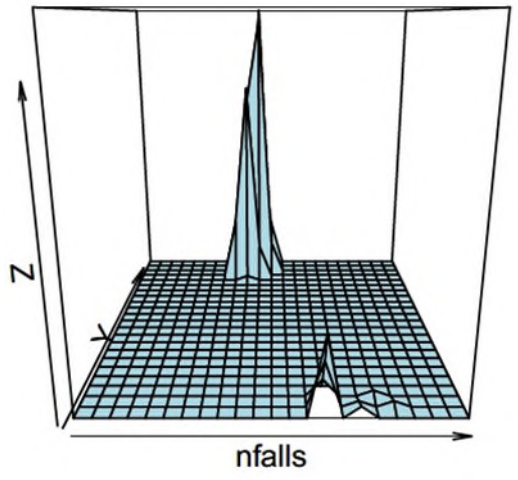


Figure 4.3 Graphique en trois dimensions de la population thmp.

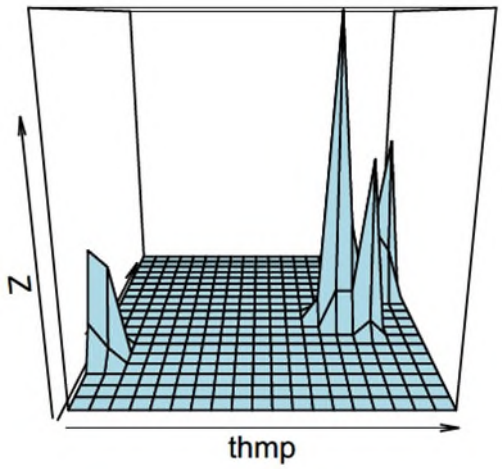


Figure 4.4 Graphique en trois dimensions de la population devt.

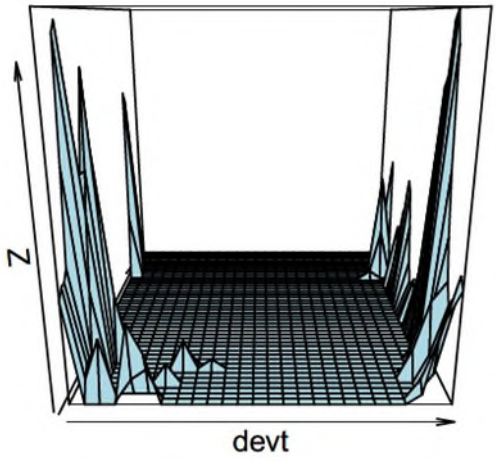


Figure 4.5 Graphique en trois dimensions de la population paris.

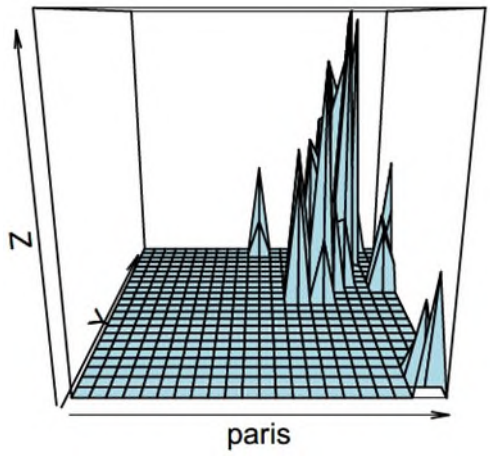
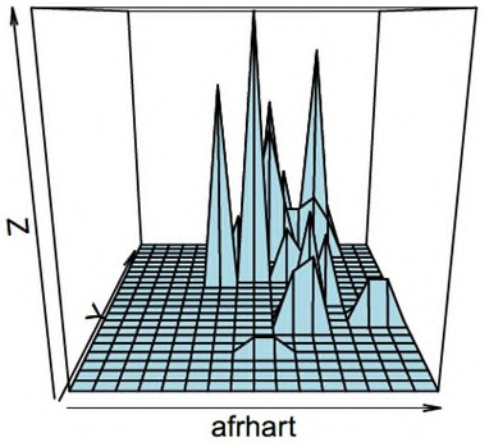


Figure 4.6 Graphique en trois dimensions de la population afrhart.



5. Les simulations

5.1 Exécution des simulations

Vers la fin de la section 2.1, nous avons expliqué notre choix de ne tenir compte que des échantillons par EGA qui comportaient au moins trois comptes supérieurs à zéro. Pour cette raison, dans nos simulations, nous tiendrons compte uniquement des échantillons où $n_b > 2$. Autrement dit, notre fréquence de couverture est une fréquence conditionnelle; elle est conditionnelle au fait de trouver au moins trois comptes supérieurs à zéro. Donc, les résultats présentés dans les tableaux 5.1 à 5.8 et 5.10 à 5.13, pour l'estimateur d'échantillonnage en grappes adaptatif et notre estimateur quasi bayésien, soit EGA et BAY respectivement, sont conditionnels. Pour chaque méthode, les tableaux donnent la valeur moyenne d'un estimateur, « Est », son biais relatif moyen, « Rbias », son erreur absolue moyenne, « Aberr », la borne inférieure moyenne de son estimation par intervalle, « Lowbd », la longueur moyenne de son intervalle, « Len », et la fréquence à laquelle sa borne supérieure était plus grande que T_b.

Dans nos simulations d'échantillonnage en grappes adaptatif, nous avons utilisé deux tailles d'échantillon initial différentes, soit 10 % et 20 % de la taille de la population. Nous présenterons, dans le cas présent, uniquement les résultats des 10 %. La façon de comparer les deux méthodes est essentiellement la même pour les deux tailles d'échantillon initial différentes, mais bien entendu, elles donnent toutes les deux de meilleurs résultats pour la taille d'échantillon initial la plus grande.

Finalement, on pourrait s'inquiéter de ce qu'il adviendrait si la taille réelle de D_b était légèrement plus grande que $N/10$ et, par conséquent, ne respecterait pas notre définition d'une espèce rare. En pratique, il est peu vraisemblable que la valeur de n_b observée soit plus grande que cette limite. Toutefois, pour permettre cette possibilité, lors des simulations, nous avons défini notre distribution *a priori* sur les entiers entre $0{,}01\,N$ et $0{,}15\,N$. Les résultats diffèrent peu des résultats qui auraient été obtenus si l'ensemble des entiers possibles était les entiers non négatifs inférieurs ou égaux à $0{,}1\,N$. Cette situation survient parce que,

alors que nous nous éloignons de n_b, la distribution *a posteriori* diminue très rapidement et les points qui sont éloignés contribuent peu à la probabilité. Par conséquent, même si nous avons créé notre distribution « *a posteriori* » en pensant à la borne supérieure la plus petite, le résultat est presque aussi bon lorsque nous prenons une borne supérieure beaucoup plus grande et imprécise. Ainsi, pour utiliser notre méthode, il n'est pas nécessaire de faire une bonne supposition de la borne supérieure de θ.

5.2 Résultats pour les six populations

Il y a plusieurs éléments à souligner à propos des résultats de la simulation présentés dans les tableaux 5.1 à 5.6. Pour toutes les populations, l'estimateur BAY a une erreur absolue moyenne plus petite que l'estimateur par EGA, parfois même dramatiquement moindre. En général, les bornes supérieures de BAY donnent de meilleurs résultats que les bornes supérieures d'EGA. La population « afrbuf » est le seul cas où la borne supérieure d'EGA est élevée. La borne supérieure de BAY est trop grande. Pour la population « nfalls », les deux méthodes ont un comportement assez similaire. Pour les populations « thmp », « devt » et « paris », la borne supérieure BAY est la meilleure de toute évidence. La seule population pour laquelle la fréquence de couverture diminue sous 0,90 est la population « afrhart ». Dans ce cas, la fréquence de couverture est seulement de 0,892, mais sa borne supérieure était beaucoup plus petite que la borne d'EGA, qui avait une fréquence de couverture de 0,861. Une partie des renseignements des tableaux est présentée sous forme graphique dans la figure 5.1.

Tableau 5.1
Population afrbuf.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
EGA	489	0,464	187,2	195	974	0,98
BAY	334	-0,000	69,3	195	1 644	0,98

Notes : Pour 1 000 échantillons aléatoires simples de taille 39, 660 échantillons comportaient au moins trois comptes supérieurs à zéro. Le nombre total moyen de valeurs positives observées était 6,96.

Tableau 5.2
Population nfalls.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
EGA	448	0,217	198,9	319	706	0,924
BAY	447	0,214	126,5	319	724	0,998

Notes : Pour 1 000 échantillons aléatoires simples de taille 53, 819 échantillons comportaient au moins trois comptes supérieurs à zéro. Le nombre total moyen de valeurs positives observées était 13,57.

Tableau 5.3
Population thmp.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
EGA	211	0,112	87,9	118	341	0,921
BAY	157	-0,173	45,7	118	274	1,000

Notes : Pour 1 000 échantillons aléatoires simples de taille 40, 887 échantillons comportaient au moins trois comptes supérieurs à zéro. Le nombre total moyen de valeurs positives observées était 12,98.

Tableau 5.4
Population devt.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
EGA	874	0,007	224	723	708	0,942
BAY	857	-0,012	89	723	548	0,972

Notes : Pour 1 000 échantillons aléatoires simples de taille 117, 1 000 échantillons comportaient au moins trois comptes supérieurs à zéro. Le nombre total moyen de valeurs positives observées était 65,98.

Tableau 5.5
Population paris.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
EGA	1 119	0,006	389,5	970	1 119	0,909
BAY	1 114	0,002	69,2	969	603	0,984

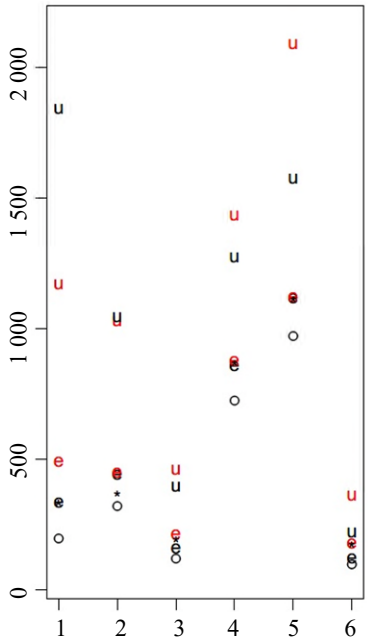
Notes : Pour 1 000 échantillons aléatoires simples de taille 53, 983 échantillons comportaient au moins trois comptes supérieurs à zéro. Le nombre total moyen de valeurs positives observées était 34,19.

Tableau 5.6
Population afrhart.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
EGA	175	0,021	78,4	96	261	0,861
BAY	119	-0,305	52,4	96	120	0,892

Notes : Pour 951 échantillons aléatoires simples de taille 39, 968 échantillons comportaient au moins trois comptes supérieurs à zéro. Le nombre total moyen de valeurs positives observées était 18,74.

Figure 5.1 Représentation graphique de certaines données des tableaux 5.1 à 5.6 pour les six populations, soit afrbuf(1), nfalls(2), thmp(3), devt(4), paris(5) et afrhart(6).



Notes : Le chiffre entre parenthèses correspond à leur emplacement sur l'axe horizontal. Le vrai total de population est indiqué par ⋆. L'estimateur par EGA et sa borne supérieure sont désignés par e et u. Pour l'estimateur BAY, ces quantités sont en noir et la borne inférieure commune indiquée par o.

Pour les six populations du tableau 4.1, celui pour lequel l'estimateur BAY avait le biais négatif le plus extrême, soit -0,305, était la population « afrbuf ». En observant de plus près cette population « afrbuf », nous constatons qu'elle présente 33 unités de plus de zéro dans 9 réseaux, et une valeur moyenne de 5,18. Les trois valeurs les plus grandes dans la population sont 17, 15 et 20. Les deux dernières apparaissent dans un réseau de taille un, alors que la première apparaît dans le réseau le plus grand qui comporte 16 unités. La plupart des valeurs restantes de la population sont de 5 ou moins. La moyenne de ces 30 unités restantes est de 3,97. Pour un choix de $\alpha = 1$ et $\beta = 90,$ la valeur moyenne de notre estimateur de $\theta = 33$ était 23,5. Ce résultat aide à expliquer le grand biais négatif de notre estimateur pour cette population. Pour améliorer son rendement, il faudrait augmenter l'estimation de θ en faisant un choix différent de distribution *a priori*. Plus généralement, le fait d'avoir de bons renseignements *a priori* à propos de la taille de θ peut mener à des résultats améliorés.

Il n'est pas surprenant que l'estimateur par EGA soit biaisé vers le haut. Ce résultat s'explique par le fait que nous ignorons tous les échantillons par EGA qui ont moins de trois comptes supérieurs à zéro. S'ils étaient inclus, alors l'estimateur par EGA serait toujours sans biais. Par ailleurs, notre estimateur BAY peut avoir un biais vers le haut ou vers le bas. Il peut être biaisé vers le haut lorsque nous surestimons la taille de D_b. Il peut être biaisé vers le bas si nous sous-estimons la taille de D_b ou s'il y a un très petit réseau où les comptes sont beaucoup plus grands que les comptes dans les réseaux restants. Toutefois, ce dernier cas est également problématique pour l'estimateur par EGA. Nous présenterons une explication plus détaillée à ce sujet à la section 5.3.

Dans le tableau 4.1, nous pouvons observer que la population « paris » présente 43 unités ayant un compte supérieur à zéro dans 6 réseaux, et une valeur moyenne de 25,9. En vérifiant les réseaux, nous constatons qu'un seul réseau contient 31 unités, dont la valeur moyenne est 30,0. Le tableau 5.5 indique qu'en moyenne, nous avons observé 39,4 unités. Cela signifie que l'échantillon par EGA contenait presque toujours les unités de ce réseau. Cette observation aide à expliquer les excellents résultats obtenus pour cette population.

Une situation similaire se produit pour la population « devt ». Dans cette population, il y a 10 réseaux contenant 85 unités ayant un compte supérieur à zéro. Le total de ces unités est de 868. Les deux plus grands réseaux contiennent 31 et 23 unités, respectivement, et leur moyenne respective est de 11,3 et 12,8. Le réseau suivant le plus grand contient 13 unités et tous les autres ont 5 unités ou moins. D'après le tableau 5.4, nous constatons que le nombre moyen d'unités dans les échantillons par EGA finaux était de 65. Cela veut dire que la plupart des échantillons par EGA contenaient les deux plus grands réseaux. Ce résultat n'est pas surprenant, mais le fait que ces deux plus grands réseaux sont de bons échantillons représentatifs de D_b explique le bon rendement de nos estimateurs pour cette population.

Dans le tableau 4.1, nous constatons que, pour nos six populations, le rapport $\theta/N,$ soit la proportion d'unités ayant des comptes positifs dans la population, s'étend de 0,038 à 0,084. Il convient de rappeler que, pour un échantillon par EGA, le facteur ajusté $\lambda,$ défini dans l'équation (3.7), a été introduit pour augmenter notre estimation de la variance lorsque le nombre de comptes supérieurs à zéro dans l'échantillon par EGA était plutôt petit. Pour les populations « afrbuf » et « nfalls », les populations ayant les plus petites valeurs

de rapport θ/N, la valeur moyenne de λ était de 1,42 et 1,20 respectivement, alors que pour les deux plus grandes populations, « devt » et « paris », ces moyennes étaient de 1,01 et 1,02 respectivement. Ces résultats démontrent que notre choix de λ fonctionne comme prévu.

5.3 Six autres populations

Comme nous l'avons mentionné, nous avons choisi notre estimateur à partir d'études de simulation menées sur ces six populations. Ces populations ont été retenues parce qu'elles représentent une variété de situations pour lesquelles l'EGA serait utilisé. Un lecteur pourrait s'inquiéter du fait que notre estimateur était trop dépendant des populations que nous avons utilisées dans notre étude, même si ces populations étaient très différentes. Dans une tentative de démontrer le contraire, nous présentons maintenant six nouvelles populations qui n'ont pas été utilisées dans l'élaboration de notre estimateur. Nous avons construit les deux premières populations sur une grille de 400 carrés comportant seulement deux réseaux. Le premier réseau comptait trois membres ayant les valeurs 50, 60 et 70. Le deuxième réseau comptait douze membres ayant les valeurs 14, 15 et 16 qui apparaissaient chacune quatre fois. Supponsons que « ref1 » désigne cette population. Supponsons que « ref2 » désigne la population où les valeurs 14, 15 et 16 apparaissent dans le plus petit réseau et les valeurs 50, 60 et 70 apparaissent chacune quatre fois dans le plus grand réseau. Les tableaux 5.7 et 5.8 présentent les résultats pour 1 000 échantillons de taille 40, pour les deux populations. Le même ensemble d'échantillons a été produit pour les deux populations.

Nous constatons que dans les 798 échantillons ayant au moins deux comptes supérieurs à zéro, le nombre moyen de comptes observés était de 11,67. Cela signifie que dans la majorité de ces échantillons, seul le plus grand réseau a été observé. Mentionnons toutefois que nos résultats pour la population « ref1 » sont assez bons, même si de tels échantillons contiennent seulement les plus petits comptes. Notre estimation moyenne de θ était de 15,93, une légère surestimation. L'observation du plus grand réseau nous aide à compenser le fait que seulement les unités comprenant le plus petit nombre de comptes se trouvent dans l'échantillon. Pour la population « ref2 », cependant, le fait d'avoir les comptes les plus grands dans le réseau le plus grand signifie que notre estimateur est biaisé vers le haut, mais moins que l'estimation par EGA.

Tableau 5.7
Population ref1.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
EGA	464	0,290	269,6	223	892	1
BAY	330	-0,084	129,8	223	704	1

Notes : Pour 1 000 échantillons aléatoires simples de taille 40, 789 échantillons comportaient au moins trois comptes supérieurs à zéro. Le nombre total moyen de valeurs positives observées était 11,67.

Tableau 5.8
Population ref2.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
EGA	957	0,251	444	653	1 574	0,885
BAY	865	0,131	244	653	962	1,000

Notes : Pour 1 000 échantillons aléatoires simples de taille 40, 789 échantillons comportaient au moins trois comptes supérieurs à zéro. Le nombre total moyen de valeurs positives observées était 11,67.

Selon un raisonnement similaire à l'analyse de la population « ref2 » susmentionnée, nous pouvons expliquer le biais vers le haut de notre estimateur pour la population « nfalls » du tableau 5.2. Cette population comprend sept réseaux contenant un total de $\theta = 20$ unités. Dans le cas présent, un réseau contient 13 des unités et toutes les unités les plus importantes également. Le nombre moyen de comptes observés supérieurs à zéro était de 13,57, de sorte que le réseau contenant 13 unités était presque toujours dans l'échantillon. La valeur moyenne de notre estimateur pour θ était de 19,5, ce qui est assez bon. Toutefois, elle ne peut pas compenser le fait que les échantillons par EGA contenaient presque toujours les comptes les plus grands.

Rappelons que $\hat{T}_{ac}$ serait non biaisé si nous utilisions tous les échantillons possibles plutôt que seulement ceux qui ont au moins trois comptes supérieurs à zéro. Lorsque tous les comptes observés sont à zéro, alors l'estimation est de zéro. Cet événement se produira avec une probabilité positive, et sera une sous-estimation, sauf lorsque la population totale est effectivement de zéro. Pour compenser cette sous-estimation, $\hat{T}_{ac}$ fait une surestimation lorsque l'échantillon contient « beaucoup » de voisinages dont les comptes sont supérieurs à zéro. Ce résultat aide également à expliquer la raison pour laquelle BAY réussit mieux qu'EGA.

Pour le constater, nous observerons plus étroitement les deux populations « ref1 » et « ref2 ». D'abord, prenons les trois cas où seulement le deuxième réseau était soit observé une fois, deux fois ou trois fois. Pour la population « ref1 », dont le total est 360, les estimations correspondantes BAY (EGA) allaient de 226 (150) à 257 (450). Puis, prenons les trois cas où seulement le premier réseau était observé une fois, et le deuxième réseau était observé une fois, deux fois ou trois fois. Pour ces trois cas, les estimations allaient de 446 (750) à 500 (1 050). Pour la population « ref2 », dont la population totale est de 765, les valeurs correspondantes des estimations pour les trois premiers cas allaient de 903 (600) à 1 029 (1 800) et pour le deuxième ensemble de cas, les valeurs correspondantes des estimations allaient de 949 (750) à 1 063 (1 950). Soulignons que la fourchette de valeurs de l'estimateur par EGA est beaucoup plus grande que celle de l'estimateur BAY. Ce résultat explique également la raison pourquoi dans toutes nos simulations, $\hat{T}_{ac}$ est biaisé vers le haut. En effet, c'est parce que nous ignorons tous les échantillons qui ont moins de trois comptes supérieurs à zéro.

Nous décrirons maintenant les quatre dernières nouvelles populations, soit « mich », « rome », « afrdeer » et « suspt ». Les renseignements sommaires sur les quatre populations sont présentés au tableau 5.9. Les graphiques des quatre populations sont présentés dans les figures 5.2 à 5.5.

Tableau 5.9
Renseignements sommaires sur les nouvelles populations, soit « rome », « mich », « afrdeer » et « sunspt ».

Population	N	N_{ntw}	θ	θ/N	$y_{\max}$	T_b	T_b/θ
rome	1 600	16	48	0,03	495	7 927	165,1
mich	400	10	20	0,05	3 577	24 740	1 237
afrdeer	391	13	76	0,19	140	1 309	16,6
sunspt	500	14	27	0,05	98	434	16,1

Notes : Il convient de rappeler que N correspond à la taille de la population, θ correspond au nombre d'unités supérieures à zéro et que T_b est leur somme. Nous supposons que N_{ntw} est le nombre de réseaux dans la population et que $y_{\max}$ est la valeur maximale de y dans la population.

Figure 5.2 Graphique en trois dimensions de la population rome.

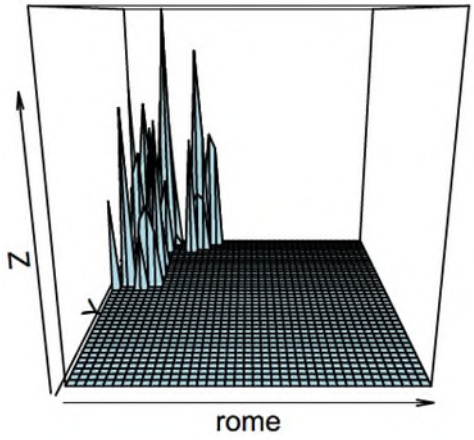


Figure 5.3 Graphique en trois dimensions de la population mich.

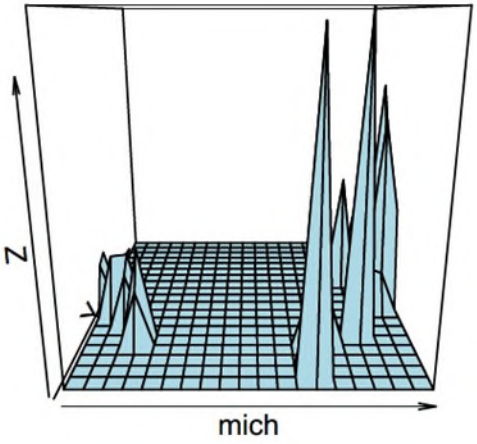


Figure 5.4 Graphique en trois dimensions de la population afrdeer.

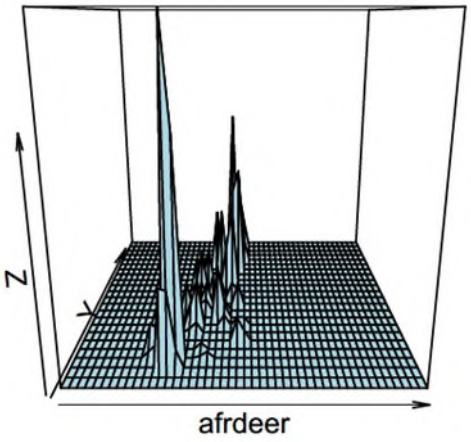
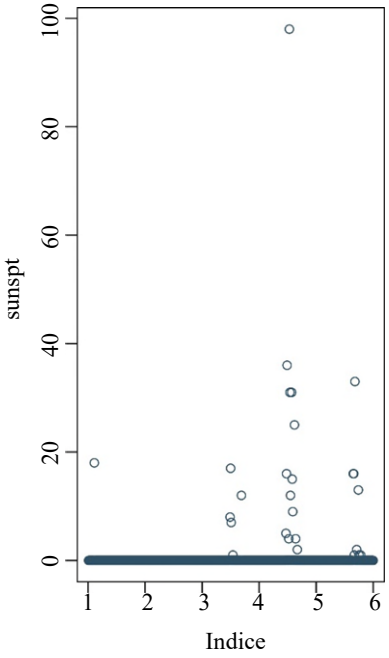


Figure 5.5 Graphique de la population sunspt.



La première population repose sur une grille de 40 par 40 à Rome, en Italie. La deuxième repose sur une grille de 20 par 20 dans l'état du Michigan, aux États-Unis. Nous désignons ces deux populations par « rome » et « mich », respectivement.

Pour la population « rome », nous observons 16 réseaux qui contiennent 48 comptes supérieurs à zéro. Quatre des réseaux, ayant une fourchette de taille de 6 à 11, contiennent 34 des 48 comptes supérieurs à zéro. La moyenne de ces réseaux s'étend sur une fourchette de 139,3 à 219. La moyenne de ces quatre moyennes est de 176,8, ce qui est assez proche de 165,1, la moyenne de tous les comptes supérieurs à zéro. Les 12 autres réseaux contiennent une ou deux valeurs et la majorité d'entre elles sont plutôt petites. Les deux plus grands ont un compte de 335 et de 301. Ces résultats expliquent l'excellent comportement de notre estimateur BAY dans le tableau 5.10. Et il y a une forte probabilité que l'échantillon par EGA contienne un des quatre plus grands réseaux.

Tableau 5.10
Population rome.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
EGA	8 206	0,035	2 793	4 044	11 107	0,923
BAY	7 459	-0,059	1 731	4 044	12 673	1,000

Notes : Pour 1 000 échantillons aléatoires simples de taille 160, 986 échantillons comportaient au moins trois comptes supérieurs à zéro. Le nombre total moyen de valeurs positives observées était 22,81.

Pour la population « mich », nous observons 10 réseaux qui contiennent 20 comptes supérieurs à zéro. De ces comptes, 11 se trouvent dans 6 réseaux où le compte maximal est de 695 et les six plus petits se trouvent dans une fourchette allant de 105 à 190. Le plus grand réseau contient 5 comptes et sa moyenne est de 969,2. Le plus grand compte est 3 077 et se trouve dans un réseau de taille un. Donc, bien qu'il n'y ait que quelques unités où les comptes sont supérieurs à zéro, les comptes réels peuvent être assez grands. Nous incluons cet exemple pour observer ce qui pourrait se produire dans un tel cas. Le grand biais négatif de l'estimateur BAY, dans le tableau 5.11, survient parce que le plus grand compte est un réseau de taille un. Même avec un tel biais, il présente une erreur absolue moyenne beaucoup plus petite que l'estimateur par EGA. Les bornes supérieures des deux estimateurs ne donnent toutefois pas de bons résultats. Il est difficile d'obtenir une borne supérieure sensée lorsqu'il y a un réseau comprenant une ou deux valeurs qui sont beaucoup plus grandes que les autres comptes de la population. Les deux produiront une sous-estimation si les grands comptes ne sont pas inclus dans l'échantillon et une surestimation s'ils le sont. À moins de disposer de renseignements supplémentaires *a priori*, nous croyons qu'il est très difficile d'obtenir des estimations sensées sans faire de l'échangeabilité approximative entre les comptes dans les réseaux lorsque nous observons uniquement un très petit nombre de comptes supérieurs à zéro.

Tableau 5.11
Population mich.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
EGA	32 657	0,320	14 333	8 065	63 114	0,964
BAY	16 805	-0,321	8 602	8 065	64 281	1,000

Notes : Pour 1 000 échantillons aléatoires simples de taille 40, 686 échantillons comportaient au moins trois comptes supérieurs à zéro. Le nombre total moyen de valeurs positives observées était 6,49.

Pour la nouvelle population suivante, nous avons étudié une troisième population donnée dans Gattone et coll. (2016). Nous désignons cette population « afrdeer », et les renseignements sommaires sont présentés dans le tableau 5.9. La majorité des unités appartiennent aux trois plus grands réseaux qui sont de taille 13, 17 et 21. Leur moyenne respective est de 29,3, 8,8 et 25,8. Le réseau le plus grand suivant est de taille 5 et les valeurs restantes sont majoritairement petites. La valeur la plus grande dans ces autres réseaux est de 34. Soulignons que, pour cette population, la valeur réelle de θ est 76, de sorte que le rapport $\theta/N = 76/391 = 0{,}19$ est supérieur à 0,15, soit la borne supérieure utilisée dans la définition de notre distribution *a priori*.

Donc, la question qui se pose est la suivante : de quelle façon notre estimateur fonctionne-t-il dans ce cas ? Les résultats de cette simulation sont présentés dans le tableau 5.12. Nous constatons que, comme dans les autres exemples, nous obtenons des résultats raisonnablement bons, voire meilleurs que l'estimateur par EGA. Ces résultats s'expliquent par le fait que les échantillons par EGA ont tendance à comprendre les réseaux les plus grands et que les réseaux les plus petits ont tendance à comprendre les plus petits comptes.

Tableau 5.12
Population afrdeer.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
EGA	1 307	-0,001	406	962	1 304	0,917
BAY	1 213	-0,073	214	962	1 269	0,970

Notes : Pour 1 000 échantillons aléatoires simples de taille 39, 1 000 échantillons comportaient au moins trois comptes supérieurs à zéro. Le nombre total moyen de valeurs positives observées était 49,76.

À la fin de la section 3.1, nous avons souligné que notre distribution *a posteriori* pour θ ne dépend pas du fait que nos unités de base sont des carrés et que la population est un rectangle. Toutefois, il est important que les unités de bases correspondent approximativement à la même taille et que les voisins puissent être définis d'une manière sensée. Pour démontrer ce point, nous examinons maintenant une population définie sur un ensemble d'intervalles successifs de longueur égale. Les voisins d'un intervalle sont simplement les deux intervalles adjacents, à gauche et à droite, à l'exception des points d'extrémité qui ont seulement un voisin. Pour obtenir les comptes, nous utilisons la population de taches solaires qui se trouve dans R (R Core Team, 2023). Il s'agit de la moyenne mensuelle du nombre relatif de taches solaires, soit une fourchette de 1 749 à 1 983, ayant le comportement de regroupement qui devrait rendre l'EGA approprié. Nous avons pris les 500 premières valeurs des taches solaires et nous avons soustrait 141 à chacune. Les valeurs négatives obtenues ont été rapportées à zéro. Les renseignements sommaires de cette nouvelle population, « sunspt », sont présentés au tableau 5.9. Dans la population « sunspt », il y a 14 réseaux : 8 de taille 1; 1 de taille 2; 3 de taille 3 et 2 de taille 4; donc, thêta = 27. La fourchette des 27 comptes supérieurs à zéro varie de 1 à 98. Le plus grand compte suivant est 36. Le plus grand compte apparaît dans un réseau de taille 4 ayant une moyenne de 36,25 qui est bien plus grande que $T_b/\theta =$ 16,1. Nous observons, dans le tableau 5.13, que le comportement des deux estimateurs est très similaire à celui que nous avons observé dans nos deux exemples dimensionnels. Le présent exemple démontre que notre approche pourrait être utile lors de l'étude de données longitudinales qui satisfont à nos hypothèses de rareté et de regroupement.

Tableau 5.13
Population sunspt.

	Est	Rbias	Abserr	Lowbd	Len	Freqcov
EGA	497	0,144	217,8	140	930	0,952
BAY	304	-0,299	159,3	140	1 369	0,999

Notes : Pour 1 000 échantillons aléatoires simples de taille 50, 874 échantillons comportaient au moins trois comptes supérieurs à zéro. Le nombre total moyen de valeurs positives observées était 7,29.

Les trois graphiques dimensionnels des populations « mich », « rome » et « afrdeer » sont présentés dans les figures 5.2, 5.3 et 5.4. Un graphique de « sunspt » est présenté à la figure 5.5.

6. Code R pour calculer notre estimateur

Dans le cas présent, nous avons défendu un estimateur quasi bayésien en particulier lorsqu'il y a peu de renseignements sur la population d'intérêt, à l'exception de la rareté de l'espèce d'intérêt. Mais en fait, nous

avons défini une famille entière d'estimateurs possibles en tenant compte de différents choix de paramètres dans la densité bêta et différents choix de définition de l'espace du paramètre Θ. L'un ou l'autre de ces estimateurs possibles est très simple à calculer en R (R Core Team, 2023). Nous présentons ci-dessous une fonction R qui calcule une de ces estimations pour le total de la population, ainsi que la borne supérieure et la borne inférieure. Il suffit de préciser l'espace de paramètre Θ, la distribution bêta qui sert à définir la distribution *a priori* sur Θ et certains des renseignements d'un échantillon par EGA complet. Il n'est pas nécessaire de connaître les arêtes des réseaux observés.

Plus formellement, voici ce qui doit être défini :

- n, la taille de l'échantillon aléatoire initial;
- nbiginrs, le nombre de comptes supérieurs à zéro dans l'échantillon aléatoire initial;
- bigsmp, tous les comptes supérieurs à zéro dans l'échantillon par EGA complet;
- bds, utilisé pour définir l'espace de paramètre Θ;
- alp et bet, les paramètres de la distribution bêta utilisée pour définir notre distribution *a priori*;
- N, la taille de la population.

Puis, nous donnons le code de la fonction

```
qbay<-function(n,nbiginrs,bigsmp,bds,alp,bet,N)
  {
     klw<-floor(bds[1]*N)
     kup<-ceiling(bds[2]*N)
     theta<-klw:kup
     nbigsmp<-length(bigsmp)
     dtheta<-theta[theta>=nbigsmp]
     ntheta<-length(dtheta)
     llike<-lchoose(dtheta,nbiginrs) + lchoose(N-dtheta,n-nbiginrs)
     lprior<-log(dbeta(dtheta/N,alp,bet))
     dum<-lprior + llike
     post<-rep(0,ntheta)
     for(i in 1:ntheta){
       post[i]<-1/sum(exp(dum-dum[i]))
     }
     pstmnth<-sum(dtheta*post)
     est<-pstmnth*mean(bigsmp)#the point estimate in equation (5.6)
     pst2ndmnth<-sum(dtheta^2*post)
     pstvrth<-pst2ndmnth - pstmnth^2
     nbig<-length(bigsmp)
     mnbg<-mean(bigsmp)
```

```
    lwbd<-nbig*mnbg
    vr1<-(mnbg)^2*pstvrth
    nb<-length(bigsmp)
    d1<-var(bigsmp)
    d2<-sum(post*(dtheta-nb)^2)
    vr2<-d1*d2
    vr<-vr1+vr2             #the variance in equation (5.7)
    lam<-10^(2.5/nbig)      #the adjustment factor in equation (5.8)
    upbd<-est+sqrt(vr)*lam*1.96
    ans<-c(est,lwbd,upbd)
    return(ans)
}
```

Cette fonction permet d'explorer un meilleur choix d'un estimateur quasi bayésien lorsque des renseignements supplémentaires sur la distribution *a priori* sont connus à propos de la population. Mentionnons qu'il est également facile de changer le facteur d'ajustement, λ, qui apparaît dans l'équation (3.7). Dans la prochaine section, nous aborderons brièvement certaines extensions possibles.

7. Extensions possibles

Thompson (1990) a présenté l'EGA pour les applications en biologie où l'objet d'étude était les espèces rares qui ont tendance à apparaître en grappes. Comme nous l'avons souligné dans la section 2.2, nous n'avons pas pu trouver de définition de « rare » dans la littérature. Toutefois, comme l'ont démontré nos simulations, notre approche convient bien à une vaste gamme de situations de rareté. Notre distribution *a priori* pour le nombre de comptes supérieurs à zéro ne dépend pas de leur ampleur. D'un certain angle, ce raisonnement est sensé parce que la notion de rareté peut dépendre de l'espèce à l'étude. Dans certains cas, nous pourrions disposer de bons renseignements *a priori* à propos du nombre de comptes supérieurs à zéro. Dans ce cas, il serait possible de sélectionner différentes valeurs de α et de β dans la distribution *a priori* afin de mieux refléter ces renseignements.

D'après nos connaissances, il n'y a pas de définition officielle pour le terme « regroupement » dans la littérature. Une forme extrême de regroupement serait où seulement une unité contient la totalité des éléments T_b de l'espèce d'intérêt. De toute évidence, cela ne correspond pas à l'esprit de l'EGA. Il nous semble que l'EGA suppose qu'un compte concentré aussi important est impossible. L'EGA présume plutôt qu'un compte aussi important aurait tendance à s'étendre aux carrés voisins et à former un réseau. Comme nous l'avons vu dans notre description de la population « thmp », au début de la section 4, ses trois réseaux ont cette forme. Lorsque la taille des réseaux tend à être plus grande et que les comptes dans un réseau sont représentatifs de l'ensemble des comptes dans la population, nous aurons de bons résultats.

Lorsque ce cas n'est pas avéré et qu'il y a un compte important par rapport au reste dans un petit réseau, alors les bornes supérieures de l'estimateur par EGA et de l'estimateur BAY ne donnent pas de bons

résultats. Nous l'avons constaté pour les populations « afrbuf » et « mich ». Notre choix de facteur d'ajustement λ dans l'équation (3.7) repose sur une hypothèse de regroupement implicite relativement à la portée des valeurs possibles pour les comptes et sur la vraisemblance d'avoir un petit réseau qui contiendrait des comptes extrêmement importants. Dans certains cas, il est possible que nous disposions de renseignements à propos de la définition d'un « grand » compte pour la population et de la proportion de ces « grands » comptes dans la population. Selon la taille des comptes dans l'échantillon par EGA, il sera possible de laisser le facteur d'ajustement, λ, dépendre de la valeur des comptes dans l'échantillon et des renseignements *a priori*. On pourrait aussi affaiblir notre hypothèse de l'échangeabilité approximative et utiliser les renseignements *a priori* pour remplacer la moyenne des comptes observés par un estimateur différent. Les renseignements *a priori* pourraient améliorer notre estimateur, mais ils donneront lieu à de moins bons estimateurs lorsqu'ils sont incorrects. Il faudrait entreprendre d'autres études pour déterminer des façons d'utiliser les connaissances *a priori* supplémentaires d'une manière plus formelle.

Rapley et Welsh (2008) ont élaboré des modèles intéressants pour l'étude du type de populations sur lesquelles l'EGA est utilisé. Ils ont notamment modélisé la façon de former les réseaux et les arêtes. Notre approche est beaucoup plus simple, parce que nous ignorons les réseaux et ne prenons que le nombre d'unités non vides dans la population. La distribution *a priori* que nous proposons pour la taille de D_b est très similaire à une distribution *a priori* qui est utilisée dans un contexte légèrement différent de leur modèle. Dans le cas présent, nous souhaitions démontrer qu'il est possible d'améliorer l'approche par EGA classique sans faire d'hypothèses de modèle. Il serait intéressant de comparer notre approche aux approches bayésiennes du problème. La question de savoir ce que l'approche bayésienne pourrait apporter de plus, par rapport à ce que nous avons présenté, lorsque nous disposons de bons renseignements *a priori* nécessite une étude plus poussée.

Rapley et Welsh ont également souligné que l'échantillonnage pourrait être fait d'une manière séquentielle. Selon cette méthode, une unité est sélectionnée au hasard dans la population. Si son compte est supérieur à zéro, il faut alors observer toutes les unités de son réseau et de ses arêtes. À chaque étape, nous avons sélectionné seulement une unité au hasard dans les unités non observées restantes. Nous continuons de cette manière jusqu'à ce que notre règle d'arrêt nous indique que l'échantillonnage est terminé. Notre approche peut être étendue à de tels plans d'échantillonnage séquentiel. Présumant que nous connaissons l'ordre dans lequel l'échantillon est pris, alors la forme de la fonction de vraisemblance changerait, mais il serait possible d'utiliser la même distribution *a priori* pour θ. D'un point de vue théorique, l'échantillonnage séquentiel est un choix sensé, mais il n'est pas certain que ce soit pratique dans bon nombre des problèmes pour lesquels l'EGA serait utilisé.

8. Conclusions

L'échantillonnage en grappes adaptatif a été proposé pour améliorer l'efficacité de l'échantillonnage de populations ayant un type particulier de structure. Il convient lorsque les statisticiens savent que la population ne comporte que quelques cellules ayant une valeur de y supérieure à zéro et que ces cellules

ont tendance à apparaître en grappes. Il s'agit d'une approche intéressante qui a été largement adoptée lors de l'étude de populations biologiques sur le terrain. Nous avons démontré, toutefois, qu'étant axée sur la recherche d'un estimateur qui est sans biais par rapport au plan, cette approche n'a pas tenu compte de certains des renseignements connus. Dans le cas présent, nous avons proposé, pour ce problème, une approche quasi bayésienne qui exploite ces renseignements. Nous avons démontré que notre estimation ponctuelle et notre limite de confiance supérieure de 95 % pour le total de la population produisaient de bien meilleurs résultats que l'approche classique.

Remerciements

Nous souhaitons remercier Gabriel Mersy pour son aide sur l'utilisation du progiciel *R* rgbif.

Bibliographie

Chamberlain, S., Ram, K., Mcglinn, D. et Barve, V. (2019). *rgbif: Interface to the Global Biodiversity Information Facility API*. https://CRAN.R-project.org/package=rgbif.

Cochran, W. (1977). *Sampling Techniques* (third ed.). New York: John Wiley & Sons, Inc.

Dryver, A., et Chao, C. (2007). Ratio estimators in adaptive cluster sampling. *Environmetrics*, 18, 607-620.

Gattone, S., Mohamed, E. et Di Battista, T. (2016). Adaptive cluster sampling with clusters selected without replacement and stopping rule. *Environmental and Ecological Statistics*, 23, 453-468.

Ghosh, J.K. (1988). Statistical Information and Likelihood: A Collection of Critical Essays by D. Basu. New York: Springer-Verlag.

Goncalves, K.C.M., et Moura, F.A.S. (2016). A mixture model for rare and clustered populations under adaptive cluster sampling. *Bayesian Analysis*, 11, 519-544.

Latpate, R., Kshirsagar, J., Gupta, V.K. et Chandra, G. (2021). Adaptive cluster sampling. *Advanced Samping Methods*, 125-156. Springer Singapore.

Meeden, G. (1999). Interval estimators for the population mean for skewed distributions with a small sample size. *Journal of Applied Statistics*, 26, 81-96.

Nolau, I., Goncalves, K.C.M. et Pereira, J.B.M. (2022). Model-based inference for rare and clustered populations from adaptive cluster sampling using auxiliary variables. *Journal of Survey Statistics and Methodology*, 10, 439-465.

Pacifici, K., Reich, B., Dorazio, R. et Conroy, M. (2016). Occupancy estimation for rare species, using a spatially-adaptive design. *Methods in Ecology and Evolution*, 7, 285-293.

R Core Team (2023). *R: A Language and Evironment for Statistical Computing*. R Foundation for Statistical Computing. http://R-project.org.

Rapley, V.E., et Welsh, A.H. (2008). Model-based inferences from adaptive cluster sampling. *Bayesian Analysis*, 3, 717-736.

Savitsky, T., et Toth, D. (2014). Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10, 1677-1708.

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

Thompson, S. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85, 1050-1059.

Thompson, S. (2012). *Sampling* (third ed.). Hoboken, New Jersey: Wiley.

Turk, P., et Borkowski, J. (2005). A review of adaptive cluster sampling: 1990-2003. *Ecological and Environmental Statistics*, 12, 55-94.

Techniques d'enquête, décembre 2024
Vol. 50, nº 2, pp. 259-285
Statistique Canada, nº 12-001-X au catalogue

Inférence à l'aide de l'échantillonnage s'appuyant sur les probabilités de réponse estimées par calage

Caren Hasler[1]

Résumé

Une solution permettant de corriger le biais de non-réponse consiste à multiplier les poids de sondage des répondants par l'inverse des probabilités de réponse estimées afin de compenser la non-réponse. Le maximum de vraisemblance et le calage sont deux approches qui peuvent être utilisées pour obtenir des probabilités de réponse estimées. Nous étudions un cadre commun permettant de comparer ces deux approches. Nous élaborons une étude asymptotique du comportement de l'estimateur résultant lorsque le calage est appliqué. Un modèle de régression logistique des probabilités de réponse est supposé. Les données manquantes au hasard et les données sans mise en grappes sont présumées. Les trois principales contributions de ce travail sont les suivantes : 1) nous démontrons que les estimateurs s'appuyant sur les probabilités de réponse estimées par calage sont asymptotiquement équivalents aux estimateurs sans biais et qu'un gain d'efficacité est obtenu lors de l'estimation des probabilités de réponse par calage, par rapport à l'estimateur s'appuyant sur les probabilités de réponse réelles; 2) nous démontrons que les estimateurs s'appuyant sur les probabilités de réponse estimées par calage sont doublement robustes à une mauvaise spécification des modèles sous-jacents et expliquons la raison pour laquelle la double robustesse n'est pas garantie lorsque le maximum de vraisemblance est appliqué; 3) nous soulignons les problèmes liés à l'estimation des probabilités de réponse, à savoir l'existence d'une solution aux équations estimantes, les problèmes de convergence et les poids extrêmes. Nous présentons les résultats d'une étude par simulation pour illustrer ces éléments.

Mots-clés : Ajustement de la pondération; estimation à deux degrés; estimation du maximum de vraisemblance; non-réponse.

1. Introduction

Dans les cas de réponse complète, l'estimateur de Horvitz-Thompson (HT) est sans biais (Horvitz et Thompson, 1952). Dans les cas de non-réponse, toutefois, cet estimateur n'est pas disponible. La non-réponse peut être considérée comme une deuxième phase de l'enquête, où le mécanisme qui produit la non-réponse, appelé *mécanisme de réponse*, est inconnu (Oh et Scheuren, 1983; Särndal et Swensson, 1987). Si les probabilités de réponse étaient connues, un estimateur à deux degrés s'appuyant sur les probabilités de réponse comme les probabilités d'inclusion de la deuxième phase serait sans biais. Malheureusement, dans la pratique, les probabilités de réponse sont inconnues. Une solution pour corriger le biais de non-réponse consiste à postuler un modèle pour les probabilités de réponse, à estimer ces probabilités en fonction du modèle postulé et à inclure ces probabilités de réponse estimées en lieu et place des vraies probabilités de réponse dans l'estimateur à deux degrés. L'estimateur qui en résulte est appelé *estimateur ajusté par pondération pour la non-réponse (EAPNR) à deux degrés* ou *estimateur empirique par double expansion*. Särndal et Lundström (2005) ainsi que Haziza et Beaumont (2017) présentent des vues d'ensemble de certains EAPNR et des systèmes de pondération ajustés pour la non-réponse.

1. Caren Hasler, Institut de Statistique, Université de Neuchâtel, Av. de Bellevaux 51, 2000 Neuchâtel. Courriel : caren.hasler@unine.ch.

L'estimation du maximum de vraisemblance (EMV) et le calage (Deville et Särndal, 1992) sont deux approches possibles pour les EAPNR. Dans la première approche, un modèle pour les probabilités de réponse, comme le modèle de régression logistique, est postulé (Cassel, Särndal et Wretman, 1983; Ekholm et Laaksonen, 1991). Les paramètres du modèle sont estimés par la méthode d'EMV et les probabilités de réponse estimées sont obtenues en insérant les paramètres estimés dans le modèle choisi. Dans la deuxième approche, les poids de sondage sont modifiés de sorte que l'EAPNR résultant pour des variables auxiliaires est égal à l'estimateur de HT de ces variables auxiliaires (calage au niveau de l'échantillon complet) ou à leur total sur la population (calage au niveau de la population). Les poids de calage peuvent être considérés comme les poids de sondage multipliés par l'inverse des probabilités de réponse estimées. D'après nos connaissances, le premier auteur ayant suggéré l'utilisation de ce qui serait plus tard appelé la pondération par calage pour estimer les probabilités de réponse est Folsom (1991), suivi peu après par Deville et Dupont (1993) et Dupont (1993). Lundström et Särndal (1999) ont poursuivi l'étude de l'estimateur ponctuel et de l'estimateur de variance pour les deux niveaux de calage, à savoir le calage au niveau de l'échantillon et au niveau de la population.

La première approche est étudiée en profondeur dans Kim et Kim (2007), qui présente les propriétés asymptotiques de l'EAPNR dans le cadre d'un modèle de réponse général. Les deux principaux résultats de leur article sont les suivants : 1) l'EAPNR s'appuyant sur les probabilités de réponse estimées par maximum de vraisemblance est asymptotiquement équivalent à un estimateur sans biais; 2) un gain d'efficacité est obtenu lorsque les probabilités de réponse sont estimées par maximum de vraisemblance par rapport à l'estimateur s'appuyant sur les probabilités de réponse réelles. Le deuxième résultat est également présenté par Beaumont (2005) dans le cadre du modèle de réponse logistique.

La deuxième approche peut être divisée en deux niveaux : le calage au niveau de l'échantillon et le calage au niveau de la population. L'EAPNR obtenu lorsque les probabilités de réponse sont estimées par calage au niveau de l'échantillon est un cas particulier de l'estimateur ajusté par le score de propension de Kim et Riddles (2012). Ces auteurs ont élaboré les propriétés asymptotiques de cet estimateur dans un cadre théorique différent de celui examiné dans Kim et Kim (2007). Cet estimateur est également étudié dans Iannacchione, Milne et Folsom (1991), qui porte sur les aspects pratiques de l'estimation ajustée par pondération pour la non-réponse à l'aide du calage au niveau de l'échantillon. Aucune théorie n'y est présentée.

L'objectif principal des deux approches est de réduire le biais de non-réponse et, si possible, la variance des estimateurs de population. La deuxième approche, le calage, assure également la cohérence entre les totaux de population estimés et des totaux connus, ce qui n'est pas le cas de la première approche, l'EMV. Toutefois, la deuxième approche, c'est-à-dire l'estimation directe des probabilités de réponse par calage, appelée *approche en une étape*, est parfois critiquée, puisqu'elle tend à produire des estimations biaisées lorsque le modèle de réponse n'est pas spécifié correctement (Haziza et Lesage, 2016). Une solution à ce problème consiste à estimer la probabilité de réponse par maximum de vraisemblance, puis à appliquer le calage sur les poids corrigés au moyen des probabilités de réponse estimées pour assurer la cohérence entre les totaux estimés et les totaux connus. Cette solution s'appelle *approche en deux étapes*. Le lecteur est invité à consulter Haziza et Lesage (2016) et Haziza et Beaumont (2017), à la page 222, pour obtenir une analyse de ces deux approches.

Dans le présent article, nous étudions l'EMV et l'approche en une étape de calage pour l'ajustement de la pondération pour la non-réponse. Nous nous appuyons sur Kim et Kim (2007) et élaborons les propriétés asymptotiques de l'EAPNR obtenu au moyen du calage au niveau de l'échantillon et de la population. Pour la première fois, un cadre théorique commun est étudié pour les deux approches d'estimation ajustée par pondération pour la non-réponse, à savoir l'EMV et le calage. Ce cadre nous permet de comparer le comportement asymptotique des EAPNR résultants en terme de biais et de variance. Nous postulons un modèle de régression logistique pour les probabilités de réponse. Nous supposons que les données sont manquantes au hasard (voir Rubin (1976) pour obtenir une définition détaillée) et sans mise en grappes. Les deux principaux résultats théoriques obtenus sont : 1) les EAPNR s'appuyant sur les probabilités de réponse estimées par calage sont asymptotiquement équivalents à des estimateurs sans biais; 2) un gain d'efficacité est obtenu lorsque les probabilités de réponse sont estimées par calage par rapport à l'estimateur s'appuyant sur les probabilités de réponse réelles. Ces résultats sont valides pour les deux niveaux de calage.

L'étude de la double robustesse des EAPNR est une autre contribution principale de la présente étude. En effet, les deux approchent présument, implicitement ou explicitement, deux modèles : 1) un modèle qui lie la variable d'intérêt et les variables auxiliaires, appelé *modèle de superpopulation*; 2) un modèle pour les probabilités de réponse, appelé *modèle de réponse*. Nous démontrons que les EAPNR s'appuyant sur les probabilités de réponse estimées par calage sont doublement robustes. Ces estimateurs sont consistants même si l'un des deux modèles mentionnés précédemment est mal spécifié. Nous expliquons également la raison pour laquelle la double robustesse de l'EAPNR s'appuyant sur les probabilités de réponse estimées par maximum de vraisemblance n'est pas garantie. D'après nos connaissances, seuls Kott et Liao (2012) ont analysé la double robustesse de l'EAPNR par calage dans les enquêtes par échantillonnage probabiliste. Dans leur article, ils mettent l'accent sur une forme exponentielle des probabilités de réponse. Finalement, le présent travail apporte une dernière contribution importante : une analyse portant sur les problèmes de convergence et les poids extrêmes. En effet, il pourrait arriver que les équations estimantes utilisées pour obtenir les probabilités de réponse estimées n'admettent pas de solution ou qu'une solution à ces équations existe, mais que les poids résultants, soit l'inverse des probabilités de réponse estimées, soient très grands. Nous illustrons ce phénomène. Les résultats d'une étude par simulation confirment les résultats théoriques et les considérations pratiques présentés. Une version plus longue de l'article comportant des éléments techniques et pratiques supplémentaires est accessible sur ArXiv à partir du lien suivant :

http://doi.org/10.48550/arXiv.2202.03897 (en anglais).

Le présent article est structuré comme suit : la section 2 contient des éléments de notation et d'importants concepts. Dans la section 3, nous présentons les deux approches d'estimation des probabilités de réponse. Dans la section 4, nous décrivons certaines propriétés asymptotiques des EAPNR d'intérêt. L'annexe de la version plus longue de l'article (Hasler, 2023) contient certains éléments techniques. Dans la section 5, nous analysons la double robustesse par rapport à l'erreur de spécification du modèle. Dans la section 6, nous présentons la variance et l'estimation de la variance des EAPNR d'intérêt. La section 7 contient les résultats d'une étude par simulation. Une analyse conclut l'article dans la section 8.

2. Cadre

Supposons une population finie $U = \{1, 2, \ldots, i, \ldots, N\}$ de taille N. Un vecteur de v variables auxiliaires $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iv})$ est associé à une unité générique i. Nous présumons que la première variable auxiliaire est constante et égale à 1. Le paramètre d'intérêt est le total sur la population

$$Y = \sum_{i \in U} y_i,$$

pour une variable d'intérêt y. Un échantillon s de taille n est sélectionné dans U au moyen d'un plan d'échantillonnage probabiliste non informatif $p(\cdot)$ dans le but d'observer y_i pour $i \in s$. Un échantillon aléatoire S est une variable aléatoire étant telle que $\Pr(S = s) = p(s)$. L'échantillon aléatoire est également défini à l'aide d'une variable indicatrice $(a_i \mid i \in U)^\top$, où a_i vaut 1 si l'unité i est dans l'échantillon et 0 sinon. Considérons que

$$\pi_i = \Pr(i \in S) = \sum_{s \subset U;\, s \ni i} p(s),$$

est la probabilité d'inclusion d'ordre un de l'unité i, et supposons que $\pi_i > 0$ pour tout $i \in U$. Soient $\mathrm{E}_p(\cdot)$ et $\mathrm{V}_p(\cdot)$ l'espérance et la variance calculées par rapport au plan d'échantillonnage $p(\cdot)$. En cas de réponse complète, l'estimateur de HT (Horvitz et Thompson, 1952)

$$\hat{Y}_\pi = \sum_{i \in S} \frac{y_i}{\pi_i} \tag{2.1}$$

est sans biais par rapport au plan pour Y, c'est-à-dire $\mathrm{E}_p(\hat{Y}) = Y$.

En cas de non-réponse, chaque unité échantillonnée $i \in S$ est catégorisée comme étant soit *répondant* soit *non-répondant* selon que y_i est observé ou manquant. Considérons le vecteur de variables indicatrices de réponse $(r_i \mid i \in S)^\top$, où r_i prend la valeur de 1 si y_i est observé et de 0 s'il est manquant et $p_i = \Pr(r_i = 1 \mid i \in S)$ la probabilité de réponse d'une unité échantillonnée i. L'ensemble des répondants s'écrit $S_r = \{i \in S \mid r_i = 1\}$ et sa taille est n_r. En présence de non-réponse, l'estimateur de HT de l'équation (2.1) n'est pas disponible. On pourrait envisager d'estimer le total Y au moyen de l'estimateur *à deux degrés* (ou *par double expansion*)

$$\hat{Y}_p = \sum_{i \in S_r} \frac{y_i}{\pi_i p_i}, \tag{2.2}$$

si les probabilités de réponses p_i étaient connues et strictement positives pour tout $i \in S$. Cet estimateur est sans biais, puisque

$$\mathrm{E}_p\left\{E_q(\hat{Y}_p \mid S)\right\} = Y,$$

où $q(\cdot \mid S)$ est la distribution de probabilité de S_r étant donné un échantillon S. L'indice q indique que l'espérance est calculée par rapport à la distribution de probabilité $q(\cdot \mid S)$. Dans la pratique, les probabilités

de réponse sont inconnues. Pour régler ce problème, un modèle pour les probabilités de réponse, appelé *modèle de réponse*, est postulé. Les probabilités de réponse sont estimées à l'aide de ce modèle, qui produit les probabilités de réponse estimées $\hat{p}_i$, et l'*EAPNR* (ou l'*estimateur empirique par double expansion*)

$$\hat{Y}_{\hat{p}} = \sum_{i \in S_r} \frac{y_i}{\pi_i \hat{p}_i} \tag{2.3}$$

est utilisé. Les probabilités de réponse sont estimées à l'aide de $\hat{p}_i = f(\mathbf{x}_i; \hat{\boldsymbol{\lambda}})$ pour un certain modèle $f(\mathbf{x}_i; \boldsymbol{\lambda})$ et un certain estimateur $\hat{\boldsymbol{\lambda}}$ de $\boldsymbol{\lambda}$. Un modèle couramment utilisé pour les probabilités de réponse est le modèle de régression logistique

$$p_i = f(\mathbf{x}_i; \boldsymbol{\lambda}) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\lambda})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\lambda})} = \frac{1}{1 + \exp(-\mathbf{x}_i^\top \boldsymbol{\lambda})}, \tag{2.4}$$

où $\boldsymbol{\lambda}$ est un vecteur de paramètres à estimer. Deux méthodes d'estimation disponibles sont le maximum de vraisemblance et le calage (voir la section 3). Il convient de souligner qu'il y a des façons d'utiliser la pondération par calage pour ajuster la non-réponse autres que celles reposant sur un modèle de réponse logistique comme présenté ci-dessus. Par exemple, d'autres méthodes s'appuient sur une fonction linéaire ou logit qui limite les probabilités de réponse à des valeurs entre 0 et 1. Pour obtenir de plus amples renseignements, veuillez consulter les articles de Deville et Särndal (1992), de Deville, Särndal et Sautory (1993), et de Haziza et Beaumont (2017), entre autres. Dans le présent article, nous nous concentrons sur le modèle de régression logistique de l'équation (2.4).

Certaines hypothèses requises pour le mécanisme de réponse sont les suivantes :

(R1) : Les unités répondent indépendemment les unes des autres, c'est-à-dire

$$\Pr(i, j \in S_r \mid i, j \in S) = p_i p_j.$$

(R2) : Les probabilités de réponse admettent une limite inférieure, c'est-à-dire qu'il existe une constante $c > 0$ telle que $p_i > c$ pour tout $i \in U$.

(R3) : Les probabilités de réponse sont $p_i = f(\mathbf{x}_i, \boldsymbol{\lambda}^0)$ où la fonction f est définie à l'équation (2.4) pour un certain vecteur de paramètres inconnu $\boldsymbol{\lambda}^0$.

L'hypothèse (R3) présume que les données sont manquantes au hasard (voir Rubin (1976) pour obtenir une définition détaillée). Par conséquent

$$\Pr(i \in S_r \mid i \in S, \mathbf{x}_i, y_i) = \Pr(i \in S_r \mid i \in S, \mathbf{x}_i).$$

La propension à répondre est indépendante de la variable d'intérêt lorsque les variables auxiliaires sont prises en compte. Cette hypothèse peut être erronée, dans la pratique, lorsque la propension à répondre dépend encore de la variable d'intérêt même lorsque toute l'information auxiliaire disponible a été prise en compte. Dans un tel cas, il est possible d'utiliser le calage généralisé (Deville, 2002; Kott, 2006; Lesage,

Haziza et D'Haultfoeuille, 2019; Ranalli, Matei et Neri, 2023) pour estimer les probabilités de réponse au lieu des approches présentées dans la section 3.

3. Estimation

Nous considérons deux approches pour obtenir l'EAPNR : l'EMV et le calage (Deville et Särndal, 1992). Kim et Kim (2007) ont étudié les EAPNR par l'EMV des probabilités de réponse dans le cadre d'un modèle de réponse général. Pour le modèle de régression logistique, l'EMV de $\boldsymbol{\lambda}^0$ est la solution $\hat{\boldsymbol{\lambda}}^{\text{EMV}}$ de l'équation estimante

$$Q^{\text{EMV}}(\hat{\boldsymbol{\lambda}}) = \sum_{i \in S} k_i \left\{ r_i - f(\mathbf{x}_i; \hat{\boldsymbol{\lambda}}) \right\} \mathbf{x}_i = 0. \tag{3.1}$$

Lorsque $k_i = 1,$ la solution est l'EMV habituel. Lorsque $k_i = 1/\pi_i,$ nous obtenons une équation estimante pondérée par les poids de sondage. Cette alternative est communément appelée *pseudo-maximum de vraisemblance*. L'idée est d'estimer sans biais l'équation estimante de vraisemblance de la population, puis de maximiser l'équation estimante estimée. Il est possible de faire d'autres choix pour k_i. Nous nous concentrons sur les deux choix courants mentionnés précédemment. Un gain en terme d'efficacité de l'EAPNR dans l'équation (2.3), par rapport à l'estimateur à deux degrés dans l'équation (2.2) s'appuyant sur les probabilités de réponse réelles, est obtenu lorsque $k_i = 1$ (Beaumont, 2005; Kim et Kim, 2007). Ce choix produit la meilleure estimation de $\boldsymbol{\lambda}^0$ et la meilleure estimation des probabilités de réponse. L'efficacité de l'EAPNR peut toutefois être améliorée en faisant d'autres choix pour $k_i,$ comme $k_i = 1/\pi_i$. Les sources disponibles portant sur ce choix sont très limitées. Kott (2012) a analysé ce choix et l'incidence sur l'efficacité de l'EAPNR pour le cas des groupes de réponses homogènes. Aucune théorie ou ligne directrice générale relative au choix de k_i n'ont encore été suggérées dans la littérature. Ce sujet dépasse la portée du présent article.

Deux niveaux de calage sont possibles : le calage au niveau de l'échantillon et le calage au niveau de la population. Dans le premier cas, l'estimateur par calage de $\boldsymbol{\lambda}^0$ est la solution $\hat{\boldsymbol{\lambda}}^{\text{cal},S}$ de l'équation estimante

$$Q^{\text{cal},S}(\hat{\boldsymbol{\lambda}}) = \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i f(\mathbf{x}_i; \hat{\boldsymbol{\lambda}})} - \sum_{i \in S} \frac{\mathbf{x}_i}{\pi_i} = 0. \tag{3.2}$$

L'équation estimante (3.2) est suggérée dans Iannacchione et coll. (1991). Dans le deuxième cas, l'estimateur par calage de $\boldsymbol{\lambda}^0$ est la solution $\hat{\boldsymbol{\lambda}}^{\text{cal},U}$ de l'équation estimante

$$Q^{\text{cal},U}(\hat{\boldsymbol{\lambda}}) = \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i f(\mathbf{x}_i; \hat{\boldsymbol{\lambda}})} - \sum_{i \in U} \mathbf{x}_i = 0. \tag{3.3}$$

Les deux équations estimantes (3.2) et (3.3) peuvent être résolues en utilisant un logiciel de calage pour réponse complète, comme la fonction *calib* du module *d'échantillonnage* R (Tillé et Matei, 2021).

Avec le calage au niveau de l'échantillon, le but est d'obtenir des probabilités de réponse de sorte que le total estimé ajusté par pondération pour la non-réponse de certaines variables auxiliaires est égal à leur estimateur de HT. Avec le calage au niveau de la population, le but est d'obtenir des probabilités de réponse de sorte que le total estimé ajusté par pondération pour la non-réponse de certaines variables auxiliaires est égal à leur total sur la population. Par conséquent, la première approche tente de corriger l'erreur de non-réponse. La deuxième approche tente de corriger l'erreur de non-réponse et l'erreur d'échantillonnage.

Les deux approches, l'EMV et le calage, sont appliquées dans l'article pour estimer les probabilités de réponse utilisées dans l'EAPNR de l'équation (2.3). Elles diffèrent toutefois quant à leur esprit et à l'information requise dans le processus d'estimation. L'esprit de l'EMV est de maximiser la vraisemblance des données générées par le modèle de réponse postulé. L'objectif est l'estimation des probabilités de réponse sans avoir en tête un paramètre d'intérêt spécifique. De plus, l'EMV ne présume pas explicitement un modèle de superpopulation, c'est-à-dire un modèle qui lie la variable d'intérêt et les variables auxiliaires. Toutefois, nous verrons à la section 4 que l'EMV présume un modèle de superpopulation implicite. L'esprit du calage est d'estimer le total de certaines variables auxiliaires aussi précisément que possible. Le biais de non-réponse du total de la variable d'intérêt est d'autant plus petit que la corrélation entre la variable d'intérêt et les variables auxiliaires est haute. Le calage cible donc un paramètre d'intérêt particulier, à savoir le total, et présume explicitement un modèle de superpopulation, à savoir un modèle de régression linéaire.

Les deux approches diffèrent également quant à l'information nécessaire lors du processus d'estimation. L'EMV requiert de connaître les valeurs $\mathbf{x}_i$ pour toutes les unités échantillonnées $i \in S$. Le calage au niveau de l'échantillon obtenu au moyen de l'équation estimante (3.2) requiert de connaître les valeurs $\mathbf{x}_i$ de toutes les unités répondantes $i \in S_r$ et l'estimateur de HT de $\mathbf{x}_i$ au niveau de l'échantillon. Le calage au niveau de la population obtenu au moyen de l'équation estimante (3.3) requiert de connaître les valeurs $\mathbf{x}_i$ de toutes les unités répondantes $i \in S_r$ et le total de $\mathbf{x}_i$ sur la population. Pour l'EMV et le calage au niveau de l'échantillon, aucune information n'est nécessaire à propos du $\mathbf{x}_i$ hors de l'échantillon.

Nous comparons quatre EAPNR : 1) $\hat{Y}_{\hat{p}}^{\mathrm{EMV},1}$ obtenu à l'aide des probabilités de réponse estimées au moyen de l'équation (3.1), où $k_i = 1$; 2) $\hat{Y}_{\hat{p}}^{\mathrm{EMV},1/\pi}$ obtenu à l'aide des probabilités de réponse estimées au moyen de l'équation (3.1), où $k_i = 1/\pi_i$; 3) $\hat{Y}_{\hat{p}}^{\mathrm{cal},S}$ obtenu à l'aide des probabilités de réponse estimées au moyen de l'équation (3.2); 4) $\hat{Y}_{\hat{p}}^{\mathrm{cal},U}$ obtenu à l'aide des probabilités de réponse estimées au moyen de l'équation (3.3).

4. Analyse asymptotique I

4.1 Cadre théorique

Dans la présente section, nous nous appuyons sur les résultats et les hypothèses de Kim et Kim (2007) pour obtenir certaines propriétés asymptotiques des EAPNR obtenus par calage. Nous utilisons le cadre asymptotique d'Isaki et Fuller (1982). Considérons une séquence U_N de populations finies emboîtées de taille N, où N augmente à l'infini. Considérons une séquence d'échantillons s_N sélectionnés dans U_N

selon un plan d'échantillonnage $p_N(\cdot)$. Les probabilités d'inclusion d'ordre un et d'ordre deux associées à $p_N(\cdot)$ pour des unités génériques i et j sont $\pi_{N,i}$ et $\pi_{N,ij}$, respectivement. Dans les équations qui suivent, nous omettrons l'indice N autant que possible pour simplifier la notation. Nous considérons les conditions de régularité suivantes de la séquence de plans d'échantillonnage pour assurer la consistance de l'estimateur de HT et de son estimateur de variance.

(D1) : Lorsque $N \to +\infty$, nous avons $n/N \to \pi^* \in (0,1)$,

(D2) : Pour tout N, $\pi_i > \lambda_1 > 0$ pour tout $i \in U$,

(D3) : Pour tout N, $\pi_{ij} > \lambda_2 > 0$ pour tout $i, j \in U$,

(D4) : $\limsup_{N\to+\infty} n \max_{i,j\in U, i\neq j} \left|\pi_{ij} - \pi_i\pi_j\right| < +\infty$.

La notation $\limsup$ dénote la limite supérieure. Elle est définie comme la limite de la séquence de bornes supérieures. Dans le cas de (D4), nous pouvons écrire

$$\limsup_{N\to+\infty} n \max_{i,j\in U, i\neq j} \left|\pi_{ij} - \pi_i\pi_j\right| = \lim_{N\to+\infty} \sup\left\{u_k \middle| k \geq N\right\},$$

où

$$u_k = n_k \max_{i,j\in U_k, i\neq j} \left|\pi_{k,ij} - \pi_{k,i}\pi_{k,j}\right|,$$

et n_k est la taille de s_k. L'hypothèse (D4) présume que la dépendance entre les indicatrices d'inclusion dans l'échantillon est relativement petite (Breidt et Opsomer, 2017). Intuitivement, si nous considérons

$$n \max_{i,j\in U, i\neq j} \left|\pi_{ij} - \pi_i\pi_j\right|$$

comme une mesure de dépendance entre les indicatrices d'inclusion dans l'échantillon, cette mesure ne devrait pas augmenter vers l'infini. Par exemple, cette hypothèse est vérifiée pour l'échantillonnage aléatoire simple sans remise, l'échantillonnage de Bernoulli et tout échantillonnage stratifié qui n'est pas fortement stratifié. Cette hypothèse n'est pas vérifiée pour l'échantillonnage en grappes ou les plans d'échantillonnage fortement stratifiés. La prochaine section résume les résultats de Kim et Kim (2007) relatifs à l'asymptotique de l'EAPNR lorsque le maximum de vraisemblance est appliqué pour obtenir des probabilités de réponse estimées. Dans les deux sections qui suivent, nous étendons les résultats au cas où le calage est utilisé. Dans ce qui suit, la distribution de la probabilité de référence pour la convergence est celle conjointement définie par le mécanisme d'échantillonnage et le mécanisme de réponse.

4.2 Maximum de vraisemblance

À partir du théorème 1 de Kim et Kim (2007), nous constatons que selon les conditions de régularité (D1) à (D4), les hypothèses (R2) et (R3) relatives au mécanisme de réponse, et les conditions de régularité

supplémentaires énoncées dans l'annexe de la version plus longue du présent article (Hasler, 2023), l'EAPNR $\hat{Y}_{\hat{p}}^{\mathrm{EMV}}$ satisfait

$$\frac{1}{N}\hat{Y}_{\hat{p}}^{\mathrm{EMV}} = \frac{1}{N}\hat{Y}_{\hat{p},l}^{\mathrm{EMV}} + O_p(n^{-1}),$$

où

$$\begin{aligned}\hat{Y}_{\hat{p},l}^{\mathrm{EMV}} &= \sum_{i\in S}\frac{1}{\pi_i}\left\{k_i\pi_i p_i \mathbf{x}_i^\top \boldsymbol{\gamma}_n^{\mathrm{EMV}} + \frac{r_i}{p_i}\left(y_i - k_i\pi_i p_i \mathbf{x}_i^\top \boldsymbol{\gamma}_n^{\mathrm{EMV}}\right)\right\},\\ \boldsymbol{\gamma}_n^{\mathrm{EMV}} &= \left\{\sum_{i\in S} k_i p_i(1-p_i)\,\mathbf{x}_i\mathbf{x}_i^\top\right\}^{-1}\sum_{i\in S}\frac{1-p_i}{\pi_i}\,\mathbf{x}_i y_i.\end{aligned}$$

Remarque 1. *L'EAPNR $\hat{Y}_{\hat{p}}^{\mathrm{EMV}}$ se comporte asymptotiquement comme l'estimateur linéarisé $\hat{Y}_{\hat{p},l}^{\mathrm{EMV}}$, qui est sans biais pour le total sur la population Y.*

Remarque 2. *S'il existe un vecteur $\boldsymbol{\beta}$ tel que $y_i = k_i\pi_i p_i \mathbf{x}_i^\top \boldsymbol{\beta}$ pour tout $i \in S$, alors*

$$\hat{Y}_{\hat{p},l}^{\mathrm{EMV}} = \sum_{i\in S}\frac{y_i}{\pi_i}.$$

Par conséquent, $\hat{Y}_{\hat{p}}^{\mathrm{EMV}}$ est asymptotiquement équivalent à l'estimateur de HT de l'échantillon complet dans ce cas. Lors de l'estimation des probabilités de réponse par maximum de vraisemblance, selon l'équation (3.1), nous présumons implicitement un modèle de superpopulation, c'est-à-dire que y_i est une combinaison linéaire de $k_i\pi_i p_i \mathbf{x}_i$.

4.3 Calage au niveau de l'échantillon

Résultat 1. *Présumons que la séquence de plans d'échantillonnage satisfait les hypothèses (D1) à (D4), que le mécanisme de réponse satisfait les hypothèses (R1) à (R3) et que la séquence de populations finies satisfait les conditions de régularité énoncées dans l'annexe de la version plus longue du présent article (Hasler, 2023). L'EAPNR $\hat{Y}_{\hat{p}}^{\mathrm{cal},S}$ satisfait*

$$\frac{1}{N}\hat{Y}_{\hat{p}}^{\mathrm{cal},S} = \frac{1}{N}\hat{Y}_{\hat{p},l}^{\mathrm{cal},S} + O_p(n^{-1}),$$

où

$$\begin{aligned}\hat{Y}_{\hat{p},l}^{\mathrm{cal},S} &= \sum_{i\in S}\frac{1}{\pi_i}\left\{\mathbf{x}_i^\top \boldsymbol{\gamma}_S + \frac{r_i}{p_i}\left(y_i - \mathbf{x}_i^\top \boldsymbol{\gamma}_S\right)\right\},\\ \boldsymbol{\gamma}_S &= \left(\sum_{i\in S}\frac{1-p_i}{\pi_i}\,\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}\sum_{i\in S}\frac{1-p_i}{\pi_i}\,\mathbf{x}_i y_i.\end{aligned}$$

La preuve est fournie dans l'annexe de la version plus longue du présent article (Hasler, 2023).

Remarque 3. *L'EAPNR* $\hat{Y}_{\hat{p}}^{\text{cal},S}$ *se comporte asymptotiquement comme l'estimateur linéarisé* $\hat{Y}_{\hat{p},l}^{\text{cal},S}$, *qui est sans biais pour le total sur la population* Y.

Remarque 4. *S'il existe un vecteur* $\boldsymbol{\beta}$ *tel que* $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ *pour tout* $i \in S$, *alors*

$$\hat{Y}_{\hat{p},l}^{\text{cal},S} = \sum_{i \in S} \frac{y_i}{\pi_i}.$$

Par conséquent, $\hat{Y}_{\hat{p}}^{\text{cal},S}$ *est asymptotiquement équivalent à l'estimateur de HT de l'échantillon complet dans ce cas. Lors du calage au niveau de l'échantillon selon l'équation (3.2), nous présumons un modèle de superpopulation, c'est-à-dire que* y_i *est une combinaison linéaire de* $\mathbf{x}_i$.

4.4 Calage au niveau de la population

Résultat 2. *Présumons que la séquence de plans d'échantillonnage satisfait les hypothèses (D1) à (D4), que le mécanisme de réponse satisfait les hypothèses (R2) et (R3) et que la séquence de populations finies satisfait les conditions de régularité énoncées dans l'annexe de la version plus longue du présent article (Hasler, 2023). L'EAPNR* $\hat{Y}_{\hat{p}}^{\text{cal},U}$ *satisfait*

$$\frac{1}{N}\hat{Y}_{\hat{p}}^{\text{cal},U} = \frac{1}{N}\hat{Y}_{\hat{p},l}^{\text{cal},U} + O_p(n^{-1}),$$

où

$$\begin{aligned}
\hat{Y}_{\hat{p},l}^{\text{cal},U} &= \sum_{i \in U}\left\{\mathbf{x}_i^\top \boldsymbol{\gamma}_U + \frac{a_i}{\pi_i}\frac{r_i}{p_i}\left(y_i - \mathbf{x}_i^\top \boldsymbol{\gamma}_U\right)\right\}, \\
\boldsymbol{\gamma}_U &= \left\{\sum_{i \in U}(1-p_i)\,\mathbf{x}_i \mathbf{x}_i^\top\right\}^{-1} \sum_{i \in U}(1-p_i)\,\mathbf{x}_i y_i.
\end{aligned}$$

La preuve est fournie dans l'annexe de la version plus longue du présent article (Hasler, 2023).

Remarque 5. *L'EAPNR* $\hat{Y}_{\hat{p}}^{\text{cal},U}$ *se comporte asymptotiquement comme l'estimateur linéarisé* $\hat{Y}_{\hat{p},l}^{\text{cal},U}$, *qui est sans biais pour le total sur la population* Y.

Remarque 6. *S'il existe un vecteur* $\boldsymbol{\beta}$ *tel que* $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ *pour tout* $i \in U$, *alors*

$$\hat{Y}_{\hat{p},l}^{\text{cal},U} = \sum_{i \in U} y_i.$$

Par conséquent, $\hat{Y}_{\hat{p}}^{\text{cal},U}$ *est asymptotiquement équivalent au total sur la population dans ce cas. Lors du calage au niveau de la population selon l'équation (3.3), nous présumons un modèle de superpopulation, c'est-à-dire que* y_i *est une combinaison linéaire de* $\mathbf{x}_i$.

5. Analyse asymptotique II : Double robustesse

Les résultats de la section 4 reposent sur l'hypothèse (R3). Cela signifie que ces résultats sont valides si le modèle de réponse est correctement spécifié. Dans la présente section, nous démontrons que les EAPNR obtenus par calage peuvent tout de même être consistants lorsque le modèle de réponse n'est pas spécifié correctement, pour autant qu'un modèle de superpopulation, à savoir un modèle qui lie la variable d'intérêt aux variables auxiliaires, est correctement spécifié. Nous disons, dans ce cas, que les EAPNR résultants sont doublement robustes parce que la consistance est maintenue même lorsqu'un des deux modèles, soit le modèle de réponse ou le modèle de superpopulation, est mal spécifié. Les résultats 3 et 4 plus bas énoncent cette propriété. Pour le premier résultat, deux hypothèses requises concernant le mécanisme de réponse et les probabilités de réponse estimées sont :

(R4) : Les données sont manquantes au hasard.

(R5) : Les probabilités de réponse estimées admettent une limite inférieure, c'est-à-dire qu'il existe une constante $c_1 > 0$ telle que $\hat{p}_i > c_1$ pour tout $i \in S$ et tout N.

Résultat 3. *Considérons le modèle de superpopulation* $\xi : y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$ *où* $\mathrm{E}_\xi(\varepsilon_i) = 0,\ \mathrm{E}_\xi(\varepsilon_i \varepsilon_j) = \sigma^2 \leq +\infty$ *si* $i = j$ *et 0 sinon, et l'indice* ξ *signifie que l'espérance et la variance sont calculées par rapport au modèle* ξ*. Supposons que les hypothèses (D1) à (D4), (R2), (R4) et (R5) sont vérifiées. Alors*

$$\frac{\hat{Y}_{\hat{p}}^{\mathrm{cal},U} - Y}{N} = o_{\mathbb{P}}(1),$$
$$\frac{\hat{Y}_{\hat{p}}^{\mathrm{cal},S} - Y}{N} = o_{\mathbb{P}}(1).$$

L'indice $\mathbb{P}$ *signifie que la distribution de la probabilité de référence est celle déterminée par le modèle de superpopulation, le plan d'échantillonnage et le mécanisme de réponse.*

La preuve est fournie dans l'annexe de la version plus longue du présent article (Hasler, 2023). Ce résultat indique que lorsque les probabilités de réponse sont obtenues par calage, les EAPNR résultants sont des estimateurs consistants du total sur la population. Le résultat 3 se vérifie même lorsque le modèle de réponse de l'hypothèse (R3) est mal spécifié.

Résultat 4. *Présumons que la séquence de plans d'échantillonnage satisfait les hypothèses (D1) à (D4), que le mécanisme de réponse satisfait les hypothèses (R1) à (R3) et que la séquence de populations finies satisfait les hypothèses (P1) à (P6) dans l'annexe de l'article de Hasler (2023). Alors*

$$\frac{\hat{Y}_{\hat{p}}^{\mathrm{cal},U} - Y}{N} = o_p(1),$$
$$\frac{\hat{Y}_{\hat{p}}^{\mathrm{cal},S} - Y}{N} = o_p(1).$$

La preuve est fournie dans l'annexe de la version plus longue du présent article (Hasler, 2023). Ce résultat indique que lorsque les probabilités de réponse sont obtenues par calage, les EAPNR résultants sont des estimateurs consistants du total sur la population si le modèle de réponse est spécifié correctement. Le résultat 4 se vérifie même lorsque le modèle de réponse énoncé au résultat 3 est mal spécifié. Il convient de souligner que la distribution de la probabilité du résultat 4 est celle déterminée par le plan d'échantillonnage et le mécanisme de réponse. Les deux quantités du résultat 4 sont donc aussi $o_{\mathbb{P}}(1)$.

À partir des résultats 3 et 4, nous concluons que les EAPNR obtenus par calage sont doublement robustes. En d'autres termes, ces estimateurs restent consistants même si un des deux modèles, soit le modèle de superpopulation ou le modèle de réponse, est mal spéficié. Toutefois, lorsque les probabilités de réponse sont estimées par maximum de vraisemblance, la consistance de l'EAPNR résultant n'est pas garantie sous les hypothèses énoncées dans les résultats. En effet, lorsque les probabilités de réponse sont obtenues par maximum de vraisemblance à partir de l'équation (3.1), les poids résultants peuvent ne pas être calés. Le fait que les poids sont calés joue un rôle central dans la preuve du résultat 3. Par conséquent, si la double robustesse de l'EAPNR obtenu par maximum de vraisemblance se vérifie, il est nécessaire de faire d'autres hypothèses. Ce sujet dépasse la portée du présent article.

6. Variance et estimation de la variance

Tout au long de la présente section, nous présumons que l'hypothèse (R1) se vérifie. En cas de non-réponse, nous pouvons formuler la variance d'un estimateur générique $\hat{Y}_g$ comme suit :

$$\mathrm{V}(\hat{Y}_g) = \mathrm{V}_{\mathrm{ech}}(\hat{Y}_g) + \mathrm{V}_{\mathrm{nr}}(\hat{Y}_g),$$

où les deux termes sont la *variance d'échantillonnage* et la *variance de non-réponse*, respectivement, et sont donnés par

$$\begin{aligned} \mathrm{V}_{\mathrm{ech}}(\hat{Y}_g) &= \mathrm{V}_p\left\{\mathrm{E}_q(\hat{Y}_g \mid S)\right\}, \\ \mathrm{V}_{\mathrm{nr}}(\hat{Y}_g) &= \mathrm{E}_p\left\{\mathrm{V}_q(\hat{Y}_g \mid S)\right\}. \end{aligned}$$

Selon cette décomposition, la variance de l'estimateur s'appuyant sur les probabilités de réponse réelles peut être formulée comme suit :

$$\mathrm{V}(\hat{Y}_p) = \mathrm{V}_p\left(\sum_{i\in S}\frac{y_i}{\pi_i}\right) + \mathrm{E}_p\left(\sum_{i\in S}\frac{1}{\pi_i^2}\frac{1-p_i}{p_i}\,y_i^2\right).$$

Selon la même décomposition, Kim et Kim (2007), page 507, définissent la variance de l'estimateur linéarisé $\hat{Y}_{\hat{p},l}^{\mathrm{EMV}}$ comme suit :

$$\mathrm{V}\left(\hat{Y}_{\hat{p},l}^{\mathrm{EMV}}\right) = \mathrm{V}_{\mathrm{ech}}(\hat{Y}_{\hat{p},l}^{\mathrm{EMV}}) + \mathrm{V}_{\mathrm{nr}}(\hat{Y}_{\hat{p},l}^{\mathrm{EMV}}),$$

où

$$\begin{aligned} \mathrm{V}_{\mathrm{ech}}(\hat{Y}_{\hat{p},l}^{\mathrm{EMV}}) &= \mathrm{V}_p\left(\sum_{i\in S}\frac{y_i}{\pi_i}\right), \\ \mathrm{V}_{\mathrm{nr}}(\hat{Y}_{\hat{p},l}^{\mathrm{EMV}}) &= \mathrm{E}_p\left\{\sum_{i\in S}\frac{1}{\pi_i^2}\frac{1-p_i}{p_i}(y_i - k_i\pi_i p_i \mathbf{x}_i^\top \boldsymbol{\gamma}_n^{\mathrm{EMV}})^2\right\}. \end{aligned}$$

Le premier terme est la variance de l'estimateur de HT de l'échantillon complet. Le deuxième terme disparaît s'il existe un vecteur $\boldsymbol{\beta}$ tel que $y_i = k_i\pi_i p_i \mathbf{x}_i^\top \boldsymbol{\beta}$. Cet énoncé respecte la remarque 2 de la section 4, à savoir que $\hat{Y}_{\hat{p},l}^{\mathrm{EMV}}$ correspond à l'estimateur de HT de l'échantillon complet lorsque cette relation est vérifiée.

Une décomposition similaire est obtenue lorsque le calage est appliqué. Pour obtenir plus de précisions, veuillez consulter Hasler (2023). La variance de l'estimateur linéarisé $\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}$ peut être formulée comme suit :

$$\mathrm{V}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}) = \mathrm{V}_{\mathrm{ech}}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}) + \mathrm{V}_{\mathrm{nr}}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}),$$

où

$$\begin{aligned} \mathrm{V}_{\mathrm{ech}}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}) &= \mathrm{V}_p\left(\sum_{i\in S}\frac{y_i}{\pi_i}\right), \\ \mathrm{V}_{\mathrm{nr}}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}) &= \mathrm{E}_p\left\{\sum_{i\in S}\frac{1}{\pi_i^2}\frac{1-p_i}{p_i}(y_i - \mathbf{x}_i^\top \boldsymbol{\gamma}_S)^2\right\}. \end{aligned}$$

Le premier terme est la variance de l'estimateur de HT de l'échantillon complet. Le deuxième terme disparaît s'il existe un vecteur $\boldsymbol{\beta}$ tel que $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}$. Cet énoncé respecte la remarque 4, à savoir que $\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}$ correspond à l'estimateur de HT de l'échantillon complet lorsque cette relation est vérifiée.

De même, la variance de l'estimateur linéarisé $\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}$ peut être formulée comme suit :

$$\mathrm{V}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}) = \mathrm{V}_{\mathrm{ech}}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}) + \mathrm{V}_{\mathrm{nr}}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}),$$

où

$$\begin{aligned} \mathrm{V}_{\mathrm{ech}}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}) &= \mathrm{V}_p\left\{\sum_{i\in S}\frac{1}{\pi_i}(y_i - \mathbf{x}_i^\top \boldsymbol{\gamma}_U)\right\}, \\ \mathrm{V}_{\mathrm{nr}}(\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}) &= \mathrm{E}_p\left\{\sum_{i\in S}\frac{1}{\pi_i^2}\frac{1-p_i}{p_i}(y_i - \mathbf{x}_i^\top \boldsymbol{\gamma}_U)^2\right\}. \end{aligned}$$

Le premier terme est la variance de l'estimateur de HT de l'échantillon complet des différences $y_i - \mathbf{x}_i^\top \boldsymbol{\gamma}_U$. Les premier et deuxième termes disparaissent s'il existe un vecteur $\boldsymbol{\beta}$ tel que $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}$. Cet énoncé respecte la remarque 6, à savoir que $\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}$ est égal au total sur la population, qui a une variance nulle, lorsque cette relation est vérifiée.

Remarque 7. *La variance d'échantillonnage des estimateurs linéarisés* $\hat{Y}_{\hat{p},l}^{\mathrm{EMV}}$ *et* $\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}$ *est égale à la variance d'échantillonnage de* $\hat{Y}_p$*. Leur variance de non-réponse n'est pas supérieure à celle de* $\hat{Y}_p$*. Les EAPNR* $\hat{Y}_{\hat{p}}^{\mathrm{EMV}}$ *et* $\hat{Y}_{\hat{p}}^{\mathrm{cal},S}$ *sont donc asymptotiquement équivalents à des estimateurs qui sont au moins autant efficaces que l'estimateur s'appuyant sur les probabilités de réponse réelles. Cet énoncé a été démontré dans Kim et Kim (2007) pour* $\hat{Y}_{\hat{p}}^{\mathrm{EMV}}$*; voir la page 505.*

De plus, on s'attend à ce que la variance d'échantillonnage de l'estimateur linéarisé $\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}$ *soit plus petite que la variance échantillonnale de* $\hat{Y}_p$ *pour autant que les résidus* $y_i - \mathbf{x}_i^\top \boldsymbol{\gamma}_U$ *présentent moins de variation que les valeurs* y_i*. La variance de non-réponse de* $\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}$ *n'est pas supérieure à celle de* $\hat{Y}_p$*. Par conséquent,* $\hat{Y}_{\hat{p}}^{\mathrm{cal},U}$ *est asymptotiquement équivalent à un estimateur qui est au moins aussi efficace que l'estimateur s'appuyant sur les probabilités de réponse réelles selon les conditions énoncées ci-dessus.*

En général, il semble qu'il y ait un gain d'efficacité lorsque l'on utilise les probabilités de réponse estimées par rapport aux probabilités de réponse réelles, au moins pour les populations et les échantillons suffisamment grands. Une explication possible est que l'estimation des probabilités de réponse peut être vue comme une façon de lisser les poids en utilisant un modèle approprié. Il a déjà été démontré qu'un tel lissage améliore l'efficacité de l'estimateur de HT habituel; voir Beaumont (2008), par exemple.

Remarque 8. *Faisons maintenant la comparaison de la variance des EAPNR par calage, soit* $\hat{Y}_{\hat{p}}^{\mathrm{cal},U}$ *et* $\hat{Y}_{\hat{p}}^{\mathrm{cal},S}$*. On s'attend à ce que la variance d'échantillonnage de l'estimateur linéarisé* $\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}$ *soit plus petite que la variance d'échantillonnage de l'estimateur linéarisé* $\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}$*, pour autant que les résidus* $y_i - \mathbf{x}_i^\top \boldsymbol{\gamma}_U$ *présentent moins de variation que les valeurs* y_i*. De plus, on s'attend à ce que la variance de non-réponse de* $\hat{Y}_{\hat{p},l}^{\mathrm{cal},U}$ *soit proche de celle de* $\hat{Y}_{\hat{p},l}^{\mathrm{cal},S}$*, puisque la seule différence tient au fait que le coefficient de population* $\boldsymbol{\gamma}_U$ *de la variance de non-réponse du premier est replacé par un estimateur sur l'échantillon* $\boldsymbol{\gamma}_S$ *dans le second. En pratique, cela signifie que nous nous attendons à un gain d'efficacité de l'EAPNR lors de l'estimation des probabilités de réponse par calage au niveau de la population par rapport à une telle estimation au niveau de l'échantillon.*

Nous présumons, dorénavant et jusqu'à la fin de la présente section, que les hypothèses (D1) à (D4) et (R1) à (R3), et les conditions de régularité supplémentaires énoncées dans l'annexe de la version plus longue du présent article (Hasler, 2023) sont vérifiées. Selon la décomposition de la variance présentée plus haut, l'estimateur suivant peut être utilisé pour la variance de l'EAPNR $\hat{Y}_{\hat{p}}^{\mathrm{EMV}}$ (voir Kim et Kim (2007), page 507)

$$\hat{\mathrm{V}}(\hat{Y}_{\hat{p}}^{\mathrm{EMV}}) = \hat{\mathrm{V}}_{\mathrm{ech}}(\hat{Y}_{\hat{p},l}^{\mathrm{EMV}}) + \hat{\mathrm{V}}_{\mathrm{nr}}(\hat{Y}_{\hat{p},l}^{\mathrm{EMV}}),$$

où

$$\begin{aligned}
\hat{\mathrm{V}}_{\mathrm{ech}}(\hat{Y}_{\hat{p},l}^{\mathrm{EMV}}) &= \sum_{i\in S_r} \frac{1-\pi_i}{\pi_i^2}\frac{y_i^2}{\hat{p}_i} + \sum_{i,j\in S_r, i\neq j} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j\pi_{ij}}\frac{y_i}{\hat{p}_i}\frac{y_j}{\hat{p}_j},\\
\hat{\mathrm{V}}_{\mathrm{nr}}(\hat{Y}_{\hat{p},l}^{\mathrm{EMV}}) &= \sum_{i\in S_r} \frac{1}{\pi_i^2}\frac{1-\hat{p}_i}{\hat{p}_i^2}(y_i - k_i\pi_i\hat{p}_i\mathbf{x}_i^\top\hat{\boldsymbol{\gamma}}_n^{\mathrm{EMV}})^2,\\
\hat{\boldsymbol{\gamma}}_n^{\mathrm{EMV}} &= \left\{\sum_{i\in S_r} k_i(1-\hat{p}_i)\,\mathbf{x}_i\mathbf{x}_i^\top\right\}^{-1} \sum_{i\in S_r}\frac{1}{\pi_i}\frac{1-\hat{p}_i}{\hat{p}_i}\,\mathbf{x}_i y_i.
\end{aligned}$$

Nous employons la même approche pour dériver un estimateur de variance des EAPNR $\hat{Y}_{\hat{p}}^{\text{cal},S}$ et $\hat{Y}_{\hat{p}}^{\text{cal},U}$. Nous obtenons

$$\hat{\text{V}}\,(\hat{Y}_{\hat{p}}^{\text{cal},S}) = \hat{\text{V}}_{\text{ech}}\,(\hat{Y}_{\hat{p},l}^{\text{cal},S}) + \hat{\text{V}}_{\text{nr}}\,(\hat{Y}_{\hat{p},l}^{\text{cal},S}),$$

où

$$\begin{aligned}
\hat{\text{V}}_{\text{ech}}\,(\hat{Y}_{\hat{p},l}^{\text{cal},S}) &= \sum_{i\in S_r}\frac{1-\pi_i}{\pi_i^2}\frac{y_i^2}{\hat{p}_i} + \sum_{i,\,j\in S_r;\,i\neq j}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\frac{y_i}{\hat{p}_i}\frac{y_j}{\hat{p}_j},\\
\hat{\text{V}}_{\text{nr}}\,(\hat{Y}_{\hat{p},l}^{\text{cal},S}) &= \sum_{i\in S_r}\frac{1}{\pi_i^2}\frac{1-\hat{p}_i}{\hat{p}_i^2}\left(y_i-\mathbf{x}_i^{\top}\hat{\boldsymbol{\gamma}}_n^{\text{cal}}\right)^2\\
\hat{\boldsymbol{\gamma}}_n^{\text{cal}} &= \left(\sum_{i\in S_r}\frac{1}{\pi_i}\frac{1-\hat{p}_i}{\hat{p}_i}\mathbf{x}_i\mathbf{x}_i^{\top}\right)^{-1}\sum_{i\in S_r}\frac{1}{\pi_i}\frac{1-\hat{p}_i}{\hat{p}_i}\mathbf{x}_i y_i.
\end{aligned}$$

De même, nous obtenons

$$\hat{\text{V}}\,(\hat{Y}_{\hat{p}}^{\text{cal},U}) = \hat{\text{V}}_{\text{ech}}\,(\hat{Y}_{\hat{p},l}^{\text{cal},U}) + \hat{\text{V}}_{\text{nr}}\,(\hat{Y}_{\hat{p},l}^{\text{cal},U}),$$

où

$$\begin{aligned}
\hat{\text{V}}_{\text{ech}}\,(\hat{Y}_{\hat{p},l}^{\text{cal},U}) &= \sum_{i\in S_r}\frac{1-\pi_i}{\pi_i^2}\frac{e_i^2}{\hat{p}_i} + \sum_{i,\,j\in S_r;\,i\neq j}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\frac{e_i}{\hat{p}_i}\frac{e_j}{\hat{p}_j},\\
e_i &= y_i-\mathbf{x}_i^{\top}\hat{\boldsymbol{\gamma}}_n^{\text{cal}},\\
\hat{\text{V}}_{\text{nr}}\,(\hat{Y}_{\hat{p},l}^{\text{cal},U}) &= \sum_{i\in S_r}\frac{1}{\pi_i^2}\frac{1-\hat{p}_i}{\hat{p}_i^2}e_i^2.
\end{aligned}$$

Pour obtenir plus de précisions, veuillez consulter Hasler (2023).

7. Étude par simulation

7.1 Paramètres de la simulation

Cinq populations différentes sont étudiées et obtenues comme suit. Pour chaque population, nous générons $N = 2\,000$ unités de population. Les variables auxiliaires sont les mêmes pour les cinq populations et sont formulées par $\mathbf{x}_i = (1, x_i)^{\top}$, où x_i sont des observations de variables aléatoires uniformes indépendantes et identiquement distribuées selon des paramètres, soit les limites, de 0 et de 100. Les valeurs des variables d'intérêt sont obtenues comme suit :

$$\begin{aligned}
y_{1i} &= 1\,000 + 20x + \varepsilon_{1i},\\
y_{2i} &= 1\,500 + 500\exp(-10 + 0{,}1x) + \varepsilon_{2i},\\
y_{3i} &= \begin{cases}1 & \text{avec probabilité } \phi_i,\\ 0 & \text{sinon,}\end{cases} \quad \text{où} \quad \phi_i = \begin{cases}0{,}8 & \text{si } x_i > 75,\\ 0{,}2 & \text{sinon,}\end{cases}\\
y_{4i} &= 1\,000 + \varepsilon_{4i},\\
y_{5i} &= 1\,000 + 20x + \varepsilon_{5i},
\end{aligned}$$

où $\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{4i},$ et ε_{5i} sont des observations de variables aléatoires normales indépendantes et identiquement distribuées ayant une moyenne de 0 et un écart-type de 750, 100, 750 et 50, respectivement. Dans la population 1, il y a une relation linéaire entre x et $y_1,$ et une corrélation d'environ 0,6. Dans la population 2, il y a une relation non linéaire entre x et $y_2.$ Dans la population 3, y_3 est catégorique et les valeurs sont obtenues à partir de variables aléatoires de Bernouilli indépendantes ayant un paramètre de 0,8 pour les grandes valeurs de x et un paramètre de 0,2 pour les petites valeurs de $x.$ Dans la population 4, il n'y a pas de relation entre x et $y_4.$ Dans la population 5, il y a une très forte relation linéaire entre x et $y_5,$ et une corrélation d'environ 0,99.

Deux vecteurs de probabilités de réponse sont créés comme suit :

$$
\begin{aligned}
p_{1i} &= \frac{1}{1+\exp(-\mathbf{x}_i^{\top}\boldsymbol{\lambda})}, \\
p_{2i} &= \begin{cases} 1-a_1(x_i-k_1)^2+h_1 & \text{si } a_1(x_i-k_1)^2+h_1 > 0{,}01, \\ 0{,}9 & \text{sinon,} \end{cases}
\end{aligned}
$$

où $a_1 = -0{,}0005,$ $k_1 = 25{,}79116,$ $h_1 = 0{,}9$ et $\boldsymbol{\lambda} = (-2; 0{,}04)^{\top}.$ Les deux vecteurs sont construits de sorte à donner un taux de réponse moyen pour la population d'environ 50 %. Il convient de souligner que, selon l'échantillon sélectionné, le taux de réponse moyen pour l'échantillon peut être supérieur ou inférieur à 50 %, puisque les unités ne sont pas nécessairement sélectionnées uniformément sur toutes les valeurs de $x.$ Pour le premier vecteur de probabilités de réponse, le modèle de régression logistique dans l'équation (2,4) est correctement spécifié. Pour le deuxième vecteur de probabilités de réponse, ce modèle est mal spécifié. Pour les deux vecteurs, les grandes valeurs de x tendent à avoir de grandes probabilités de réponse. La figure 7.1 illustre les cinq populations et la figure 7.2 présente les probabilités de réponse comme une fonction des valeurs de $x.$

Deux plans d'échantillonnage sont étudiés : 1) l'échantillonnage aléatoire simple sans remise où $n =$ 200 unités sont sélectionnées; 2) l'échantillonnage stratifié où deux strates sont étudiées. La première strate contient les unités ayant une valeur x inférieure à la valeur médiane de x et la deuxième strate contient les unités ayant une valeur x supérieure à la médiane. Dans la première strate, 40 unités sont sélectionnées en utilisant l'échantillonnage aléatoire simple. La fraction de sondage dans la première strate est de 4 %. Dans la deuxième strate, 160 unités sont sélectionnées en utilisant l'échantillonnage aléatoire simple. La fraction de sondage dans la deuxième strate est de 16 %.

Au total, 10 000 simulations sont exécutées, comme on l'explique ci-après, et ce pour chaque population, chaque plan d'échantillonnage et chaque vecteur de probabilités de réponse. Nous obtenons 20 scénarios. Un échantillon de taille $n =$ 200 est sélectionné conformément au plan d'échantillonnage. Un ensemble de répondants est produit au moyen du plan d'échantillonnage de Poisson selon le vecteur de probabilités de réponse. La fonction `optim` est utilisée pour résoudre les équations estimantes afin d'obtenir les paramètres du modèle de réponse présenté à la section 3. La fonction minimise le maximum de la valeur relative absolue du côté gauche des équations estimantes (3.1), (3.2) et (3.3) pour les variables auxiliaires. Nous déterminons

que l'algorithme converge si ce maximum est inférieur à 0,01. La valeur initiale du vecteur de paramètre est fixée à $(0, 0)$ de sorte que les probabilités de réponse initiales sont toutes de 1/2. Lorsque l'on compare la performance des EAPNR et leurs estimateurs de variance, seules les simulations pour lesquelles l'algorithme converge sont utilisées pour calculer les mesures de comparaison d'un estimateur donné. Le total Y est estimé au moyen des sept estimateurs énumérés ci-après.

1. $\hat{Y}$ (HT) : l'estimateur de Horvitz-Thompson. Il convient de souligner que cet estimateur n'est pas disponible, en pratique, dans le cas de non-réponse. Il sert de point de comparaison dans la présente étude par simulations
2. $\hat{Y}_p(p)$: l'estimateur s'appuyant sur les probabilités de réponse réelles dans l'équation (2.2). L'estimateur n'est pas disponible en pratique. Il sert de point de comparaison dans la présente étude par simulations.
3. $\hat{Y}_{\text{naïf}}$ (naïf) : l'estimateur qui ignore la non-réponse, soit $\hat{Y}_{\text{naïf}} = \frac{n}{n_r} \sum_{i \in S_r} \frac{y_i}{\pi_i}$.
4. $\hat{Y}_{\hat{p}}^{\text{EMV},1}$ (EMV, 1) : L'EAPNR s'appuyant sur les probabilités de réponse estimées par maximum de vraisemblance, équation (3.1), où $k_i = 1$.
5. $\hat{Y}_{\hat{p}}^{\text{EMV},1/\pi}$ (EMV, $1/\pi$) : L'EAPNR s'appuyant sur les probabilités de réponse estimées par maximum de vraisemblance, équation (3.1), où $k_i = 1/\pi_i$.
6. $\hat{Y}_{\hat{p}}^{\text{cal},U}$ (cal, U) : L'EAPNR s'appuyant sur les probabilités de réponse estimées par calage au niveau de la population, équation (3.3).
7. $\hat{Y}_{\hat{p}}^{\text{cal},S}$ (cal, S) : L'EAPNR s'appuyant sur les probabilités de réponse estimées par calage au niveau de l'échantillon, équation (3.2).

Figure 7.1 Cinq populations.

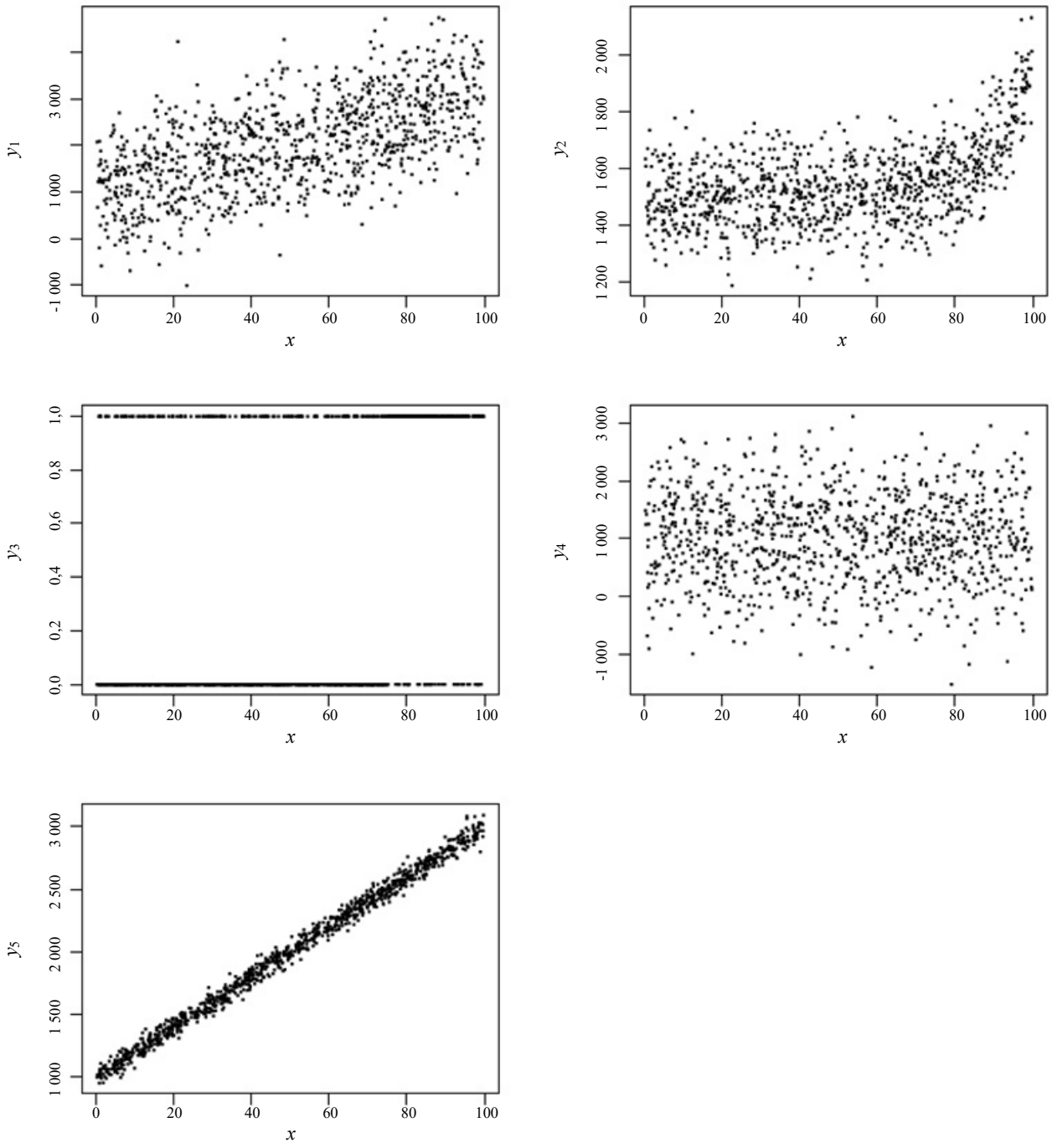
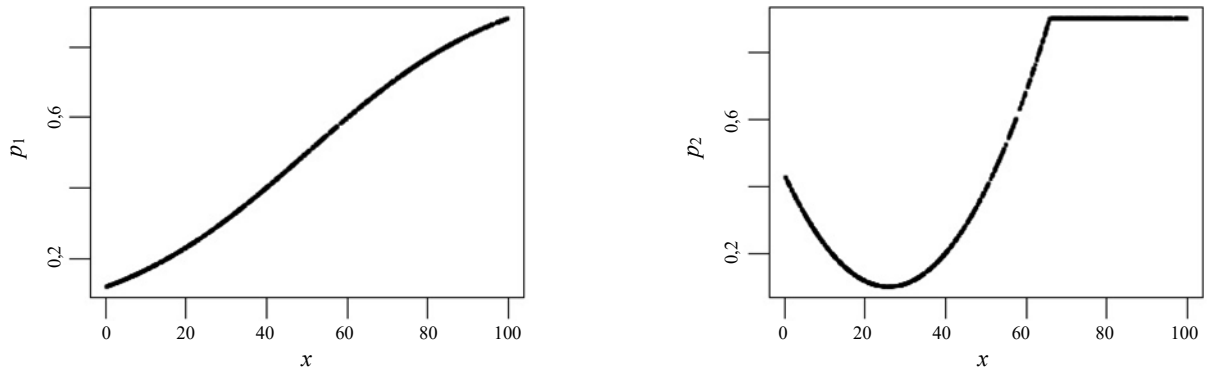


Figure 7.2 Deux vecteurs de probabilités de réponse.



7.2 Performance des estimateurs par pondération ajustée pour la non-réponse

La performance de ces estimateurs est évaluée au moyen des mesures de comparaison suivantes définies pour un estimateur générique $\hat{Y}_g$:

- Biais relatif absolu de Monte-Carlo $(|\text{BR}|)$ défini par

$$|\text{BR}| = \left|\frac{B}{Y}\right|,$$

où $\text{B} = \hat{Y}_g^{(\cdot)} - Y$, $\hat{Y}_g^{(\cdot)}$ est la moyenne de l'estimateur sur les L simulations (ou les L simulations pour lesquelles l'algorithme d'optimisation converge si $\hat{Y}_g$ est un EAPNR),

$$\hat{Y}_g^{(\cdot)} = \frac{1}{L}\sum_{\ell=1}^{L} \hat{Y}_g^{(\ell)},$$

et $\hat{Y}_g^{(\ell)}$ est l'estimateur $\hat{Y}_g$ obtenu à la ℓ^{e} simulation,

- Écart-type relatif (ETR) de Monte-Carlo défini par

$$\text{ETR} = \frac{(\text{VAR})^{1/2}}{Y},$$

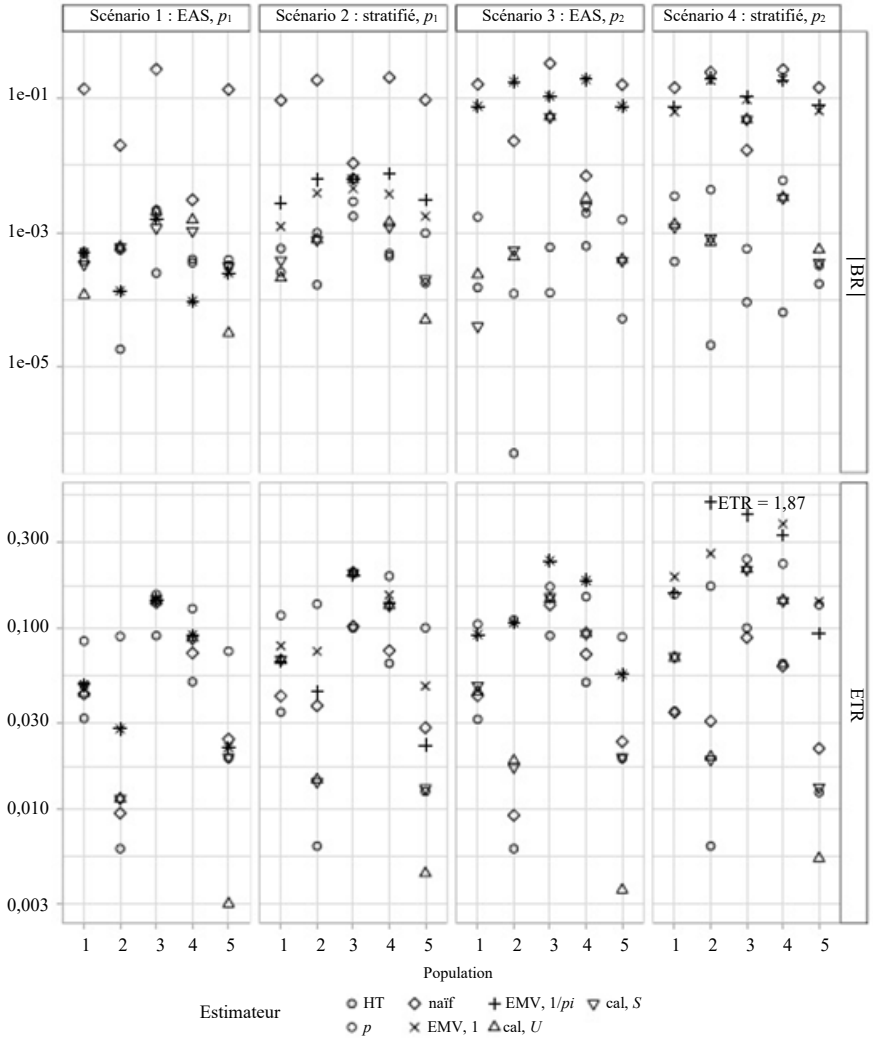
où

$$\text{VAR} = \frac{1}{L-1}\sum_{\ell=1}^{L} (\hat{Y}_g^{(\ell)} - \hat{Y}_g^{(\cdot)})^2.$$

Les résultats sont présentés à la figure 7.3. Les axes des y sont présentés en échelles logarithmiques. Pour les graphiques d'ETR, la valeur maximale sur l'axe des y est fixée à 0,5 à des fins de clarté. Un estimateur a une valeur supérieure à 0,5 dans le scénario 4, pour la population 2. Cette valeur est indiquée sur le diagramme. Dans les scénarios 1 et 2, lorsque le modèle des probabilités de réponse est spécifié correctement, les quatre EAPNR montrent un BR du même ordre que le BR de l'estimateur de HT et de l'estimateur s'appuyant sur les probabilités de réponse réelles $\hat{Y}_p$. Ces deux derniers estimateurs étant sans biais, ce résultat illustre la façon dont les quatre EAPNR sont presque sans biais; voir les remarques 1, 3 et

5. Dans les scénarios 3 et 4, lorsque le modèle des probabilités de réponse est mal spécifié, les deux EAPNR s'appuyant sur les probabilités de réponse estimées par calage montrent un BR du même ordre que le BR de l'estimateur de HT et de l'estimateur s'appuyant sur les probabilités de réponse réelles $\hat{Y}_p$. Les deux estimateurs s'appuyant sur les probabilités de réponse estimées par maximum de vraisemblance montrent un BR plus grand. Ce constat illustre l'incidence possible du calage comme protection plus solide contre la spécification incorrecte du modèle des probabilités de réponse, par rapport à l'EMV. Dans les quatre scénarios, l'estimateur naïf donne le plus grand BR.

Figure 7.3 |BR| et ETR pour les sept estimateurs, les cinq populations et les quatre scénarios.



Dans les scénarios 1 et 2, lorsque le modèle des probabilités de réponse est spécifié correctement, les quatre EAPNR montrent une variance plus petite que la variance de l'estimateur s'appuyant sur les probabilités de réponse réelles $\hat{Y}_p$. Ce constat confirme qu'un gain d'efficacité de l'estimateur du total est obtenu lorsque l'on estime les probabilités de réponse par maximum de vraisemblance ou par calage, par rapport à l'utilisation des probabilités de réponse réelles; voir la remarque 7. Dans ces deux scénarios, les

quatre EAPNR montrent un ETR du même ordre. Dans les scénarios 3 et 4, lorsque le modèle des probabilités de réponse est mal spécifié, les deux EAPNR s'appuyant sur les probabilités de réponse estimées par calage montrent un ETR plus petit que l'ETR des deux EAPNR s'appuyant sur les probabilités de réponse estimées par maximum de vraisemblance. Ce constat illustre l'incidence possible du calage comme protection plus solide contre la spécification incorrecte du modèle des probabilités de réponse, par rapport à l'EMV.

7.3 Performance des estimateurs de variance

La variance des quatre EAPNR est estimée pour chaque simulation en utilisant les formules de la section 6. La performancedes estimateurs de variance est évaluée au moyen des mesures de comparaison suivantes définies pour un estimateur générique $\hat{Y}_g$:

- Biais relatif absolu de Monte-Carlo $(|\mathrm{BR}|)$ défini par

$$|\mathrm{BR}| = \frac{|B|}{\mathrm{V}_{\mathrm{sim}}(\hat{Y}_g)},$$

où $\mathrm{V}_{\mathrm{sim}}(\hat{Y}_g)$ est la variance de $\hat{Y}_g$ sur les L simulations pour lesquelles l'algorithme d'optimisation converge, $\mathrm{B} = \hat{\mathrm{V}}(\hat{Y}_g^{(\cdot)}) - \mathrm{V}_{\mathrm{sim}}(\hat{Y}_g)$, et $\hat{\mathrm{V}}(\hat{Y}_g^{(\cdot)})$ est la moyenne de $\hat{\mathrm{V}}(\hat{Y}_g)$ pour ces L simulations,

- TC : le taux de couverture réelle de l'intervalle de confiance de 95 %, soit la proportion de simulations pour lesquelles l'intervalle de confiance de 95 % contient le total sur la population Y.

Les résultats sont présentés à la figure 7.4. Les axes des y sont présentés en échelles logarithmiques. Pour faciliter la lecture des diagrammes, quatre BR supérieurs à 1 ont été ramenés à 1 et cinq taux de couverture inférieurs à 0,5 ont été ramenés à 0,5. Dans les scénarios 1 et 2, lorsque le modèle des probabilités de réponse est spécifié correctement, le BR de l'estimateur de variance s'appuyant sur les probabilités de réponse estimées par maximum de vraisemblance tend à être inférieur au BR de l'estimateur de variance s'appuyant sur les probabilités de réponse estimées par calage. Dans les scénarios 3 et 4, lorsque le modèle des probabilités de réponse est mal spécifié, l'inverse se produit. Dans les scénarios 1 et 2, les quatre estimateurs de variance produisent un taux de couverture généralement proche de la couverture nominale de 95 %. Dans les scénarios 3 et 4, l'estimateur de variance s'appuyant sur les probabilités de réponse estimées par maximum de vraisemblance produit un taux de couverture très bas dans plusieurs cas.

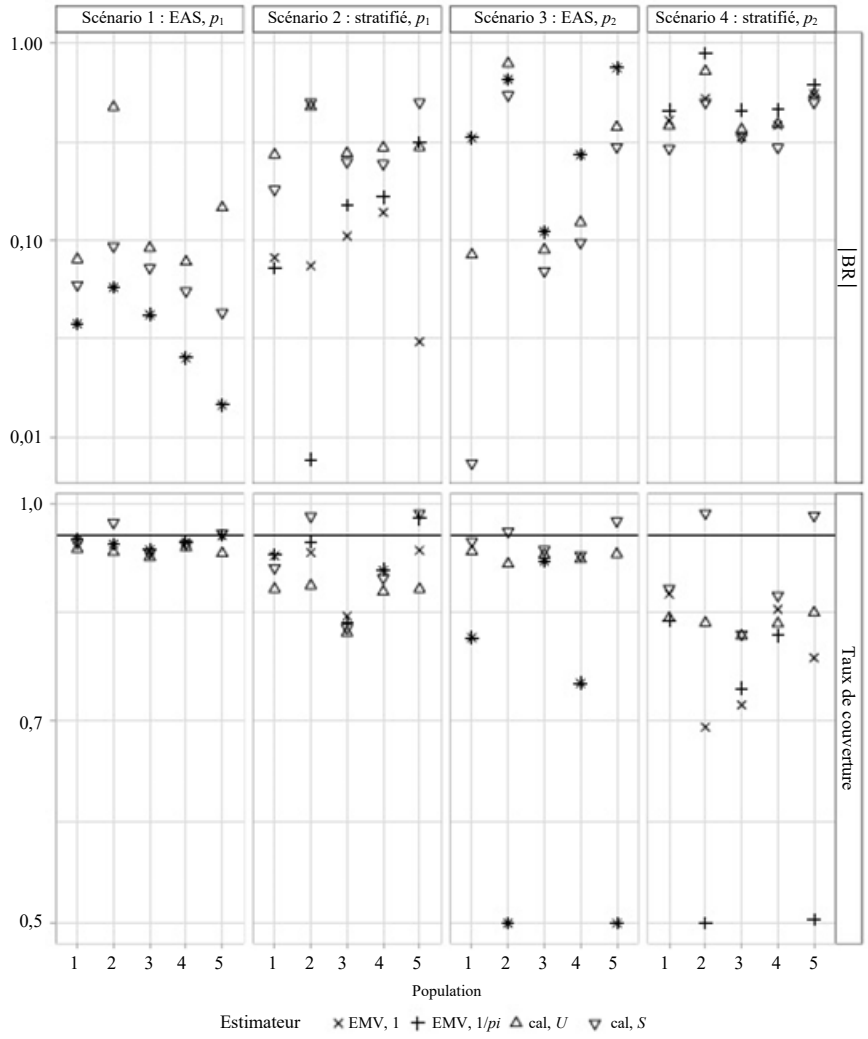
7.4 Poids et convergence

Dans certains cas, les équations estimantes utilisées pour obtenir les probabilités de réponse estimées peuvent ne pas admettre de solution. Dans certains autres cas, une solution aux équations estimantes existe, mais les poids résultants, soit l'inverse des probabilités de réponse estimées, peuvent être très grands. La section 6 de l'article d'Hasler (2023) fournit des précisions et des explications. Afin d'illustrer ces

problèmes de convergence et de poids extrêmes, les trois mesures de comparaison suivantes sont calculées pour chaque EAPNR :

- Poids maximum : le poids ajusté final $1/(\pi_i \hat{p}_i)$ le plus grand sur les 10 000 simulations,
- Erreur relative moyenne : la moyenne, sur 10 000 simulations, du maximum de l'erreur relative absolue de l'équation estimante,
- Taux de calage : la proportion des simulations pour lesquelles l'erreur relative moyenne est plus petite que le seuil de 0,01. Nous déterminons que l'algorithme converge vers une solution lorsque l'erreur relative moyenne est inférieure à ce seuil.

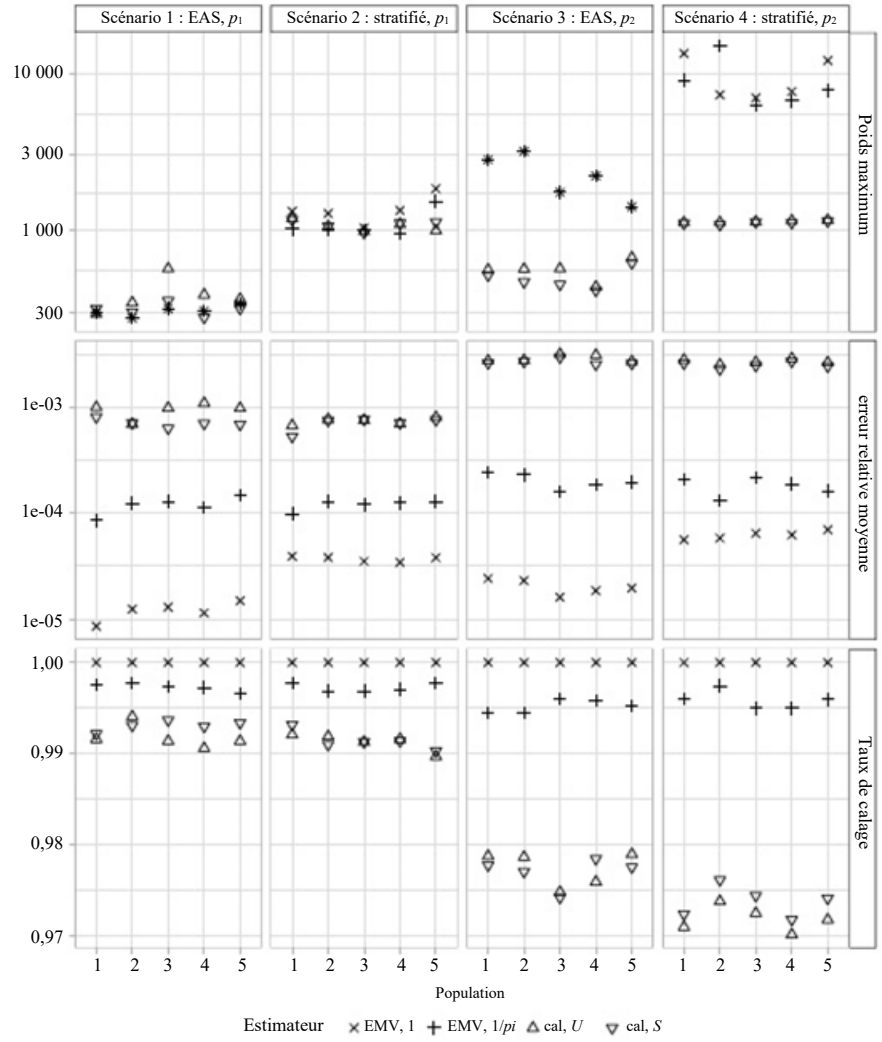
Figure 7.4 |BR| et taux de couverture pour les quatre estimateurs de variance, les cinq populations et les quatre scénarios.



Les résultats sont présentés à la figure 7.5. Les axes des y sont présentés en échelles logarithmiques. Un estimateur donne un poids maximum de plus de 400 000 dans le scénario 4. Pour faciliter la lecture des graphiques, cette valeur est réduite à 15 000. Dans les scénarios 1 et 2, lorsque le modèle des probabilités

de réponse est spécifié correctement, les quatre EAPNR donnent des poids maximaux proches les uns des autres. Aucun poids extrême n'est apparent. Dans les scénarios 3 et 4, lorsque le modèle des probabilités de réponse est mal spécifié, de très grands poids sont obtenus par l'EMV, surtout dans le scénario 4. Le calage peut offrir une protection contre les poids extrêmes lorsque le modèle de réponse est mal spécifié. Dans les quatre scénarios, l'erreur relative moyenne est plus petite avec l'EMV que le calage. Cette différence est plus grande dans les scénarios 3 et 4, lorsque le modèle des probabilités de réponse est mal spécifié. De plus, l'algorithme produit une erreur relative moyenne plus petite que le seuil de 0,01 plus souvent au moyen de l'EMV que du calage. Ce constat illustre la façon dont l'application de l'algorithme permettant d'obtenir les paramètres du modèle de réponse converge plus souvent vers une solution aux équations estimantes du maximum de vraisemblance que vers une solution aux équations estimantes du calage.

Figure 7.5 Poids maximum, erreur relative moyenne et taux de calage des quatre estimateurs ajustés par pondération pour la non-réponse, des cinq populations et des quatre scénarios.



8. Analyse

Nous nous sommes appuyés sur Kim et Kim (2007) et avons élaboré les propriétés asymptotiques de l'EAPNR lorsque le calage est appliqué pour estimer les probabilités de réponse. Pour la première fois, un cadre théorique commun est étudié pour les deux approches à d'EAPNR, à savoir l'EMV et le calage. Ce cadre nous permet de comparer le comportement asymptotique de quatre estimateurs sur les plans du biais et de la variance, selon des hypothèses communes. Nous postulons un modèle de régression logistique pour les probabilités de réponse. Nous considérons deux niveaux de calage : celui au niveau de la population et celui au niveau de l'échantillon complet. Les principaux résultats obtenus sont les suivants : 1) les EAPNR s'appuyant sur les probabilités de réponse estimées par calage sont asymptotiquement équivalents à des estimateurs sans biais; 2) un gain d'efficacité est obtenu lors de l'estimation des probabilités de réponse par calage par rapport à l'estimateur s'appuyant sur les probabilités de réponse réelles; 3) l'EAPNR s'appuyant sur les probabilités de réponse estimées par calage au niveau de la population est généralement plus efficace que l'EAPNR s'appuyant sur les probabilités de réponse estimées par calage au niveau de l'échantillon; 4) le calage pourrait offrir une meilleure protection contre une spécification incorrecte du modèle de réponse que le maximum de vraisemblance lorsqu'il est appliqué pour estimer les probabilités de réponse utilisées dans l'EAPNR; 5) nous expliquons et illustrons les problèmes relatifs à la convergence vers une solution aux équations estimantes et aux poids extrêmes. L'article porte sur l'étude et la comparaison des EAPNR obtenus par l'EMV ou par calage direct (approche en une étape). Certains auteurs suggèrent une approche en deux étapes, soit une première étape consistant à estimer les probabilités de réponse par maximum de vraisemblance pour contourner le problème des poids extrêmes, puis une deuxième étape consistant en un calage pour améliorer davantage l'efficacité de l'EAPNR; voir Haziza et Lesage (2016), et Haziza et Beaumont (2017), page 222. Cette approche dépasse la portée de la présente recherche et mérite de faire l'objet de futurs travaux.

Remerciements

La présente recherche a été soutenue par l'Office fédéral de la statistique. L'auteure remercie le professeur Yves Tillé, les deux examinateurs, le rédacteur associé et le rédacteur en chef pour leurs commentaires constructifs. Les opinions exprimées dans l'article sont uniquement celles de l'auteure et ne témoignent pas nécessairement des opinions de l'organisme ou des personnes mentionnés précédemment.

Bibliographie

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society, Series B*, 67, 445-458.

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, 95(3), 539-553.

Breidt, F.J., et Opsomer, J.D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32, 2, 190-205.

Cassel, C.M., Särndal, C.-E. et Wretman, J.H. (1983). Some uses of statistical models in connexion with the nonresponse problem. Dans *Incomplete Data in Sample Surveys*, (Éds., W.G. Madow et I. Olkin), 3, 143-160. New York: Academic Press.

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. *Actes des Journées de Méthodologie Statistique*, Paris. Insee-Méthodes.

Deville, J.-C., et Dupont, F. (1993). Non-réponse : principes et méthodes. *Actes des Journées de Méthodologie Statistique*, INSEE, Paris, 53-70.

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 procedure in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Dupont, F. (1993). Calage et redressement de la non-réponse totale : validité de la pratique courante de redressement et comparaison des méthodes alternatives pour l'enquête sur la consommation alimentaire de 1989. *Actes des Journées de Méthodologie Statistique*, INSEE, Paris, 9-42.

Ekholm, A., et Laaksonen, S. (1991). Weighting via response modeling in the finish household budget survey. *Journal of Official Statistics*, 3, 325-337.

Folsom, R.E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. *Proceedings of the Social Statistics Section*, American Statistical Association, 197-202.

Hasler, C. (2023). Inference from sampling with response probabilities estimated via calibration. Rapport technique, University of Neuchâtel. Disponible sur ArXiv à DOI: https://doi.org/10.48550/arXiv.2202.03897.

Haziza, D., et Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32(2), 206-226.

Haziza, D., et Lesage, E. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32(1), 129-145.

Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Iannacchione, V.G., Milne, J.G. et Folsom, R.E. (1991). Response probability weight adjustments using logistic regression. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 637-642.

Isaki, C.T., et Fuller, W.A. (1982). Survey design under a regression population model. *Journal of the American Statistical Association*, 77, 89-96.

Kim, J.K., et Kim, J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 35(4), 501-514.

Kim, J.K., et Riddles, M.K. (2012). Théorie concernant les estimateurs ajustés sur le score de propension dans les sondages. *Techniques d'enquête*, 38, 2, 171-180. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2012002/article/11754-fra.pdf.

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 2, 149-160. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006002/article/9547-fra.pdf.

Kott, P.S. (2012). Pourquoi les poids de sondage devraient être intégrés dans la correction de la non-réponse totale fondée sur des groupes de réponse homogènes. *Techniques d'enquête*, 38, 1, 103-107. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2012001/article/11689-fra.pdf.

Kott, P.S., et Liao, D. (2012). Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. *Survey Research Methods*, 6(2), 105-111.

Lesage, E., Haziza, D. et D'Haultfoeuille, X. (2019). A cautionary tale on instrumental calibration for the treatment of nonignorable unit nonresponse in surveys. *Journal of the American Statistical Association*, 114(526), 906-915.

Lundström, S., et Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.

Oh, H.L., et Scheuren, F. (1983). Weighted adjustment for nonresponse. Dans *Incomplete Data in Sample Survey*, (Éds., W.G. Madow, H. Nisselson et I. Olkin), 2, 143-184. New York: Academic Press.

Ranalli, M., Matei, A. et Neri, A. (2023). Generalised calibration with latent variables for the treatment of unit nonresponse in sample surveys. *Statistical Methods and Applications*, 32(1), 169-195.

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

Särndal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons, Inc.

Särndal, C.-E., et Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *Revue Internationale de Statistique*, 55(3), 279-294.

Tillé, Y., et Matei, A. (2021). *sampling: Survey Sampling*. R package version 2.9.

Techniques d'enquête, décembre 2024
Vol. 50, nº 2, pp. 287-312
Statistique Canada, nº 12-001-X au catalogue

Calage assoupli de poids d'enquête

Nicholas T. Longford[1]

Résumé

Les enquêtes sur la population sont désormais rarement analysées indépendamment d'informations auxiliaires prenant souvent la forme de chiffres de population, de totaux et d'autres sommaires. Le calage, ou l'étalonnage, selon lequel les totaux d'échantillon pondérés des variables auxiliaires sont mis en correspondance avec leurs totaux (connus) de population, est largement appliqué. Des méthodes d'ajustement des poids permettant de respecter ces contraintes font intervenir des procédures itératives aux propriétés d'échantillon fini inconnues. Nous élaborons une autre méthode selon laquelle les poids sont calés en minimisant une fonction quadratique, ne nécessitant aucune itération et fournissant une solution unique. La priorité relative de chaque contrainte est représentée par un paramètre d'ajustement. Les propriétés des poids et de l'estimateur par calage, comme fonctions de ces paramètres, sont examinées analytiquement et par simulations. Un lien entre la méthode proposée et le calage ridge est établi.

Mots-clés : Ajustements de poids; échantillonnage; étalonnage; information auxiliaire; priorité; régression ridge.

1. Introduction

Le calage, ou l'étalonnage, est généralement considéré comme un complément indispensable de l'estimation de sommaires de population dans des enquêtes à grande échelle menées dans un environnement au sein duquel d'autres sources de données fournissent de l'information auxiliaire. Une telle information présente le potentiel de rendre l'estimation plus efficace ou qu'elle acquière d'autres propriétés ou d'autres attributs utiles. Le calage joue un rôle important dans la compensation des imperfections du plan de sondage et de sa mise en œuvre, comme des lacunes de la base de sondage et la non-réponse.

Une documentation approfondie porte sur le calage; l'étude de Deville et Särndal (1992) est largement considérée comme un jalon, renforcée par Lundström et Särndal (1999) pour son application dans le contexte des statistiques officielles modernes. Estevao et Särndal (2006) ainsi que Särndal (2007) passent en revue les développements ultérieurs. Des examens plus récents de la documentation ont été réalisés par Kim et Park (2010), Brick (2013), Wu et Lu (2016) ainsi que par Lohr et Raghunathan (2017). La monographie de Tillé (2020) présente un traitement complet du sujet. Devaud et Tillé (2019) proposent une évaluation de l'incidence de Deville et Särndal (1992) sur l'échantillonnage et les statistiques officielles en particulier. Davies (2018) passe en revue un vaste éventail de méthodes de calage. Dans sa terminologie, le calage strict (« hard calibration ») fait référence à l'optimisation, tout en satisfaisant un ensemble de contraintes sans possibilité d'écart. Nous préférons un calage souple (« soft calibration ») selon lequel on recherche un compromis parmi les contraintes et les objectifs de la correction de la pondération et de l'estimation postérieure.

1. Nicholas T. Longford, SNTL Statistics Research and Consulting, Londres, Royaume-Uni, 3 Badgers Walk, Whyteleafe CR3 0AS, Surrey, Royaume-Uni. Courriel : sntlnick@sntl.co.uk.

Des approches fondées sur un modèle ou assistées par un modèle ont permis de trouver un terrain propice dans l'échantillonnage d'enquête en général et dans le calage de poids en particulier. La régression ordinaire et ses diverses généralisations ont été largement appliquées; voir Haziza et Beaumont (2017) et les références qu'ils fournissent. Une telle approche est motivée par une régression ridge (Hoerl et Kennard, 1970), selon laquelle l'impératif de respecter les contraintes d'étalonnage est modéré pour promouvoir la stabilité de la solution (poids de calage), afin d'éviter d'importantes corrections et des valeurs inacceptables et extrêmes ainsi que de réduire la dispersion des poids. Beaumont et Bocci (2008) établissent un lien entre la proposition initiale de Chambers (1996) et l'approche fondée sur la vraisemblance de Chen, Sitter et Wu (2002).

Le calage est particulièrement ardu lorsqu'un grand nombre d'informations auxiliaires sont disponibles et que de nombreux totaux de population doivent être mis en correspondance. Cardot, Goga et Shehzad (2017) ainsi que Vera, Sánchez Zuleta et Rueda (2023) abordent cet enjeu en projetant les données auxiliaires sur un sous-espace gérable pour lequel des méthodes établies peuvent être appliquées. Voir également Dagdoug, Goga et Haziza (2023) pour obtenir une approche assistée par un modèle.

Nous présentons une méthode d'étalonnage selon laquelle l'objectif de mise en correspondance des contraintes de calage est assoupli en une réduction des écarts entre les sommaires d'échantillon et leurs cibles selon certaines priorités précisées. L'algorithme que nous élaborons répond au même ensemble d'enjeux que les méthodes de calage établies, mais permet d'intégrer les priorités de l'analyste (ou celles de son client) de façon souple et transparente. L'algorithme ne nécessite aucune hypothèse associée à un modèle, mais est étroitement lié à un calage ridge, dans le cadre duquel un modèle est sous-entendu (Chambers, 1996; Rao et Singh, 2009). Les priorités s'avèrent très semblables aux réciproques des coefficients ridge. Cette approche présente des points communs avec la méthode de Guggemos et Tillé (2010), qui combinent un calage strict à une pénalisation et la considèrent comme une alternative fondée sur le plan aux procédures fondées sur des modèles mixtes. Notre approche est fondée entièrement sur la pénalisation, mais en l'associant à un calage ridge, nous établissons un lien avec des modèles linéaires.

L'algorithme est peu exigeant du point de vue du calcul et une partie de ses propriétés sont obtenues analytiquement. La spécification des priorités comme paramètres d'ajustement peut sembler être une contrainte supplémentaire; toutefois, ces paramètres facilitent un contrôle du processus de calage qui n'est pas possible dans le cadre de certaines méthodes établies.

Pour obtenir une approche computationnelle similaire dans un contexte n'ayant aucun rapport (c'est-à-dire établir un équilibre dans une inférence causale), voir Longford (2024). Elle remplace l'objectif d'atteindre un équilibre d'une qualité précisée pour deux groupes de traitement dans un ensemble de variables contextuelles, le meilleur équilibre pouvant être atteint selon les priorités de l'analyste, c'est-à-dire l'urgence ou l'importance relative de réduire le déséquilibre pour chaque variable contextuelle.

Le reste de la présente section définit la notation et présente le contexte analytique du problème. À la section suivante, nous formulons le problème, nous le résolvons et nous établissons un lien avec le calage ridge. La troisième section porte sur la configuration des paramètres d'ajustement. La quatrième section

illustre la méthode au moyen d'exemples. La cinquième section présente une étude par simulation démontrant l'efficacité de l'estimateur proposé et permet d'examiner la valeur de l'information auxiliaire et une bonne mise en œuvre du plan de sondage. La section de conclusion résume la méthode, ses points forts et tout son potentiel, et expose des enjeux non résolus.

1.1 Notation et contexte

Dans une enquête sur la population reposant sur un plan de sondage particulier, supposons un estimateur du total de la population d'une variable y, linéaire dans le vecteur de ses valeurs observées $\mathbf{y} = (y_1, \ldots, y_n)^\top$. Par exemple, lorsque la taille de l'échantillon n est fixe, il peut s'agir de l'estimateur Horvitz-Thompson (Horvitz et Thompson, 1952), $\hat{\theta}_{\mathrm{HT}} = \mathbf{w}^\top \mathbf{y}$, où $\mathbf{w} = (w_1, \ldots, w_n)^\top$ est le vecteur des poids d'échantillonnage, égal aux réciproques des probabilités de l'inclusion des sujets dans l'échantillon. Ces probabilités sont définies par le plan et potentiellement ajustées une fois l'échantillon tiré.

Supposons que les totaux de population t_k sont connus pour une ou plusieurs variables X_k, $k = 1, \ldots, K$. Soit $\mathbf{X}^\circ$, leur collecte. Nous utilisons également $\mathbf{X}^\circ$ pour la matrice $n \times K$ formée par les valeurs de ces variables dans l'échantillon; $\mathbf{x}_k$ désigne la colonne k de $\mathbf{X}^\circ$. Pour simplifier l'explication, nous supposons que toutes les variables X_k dans $\mathbf{X}^\circ$ sont soit ordinales (continues) soit binaires. Une variable discrète avec $H \geq 2$ catégories est représentée dans $\mathbf{X}^\circ$ par H variables binaires (fictives). La singularité de $\mathbf{X}^\circ$ n'entraînera aucun enjeu. Nous réservons l'indice 0 pour l'ordonnée à l'origine, $\mathbf{x}_0 = (1, \ldots, 1)^\top$, et désignons $\mathbf{X} = (\mathbf{x}_0, \mathbf{X}^\circ)$.

Le calage est défini comme la transformation des poids, $\mathbf{u} = C(\mathbf{w})$, selon laquelle le total pondéré $\mathbf{x}_k^\top \mathbf{u}$ correspond au total de population t_k pour chaque variable k. Cela signifie que le calage fait en sorte que $\mathbf{X}^{\circ\top}\mathbf{u} = \mathbf{t}^\circ$, où $\mathbf{t}^\circ = (t_1, \ldots, t_K)^\top$. Les transformations des variables initiales, y compris les interactions (produits), peuvent être ajoutées à $\mathbf{X}^\circ$ lorsque les totaux de population sont connus ou sont estimés selon un niveau d'erreur négligeable. D'autres sommaires de population, comme des variances et des quantiles, peuvent également être mis en correspondance.

Nous écrivons $C(\mathbf{w}) = C(\mathbf{w}; \mathbf{X}^\circ, \mathbf{t}^\circ)$ pour désigner l'information auxiliaire intervenant; cela est utile lorsque nous envisageons les variables à inclure dans $\mathbf{X}^\circ$. Nous pouvons qualifier l'estimateur $\hat{\theta}$ de manière similaire, en écrivant $\hat{\theta}(\mathbf{u})$ et $\hat{\theta}\{C(\mathbf{w})\}$ ou, sous une forme plus complète, $\hat{\theta}\{C(\mathbf{w}; \mathbf{X}^\circ, \mathbf{t}^\circ)\}$. Les variables dépendantes $\mathbf{y}$ ne jouent aucun rôle dans la recherche de $\mathbf{u} = C(\mathbf{w})$. Par conséquent, tant que $\mathbf{y}$ n'est pas inspecté avant de choisir un calage particulier $C(\mathbf{w})$, les propriétés de l'estimateur $\hat{\theta}(\mathbf{u})$ peuvent être évaluées sans tenir compte de la façon dont $\mathbf{u}$ a été obtenu. Aucune préoccupation ne survient pour ce qui est de tirer parti du hasard ou de la surexploitation des données, même si le calage est examiné à l'aide de plusieurs matrices $\mathbf{X}^\circ$ et des paramètres intervenant dans C. Bien sûr, les propriétés de $\hat{\theta}(\mathbf{u})$ dépendent de $\mathbf{u}$.

Dans une certaine perspective, lorsque l'erreur d'estimation $\hat{\theta} - \theta$ et son sommaire stochastique, comme le biais ou l'erreur quadratique moyenne (EQM), sont la seule préoccupation, le calage présente une qualité distinctement superficielle. Dans une autre, prédominante en pratique, ce calage est essentiel pour la

crédibilité des estimations, même au détriment d'un certain biais et de l'inflation de l'EQM. Ce n'est que si cette inflation est substantielle, ou que les poids $C(\mathbf{w})$ sont bien plus dispersés que $\mathbf{w}$, qu'une improvisation est requise. Cela a souvent lieu pour des valeurs de K relativement grandes, lorsqu'il existe de nombreuses variables auxiliaires et des contraintes qui y sont associées. Ce problème est généralement résolu en supprimant des variables dans $\mathbf{X}^\circ$.

Nous considérons la dichotomie d'inclure ou d'exclure une variable du processus de calage comme étant trop rigide et élaborons une approche selon laquelle des propriétés attribuées aux contraintes de calage reflètent l'importance ou l'urgence de mettre en correspondance le total d'échantillon pondéré de la variable auxiliaire X_k avec sa cible (de population) t_k. Une priorité est également attribuée à d'autres propriétés souhaitables : faible modification des poids par l'ajustement, préférence pour une plus faible dispersion des éléments de $\mathbf{u}$ et aversion contre une modification du total des poids. En bref, les contraintes établies de correspondance exacte (sans écart) sont remplacées par des pénalités d'écarts. Ces pénalités offrent une marge de manœuvre ou une certaine liberté; elles font intervenir des coefficients définis par l'utilisateur qui quantifient la priorité relative des contraintes.

Notre formulation du problème mène à une optimisation quadratique ayant une solution analytique. Elle ne fait intervenir aucune itération et aucune inversion (numérique) de matrice importante, malgré la correspondance de nombreuses marges pour des données à grande échelle. Pour un ensemble de paramètres d'ajustement, appelés *priorités*, la solution est unique et sa dépendance envers ces priorités est facile à examiner, sans nécessiter de théorie asymptotique et d'expérimentation étendue à l'aide d'un algorithme de type boîte noire.

L'interprétation naturelle de ces priorités est l'importance ou l'urgence de chaque contrainte de calage. Plus précisément, ces contraintes ne sont pas exactement satisfaites, comme cela est prévu dans un calage strict, mais chaque écart $\delta_k = \mathbf{x}_k^\top \mathbf{u} - t_k$, un élément de $\boldsymbol{\delta}^\circ = \mathbf{X}^{\circ\top}\mathbf{u} - \mathbf{t}^\circ$, peut être rendu négligeable, $\delta_k \doteq 0$, en attribuant à la priorité correspondante une valeur suffisamment élevée. À l'autre extrémité, une priorité nulle pour la variable k est équivalente à l'abandon de la colonne $\mathbf{x}_k$ dans $\mathbf{X}^\circ$. Une telle souplesse peut être considérée comme une distraction, imposant le fardeau de déclarer les priorités et de devoir justifier ce choix dans un rapport ultérieur. Toutefois, elle offre l'occasion d'intégrer à l'analyse la perspective, les jugements de valeur, les renseignements et les compétences du client. De plus, un ou plusieurs écarts aberrants $|\delta_k|$ peuvent être réduits en augmentant les priorités correspondantes, possiblement au prix d'une augmentation de certains autres écarts.

2. Optimisation sans contrainte

Soit $\mathbf{0}_K$, le vecteur de zéros de longueur K. On laisse tomber l'indice K lorsque la longueur du vecteur est évidente dans le contexte. Nous utilisons les symboles $\mathbf{1}_n$ et $\mathbf{1}$ de façon similaire pour le vecteur de uns, et $\mathbf{I}$ pour la matrice identité. Pour une population de taille (finie) N, nous considérons un échantillon de taille (fixe ou aléatoire) $n \ll N$ ayant des vecteurs d'observations $\mathbf{y}$ et des poids de base $\mathbf{w}$.

Une approche typique en matière de calage impose la contrainte $\delta_k = 0$ ou précise une limite supérieure Δ_k sur $|\delta_k|$ pour chaque $k = 1, \ldots, K$. Ces contraintes peuvent être remplacées par une seule limite supérieure pour la somme des carrés $\boldsymbol{\delta}^{\circ\top}\boldsymbol{\delta}^{\circ} = \delta_1^2 + \cdots + \delta_K^2$. Pour des valeurs de K plus grandes, nous pouvons distinguer des variables k pour lesquelles la correspondance, $\delta_k = 0$, est plus importante que pour le reste. De plus, un coefficient de priorité distinct peut être attribué à chaque variable, ou les variables peuvent être réparties dans des ensembles à coefficients constants au sein de ces ensembles. Les carrés dans la somme $\sum_k \delta_k^2$ peuvent être associés à des poids, ce qui impose une limite supérieure à $\boldsymbol{\delta}^{\circ\top}\mathbf{P}^{\circ}\boldsymbol{\delta}^{\circ} = \sum_k p_k^{\circ}\delta_k^2$, les priorités $p_k^{\circ} > 0$ étant définies par l'analyste; $\mathbf{P}^{\circ}$ est la matrice diagonale ayant $p_1^{\circ}, \ldots, p_K^{\circ}$ sur sa diagonale.

Ces façons d'assouplir les contraintes de calage et d'introduire des priorités pour réduire les écarts motivent notre proposition. Nous attribuons une priorité non négative p_k° à chaque variable k et envisageons d'abord de trouver le minimum de la fonction

$$F^{\circ}(\mathbf{u}; \mathbf{w}) = \sum_{k=1}^{K} p_k^{\circ}\delta_k^2,$$

selon les contraintes d'une faible déviation de $\mathbf{u}$ par rapport à $\mathbf{w}$ et du total des poids correspondant à la taille de la population N; $\mathbf{u}^{\top}\mathbf{1}_n = N$. Lorsque N n'est pas connu, nous le remplaçons par son estimation $\mathbf{w}^{\top}\mathbf{1}_n$. Le paramètre p_k° est confondu avec l'échelle de $\mathbf{x}_k$; le remplacement de $\mathbf{x}_k$ par $c^{-1}\mathbf{x}_k$ est compensé par le remplacement de p_k° par $c^2 p_k^{\circ}$. Pour éviter l'ambiguïté associée, nous supposons que chaque $\mathbf{x}_k$ est normalisé, c'est-à-dire transformé de façon linéaire pour présenter une moyenne et une variance d'unité nulle; t_k est ajusté en conséquence. Une autre option est de conserver les valeurs $\mathbf{x}_k$ et t_k et de remettre à l'échelle la priorité correspondante.

Nous réduisons la contrainte $\mathbf{u}^{\top}\mathbf{1} = t_0$, où $t_0 = N$ ou $t_0 = \mathbf{w}^{\top}\mathbf{1}$, et l'intégrons dans la fonction objectif F°. Nous définissons la priorité p_0° par l'importance attribuée à δ_0^2 atteignant une faible valeur. La fonction adaptée F° comprend le nouveau terme $p_0^{\circ}\delta_0^2$; c'est-à-dire que sa sommation est désormais de $k = 0$ à K. Cela correspond à attribuer $\mathbf{x}_0 = \mathbf{1}_n$ à $\mathbf{X}^{\circ}$ comme autre colonne, ce qui forme ainsi la matrice $\mathbf{X}$.

Nous attribuons la priorité S à une faible dispersion des poids $\mathbf{u}$, motivée par le souhait d'avoir une faible variance de $\hat{\theta}(\mathbf{u}) = \mathbf{u}^{\top}\mathbf{y}$, et R au souhait d'avoir une faible modification des poids, ce qui est indirectement associé à une réduction du biais, en préférant que $\hat{\theta}(\mathbf{u})$ demeure proche de l'estimateur $\hat{\theta}(\mathbf{w})$, qui serait sans biais si les poids $\mathbf{w}$ étaient corrects et n était fixe. Il convient de noter qu'une faible variance ne peut pas être synonyme d'efficacité.

Au lieu de $F^{\circ}(\mathbf{u})$, nous trouvons le minimum (sans contrainte) de la fonction

$$\begin{aligned} F(\mathbf{u}; \mathbf{w}) &= \sum_{k=0}^{K} p_k^{\circ}\delta_k^2 + R(\mathbf{u}-\mathbf{w})^{\top}(\mathbf{u}-\mathbf{w}) + S\left(\mathbf{u}^{\top}\mathbf{u} - \frac{1}{n}\mathbf{u}^{\top}\mathbf{1}\mathbf{1}^{\top}\mathbf{u}\right). \\ &= (R+S)\,\mathbf{u}^{\top}\mathbf{u} - \frac{1}{n}S\,\mathbf{u}^{\top}\mathbf{1}\mathbf{1}^{\top}\mathbf{u} + \sum_{k=0}^{K} p_k^{\circ}\delta_k^2 - 2R\,\mathbf{u}^{\top}\mathbf{w} + R\,\mathbf{w}^{\top}\mathbf{w}. \end{aligned}$$

Puisque l'optimisation de cF pour des constantes positives c constitue des problèmes identiques, aucune généralité n'est perdue en supposant que $R+S=1$. En ayant cette convention et en développant

$$p_k^\circ \delta_k^2 \;=\; p_k^\circ \mathbf{u}^\top \mathbf{x}_k \mathbf{x}_k^\top \mathbf{u} - 2p_k^\circ t_k \mathbf{u}^\top \mathbf{x}_k + p_k^\circ t_k^2$$

pour $k=0,\ldots,K,$ nous pouvons exprimer F de façon compacte sous la forme

$$F(\mathbf{u}) \;=\; \mathbf{u}^\top \mathbf{H}\mathbf{u} - 2\mathbf{u}^\top \mathbf{s} + D,$$

où

$$\begin{aligned} \mathbf{H} &= \mathbf{I}_n + \sum_{k=0}^{K} p_k \mathbf{x}_k \mathbf{x}_k^\top = \mathbf{I}_n + \mathbf{X}\mathbf{P}\mathbf{X}^\top \\ \mathbf{s} &= R\mathbf{w} + (1-R)\,\frac{t_0}{n}\,\mathbf{1}_n + \sum_{k=0}^{K} p_k t_k \mathbf{x}_k = \mathbf{w}_R + \mathbf{X}\mathbf{P}\mathbf{t}; \end{aligned} \tag{2.1}$$

$\mathbf{w}_R = R\mathbf{w} + (1-R)\,n^{-1}t_0\mathbf{1}_n$, $p_0 = p_0^\circ - \frac{1}{n}S$, $p_k = p_k^\circ$ pour $k=1,\ldots,K$, $\mathbf{P}=\operatorname{diag}(\mathbf{p})$, où $\mathbf{p}=(p_0, p_1, \ldots, p_K)$ et $D = \mathbf{t}^\top \mathbf{P}\mathbf{t} + R\mathbf{w}^\top \mathbf{w} + (1-R)\,t_0^2/n$ est un scalaire positif non pertinent pour ce qui suit. Le minimum de $F(\mathbf{u})$ est atteint pour $\mathbf{u}=\mathbf{H}^{-1}\mathbf{s}$, et le minimum atteint est $F(\mathbf{u}) = D - \mathbf{s}^\top \mathbf{H}^{-1}\mathbf{s}$. L'estimateur de calage de θ est $\hat{\theta}(\mathbf{u}) = \mathbf{y}^\top \mathbf{u} = \mathbf{y}^\top \mathbf{H}^{-1}\mathbf{s}$. La troisième section porte sur la façon de définir les valeurs des priorités p_k et R.

2.1 Inversion de H

Pour un échantillon de grande taille n, l'inversion numérique de $\mathbf{H}$ peut sembler être une difficulté computationnelle non triviale. Nous appliquons un algorithme récursif qui exploite la forme de $\mathbf{H}$ comme la somme d'une matrice facile à inverser, $\mathbf{I}$, et un, $\mathbf{X}\mathbf{P}\mathbf{X}^\top$, de faible rang (relativement); $K+1 \ll n$ au plus. Il convient de noter, en passant, que la singularité de la matrice $\mathbf{X}=(\mathbf{x}_0, \mathbf{X}^\circ)$ ne soulève aucun enjeu computationnel, même si cela peut avoir des conséquences sur l'interprétation des priorités dans $\mathbf{p}$. Par exemple, si $\mathbf{x}_k = \mathbf{x}_l$ pour certaines valeurs $k \neq l$, alors la variable commune est plus adéquatement associée à la priorité $p_k + p_l$. De plus, un conflit apparaît lorsque $\mathbf{x}_0 = \mathbf{1}_n$ est égal au total de l'ensemble des colonnes de $\mathbf{X}^\circ$ qui correspondent aux indicateurs pour une variable catégorique (avant la normalisation), mais le total des cibles correspondantes (chiffres de population) t_k diffère de t_0. Un tel conflit n'a pas d'incidence sur l'algorithme.

Soit $\mathbf{H}_{-1} = \mathbf{I}_n$ et $\mathbf{H}_k = \mathbf{H}_{k-1} + p_k \mathbf{x}_k \mathbf{x}_k^\top$, $k=0,\ldots,K$, de sorte que $\mathbf{H}=\mathbf{H}_K$. Nous obtenons l'identité

$$\mathbf{H}_k^{-1} \;=\; \mathbf{H}_{k-1}^{-1} \;-\; \frac{p_k}{1+p_k \mathbf{x}_k^\top \mathbf{H}_{k-1}^{-1}\mathbf{x}_k}\;\mathbf{H}_{k-1}^{-1}\mathbf{x}_k\mathbf{x}_k^\top\mathbf{H}_{k-1}^{-1}.$$

Sa validité est facile à vérifier en évaluant le produit des expressions pour $\mathbf{H}_k$ et $\mathbf{H}_k^{-1}$. Dans l'évaluation récursive de $\mathbf{u}=\mathbf{H}^{-1}\mathbf{s}$, nous n'avons pas à former de matrice $\mathbf{H}_k$ ou $\mathbf{H}_k^{-1}$, car nous avons uniquement besoin des vecteurs $\mathbf{h}_{kl} = \mathbf{H}_k^{-1}\mathbf{x}_l$ et $\mathbf{h}_{k,\mathrm{w}} = \mathbf{H}_k^{-1}\mathbf{w}$. Pour le premier, nous avons les identités

$$\mathbf{h}_{kl} \;=\; \mathbf{h}_{k-1,l} \;-\; \frac{p_k \mathbf{x}_k^\top \mathbf{h}_{k-1,l}}{1+p_k \mathbf{x}_k^\top \mathbf{h}_{k-1,k}} \;\mathbf{h}_{k-1,k},$$

et pour le deuxième, nous avons les mêmes identités, mais l'indice l est remplacé par w. Il convient de noter que chaque dénominateur $1+p_k \mathbf{x}_k^\top \mathbf{H}_{k-1}^{-1} \mathbf{x}_k$ est positif. La solution

$$\mathbf{u} \;=\; R\mathbf{h}_{K,\mathrm{w}} \;+\; (1-R)\;\frac{t_0}{n}\;\mathbf{h}_{K0} \;+\; \sum_{k=0}^{K} p_k t_k \mathbf{h}_{Kk},$$

est une combinaison linéaire des vecteurs $\mathbf{h}_{Kk} = \mathbf{H}^{-1}\mathbf{x}_k$ et $\mathbf{h}_{K,\mathrm{w}} = \mathbf{H}^{-1}\mathbf{w}$. En résumé, il existe une solution unique $\mathbf{u}$ et elle est uniquement évaluée par des opérations sur des vecteurs de longueur n. La matrice $\mathbf{H}$ ne fait intervenir ni R ni $\mathbf{w}$. Le vecteur $\mathbf{s}$ est une fonction linéaire à la fois pour R et $\mathbf{w}$; par conséquent, il en est de même pour la solution $\mathbf{u}$. Pour $R=0$, $\mathbf{u}$ ne dépend pas de $\mathbf{w}$.

2.2 Lien avec le calage ridge

Dans la présente section, nous montrons que les priorités $\mathbf{p}$ ont un rôle semblable aux coûts de la méthode de ridge dans Chambers (1996), équation (10), même si la fonction F intègre toutes les contraintes et tous les objectifs du calage, et fait donc intervenir des paramètres supplémentaires. Tout comme la méthode ridge peut être considérée comme un compromis entre l'application d'aucune méthode ridge et d'une méthode ridge infinie, notre proposition est un compromis entre $\mathbf{p}=\mathbf{0}$ et un $\mathbf{p}$ infiniment grand. Les deux approches donnent des estimateurs pouvant être interprétés comme des estimateurs de rétrécissement.

Dans l'équation (2.1), $\mathbf{w}_R$ est une combinaison convexe de $\mathbf{w}$ et de t_0/n, interprétée comme un rétrécissement de $\mathbf{w}$ vers sa moyenne ou son espérance. On écrit $\mathbf{w}_R = \mathbf{X}\mathbf{v}_R + \boldsymbol{\varepsilon}_R$, où $\mathbf{v}_R = (\mathbf{X}^\top\mathbf{X})^{-}\mathbf{X}^\top\mathbf{w}_R$ est un vecteur de projection et $\boldsymbol{\varepsilon}_R$ est tel que $\mathbf{X}^\top\boldsymbol{\varepsilon}_R = \mathbf{0}$. Dans ce cas-ci, $(\mathbf{X}^\top\mathbf{X})^{-}$ est une inverse généralisée de $\mathbf{X}^\top\mathbf{X}$. Son caractère non unique est résolu ci-dessous. Nous avons la décomposition analogue $\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\omega}$, où $\boldsymbol{\gamma} = (\mathbf{X}^\top\mathbf{X})^{-}\mathbf{X}^\top\mathbf{y}$, de sorte que $\mathbf{X}^\top\boldsymbol{\omega} = \mathbf{0}$. Les vecteurs $\mathbf{v}_R$ et $\boldsymbol{\gamma}$ sont, respectivement, des ajustements par les moindres carrés pour $\mathbf{w}_R$ et $\mathbf{y}$ en termes de $\mathbf{X}$. Nous ne supposons aucun aspect de validité des régressions ordinaires implicites $(\mathbf{w}_\mathrm{R} \mid \mathbf{X})$ et $(\mathbf{y} \mid \mathbf{X})$. Il convient de noter que les résidus $\boldsymbol{\varepsilon}_R$ et $\boldsymbol{\omega}$ sont uniques, et $\boldsymbol{\varepsilon}_R^\top\mathbf{y} = \boldsymbol{\varepsilon}_R^\top\boldsymbol{\omega}$.

Soit $\mathbf{e}_0 = (1, 0, \ldots, 0)^\top$, le vecteur comprenant une unité et K zéros. Pour une matrice $\mathbf{A}$ de $K+1$ colonnes, $\mathbf{A}\mathbf{e}_0$ est égal à sa première colonne. Du fait de l'orthogonalité, $\mathbf{X}^{\circ\top}\mathbf{x}_0 = \mathbf{0}$, qui est organisé par normalisation, nous obtenons l'identité $(\mathbf{X}^\top\mathbf{X})^{-}\mathbf{X}^\top\mathbf{1}_n = n(\mathbf{X}^\top\mathbf{X})^{-}\mathbf{e}_0 = \mathbf{e}_0$. Ainsi

$$\mathbf{v}_R \;=\; R\left(\mathbf{X}^\top\mathbf{X}\right)^{-}\mathbf{X}^\top\mathbf{w} \;+\; \frac{(1-R)\,t_0}{n}\;\mathbf{e}_0, \tag{2.2}$$

et par sa substitution dans l'équation (2.1),

$$\mathbf{u} \;=\; \mathbf{H}^{-1}\mathbf{s} \;=\; \left(\mathbf{I} \;+\; \mathbf{X}\mathbf{P}\mathbf{X}^\top\right)^{-1}\left\{\mathbf{X}\left(\mathbf{P}\mathbf{t} \;+\; \mathbf{v}_R\right) \;+\; \boldsymbol{\varepsilon}_R\right\}.$$

L'identité $(\mathbf{I}+\mathbf{XPX}^\top)^{-1}=\mathbf{I}-\mathbf{X}\left(\mathbf{P}^{-1}+\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ suppose que

$$\begin{aligned}\mathbf{u} &= \mathbf{X}\left\{\mathbf{I}-\left(\mathbf{P}^{-1}+\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\right\}\left(\mathbf{Pt}+\boldsymbol{\nu}_R\right)+\left\{\mathbf{I}-\mathbf{X}\left(\mathbf{P}^{-1}+\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right\}\boldsymbol{\varepsilon}_R\\ &= \mathbf{X}\left(\mathbf{P}^{-1}+\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{t}+\mathbf{P}^{-1}\boldsymbol{\nu}_R\right)+\boldsymbol{\varepsilon}_R.\end{aligned}$$

Par conséquent, l'estimateur de calage est $\hat\theta(\mathbf{u})=\hat\theta_1+\hat\theta_2(\mathbf{w})+\hat\theta_3(\mathbf{w})$, où

$$\begin{aligned}\hat\theta_1 &= \mathbf{t}^\top\left(\mathbf{P}^{-1}+\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}\\ \hat\theta_2(\mathbf{w}) &= \boldsymbol{\nu}_R^\top\mathbf{P}^{-1}\left(\mathbf{P}^{-1}+\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}\\ \hat\theta_3(\mathbf{w}) &= \boldsymbol{\varepsilon}_R^\top\boldsymbol{\omega}.\end{aligned}\qquad(2.3)$$

Dans ce cas-ci, $\hat\theta_1$ est un estimateur du total de la population θ fondé sur la prédiction de la régression ridge par le modèle linéaire $(\mathbf{y}\mid\mathbf{X})$. Sans faire intervenir les poids $\mathbf{w}$ et selon la méthode ridge, $\hat\theta_1$ est sans biais uniquement lorsque les biais dus à l'absence de pondération et à la méthode ridge s'annulent par hasard. Lorsque $\boldsymbol{\varepsilon}_R^\top\boldsymbol{\omega}=0$, $\hat\theta(\mathbf{u})$ prend la forme d'un calage ridge (Chambers, 1996; Goga et Shehzad, 2010), où $\mathbf{P}$ joue le rôle de la matrice (diagonale) de coût et $\mathbf{t}$ est corrigé par $\mathbf{P}^{-1}\boldsymbol{\nu}_R$. Toutefois, les produits croisés $\mathbf{X}^\top\mathbf{X}$ et $\mathbf{X}^\top\mathbf{y}$ sont évalués sans les poids $\mathbf{w}$.

À mesure que $\mathbf{P}^{-1}$ converge vers la matrice zéro, qui correspond à un intérêt décroissant dans la déviation de $\mathbf{u}$ par rapport à $\mathbf{w}$ ainsi que dans la dispersion de $\mathbf{u}$, $\hat\theta(\mathbf{u})$ se rapproche de $\mathbf{t}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}+\boldsymbol{\varepsilon}_R^\top\boldsymbol{\omega}$ lorsque $\mathbf{X}$ est non singulier. Cela se réduit au prédicteur par les moindres carrés lorsque $\boldsymbol{\varepsilon}_R^\top\boldsymbol{\omega}=0$. Si $\mathbf{X}$ présente un rang déficient et que toutes les priorités dans $\mathbf{p}$ sont grandes, alors $\mathbf{P}^{-1}+\mathbf{X}^\top\mathbf{X}$ a une ou plusieurs petites valeurs propres, et donc $\hat\theta(\mathbf{u})$ est instable.

Comme chaque priorité dans $\mathbf{p}$ converge vers zéro, $\hat\theta(\mathbf{u})$ converge vers l'estimateur de rétrécissement $\mathbf{w}_R^\top\mathbf{y}=R\mathbf{w}^\top\mathbf{y}+(1-R)\,t_0\,\bar y$, où $\bar y$ est la moyenne d'échantillon (non pondérée) de y. Pour $R=0$, il s'agit de l'estimateur trivial (non pondéré) $t_0\bar y$ et pour $R=1$, il s'agit de la moyenne pondérée $\mathbf{w}^\top\mathbf{y}$. Il convient de noter que les éléments de $\mathbf{p}$ (la diagonale de $\mathbf{P}$) ne sont pas définis par les considérations habituelles de la régression ridge visant à réduire la variance d'échantillonnage en échange d'un faible biais. Introduire ces considérations n'est pas simple, du fait de la contribution de $\hat\theta_2+\hat\theta_3$ à $\hat\theta$.

Soit $\hat{\mathbf{y}}$, le vecteur de projection $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{y}$ et $\hat{\mathbf{y}}_{\mathrm{P}}$, sa version ridge $\mathbf{X}(\mathbf{P}^{-1}+\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. Alors $\hat\theta_3(\mathbf{w})=R\boldsymbol{\varepsilon}_1^\top\boldsymbol{\omega}=R\mathbf{w}^\top(\mathbf{y}-\hat{\mathbf{y}})$ et, selon l'équation (2.2),

$$\begin{aligned}\hat\theta_2(\mathbf{w}) &= R\mathbf{w}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^-\mathbf{P}^{-1}\left(\mathbf{P}^{-1}+\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}\\ &\quad+\frac{(1-R)\,t_0}{n}\,\mathbf{e}_0^\top\mathbf{P}^{-1}\left(\mathbf{P}^{-1}+\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}\\ &= R\mathbf{w}^\top\mathbf{X}\left\{\left(\mathbf{X}^\top\mathbf{X}\right)^--\left(\mathbf{P}^{-1}+\mathbf{X}^\top\mathbf{X}\right)^{-1}\right\}\mathbf{X}^\top\mathbf{y}+\frac{1-R}{1+np_0}\,\frac{t_0}{n}\,\mathbf{y}^\top\mathbf{1}_n\\ &= R\mathbf{w}^\top\left(\hat{\mathbf{y}}-\hat{\mathbf{y}}_{\mathrm{P}}\right)+(1-R)\,\frac{t_0\bar y}{1+np_0}.\end{aligned}$$

Par conséquent, le total $\hat{\theta}_2(\mathbf{w})+\hat{\theta}_3(\mathbf{w})$ est

$$\hat{\theta}_{23} = R\mathbf{w}^\top\left(\mathbf{y}-\hat{\mathbf{y}}_{\mathrm{P}}\right)+(1-R)\,\frac{t_0\overline{y}}{1+np_0}.$$

Il s'agit d'une fonction linéaire de $\mathbf{w}$ et de R, rétrécissant le total pondéré des erreurs de prédiction, $\mathbf{w}^\top(\mathbf{y}-\hat{\mathbf{y}}_{\mathrm{P}})$, vers une quantité ne faisant pas intervenir $\mathbf{X}$. La colonne 0 de $\mathbf{P}^{-1}+\mathbf{X}^\top\mathbf{X}$ est égale à $\left(1/p_0+n\right)\mathbf{e}_0$ et, par conséquent, en rappelant la notation $\mathbf{X}^\circ$, $\mathbf{P}^\circ$ et $\mathbf{t}^\circ$, $\hat{\theta}(\mathbf{u})$ peut être exprimé sous la forme

$$\mathbf{t}^{\circ\top}\left(\mathbf{P}^{\circ-1}+\mathbf{X}^{\circ\top}\mathbf{X}^\circ\right)^{-1}\mathbf{X}^{\circ\top}\mathbf{y}+R\mathbf{w}^\top\left(\mathbf{y}-\hat{\mathbf{y}}_{\mathrm{P}}\right)+\left(1-\frac{R}{1+np_0}\right)t_0\overline{y}.$$

Il est utile de présenter $\hat{\theta}(\mathbf{u};R)$ comme l'interpolation linéaire de

$$\begin{aligned}\hat{\theta}(\mathbf{u};R=0) &= \frac{t_0\,\overline{y}}{1+np_0}+\mathbf{t}^\top\hat{\boldsymbol{\beta}}\\ &= t_0\,\overline{y}+\mathbf{t}^{\circ\top}\hat{\boldsymbol{\beta}}^\circ\\ \hat{\theta}(\mathbf{u};R=1) &= \mathbf{w}^\top\mathbf{y}-\boldsymbol{\delta}_{\mathrm{w}}^\top\hat{\boldsymbol{\beta}},\end{aligned}$$

où $\hat{\boldsymbol{\beta}}=\left(\mathbf{P}^{-1}+\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$, $\hat{\boldsymbol{\beta}}^\circ=(\mathbf{P}^{\circ-1}+\mathbf{X}^{\circ\top}\mathbf{X}^\circ)^{-1}\mathbf{X}^\top\mathbf{y}$ et $\boldsymbol{\delta}_{\mathrm{w}}=\mathbf{X}^\top\mathbf{w}-\mathbf{t}$ est le vecteur de divergence évalué au moyen des poids $\mathbf{w}$. Les vecteurs de dimension $K\times 1$ $\mathbf{t}^\circ$ et $\hat{\boldsymbol{\beta}}^\circ$ sont formés à partir de $\mathbf{t}$ et $\hat{\boldsymbol{\beta}}$ en supprimant leurs premiers éléments respectifs. Ainsi, $\hat{\theta}(\mathbf{u};R=0)$ peut être décrit comme l'estimateur trivial $t_0\,\overline{y}$ corrigé par la prédiction ridge à $\mathbf{t}^\circ$. Pour $R\geq 0$,

$$\hat{\theta}(\mathbf{u};R=r) = \hat{\theta}(\mathbf{u};R=0)+r\left\{\mathbf{w}^\top\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)-\frac{t_0\,\overline{y}}{1+np_0}\right\}; \tag{2.4}$$

autrement dit, $\hat{\theta}(\mathbf{u};R=0)$ est corrigé par un multiple de R de l'estimateur de l'erreur sur le total de population dans la prédiction de $\mathbf{y}$ et une fraction de $t_0\,\overline{y}$. Définir la valeur de R importe davantage lorsque les erreurs de prédiction $\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}=\mathbf{y}-\hat{\mathbf{y}}$ sont importantes et corrélées aux poids de base $\mathbf{w}$. Il convient de noter que $\hat{\theta}$ est évalué sans utiliser $(\mathbf{X}^\top\mathbf{X})^-$.

Le biais de $\hat{\theta}$ est une fonction linéaire de R; il est donc minimiser soit pour $R=0$ ou 1. Même si $\hat{\theta}$ ne fait pas intervenir $\mathbf{w}$ lorsque $R=0$, il aurait le biais le plus faible lorsque l'effet des poids est bien estimé par le modèle de prédiction. C'est le cas lorsque les poids sont construits uniquement à l'aide de variables dans $\mathbf{X}$, pour lesquelles les totaux de population sont disponibles. Nous confirmons cela dans des simulations à la cinquième section.

À l'annexe A, nous obtenons et examinons une expression pour $\hat{\theta}$ dans l'équation (2.4) lorsque le calage a lieu sur une variable catégorique unique (et la taille de la population). La décomposition de l'estimateur est $\hat{\theta}=\hat{\theta}^{(1)}+\hat{\theta}^{(Z)}$, où $\hat{\theta}^{(1)}$ dépend de $\boldsymbol{\mu}$, de $\mathbf{p}$ et de $\mathbf{t}$ uniquement par l'entremise de p_0 et de t_0, et $\hat{\theta}^{(Z)}$

ne dépend pas de p_0. Alors, $\hat{\theta}^{(Z)} = 0$ lorsque $p_1 = \ldots = p_K = 0$. Dans ce cas, lors d'un calage uniquement sur la taille de la population,

$$\hat{\theta} = \hat{\theta}^{(1)} = R\left(\mathbf{w}^\top \mathbf{y} - \frac{n\,p_0}{1+n\,p_0}\,\mathbf{w}^\top \mathbf{1}_n \bar{y}\right) + \left(1 - \frac{R}{1+n\,p_0}\right) t_0\,\bar{y}.$$

Pour $R = 0$, nous obtenons $\hat{\theta} = t_o \bar{y}$ pour tout p_0. Pour $R = 1$, $\hat{\theta} = \mathbf{w}^\top \mathbf{y} + np_0/(1+np_0) \times (t_0 - \mathbf{w}^\top \mathbf{1}_n)\,\bar{y}$, qui converge vers l'ajustement « évident » $\mathbf{w}^\top \mathbf{y} + (t_0 - \mathbf{w}^\top \mathbf{1}_n)\,\bar{y}$ lorsque $p_0 \to +\infty$.

3. Établissement des priorités

Dans certaines applications, les contraintes sont de même importance pour toutes les variables, mais seulement après la prise en compte des dispersions de ces variables. Comme on l'a énoncé plus tôt, nous supposons que toutes les variables dans $\mathbf{X}^\circ$ sont normalisées. Établir les priorités $p_1, \ldots, p_K$ par rapport à une constante commune est un point de départ ou une valeur par défaut raisonnable. Pour de grandes valeurs de K, nous pouvons définir un petit nombre de groupes de variables et attribuer une priorité commune au sein de chaque groupe. Par exemple, ces groupes de variables peuvent être associés à des variables catégoriques distinctes. L'ordonnée à l'origine $\mathbf{x}_0 = \mathbf{1}$ a un statut particulier au sein des variables dans $\mathbf{X}$. La priorité élevée p_0 correspond au souhait d'avoir un petit $|\delta_0|$. Lorsque $t_0 = N$ et qu'une valeur suffisamment grande est définie pour p_0, $\mathbf{u}^\top \mathbf{1} \doteq N$.

Par construction, $0 < R < 1$, donc $R = 0{,}5$ peut être une valeur par défaut. Une approche plus approfondie évaluerait l'importance relative d'une petite variation (R) et d'une faible variance (S) des poids de calage. Dans le cas de valeurs n plus grandes, les préoccupations en matière de biais deviennent importantes, donc un S plus petit et un $R = 1 - S$ plus grand sont adéquats. Attribuer à R une très petite valeur est généralement déconseillé, car $\hat{\theta}$ dépend alors uniquement faiblement de $\mathbf{w}$. Réduire la différence $\mathbf{u} - \mathbf{w}$ et la dispersion des poids sont deux manières d'éviter des poids extrêmes (très grands et négatifs). L'importance de ces objectifs augmente en réduisant les priorités, par exemple de $\mathbf{p}$ à $c\,\mathbf{p}$ pour $0 < c < 1$.

Nous convenons que ces lignes directrices sont plutôt vagues et incomplètes. Toutefois, la simplicité computationnelle de la solution nous permet d'examiner un éventail de paramètres plausibles de $\mathbf{p}$ et de R, en particulier lorsque $\mathbf{p}$ fait uniquement intervenir quelques valeurs distinctes. Par exemple, lorsque $\mathbf{X}^\circ$ est fondé sur une variable discrète unique ayant H catégories, alors $p_1 = \ldots = p_H$, et nous obtenons uniquement trois paramètres d'ajustement : p_0, p_1 et R. Dans le cas de deux variables discrètes, quatre paramètres interviennent. Lorsque le tableau complet à deux entrées des marges de population pour ces deux variables est disponible, faire correspondre les marges univariées est généralement plus important que de faire correspondre les sous-totaux à deux entrées.

La dépendance de l'écart ajusté $\delta_k = \mathbf{x}_k^\top \mathbf{u} - t_k$ sur p_k peut être examinée analytiquement. Par différentiation matricielle, nous obtenons l'identité

$$\begin{aligned} \frac{\partial \delta_l^2}{\partial p_k} &= -2\delta_l \left(\mathbf{x}_l^\top \mathbf{H}^{-1} \frac{\partial \mathbf{H}}{\partial p_k} \mathbf{H}^{-1} \mathbf{s} - \mathbf{x}_l^\top \mathbf{H}^{-1} \frac{\partial \mathbf{s}}{\partial p_k} \right) \\ &= -2\delta_k \delta_l \mathbf{x}_k^\top \mathbf{H}^{-1} \mathbf{x}_l, \end{aligned} \tag{3.1}$$

et son cas particulier,

$$\frac{\partial \delta_k^2}{\partial p_k} = -2\delta_k^2 \mathbf{x}_k^\top \mathbf{H}^{-1} \mathbf{x}_k < 0,$$

de sorte que

$$\frac{\partial \log\left(\delta_k^2\right)}{\partial p_k} = -2\, \mathbf{x}_k^\top \mathbf{H}^{-1} \mathbf{x}_k.$$

Par conséquent, $|\delta_k|$ est une fonction décroissante de p_k, conformément à la motivation de p_k comme importance relative de réduire $|\delta_k|$. Toutefois, la diminution est plus lente lorsque $|\delta_k|$ est plus petit;

$$\frac{\partial^2 \log\left(\delta_k^2\right)}{\partial p_k^2} = 2\left(\mathbf{x}_k^\top \mathbf{H}^{-1} \mathbf{x}_k\right)^2,$$

par conséquent, δ_k^2 est une fonction logarithmiquement convexe. Selon une interprétation vague, essayer de supprimer un petit écart $|\delta_k|$ nécessite une variation importante de p_k, ce qui pourrait augmenter d'autres écarts $|\delta_l|$.

La solution $\mathbf{u}$ est une fonction linéaire de R n'intervenant que dans $\mathbf{s}$;

$$\frac{\partial \mathbf{u}}{\partial R} = \mathbf{H}^{-1}\left(\mathbf{w} - \frac{t_0}{n}\mathbf{1}_n\right),$$

et

$$\frac{\partial \delta_k}{\partial R} = \mathbf{x}_k^\top \mathbf{H}^{-1}\left(\mathbf{w} - \frac{t_0}{n}\,\mathbf{1}_n\right). \tag{3.2}$$

L'expression de $\mathbf{s}$ dans l'équation (2.1), ainsi que l'intention initiale, suppose que R agit comme facteur de rétrécissement, en réduisant la déviation des poids $\mathbf{u}$ par rapport à $\mathbf{w}$ et $1-R$ comme réduction de leur dispersion. En effet, lorsqu'aucune information auxiliaire n'est disponible et que $\mathbf{u}^\top \mathbf{1}_n$ correspond à $\mathbf{w}^\top \mathbf{1}_n$, $\mathbf{H} = \mathbf{I}_n$ et la solution est la combinaison convexe $\mathbf{w}_R$ présentée à l'équation (2.1). Dans ce cas, nous obtenons une expression simple, $\mathrm{E}(\mathbf{w}_R^\top \mathbf{y}) - \theta = -(1-R)\,\mathrm{cov}(w, y)$, pour le biais de l'estimateur calé lorsque les poids $\mathbf{w}$ sont corrects et que $\mathbf{w}^\top \mathbf{y}$ est sans biais.

Le biais de $\hat{\theta}(\mathbf{u})$ est probablement le plus important lorsqu'une valeur nulle est attribuée à R, lorsque $\mathbf{u}$ ne dépend pas de $\mathbf{w}$. Parallèlement, $\delta_0 = 0$ uniquement lorsque $R = 0$, sauf si $\mathbf{w}^\top \mathbf{1} = t_0$; la preuve se

trouve à l'annexe B. Ainsi, se concentrer sur un seul objectif de calage aux dépens d'autres peut être mal avisé.

4. Exemples

Dans la présente section, nous démontrons la prise en compte des écarts $\boldsymbol{\delta} = (\delta_0, \ldots, \delta_K)^\top$ et de la dispersion des poids calés $\mathbf{u}$ par les priorités $\mathbf{p}$ et le paramètre R. Nous utilisons une population synthétique de taille $N = 120\,000$ au sein de laquelle nous tirons un échantillon de taille attendue $\mathrm{E}(n) \doteq 1\,000$, conformément à un plan de sondage ayant des probabilités inégales et des inclusions indépendantes dans l'échantillon, sans remplacement. Dans la population, nous avons une variable discrète Z comptant six catégories et une variable continue X.

Les six sous-populations désignées par Z ont des tailles allant de 13 000 à 26 000. Les valeurs de la seule variable contextuelle continue X sont générées sous forme d'échantillons aléatoires, un dans chaque catégorie k, de distributions gamma de formes ξ_k et d'un taux commun de 5,0, où $\boldsymbol{\xi} =$ (6,4; 6,7; 6,1; 6,6; 6,9; 6,4), de sorte que les moyennes au sein des catégories de X se situent dans la fourchette 1,22 − 1,28 et leurs écarts-types, dans la fourchette 0,244 – 0,276. La variable dépendante Y est générée par $\frac{1}{2}X + \exp(\zeta)$, où ζ comprend des échantillons aléatoires indépendants tirés de distributions normales de moyennes ξ_k dans la catégorie k et des écarts-types (0,05; 0,07; 0,04; 0,06; 0,08; 0,05). Le scénario complet a été sélectionné de façon arbitraire, afin de générer un ensemble de données au sein duquel les distributions de X et de Y diffèrent selon les catégories de Z et sont biaisées vers la droite, et X et Y sont modérément corrélés globalement (leur corrélation est de 0,30) et au sein des six catégories. Le tableau 4.1 présente les sommaires pertinents de X, de Y et de Z.

Tableau 4.1
Sommaires de population des variables X, Y et Z.

	Catégorie (k) de Z						
	1	2	3	4	5	6	Tous
$t_1 \ldots t_6$	25 000	15 000	22 000	26 000	13 000	19 000	120 000
$\sum X$	31 949	20 053	26 710	34 329	17 890	24 421	155 352
$\bar{X}$ (écart-type)	1,28 (0,51)	1,34 (0,52)	1,21 (0,49)	1,32 (0,51)	1,38 (0,52)	1,29 (0,51)	1,29 (0,51)
$\sum Y$	117 449	92 376	79 480	146 142	96 201	89 336	620 984
$\bar{Y}$ (écart-type)	4,70 (0,32)	6,16 (0,47)	3,61 (0,27)	5,62 (0,39)	7,40 (0,60)	4,70 (0,33)	5,17 (1,18)

4.1 Plan de sondage

Pour tirer un échantillon de cette population, nous envisageons un plan de sondage qui serait planifié et dont les poids de base sont calculés. Dans ce plan de sondage, les probabilités d'inclusion sont constantes au sein des six catégories, qui sont fixées à 0,0075, à 0,0093, à 0,0070, à 0,0090, à 0,0110 et à 0,0075. Les inclusions sont mutuellement indépendantes.

Nous introduisons du réalisme dans la collecte de données de l'enquête simulée en générant les « vraies » probabilités d'inclusion reflétant la mise en œuvre imparfaite de l'enquête. Les probabilités de plan de sondage sont multipliées par un échantillon aléatoire provenant de la distribution log-normale fondée sur $\mathcal{N}(0, \sigma^2)$. Aucune imperfection ne correspond à $\sigma = 0$ et à un σ plus grand, avec des écarts plus importants par rapport aux poids de base menant à une plus grande incidence des imperfections dans la réalisation de l'enquête. En d'autres termes, $\hat{\theta}(\mathbf{w})$ est biaisé, car les poids de base $\mathbf{w}$ ont été déformés lors de la mise en œuvre du plan de sondage, ce qui a mené à un ensemble modifié (perturbé) de poids $\mathbf{w}^{\dagger}$. En appliquant $\hat{\theta}\{C(\mathbf{w})\}$, notre objectif est de résoudre ce problème. En pratique, les poids modifiés $\mathbf{w}^{\dagger}$ ne peuvent pas être récupérés ni estimés; toutefois de l'information (ou une opinion bien fondée) peut exister quant au degré de cette perturbation. Dans notre modèle, cette perturbation est caractérisée par la variance σ^2 des changements des poids logarithmiques ou, de manière équivalente, des probabilités logarithmiques.

Dans la présente section, nous utilisons un échantillon tiré où $\sigma = 0{,}1$. Cela équivaut à une perturbation substantielle; les écarts-types des poids perturbés dans cet échantillon sont de 13,3, de 10,4, de 14,4, de 11,5, de 9,7 et de 12,3 dans les catégories respectives 1, …, 6, alors que les poids de base sont constants au sein des catégories. Le poids de base moyen est $\bar{w} = 122{,}44$. La moyenne des poids perturbés est $\bar{w}^{\dagger} = 123{,}22$.

4.2 Calage

Nous calons un échantillon unique et nous nous intéressons uniquement aux écarts $\boldsymbol{\delta}$. Nous utilisons les valeurs des paramètres d'ajustement

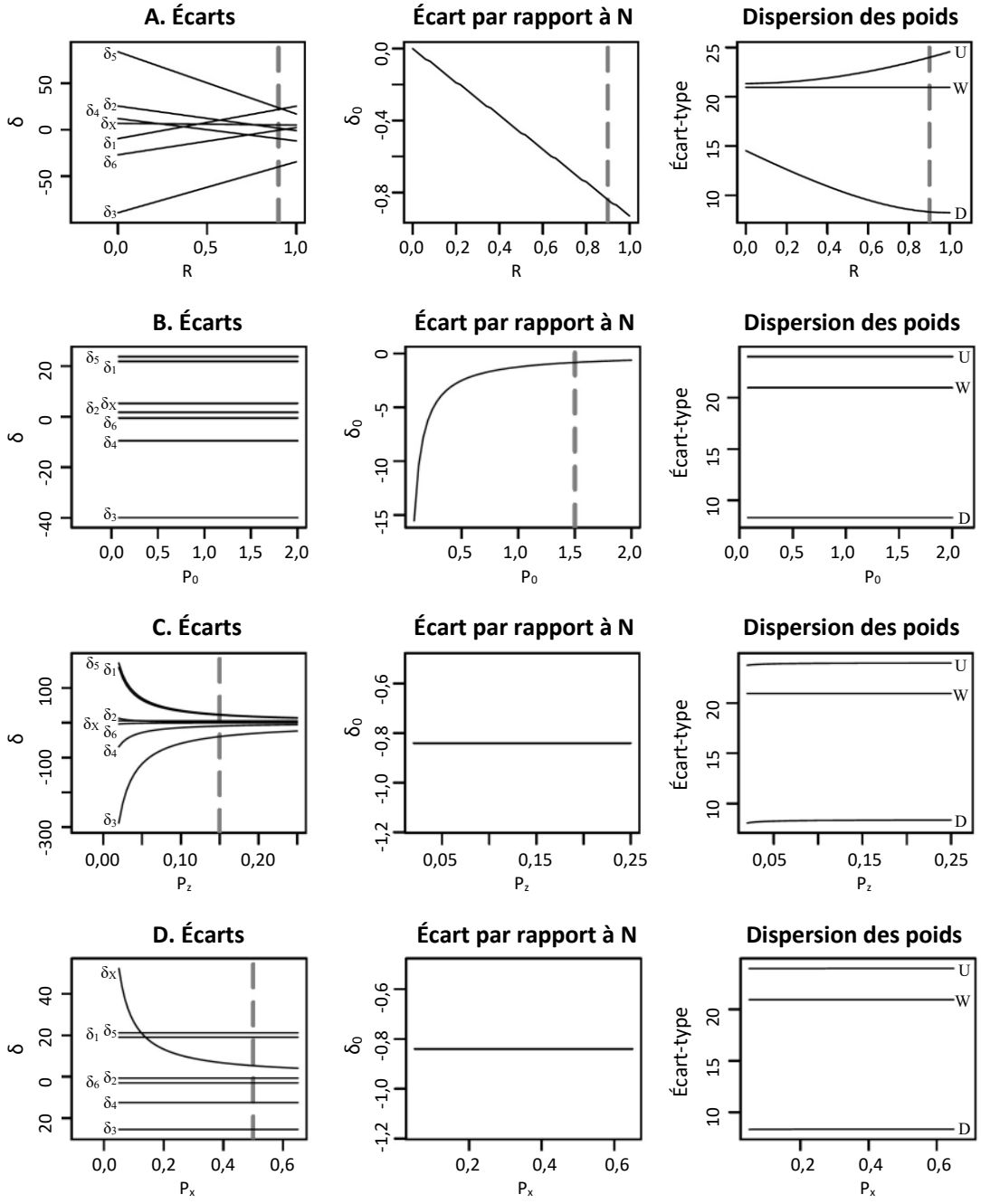
$$p_0 = 1{,}5;\ p_1 = \ldots = p_6 = 0{,}15;\ p_7 = 0{,}5;\ R = 0{,}9; \tag{4.1}$$

et les appelons le scénario de référence. Nous utilisons la notation p_{X} pour p_7 et δ_{X} pour δ_7. Lorsque $p_1 = \ldots = p_6$, nous désignons par p_{Z} la valeur commune.

Nous appliquons le premier calage au moyen de valeurs de R situées dans la fourchette $(0, 1)$ et du scénario de référence pour $\mathbf{p}$. Les écarts, comme fonctions de R, sont représentés graphiquement dans la rangée A en haut de la figure 4.1. Dans la colonne de gauche, les écarts sont représentés graphiquement pour X et les six catégories de Z. Le tracé confirme que ces fonctions $\delta_k(R)$ sont linéaires; voir l'équation (2.4). Les écarts sont davantage dispersés pour $R = 0$ que pour $R = 1$, même si certaines des fonctions $\delta_k(R)$ franchissent le zéro à $R \in (0, 1)$.

La fonction $\delta_0(R)$ est représentée graphiquement dans une colonne séparée, car une échelle bien plus étroite est nécessaire pour l'axe vertical. La correspondance est parfaite, $\delta_0 = 0$, pour $R = 0$, conformément à la preuve de l'annexe B. Les écarts-types de $\mathbf{u}$ et de $\mathbf{u} - \mathbf{w}$, indiqués par les symboles respectifs U et D, sont représentés dans la colonne de droite, tout comme l'écart-type de $\mathbf{w}$ (W), qui, bien sûr, est constant. L'écart-type de $\mathbf{u}$ augmente et l'écart-type de $\mathbf{u} - \mathbf{w}$ diminue avec R. En fonction de ces trois tracés, nous choisirions une grande valeur de R pour réduire les valeurs de δ_k globalement, malgré un petit sacrifice de δ_0. Nous choisissons $R = 0{,}9$, qui est indiqué par les tirets verticaux.

Figure 4.1 Écarts δ et écarts-types des poids en tant que fonctions du paramètre $R \in (0,1)$ à la rangée A (en haut), de $p_0 \in (0,2)$ à la rangée B, de $p_Z \in$ (0,02; 0,25) à la rangée C et de $p_X \in$ (0,05; 0,65) à la rangée D; tous adaptés du scénario de référence fourni (4.1), indiqué sur les tracés par de longs tirets verticaux.



À la rangée B, les écarts sont représentés graphiquement comme des fonctions de p_0 dans la fourchette $(0, 2)$; les autres paramètres étant fournis par le scénario de référence. À l'exception de $\delta_0(p_0)$, chaque écart et chaque écart-type est très proche d'une constante. La fonction $|\delta_0(p_0)|$ diminue rapidement pour de petites valeurs de p_0 et converge vers zéro très lentement pour de grandes valeurs de p_0, comme cela est présenté après l'équation (3.1). Selon ce tracé, nous attribuons la valeur 1,5 à p_0.

À la rangée C, la valeur de $p_Z = p_1 = \ldots = p_6$ varie dans la fourchette (0,02, 0,25) et les autres paramètres sont maintenus à leurs valeurs de référence. Les écarts $\delta_1, \ldots, \delta_6$ se rapprochent rapidement de zéro pour de petites valeurs de p_Z et convergent lentement vers zéro pour de grandes valeurs de p_Z. Les autres fonctions de p_Z, δ_0 et δ_X, ainsi que les deux écarts-types de la colonne de droite, se situent tous dans une fourchette très étroite et leur courbure est uniquement importante pour de très petites valeurs de p_Z. Nous établissons que $p_Z = 0{,}15$. Puisque $|\delta_3|$, qui est égal à 39,9 lorsque $p_Z = 0{,}15$, est plus grand que les écarts pour les cinq autres catégories, nous augmentons p_3 à 0,25. Maintenant, $\delta_3 = -25{,}4$, ce qui est bien plus petit en valeur absolue. Le deuxième écart en importance correspond à la cinquième catégorie, $\delta_5 = 21{,}2$, qui était de 23,7 selon le scénario initial.

Les colonnes de la rangée D (en bas) présentent le scénario comprenant une variation de p_X, alors que les autres paramètres sont maintenus à leurs valeurs de référence, à l'exception de $p_3 = 0{,}25$. Le diagramme confirme que $|\delta_X(p_X)|$ diminue rapidement pour de petites valeurs de p_X et plus lentement lorsqu'il se rapproche de zéro pour de grandes valeurs de p_X. Les autres écarts et les deux écarts-types dépendent très faiblement de p_X. Nous établissons que $p_X = 0{,}5$.

En résumé, la figure 4.1 illustre que nous pouvons réduire les écarts absolus $|\delta_k|$ en accroissant la priorité correspondante p_k. De plus, en modifiant R, nous pouvons échanger une faible dispersion de $\mathbf{u}$ (ce qui est essentiel du point de vue de l'efficacité) contre une petite modification des poids (faible dispersion de $\mathbf{u} - \mathbf{w}$), indirectement liée au biais. La figure 4.1 confirme qu'éliminer un petit écart nécessite une augmentation substantielle de la priorité correspondante. Elle indique qu'une telle augmentation influerait seulement légèrement sur les autres écarts. L'équation (3.1) et l'analyse s'y rapportant laissent croire que cela n'est pas le cas en général, en particulier s'il y a de nombreuses variables dans $\mathbf{X}$, certaines étant hautement corrélées, lorsque plusieurs écarts sont non triviaux.

5. Simulations

Dans la présente section, nous étudions le biais empirique et la racine carrée de l'erreur quadratique moyenne (REQM) de l'estimateur de calage $\hat{\theta}\{C(\mathbf{w}; \mathbf{X}, \mathbf{t})\} = \mathbf{u}^\top \mathbf{y}$ pour plusieurs scénarios ($\mathbf{p}$ et R) et niveaux de mise en œuvre imparfaite du plan de sondage, caractérisé par le paramètre de perturbation σ. Nous évaluons également la contribution à la réduction de la REQM d'une variable utilisée dans le calage. Les simulations que nous décrivons font intervenir des ensembles de 1 000 répliques. Nous avons vérifié que ce nombre était suffisant en comparant les résultats avec leur réexécution en utilisant 2 000 répliques pour certains cas. Nous utilisons la même population ($N = 120\,000$), un plan de sondage planifié où $\mathrm{E}(n) \doteq 1\,000$ et le processus de perturbation des poids de base pour représenter la mise en œuvre imparfaite du plan de sondage, comme cela est décrit à la quatrième section. Le biais et la REQM sont représentés sous forme de fonctions de R, à l'aide de l'interpolation fondée sur l'équation (2.4). Ces fonctions sont évaluées pour $\sigma = 0, 0{,}02, \ldots, 0{,}10$ et quelques vecteurs $\mathbf{p}$.

La figure 5.1 présente, sous forme de lignes continues noires, le biais et la REQM de $\hat{\theta}$ en tant que fonctions de $R \in (0,1)$ pour le scénario de référence de $\mathbf{p}$ et $\sigma = 0; 0{,}02; \ldots, 0{,}1$, comme cela est indiqué dans la marge de droite. Pour chaque valeur de σ, le biais est le plus petit pour $R = 1$. Les fonctions de biais sont parallèles et, pour $\sigma \geq 0{,}02$, ne diffèrent que légèrement. Les fonctions de REQM présentent une

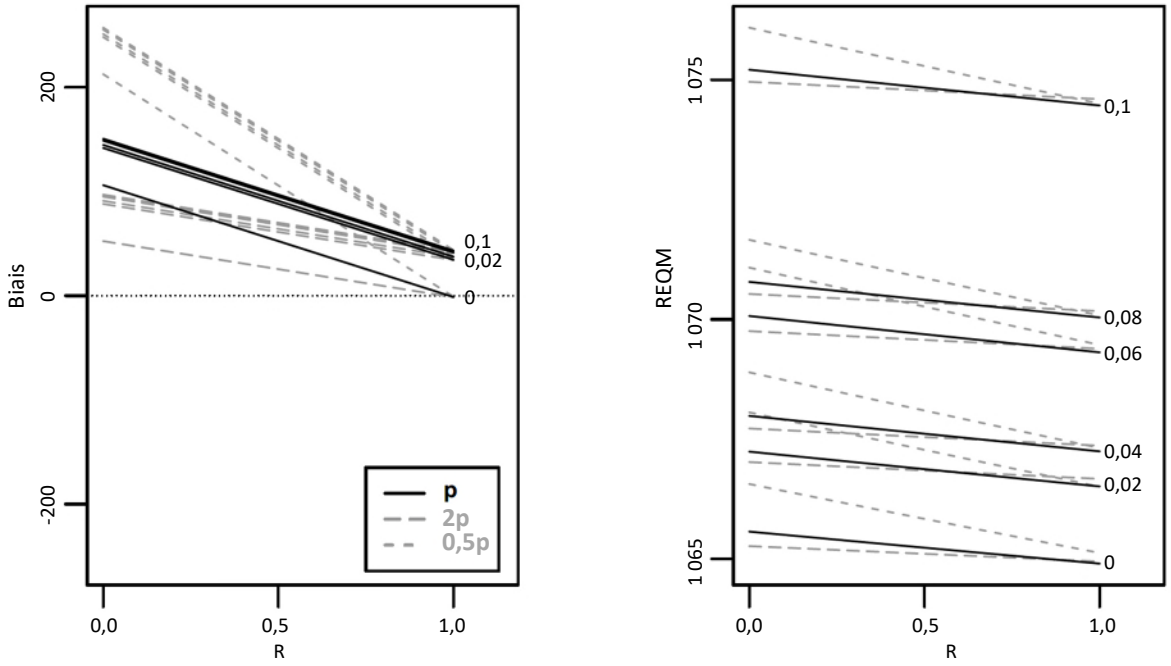
courbure très faible, sont pratiquement parallèles et atteignent également leurs plus faibles valeurs pour $R = 1$. La REQM augmente avec σ pour tous les R, mais inégalement. Les fonctions de biais et de REQM pour les priorités $2\mathbf{p}$ et $\frac{1}{2}\mathbf{p}$ sont, respectivement, représentées par des lignes grises à tirets longs et courts. Toutes les fonctions ont des pentes plus abruptes pour $\frac{1}{2}\mathbf{p}$ et des pentes plus planes pour $2\mathbf{p}$, et diffèrent uniquement légèrement de leurs homologues pour la valeur de référence de $\mathbf{p}$ à $R = 1$. Le tableau 5.1 présente les valeurs de REQM à $R = 0$ et à 1 sous forme tabulaire, ainsi que certains résultats dont il est question ci-dessous.

Tableau 5.1
REQM des estimateurs calés du total de population de Y. La première colonne indique les variables sur lesquelles les probabilités de plan de sondage sont fondées et, entre crochets en dessous, les variables utilisées dans le calage.

Plan de sondage [calage]		σ					
		0	**0,02**	**0,04**	**0,06**	**0,08**	**0,10**
		p (référence)					
	$R = 0$	1 065,6	1 067,2	1 068,0	1 070,1	1 070,8	1 075,2
	$R = 1$	1 064,9	1 066,5	1 067,2	1 069,3	1 070,0	1 074,5
		$2\mathbf{p}$					
	$R = 0$	1 065,3	1 067,0	1 067,7	1 069,8	1 070,5	1 075,0
	$R = 1$	1 064,9	1 066,7	1 067,4	1 069,4	1 070,2	1 074,6
		$\frac{1}{2}\mathbf{p}$					
	$R = 0$	1 066,6	1 068,1	1 068,9	1 071,1	1 071,7	1 076,1
	$R = 1$	1 065,1	1 066,5	1 067,3	1 069,5	1 070,1	1 074,5
Z, X		$p_3 = 0{,}25$					
$[Z, X]$	$R = 0$	1 064,9	1 066,5	1 067,2	1 069,3	1 070,0	1 074,5
	$R = 1$	1 065,6	1 067,2	1 068,0	1 070,1	1 070,8	1 075,2
		$p_0 = 15$					
	$R = 0$	1 064,9	1 066,5	1 067,2	1 069,3	1 070,0	1 074,5
	$R = 1$	1 065,6	1 067,2	1 068,0	1 070,1	1 070,8	1 075,2
		Estimateurs ridge					
	$\hat{\theta}_1$	1 082,5	1 100,4	1 103,7	1 103,1	1 103,1	1 105,4
	$\tilde{\theta}_W$	1 082,3	1 100,6	1 103,8	1 103,1	1 103,3	1 105,5
		p (référence)					
Z, X	$R = 0$	1 431,8	1 440,9	1 443,0	1 445,7	1 449,0	1 449,1
$[Z]$	$R = 1$	1 431,3	1 440,3	1 442,3	1 445,1	1 448,4	1 448,4
		p (référence)					
	$R = 0$	1 397,2	1 403,1	1 403,9	1 405,5	1 408,2	1 408,1
	$R = 1$	1 378,6	1 384,9	1 385,5	1 386,9	1 389,4	1 389,3
		$2\mathbf{p}$					
	$R = 0$	1 396,7	1 402,6	1 403,3	1 404,9	1 407,6	1 407,4
	$R = 1$	1 378,5	1 384,8	1 385,4	1 386,7	1 389,2	1 389,1
Z, X, U		$\frac{1}{2}\mathbf{p}$					
$[Z, X]$	$R = 0$	1 398,5	1 404,7	1 405,5	1 407,2	1 409,9	1 409,9
	$R = 1$	1 379,1	1 385,6	1 386,2	1 387,6	1 390,1	1 390,1
		Estimateurs ridge					
	$\hat{\theta}_1$	1 445,4	1 461,8	1 463,9	1 468,6	1 473,5	1 479,5
	$\tilde{\theta}_W$	1 426,6	1 446,8	1 448,7	1 452,7	1 457,8	1 463,6

La figure 5.1 montre que la REQM dépend de R et du facteur c dans les priorités $c\mathbf{p}$ uniquement légèrement et bien moins qu'elle ne dépend de σ. Pour un R bien plus petit qu'une unité, le biais est une fraction non triviale de la REQM, mais la pente du biais est en grande partie améliorée dans la REQM. Par exemple, si la variance échantillonnale de $\hat{\theta}$ était égale à $1\,065^2$ pour tout $R \in (0,1)$ et que le biais diminuait de 100 à $R = 0$ jusqu'à zéro à $R = 1$, alors la REQM atteindrait des valeurs de 1 069,7 et de 1,065.0 à $R = 0$ et à 1, respectivement, soit une différence de 4,7. Les valeurs empiriques correspondantes, 1 065,6 et 1 064,9, sont uniquement différentes de 0,7.

Figure 5.1 Biais et REQM en tant que fonctions du paramètre R pour le scénario de référence et la perturbation σ = 0; 0,02; …, 0,1, comme cela est indiqué dans la marge de droite; scénarios comprenant les valeurs de référence de $\mathbf{p}$ (lignes continues noires), de $2\mathbf{p}$ (longs tirets gris) et de $\frac{1}{2}\mathbf{p}$ (courts tirets gris). Calage sur $\boldsymbol{Z}$ et $\boldsymbol{X}$.



Nous comparons nos estimateurs de calage à deux autres options fondées sur une régression ridge,

$$\hat{\theta}_1 = \mathbf{t}^\top \left(\mathbf{P}^{-1} + \mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$
$$\tilde{\theta}_W = \mathbf{t}^\top \left(\lambda \mathbf{P}^{-1} + \mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y},$$

où $\mathbf{W}$ est la matrice diagonale dont $\mathbf{w}$ est sur sa diagonale et $\lambda = n/N$. Nous avons choisi $\hat{\theta}_1$, car il s'agit d'un élément de la décomposition de $\hat{\theta}$ dans l'équation (2.3); $\tilde{\theta}_W$ est motivé par Chambers (1996).

Ces estimateurs ridge présentent des biais bien plus grands, se situant dans une fourchette de 300 – 500, mais leurs REQM sont supérieures aux REQM de nos estimateurs de calage par de bien plus faibles marges; les REQM de $\hat{\theta}_1$ et de $\tilde{\theta}_W$ se situent dans la fourchette de 1 082,3 – 1 105,5 pour $\sigma \in (0; 0{,}1)$; des précisions

sont fournies en bas du premier bloc du tableau 5.1. Pour tout scénario donné, les REQM de $\hat{\theta}_1$ et de $\tilde{\theta}_W$ diffèrent de moins de 0,1. L'estimateur pondéré $\mathbf{w}^\top\mathbf{y}$ et l'estimateur de Hájek $N\mathbf{w}^\top\mathbf{y}\,/\mathbf{w}^\top\mathbf{1}$ présentent de faibles biais, mais leurs REQM sont bien plus importantes et dépassent 2 000.

En accroissant p_3 pour le faire passer de 0,15 à 0,25, comme on le suggère dans l'analyse de la figure 4.1, les fonctions de biais et de REQM sont modifiées seulement légèrement. Le biais est réduit de jusqu'à 0,05, mais contrairement aux attentes, la REQM augmente pour toutes les valeurs de σ de 0,1 au plus. Lorsque l'on augmente p_0 pour le faire passer de 1,5 à 15, p_3 étant défini à 0,15, le biais et la REQM sont modifiés de manière imperceptible, soit de moins de 0,01.

En conclusion, la figure 5.1 laisse croire que le calage accroît l'efficacité de $\hat{\theta}$ et que l'estimateur n'est pas très sensible au choix de paramètres d'ajustement, même si les écarts $\boldsymbol{\delta}$ sont sensibles à ce choix. Dans tous les scénarios utilisés, $R=1$ est le choix optimal; autrement dit, le critère de faible variation $\|\mathbf{u}-\mathbf{w}\|$ devrait être ignoré. Il entraîne également un faible biais bien inférieur à ce qu'il est pour $R=0$. Cela ne s'applique pas en général. Un S positif, c'est-à-dire $R<1$, est utile lorsque les poids sont bien plus dispersés et non hautement corrélés avec la variable dépendante. Dans les petits échantillons, la réduction de la variance est relativement plus importante que la réduction du biais; alors, S devrait être défini en une valeur positive plus grande.

La valeur d'une marge t_k peut être quantifiée par calage en omettant la variable correspondante X_k ou en y attribuant une très faible priorité. En retirant la marge t_X du calage, les fonctions de REQM augmentent substantiellement; elles se situent dans la fourchette (1 431,8 – 1 449,1) pour $\sigma\in(0,\ 0{,}1)$ à $R=0$ et sont inférieures de 0,5 à 0,6 à $R=1$. Des précisions sont fournies dans le bloc médian du tableau 5.1.

Le plan de sondage et le scénario de calage examinés jusqu'à présent sont déraisonnablement appropriés, en ce que les poids de base $\mathbf{w}$ dépendent uniquement des variables pour lesquelles les totaux de population dans $\mathbf{t}$ sont disponibles et donc $\hat{\theta}_3=0$ dans l'équation (2.3). Nous générons un scénario plus réaliste en ajoutant une variable U corrélée à la variable dépendante Y, et définissons des probabilités de plan de sondage qui, outre Z, dépendent également de U. En particulier, $\mathbf{U}$ est généré comme un échantillon aléatoire de la distribution log-normale fondée sur $\mathcal{N}$ (1; 0,4), de sorte que sa moyenne soit de 2,94 et son écart-type, de 1,23. De plus, les probabilités de plan de sondage sont établies à

$$\boldsymbol{\pi}_U = \boldsymbol{\pi}_0 + \frac{0{,}004\,(\mathbf{U}-\bar{U})}{\max(\mathbf{U})-\min(\mathbf{U})},$$

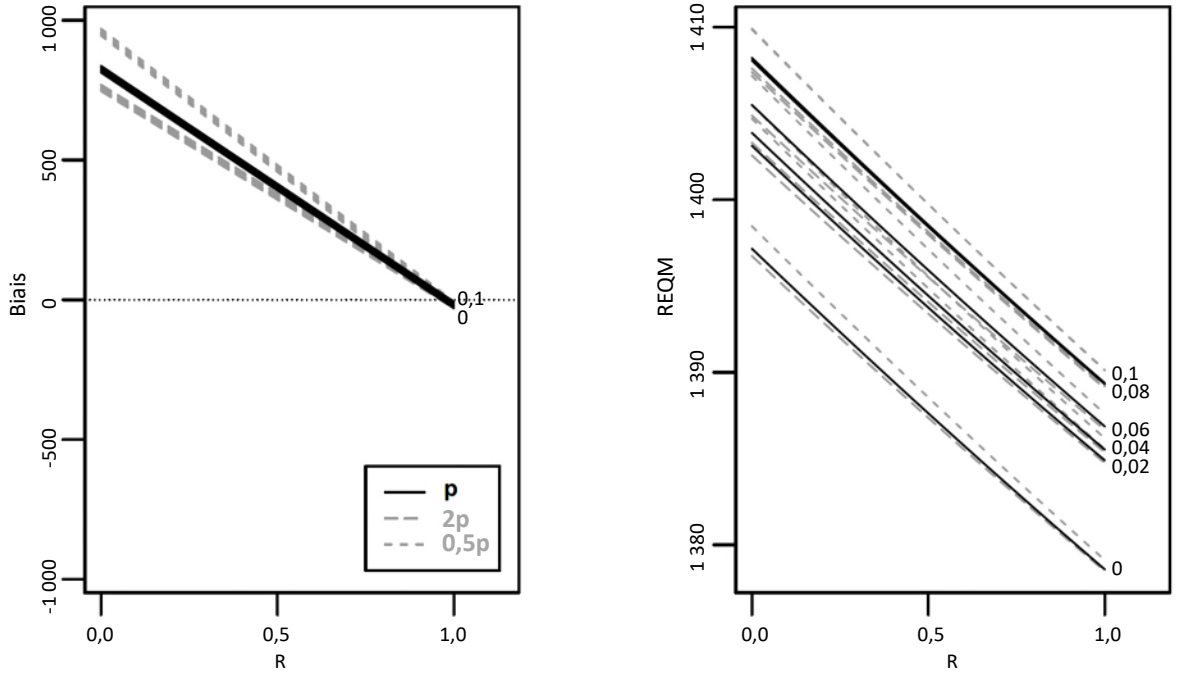
où $\boldsymbol{\pi}_0$ sont les probabilités de plan de sondage utilisées jusqu'à présent. Les probabilités $\boldsymbol{\pi}_U$ sont normalisées pour que le total de population soit égal à 1 000. La « nouvelle » variable dépendante est établie à $Y+\frac{1}{5}U$. La cible correspondante est $(\mathbf{Y}+\frac{1}{5}\mathbf{U})^\top\mathbf{1}_N = 691\ 516{,}8$. Les probabilités de plan de sondage $\boldsymbol{\pi}_U$ mises à l'échelle pour donner un échantillon de taille $n \doteq 1\ 000$ présentent des écarts-types d'environ $2{,}0\times10^{-4}$ au sein des six catégories de Z. (La probabilité moyenne est $n/N = 8{,}33\times10^{-4}$.)

La figure 5.2 présente les résultats à l'aide de la même disposition que la figure 5.1. Elle montre que les fonctions de REQM atteignent des valeurs bien plus grandes que dans le scénario sans variable U; les

fonctions de REQM atteignent leurs minimums pour $R = 1$, mais les pentes de ces fonctions sont bien plus abruptes que dans le scénario initial.

Les estimateurs ridge $\hat{\theta}_{\text{I}}$ et $\tilde{\theta}_{\text{W}}$ sont moins efficaces que $\hat{\theta}$, même si $\tilde{\theta}_{\text{W}}$ est désormais sensiblement plus efficace que $\hat{\theta}_{\text{I}}$; leurs REQM diffèrent d'entre 14 (pour $\sigma = 0{,}1$) et 19 ($\sigma = 0$); voir le bas du dernier bloc du tableau 5.1. Le biais de $\hat{\theta}_{\text{I}}$ se situe dans la fourchette (330, 370), ce qui est comparable au biais de $\hat{\theta}$ à $R = 0{,}6$. Le biais de $\tilde{\theta}_{\text{W}}$ est bien plus petit et se situe dans la fourchette $(-15, 24)$. Il est clair que le fait de ne pas faire intervenir de poids $\mathbf{w}$ est un handicap de $\hat{\theta}_{\text{I}}$ lorsque le modèle de régression $(y \mid X)$ est déficient.

Figure 5.2 Biais et REQM en tant que fonctions du paramètre R pour le scénario de référence et la perturbation $\sigma = 0; 0{,}02; \ldots, 0{,}1$, comme cela est indiqué dans la marge de droite. Scénarios ayant les valeurs de référence de $\mathbf{p}$ (lignes continues noires), de $2\mathbf{p}$ (longs tirets gris) et de $\frac{1}{2}\mathbf{p}$ (courts tirets gris). Calage sur Z, X et U.



Les simulations démontrent que notre estimateur de calage $\hat{\theta}$ est plus efficace que les deux estimateurs ridge $\hat{\theta}_{\text{I}}$ et $\tilde{\theta}_{\text{W}}$, mais que les gains d'efficacité sont relativement modestes. Même si les priorités $\mathbf{p}$ peuvent être définies afin de réduire au minimum les écarts dans $\boldsymbol{\delta}$, un choix attentif n'est pas toujours récompensé par une plus grande efficacité. Toutefois, le biais et l'efficacité sont relativement insensibles au scénario de $\mathbf{p}$. Nous avons obtenu des résultats très similaires en utilisant des plans de sondage comprenant une taille d'échantillon fixe et une stratification sur Z.

Nous avons relevé deux facteurs ayant une grande incidence sur l'efficacité de $\hat{\theta}$: l'imperfection de la mise en œuvre du plan de sondage, influencée dans les simulations par le paramètre σ, et le caractère complet du vecteur $\mathbf{t}$. Cela signifie que l'efficacité est améliorée par une bonne mise en œuvre du plan de

sondage et par un calage sur toutes les variables sur lesquelles le plan de sondage est fondé, ou par l'élaboration de poids de base reposant uniquement sur des variables dont les totaux de population sont connus. Les conclusions fondées sur nos simulations ne justifient pas une généralisation aux scénarios comprenant un vecteur de calage $\mathbf{t}$ bien plus étendu et des plans de sondage complexes.

6. Discussion et conclusion

Nous avons présenté une méthode de calage mettant en correspondance les marges prescrites (tailles et totaux de population et de sous-population) faisant l'objet de petits écarts, tout en tenant compte de la préoccupation d'efficacité. L'impératif de correspondance exacte et la dichotomie de faire correspondre ou d'ignorer chaque marge disponible sont remplacés par un ensemble de coefficients de priorité quantifiant l'importance ou l'urgence de faire correspondre chaque marge, tout en limitant la correction des poids de base et la préférence pour des poids moins dispersés. L'algorithme de calage des poids ne fait intervenir aucune itération ni l'utilisation d'importantes matrices; il peut donc être appliqué de multiples fois, à la recherche du meilleur compromis entre les contraintes concurrentes.

L'estimateur de calage $\hat{\theta}$ est lié à un prédicteur de régression ridge. Nous avons obtenu deux décompositions de cet estimateur. L'une de la somme d'un prédicteur de régression ridge ne faisant pas intervenir les poids d'échantillonnage, d'une correction de biais ayant également la forme d'une prédiction ridge et d'un terme résiduel qui disparaît dans certains scénarios appropriés. L'autre décomposition mène à une interpolation linéaire entre les valeurs de $\hat{\theta}$ pour $R = 0$ et 1; les deux faisant intervenir des expressions simples.

Notre examen analytique et par simulations indique que la réduction d'un écart absolu $|\delta_k|$ est atteinte par une petite modification de la priorité p_k lorsque l'écart est important, mais p_k doit être augmenté de façon substantielle lorsque δ_k est proche de zéro. Cela laisse croire que l'application du calage assoupli après un examen attentif des scénarios des coefficients de priorité peut être bien plus constructive que le choix binaire de faire correspondance exactement la marge disponible ou de l'ignorer. Dans nos expérimentations au moyen de scénarios relativement simples, nous avons montré que les écarts $\boldsymbol{\delta}$ sont faciles à contrôler, mais qu'un contrôle raffiné n'est pas nécessairement récompensé par une plus grande efficacité de l'estimateur de calage. Toutefois, l'efficacité est vraiment insensible au scénario détaillé du vecteur des priorités $\mathbf{p}$ et R.

Nous avons relevé trois facteurs influençant fortement l'efficacité de $\hat{\theta}$. Dans le premier, le degré ou la mesure de l'imperfection de la mise en œuvre du plan de sondage n'est pas une surprise; même si l'objectif du calage est de régler ce problème. Cela peut être le cas dans une certaine mesure, mais le calage n'offre pas une compensation complète. Le deuxième est la disponibilité des totaux de population pour les variables sur lesquelles les poids de sondage $\mathbf{w}$ sont fondés. Le troisième est la variance résiduelle du modèle linéaire liant la variable dépendante y aux variables dans $\mathbf{X}$. Bien sûr, les conclusions doivent être confirmées dans des scénarios et des plans de sondage plus complexes ainsi qu'au moyen de renseignements sur les totaux de population.

Notre méthode de calage est entièrement sans modèle, mais n'exclut pas les adaptations faisant intervenir des modèles. Par exemple, les poids de base $\mathbf{w}$ peuvent être d'abord corrigés par une méthode fondée sur un modèle ou assistée par un modèle, et les poids obtenus peuvent être soumis à un calage assoupli. Les priorités $\mathbf{p}$ (et peut-être également le coefficient R) peuvent être définies par des considérations liées à des modèles, en plus du discernement d'experts présumé initialement.

Les décompositions obtenues font intervenir des coefficients et des prédicteurs de régression ridge fondés sur des modèles implicites. Une avenue à examiner à l'avenir est de savoir si et quand il est avantageux de replacer ces termes à l'aide de considérations fondées sur un modèle. Un autre défi est d'élaborer des façons d'incorporer des propriétés connues et conjecturées de la ou des variables dépendantes dans le calage. Dans des enquêtes menées régulièrement, elles peuvent être fondées sur des renseignements obtenus lors de cycles précédents. Dans des scénarios simples, nous avons relevé que la corrélation des poids et des variables dépendantes était un facteur important. Nous conjecturons que ces corrélations sont également importantes dans des scénarios plus complexes.

L'adaptation de la méthode présentée à des scénarios où l'information auxiliaire est imprécise (West et Little, 2012; Opsomer et Erciulescu, 2021) est un autre défi non résolu. La distance du khi carré des poids utilisés dans l'optimisation peut être remplacée par d'autres options, possiblement au prix d'un algorithme (itératif) plus complexe. Une exception est la distance du khi carré pour les poids logarithmiques (ou une autre transformation des poids), qui peut être motivée par la référence à des corrections multiplicatives des poids.

Notre intention n'était pas de proposer un modèle pour la variable dépendante, mais nous avons obtenu un estimateur associé à un modèle. Cela indique que nous devrions réduire l'accent mis sur la validité du modèle dans une approche fondée sur un modèle et nous concentrer sur l'exploitation de toute l'information disponible. La compensation dans la régression ridge à laquelle l'estimateur de calage est associé peut être considérée de façon similaire et définie par les considérations du compromis entre le biais et la variance. L'engagement envers une approche ou un paradigme est peut-être inférieur au fait de faire de bonnes combinaisons ou de trouver de bons compromis, afin d'exploiter les points forts de chaque approche ou de chaque paradigme et d'améliorer leurs faiblesses.

Tous les calculs décrits dans le présent article sont effectués dans R. Le code créé peut être obtenu auprès de l'auteur sur demande.

Annexe

A. Calage sur une seule variable catégorique

Dans le présent annexe, nous obtenons une expression de l'estimateur $\hat{\theta}$ lors d'un calage effectué uniquement sur la taille de la population t_0 et les tailles des sous-populations t_k des catégories $k = 1, \ldots, K$ d'une variable discrète Z. Nous considérons comme étant négligeable la probabilité qu'une catégorie ne figure pas dans l'échantillon. Le point de départ est l'équation (2.4), selon laquelle nous devons évaluer

$\mathbf{w}^\top \mathbf{X}\hat{\boldsymbol{\beta}}$ et $\mathbf{t}^\top\hat{\boldsymbol{\beta}}$, où $\hat{\boldsymbol{\beta}} = (\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X})^{-1}$ et $\mathbf{X}$ est une matrice de variables auxiliaires, une fois ses colonnes $1, \ldots, K$ normalisées.

La matrice $n \times (K+1)$ initiale des données auxiliaires, désignée par $\mathbf{X}_{\text{or}}$, comprend $\mathbf{1}_n$ dans la colonne 0 et l'indicateur de catégorie k dans la colonne $k = 1, \ldots, K$. Soit $\boldsymbol{\mu} = (0, \mu_1, \ldots, \mu_K)^\top$, le vecteur des proportions d'échantillon des catégories de Z, précédé d'un zéro, $s_0^2 = 1$ et $s_k^2 = \mu_k(1-\mu_k)$ pour $k = 1, \ldots, K$, et $\mathbf{S} = \text{diag}(1, s_1, \ldots, s_K)$. Donc, la normalisation est la transformation $\mathbf{X} = (\mathbf{X}_{\text{or}} - \mathbf{1}_n\boldsymbol{\mu}^\top)\,\mathbf{S}^{-1}$; elle laisse la colonne 0 intacte. Puisque les autres colonnes de $\mathbf{X}_{\text{or}}$ sont des indicateurs orthogonaux par paires, $\hat{\boldsymbol{\beta}} = (\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ est exprimé en fonction de $\mathbf{X}_{\text{or}}$, de $\boldsymbol{\mu}$ et de $\mathbf{S}$ sous la forme

$$\hat{\boldsymbol{\beta}} = \mathbf{S}\left\{\mathbf{S}\mathbf{P}^{-1}\mathbf{S} + n\,\text{diag}\left(\boldsymbol{\mu}^{(1)}\right) - n\boldsymbol{\mu}\boldsymbol{\mu}^\top\right\}^{-1}\left(\mathbf{X}_{\text{or}} - \mathbf{1}_n\boldsymbol{\mu}^\top\right)^\top\mathbf{y},$$

où $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu} + \mathbf{e}_0$; l'élément 0 de $\boldsymbol{\mu}^{(1)}$ est égal à 1 et ses autres éléments sont μ_k. Soit $\boldsymbol{\Pi} = \mathbf{S}\mathbf{P}^{-1}\mathbf{S} + n\,\text{diag}\,(\boldsymbol{\mu}^{(1)})$; ses éléments diagonaux sont $1/p_0 + n$ et $s_k^2/p_k + n\mu_k$, $k = 1, \ldots, K$. L'inverse intervenant dans $\hat{\boldsymbol{\beta}}$ est

$$\left(\boldsymbol{\Pi} - n\boldsymbol{\mu}\boldsymbol{\mu}^\top\right)^{-1} = \boldsymbol{\Pi}^{-1} + \frac{n}{G}\,\boldsymbol{\Pi}^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^\top\,\boldsymbol{\Pi}^{-1}$$

où, puisque $\boldsymbol{\mu}^\top\mathbf{1} = 1$,

$$\begin{aligned} G &= 1 - n\,\boldsymbol{\mu}^\top\,\boldsymbol{\Pi}^{-1}\boldsymbol{\mu} \\ &= \sum_{k=1}^{K}\left(\mu_k - \frac{n\,p_k\mu_k^2}{\mu_k(1-\mu_k) + n\,p_k\mu_k}\right) = \sum_{k=1}^{K} g_k s_k^2, \end{aligned}$$

et $g_k = 1/(1 - \mu_k + n\,p_k)$, $k = 1, \ldots, K$.

Soit $\overline{\mathbf{Y}} = (\bar{y}, \bar{y}_1\mu_1, \ldots, \bar{y}_K\mu_K)^\top$, où $\bar{y}$ est la moyenne d'échantillon de y et $\bar{y}_k$ est la moyenne de sous-échantillon de y dans la catégorie k de Z. Désignons $\overline{\mathbf{W}} = (\bar{w}, \bar{w}_1\mu_1, \ldots, \bar{w}_K\mu_K)^\top$ de façon similaire. Nous obtenons les identités $(\boldsymbol{\Pi} - n\boldsymbol{\mu}\boldsymbol{\mu}^\top)^{-1}\boldsymbol{\mu} = \frac{1}{G}\,\boldsymbol{\Pi}^{-1}\boldsymbol{\mu}$ et $\mathbf{a}^\top\boldsymbol{\Pi}^{-1}\boldsymbol{\mu} = \sum_{k=1}^{K} a_k p_k g_k$ pour tout vecteur $\mathbf{a} = (a_0, a_1, \ldots, a_K)^\top$. Alors,

$$\begin{aligned} \mathbf{w}^\top\mathbf{X}\hat{\boldsymbol{\beta}} &= n\,\mathbf{w}^\top\left(\mathbf{X}_{\text{or}} - \mathbf{1}_n\boldsymbol{\mu}^\top\right)\left(\boldsymbol{\Pi} - n\boldsymbol{\mu}\boldsymbol{\mu}^\top\right)^{-1}\left(\overline{\mathbf{Y}} - \bar{y}\boldsymbol{\mu}\right) \\ &= n^2\left(\overline{\mathbf{W}} - \bar{w}\boldsymbol{\mu}\right)^\top\boldsymbol{\Pi}^{-1}\left(\overline{\mathbf{Y}} - \bar{y}\boldsymbol{\mu}\right) \\ &\quad + \frac{n^3}{G}\left(\overline{\mathbf{W}} - \bar{w}\boldsymbol{\mu}\right)^\top\boldsymbol{\Pi}^{-1}\boldsymbol{\mu} \times \left(\overline{\mathbf{Y}} - \bar{y}\boldsymbol{\mu}\right)^\top\boldsymbol{\Pi}^{-1}\boldsymbol{\mu} \\ &= \frac{n^2 p_0}{1 + n\,p_0}\bar{w}\bar{y} + n^2\sum_{k=1}^{K} g_k p_k \mu_k\left(\bar{w}_k - \bar{w}\right)\left(\bar{y}_k - \bar{y}\right) \\ &\quad + \frac{n^3}{G}\sum_{k=1}^{K} g_k p_k \mu_k\left(\bar{w}_k - \bar{w}\right) \times \sum_{k=1}^{K} g_k p_k \mu_k\left(\bar{y}_k - \bar{y}\right). \end{aligned}$$

(Le signe × est ajouté pour souligner que la multiplication s'applique à deux scalaires.) La même séquence d'opérations donne l'identité

$$\begin{aligned}\mathbf{t}^\top\hat{\boldsymbol{\beta}} &= n\left(\mathbf{t}_{\text{or}} - t_0\boldsymbol{\mu}\right)^\top\left(\boldsymbol{\Pi} - n\boldsymbol{\mu}\boldsymbol{\mu}^\top\right)^{-1}\left(\overline{\mathbf{Y}} - \overline{y}\,\boldsymbol{\mu}\right)\\ &= \frac{n\,p_0}{1+n\,p_0}t_0\,\overline{y} + n\,t_0\sum_{k=1}^{K} g_k p_k\left(\frac{t_k}{t_0} - \mu_k\right)\left(\overline{y}_k - \overline{y}\right)\\ &\quad + \frac{n^2 t_0}{G}\sum_{k=1}^{K} g_k p_k\left(\frac{t_k}{t_0} - \mu_k\right) \times \sum_{k=1}^{K} g_k p_k\left(\overline{y}_k - \overline{y}\right).\end{aligned}$$

Par conséquent, après substitution à l'équation (2.4), $\hat{\theta} = \hat{\theta}^{(1)} + \hat{\theta}^{(Z)}$, où

$$\begin{aligned}\hat{\theta}^{(1)} &= R\left(\mathbf{w}^\top\mathbf{y} - \frac{n\,p_0}{1+n\,p_0}\,n\,\overline{w}\,\overline{y}\right) + \left(1 - \frac{R}{1+n\,p_0}\right)t_0\,\overline{y}\\ \hat{\theta}^{(Z)} &= n\sum_{k=1}^{K} g_k p_k \Lambda_k\left(\overline{y}_k - \overline{y}\right) + \frac{n^2}{G}\sum_{k=1}^{K} g_k p_k \Lambda_k \times \sum_{k=1}^{K} g_k p_k\left(\overline{y}_k - \overline{y}\right)\end{aligned} \tag{6.1}$$

et $\Lambda_k = t_k - n\,R\,\overline{w}_k - \mu_k(t_0 - n\,R\,\overline{w}) = t_k - t_0\mu_k - nR(\overline{w}_k - \overline{w}\mu_k)$. Il convient de noter que $\hat{\theta}^{(1)}$ dépend de $\boldsymbol{\mu}$, de $\mathbf{p}$ et de $\mathbf{t}$ uniquement par l'intermédiaire de p_0 et de t_0, alors que $\hat{\theta}^{(Z)}$ ne dépend pas de p_0. En fait, $\hat{\theta}^{(Z)} = 0$ lors d'un calage effectué uniquement sur la taille de la population, lorsque $p_1 = \ldots = p_K = 0$. Nous avons $\hat{\theta}^{(Z)} = 0$ également lorsque $\Lambda_1 = \ldots = \Lambda_K = 0$. Un exemple d'une telle réduction se présente lorsque les poids sont constants, $\mathbf{w} = \overline{w}\,\mathbf{1}_n$, et que les tailles de sous-échantillon au sein des K catégories sont fixes, $t_k = t_0\,\mu_k$. En général, $\hat{\theta}^{(Z)}$ peut être considéré comme une correction de $\hat{\theta}^{(1)}$ pour les priorités associées à Z.

L'estimateur $\hat{\theta} = \hat{\theta}^{(1)} + \hat{\theta}^{(Z)}$ dépend de $\mathbf{w}$ et de $\mathbf{y}$ uniquement par l'intermédiaire de $\mathbf{w}^\top\mathbf{y}$ et des K paires de moyennes $\overline{w}_k$ et $\overline{y}_k$ au sein des catégories. Le produit $g_k p_k = p_k/(1 - \mu_k + np_k)$ est une fonction croissante de p_k dont les limites respectives sont de 0 à $p_k = 0$ et de $1/n$ lorsque $p_k \to +\infty$. Le dénominateur G est une fonction décroissante de chaque p_k. Il atteint son maximum de 1 pour $p_1 = \ldots = p_K = 0$ et sa limite est zéro lorsque toutes les K priorités p_k divergent vers $+\infty$. Dans le dernier cas, $\hat{\theta}$ devient instable. Toutefois, en pratique, la valeur attribuée à p_0 est supérieure à celle des autres priorités; ce cas présente donc peu de pertinence pratique.

Lorsque p_0 diverge vers $+\infty$, $\hat{\theta}^{(1)}$ converge vers

$$R\,(\mathbf{w}^\top\mathbf{y} - n\,\overline{w}\,\overline{y}) + t_0\,\overline{y} = (n-1)\,R\,\text{cov}\,(w, y) + t_0\,\overline{y}\,.$$

Cela confirme que les poids sont particulièrement importants lorsqu'ils sont fortement corrélés avec la variable dépendante.

B. $R = 0$ sous-entend que $\delta_0 = 0$

Dans la décomposition $\mathbf{w}_R = \mathbf{X}\boldsymbol{\nu}_R + \boldsymbol{\varepsilon}_R$ pour $\mathbf{w}_0$, nous avons $\boldsymbol{\nu}_0 = \mathbf{e}_0$ et $\boldsymbol{\varepsilon}_R = \mathbf{0}$. Ainsi,

$$\mathbf{u} = \mathbf{X}\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{t} + \frac{t_0}{n}\,\mathbf{P}^{-1}\mathbf{e}_0\right).$$

Du fait de la normalisation, $\mathbf{X}^\top \mathbf{1}_n = n\mathbf{e}_0$ et la colonne 0 de $\mathbf{X}^\top\mathbf{X}$ est également égale à $n\mathbf{e}_0$. En outre, $\mathbf{P}^{-1}\mathbf{e}_0 = p_0^{-1}\mathbf{e}_0$; donc,

$$\begin{aligned}\mathbf{u}^\top \mathbf{1} &= \left(\mathbf{t} + \frac{t_0}{np_0}\mathbf{e}_0\right)^\top \left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{1} \\ &= \frac{n}{p_0^{-1}+n}\left(\mathbf{t} + \frac{t_0}{np_0}\,\mathbf{e}_0\right)^\top \mathbf{e}_0 = \mathbf{t}_0.\end{aligned}$$

Par conséquent, $\delta_0 = \mathbf{u}^\top\mathbf{1} - t_0 = 0.$ Nous analysons le moment où la pente de δ_0 (fonction linéaire de R) disparaît. Selon l'équation (3.2), cette pente est

$$\begin{aligned}\frac{\partial\delta_0}{\partial R} &= \mathbf{e}_0^\top\mathbf{X}^\top\left\{\mathbf{I}_n - \mathbf{X}\left(\mathbf{P}^{-1}+\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right\}\left(\mathbf{w} - \frac{t_0}{n}\mathbf{1}_n\right) \\ &= \mathbf{e}_0^\top\left\{\mathbf{I} - \mathbf{X}^\top\mathbf{X}\left(\mathbf{P}^{-1}+\mathbf{X}^\top\mathbf{X}\right)^{-1}\right\}\mathbf{X}^\top\left(\mathbf{w} - \frac{t_0}{n}\mathbf{1}_n\right) \\ &= \mathbf{e}_0^\top\mathbf{P}^{-1}\left(\mathbf{P}^{-1}+\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{X}^\top\mathbf{w} - \frac{t_0}{n}\mathbf{X}^\top\mathbf{1}_n\right) \\ &= \frac{1}{p_0}\,\frac{1}{p_0^{-1}+n}\,\mathbf{e}_0^\top\left(\mathbf{X}^\top\mathbf{w} - t_0\mathbf{e}_0\right) \\ &= \frac{1}{1+np_0}\left(\mathbf{w}^\top\mathbf{1} - t_0\right).\end{aligned}$$

Ainsi, $\delta_0 = 0$ lorsque $R = 0$ ou $\mathbf{w}^\top\mathbf{1} = t_0,$ comme cela est indiqué à la fin de la troisième section. Il convient de noter que dans la plupart des enquêtes dont le plan de sondage est complexe et la taille de l'échantillon est non triviale (aléatoire), la probabilité que $\mathbf{w}^\top\mathbf{1} = t_0$ est très faible sinon nulle.

Bibliographie

Beaumont, J.-F., et Bocci, C. (2008). Another look at ridge calibration. *Metron*, 66, 5-20.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29, 323-353.

Cardot, H., Goga, C. et Shehzad, M.-A. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Statistica Sinica*, 27, 243-260.

Chambers, R. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

Chen, J., Sitter, R.R. et Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89, 230-237.

Dagdoug, M., Goga, C. et Haziza, D. (2023). Model-assisted estimation in high-dimensional settings for survey data. *Journal of Applied Statistics*, 50, 761-785.

Davies, G. (2018). *Examination of Approaches to Calibration in Survey Sampling.* Thèse de doctorat, Cardiff University, Royaume-Uni.

Devaud, D., et Tillé, Y. (2019). Deville and Särndal's calibration: Revisiting a 25-years-old successful optimization problem. *Test*, 28, 1033-1065.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of American Statistical Association*, 87, 1013-1020.

Estevao, V.M., et Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *Revue Internationale de Statistique*, 74, 127-147.

Goga, C., et Shehzad, M.A. (2010). Overview of ridge estimators in survey sampling. Université de Bourgogne, Dijon, France.

Guggemos, F., et Tillé, Y. (2010). Penalized calibration in survey sampleing: Design based estimation assisted by mixed model. *Journal of Statistical Planning and Inference*, 140, 3199-3212.

Haziza, D., et Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32, 206-226.

Hoerl, A.E., et Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67.

Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47, 663-685.

Kim, J.K., et Park, M. (2010). Calibration estimation in survey sampling. *Revue Internationale de Statistique*, 78, 21-39.

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

Longford, N.T. (2024). Statistical balancing as an unconstrained optimisation problem. Soumis.

Lundström, S., et Särndal, C.-E. (1999). Calibration as a standard method for the treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.

Opsomer, J.D., et Erciulescu, A.L. (2021). Estimation de la variance par répliques après calage fondé sur l'échantillon. *Techniques d'enquête*, 47, 2, 287-301. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021002/article/00006-eng.pdf.

Rao, J.N.K., et Singh, A.C. (2009). Range restricted weight calibration for survey data using ridge regression. *Pakistan Journal of Statistics*, 25, 371-383.

Särndal, C.-E. (2007). La méthode de calage dans la théorie et la pratique des enquêtes. *Techniques d'enquête*, 33, 2, 113-135. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2007002/article/10488-fra.pdf.

Tillé, Y. (2020). *Sampling and Estimation from Finite Populations.* New York: John Wiley & Sons, Inc.

Vera, J.F., Sánchez Zuleta, C.C. et Rueda, M. (2023). A unified approach based on multidimensional scaling for calibration estimation in surveys with qualitative auxiliary information. *Statistical Methods in Medical Research*, 32, 760-772.

West, B.T., et Little, R.J.A. (2012). Non-response adjustment of survey estimates based on auxiliary variables subject to error. *Journal of the Royal Statistical Society*, Series A, 176, 211-225.

Wu, C., et Lu, W.W. (2016). Calibration weighting methods for complex surveys. *Revue Internationale de Statistique*, 84, 21-39.

Techniques d'enquête, décembre 2024
Vol. 50, nº 2, pp. 313-329
Statistique Canada, nº 12-001-X au catalogue

Distribution *a priori* gamma hiérarchique pour la modélisation des effets aléatoires dans l'estimation sur petits domaines

Xueying Tang et Liangliang Zhang[1]

Résumé

L'estimation sur petits domaines est de plus en plus populaire auprès des statisticiens d'enquêtes. Étant donné que les estimations directes sur petits domaines comportent habituellement des erreurs-types élevées, des approches fondées sur des modèles sont souvent adoptées pour emprunter de l'information entre domaines. Les modèles d'estimation sur petits domaines reposent souvent sur des covariables pour coupler différents domaines et des effets aléatoires pour tenir compte de la variation supplémentaire. Des études récentes ont montré que les effets aléatoires ne sont pas nécessaires pour tous les domaines, de sorte que des distributions *a priori* en rétrécissement global-local (GL) ont été introduites pour modéliser efficacement la parcimonie des effets aléatoires. Le comportement des distributions *a priori* GL relatif à la queue varie, et leur rendement diffère selon les différents niveaux de parcimonie des effets aléatoires. Donc, il faut adapter le modèle à différents choix de distributions *a priori*, puis choisir celle qui convient le mieux en fonction du critère d'information de déviance ou d'autres paramètres d'évaluation. Dans le présent document, nous proposons une distribution *a priori* souple pour la modélisation des effets aléatoires dans l'estimation sur petits domaines. Les hyperparamètres de la distribution *a priori* déterminent le comportement de la queue et peuvent être estimés dans un cadre entièrement bayésien. Par conséquent, le modèle qui en résulte est adapté à la parcimonie des effets aléatoires sans ajustement répétitif. Nous démontrons le rendement de la distribution *a priori* proposée par des simulations et des applications réelles.

Mots-clés : Distributions *a priori* gamma normales; distributions *a priori* globales-locales; modèle de Fay-Herriot; Rétrécissement adaptatif.

1. Introduction

L'estimation sur petits domaines vise à produire des estimations fiables de statistiques cruciales à un niveau géographique plus précis ou pour une petite sous-population. Les résultats fournissent souvent des renseignements importants pour l'élaboration des politiques publiques et l'affectation des ressources. Un exemple d'estimation sur petits domaines est le programme Small Area Income and Poverty Estimation [estimations sur petits domaines du revenu et de la pauvreté] mené par le Bureau du recensement des États-Unis. L'objectif du programme est de fournir des estimations liées au revenu et à la pauvreté à divers échelons, y compris les comtés et les districts scolaires, à partir des données recueillies dans le cadre de l'American Community Survey. Les petits domaines et les petits sous-groupes sont souvent associés à de petites tailles d'échantillon dans une enquête, ce qui fait que les estimations directes comportent des erreurs-types et des coefficients de variation élevés. Par conséquent, les approches fondées sur des modèles sont souvent utilisées pour produire des estimations avec une précision souhaitable en empruntant de l'information entre petits domaines. Les modèles d'estimation sur petits domaines sont souvent classés en modèles au niveau de l'unité et en modèles au niveau du domaine. Le présent article porte sur cette dernière

1. Xueying Tang, Department of Mathematics, University of Arizona, 617 N. Santa Rita Avenue, Tucson, AZ 85721, États-Unis. Courriel : xytang@arizona.edu; Liangliang Zhang, Department of Population and Quantitative Health Sciences, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, États-Unis. Courriel : lxz716@case.edu.

catégorie et nous renvoyons les lecteurs à Pfeffermann (2013), à Rao et Molina (2015) et à Sugasawa et Kubokawa (2020) pour consulter des analyses détaillées des modèles d'estimation sur petits domaines.

Dans les modèles au niveau du domaine, l'estimation directe de chaque petit domaine est souvent rendue par la somme de la moyenne des petits domaines et de l'erreur d'échantillonnage. On suppose souvent que les erreurs d'échantillonnage sont des variables aléatoires normales indépendantes ayant une moyenne nulle et des variances connues. Les moyennes de petits domaines sont ensuite décomposées en effets fixes et en effets aléatoires propres à un domaine. La partie à effet fixe repose sur l'information auxiliaire provenant des dossiers administratifs et du Recensement de la population comme covariables pour coupler différents petits domaines, tandis que les effets aléatoires caractérisent la variation des moyennes de petits domaines qui n'est pas prise en compte par les effets fixes. Le modèle au niveau des domaines le plus célèbre dans la littérature sur l'estimation sur petits domaines est le modèle de Fay-Herriot (FH) [Fay et Herriot, 1979], où les effets aléatoires sont présumés être des distributions normales indépendantes ayant une moyenne nulle et une variance inconnue commune. Ces hypothèses sont pratiques pour l'analyse théorique et l'application en pratique, ce qui fait du modèle de FH l'un des modèles les plus populaires pour l'estimation sur petits domaines.

Malgré leur caractère pratique, les hypothèses du modèle de FH ont été remises en question, car elles ne sont pas souvent respectées dans la pratique. Diverses extensions ont été réalisées pour assouplir les hypothèses et améliorer davantage le rendement du modèle. Parmi les exemples de ces travaux, mentionnons Datta et Lahiri (1995), Li et Lahiri (2007), Ybarra et Lohr (2008), Fabrizi et Trivisano (2010) et Porter, Wikle et Holan (2015). Une question qui a été soulevée récemment est de savoir si l'inclusion d'effets aléatoires pour tous les domaines est nécessaire. L'exploration de la question commence par Datta, Hall et Mandal (2011), qui ont conçu une procédure de vérification des hypothèses pour déterminer si l'élimination des effets aléatoires est appropriée. L'hypothèse nulle est que la variance des effets aléatoires est égale à zéro. La vérification est fondée sur la qualité de l'ajustement du modèle à effet fixe et fonctionne bien s'il existe un petit nombre ou un nombre modéré de petits domaines. Cependant, lorsque le nombre de petits domaines est élevé, l'hypothèse nulle est souvent rejetée en raison de l'écart important entre l'effet fixe et l'estimation directe dans quelques domaines. Sur la base de ces observations, Datta et Mandal (2015) ont proposé d'utiliser des distributions *a priori* de type *spike-and-slab* pour modéliser les effets aléatoires. Dans ce modèle, on suppose que la distribution des effets aléatoires est un mélange d'une masse ponctuelle à zéro (la composante *spike*) et d'une distribution normale à moyenne nulle (la composante *slab*). La composante *spike* permet d'éliminer les effets aléatoires dans les domaines où cela est approprié, et la composante *slab* caractérise les effets aléatoires non nuls. Cette idée est approfondie davantage par Chakraborty, Datta et Mandal (2016), qui utilisent un mélange de deux distributions normales ayant des variances différentes pour modéliser les effets aléatoires.

Plus récemment, Tang, Ghosh, Ha et Sedransk (2018) ont proposé d'utiliser des distributions *a priori* en rétrécissement global-local (GL) pour décrire les effets aléatoires à l'aide de différentes structures de parcimonie. Le modèle suppose toujours que les effets aléatoires suivent des distributions normales

indépendantes à moyenne nulle, mais les variances sont propres au domaine. Chaque variance correspond au produit d'un paramètre local propre à un domaine et d'un paramètre global partagé entre de petits domaines. Un petit paramètre global tend à réduire toutes les estimations directes à l'estimateur synthétique pour tenir compte des effets aléatoires faibles ou proches de zéro, tandis qu'un grand paramètre local compense le rétrécissement pour les domaines qui ont besoin d'un effet aléatoire important. Les choix possibles de distributions *a priori* pour les paramètres locaux comprennent une large gamme de distributions à queue lourde telles que les distributions *a priori* de Laplace (Park et Casella, 2008), les distributions *a priori* de type *horseshoe* (Carvalho, Polson et Scott, 2009) et les distributions *a priori* bêta à trois paramètres (Armagan, Clyde et Dunson, 2011). La flexibilité des distributions *a priori* locales permet au modèle GL de caractériser les effets aléatoires dans divers contextes. Tang et coll. (2018) ont montré que le rendement du modèle GL est souvent meilleur que celui du modèle de type *spike-and-slab*.

Le rendement exceptionnel du modèle GL repose sur le choix approprié des distributions *a priori* pour les paramètres locaux. Les distributions *a priori* sont souvent classées en distributions *a priori* à queue exponentielle et en distributions *a priori* à queue polynomiale dans l'analyse théorique. Il a été démontré que les distributions *a priori* à queue polynomiale conviennent mieux lorsque les effets aléatoires sont minimes dans la majorité des domaines et que les distributions *a priori* à queue exponentielle sont plus appropriées dans le cas contraire. L'utilisation d'une distribution *a priori* inappropriée peut entraîner des résultats indésirables, tels qu'une faible exactitude de l'estimation ou un faible taux de couverture des intervalles de crédibilité. Étant donné que la structure sous-jacente des effets aléatoires est inconnue pour un ensemble de données particulier, les méthodes de sélection des distributions *a priori* basées sur les données sont cruciales pour l'application du modèle GL dans la pratique. Tang et coll. (2018) ont utilisé le critère d'information de déviance (CID) [Spiegelhalter, Best, Carlin et Van Der Linde, 2002] à cette fin. Cette méthode permet souvent de sélectionner une distribution *a priori* raisonnable. Cependant, le calcul des CID et d'autres critères de sélection des modèles nécessite plusieurs ajustements des modèles, chaque fois en ayant un choix différent de distributions *a priori* locales. Cela nécessite des ressources informatiques considérables, en particulier lorsque le nombre de petits domaines est élevé. En outre, la quantification de l'incertitude basée uniquement sur le modèle sélectionné peut sous-estimer la variation des estimations, car elle ne prend pas en compte la variation générée par les différents modèles.

Dans le présent article, nous proposons un nouveau modèle pour les effets aléatoires du modèle au niveau du domaine. Ce modèle s'adapte à divers niveaux de parcimonie et à diverses structures d'effets aléatoires, comme le modèle GL, mais ne nécessite pas d'ajustements répétés pour la sélection de la distribution *a priori*. Comme pour le modèle GL, nous supposons que les effets aléatoires sont distribués normalement au moyen de variances propres au domaine. Une distribution *a priori* gamma est ensuite appliquée aux variances. En ayant des choix différents de paramètres de forme et de taux, la distribution *a priori* peut présenter une queue exponentielle ou (presque) polynomiale et s'adapter à différents niveaux et différentes structures de parcimonie des effets aléatoires d'une manière semblable au modèle GL. Étant donné que le comportement de la queue de la distribution *a priori* est indexé par les hyperparamètres, le problème de

sélection de la distribution *a priori* la plus appropriée devient un problème d'estimation des hyperparamètres. Nous appliquons en outre des hyperdistributions *a priori* aux paramètres de forme et de taux de la distribution gamma afin d'estimer les hyperparamètres dans un cadre entièrement bayésien. De cette manière, il est possible d'éviter l'ajustement de modèles au moyen de distributions *a priori* différentes, et la variation qu'apportent les différents modèles peut également être prise en compte.

Le reste de l'article est structuré de la manière suivante. À la section 2, nous décrivons le modèle gamma hiérarchique permettant d'obtenir un rétrécissement adaptatif dans l'estimation sur petits domaines et son lien avec certains modèles existants. La section 3 décrit un algorithme de Monte Carlo par chaîne de Markov permettant de tirer des échantillons *a posteriori*. À la section 4, le rendement du modèle proposé est démontré par des études par simulation, et des applications à deux ensembles de données réels sont présentées à la section 5. Nous terminons par la conclusion à la section 6.

2. Modèle

Supposons que $y_1,\ldots,y_n$ désigne les estimations directes des moyennes de petits domaines $\theta_1,\ldots,\theta_m$ de m petits domaines. Nous supposons que

$$y_i = \theta_i + \varepsilon_i, \text{ et } \theta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + u_i, \text{ pour } i = 1,\ldots,m, \tag{2.1}$$

où $\varepsilon_1,\ldots,\varepsilon_m$ sont des erreurs d'échantillonnage indépendantes selon que $\varepsilon_i \sim N(0,D_i)$ et que D_i est une variance de l'erreur connue, $\mathbf{x}_i$ est un vecteur de dimension p de variables auxiliaires, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^\top$ est le vecteur de coefficient correspondant et $u_1,\ldots,u_m$ sont des effets aléatoires indépendants propres au domaine. Les effets aléatoires caractérisent la variation dans θ_i qui ne peut être expliquée par les variables auxiliaires. Tout au long du document, nous utilisons la notation concise suivante pour désigner les composantes du modèle : $\mathbf{y} = (y_1,\ldots,y_m)^\top$, $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_m)^\top$, $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_m)^\top$, $\mathbf{u} = (u_1,\ldots,u_m)^\top$ et $\mathbf{D} = \text{diag}\{D_1,\ldots,D_m\}$. La notation $N(\mu,\sigma^2)$ représente la distribution normale où μ est la moyenne et σ^2 est la variance.

Nous supposons que les effets aléatoires suivent ce qui suit :

$$\begin{aligned} u_i \mid \sigma_i^2 &\sim \text{N}(0,\sigma_i^2) \\ \sigma_i^2 \mid a,b &\sim \text{Gamma}(a,b) \end{aligned} \tag{2.2}$$

où $0 < a < 1$ et $b > 0$ sont deux hyperparamètres et $\text{Gamma}(a,b)$ désigne la distribution gamma selon le paramètre de forme a et le paramètre de taux b. La fonction de densité de probabilité de $\text{Gamma}(a,b)$ est $\pi(x \mid a,b) = (b^a / \Gamma(a))\, x^{a-1} \exp(-bx)$. Contrairement au modèle classique de FH, où les effets aléatoires suivent des distributions normales indépendantes ayant une variance commune σ^2 comme

$$u_i \mid \sigma^2 \sim N(0,\sigma^2), \tag{2.3}$$

nous attribuons des variances distinctes pour les effets aléatoires de différents domaines et nous appliquons une distribution gamma aux paramètres de variance. De façon marginale, la structure de mélange d'échelles de (2.2) suppose une distribution à queue lourde sur u_i. De plus, nous limitons

l'hyperparamètre a à l'intervalle $(0,1)$ de sorte que la distribution marginale de u_i ait une masse de probabilité importante autour de zéro. Ces caractéristiques de u_i permettent au modèle de saisir la variation élevée des effets aléatoires qui se produit habituellement lorsque le nombre de petits domaines est élevé.

La structure de mélange d'échelles est également utilisée dans le modèle GL proposé par Tang et coll. (2018). Ainsi

$$\begin{aligned} u_i \mid \lambda_i^2, \tau^2 &\sim N(0, \lambda_i^2 \tau^2), \\ \lambda_i^2 \sim \pi_L(\lambda_i^2), \quad \tau^2 &\sim \pi_G(\tau^2), \end{aligned} \tag{2.4}$$

où λ_i^2 et τ^2 correspondent aux paramètres local et global, respectivement, et π_L et π_G désignent leurs distributions *a priori* respectives. Notre modèle est semblable au modèle GL, en ce sens que les deux modèles supposent des variances d'effets aléatoires propres à un domaine et appliquent des hyperdistributions *a priori* aux paramètres de variance. Bien que nous n'incluions pas explicitement un paramètre global dans (2.4), $1/b$ est un paramètre d'échelle de λ_i et donc u_i joue le rôle du paramètre global. Notre modèle peut être réécrit en tant que modèle GL où $\lambda_i^2 \sim \text{Gamma}(a,1)$, ce qui est l'un des choix de π_L envisagés dans Tang et coll. (2018). Cependant, a est traité comme un hyperparamètre à estimer dans notre modèle alors qu'il s'agit d'une constante connue dans le modèle GL. Dans notre modèle, la lourdeur de la queue de $\pi(\sigma_i^2)$ varie en fonction des valeurs de a et de b. Si a est proche de 1, la composante exponentielle $\exp(-b\sigma_i^2)$ de la densité gamma domine. Si b est proche de zéro, alors le terme polynomial $(\sigma_i^2)^{a-1}$ domine. Dans le modèle GL, les auteurs de Tang et coll. (2018) ont divisé les choix de π_L entre les distributions *a priori* à queue polynomiale et les distributions *a priori* à queue exponentielle. Ils ont montré que les deux groupes de distributions *a priori* avaient leurs propres scénarios offrant un meilleur rendement en ce qui concerne l'estimation des moyennes de petits domaines. Les distributions *a priori* à queue polynomiale donnent de meilleurs résultats lorsque seulement quelques domaines ont besoin d'effets aléatoires, tandis que les distributions *a priori* à queue exponentielle conviennent mieux lorsque plus de domaines ont besoin d'effets aléatoires. En envisageant une distribution *a priori* gamma sur σ_i^2 et que a et b varient, notre modèle unifie les distributions *a priori* à queue polynomiales et les distributions *a priori* à queue exponentielle pour λ_i^2 dans le modèle GL. Le problème lié au choix de π_L dans le modèle GL devient le problème lié à l'estimation de a et de b dans le modèle proposé.

Pour estimer a et b, nous envisageons un cadre entièrement bayésien et appliquons les hyperdistributions *a priori* sur a et b. Plus précisément, nous considérons

$$a \sim \text{Uniforme}(0,1), \quad b \sim \text{Gamma}(s_0, t_0), \tag{2.5}$$

où s_0 et t_0 sont fixés à de petites valeurs positives. Puisque $1/b$ est analogue au paramètre global, la distribution *a priori* gamma sur b est semblable à une distribution *a priori* gamma inverse faiblement informative sur τ^2 dans le modèle GL. Bien que d'autres choix pour les hyperdistributions *a priori* soient possibles, nous choisissons la distribution *a priori* gamma pour des raisons de commodité, car il s'agit d'une distribution conjuguée conditionnelle. Conformément à Tang et coll. (2018), nous établissons $s_0 = t_0 = 10^{-10}$. En raison de la structure hiérarchique de (2.2) et de (2.5), nous appelons notre modèle le modèle

gamma hiérarchique (GH). Cette distribution *a priori* est étroitement liée à la distribution *a priori* gamma normale dans le contexte de la sélection de variables (Griffin et Brown, 2005, 2010).

La spécification du modèle s'achève par une distribution *a priori* sur $\boldsymbol{\beta}$. Conformément à la littérature sur l'estimation bayésienne sur petits domaines, nous envisageons une distribution *a priori* uniforme

$$\pi(\boldsymbol{\beta}) \propto 1. \tag{2.6}$$

Bien que cette distribution *a priori* soit inappropriée, on peut facilement démontrer que la distribution *a posteriori* résultante est appropriée dans des conditions de régularité mineures.

Théorème 1. *La distribution a posteriori du modèle précisé par (2.1), (2.2), (2.5) et (2.6) est appropriée si* $\text{rang}(\mathbf{X}) = p < m$.

La démonstration du théorème figure dans l'annexe.

3. Calcul

Dans le modèle GH proposé dans (2.1), (2.2), (2.5) et (2.6), la densité *a posteriori* de $\boldsymbol{\beta}$, $\mathbf{u}$, $\boldsymbol{\sigma}^2 = (\sigma_1^2,\ldots,\sigma_m^2)^\top$, a et b est

$$\begin{aligned}\pi(\boldsymbol{\beta},\mathbf{u},\boldsymbol{\sigma}^2,a,b\,|\,\mathbf{y}) \quad &\propto \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{u}-\mathbf{X}\boldsymbol{\beta})^\top \mathbf{D}^{-1}(\mathbf{y}-\mathbf{u}-\mathbf{X}\boldsymbol{\beta})\right] \\ &\quad \times \prod_{i=1}^{m}\left\{(\sigma_i^2)^{-1/2}\exp\left(-\frac{u_i^2}{2\sigma_i^2}\right)\frac{b^a}{\Gamma(a)}(\sigma_i^2)^{a-1}\exp(-b\sigma_i^2)\right\} \\ &\quad \times b^{s_0-1}\exp(-t_0 b).\end{aligned} \tag{3.1}$$

Nous utilisons l'échantillonnage de Gibbs (Gelfand et Smith, 1990) pour tirer des échantillons de la distribution *a posteriori*. À cette fin, il est facile de trouver les distributions conditionnelles complètes comme suit :

- $\boldsymbol{\beta}\,|\,\mathbf{y},\mathbf{u},\boldsymbol{\sigma}^2,a,b \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}},\boldsymbol{\Sigma}_{\boldsymbol{\beta}})$, où $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\mathbf{X}^\top\mathbf{D}^{-1}\mathbf{w}$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{D}^{-1}\mathbf{X})^{-1}$ et $\mathbf{w} = \mathbf{y}-\mathbf{u}$;
- $\mathbf{u}\,|\,\mathbf{y},\boldsymbol{\beta},\boldsymbol{\sigma}^2,a,b \sim N(\boldsymbol{\mu}_{\mathbf{u}},\boldsymbol{\Sigma}_{\mathbf{u}})$, où $\boldsymbol{\mu}_{\mathbf{u}} = (\mathbf{I}-\mathbf{B})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$, $\boldsymbol{\Sigma}_{\mathbf{u}} = (\mathbf{I}-\mathbf{B})\mathbf{D}$, $\mathbf{I}_m$ est la matrice d'identité dimensionnelle m et $\mathbf{B} = \text{diag}(B_1,\ldots,B_m)$ selon $B_i = D_i/(D_i+\sigma_i^2)$;
- $\pi(\boldsymbol{\sigma}^2\,|\,\mathbf{y},\boldsymbol{\beta},\mathbf{u},a,b) \propto \prod_{i=1}^{m}(\sigma_i^2)^{a-3/2}\exp\left[-(u_i^2/2\sigma_i^2)-b\sigma_i^2\right]$;
- $b\,|\,\mathbf{y},\boldsymbol{\beta},\mathbf{u},\boldsymbol{\sigma}^2,a \sim \text{Gamma}\,(am+s_0,\sum_{i=1}^{m}\sigma_i^2+t_0)$;
- $\pi(a\,|\,\mathbf{y},\boldsymbol{\beta},\mathbf{u},\boldsymbol{\sigma}^2,a) \propto \frac{b^{ma}}{\Gamma(a)^m}\left(\prod_{i=1}^{m}\sigma_i^2\right)^{a-1}$.

Par conséquent, il est simple d'échantillonner $\boldsymbol{\beta}$, $\mathbf{u}$ et b à partir de leur distribution conditionnelle complète respective. Pour $\boldsymbol{\sigma}^2$, ses éléments peuvent être échantillonnés indépendamment. Constatant que la fonction de densité de probabilité d'une distribution inverse gaussienne généralisée (représentée par $\text{IGG}(\eta,\chi,\psi)$) est $f(x\,|\,\eta,\chi,\psi) \propto x^{\eta-1}\exp[-\chi/(2x)-\psi x/2]$, nous pouvons échantillonner σ_i^2 à partir de la distribution

IGG $(a-1/2, u_i^2, 2b)$. La distribution conditionnelle complète de a n'est pas une distribution commune. Nous utilisons une étape d'échantillonnage par tranche (Neal, 2003) pour tirer des échantillons de a dans l'échantillonneur de Gibbs.

4. Simulations

4.1 Scénarios

Dans la présente section, nous examinons le rendement du modèle proposé pour les ensembles de données simulés. Les ensembles de données sont générés à partir du modèle (2.1). Nous envisageons trois choix pour le nombre de petits domaines m = 100, 500, 1 000. Pour chaque domaine, le vecteur de covariable se compose d'une valeur de 1 et d'un élément généré aléatoirement à partir de $N(10, 2)$. Le vecteur du coefficient de covariable est fixé à $\boldsymbol{\beta} = (20, 1)^\top$. Les valeurs possibles de la variance de l'erreur D_i sont 0,5; 1; 1,5; …, 5. Chaque valeur est attribuée au même nombre de domaines dans chaque ensemble de données. Nous considérons cinq scénarios pour générer les effets aléatoires u_i :

i) Normal : $u_i \sim N(0, 4)$,

ii) Mélange 0,2 : $u_i \sim \delta_i N(0, 25), \delta_i \sim \text{Ber}(0{,}2)$,

iii) Mélange 0,5 : $u_i \sim \delta_i N(0, 25), \delta_i \sim \text{Ber}(0{,}5)$,

iv) Mélange 0,8 : $u_i \sim \delta_i N(0, 25), \delta_i \sim \text{Ber}(0{,}8)$,

v) Distribution t de Student : $u_i \sim t_3$,

où Ber(p) désigne la distribution de Bernoulli selon une moyenne p. Pour faciliter la comparaison, les effets aléatoires générés à partir des scénarios ii) à v) sont remis à l'échelle pour avoir le même écart-type que ceux générés à partir du scénario i).

Nous générons 100 ensembles de données pour chaque combinaison de m et le scénario pour u_i. Le modèle GH proposé est adapté à chaque ensemble de données. Les échantillons *a posteriori* sont obtenus à l'aide de l'échantillonneur de Gibbs décrit à la section 3. L'échantillonneur est exécuté pour 20 000 itérations, et la première moitié est éliminée pour le rodage. Les moyennes de petits domaines θ_i sont estimées par les moyennes *a posteriori* de l'échantillon correspondantes. À des fins de comparaison, nous utilisons également le modèle de FH et le modèle GL pour estimer les moyennes de petits domaines. Pour le modèle GL, nous envisageons deux choix de distributions *a priori* pour les paramètres locaux λ_i^2, soit la distribution *a priori* de type *horseshoe* (HS) $\pi_L(\lambda_i^2) \propto (\lambda_i^2)^{-1/2}(1+\lambda_i^2)^{-1}$ (Carvalho, Polson et Scott, 2010) et la distribution *a priori* de Laplace (LA) $\pi_L(\lambda_i^2) = \exp(-\lambda_i^2)$. Ces distributions sont choisies en tant que représentantes des distributions *a priori* à queue polynomiale et exponentielle, respectivement.

Nous utilisons l'écart absolu moyen (EAM) et l'écart quadratique moyen (EQM) pour quantifier la différence entre les moyennes de petits domaines estimées et les valeurs réelles. Les deux critères sont définis comme suit :

$$\text{EAM} = \sum_{i=1}^{m} |\hat{\theta}_i - \theta_i|, \text{ et EQM} = \sum_{i=1}^{m} (\hat{\theta}_i - \theta_i)^2. \quad (4.1)$$

Nous construisons également les intervalles empiriques de crédibilité de 95 % pour θ_i et calculons les taux de couverture sur 100 ensembles de données afin d'examiner la quantification de l'incertitude.

4.2 Résultats

Les principaux résultats de notre étude par simulation sont présentés aux figures 4.1 à 4.3. En outre, les figures 4.1 et 4.2 fournissent l'EAM et l'EQM des moyennes de petits domaines estimées à partir de différents modèles et dans différents contextes de génération d'effets aléatoires. La figure 4.3 présente la couverture moyenne des intervalles empiriques de crédibilité de 95 % pour tous les petits domaines. Ces figures montrent que, parmi les modèles que nous envisageons, le modèle GH présente le rendement le plus robuste en ce qui concerne l'exactitude de l'estimation et la quantification de l'incertitude. Il convient de noter que le modèle qui produit les mesures d'écart les plus faibles varie selon les différents scénarios. Bien que le modèle GH ne soit pas toujours le meilleur modèle pour ce qui est des deux mesures d'écart, son rendement est souvent proche de celui du meilleur modèle, peu importe le nombre de petits domaines et les scénarios de génération d'effets aléatoires. Pour les modèles de FH, de type HS et de LA, bien que chacun d'eux produise les plus petits EAM et EQM dans certains contextes, leur rendement pourrait être bien pire que celui du modèle GH dans d'autres contextes. Par exemple, le modèle de FH présente un meilleur rendement selon le scénario « normal », et le modèle de LA présente un meilleur rendement selon le scénario « mélange 0,8 ». Cependant, selon le scénario « mélange 0.2 », ils produisent des EAM et des EQM plus élevés que le modèle de type HS et le modèle GH. Le modèle de type HS présente un meilleur rendement selon le scénario « mélange 0,2 », mais produit les mesures d'écart les plus élevées selon le scénario « normal » et le scénario « mélange 0,8 ». De plus, la couverture des intervalles de crédibilité du modèle de type HS est considérablement plus faible que celle des autres modèles et que le taux de couverture nominal selon les deux scénarios.

Le rendement robuste du modèle GH est obtenu par le choix adaptatif des hyperparamètres a et b. La figure 4.4 présente les estimations de la moyenne *a posteriori* $\hat{a}$ et $\hat{b}$ des hyperparamètres. Selon les trois scénarios « mélange », $\hat{a}$ et $\hat{b}$ augmentent à mesure que la proportion d'effets aléatoires non nuls augmente. Dans le scénario le moins dense (mélange 0,2), $\hat{b}$ est près de zéro et $\hat{a}$ est considérablement plus petit que un, ce qui indique que la composante polynomiale de la densité gamma de σ_i^2 joue un rôle essentiel dans la description des effets aléatoires. Dans le scénario le plus dense (mélange 0,8), $\hat{a}$ est proche de un, ce qui signifie que la composante exponentielle joue un rôle essentiel. Tang et coll. (2018) ont montré que, dans le modèle GL, les distributions *a priori* locales à queue polynomiale sont plus efficaces pour caractériser les petits effets aléatoires, tandis que les distributions *a priori* à queue exponentielle sont plus efficaces pour caractériser les effets aléatoires importants. Nos résultats correspondent aux leurs.

Figure 4.1 Écart absolu moyen des moyennes de petits domaines estimées à partir de différents modèles.

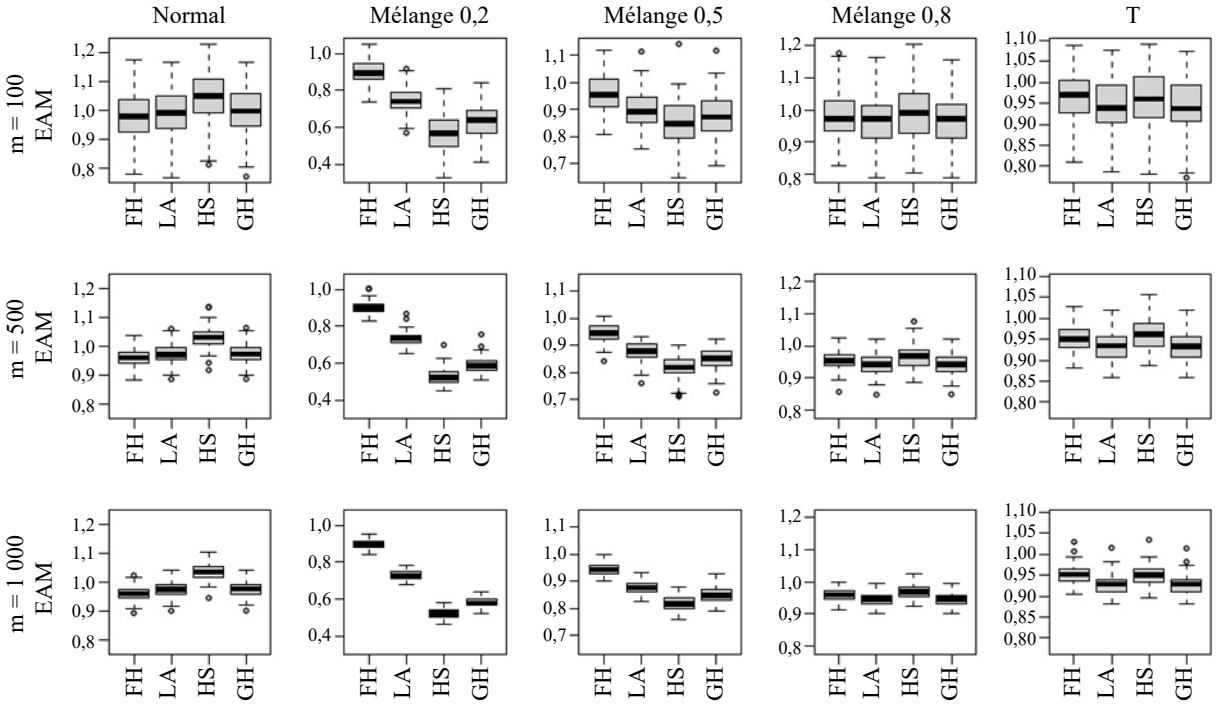


Figure 4.2 Écart quadratique moyen des moyennes de petits domaines estimées à partir de différents modèles.

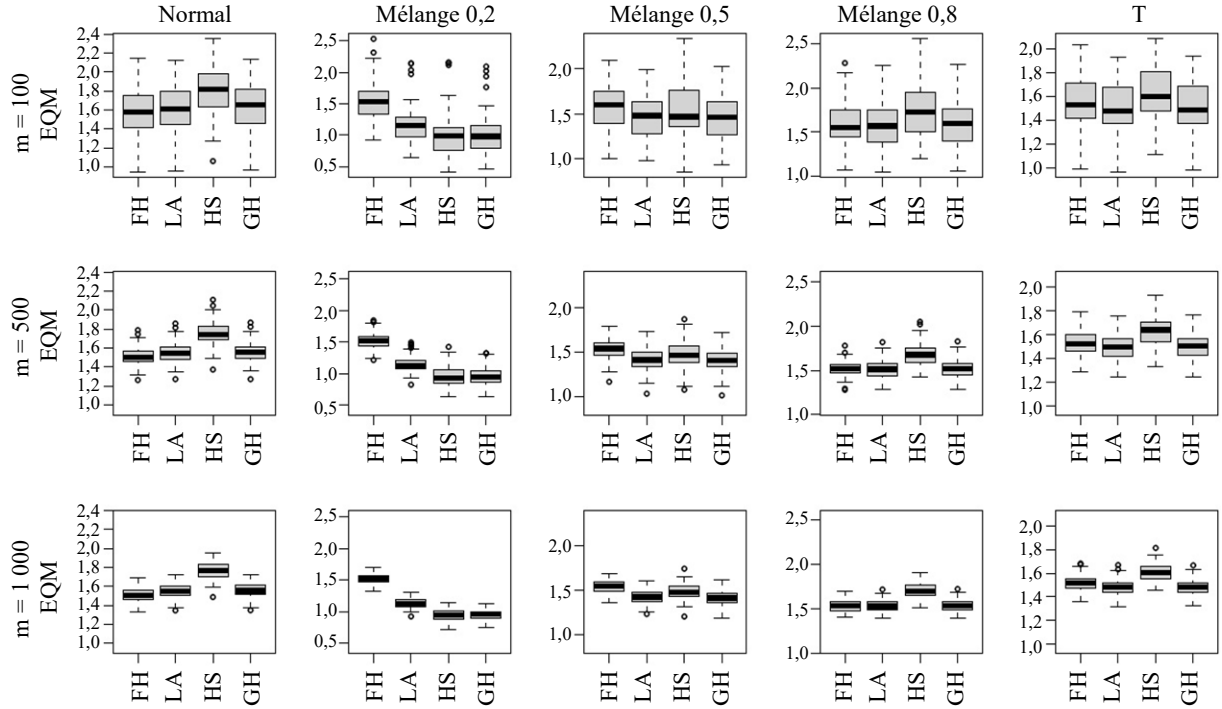


Figure 4.3 Couverture moyenne des intervalles empiriques de crédibilité de 95 % des moyennes de petits domaines à partir de différents modèles.

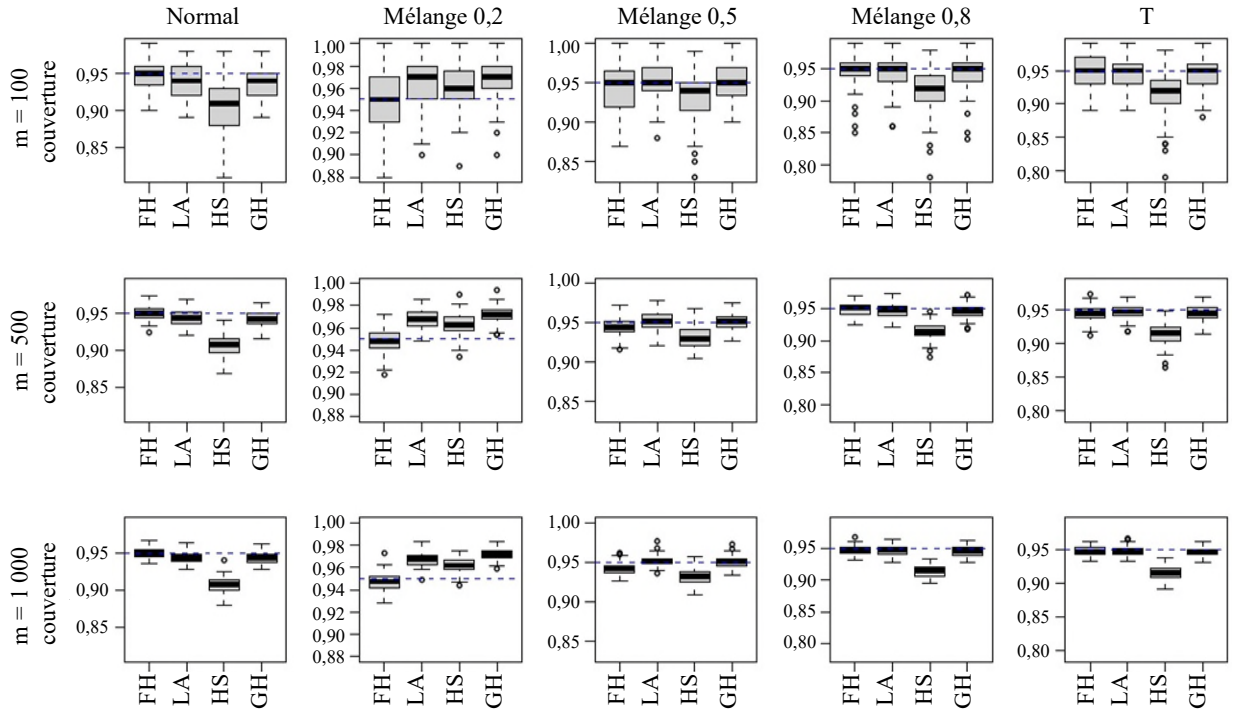
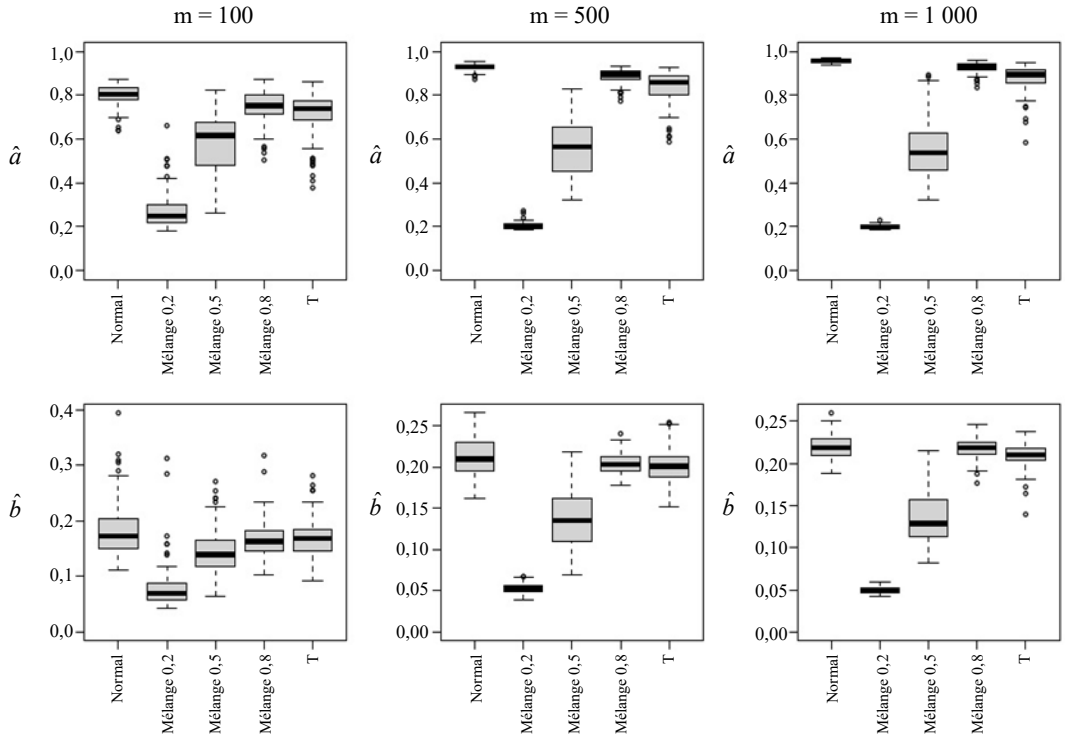
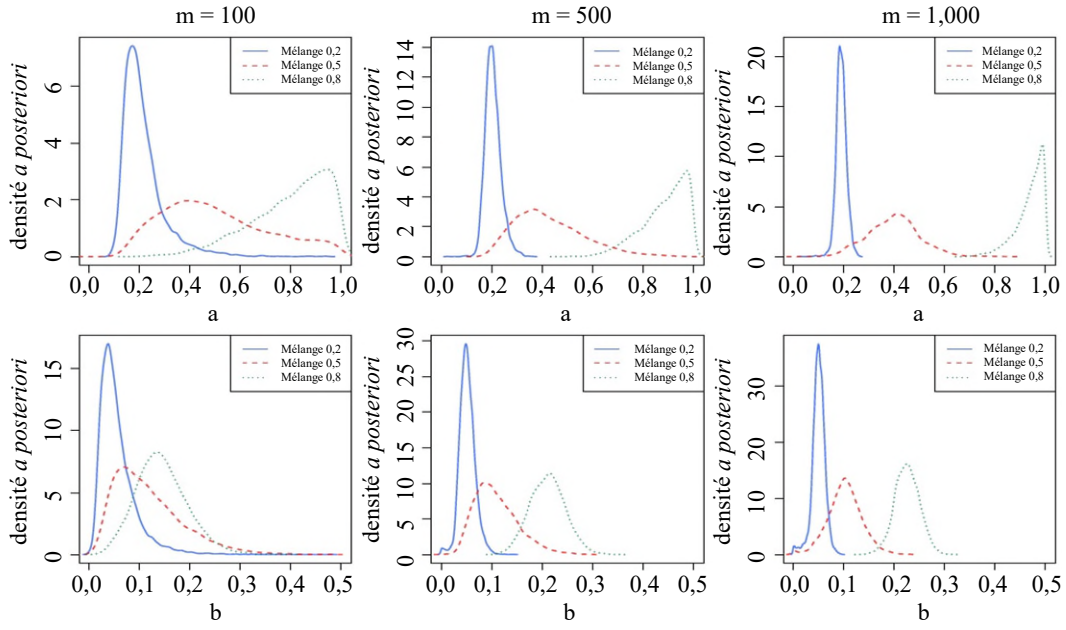


Figure 4.4 Hyperparamètres estimés *a* et *b* dans le modèle gamma hiérarchique.



La figure 4.5 présente la densité *a posteriori* de a et de b dans différents scénarios pour des ensembles de données sélectionnés. Pour a et b, la distribution *a posteriori* s'appuie sur des valeurs plus grandes pour les scénarios ayant une proportion plus élevée d'effets aléatoires non nuls.

Figure 4.5 Densité *a posteriori* de a et de b dans le modèle gamma hiérarchique pour un ensemble de données type.



Les figures 4.4 et 4.5 montrent également que la variation dans les deux hyperparamètres diminue à mesure que le nombre de petits domaines m augmente. Un nombre accru de petits domaines génère plus d'effets aléatoires et, par conséquent, plus d'information pour caractériser la distribution des effets aléatoires.

Le modèle GH permet d'obtenir un rendement robuste sans faire de grands sacrifices en coûts de calcul. Le tableau 4.1 présente le temps nécessaire pour obtenir des estimations sur petits domaines au moyen de différents modèles. En raison de l'effort supplémentaire pour estimer les hyperparamètres a et b, l'ajustement du modèle GH prend souvent plus de temps que l'ajustement des modèles GL. Cependant, l'augmentation n'est souvent pas supérieure à 15 %. Aussi, pour les modèles GL, comme il a été mentionné dans Tang et coll. (2018), il est possible d'utiliser le critère d'information de déviance (Spiegelhalter et coll., 2002) pour sélectionner la distribution *a priori* la plus appropriée pour le paramètre local. Pour ce faire, il faut ajuster les modèles GL selon plusieurs distributions *a priori* différentes pour les paramètres locaux, ce qui multiplie les coûts de calcul tout en permettant d'obtenir des résultats semblables à ceux générés par le modèle GH.

Tableau 4.1
Temps de calcul moyen (écart-type) en secondes sur 100 ensembles de données en fonction de divers scénarios.

m	Modèle	Scénario des effets aléatoires				
		Normal	Mélange 0,2	Mélange 0,5	Mélange 0,8	Distribution t de Student :
100	FH	2,78 (0,11)	2,76 (0,16)	2,74 (0,11)	2,75 (0,12)	2,77 (0,13)
	LA	16,37 (0,80)	16,32 (0,82)	16,34 (0,87)	16,28 (0,75)	16,24 (0,79)
	HS	17,00 (0,79)	17,01 (0,97)	17,00 (0,82)	16,92 (0,76)	16,94 (0,84)
	GH	17,43 (0,87)	17,72 (0,80)	17,61 (0,94)	17,44 (0,77)	17,50 (0,82)
500	FH	3,96 (0,17)	3,92 (0,18)	3,95 (0,22)	3,90 (0,17)	3,93 (0,17)
	LA	70,54 (3,84)	70,52 (3,76)	70,30 (3,61)	70,27 (3,66)	70,29 (3,62)
	HS	73,15 (3,76)	73,17 (3,76)	72,97 (3,56)	72,92 (3,56)	72,95 (3,39)
	GH	77,41 (3,87)	79,69 (4,12)	78,43 (4,00)	77,25 (3,57)	77,52 (3,79)
1 000	FH	5,33 (0,20)	5,32 (0,21)	5,29 (0,20)	5,27 (0,20)	5,22 (0,24)
	LA	138,12 (7,22)	138,06 (7,29)	137,38 (7,09)	137,03 (7,00)	134,73 (7,06)
	HS	144,01 (7,65)	143,69 (7,84)	142,83 (7,52)	142,66 (7,51)	138,57 (6,83)
	GH	152,88 (7,42)	157,93 (7,68)	154,94 (7,67)	152,45 (7,34)	151,80 (8,23)

5. Analyse de données réelles

Dans la présente section, nous estimons les taux de pauvreté au niveau de l'État et du comté aux États-Unis en utilisant le modèle proposé. Les deux ensembles de données que nous analysons proviennent de Datta et Mandal (2015) et de Tang et coll. (2018). Le premier ensemble de données concerne le ratio de pauvreté chez les enfants de 5 à 17 ans au niveau de l'État. Outre les estimations directes obtenues à partir de la Current Population Survey de 1999, l'ensemble de données comprend également le nombre d'exemptions pour enfant, le taux de non déclarants de l'Internal Revenue Service et les résidus de la régression des taux de pauvreté du recensement de 1989 sur les deux variables précédentes. Nous prenons en compte les trois variables comme covariables et une ordonnée à l'origine dans notre modèle. Dans le deuxième ensemble de données, nous avons des estimations directes des taux de pauvreté sur une période de cinq ans (2007 à 2011) regroupés au niveau du comté pour 3 141 comtés tirés de l'American Community Survey. Le taux de participation aux programmes de bons alimentaires est utilisé comme covariable en plus de l'ordonnée à l'origine.

En plus du modèle proposé, nous ajustons également le modèle de FH et le modèle GL aux distributions *a priori* de type HS et de LA pour chaque ensemble de données à des fins de comparaison. Le critère d'information de déviance (CID; Spiegelhalter et coll., 2002) est utilisé pour comparer l'ajustement du modèle; les valeurs sont présentées au tableau 5.1. Dans l'analyse au niveau de l'État, conformément aux principes énoncés par Datta et Mandal (2015) et Tang et coll. (2018), nous traitons les ratios de pauvreté réconciliés par le ratio au niveau de l'État obtenus à partir du recensement de 2000 comme étant les moyennes « réelles » de petits domaines, et nous mesurons les erreurs des valeurs estimées à l'aide de l'EAM et de l'EQM définis à (4.1). Les résultats sont également présentés au tableau 5.1.

Pour l'estimation au niveau de l'État, le modèle GH convient le mieux en ce qui concerne le CID. Les estimations médianes *a posteriori* des paramètres a et b dans le modèle GH sont, respectivement, de 0,53 et de 0,22. L'estimation relativement faible de b laisse entendre que la composante polynomiale dans la

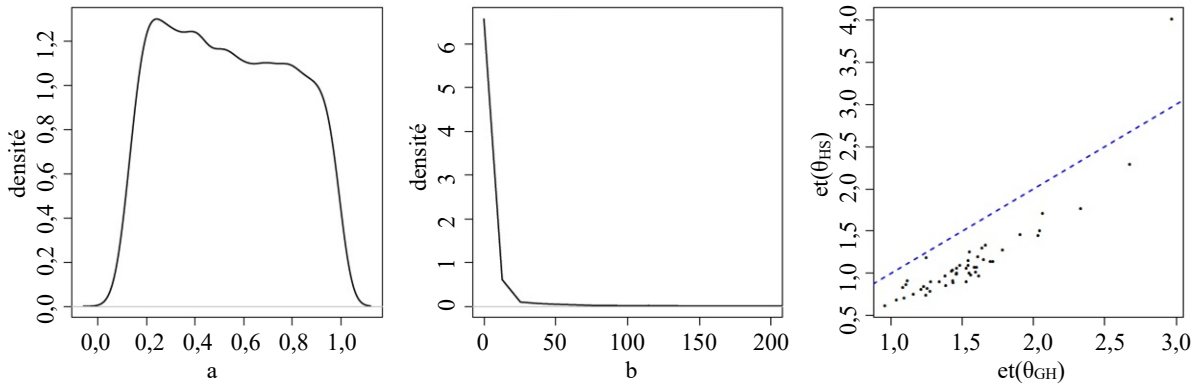
distribution *a priori* des effets aléatoires est importante. Cela concorde avec les résultats du CID, selon lesquels la distribution *a priori* de type HS est préférable à celle de LA dans les modèles GL. Les valeurs des deux mesures d'écart indiquent également que les modèles GH et de type HS offrent un rendement semblable, le modèle GH entraînant des erreurs légèrement plus importantes. Il convient de noter que le modèle de LA produit les plus petites erreurs, surtout en ce qui a trait à l'EQM, bien que le CID ne le favorise pas.

Tableau 5.1
Rendement de divers modèles relativement aux données au niveau de l'État et au niveau du comté.

	Mesure	GH	FH	HS	LA
Niveau de l'État	CID	271,52	273,29	273,09	275,92
	EAM	1,05	1,19	1,01	0,99
	EQM	2,19	2,55	2,04	1,68
Niveau du comté	CID	-15 946,23	-15 883,34	-15 751,12	-15 946,96

Les graphiques de gauche et du centre à la figure 5.1 présentent les densités *a posteriori* de a et de b. Les deux distributions montrent de fortes variations. La densité de a ne varie pas beaucoup de 0,1 à 0,9. La densité de b présente une queue très longue, bien qu'une partie importante de la probabilité soit répartie autour de zéro. La variation prononcée est principalement attribuable au nombre limité de petits domaines ($m = 51$), ce qui laisse supposer un degré élevé d'incertitude dans la détermination d'un modèle approprié. Cela se reflète également dans les valeurs du CID rapprochées pour différents modèles. De plus, l'incertitude des hyperparamètres a une incidence sur la variabilité des moyennes de petits domaines. Le graphique de droite à la figure 5.1 montre un tracé des écarts-types *a posteriori* des moyennes de petits domaines obtenues à partir du modèle de type HS par rapport à celles obtenues à partir du modèle GH. En raison de l'incertitude dans a et b, le modèle GH produit un écart-type *a posteriori* plus élevé dans la plupart des petits domaines.

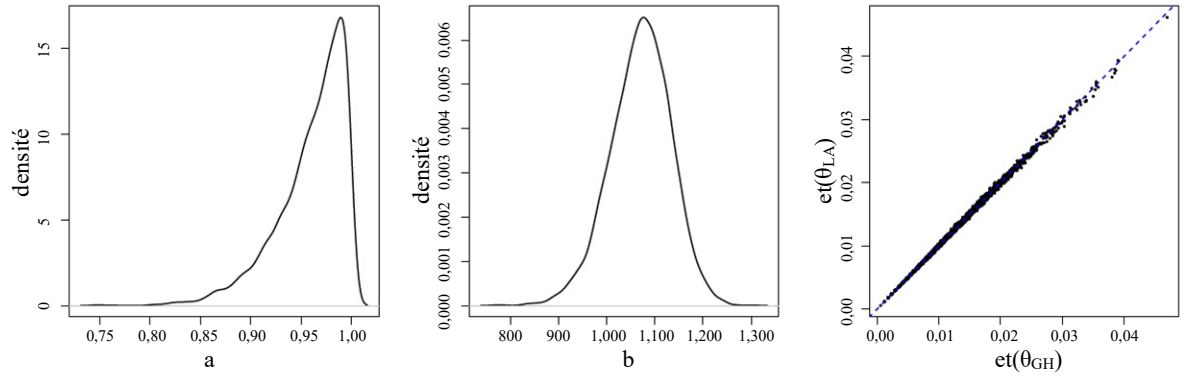
Figure 5.1 Résultats des estimations au niveau de l'État : densités *a posteriori* du paramètre a (gauche); densités *a posteriori* du paramètre b (centre); écarts-types *a posteriori* produits à partir des modèles GH et de type HS (droite).



Pour l'estimation au niveau du comté, Tang et coll. (2018) ont montré que la distribution *a priori* de LA a obtenu les meilleurs résultats parmi d'autres choix de distribution *a priori* dans le modèle GL. Les valeurs

du CID présentées au tableau 5.1 indiquent que le modèle GH proposé obtient un ajustement semblable à celui du modèle de LA. En fait, dans le modèle GH, la médiane *a posteriori* de a est de 0,97, ce qui est très proche de un, et la médiane *a posteriori* de b est de 1 073,25, ce qui signifie que la distribution *a priori* GH ressemble à la distribution *a priori* de LA à queue exponentielle dans ce cas. De plus, en raison d'un nombre beaucoup plus élevé de petits domaines dans les données au niveau du comté, les densités *a posteriori* de a et de b indiquées dans le graphique de gauche et le graphique du centre à la figure 5.2 présentent une variation inférieure par rapport aux variations issues de l'estimation au niveau de l'État. Comme le montre le graphique de droite à la figure 5.2, les écarts-types *a posteriori* des moyennes de petits domaines obtenues à partir des modèles GH et de LA sont également rapprochés.

Figure 5.2 Résultats des estimations au niveau du comté : densités *a posteriori* du paramètre a (gauche); densités *a posteriori* du paramètre b (centre); écarts-types *a posteriori* produits à partir des modèles GH et de type HS (droite).



Les ensembles de données et le code R permettant de produire les résultats de la présente section sont accessibles à l'adresse suivante : https://github.com/xytang/HGSAE.

6. Conclusion

Dans le présent article, nous avons proposé un modèle gamma hiérarchique (GH) pour les effets aléatoires dans l'estimation sur petits domaines. Il suppose que les effets aléatoires suivent un mélange d'échelles de distributions normales, la distribution de mélange étant gamma. Les hyperdistributions *a priori* sont ensuite appliquées au paramètre de forme et de taux de la distribution gamma. Au moyen de simulations et d'analyses de données réelles, nous avons démontré que le modèle proposé est capable de caractériser les effets aléatoires hétérogènes dans de petits domaines en tant que modèle global-local (GL) sans avoir à ajuster le modèle plusieurs fois pour choisir les distributions *a priori* convenant le mieux aux paramètres locaux.

Le modèle GH peut être considéré comme un mélange de différents modèles GL. En raison de cette formulation, la variation *a posteriori* des moyennes de petits domaines obtenues à partir du modèle GH, intègre, dans une certaine mesure, l'incertitude du modèle (c'est-à-dire la variation des résultats de l'estimation à partir de différents modèles). Lorsque le nombre de petits domaines est limité, l'information

disponible pour les hyperparamètres a et b est souvent insuffisante, ce qui accroît l'incertitude du modèle. Par conséquent, la variabilité des moyennes de petits domaines est plus prononcée par rapport au modèle GL selon un choix donné de distributions *a priori* pour les paramètres locaux. Dans cette optique, il est recommandé d'opter pour le modèle proposé lorsque le nombre de petits domaines est élevé afin d'éviter une variance d'estimation importante.

Le présent article porte sur les modèles au niveau du domaine pour l'estimation sur petits domaines. L'élaboration d'une approche similaire pour les modèles au niveau de l'unité représente une piste envisageable pour l'avenir.

Annexe

A. Démonstration du théorème 1

Dans le modèle proposé, la densité *a posteriori* peut s'écrire comme suit :

$$\begin{aligned}\pi(\boldsymbol{\beta},\mathbf{u},\boldsymbol{\sigma}^2,a,b\,|\,\mathbf{y}) = \quad & K\prod_{i=1}^{m}\exp\left[-\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}-u_i)^2}{2D_i}\right]\prod_{i=1}^{m}(2\pi\sigma_i^2)^{-1/2}\exp\left(-\frac{u_i^2}{2\sigma_i^2}\right)\\ & \times\prod_{i=1}^{m}\left[\frac{b^a}{\Gamma(a)}(\sigma_i^2)^{a-1}\exp(-b\sigma_i^2)\right]b^{s_0-1}\exp(-bt_0),\end{aligned} \tag{A.1}$$

où K est une constante générique qui ne dépend pas de $\boldsymbol{\beta}$, $\mathbf{u}$, $\boldsymbol{\sigma}^2$, a et b. Il suffit de montrer l'intégrale de $\pi(\boldsymbol{\beta},\mathbf{u},\boldsymbol{\sigma}^2,a,b\,|\,\mathbf{y})$ par rapport à $\boldsymbol{\beta},\mathbf{u},\boldsymbol{\sigma}^2,a,$ et b est fini.

Tout d'abord, considérons l'intégration en ce qui concerne $\boldsymbol{\beta}$. Supposons que $\mathbf{z}=(z_1,\ldots,z_m)^\top$ selon $z_i=y_i-u_i$ pour $i=1,\ldots,m$. Étant donné que $\mathbf{X}$ est de plein rang-colonne, nous pouvons définir $\hat{\boldsymbol{\beta}}=(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{z}}$. Notons que

$$\prod_{i=1}^{m}\exp\left[-\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}-u_i)^2}{2D_i}\right]=\exp\left[-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top(\mathbf{X}^\top\mathbf{D}\mathbf{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})-\frac{1}{2}\tilde{\mathbf{z}}^\top(\mathbf{I}-\mathbf{P})\tilde{\mathbf{z}}\right],$$

où $\mathbf{P}=\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top$. L'intégration de $\pi(\boldsymbol{\beta},\mathbf{u},\boldsymbol{\sigma}^2,a,b\,|\,\mathbf{y})$ par rapport à $\boldsymbol{\beta}$ donne

$$\begin{aligned}\pi(\mathbf{u},\boldsymbol{\sigma}^2,a,b\,|\,\mathbf{y}) \quad & = K\exp\left[-\frac{1}{2}\tilde{\mathbf{z}}^\top(\mathbf{I}-\mathbf{P})\tilde{\mathbf{z}}\right]\prod_{i=1}^{m}(2\pi\sigma_i^2)^{-1/2}\exp\left(-\frac{u_i^2}{2\sigma_i^2}\right)\\ & \quad\times\prod_{i=1}^{m}\left[\frac{b^a}{\Gamma(a)}(\sigma_i^2)^{a-1}\exp(-b\sigma_i^2)\right]b^{s_0-1}\exp(-bt_0)\\ & \le K\prod_{i=1}^{m}(2\pi\sigma_i^2)^{-1/2}\exp\left(-\frac{u_i^2}{2\sigma_i^2}\right)\prod_{i=1}^{m}\left[\frac{b^a}{\Gamma(a)}(\sigma_i^2)^{a-1}\exp(-b\sigma_i^2)\right]\\ & \quad\times b^{s_0-1}\exp(-bt_0).\end{aligned} \tag{A.2}$$

Il convient de noter que l'expression dans les deux dernières lignes de (A.2) donne la distribution *a priori* conjointe de $\mathbf{u}$, $\boldsymbol{\sigma}^2$, a et b. Étant donné que la distribution *a priori* est appropriée, la démonstration est maintenant terminée.

Bibliographie

Armagan, A., Clyde, M. et Dunson, D.B. (2011). Generalized beta mixtures of Gaussians. *Advances in Neural Information Processing Systems*, 523-531.

Carvalho, C.M., Polson, N.G. et Scott, J.G. (2009). Handling sparsity via the horseshoe. *International Conference on Artificial Intelligence and Statistics*, 73-80.

Carvalho, C.M., Polson, N.G. et Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2), 465-480.

Chakraborty, A., Datta, G.S. et Mandal, A. (2016). A two-component normal mixture alternative to the Fay-Herriot model. *Statistics in Transition New Series*, 17(1), 67-90.

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

Datta, G.S., et Mandal, A. (2015). Small area estimation with uncertain random effects. *Journal of the American Statistical Association*, 110(512), 1735-1744.

Datta, G.S., Hall, P. et Mandal, A. (2011). Model selection by testing for the presence of small-area effects, and application to area-level data. *Journal of the American Statistical Association*, 106(493), 362-374.

Fabrizi, E., et Trivisano, C. (2010). Robust linear mixed models for small area estimation. *Journal of Statistical Planning and Inference*, 140(2), 433-443.

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.

Gelfand, A.E., et Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410), 398-409.

Griffin, J.E., et Brown, P.J. (2005). Alternative prior distributions for variable selection with very many more variables than observations. Rapport technique, University of Warwick.

Griffin, J.E., et Brown, P.J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1), 171-188.

Li, Y., et Lahiri, P. (2007). Robust model-based and model-assisted predictors of the finite population total. *Journal of the American Statistical Association*, 102(478), 664-673.

Neal, R.M. (2003). Slice sampling. *The Annals of Statistics*, 31(3), 705-767.

Park, T., et Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482), 681-686.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28(1), 40-68.

Porter, A.T., Wikle, C.K. et Holan, S.H. (2015). Small area estimation via multivariate fay-herriot models with latent spatial dependence. *Australian & New Zealand Journal of Statistics*, 57(1), 15-29.

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

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. et Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, 64(4), 583-639.

Sugasawa, S., et Kubokawa, T. (2020). Small area estimation with mixed models: a review. *Japanese Journal of Statistics and Data Science*, 3, 693-720.

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*, 113(524), 1476-1489.

Ybarra, L.M.R., et Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95(4), 919-931.

Techniques d'enquête, décembre 2024
Vol. 50, nº 2, pp. 331-350
Statistique Canada, nº 12-001-X au catalogue

Estimation fondée sur un modèle des domaines petits et vides dans l'analyse des données d'enquête à l'aide de contraintes d'ordre

Xiyue Liao, Mary C. Meyer et Xiaoming Xu[1]

Résumé

Des travaux récents sur l'estimation des domaines d'enquête ont montré que l'intégration *a priori* d'hypothèses sur l'ordonnancement des moyennes des domaines de population réduit la variance des estimateurs et fournit des intervalles de confiance plus petits ayant une bonne couverture. Dans le présent document, nous montrons dans quelle mesure les hypothèses d'ordonnancement partiel permettent une estimation fondée sur un modèle des moyennes d'échantillon dans des domaines pour lesquels la taille de l'échantillon est nulle, lorsque les estimations de la variance et les intervalles de confiance sont conservateurs. Les restrictions d'ordre peuvent également améliorer considérablement l'estimation et l'inférence dans les domaines de petite taille. Des exemples de données d'enquêtes bien connues démontrent l'utilité de ces méthodes. Le code permettant de mettre en œuvre les exemples à l'aide du paquet `R csurvey` est fourni en annexe.

Mots-clés : Enquête; estimation de petit domaine; isotonique; moyenne d'un domaine; R; restrictions d'ordre.

1. Contexte et introduction

Considérons une population finie ayant les étiquettes $U = \{1,\ldots,N\}$ et supposons que U_d, $d = 1,\ldots,D$ désignent une partition de la population en domaines où U_d possède N_d éléments. Pour une variable à l'étude y, supposons que l'on souhaite estimer les moyennes des domaines de population

$$\bar{y}_{U_d} = \frac{\sum_{k \in U_d} y_k}{N_d}$$

pour chaque d, et fournir des inférences telles que des intervalles de confiance pour chacun d'entre eux $\bar{y}_{U_d}$. Selon le plan d'enquête, un échantillon $s \subset U$ est choisi; supposons $s_d = s \cap U_d$ pour $d = 1,\ldots,D$. L'estimateur de Hájek standard $\tilde{\mathbf{y}}_s = (\tilde{y}_{s_1},\ldots,\tilde{y}_{s_D})^\top$ des moyennes des domaines de la population est une moyenne pondérée des observations de l'échantillon dans chaque domaine d. Plus précisément :

$$\tilde{\mathbf{y}}_{s_d} = \frac{\sum_{i \in s_d} y_i / \pi_i}{\sum_{i \in s_d} 1/\pi_i},$$

où π_i est la probabilité d'échantillonnage pour l'élément de population i^{e} calculée à partir du plan d'échantillonnage (voir Särndal, Swensson et Wretman (1992), page 185).

Comme l'estimation pour chaque domaine ne repose que sur les observations à l'intérieur d'un domaine, un échantillon de domaine de petite taille se traduit par des estimateurs peu fiables dans ce domaine. Les

1. Xiyue Liao, Mary C. Meyer et Xiaoming Xu, San Diego State University, Colorado State University et Duke University. Courriel : xliao@sdsu.edu.

méthodes traditionnelles d'estimation pour les domaines de petite taille s'appuient sur des observations dans d'autres domaines pour fournir davantage de renseignements pour les domaines à échantillons de petite taille. L'estimateur de Fay-Herriot y parvient en imposant un modèle paramétrique sur les moyennes de domaine, avec un effet aléatoire sur domaine pour tenir compte de l'écart des moyennes de domaine par rapport au modèle global supposé. Voir Rao et Molina (2015) et Pfeffermann (2013) pour un traitement complet des méthodes d'estimation de petits domaines.

Nous examinerons les hypothèses *a priori* sur les moyennes de domaine de population qui peuvent être exprimées sous la forme d'un ordonnancement partiel de domaines. Par exemple, dans une enquête sur le milieu de travail, nous pourrions supposer que le salaire moyen augmente avec le rang de l'emploi, au sein d'un type d'emploi et d'un lieu. Dans une étude environnementale, il peut être raisonnable de supposer que la quantité de pollution diminue à mesure que l'on s'éloigne de la source. Les ordres imposés dans les exemples de la section 4 comprennent l'hypothèse selon laquelle les résultats des tests diminuent avec l'augmentation de la pauvreté et que le taux moyen de cholestérol augmente avec l'âge et le tour de taille du sujet. Les ordres permettent d'échanger des renseignements entre domaines sans modélisation paramétrique.

Supposons que $S = \{1,\ldots,D\}$ énumère les domaines; un ordonnancement partiel de S est précisé par une relation binaire $\preceq$, de sorte que pour d_i et $d_j \in S$, l'expression $d_i \preceq d_j$ signifie que nous supposons $\overline{y}_{U_i} \leq \overline{y}_{U_j}$. L'ordonnancement partiel doit avoir les propriétés suivantes : réflexif ($d \preceq d$ pour tout $d \in S$), antisymétrique (si $d_i, d_j \in S$, $d_i \preceq d_j$ et $d_j \preceq d_i$, alors $\overline{y}_{U_i} = \overline{y}_{U_j}$), et transitif (pour $d_i, d_j, d_k \in S$, si $d_i \preceq d_j$ et $d_j \preceq d_k$, alors $d_i \preceq d_k$). Un ordonnancement complet possède la propriété supplémentaire selon laquelle toutes les paires de points dans S sont comparables (si $d_i, d_j \in S$, alors soit $d_i \preceq d_j$, soit $d_j \preceq d_i$, soit les deux). Les ordres intéressants pour l'estimation de la moyenne de domaine d'enquête comprennent les ordonnancements complets, les ordres dans des grilles de domaines et les ordres par blocs.

Wu, Meyer et Opsomer (2016) ont considéré un ordonnancement complet sur la séquence des moyennes de domaine, en appliquant l'algorithme *Pooled Adjacent Violators* (Brunk, 1958) pour l'estimation de la moyenne de domaine. Ils ont obtenu des intervalles de confiance moins larges sans sacrifier la couverture, par rapport aux estimateurs qui ne tiennent pas compte de l'ordre. Oliva-Aviles, Meyer et Opsomer (2020) ont élaboré une méthodologie pour les ordonnancements partiels ainsi que des contraintes linéaires plus générales à appliquer aux domaines.

Un ordonnancement partiel peut être imposé aux estimateurs des moyennes de domaine à l'aide de contraintes d'inégalité linéaires sous la forme d'une matrice de contraintes $m \times D$ $\mathbf{A}$, et l'estimateur contraint $\tilde{\boldsymbol{\theta}}$ du vecteur de moyenne de domaine est trouvé en minimisant

$$\min_{\theta} (\tilde{\mathbf{y}}_s - \boldsymbol{\theta})^\top \mathbf{W}_s (\tilde{\mathbf{y}}_s - \boldsymbol{\theta}) \text{ de façon que } \mathbf{A}\boldsymbol{\theta} \geq \mathbf{0}. \tag{1.1}$$

La matrice de poids $\mathbf{W}_s$ est diagonale avec l'élément i^{e} $\hat{N}_i / \hat{N}$, où $\hat{N}_i = \sum_{i \in s_d} 1/\pi_i$ et $\hat{N} = \sum_{i=1}^{D} \hat{N}_i$.

Pour voir un exemple simple de matrice de contraintes, considérons cinq domaines ayant un ordonnancement complet, où nous supposons $\overline{y}_{U_1} \le \overline{y}_{U_2} \le \overline{y}_{U_3} \le \overline{y}_{U_4} \le \overline{y}_{U_5}$. Il peut s'agir du taux moyen de cholestérol pour cinq groupes d'âge, ou du salaire moyen pour chaque catégorie d'employés. La matrice de contraintes est la suivante :

$$\mathbf{A} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}.$$

Pour les ordonnancements complets sur les domaines D, la matrice de contraintes est $(D-1)\times D$. Pour obtenir un exemple d'ordonnancement partiel, supposons que nous considérions cinq groupes d'âge pour les femmes et les hommes et que nous voulions toujours ordonner le taux de cholestérol par âge à l'intérieur des sexes, mais sans avoir d'ordre entre les sexes. Si les cinq premiers domaines représentent les cinq groupes d'âge des femmes et les domaines 6 à 10, les groupes d'âge des hommes, la matrice de contraintes 8×10 est la suivante :

$$\mathbf{A} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

où les quatre premières lignes définissent les contraintes pour les groupes d'âge des femmes et les quatre dernières lignes, les contraintes pour les groupes d'âge des hommes. L'ordonnancement n'est pas complet, car, par exemple, le domaine 2 n'est pas comparable au domaine 8.

Les hypothèses *a priori* sur l'ordonnancement peuvent être vérifiées à l'aide du critère d'information du cône (CIC) élaboré par Oliva-Aviles, Meyer et Opsomer (2019). Le CIC est similaire à un critère d'information bien connu, le critère d'Akaike, en ce sens qu'il s'agit d'une mesure de la qualité de l'ajustement, avec une pénalité pour les degrés de liberté effectifs; il est fourni par le paquet `csurvey` (voir le code dans les annexes A et B). Le CIC est indiqué pour les estimateurs de Hajék contraints et non contraints; si le CIC est plus petit pour l'estimateur contraint, cela prouve que les hypothèses d'ordre sont correctes.

La solution de (1.1) est la projection pondérée de $\tilde{\mathbf{y}}_s$ sur le sous-ensemble C de $\mathbb{R}^D$ défini par $\mathbf{A}$:

$$C = \left\{ \boldsymbol{\theta} \in \mathbb{R}^D : \mathbf{A}\boldsymbol{\theta} \geq \mathbf{0} \right\}. \tag{1.2}$$

Ce sous-ensemble C de $\mathbb{R}^D$ est un *cône,* car pour tout $\boldsymbol{\theta} \in C$ et tout $a \geq 0$, $a\boldsymbol{\theta} \in C$. Oliva-Aviles, Meyer et Opsomer (2020) ont expliqué la façon dont une telle projection conique conduit à une mise en commun des données entre les domaines, ce qui permet d'obtenir des estimateurs plus précis. L'estimateur des moyennes de domaine ayant des contraintes d'ordre est construit par une mise en commun optimale des domaines sur lesquels les estimateurs non contraints ne respectent pas l'ordre, et la mise en commun réduit la variance estimée puisque les moyennes portent sur un plus grand nombre d'observations. Ils ont construit un estimateur de covariance basé sur la mise en commun observée et ont montré de quelle façon cet estimateur produisait des intervalles de confiance plus petits et ayant une bonne couverture. Xu, Meyer et Opsomer (2021) ont élaboré un estimateur de la variance fondé sur un mélange de matrices de covariance, que nous appellerons estimateur de covariance de mélange. Au lieu de construire l'estimateur de covariance au moyen du regroupement observé, l'estimateur de covariance de mélange reconnaît qu'un autre ensemble de données peut donner lieu à un regroupement différent; il repose donc sur une moyenne pondérée des matrices de covariance pour tous les regroupements possibles. Les deux auteurs ont présenté une théorie sur les grands échantillons et ont démontré que l'estimateur de covariance de mélange améliore la couverture des intervalles de confiance tout en conservant des longueurs d'intervalle plus petites.

Dans le présent article, nous poussons plus loin les travaux précédents en fournissant une estimation et une inférence pour les cellules vides qui ne sont pas à la « frontière » de l'ordre supposé. Nous généralisons ces idées en utilisant une méthode permettant d'imposer des contraintes d'ordre sur les limites de confiance supérieures et inférieures, ce qui permet d'obtenir des intervalles de confiance plus petits dans des domaines où la taille de l'échantillon est petite. Nous montrons que la couverture des intervalles de confiance ajustés est au moins aussi bonne que celle des intervalles de confiance originaux fournis par la matrice de covariance du mélange.

Dans la section 2, nous proposons une méthode simple pour estimer les moyennes d'un domaine (et fournir des intervalles de confiance) lorsque la taille de l'échantillon dans le domaine est nulle et qu'un ordonnancement partiel est supposé. Il s'agit d'un estimateur fondé sur un modèle qui n'intègre pas de modèle paramétrique pour faciliter l'estimation. Dans la section 3, nous imposons l'ordre supposé sur les limites supérieures et inférieures de l'intervalle de confiance, ce qui conduit à des longueurs d'intervalle de confiance plus petites pour les domaines ayant des échantillons de taille réduite. Nous montrons également que la méthode fournit des intervalles de confiance valides et nous fournissons quelques simulations visant à comparer les intervalles de confiance proposés avec ceux de l'estimateur de Hájek non contraint et de l'estimateur de Fay-Herriot. Des exemples d'estimation de petits domaines reposant sur des ensembles de données bien connus sont fournis à la section 4, et une analyse est proposée à la section 5. Les méthodes sont disponibles dans le paquet `csurvey`; l'annexe contient le code permettant de reproduire les résultats des exemples de la section 4.

2. Estimation pour les domaines vides dans un ordonnancement partiel

L'un des principaux avantages de la mise en œuvre de contraintes d'inégalité valides est que les données sont mises en commun entre les domaines pour construire les estimations, alors que pour l'estimateur non contraint fondé sur un modèle, chaque estimateur de domaine repose uniquement sur les observations de ce domaine. Lorsque la taille de l'échantillon dans un domaine est petite, les estimations de la variance fondées sur un modèle sans contraintes ne sont pas fiables, et pour les cellules ne comportant qu'une seule observation, ou pour les cellules vides, la variance peut être impossible à estimer sans utiliser des hypothèses supplémentaires, comme dans certaines méthodes d'estimation pour petits domaines. Voir la Rao et Molina (2015) (préface) pour obtenir une analyse intéressante des raisons pour lesquelles les méthodes d'estimation des petits domaines sont nécessaires et sur les cas où elles le sont. Les contraintes d'ordre, si elles sont appropriées, permettent une approche entièrement fondée sur un modèle pour l'estimation des moyennes dans les domaines où la taille de l'échantillon est faible.

Si un domaine d n'a pas d'observations, il ne peut pas être inclus dans l'estimation fondée sur (1.1), mais si nous pouvons supposer certaines contraintes d'inégalité qui supposent $\overline{y}_{U_d}$, une estimation pour $\overline{y}_{U_d}$ peut être fournie avec un intervalle de confiance conservateur. À titre d'exemple, supposons qu'il y a 20 domaines ayant un ordre non décroissant simple, et que seul le domaine 16 n'a pas d'observations. Nous pouvons obtenir des estimations et des intervalles de confiance pour les 19 domaines non vides et affirmer que la moyenne du 16e domaine de population ne doit pas dépasser le 17e, mais doit être au moins aussi grande que celle du 15e. La valeur inférieure de l'intervalle de confiance pour la moyenne du 15e domaine, combinée à la valeur supérieure de l'intervalle pour la moyenne du 17e domaine, fournit un intervalle de confiance prudent pour la moyenne du 16e domaine. L'estimateur de la moyenne du 16e domaine peut être considéré comme le centre de son intervalle de confiance ou comme le plus proche possible du centre, tout en respectant les contraintes. S'il existe des domaines vides consécutifs, les domaines limites non vides produisent les intervalles de confiance.

Pour des contraintes plus complexes, nous pouvons appliquer la même idée d'obtenir des limites supérieures et inférieures pour les domaines vides, en utilisant les estimations et les intervalles de confiance calculés à partir des observations dans les autres domaines. Supposons que D' est le nombre de domaines et que m' est le nombre total de contraintes imposées aux moyennes des domaines, à l'aide d'une matrice de contraintes $m' \times D'$ $\mathbf{A}'$. Supposons que D est le nombre de domaines non vides et que m est le nombre de contraintes imposées aux moyennes des domaines non vides; nous pouvons alors obtenir une matrice de contraintes $m \times D$ $\mathbf{A}$ en modifiant $\mathbf{A}'$. Grâce à ces valeurs, l'estimateur contraint $\tilde{\boldsymbol{\theta}}$ ainsi que la matrice de covariance du mélange $D \times D$ peuvent être obtenus pour les domaines non vides D. Pour un domaine vide d, nous examinons la d^{e} colonne de $\mathbf{A}'$ pour trouver des domaines non vides d_1 et d_2 de telle sorte que $\overline{y}_{U,d_1} \leq \overline{y}_{U,d} \leq \overline{y}_{U,d_2}$. Si deux domaines de ce type n'existent pas, comme dans le cas d'un domaine « en coin », la moyenne du domaine ne peut pas être estimée et l'intervalle de confiance ne sera pas borné à l'une des extrémités. S'il existe au moins une valeur d_1 telle que $\overline{y}_{U,d_1} \leq \overline{y}_{U,d}$, nous examinons les intervalles de confiance de toutes ces valeurs d_1 et choisissons la limite inférieure la plus élevée comme limite inférieure

de l'intervalle de confiance du domaine vide. De même, s'il existe au moins une valeur d_2 pour laquelle $\overline{y}_{U,d} \leq \overline{y}_{U,d_2}$, nous examinons les intervalles de confiance de toutes ces valeurs d_2 et choisissons la plus petite limite supérieure comme limite supérieure de l'intervalle de confiance du domaine vide.

3. Ordonnancement des limites de l'intervalle de confiance

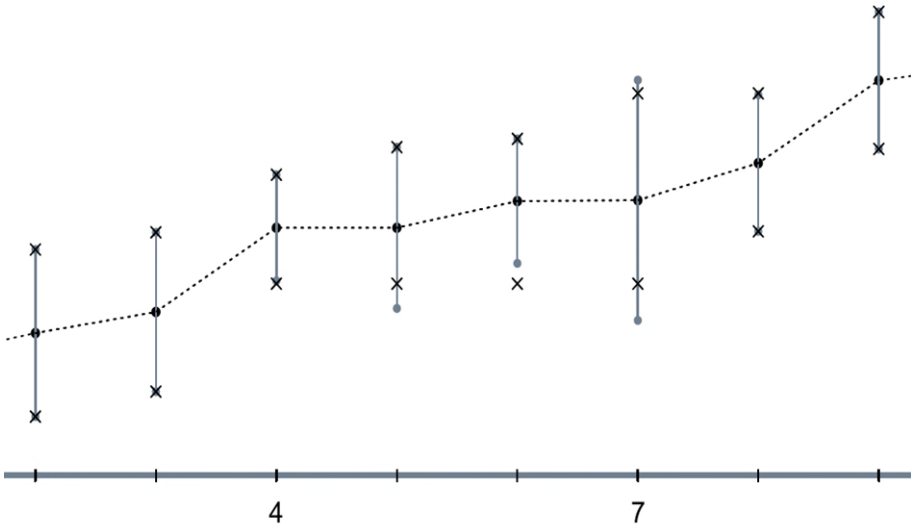
Pour les méthodes avec contraintes, le problème des échantillons de petite taille dans les domaines est atténué parce que l'estimateur de la moyenne du domaine et l'estimateur de la variance du mélange utilisent tous deux des données provenant d'autres domaines. Les ajustements des intervalles de confiance proposés ici, en imposant des contraintes d'ordre sur les limites de confiance supérieures et inférieures ainsi que sur les estimations de domaine moyennes, permettent un meilleur échange de l'information entre les domaines.

Les contraintes d'ordre doivent également être vérifiées dans les limites supérieures et inférieures de l'intervalle de confiance. Pour s'en convaincre, il suffit de considérer le cas où la moyenne du domaine 6, par exemple, est censée ne pas être supérieure à la moyenne du domaine 7. La limite supérieure de l'intervalle de confiance pour le domaine 7 indique un niveau de confiance selon lequel la moyenne de la population du domaine est inférieure à cette limite, de sorte que ce niveau de confiance devrait également s'appliquer à la moyenne de la population (plus petite) pour le domaine 6.

Les intervalles de confiance fournis par la matrice de covariance du mélange ne satisfont pas nécessairement aux contraintes. Rappelons que cette matrice de covariance estimée est une moyenne pondérée de matrices de covariance linéaires et que les limites supérieures et inférieures calculées à l'aide de cette estimation de la covariance du mélange ne suivent pas nécessairement l'ordonnancement partiel supposé pour les domaines. Cela est illustré dans la figure 3.1, où le domaine 7 a une petite taille d'échantillon par rapport aux domaines environnants. Les moyennes estimées du domaine sont indiquées par des points noirs, alors que les points gris correspondent aux limites des intervalles de confiance non ajustés calculés à partir de la matrice de covariance du mélange. Dans l'exemple de la figure, la variance estimée pour le domaine 7 est plus importante, ce qui fait que la limite inférieure est inférieure à celle du domaine 6. Cependant, nous supposons que la moyenne du 7e domaine de population est au moins aussi importante que la moyenne du 6e domaine de population et nous ajustons donc les limites de confiance de manière à ce qu'elles satisfassent aux contraintes d'ordre.

De façon plus générale, si $\tilde{\mathbf{u}}$ représente la limite supérieure des intervalles de confiance calculés à partir des variances données par l'estimateur de covariance du mélange, nous projetons $\tilde{\mathbf{u}}$ sur le cône (1.2), en utilisant les tailles d'échantillon (ou les tailles d'échantillon effectives) comme pondérations. La projection $\hat{\mathbf{u}}$ est le nouvel ensemble de limites supérieures satisfaisant aux contraintes. De même, $\hat{\ell}$ est le nouvel ensemble de limites inférieures obtenu par projection pondérée des limites inférieures initiales $\tilde{\ell}$ sur le cône de contrainte. Les marques × de la figure 3.1 indiquent les nouvelles limites supérieures et inférieures soumises à des contraintes d'ordre. Les intervalles de confiance pour les domaines à faible taille d'échantillon subissent un ajustement plus marqué que ceux des domaines de plus grande taille, ce qui permet d'obtenir une estimation plus précise pour les petits domaines.

Figure 3.1 Exemple d'ajustement de la limite de confiance pour s'assurer que les limites respectent les contraintes d'ordre.



Notes : Les estimations de la moyenne du domaine sont indiquées par des points noirs, les limites de confiance fournies par la matrice de covariance du mélange sont indiquées par des points gris et les limites de confiance ajustées sont indiquées par des ×.

Les nouvelles limites supérieures $\hat{\mathbf{u}}$ sont la projection pondérée de $\tilde{\mathbf{u}}$ sur C. Autrement dit, $\hat{\mathbf{u}}$ minimise $\sum_{d=1}^{D} n_d(\tilde{u}_d - u_d)^2$ sur $\mathbf{u} \in C$, où n_d est le nombre d'observations dans le domaine $d \in S$. Supposons que $A_c \subseteq S$ est défini comme $\{d \in S : \hat{u}_d = c\}$. Le lemme suivant est similaire au théorème 1.3.5 de Robertson, Wright et Dykstra (1988) et nous indique que tout $\hat{u}_d$ est une moyenne pondérée d'un sous-ensemble de $\tilde{u}_1,\ldots,\tilde{u}_D$.

Lemme 1. *Si A_c n'est pas vide, alors*

$$c = \frac{\sum_{d \in A_c} \tilde{u}_d n_d}{\sum_{d \in A_c} n_d}.$$

Preuve : Supposons que $L_c = \{d : \hat{u}_d < c\}$ et $U_c = \{d : \hat{u}_d > c\}$. Supposons que c_1 est le maximum de $\hat{u}_j$ pour $j \in L_c$, et c_2 le minimum de $\hat{u}_j$ pour $j \in U_c$ alors $c_1 < c < c_2$. Écrivons

$$\sum_{d=1}^{D} [\hat{u}_d - \tilde{u}_d]^2 n_d = \sum_{d \in L_c} [\hat{u}_d - \tilde{u}_d]^2 n_d + \sum_{d \in A_c} [c - \tilde{u}_d]^2 n_d + \sum_{d \in U_c} [\hat{u}_d - \tilde{u}_d]^2 n_d,$$

et notons que la valeur de c qui minimise le terme moyen est donnée dans le lemme. Si le résultat n'était pas vrai, le terme moyen pourrait être réduit en déplaçant c vers le haut ou vers le bas, vers la moyenne pondérée donnée de $\tilde{u}_d$, et si nous restons à l'intérieur de (c_1, c_2), la fonction satisferait toujours aux contraintes.

Xu, Meyer et Opsomer (2021) ont démontré que les limites de confiance $(\tilde{\ell}_d, \tilde{u}_d)$ obtenues à partir de la matrice de covariance du mélange ont une couverture asymptotique correcte si les contraintes sont strictes.

Si pour tout d, $\hat{u}_d \geq \tilde{u}_d$, la couverture supérieure de $\bar{y}_{U_d}$ est au moins aussi bonne que pour l'intervalle non ajusté. De même, si $\hat{\ell}_d \leq \tilde{\ell}_d$, alors $\hat{\ell}_d \leq \bar{y}_{U_\ell}$ si $\tilde{\ell}_d \leq \bar{y}_{U_\ell}$. La couverture des intervalles de confiance ajustés lorsque $\hat{u}_d < \tilde{u}_d$ ou $\hat{\ell}_d > \tilde{\ell}_d$ est abordée ci-dessous.

Théorème 1. *Pour $d \in S$ de sorte que $\hat{u}_d < \tilde{u}_d$, prenons $c = \hat{u}_d$ et définissons $A_c^+ = \{j \in S: \hat{u}_j = c \text{ et } d \preceq j\}$. Alors A_c^+ a au moins deux éléments, et si pour tout $j \in A_c^+$ nous avons $\tilde{u}_j \geq \bar{y}_{U_j}$, cela garantit $\hat{u}_d \geq \bar{y}_{U_d}$. De même, pour $d \in S$ de sorte que $\hat{\ell}_d > \tilde{\ell}_d$, prenons $c = \hat{\ell}_d$ et définissons $A_c^- = \{j \in S: \hat{\ell}_j = c \text{ et } j \preceq d\}$. Alors A_c^- comporte au moins deux éléments, et si pour tout $j \in A_c^-$ nous avons $\tilde{\ell}_j \leq \bar{y}_{U_j}$, cela garantit $\hat{\ell}_d \leq \bar{y}_{U_d}$.*

Preuve : Supposons que pour un domaine d, $\hat{u}_d \neq \tilde{u}_d$. Si $\hat{u}_d > \tilde{u}_d$, alors certainement $\tilde{u}_d \geq \bar{y}_{U_d} \Rightarrow \hat{u}_d \geq \bar{y}_{U_d}$. Si $\hat{u}_d < \tilde{u}_d$, supposons $c = \hat{u}_d$. Nous savons par le lemme 1 que A_c compte plus d'un élément. Il existe au moins un $j \in A_c^+$ de sorte que $\tilde{u}_j < \tilde{u}_d$; si ce n'est pas le cas, alors $\hat{u}_d$ pourrait être plus grand et plus proche de $\tilde{u}_d$ sans modifier l'autre $\hat{u}_j$. Parce que $\bar{y}_{U_d} \leq \bar{y}_{U_j}$ et $\tilde{u}_j < \hat{u}_j = \hat{u}_d$,

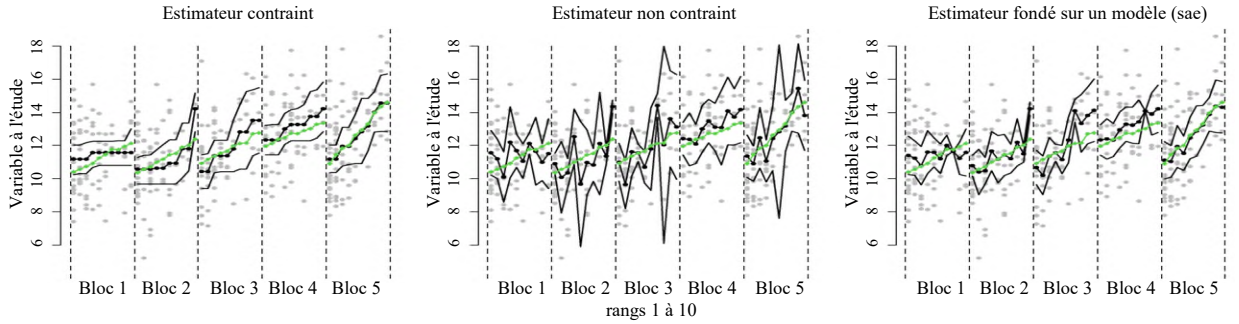
$$\tilde{u}_j \geq \bar{y}_{U_j} \Rightarrow \hat{u}_j \geq \bar{y}_{U_j} \Rightarrow \hat{u}_d \geq \bar{y}_{U_j} \Rightarrow \hat{u}_d \geq \bar{y}_{U_d}.$$

La preuve de la limite inférieure est similaire.

Pour démontrer l'efficacité des méthodes des domaines vides et petits, nous avons généré une population d'une taille de 40 000 en utilisant un vecteur $\boldsymbol{\mu}$ de 50 moyennes de domaine, de sorte que chaque domaine présente 800 valeurs de population provenant d'une distribution normale ayant une moyenne de μ_i et une variance de 4. Supposons que les unités de la population simulée soient des travailleurs dans un certain domaine et que les valeurs représentent le logarithme des salaires. Les 50 domaines comptent 10 catégories d'emplois et 5 lieux. Les domaines les moins bien classés comptent un plus grand nombre de travailleurs. Nous supposons qu'à l'intérieur de chaque lieu, les salaires augmentent en fonction du rang, et nous imposons un ordre en bloc aux lieux : pour chaque rang, le salaire moyen du lieu 4 est supérieur au salaire moyen du rang correspondant dans n'importe lequel des lieux 1, 2 et 3, et les salaires moyens des rangs du lieu 5 sont également supérieurs à ceux des trois premiers lieux. Cependant, aucun ordre n'est imposé à l'intérieur des lieux 1, 2 et 3; de même, nous n'imposons pas d'ordre sur les salaires moyens des lieux 4 et 5. Un plan stratifié est utilisé pour échantillonner la population.

Une taille d'échantillon de $n = 400$ de cette population est représentée par les points gris dans la figure 3.2. Les moyennes réelles de la population sont représentées par les losanges, qui sont presque linéaires dans la population finie. Les estimations sont représentées par des points noirs et les intervalles de confiance à 95 % sont également indiqués. L'estimateur contraint est comparé à l'estimateur de Hajék non contraint et à l'estimation de Fay-Herriot fournie par le paquet `sae` en utilisant le rang et le bloc comme prédicteurs ordinaux et nominaux, respectivement. La taille moyenne de l'échantillon est de 8, mais certains domaines correspondant à des rangs plus élevés sont plus susceptibles d'avoir des tailles d'échantillon plus petites. Les intervalles de confiance de l'estimateur contraint semblent avoir une meilleure couverture et une petite longueur.

Figure 3.2 Un ensemble de données simulées provenant d'une population de travailleurs où la variable à l'étude est log(salaire).

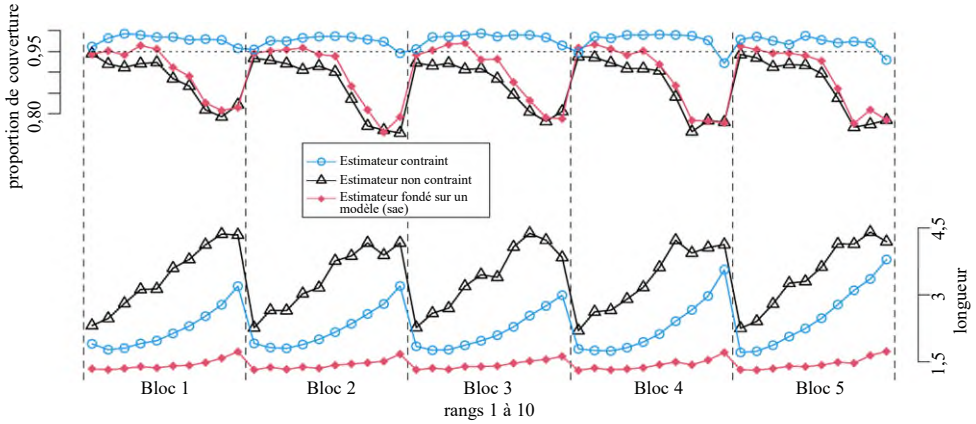


Notes : Pour l'estimateur contraint, on suppose que le salaire augmente dans les rangs 1 à 10 au sein de chaque lieu et que les salaires suivent un ordre séquentiel entre les lieux, les lieux 4 et 5 ayant des salaires moyens plus élevés que les lieux 1, 2 et 3. Les moyennes de la population sont représentées par des losanges plus clairs, tandis que les estimations sont représentées par des points noirs. Les intervalles de confiance approximatifs à 95 % sont représentés par les lignes. L'estimateur non contraint donné par `survey` et l'estimateur de Fay-Herriot donné par `sae` sont également présentés.

Pour démontrer la couverture et la précision des estimations limitées par l'ordre avec un échantillon de petite taille, nous avons échantillonné 1 000 ensembles de données de taille 400, en utilisant le plan stratifié. Pour chaque ensemble de données, nous avons calculé les estimations contraintes, les estimations de Hajék, les estimations de Fay-Herriot et les intervalles de confiance à 95 % pour chacune d'entre elles. Pour chacun des 50 domaines, nous avons déterminé la proportion d'ensembles de données pour lesquels l'intervalle de confiance tient compte de la moyenne de la population, et nous avons également déterminé la longueur des intervalles de confiance. Les taux de couverture et les longueurs d'intervalle sont résumés pour les deux estimateurs dans la figure 3.3. L'estimation de Hajék ne pouvant être calculée dans des domaines vides, le pourcentage de couverture indiqué est limité aux ensembles de données comportant des domaines non vides. De même, le pourcentage de couverture de l'estimateur fondé sur un modèle correspond aux ensembles de données simulées pour lesquelles l'estimateur peut être calculé.

Pour l'estimateur Hajék, les intervalles pour les rangs plus élevés offrent une faible couverture au sein de chaque bloc, en raison de la taille réduite des échantillons. L'estimateur `sae` fondé sur un modèle constitue une amélioration dans la mesure où la couverture est plus élevée que pour l'estimateur non contraint et où les longueurs sont considérablement plus petites. Toutefois, la couverture `sae` ne se rapproche de l'objectif que pour les rangs 1 à 5, pour lesquels la population et la taille de l'échantillon sont plus importants. La couverture de l'estimateur contraint par l'ordre est encore bonne pour ces domaines, parce que les données des domaines ayant des tailles d'échantillon plus importantes sont utilisées. La longueur des intervalles contraints est systématiquement inférieure à celle de l'estimateur de Hajék, tandis que la couverture est systématiquement bonne. Le code complet permettant de produire ces résultats de simulation est disponible dans le matériel supplémentaire.

Figure 3.3 Probabilités de couverture (en haut) et longueur des intervalles (en bas) pour des intervalles de confiance à 95 % pour 50 moyennes de domaine ayant $n = 400$, en supposant que la moyenne de domaine augmente avec le rang et que les blocs 4 et 5 des moyennes plus élevées que les blocs 1, 2 et 3.



Notes : Les échantillons des rangs plus élevés sont plus petits, de sorte que la couverture est médiocre en l'absence de contraintes.

4. Applications

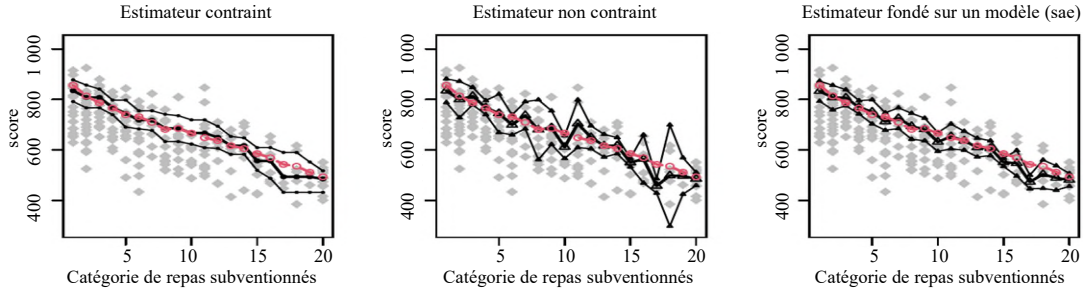
Le premier exemple repose sur un ordonnancement complet. L'ensemble de données `apipop` du paquet `R survey` contient des données de tests normalisées provenant de $N = 6\,194$ écoles californiennes. Nous utiliserons cet ensemble de données comme population à partir de laquelle nous procéderons à un échantillonnage, afin de comparer les performances de l'estimateur contraint à celles de l'estimateur non contraint standard. Comme nous connaissons les moyennes des domaines de population, nous pouvons comparer les erreurs d'ajustement et déterminer les proportions de couverture pour les intervalles de confiance.

Supposons que l'on s'intéresse à la moyenne des résultats des tests normalisés de l'école, appelée `api00` dans l'ensemble des données, et à la manière dont elle peut être liée à une mesure de l'aisance. La variable `meals` représente la proportion d'élèves de l'école qui ont droit à des repas subventionnés; nous classons cette mesure en 20 niveaux de cinq points de pourcentage chacun et supposons que le score moyen au test diminue à mesure que la proportion d'élèves ayant droit à des repas subventionnés augmente.

Les observations pour un seul échantillon de taille $n = 240$ sont représentées par des losanges gris dans la figure 4.1. L'échantillon est stratifié par type d'école, la taille étant de 60 pour les écoles élémentaires et les écoles intermédiaires, et de 120 pour les écoles secondaires. Les moyens du domaine de la population sont représentés par des cercles reliés par des lignes en pointillés. À gauche, les estimations moyennes du domaine contraintes à être décroissantes sont représentées par les points reliés par des lignes pleines, et les limites de l'intervalle de confiance à 95 % sont également indiquées. Au centre, les estimateurs de la moyenne du domaine sans contraintes (Hajék) sont présentés avec leurs intervalles de confiance. Les longueurs des intervalles de confiance pour les estimateurs contraints sont plus petites et, pour cet échantillon, ces intervalles de confiance tiennent compte de toutes les moyennes du domaine de population. À droite se trouvent les estimations et les intervalles de confiance pour l'estimateur de Fay Herriot fondé sur un modèle, tel que fourni par l'ensemble `sae`. Les longueurs des intervalles de confiance sont les plus

petites pour cet estimateur, mais les intervalles ne tiennent pas compte de toutes les moyennes de la population.

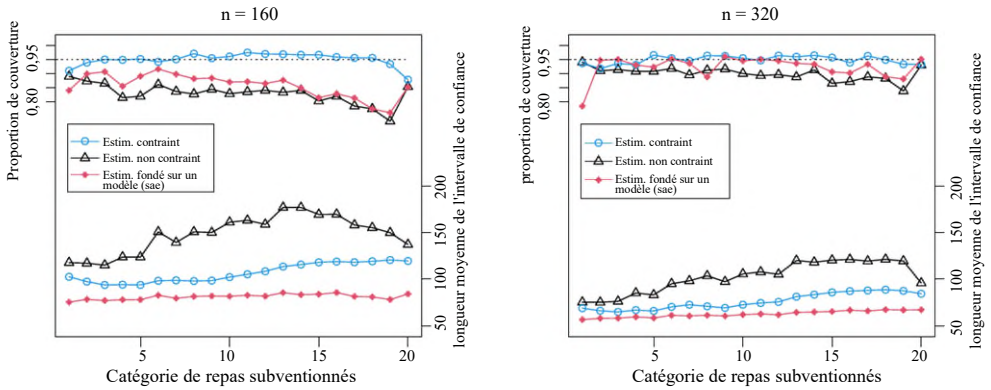
Figure 4.1 Moyennes par domaine et intervalles de confiance pour un échantillon stratifié dans l'ensemble de données `R api`, $n = 240$.



Notes : Les valeurs de la population sont représentées par des cercles et les valeurs de l'échantillon, par des losanges gris. Les intervalles de confiance à 95 % sont indiqués.

Les taux de couverture et les longueurs d'intervalle réels pour différentes tailles d'échantillon peuvent être établis par des échantillonnages répétés. Nous utilisons le même plan d'échantillonnage et, pour chacun des 1 000 échantillons, nous déterminons les estimations du domaine de population et leurs intervalles de confiance à 95 %. Les longueurs moyennes des intervalles de confiance et les proportions de couverture sont présentées à la figure 4.2, où il est indiqué que les taux de couverture pour les estimations non contraintes sont inférieurs à l'objectif, la couverture n'étant raisonnable que pour les échantillons de grande taille. Nous avons choisi $n = 160$ pour un échantillon de taille « moyenne » ayant une moyenne de huit observations par domaine, et $n = 320$, pour un échantillon de « grande » taille. Pour la taille d'échantillon $n = 160$, les estimateurs non contraints ont une couverture médiocre, qui s'améliore pour $n = 320$. L'estimateur contraint présente toutefois des proportions de couverture souvent supérieures à l'objectif. En outre, les longueurs des intervalles contraints sont systématiquement inférieures à celles de l'estimateur de Hajék.

Figure 4.2 Probabilités de couverture et longueur des intervalles pour des tailles d'échantillons stratifiés de 160 et de 320 pour 20 domaines.

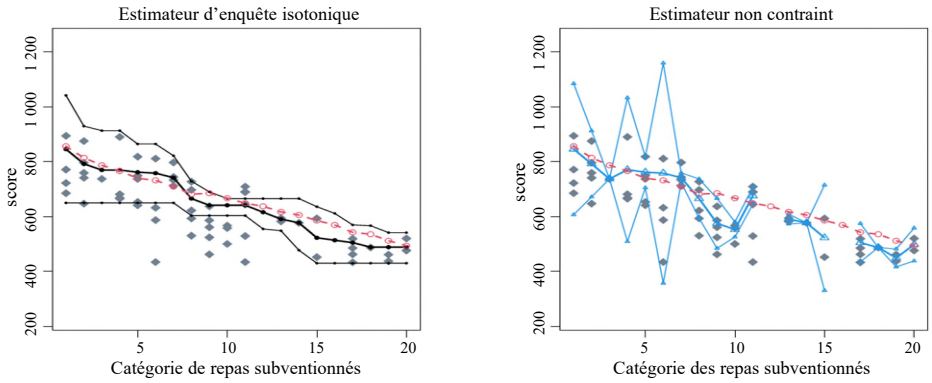


Notes : Les estimateurs contraints ont systématiquement une meilleure couverture et des longueurs plus faibles que les estimateurs non contraints.

Nous choisissons ensuite un échantillon stratifié unique de $n = 60$: taille de 15 pour les écoles élémentaires et les écoles intermédiaires, et de 30 pour les écoles secondaires. Cette taille d'échantillon est suffisamment petite pour que l'on obtienne des domaines vides et des domaines de petite taille pour chacun des échantillons. L'échantillon présenté à la figure 4.3 comporte trois domaines vides et deux domaines ne comportant qu'une seule observation. Pour les estimateurs non contraints, il n'est pas possible d'obtenir des estimations fondées sur le modèle pour les domaines vides, et les intervalles de confiance ne sont pas fiables pour les domaines dans lesquels la taille de l'échantillon est petite. Pour l'estimateur contraint, les données provenant d'autres domaines sont utilisées pour les estimateurs des moyennes du domaine qui sont vides ou qui ont un petit nombre d'observations, ce qui permet d'obtenir des intervalles de confiance valides d'une longueur raisonnable. Les résultats des simulations répétées à $n = 60$ sont présentés dans la figure 4.4. Pour cette taille d'échantillon, il y a en moyenne trois observations par domaine, mais la proportion de couverture reste bonne.

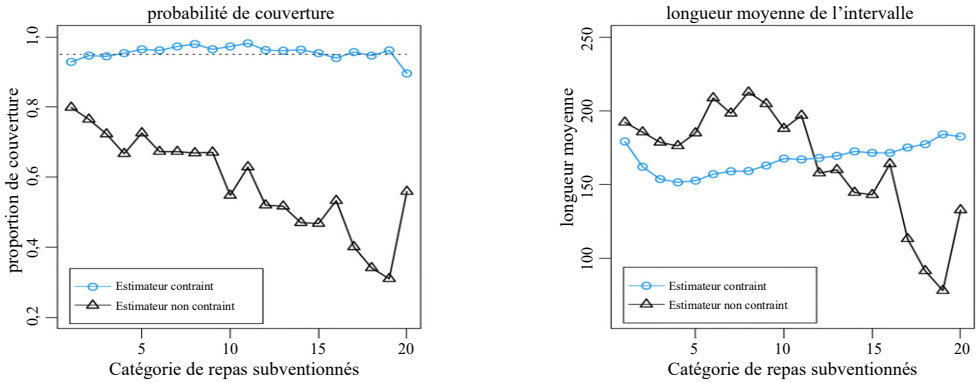
Pour donner un exemple de variable d'étude binaire, nous utilisons les données de l'étude de la National Health and Nutrition Examination Survey (NHANES), qui fournit des données sur la santé pour un échantillon de la population américaine et qui est accessible au public à l'adresse http://www.cdc.gov/nchs/nhanes.htm. Il y a $n = 1\,680$ observations comportant des enregistrements complets pour le taux de cholestérol, l'âge, la taille et le tour de taille pour les adultes âgés de 21 à 40 ans; nous les utiliserons pour démontrer un ordonnancement partiel en estimant la proportion de la population ayant un taux de cholestérol supérieur à 200, en fonction de l'âge, du tour de taille et du sexe. Le tour de taille est divisé par la taille afin de mesurer la circonférence relative, puis réparti dans quatre niveaux, le niveau 1 étant le plus petit et le niveau 4 le plus grand. Il y a 160 domaines qui représentent des combinaisons d'âge, de taille et de sexe pour une grille de domaines $20 \times 4 \times 2$. Le nombre de domaines est important pour cette taille d'échantillon, ce qui fait que de nombreux domaines ont moins de cinq observations. En l'absence d'hypothèses sur l'ordre, les domaines devraient être regroupés pour obtenir une estimation et une inférence fiables. Il est toutefois raisonnable de supposer que la probabilité d'avoir un taux de cholestérol élevé augmente avec l'âge et le tour de taille. Le sous-ensemble de données de la NHANES utilisé ici est inclus dans l'objet `nhdat` du paquet `csurvey`.

Figure 4.3 Moyennes par domaine et intervalles de confiance pour un échantillon stratifié dans l'ensemble de données `R api`, $n = 60$.



Notes : Les moyennes du domaine de la population sont représentées par des cercles et les valeurs de l'échantillon, par des losanges gris. La taille de l'échantillon est trop petite pour permettre une estimation fondée sur un modèle sans contraintes.

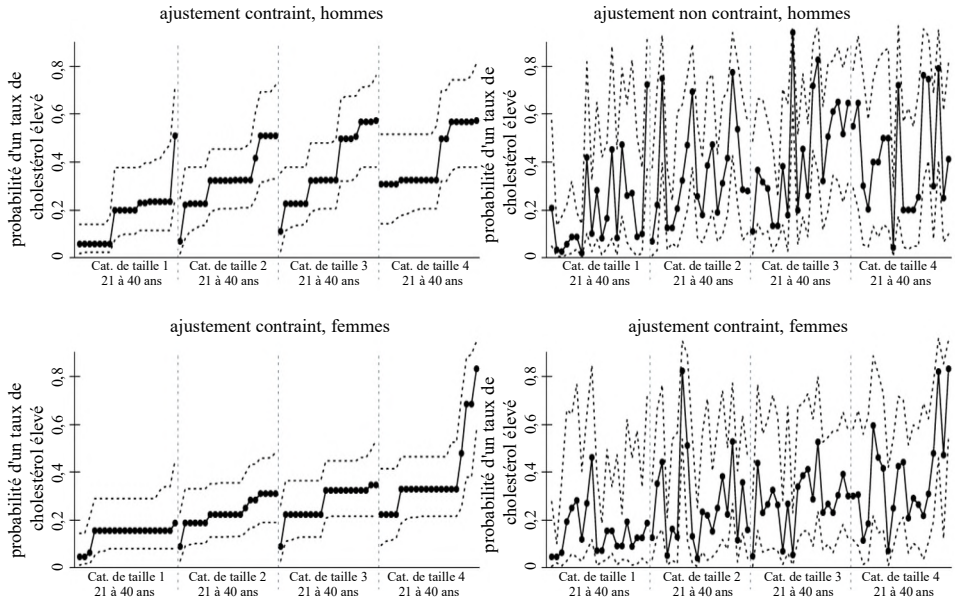
Figure 4.4 Probabilités de couverture et longueur des intervalles, pour 1 000 simulations à partir de l'ensemble de données `apipop` ayant une taille d'échantillon de $n = 60$.



Notes : Pour l'estimateur non contraint, la longueur moyenne et la probabilité de couverture sont calculées sur les seuls domaines non vides.

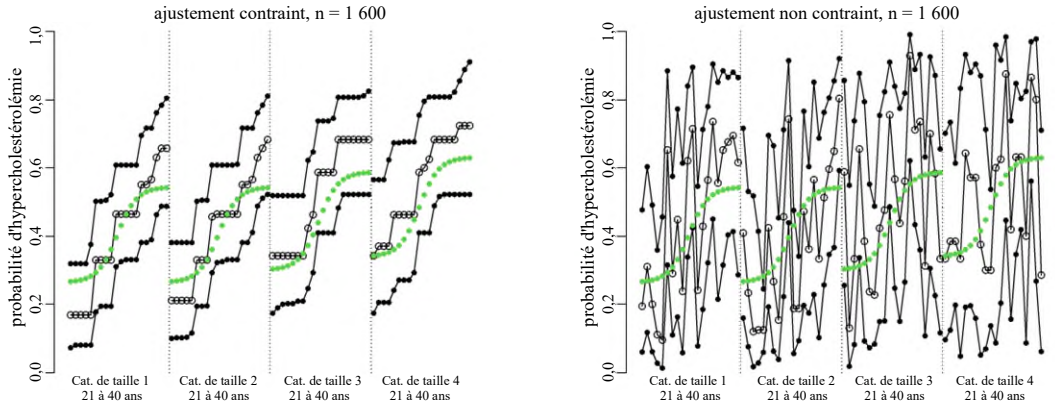
Les estimations et les intervalles de confiance à 95 % pour les 160 domaines sont présentés dans la figure 4.5, où l'on constate que les estimations contraintes sont plus stables et tendent à avoir des intervalles de confiance plus petits que les estimateurs de Hájek non contraints. (Nous n'avons pas inclus les estimateurs de Fay-Herriot à titre de comparaison parce qu'une réponse binaire n'est pas aussi simple à mettre en œuvre avec des échantillons de petite taille.) Bien que, dans ce cas, nous ne connaissions pas les proportions des domaines de la population, il est peu probable qu'elles augmentent et baissent au fur et à mesure que l'âge augmente, au sein d'une catégorie de taille. Les augmentations et les baisses observées pour l'estimateur non contraint sont probablement le résultat d'un hasard, en raison de la petite taille des échantillons dans les 160 domaines.

Figure 4.5 Estimations de la probabilité d'hypercholestérolémie pour $D = 160$ domaines, avec des intervalles de confiance à 95 % en fonction de l'âge, du tour de taille et du sexe, au moyen des données de la National Health and Nutrition Examination Survey formant un échantillon de $n = 1\,680$ observations.



Pour vérifier la méthode, nous avons procédé à des simulations afin de comparer les taux de couverture et les longueurs d'intervalle pour les estimations contraintes et non contraintes des estimations de probabilité. Nous avons 160 domaines comme dans l'exemple de la NHANES, et des tailles d'échantillon de $n =$ 1 600 et $n =$ 4 800. Nous précisons les probabilités réelles d'hypercholestérolémie comme il est indiqué (pour l'un des sexes) dans la figure 4.6 (pour l'autre sexe, les probabilités sont légèrement plus élevées). La taille de l'échantillon de $n =$ 1 600 est trop petite pour obtenir des estimations raisonnables pour l'estimateur non contraint traditionnel; les chercheurs regrouperaient les domaines pour obtenir des tailles d'échantillon plus importantes et des variances plus faibles.

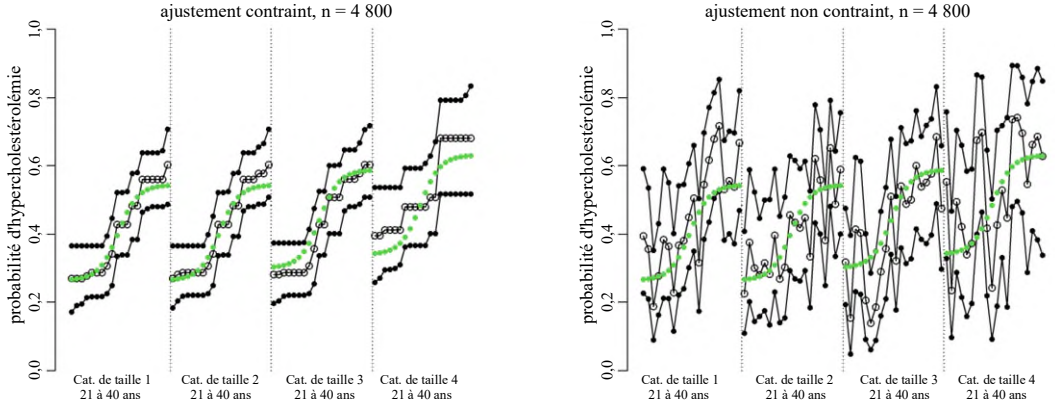
Figure 4.6 Données de la National Health and Nutrition Examination Survey simulées avec les probabilités d'hypercholestérolémie réelles représentées par les formes sigmoïdales.



Notes : Les estimations de gauche pour $D =$ 80 domaines sont contraintes d'augmenter avec le tour de taille et l'âge, tandis que les estimations de droite ne sont pas contraintes.

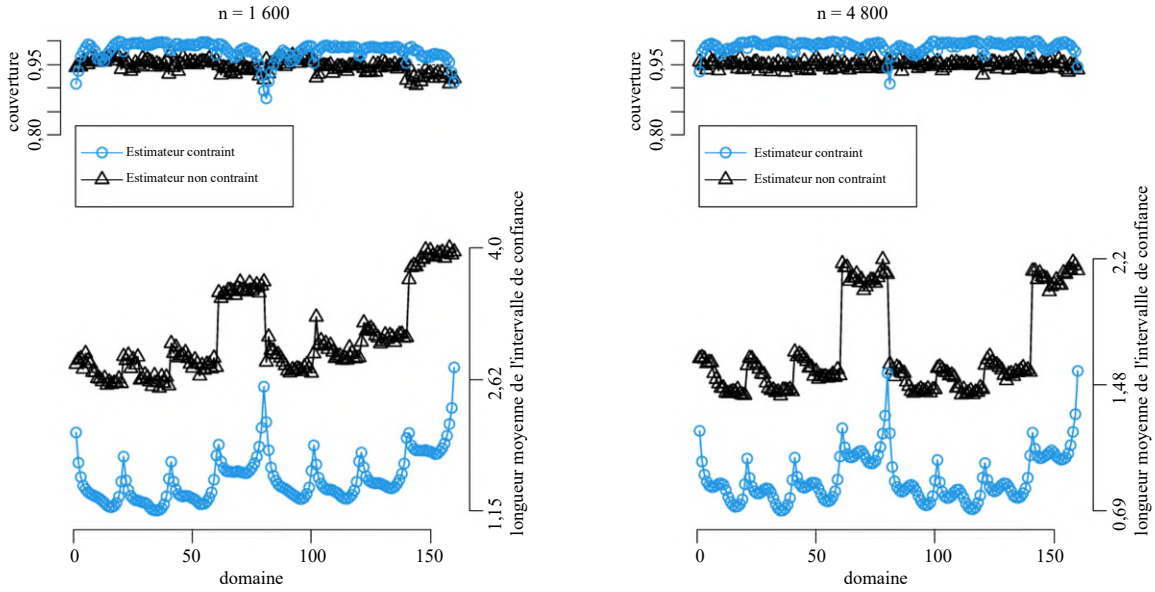
Dans le cas de l'ensemble de données simulées présenté dans la figure 4.7, la taille de l'échantillon est plus importante. Les variations ne sont pas aussi radicales que pour l'échantillon de plus petite taille, mais l'estimateur contraint produit toujours des estimations plus raisonnables et des intervalles de confiance plus petits. Pour la taille réduite de l'échantillon, de nombreuses tailles d'échantillon de domaine sont petites, trop petites pour que l'ajustement non contraint génère un grand volume de données. L'imposition de ces contraintes permet d'obtenir des estimations plus précises et des intervalles de confiance plus étroits. Les résultats des simulations de la figure 4.8 montrent les proportions de couverture et les longueurs des intervalles de confiance pour le logarithme du risque d'hypercholestérolémie dans chacun des 160 domaines; l'imposition des contraintes entraîne une réduction spectaculaire de la longueur des intervalles de confiance. L'estimateur contraint donne de meilleurs résultats pour les domaines qui se situent au « milieu » de l'ordonnancement partiel, c'est-à-dire si la moyenne du domaine est contrainte d'être à la fois inférieure à la moyenne principale de certains domaines et supérieure à celle d'autres domaines. Les domaines « périphériques » offrent une couverture légèrement inférieure à la cible et des intervalles de confiance plus larges.

Figure 4.7 Données de la National Health and Nutrition Examination Survey simulées avec les probabilités d'hypercholestérolémie réelles représentées par les formes sigmoïdales.



Notes : Les estimations de gauche sont contraintes d'augmenter avec le tour de taille et l'âge, tandis que les estimations de droite ne sont pas contraintes. Cet ensemble de données plus important permet d'obtenir des estimations plus précises.

Figure 4.8 Proportions de couverture et longueurs des intervalles de confiance pour les logarithmes du risque d'hypercholestérolémie, pour $D =$ 160 domaines, sur 1 000 ensembles de données simulées.



Notes : Dans le cas des estimations contraintes, les longueurs sont systématiquement plus petites.

5. Discussion

Nous avons introduit une nouvelle méthode pour l'estimation et l'inférence des moyennes de domaine fondées sur un modèle ayant une taille d'échantillon petite ou nulle, en supposant des contraintes d'inégalité *a priori*. L'estimation et l'inférence des moyennes de domaine à partir de données d'enquête peuvent être considérablement améliorées si l'on utilise des estimateurs fondés sur les ordres naturels et l'estimateur de

la variance de mélange; l'amélioration est d'autant plus importante que la taille de l'échantillon est petite. Ces estimateurs reposent sur des données provenant d'autres domaines dans le cadre d'une approche fondée sur un modèle. Les méthodes contraintes ont été introduites par Wu, Meyer et Opsomer (2016) et Oliva-Aviles, Meyer et Opsomer (2020), qui ont souligné que les hypothèses d'ordre étaient imposées à une « super-population » imaginaire ou à un mécanisme qui génère la population finie, de sorte que la population finie elle-même peut ne pas satisfaire exactement à l'ordre. Ils ont démontré par des simulations que si la population contient de petites déviations de l'ordre (comme dans l'exemple de l'école californienne), l'inférence est encore améliorée par rapport à l'estimateur non contraint. Nous avons perfectionné ces méthodes, en fournissant des estimations et des intervalles de confiance fiables pour des échantillons de petite taille ou des cellules vides. Les simulations montrent que les intervalles de confiance calculés à l'aide des méthodes proposées offrent une bonne couverture par rapport à l'estimateur de Hajék standard et à l'estimateur de Fay-Herriot et que la longueur des intervalles de confiance est inférieure à celle de l'estimateur de Hajék. Le paquet `csurvey` emploie ces méthodes et permet aux utilisateurs de préciser des ordonnancements sur des grilles de domaines et d'obtenir des estimations et des intervalles de confiance pour les moyennes des domaines de population. L'utilité de ces méthodes a été démontrée à l'aide d'ensembles de données d'enquête bien connus.

Remerciements

Ce travail a été partiellement financé par la subvention MMS 1533804 de la Fondation nationale des sciences.

Annexe

A. Code pour l'exemple des données sur les écoles californiennes

L'ensemble de données `api` du paquet `survey` contient des données sur les écoles primaires, intermédiaires et secondaires de Californie. L'unité est l'école et, pour cet exemple, nous nous intéressons à la moyenne des résultats des tests normalisés pour l'année 2000, `api00`. Nous nous attendons à ce que les scores moyens diminuent sur 20 niveaux de la variable `meals`, la proportion d'étudiants bénéficiant d'un repas gratuit ou à prix réduit.

```
mcat = apipop$meals
M = 20
for(i in 1:M){mcat[trunc(apipop$meals / 5) + 1 == i] = i}
mcat[mcat == 100] = M
mcat = as.factor(mcat)
```

À des fins de comparaison, nous calculons les véritables moyennes du domaine de la population :

```
tsc = 1:M
for(i in 1:M){tsc[i] = mean(apipop$api00[mcat == i])}
```

La variable `stype` indique le type d'école; nous choisirons un échantillon stratifié sur la base de cette variable. La variable `snum` est le numéro d'identification de l'école; le code suivant vise à choisir un échantillon aléatoire simple dans chaque type d'école, de 60 dans les écoles primaires et les écoles intermédiaires, respectivement et de 120 dans les écoles secondaires.

```
nsp = c(60, 60, 120)
es    =    sample(apipop$snum[apipop$stype    ==    "E"    &!is.na(apipop$avg.ed)
&!is.na(apipop$api00)], nsp[1])
ms    =    sample(apipop$snum[apipop$stype    ==    "M"    &!is.na(apipop$avg.ed)
&!is.na(apipop$api00)], nsp[2])
hs    =    sample(apipop$snum[apipop$stype    ==    "H"    &!is.na(apipop$avg.ed)
&!is.na(apipop$api00)], nsp[3]) sid = c(es, ms, hs)
```

Les pondérations de probabilité et la correction de la population finie sont calculées ensuit; 6 194 est le nombre total d'écoles dans la base de données; il y a 4 421 écoles élémentaires, 1 018 écoles intermédiaires et 755 écoles secondaires.

```
pw = 1:6194 * 0 + 4421 / nsp[1]
pw[apipop$stype == "M"] = 1018 / nsp[2]
pw[apipop$stype == "H"] = 755 / nsp[3]
fpc = 1:6194 * 0 + 4421
fpc[apipop$stype == "M"] = 1018
fpc[apipop$stype == "H"] = 755
```

Le plan est précisé à l'aide des fonctions `svydesign` et `as.svrepdesign` dans l'ensemble `survey`.

```
strsamp = cbind(apipop, mcat, pw, fpc)[sid, ]
dstrat = svydesign(ids = ~snum, strata = ~stype, fpc = ~fpc, data = strsamp,
weight = ~pw)
rds = as.svrepdesign(dstrat, type = "JKn")
```

Pour obtenir plus de renseignements sur la spécification du modèle, voir Lumley (2004), Lumley (2010) et Lumley (2023).

Pour obtenir l'estimation moyenne du domaine contraint proposée, nous utilisons la fonction `csvy` du paquet `csurvey`. Dans cet exemple, la fonction `decr` est utilisée pour contraindre les moyennes de domaine de `api00` à être décroissantes pour des valeurs plus importantes de `mcat`. Les arguments de la fonction `csvy` sont similaires à ceux requis par la fonction `svyglm` du paquet `R survey`. Un argument supplémentaire est `nD`, qui précise le nombre total de domaines dans un ensemble de données. L'utilisateur doit fournir cet argument pour que la fonction `csvy` puisse générer l'estimation et l'inférence pour les domaines vides.

```
ans = csvy(api00 ~ decr(mcat), design = rds, nD = M)
```

La valeur du critère d'information du cône (CIC) pour l'estimateur contraint et non contraint peut être extraite comme suit :

```
ans$CIC
ans$CIC.un
```

Un CIC plus petit indique une meilleure adéquation.

La fonction `confint` peut être utilisée pour extraire les intervalles de confiance pour les estimations de la moyenne du domaine à partir de l'objet `ans`. Les fonctions `svyby` et `svymean` du paquet `survey` sont utilisées pour obtenir l'estimation moyenne du domaine contraint ainsi que l'erreur-type.

```
cstr = confint(ans, level = 0.95, type = "link")
unc = svyby(formula = ~api00, by = ~mcat, design = rds, FUN = svymean, covmat =
TRUE)
```

La fonction `mseFH` du paquet `sae` est utilisée pour obtenir l'estimation de Fay-Herriot ainsi que l'erreur-type. Nous devons fournir l'estimateur non contraint et l'erreur-type de la fonction `svyby` du paquet `survey` comme valeurs d'entrée de la fonction `mseFH`.

```
mhatu = unc$y
seu = unc$se
ysae = mhatu
doms = expand.grid(1:10,1:5)
x1sae = doms[,1]
x2sae = doms[,2]
anss = mseFH(ysae ~ x1sae*factor(x2sae), vardir = seu^2)
```

La fonction `ebp` du paquet `emdi` est utilisée pour obtenir l'estimation de la meilleure prédiction empirique par mol10 ainsi que l'erreur-type de bootstrap paramétrique. Nous devons fournir un ensemble de données de population et un ensemble de données d'échantillon. Nous devons également préciser le nom d'une variable qui indique les domaines dans les données de la population et les données de l'échantillon.

```
emdi_model = ebp(fixed = y ~ x1*factor(x2), pop_data = pop,
      pop_domains = "domain", smp_data = sample.stsi, smp_domains = "domain",
      MSE = TRUE, seed = NULL, na.rm = TRUE)
```

B. Code pour l'exemple de données de la National Health and Nutrition Examination Survey

L'ensemble de données `nhdat` du paquet `csurvey` est un sous-ensemble des données recueillies dans le cadre de la National Health and Nutrition Examination Survey (NHANES), qui combine des interviews en personne et des examens physiques pour produire un ensemble de données complet à partir d'un échantillon probabiliste de résidents des États-Unis. Le paquet `nhdat` comprend des observations de 1 680 sujets. Nous utilisons cet ensemble de données pour estimer la probabilité qu'une personne ait un taux de cholestérol élevé en supposant que le taux de cholestérol moyen augmente avec l'âge et le tour de taille,

mais nous n'avons pas d'ordre de grandeur pour le sexe. La variable de réponse `chol` est codée 1 si le taux de cholestérol d'une personne est supérieur à 200 mg/dl et codée 0 dans les autres cas. L'âge est catégorisé et représenté par des valeurs entières comprises entre 21 et 40. Une autre variable `wcat` catégorise le rapport entre le tour de taille et la taille d'une personne. Elle comporte quatre catégories et les trois valeurs seuils sont 0,48, 0,55 et 0,66. Une autre covariable est le sexe `(gender)`, qui est codée 1 et 2 pour l'homme et la femme, respectivement.

Après avoir importé l'ensemble des données de `csurvey`, nous utilisons la fonction `svydesign` pour préciser un plan d'échantillonnage stratifié avec `str` comme strates :

```
library(csurvey)
data(nhdat)
dstrat = svydesign(ids = ~ id, strata = ~ str, data = nhdat, weight = ~ wt)
```

Pour obtenir l'estimateur contraint, nous utilisons la fonction symbolique `incr` deux fois, soit `incr*incr`, pour préciser que la moyenne du domaine du taux de cholestérol, c'est-à-dire la probabilité d'avoir un taux de cholestérol élevé, augmente à la fois dans la catégorie `age` et dans la catégorie `wcat`, et que les effets ne sont pas censés être additifs. Lorsque la réponse est binaire, nous précisons `family = quasibinomial(link = "logit")` dans `csvy`. Ici, nous utilisons `family = quasibinomial(link = "logit")` pour la même raison afin d'éviter un avertissement concernant les nombres non entiers de succès, ce qui est recommandé par l'auteur pour la fonction `svyglm` du paquet `survey`. Enfin, le nombre total de domaines sera $M = 160$ et nous devons le fournir pour estimer les domaines vides.

```
M = 160
ans = csvy(chol ~ incr(age) * incr(wcat) * gender, design = dstrat, nD = M,
family = quasibinomial(link = "logit"))
```

La valeur du CIC pour l'estimateur contraint et non contraint peut être extraite comme suit :

```
ans$CIC
ans$CIC.un
```

Un CIC plus petit indique une meilleure adéquation.

Pour prédire la probabilité qu'une personne présentant un ensemble de caractéristiques appartienne au groupe à fort taux de cholestérol, nous appelons la fonction `predict`. Les arguments sont similaires à ceux de la fonction `predict.glm`. Par exemple, si nous voulons prédire la probabilité d'une personne pour qui `age = 40`, `wcat = 4` et `gender = 2`, nous créons un nouveau cadre de données contenant ces caractéristiques et le conférons à la fonction `predict` sous la forme suivante :

```
pred.muhat = predict(ans, newdata = data.frame(age = 40, wcat = 4, gender = 2),
      type = "response", se.fit = FALSE)
```

Bibliographie

Brunk, H.D. (1958). Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, 29(2), 437-454.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(8), 1-19.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R*. New York: John Wiley & Sons, Inc.

Lumley, T. (2023). survey: Analysis of complex survey samples. *R package*.

Oliva-Aviles, C., Meyer, M.C. et Opsomer, J.D. (2019). Checking validity of monotone domain mean estimators. *Canadian Journal of Statistics*, 47(2), 315-331.

Oliva-Aviles, C., Meyer, M.C. et Opsomer, J.D. (2020). Estimation et inférence des moyennes de domaine soumises à des contraintes qualitatives. *Techniques d'enquête*, 46, 2, 155-191. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2020002/article/00002-fra.pdf.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28(1), 40-68.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation* (2nd ed.). Hoboken, New Jersey: Wiley.

Robertson, T., Wright, F.T. et Dykstra, R.L. (1988). *Order Restricted Statistical Inference*. New York: John Wiley & Sons, Inc.

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

Wu, J., Meyer, M.C. et Opsomer, J.D. (2016). Survey estimation of domain means that respect natural orderings. *Canadian Journal of Statistics*, 44(4), 431-444.

Xu, X., Meyer, M.C. et Opsomer, J.D. (2021). Improved variance estimation for inequality-constrained domain mean estimators using survey data. *Journal of Statistical Planning and Inference*, 215, 47-71.

Techniques d'enquête, décembre 2024
Vol. 50, nº 2, pp. 351-374
Statistique Canada, nº 12-001-X au catalogue

Une approche d'estimation sur petits domaines pour concilier les différences entre deux enquêtes sur l'effort de pêche récréative

Teng Liu, F. Jay Breidt et Jean D. Opsomer[1]

Résumé

De nombreuses études sont aux prises avec un problème de comparaison d'estimations obtenues à l'aide de différentes méthodologies d'enquête, notamment des différences de base de sondage, d'instruments de mesure et de modes d'exécution. L'enjeu se présente dans les enquêtes multimodales et les enquêtes remaniées. Un remaniement majeur des processus d'enquête pourrait avoir une incidence systématique sur les estimations d'enquêtes; il est donc important de quantifier et d'ajuster de telles discontinuités entre les plans de sondage pour assurer la comparabilité des estimations au fil du temps. Nous proposons une approche d'estimation sur petits domaines pour rapprocher deux ensembles d'estimations d'enquête et l'appliquons à deux enquêtes du Marine Recreational Information Program (MRIP), qui surveille la pêche récréative le long des côtes de l'océan Atlantique et du golfe du Mexique aux États-Unis. Nous développons un modèle log-normal pour les estimations issues des deux enquêtes, en tenant compte de la dynamique temporelle par régression sur la taille de la population et les facteurs saisonniers État-par-vague et en tenant partiellement compte des propriétés de couverture changeantes par régression sur la pénétration du téléphone sans fil. À l'aide des variances de plan de sondage estimées, nous développons un modèle de régression qui est analytiquement cohérent avec le modèle de moyenne log-normal. Nous utilisons les variances de plan de sondage modélisées dans une procédure d'estimation sur petits domaines de Fay-Herriot, afin d'obtenir les meilleurs prédicteurs linéaires sans biais empiriques des estimations rapprochées de l'effort de pêche (nécessitant des prédictions pour de nouveaux ensembles de covariables), et de fournir une approximation asymptotiquement valide de l'erreur quadratique moyenne.

Mots-clés : Approximation de l'EQM; erreur de couverture; erreur non due à l'échantillonnage; modèle Fay-Herriot; modèle log-normal; MPLSBE.

1. Introduction

Pendant des décennies, le National Marine Fisheries Service (NMFS) a mené des enquêtes auprès des ménages, afin de dénombrer le nombre de sorties de pêche récréative en mer depuis la rive et par des pêcheurs dans des bateaux privés dans 17 États des États-Unis le long des côtes de l'océan Atlantique et du golfe du Mexique : Alabama, Connecticut, Delaware, Floride, Géorgie, Louisiane, Maine, Maryland, Massachusetts, Mississippi, New Hampshire, New Jersey, New York, Caroline du Nord, Rhode Island, Caroline du Sud et Virginie. La collecte de données se déroule pendant deux semaines à la fin de chaque période d'échantillon de deux mois (ou « vague »), donnant six vagues chaque année. Toutefois, des échantillons ne sont pas obtenus pour chaque vague dans chaque État; par exemple, bon nombre d'États n'ont pas d'échantillon de vague 1, ce qui révèle un effort de pêche minime en janvier et février dans ces États.

Jusqu'en 2017, le NMFS utilisait la Coastal Household Telephone Survey (CHTS; enquête téléphonique auprès des ménages côtiers) pour recueillir des données sur les sorties de pêche. La base de sondage de la CHTS était une liste de ménages résidentiels à temps plein ayant un service téléphonique filaire dans un

1. Teng Liu, Colorado State University, Fort Collins, États-Unis. Courriel : tristan.tju@gmail.com; F. Jay Breidt, NORC, Université de Chicago, Chicago, États-Unis; Jean D. Opsomer, Westat, Inc., Rockville, États-Unis.

comté côtier. Le plan de sondage était un échantillonnage aléatoire simple stratifié par État et comté. La CHTS utilisait une composition aléatoire des numéros de téléphone filaire des ménages dans les comtés côtiers. La composition aléatoire présente plusieurs limites dans ce contexte, comme l'inefficacité de reconnaître les pêcheurs à la ligne (National Research Council, 2006), la baisse du taux de réponse des enquêtes téléphoniques (Curtin, Presser et Singer, 2005) et le sous-dénombrement des pêcheurs du fait de l'augmentation des ménages ayant uniquement un service téléphonique sans fil (Blumberg et Luke, 2013). Par conséquent, après expérimentation (Andrews, Brick et Mathiowetz, 2014), le NMFS a mis en place une nouvelle enquête en 2015, la Fishing Effort Survey (FES; enquête sur l'effort de pêche).

Contrairement à la CHTS, la FES est une enquête par la poste à remplir soi-même et utilise comme base de sondage un annuaire d'adresses postales dans les États côtiers (et non seulement les comtés côtiers) desservis par le US Postal Service. Ces dernières années, de nombreuses études ont suivi ce même cheminement, passant d'un mode téléphonique à un mode d'autodéclaration (voir Olson, Smyth, Horwitz, Keeter, Lesser, Marken, Mathiowetz, McCarthy, O'Brien, Opsomer, Steiger, Sterrett, Su, Suzer-Gurtekin, Turakhia et Wagner (2020) pour un examen récent). Le plan de sondage de la FES est un échantillonnage aléatoire simple stratifié par État, proximité de la côte et état de permis de pêche; ce statut étant déterminé par la correspondance des adresses avec le registre national des pêcheurs à la ligne en eau salée (National Saltwater Angler Registry). La CHTS a été abolie après 2017, de sorte que ces deux enquêtes se sont chevauchées en 2015-2017.

La CHTS téléphonique et la FES par la poste présentent des différences méthodologiques évidentes. Ces deux enquêtes ont des propriétés de couverture différentes, puisqu'elles utilisent des bases de sondage très différentes : composition aléatoire de numéros de téléphone filaire, pour la CHTS, et échantillonnage fondé sur l'adresse avec suréchantillonnage des adresses correspondant à des pêcheurs ayant un permis, pour la FES. Elles présentent des tendances de non-réponses différentes : les taux de réponse généraux de la FES étant près de trois fois supérieurs aux taux de réponse de la CHTS (Andrews et coll., 2014). Enfin, les processus de mesure sont fondamentalement différents, du fait de la différence entre poser des questions par téléphone sur l'activité de pêche comparativement à l'autodéclaration sur un formulaire papier.

En raison, au moins en partie, de ces différences méthodologiques, il existe de grands écarts entre les estimations des sorties de pêche de la CHTS et de la FES, celles de la FES étant systématiquement supérieures. Puisque nous pensons que ni l'une ni l'autre de ces enquêtes ne reflète exactement le nombre réel de sorties de pêche, quelle que soit la raison de l'écart, il est intéressant pour les gestionnaires des pêches et les scientifiques en évaluation des stocks de pouvoir convertir les « unités » des estimations de l'enquête téléphonique en celles de l'enquête par la poste, et vice versa. Cette conversion est appelée « calage » dans ce contexte et ne doit pas être confondue avec la méthodologie de calage courante dans les enquêtes complexes (Deville et Särndal, 1992). Ce calage permet de construire une série d'estimations comparables au fil du temps.

Les données utilisées pour l'exercice de calage proviennent de la CHTS pour la plupart des États et des vagues entre 1981 et 2017, et de la FES pour les États et les vagues de 2015 à 2018. Dans ce qui suit, nous travaillons sur l'échelle des logarithmes naturels du nombre de sorties de pêche et désignons ce chiffre

logarithmique un « effort ». Pour chaque enquête, les données portent sur l'effort total estimé de la pêche de rive et l'effort total estimé de la pêche en bateau privé, ainsi que sur les variances de plan de sondage et les tailles d'échantillon estimées pour chaque État et vague disponibles.

Comme nous le présentons ci-dessous, nous formulons le problème de calage comme une méthode d'estimation sur petits domaines au niveau du domaine, au sujet de laquelle nous passons brièvement en revue de la littérature connexe. Rao et Yu (1994) proposent un modèle d'estimation sur petits domaines faisant intervenir des effets aléatoires et des erreurs d'échantillonnage autorégressifs avec une matrice de covariance arbitraire connue utilisant à la fois des séries chronologiques et des données transversales. Datta, Lahiri, Maiti et Lu (1999) utilisent un modèle de cheminement aléatoire pour la composante temporelle, avec des erreurs d'échantillonnage corrélées. Pfeffermann et Tiller (2006) ajoutent des contraintes d'étalonnage au modèle spatial en fonction de l'État avec des mesures corrélées. Boonstra, van den Brakel et Das (2021) ainsi que Boonstra et van den Brakel (2022) développent des modèles hiérarchiques bayésiens pour les séries chronologiques multiniveaux dans de petits domaines. Feder (2001) examine diverses méthodes de séries chronologiques sur des données d'enquête transversales.

La littérature sur la combinaison d'enquêtes n'est pas aussi importante que celle portant sur les enquêtes uniques. Merkouris (2010) propose un estimateur assisté par un modèle en calant des estimations par domaine comparables provenant de diverses enquêtes non répétées partageant les mêmes domaines. Lohr et Brick (2012) adoptent une approche d'enquête à double base de sondage et explorent des méthodes d'estimations sur petits domaines pour deux enquêtes lorsqu'une peut présenter un biais. Manzi, Spiegelhalter, Turner, Flowers et Thompson (2011) proposent une série de modèles hiérarchiques bayésiens, afin de combiner des estimations de prévalence de multiples sources de données présentant des biais additifs. Wang, Holan, Nandram, Barboza, Toto et Anderson (2012) combinent trois enquêtes mesurées sur différentes bases temporelles, puis élaborent un modèle hiérarchique bayésien produisant une meilleure estimation du rendement de cultures, selon les hypothèses que l'une des trois enquêtes analysées ne présente pas de biais par rapport au rendement réel. Van den Brakel, Zhang et Tam (2020) examinent différentes méthodes pour mesurer des discontinuités dues à un remaniement de processus d'enquête, en classant ces méthodes selon que la présence d'une période de chevauchement entre l'ancienne et la nouvelle enquête existe, selon la durée d'une telle période et la façon dont se déroule la transition de l'ancienne enquête vers la nouvelle. En cas de collecte de données parallèle, dans le cadre de laquelle les données sont recueillies selon l'ancien et le nouveau plan de sondage en parallèle pendant une certaine période, des méthodes fondées sur le plan dans van den Brakel (2008, 2013), des modèles d'espace d'états dans van den Brakel (2008, 2010) et des modèles d'estimation sur petits domaines dans Pfeffermann (2002, 2013) ainsi que Rao et Molina (2015) peuvent être adoptées, en fonction de la durée de l'exécution parallèle et de la taille des échantillons. D'autres études connexes comprennent celle de Raghunathan, Xie, Schenker, Parsons, Davis, Dodd et Feuer (2007), qui combine les renseignements de deux enquêtes afin de corriger des enjeux de non-couverture et de non-réponse par l'adoption d'un modèle hiérarchique bayésien en supposant l'absence de biais pour l'une des enquêtes, et celle d'Erciulescu, Opsomer et Breidt (2021), qui établit un modèle hiérarchique bayésien afin de tenir compte des écarts entre deux séries d'estimations d'enquête et produit des estimations fiables à divers niveaux d'agrégation.

À la section 2.1, nous construisons un modèle qui suppose que les estimations des enquêtes par la poste et par téléphone présentent des « cibles » sous-jacentes d'intérêt dans le calage. Les deux séries de cibles incluent un modèle de séries chronologiques classique comprenant trois composantes : tendance, variations saisonnières et terme irrégulier. Cette spécification de modèle permet le calage en avançant et en reculant dans le temps. Pour une période passée, nous pouvons prédire ce qu'aurait été l'effort en « unités par la poste » à l'aide de l'estimation de l'enquête téléphonique antérieure afin de prédire la cible par la poste. De la même façon, pour une période future, nous pouvons prédire ce qu'aurait été l'effort en « unités par téléphone » en prédisant la cible par téléphone à l'aide de l'estimation par la poste.

À la section 2.3, nous montrons que le modèle combiné pour les deux ensembles d'estimations et les cibles sous-jacentes est un modèle linéaire mixte d'un type courant dans le contexte d'estimations sur petits domaines au niveau du domaine, connu comme étant le modèle Fay-Herriot (Fay et Herriot, 1979). Dans le modèle Fay-Herriot, la norme est de traiter les variances de plan de sondage comme étant connues. Nos variances de plan de sondage reposent sur des tailles d'échantillon modérées à grandes (taille minimale $n = 39$) dans chaque État et pour chaque vague, et sont donc correctement estimées par les normes d'une estimation sur petits domaines. Une complication est que les variances par rapport au plan original se situent sur l'échelle du nombre de sorties plutôt que sur l'échelle de l'effort (nombres logarithmiques de sorties). Comme solution de rechange à une linéarisation par série de Taylor classique, nous élaborons en annexe B une nouvelle approche de transformation des variances de plan de sondage estimées qui assure une cohérence analytique entre notre modèle moyen et notre modèle de variance.

La méthodologie de Fay-Herriot dans les sections 3.1-3.2 produit de meilleurs prédicteurs linéaires sans biais empiriques (MPLSBE) de la cible par la poste et de la cible par téléphone qui représentent nos séries d'effort calibrées. Contrairement au contexte Fay-Herriot standard, les MPLSBE nécessitent une prédiction pour les nouveaux ensembles de covariables. Aux sections 3.3-3.4, nous adaptons des approximations et des estimations standard d'erreur quadratique moyenne (EQM) à cette situation non standard et évaluons leur rendement au moyen de la simulation de la section 4.1. À la section 4.2, nous appliquons les méthodes au problème de la réconciliation des estimations de l'enquête téléphonique antérieure à celle de l'enquête par la poste, et concluons par une brève discussion à la section 5.

2. Modèle

2.1 Modèle moyen

Nous nous concentrons sur un type de comportement de pêche (soit de rive ou en bateau privé), car l'élaboration du modèle est similaire dans les deux cas. Supposons que $s = 1, 2, \ldots, 17$ représentent les États des États-Unis et $t = 1, 2, \ldots$ représentent le temps en vagues de deux mois commençant en janvier-février 1981. Nous supposons que l'estimation de l'effort téléphonique $\hat{T}_{st}$ est un estimateur sans biais par rapport au plan de la « cible par téléphone » T_{st}, qui comprend l'effort réel et les effets du mode d'enquête dus à la méthodologie de l'enquête par téléphone, alors que l'estimation de l'effort par la poste $\hat{M}_{st}$ est un estimateur

sans biais par rapport au plan de la « cible par la poste » M_{st}, qui comprend l'effort réel et les effets du mode d'enquête dus à la méthodologie de l'enquête par la poste.

Nous supposons que la cible par téléphone et la cible par la poste comprennent les séries d'effort réel, que l'on suppose ensuite comprendre les tendances propres aux États, du fait, en partie, de la variation des tailles de population des États, des effets saisonniers propres aux États variant d'une vague à l'autre et de termes irréguliers qui sont des effets idiosyncratiques que la tendance régulière ou les tendances saisonnières n'expliquent pas. Nous modélisons les tendances propres aux États à l'aide d'estimations annuelles de la taille de la population au niveau des États provenant du Bureau du recensement des États-Unis (2016) sur une échelle logarithmique. Nous modélisons une tendance saisonnière générale au moyen d'indicateurs pour les vagues de deux mois et cette tendance saisonnière peut varier d'un État à l'autre. Les termes irréguliers restants, désignés par $\{\nu_{st}\}$ ci-dessous, représentent la variation réelle, que n'expliquent pas la tendance régulière et la tendance saisonnière, et que nous modélisons comme des variables aléatoires distribuées de manière indépendante et identique (iid) de variance moyenne nulle et inconnue, ψ.

Les effets du mode d'enquête présents dans les cibles par téléphone et par la poste sont des erreurs non dues à l'échantillonnage, notamment des biais potentiels dus à une erreur de couverture (population $\neq$ base de sondage), une erreur de non-réponse (échantillon $\neq$ répondants) et une erreur de mesure (réponses réelles $\neq$ réponses mesurées). Ces effets peuvent présenter leurs propres tendances et saisonnalités : par exemple, en raison de changements dans la qualité de la base de sondage au fil du temps, des variations des taux de réponse au fil des années ou des vagues, des variations de la mise en œuvre des protocoles de mesure au fil du temps, etc. Ces erreurs non dues à l'échantillonnage ne peuvent donc pas être entièrement démêlées de la série de l'effort réel (enjeu que présentent toutes les enquêtes).

Compte tenu des covariables pertinentes expliquant la variation de l'erreur de mesure, de l'erreur de non-réponse ou de l'erreur de couverture au fil du temps, les erreurs non dues à l'échantillonnage peuvent être modélisées et supprimées. Le changement de la proportion de ménages ayant uniquement un service téléphonique sans fil est une covariable potentielle d'explication des variations de l'erreur de couverture au fil du temps de l'enquête par téléphone filaire uniquement. En annexe A, nous décrivons comment nous avons créé un ensemble de proportions prédites de ménages ayant uniquement un service téléphonique sans fil, $\{w_{st}\}$, pour chaque État et vague au sein de nos données.

Les tendances ou les variations saisonnières pourraient contenir des effets du mode d'enquête. Par conséquent, nous laissons la possibilité que les tendances et variations saisonnières soient différentes pour les enquêtes par la poste comparativement aux enquêtes par téléphone; nous laissons, en particulier, la possibilité que les tendances et les variations saisonnières varient selon le degré de service sans fil.

Notre modèle combiné pour l'effort (logarithmes naturels du nombre de sorties de pêche) suppose donc que :

$$\begin{aligned}
&\text{estimation de l'effort pour l'enquête par téléphone} && = \hat{T}_{st} = T_{st} + e_{Tst} \\
&\text{cible par téléphone} && = T_{st} = \mathbf{x}_{Tst}^{\mathsf{T}}\boldsymbol{\beta} + \nu_{st} \\
&\text{estimation de l'effort pour l'enquête par la poste} && = \hat{M}_{st} = M_{st} + e_{Mst} \\
&\text{cible par la poste} && = M_{st} = \mathbf{x}_{Mst}^{\mathsf{T}}\boldsymbol{\beta} + \nu_{st}, \qquad (2.1)
\end{aligned}$$

où

- $\boldsymbol{\beta}$ est un vecteur aux coefficients de régression inconnus;
- les erreurs d'échantillonnage $\{e_{Tst}\}$ sont des variables aléatoires indépendantes $\mathcal{N}\left(0, \sigma_{Tst}^2\right)$ où les variances de plan de sondage σ_{Tst}^2 sont connues;
- les erreurs d'échantillonnage $\{e_{Mst}\}$ sont des variables aléatoires indépendantes $\mathcal{N}\left(0, \sigma_{Mst}^2\right)$ où les variances de plan de sondage σ_{Mst}^2 sont connues;
- les termes irréguliers $\{\nu_{st}\}$, représentant la variation réelle que n'expliquent pas la tendance régulière et la tendance saisonnière, sont des variables aléatoires distribuées de manière indépendante et identique (iid) $\mathcal{N}(0, \psi)$ de variance inconnue ψ;
- $\{e_{Tst}\}$, $\{e_{Mst}\}$ et $\{\nu_{st}\}$ sont mutuellement indépendants.

L'hypothèse d'indépendance des erreurs d'échantillonnage est justifiée, car l'échantillon est stratifié et des échantillons indépendants sont prélevés État par État et vague par vague. L'hypothèse de normalité est justifiée par des effets limitants centraux d'échantillons stratifiés de taille moyenne à grande pour chaque État et chaque vague (comme mentionné précédemment, la taille d'échantillon minimale est 39). De plus, nous supposons que les erreurs d'échantillonnage $\{e_{Tst}\}$ et $\{e_{Mst}\}$ sont indépendantes l'une de l'autre, car la sélection et la conduite des enquêtes par la poste et par téléphone sont indépendantes. Nous utilisons une simulation pour évaluer la sensibilité de certaines de nos méthodes à l'hypothèse de normalité sur les effets aléatoires à la section 4.1 ci-dessous. Les variances de plan de sondage $\{\sigma_{Tst}^2\}$ et $\{\sigma_{Mst}^2\}$ correspondent aux estimations d'effort (logarithmes naturels des estimations du nombre de sorties), alors que les estimations de variances de plan de sondage disponibles $\{\hat{V}_{Tst}\}$ et $\{\hat{V}_{Mst}\}$ correspondent aux estimations du nombre de sorties; nous étudions cet écart à la section 2.2 ci-dessous.

Supposons que $\nu_{st}, \nu_{st}^T, \nu_{st}^M$ désignent des effets aléatoires indépendants de moyenne nulle, où ν_{st} est mesurée dans les deux enquêtes et représente la variation réelle non expliquée par les covariables, alors que ν_{st}^T, ν_{st}^M désignent les effets aléatoires propres au mode. Nous avons tenu compte de diverses spécifications d'effets aléatoires, notamment : a) les deux enquêtes mesurent la variation réelle (ν_{st}); b) l'enquête par la poste mesure la réalité (ν_{st}), alors que l'enquête téléphonique mesure la réalité plus l'effet aléatoire de l'enquête par téléphone $(\nu_{st} + \nu_{st}^T)$; c) l'enquête téléphonique mesure la réalité (ν_{st}), alors que l'enquête par la poste mesure la réalité plus l'effet aléatoire de l'enquête par la poste $(\nu_{st} + \nu_{st}^M)$; d) chaque enquête mesure la réalité plus son propre effet aléatoire propre au mode, $(\nu_{st} + \nu_{st}^T, \nu_{st} + \nu_{st}^M)$; e) les deux enquêtes ont leur propre effet aléatoire propre au mode, indépendamment l'une de l'autre, sans variation réelle (hors des effets fixes) mesurée par l'une des deux valeurs (ν_{st}^T, ν_{st}^M). L'indépendance est présumée dans les spécifications b) et e), car tout effet corrélé dans les deux mesures devrait être la variation réelle et non une erreur de mesure.

Les deux modèles b) et c) nécessitent que la spécification d'un modèle serve de « référence », ce qui ne correspond pas à l'approche adoptée dans la présente analyse. Nous ne poursuivons donc pas l'étude de ces modèles. Le modèle d) est scientifiquement plausible, mais il s'agit du modèle le plus vaste pris en compte et il nécessite un logiciel d'estimation personnalisé pour tenir compte des deux observations bivariées au cours des périodes de chevauchement et des observations univariées au cours des périodes qui ne se chevauchent pas. Dans le cadre d'une analyse exploratoire non présentée ici, nous avons trouvé qu'il était

difficile d'ajuster ce modèle à nos données de chevauchement limitées. Le modèle e) indique que tout effet aléatoire est purement une erreur de mesure propre au mode, non lié au phénomène sous-jacent réel. Malgré cette possibilité, l'incidence d'un tel modèle est que seuls les effets fixes sont d'intérêt pour la prédiction, et même si les effets aléatoires de moyenne nulle sont mal spécifiés, nous obtiendrions tout de même des estimations raisonnables des effets fixes. Par conséquent, nous avons choisi le modèle a), dans le cadre duquel les deux enquêtes mesurent une valeur ν_{st} commune qui représente la variation réelle. Il s'agit d'une spécification standard d'une estimation sur petits domaines et cela nous permet d'obtenir des estimations à l'aide de logiciels disponibles dans le commerce pour une modélisation Fay-Herriot univariée, comme nous le décrivons ci-dessous.

Les erreurs d'échantillonnage dans notre application sont indépendantes en raison de la stratification du plan de sondage. Les effets aléatoires $\{\nu_{st}\}$ de notre spécification n'incluent pas d'autocorrélation spatiale ou temporelle. L'échelle spatiotemporelle de nos données est l'état par vague, que nous modélisons à l'aide d'effets fixes. Toute autocorrélation spatiale ou temporelle résiduelle devrait être réduite; l'autocorrélation temporelle étant susceptible d'être plus importante si une autocorrélation existe. Toutefois, nos diagnostics (voir la section 4.2) n'ont pas relevé de justification d'une autocorrélation temporelle résiduelle.

Supposons que $\boldsymbol{\beta}^\intercal = \left[\boldsymbol{\alpha}^\intercal, \boldsymbol{\mu}^\intercal, \boldsymbol{\gamma}^\intercal\right]$, alors les effets fixes $\mathbf{x}_{Tst}^\intercal\boldsymbol{\beta}$ et $\mathbf{x}_{Mst}^\intercal\boldsymbol{\beta}$ peuvent être chacun décomposés en trois composantes :

$$\begin{aligned}\mathbf{x}_{Tst}^\intercal\boldsymbol{\beta} &= \left[\mathbf{a}_{st}^\intercal, 0\cdot\mathbf{b}_{st}^\intercal, w_{st}\mathbf{c}_{st}^\intercal\right]\boldsymbol{\beta}, = \mathbf{a}_{st}^\intercal\boldsymbol{\alpha} + 0\cdot\mathbf{b}_{st}^\intercal\boldsymbol{\mu} + w_{st}\mathbf{c}_{st}^\intercal\boldsymbol{\gamma} \\ \mathbf{x}_{Mst}^\intercal\boldsymbol{\beta} &= \left[\mathbf{a}_{st}^\intercal, 1\cdot\mathbf{b}_{st}^\intercal, 0\cdot\mathbf{c}_{st}^\intercal\right]\boldsymbol{\beta} = \mathbf{a}_{st}^\intercal\boldsymbol{\alpha} + 1\cdot\mathbf{b}_{st}^\intercal\boldsymbol{\mu} + 0\cdot\mathbf{c}_{st}^\intercal\boldsymbol{\gamma},\end{aligned} \tag{2.2}$$

où le vecteur de covariable connue $\mathbf{a}_{st}$ comprend l'ordonnée à l'origine, le logarithme (taille de la population de l'État), les indicateurs d'État, les indicateurs de vague et les interactions État par logarithme (population) et État par vague. Dans notre application, les vecteurs de covariable $\mathbf{b}_{st}$ et $\mathbf{c}_{st}$ sont des sous-vecteurs de $\mathbf{a}_{st}$, en raison de la parcimonie; des détails sont fournis à la section 4.2. Par conséquent, $\mathbf{a}_{st}^\intercal\boldsymbol{\alpha}$ désigne la variation de la tendance et la variation saisonnière propre à l'État pour les données de l'enquête téléphonique, $\mathbf{a}_{st}^\intercal\boldsymbol{\alpha} + \mathbf{b}_{st}^\intercal\boldsymbol{\mu}$ désigne la variation de la tendance et la variation saisonnière propre à l'État pour les données de l'enquête par la poste, le terme d'interaction du service sans fil $w_{st}\mathbf{c}_{st}^\intercal\boldsymbol{\gamma}$ modélise l'incidence de la pénétration du service de téléphone sans fil sur la variation de la tendance et la variation saisonnière de l'enquête téléphonique sans incidence sur l'enquête par la poste. Il est facile de vérifier que tous les paramètres de ce modèle sont définis et de reconnaître des estimations sans biais dans les données disponibles.

2.2 Modèle de variance du plan de sondage

Selon les modèles de l'effort mentionnés en (2.1), les variances des erreurs d'échantillonnage par rapport à l'échelle d'origine du nombre de sorties non transformé peuvent être dérivées d'une distribution log-normale sous la forme :

$$V_{Tst} = \mathrm{Var}\left(\exp(\hat{T}_{st}) \mid T_{st}\right) = \left\{\exp(\sigma_{Tst}^2) - 1\right\}\exp\left\{2T_{st} + \sigma_{Tst}^2\right\} \tag{2.3}$$

et

$$V_{Mst} = \mathrm{Var}\left(\exp(\hat{M}_{st}) \mid M_{st}\right) = \left\{\exp(\sigma^2_{Mst}) - 1\right\} \exp\left\{2M_{st} + \sigma^2_{Mst}\right\}. \tag{2.4}$$

Nous devons estimer σ^2_{Tst} et σ^2_{Mst} (variances du plan de sondage par rapport à l'échelle de l'effort ou nombres logarithmiques de sorties), en intégrant les estimations approximativement sans biais par rapport au plan, respectivement $\hat{V}_{Tst}$ et $\hat{V}_{Mst}$ de V_{Tst} et V_{Mst}.

Nous suivons une approche étroitement liée à l'estimation de la fonction de variance généralisée (par exemple chapitre 7 de Wolter (2007)) en élaborant des modèles de régression pour les logarithmes du CV^2 (coefficients de variation au carré) empirique et en utilisant ces modèles ajustés pour produire des estimations de variances de plan de sondage sur l'échelle logarithmique, σ^2_{Tst} et σ^2_{Mst}, qui assurent la cohérence analytique entre le modèle moyen et le modèle de variance. Des détails sont fournis en annexe B. La taille des échantillons au sein des États et des vagues est grande dans notre application, nous traitons donc ces estimations comme étant fixes et connues dans ce qui suit, selon la norme des techniques d'estimation sur petits domaines que nous appliquons dans les sections suivantes.

2.3 Modèle d'estimation sur petits domaines Fay-Herriot

Soit,

$$\mathbf{x}^{\mathsf{T}}_{st} = \begin{cases} \mathbf{x}^{\mathsf{T}}_{Tst}, & \text{en l'absence d'estimation par la poste;} \\ \mathbf{x}^{\mathsf{T}}_{Mst}, & \text{en l'absence d'estimation par téléphone;} \\ (\mathbf{x}_{Tst} + \mathbf{x}_{Mst})^{\mathsf{T}} / 2, & \text{autrement.} \end{cases}$$

Alors, il est commode d'écrire :

$$\begin{aligned} Y_{st} &= \begin{cases} \hat{T}_{st}, & \text{en l'absence d'estimation par la poste;} \\ \hat{M}_{st}, & \text{en l'absence d'estimation par téléphone;} \\ \left(\hat{T}_{st} + \hat{M}_{st}\right)/2, & \text{autrement.} \end{cases} \\ &= \begin{cases} \mathbf{x}^{\mathsf{T}}_{Tst}\boldsymbol{\beta} + \nu_{st} + e_{Tst}, & \text{en l'absence d'estimation par la poste;} \\ \mathbf{x}^{\mathsf{T}}_{Mst}\boldsymbol{\beta} + \nu_{st} + e_{Mst}, & \text{en l'absence d'estimation par téléphone;} \\ (\mathbf{x}_{Tst} + \mathbf{x}_{Mst})^{\mathsf{T}}\boldsymbol{\beta}/2 + \nu_{st} + (e_{Tst} + e_{Mst})/2, & \text{autrement.} \end{cases} \\ &= \mathbf{x}^{\mathsf{T}}_{st}\boldsymbol{\beta} + \nu_{st} + e_{st}. \end{aligned} \tag{2.5}$$

Ce modèle suit alors exactement la structure de modèle linéaire mixte de Fay et Herriot (1979), selon lequel les estimations directes Y_{st} sont égales au modèle de régression, plus l'effet aléatoire ν_{st}, plus l'erreur d'échantillonnage à variance de plan de sondage « connue », exprimée sous la forme :

$$D_{st} = \begin{cases} \sigma^2_{Tst}, & \text{en l'absence d'estimation par la poste;} \\ \sigma^2_{Mst}, & \text{en l'absence d'estimation par téléphone;} \\ \dfrac{1}{4}\left(\sigma^2_{Tst} + \sigma^2_{Mst}\right), & \text{autrement.} \end{cases}$$

Il peut sembler naturel d'utiliser une combinaison convexe autre que $(1/2, 1/2)$ pour exprimer des variances inégales dans les deux sources de données. L'établissement d'une moyenne simple pourrait

entraîner la perte non négligeable de renseignements pour la prédiction de ν_{st}; cependant, notre objectif étant le calage, la prédiction de ν_{st} n'est pas requise au cours de la période de chevauchement (nous disposons des observations des enquêtes par téléphone et par la poste et n'avons pas besoin de conversion d'unité). La seule contribution de ces observations de chevauchement est par conséquent pour l'estimation des paramètres de régression $\boldsymbol{\beta}$ et de variance de l'effet aléatoire ψ. L'établissement d'une moyenne des résultats des estimations des enquêtes par téléphone et par la poste entraîne une faible perte de renseignements pour l'estimation paramétrique, puisque nous remplaçons deux observations corrélées par une seule; cependant, cette approche simple permet l'utilisation d'un logiciel standard pour calculer l'estimation.

3. Méthodes

3.1 Estimation pour le modèle Fay-Herriot

Définissons $\mathcal{A} = \{(s,t) : Y_{st} \text{ ne manque pas}\}$ comme étant l'ensemble de tous les États par combinaisons année-vague pour lesquelles nous disposons d'une estimation de l'une ou l'autre des enquêtes. Supposons que m désigne la taille de l'ensemble $\mathcal{A}$. Définissons $\mathbf{X} \coloneqq [\mathbf{x}_{st}^{\mathsf{T}}]_{(s,t)\in\mathcal{A}}$, $\mathbf{Y} \coloneqq [Y_{st}]_{(s,t)\in\mathcal{A}}$. Nous obtenons

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + [\nu_{st}]_{(s,t)\in\mathcal{A}} + [e_{st}]_{(s,t)\in\mathcal{A}}.$$

Alors $\boldsymbol{\Sigma}(\psi) \coloneqq \mathrm{Var}(\mathbf{Y}) = \mathrm{diag}\{\psi + D_{st}\}_{(s,t)\in\mathcal{A}}$. Si ψ était connue, le meilleur estimateur linéaire sans biais (MELSB) de $\boldsymbol{\beta}$ serait

$$\tilde{\boldsymbol{\beta}}_{\psi} = \left\{\mathbf{X}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\psi)\,\mathbf{X}\right\}^{-1}\mathbf{X}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\psi)\,\mathbf{Y}. \tag{3.1}$$

Puisque ψ n'est pas connue, nous la remplaçons par un estimateur convergent pour obtenir

$$\hat{\boldsymbol{\beta}} = \left\{\mathbf{X}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\hat{\psi})\,\mathbf{X}\right\}^{-1}\mathbf{X}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\hat{\psi})\,\mathbf{Y}. \tag{3.2}$$

Nous utiliserons l'estimation du maximum de vraisemblance restreint (REML) $\hat{\psi}$, sauf indication contraire.

3.2 Prédiction

Dans le contexte de modèle Fay-Herriot classique, il est intéressant de prédire

$$\mathbf{x}_{st}^{\mathsf{T}}\boldsymbol{\beta} + \nu_{st}$$

à partir de (2.5). Dans notre contexte, nous cherchons toutefois à prédire

$$\phi_{st} = \mathbf{z}_{st}^{\mathsf{T}}\boldsymbol{\beta} + \nu_{st}, \tag{3.3}$$

où $\mathbf{z}_{st}$ peut ne pas être égal à $\mathbf{x}_{st}$. Comme nous l'avons mentionné dans l'introduction, la conversion des « unités » d'un mode en unités d'un autre mode présente un intérêt. Pour convertir l'estimation d'une enquête téléphonique antérieure en unités d'une enquête par la poste, nous pouvons utiliser

$$\mathbf{z}_{st}^{\mathsf{T}} = \mathbf{x}_{Mst}^{\mathsf{T}} = \left[\mathbf{a}_{st}^{\mathsf{T}}, \mathbf{b}_{st}^{\mathsf{T}}, \mathbf{0}^{\mathsf{T}}\right]$$

afin de prédire la cible par la poste M_{st}. Pour convertir l'estimation d'une future enquête par la poste en unités d'une enquête historique par téléphone, nous pouvons utiliser

$$\mathbf{z}_{st}^{\mathsf{T}} = \left[\mathbf{a}_{st}^{\mathsf{T}}, \mathbf{0}^{\mathsf{T}}, \mathbf{0}^{\mathsf{T}}\right]$$

afin de prédire la cible par téléphone, corrigée pour tenir compte de l'effet du service sans fil : $T_{st} - w_{st}\mathbf{c}_{st}^{\mathsf{T}}\boldsymbol{\gamma} = \mathbf{a}_{st}^{\mathsf{T}}\boldsymbol{\alpha} + \nu_{st}$.

Supposons que $\boldsymbol{\lambda}_{st}$ désigne un vecteur $m \times 1$ avec 1 en position $(s,t)^{\text{e}}$ et 0 ailleurs. Dans un contexte de normalité, il est bien connu que le meilleur prédicteur quadratique moyen de ϕ_{st} dans (3.3) est

$$\phi_{st}\left(\boldsymbol{\beta}, \psi\right) = \mathbf{z}_{st}^{\mathsf{T}}\boldsymbol{\beta} + \psi\boldsymbol{\lambda}_{st}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\psi)\left(\mathbf{Y} - X\boldsymbol{\beta}\right), \tag{3.4}$$

ce qui est uniquement possible si $\boldsymbol{\beta}$ et ψ sont connus. Si seule ψ est connue, le meilleur prédicteur linéaire sans biais (MPLSB)

$$\phi_{st}\left(\tilde{\boldsymbol{\beta}}_{\psi}, \psi\right) = \mathbf{z}_{st}^{\mathsf{T}}\tilde{\boldsymbol{\beta}}_{\psi} + \psi\boldsymbol{\lambda}_{st}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\psi)\left(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_{\psi}\right) \tag{3.5}$$

est obtenu en intégrant le MELSB de (3.1) dans (3.4). Enfin, si ni $\boldsymbol{\beta}$ ni ψ ne sont connus, le meilleur prédicteur linéaire sans biais empirique (MPLSBE) peut être obtenu en substituant un estimateur convergent de ψ dans (3.5) :

$$\phi_{st}\left(\hat{\boldsymbol{\beta}}, \hat{\psi}\right) = \mathbf{z}_{st}^{\mathsf{T}}\hat{\boldsymbol{\beta}} + \hat{\psi}\,\boldsymbol{\lambda}_{st}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\hat{\psi})\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right), \tag{3.6}$$

où $\hat{\boldsymbol{\beta}}$ est donné par la formule (3.2). Ces MPLSBE sont les valeurs calibrées proposées de l'échelle logarithmique.

3.3 Approximation de l'erreur quadratique moyenne

L'approximation de l'erreur quadratique moyenne a été abondamment étudiée; voir Jiang et Lahiri (2006) pour un excellent examen. Notre approche de prédiction est légèrement atypique, car notre prédiction porte sur de nouveaux ensembles de covariables lors de la conversion de nos estimations d'« unités » d'un mode aux unités d'un autre mode. On peut montrer que

$$\begin{aligned}
\text{EQM}\left\{\phi_{st}\left(\hat{\boldsymbol{\beta}}, \hat{\psi}\right)\right\} &= E\left[\left\{\phi_{st}\left(\hat{\boldsymbol{\beta}}, \hat{\psi}\right) - \phi_{st}\right\}^2\right] \\
&= E\left[\left\{\phi_{st}\left(\boldsymbol{\beta}, \psi\right) - \phi_{st}\right\}^2\right] + E\left[\left\{\phi_{st}\left(\tilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \phi_{st}\left(\boldsymbol{\beta}, \psi\right)\right\}^2\right] \\
&\quad + E\left[\left\{\phi_{st}\left(\hat{\boldsymbol{\beta}}, \hat{\psi}\right) - \phi_{st}\left(\tilde{\boldsymbol{\beta}}_{\psi}, \psi\right)\right\}^2\right] \\
&= \dot{g}_{1st}(\psi) + \dot{g}_{2st}(\psi) + \dot{g}_{3st}(\psi) + o\left(m^{-1}\right),
\end{aligned} \tag{3.7}$$

où m est le nombre de petits domaines

$$\dot{g}_{1st}(\psi) = \frac{\psi D_{st}}{\psi + D_{st}},$$

$$\dot{g}_{2st}(\psi) = \left(\frac{\psi(\mathbf{z}_{st} - \mathbf{x}_{st})^{\mathsf{T}} + D_{st}\mathbf{z}_{st}^{\mathsf{T}}}{\psi + D_{st}} \right) \left[\sum_{u \in \mathcal{A}} (\psi + D_u)^{-1} \mathbf{x}_u \mathbf{x}_u^{\mathsf{T}} \right]^{-1}$$

$$\times \left(\frac{\psi(\mathbf{z}_{st} - \mathbf{x}_{st})^{\mathsf{T}} + D_{st}\mathbf{z}_{st}^{\mathsf{T}}}{\psi + D_{st}} \right)^{\mathsf{T}},$$

et

$$\dot{g}_{3st}(\psi) = \frac{2D_{st}^2}{(\psi + D_{st})^3} \frac{1}{\sum_{u \in \mathcal{A}} (\psi + D_u)^{-2}}.$$

Les termes de cette approximation d'EQM peuvent être obtenus des résultats de la section 5.3 de Rao et Molina (2015).

3.4 Estimation de l'erreur quadratique moyenne

Comme à la section 5.3 de Rao et Molina (2015), un estimateur approximativement sans biais de l'approximation de l'EQM dans (3.7) est donné par

$$\text{eqm}\left\{\phi_{st}\left(\hat{\boldsymbol{\beta}}, \hat{\psi}\right)\right\} = \dot{g}_{1st}(\hat{\psi}) + \dot{g}_{2st}(\hat{\psi}) + 2\dot{g}_{3st}(\hat{\psi}). \tag{3.8}$$

Nous évaluons la qualité de l'approximation asymptotique (3.7) et de son estimateur (3.8) par une simulation à la section 4.1.

3.5 Prédiction sur l'échelle d'origine

Pour calculer les prédicteurs sur l'échelle d'origine, nous procédons à une rétro-transformation en faisant une exponentiation du MPLSBE à partir de (3.6) et corrigeons pour la non-linéarité de la rétro-transformation à l'aide de l'EQM estimée dans (3.8) :

$$\widehat{\exp(\phi_{st})} = \exp\left[\phi_{st}\left(\hat{\boldsymbol{\beta}}, \hat{\psi}\right) + \frac{1}{2}\text{eqm}\left\{\phi_{st}\left(\hat{\boldsymbol{\beta}}, \hat{\psi}\right)\right\}\right], \tag{3.9}$$

qui est un estimateur du meilleur prédicteur quadratique moyen selon le modèle normal, et une correction normalisée même sans l'hypothèse de normalité. La rétro-transformation de Slud et Maiti (2006) utilise le terme principal, $\dot{g}_{1st}(\hat{\psi})$, de (3.8) et est pratiquement identique à (3.9) dans notre application, parce que l'incidence de l'estimation paramétrique est réduite.

4. Résultats empiriques

4.1 Simulation

Dans la présente section, nous explorons le rendement de notre approximation d'EQM de second ordre et de l'EQM estimée dans un environnement reproduisant l'enjeu de rapprochement de la présente étude. Nous utilisons un sous-ensemble des données d'origine comprenant le mode de pêche de rive pour les 17 États et sept années : les trois années de chevauchement (2015, 2016, 2017) et quatre années antérieures

(1985, 1995, 2005, 2010). Selon le plan de sondage, des combinaisons d'États et de vagues sont manquantes (par exemple janvier-février pour le Maine). On compte 607 estimations pour l'enquête téléphonique et 258 estimations pour l'enquête par la poste, dont 257 concernent les mêmes combinaisons d'États et de vagues que l'enquête téléphonique. Par conséquent, le nombre de petits domaines dans cette configuration est $m = 607 + 258 - 257 = 608$. Nous tirons les valeurs de service sans fil et les chiffres de population des données réelles.

Nous utilisons les covariables indiquées plus haut pour créer la matrice du plan d'expérience et la multiplions par les estimations de paramètre provenant du modèle final ajusté à toutes les données du mode de pêche de rive comme décrit dans la section 4.2, pour obtenir les effets fixes comme dans (3.3).

Comme Datta, Rao et Smith (2005), nous considérons trois répartitions simulant les effets aléatoires normalisés :

- $\{\psi^{-1/2}\nu_{st}\}$ iid $\mathcal{N}(0,1)$;
- $\{\psi^{-1/2}\nu_{st}\}$ iid Laplace$(0,\ 1/\sqrt{2})$;
- $\{\psi^{-1/2}\nu_{st}\}$ iid centered Exponential(1) (variables aléatoires exponentielles centrées sur la moyenne nulle).

Selon chaque répartition, $\mathrm{E}[\nu_{st}] = 0$ et $\mathrm{Var}(\nu_{st}) = \psi$. Nous choisissons ψ = 0,12, à nouveau d'après l'ajustement du modèle. Nous obtenons les efforts réels en ajoutant les effets aléatoires aux effets fixes selon (3.3).

Nous considérons trois tendances différences pour les variances de plan de sondage $\{D_{st}\}$. Tout d'abord, nous utilisons les variances de plan de sondage modélisées de la section 2.2 comme variances de plan de sondage réelles pour créer la tendance (b). Nous considérons deux configurations supplémentaires, en multipliant la tendance (b) par 0,5 pour obtenir la tendance (a) et multiplions la tendance (b) par 2.0 pour obtenir la tendance (c). Les erreurs d'échantillonnage simulées $\{e_{st}\}$ dans (2.5) sont ensuite générées indépendamment sous forme $\mathcal{N}(0, D_{st})$ pour chaque tendance.

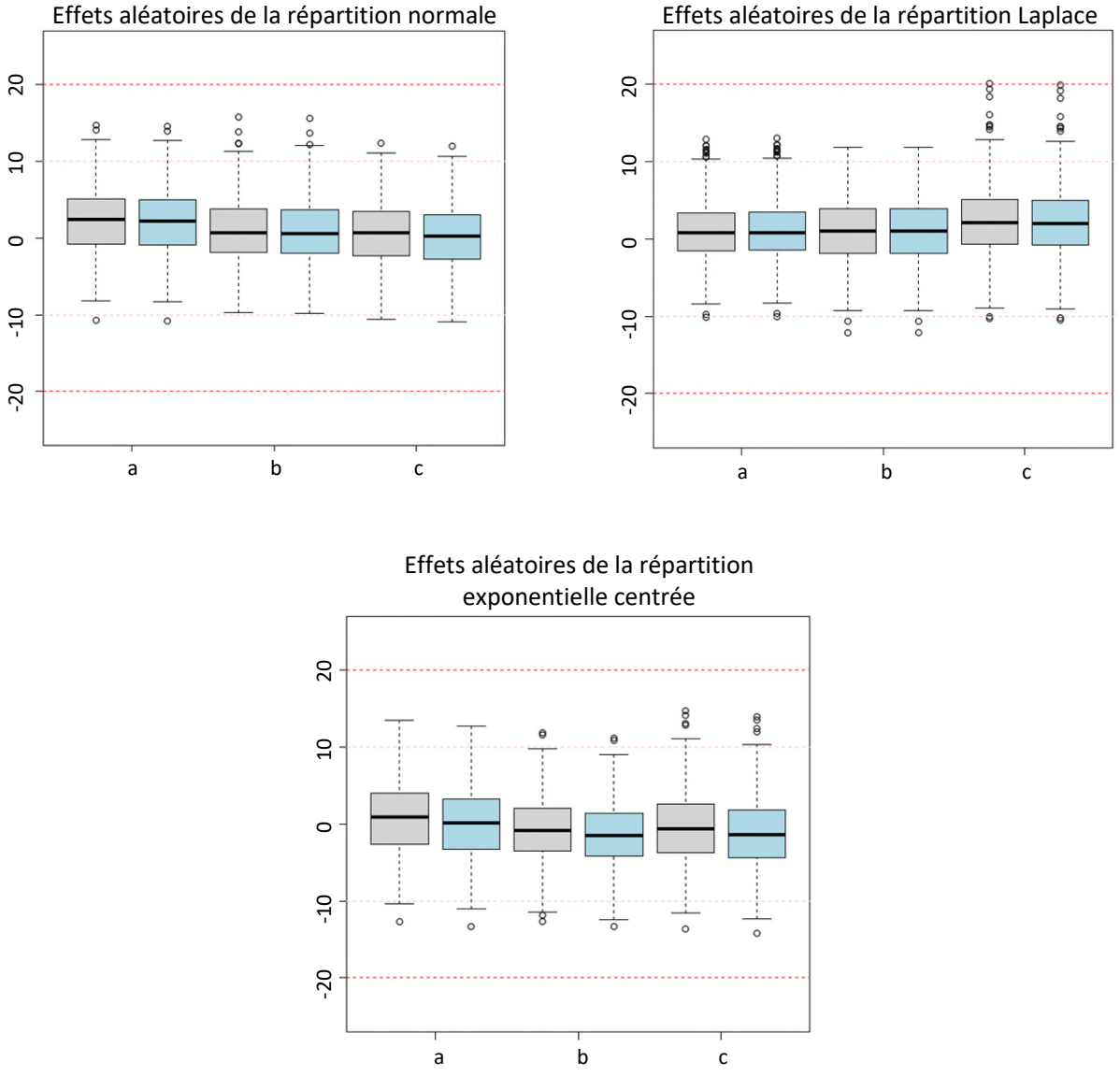
Pour chaque combinaison de tendance de variance d'échantillonnage et de répartition d'effet aléatoire, nous générons 1 000 ensembles de données à partir du modèle (2.5). Pour chaque ensemble de données simulé, nous utilisons le paquet `R sae` de Molina et Marhuenda (2015) pour calculer $\hat{\psi}$ au moyen du maximum de vraisemblance restreint REML et $\hat{\boldsymbol{\beta}}$. Nous calculons les MPLSBE selon (3.6) pour les cibles par la poste $\{M_{st}\}$, nous calculons par approximation leur EQM à l'aide de (3.7) et estimons leur EQM à l'aide de (3.8). Nous comparons ensuite ces approximations et les estimations aux EQM réelles (Monte Carlo) pour les 1 000 réalisations simulées.

Le modèle de simulation est similaire au modèle final sélectionné à la section 4.2 ci-dessous, à l'exception du retrait de certaines combinaisons inexistantes d'État et de vague dans le sous-ensemble de covariables.

La figure 4.1 présente les boîtes à moustaches de l'erreur relative (en pourcentage) de l'approximation d'EQM (3.7) et de la moyenne de Monte Carlo de l'estimateur d'EQM (3.8) par rapport à l'EQM de Monte Carlo évaluée avec 1 000 répliques. Chaque boîte à moustaches comprend l'erreur relative des 608 combinaisons d'États et de vagues pour les effets aléatoires de la répartition normale, Laplace ou exponentielle

centrée, selon la tendance d'erreur d'échantillonnage (a), (b) ou (c) décrite ci-dessus. Comme prévu, l'estimateur de l'EQM est pratiquement sans biais pour l'approximation de l'EQM dans tous les cas, donc les boîtes à moustaches sont pratiquement indiscernables dans chaque configuration. Dans l'ensemble des configurations, l'approximation de l'EQM est proche de l'EQM réelle (comme mesurée par Monte Carlo); par conséquent, la plupart des erreurs relatives sont proches de zéro et à $\pm 10\,\%$ des lignes de référence, avec quelques valeurs supérieures à $\pm 10\,\%$, mais inférieures à $\pm 20\,\%$.

Figure 4.1 Boîte à moustaches de l'erreur relative (en pourcentage) de l'approximation d'EQM (en gris clair) et de la moyenne de Monte Carlo de l'estimateur d'EQM (en bleu clair) par rapport à l'EQM de Monte Carlo évaluée avec 1 000 répliques. Chaque boîte à moustaches comprend l'erreur relative des 608 combinaisons d'États et de vagues pour les effets aléatoires de la répartition normale, Laplace ou exponentielle centrée, selon la tendance d'erreur d'échantillonnage (a), (b) ou (c) décrite dans le texte. Des lignes horizontales à $\pm 10\,\%$ et $\pm 20\,\%$ sont tracées à titre de référence.



4.2 Calage des estimations de la CHTS et de la FES

Pour les données décrites à la section 1, nous utilisons le paquet `R sae` (Molina et Marhuenda, 2015), afin d'ajuster plusieurs modèles au moyen du maximum de vraisemblance pour la pêche de rive et la pêche en bateau privé, et nous comparons les modèles selon leurs valeurs AIC (critère d'information d'Akaike). Le plus petit modèle envisagé comprend l'ordonnée à l'origine, le log(population), les indicateurs d'État, les indicateurs de vague, l'interaction État par log(population) et l'interaction État par vague. Autrement dit, le plus petit modèle ne comprend aucune différence attribuable à la méthodologie de l'enquête et supprime plutôt les termes $\mathbf{b}_{st}^{\mathrm{T}}\boldsymbol{\mu}$ et $w_{st}\mathbf{c}_{st}^{\mathrm{T}}\boldsymbol{\gamma}$ de (2.1). Le plus grand modèle envisagé ajoute le service sans fil et ses interactions avec le log(population), les indicateurs d'État, les indicateurs de vague et l'interaction État par log(population), associé à un indicateur de présence d'une estimation d'enquête par la poste et des interactions des indicateurs de l'enquête par la poste avec le log(population), les indicateurs d'État et les indicateurs de vague.

Le plus grand modèle ajoute deux principaux effets et sept interactions au plus petit modèle. Nous avons considéré 80 sous-modèles entre le plus petit et le plus grand, chacun commençant à partir du plus petit modèle plus les effets du service sans fil et les effets principaux de l'enquête par la poste. Les six interactions bidirectionnelles ont ensuite été incluses ou exclues, menant à $2^6 = 64$ modèles possibles, et $2^4 = 16$ modèles supplémentaires ont été envisagés en incluant l'interaction tridirectionnelle, le service sans fil par État par log(population) et les interactions bidirectionnelles correspondantes, service sans fil par État et service sans fil par log(population). Cela a produit un total de 80 sous-modèles à prendre en considération.

Nous utilisons les données antérieures à 2018 comme données d'entraînement, avec une taille d'échantillon de $m =$ 3 174 pour la pêche de rive et de $m =$ 3 164 pour la pêche en bateau privé. Les cinq meilleurs modèles et modèles de référence supplémentaires sont présentés au tableau 4.1 pour la pêche de rive et au tableau 4.2 pour la pêche en bateau privé. L'ordre des tableaux repose sur les valeurs AIC et les meilleurs modèles se retrouvent en premier. Les modèles qui ne tiennent pas compte de certaines différences de mode d'enquête (le plus grand moins toute l'enquête par la poste, le plus grand moins tout le service sans fil) ou toutes les différences (le plus petit) du mode d'enquête ne sont pas compétitifs par rapport aux modèles comprenant ces facteurs. Le plus grand modèle envisagé est relativement compétitif; les meilleurs modèles abandonnant un petit nombre d'interactions par rapport au plus grand modèle.

Même s'il ne s'agit pas du meilleur modèle en termes d'AIC pour la pêche de rive ou en bateau privé, le plus grand modèle moins l'interaction de l'enquête par la poste par log(population) et de l'enquête par la poste par État est au cinquième rang des meilleurs dans les deux cas. Il est pratique du point de vue opérationnel d'utiliser un modèle commun pour les deux rapprochements; ce modèle particulier est en outre également pratique, car, en extrapolant dans le temps, il ne fait intervenir que des variations au niveau de la vague une fois l'effet du service sans fil dissipé.

En utilisant les modèles ajustés, nous avons effectué des diagnostics pour évaluer l'importance potentielle de l'autocorrélation temporelle (vague par vague) dans notre spécification des effets aléatoires. Nous avons soustrait les effets fixes estimés et avons calculé les covariances empiriques pour chacun des décalages d'un à six au sein de chaque État. Ces covariances empiriques incluraient toute covariance due à

des effets aléatoires corrélés, mais aucune covariance découlant des erreurs d'échantillonnage indépendantes. Nous avons également calculé une version d'une statistique (somme normalisée des autocorrélations au carré) de Ljung et Box (1978) pour la pêche de rive et la pêche en bateau privé dans chaque État. Nous avons comparé chaque statistique calculée avec une répartition nulle (aucune autocorrélation) obtenue par le rééchantillonnage des valeurs résiduelles. Parmi les 34 statistiques, 11 étaient significatives au niveau 0,05 conventionnel, les valeurs d'autocorrélation de premier ordre variant de -0,11 à 0,29. Ces valeurs estimées étant petites et incohérentes selon les États et les modes de pêche, nous n'avons pas approfondi les effets aléatoires autocorrélés dans notre modélisation.

Tableau 4.1
Prédiction EQM, AIC et nombre de paramètres d'effets fixes hors échantillon pour divers modèles ajustés aux estimations d'effort pour la pêche de rive. Voir la description du plus grand modèle dans le texte.

Modèle le plus grand moins les termes ci-après	EQM	AIC	Paramètres
poste : log(population), poste : État, sans fil : vague	0,0837	4 564,28	152
poste : État, sans fil : vague	0,0899	4 564,69	153
poste : log(population) et sans fil : vague	0,1350	4 564,86	168
sans fil : vague	0,1354	4 566,85	169
poste : log(population) et poste : État	0,0840	4 570,45	157
rien (le plus grand)	0,1343	4 573,28	174
interactions poste	0,2104	4 580,51	152
interactions sans fil	0,3694	4 719,05	136
toutes les interactions	0,3341	4 742,84	124
tout sans fil	0,4745	4 758,73	145
tout poste	1,9466	4 838,73	151
tout poste et tout sans fil (le plus petit)	2,7443	5 106,70	122

Tableau 4.2
Prédiction EQM, AIC et nombre de paramètres d'effets fixes hors échantillon pour divers modèles ajustés aux estimations d'effort pour la pêche en bateau privé. Voir la description du plus grand modèle dans le texte.

Modèle le plus grand moins les termes ci-après	EQM	AIC	Paramètres
rien (le plus grand)	0,2068	3 314,55	174
poste : log(population)	0,2124	3 314,56	173
poste : log(population) et sans fil : vague	0,2163	3 316,42	168
sans fil : vague	0,2241	3 316,47	169
poste : log(population) et poste : État	0,2050	3 322,73	157
poste : État	0,1910	3 323,00	158
interactions poste	0,2272	3 362,27	152
tout poste	0,7046	3 501,23	151
interactions sans fil	0,4004	3 520,33	136
toutes les interactions	0,4615	3 646,78	114
tout sans fil	0,5421	3 750,03	135
tout poste et tout sans fil (le plus petit)	1,2677	3 901,82	112

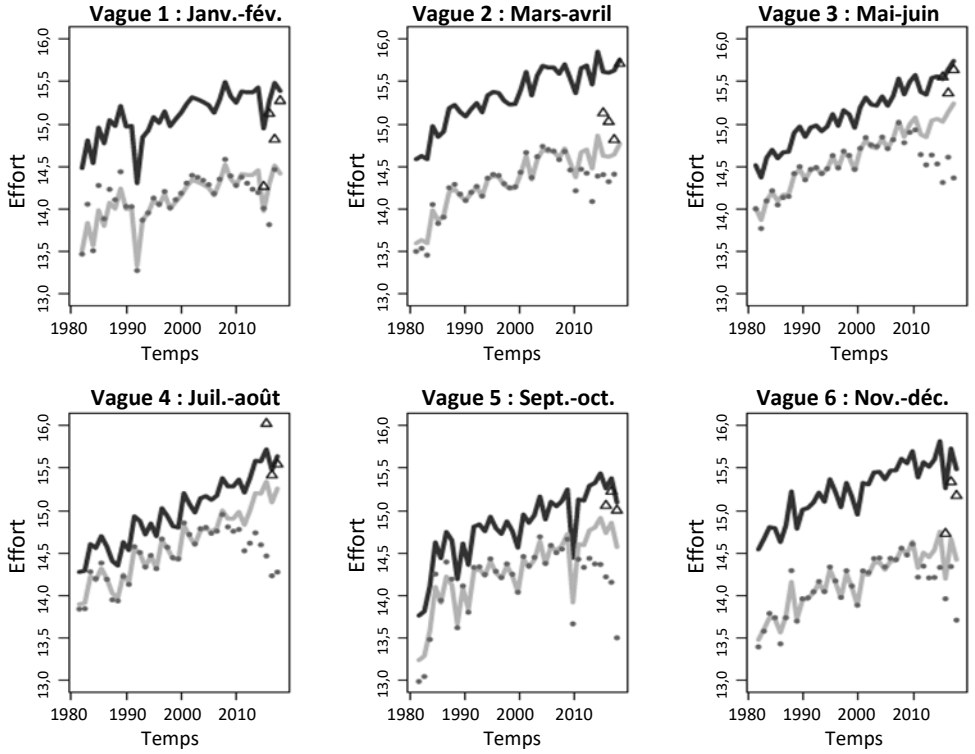
Nous utilisons les deux premières vagues de 2018 comme données hors échantillon aux fins de prédiction; des données ne sont pas disponibles pour tous les États compris dans ces vagues, entraînant 18 observations hors échantillon pour la pêche de rive et 18 pour la pêche en bateau privé. (Les données de 2018 proviennent de l'enquête par la poste uniquement et ont été sélectionnées pour la prédiction hors échantillon, car la prédiction pour l'enquête par la poste est le cas d'utilisation le plus intéressant.) Le modèle sélectionné présente l'EQM hors échantillon la plus basse pour la pêche en bateau privé et se retrouve ex

æquo (à trois décimales près) avec le résultat le plus bas pour la pêche de rive. Enfin, le modèle sélectionné est l'un des plus parcimonieux parmi les meilleurs modèles. Nous choisissons donc ce modèle comme modèle final pour les deux modes de pêche et le réajustons à l'aide du REML pour estimer la variance inconnue ψ. Nous calculons ensuite les MPLSBE de la cible par la poste $\{M_{st}\}$ pour tous les États et toutes les vagues.

Ce modèle emprunte efficacement les forces et réduit la variance des estimations directes. Pour la pêche de rive, la variance moyenne de plan de sondage est 0,0792 et l'EQM d'échantillon moyenne est 0,0445; pour la pêche en bateau privé, ces valeurs sont respectivement 0,0789 et 0,0317.

La figure 4.2 présente un exemple d'estimation de l'effort par vague pour la pêche en bateau privé pour la Floride. Dans chaque sous-figure, nous présentons les estimations ponctuelles de l'enquête par téléphone ($\hat{T}_{st}$: points gris) et de l'enquête par la poste ($\hat{M}_{st}$: triangles vides). La courbe noire représente les MPLSBE de la cible par la poste, M_{st}. La courbe grise représente les MPLSBE de la cible par téléphone corrigée pour tenir compte de l'effet du service sans fil, $T_{st} - w_{st}\mathbf{c}_{st}^{\top}\boldsymbol{\gamma}$. Les courbes au sein de chaque sous-figure présentent la tendance de l'effort au fil du temps, selon chaque mode d'enquête. Les courbes de toutes les sous-figures montrent la tendance saisonnière d'une vague à l'autre, atteignant un sommet pendant les mois d'été pour l'enquête par téléphone (même si la tendance saisonnière n'est pas forte pour la Floride et est difficile à discerner pour l'enquête par la poste).

Figure 4.2 Estimation de l'effort par vague pour la pêche en bateau privé en Floride. Les points gris sont les estimations de l'effort pour l'enquête par téléphone $\{\hat{T}_{st}\}$ et les triangles vides sont les estimations de l'effort pour l'enquête par la poste $\{\hat{M}_{st}\}$. La courbe noire représente les MPLSBE de l'effort pour l'enquête par la poste, M_{st}. La courbe grise représente les MPLSBE de l'effort pour l'enquête par téléphone corrigé pour tenir compte de l'effet du service sans fil, $T_{st} - w_{st}\mathbf{c}_{st}^{\top}\boldsymbol{\gamma}$.



On peut considérer les MPLSBE comme des versions lissées des estimations ponctuelles. La courbe grise passe par les points gris ou est une version rétrécie des points gris avant 2010. La courbe grise diverge des estimations ponctuelles après 2010, ce qui reflète l'effet du service sans fil sur la couverture de l'enquête par téléphone. Pour chaque vague, on observe un éloignement positif de la courbe grise vers la courbe noire, ce qui montre la différence sous-jacente entre les modes d'enquête par téléphone et par la poste.

5. Discussion

La méthodologie proposée tient compte de diverses sources de variation dans la série d'effort de chaque enquête, y compris la tendance, la saisonnalité et des termes irréguliers de la série d'effort réelle, ainsi que les effets du mode d'enquête dans les deux séries. Ce modèle suppose que les différences dans les erreurs de mesure et de non-réponse entre les deux enquêtes seraient stables au fil du temps, alors que les variations de l'erreur de couverture au fil du temps attribuable à l'augmentation des ménages ayant uniquement un service sans fil sont explicitement modélisées. De plus, cette méthodologie tient compte de l'incertitude attribuable à l'erreur d'échantillonnage, à l'aide d'une nouvelle approche permettant d'assurer la cohérence analytique de la mise en correspondance entre les variances de plan de sondage estimées sur l'échelle d'origine et celles estimées sur l'échelle logarithmique.

Comme nous le formulons dans la présente étude, la méthodologie de rapprochement s'avère suivre une procédure standard bien établie : l'estimation sur petits domaines de Fay-Herriot. Cela signifie que les valeurs calibrées s'avèrent être les meilleurs prédicteurs linéaires sans biais empiriques dans le cadre d'un modèle linéaire mixte utilisant des techniques fondées sur la vraisemblance. Cette méthode est suffisamment souple pour fournir des valeurs calibrées optimales pour différents enjeux : prédire les cibles d'enquête par la poste pour des points dans le temps avec des données d'enquête téléphonique uniquement, ou prédire des cibles d'enquête par téléphone pour des points dans le temps avec des données provenant uniquement d'une enquête par la poste, par exemple.

L'incertitude est quantifiée par une approximation d'erreur quadratique moyenne de MPLSBE pour de nouveaux ensembles de covariables adaptant des méthodes existantes dans la documentation. Les résultats de la simulation indiquent que l'approximation de l'erreur quadratique moyenne et son estimateur sont extrêmement précis pour les types de tailles d'échantillon et d'erreurs d'échantillonnage présents dans les données de calage. La méthodologie est facilement mise en œuvre avec un logiciel standard.

À mesure que la collecte de données se poursuit avec la nouvelle méthodologie par la poste, davantage de données seront disponibles pour explorer d'autres spécifications possibles du modèle de calage. D'autres spécifications d'effets aléatoires seraient d'un intérêt particulier, comme le modèle (d) décrit dans l'introduction, l'autocorrélation temporelle, l'hétéroscédasticité entre les États ou certaines combinaisons de ces caractéristiques. L'approche de calage de base ne serait cependant pas fondamentalement modifiée pour ces autres spécifications.

Remerciements

Nous remercions les scientifiques de NOAA Fisheries, Rob Andrews et John Foster, pour leur aide relativement à l'énonciation du problème et tous les aspects de la compilation des données. L'associé de NOAA, Ryan Kitts-Jensen, et Mike Brick, statisticien de Westat, ont contribué, de manière utile, à la discussion.

Annexe

A. Proportions des ménages ayant un service sans fil uniquement

L'évolution de la proportion de ménages ayant uniquement un service téléphonique sans fil est une covariable potentielle d'explication des variations de l'erreur de couverture au fil du temps de l'enquête par téléphone filaire uniquement. Dans le cas présent, nous décrivons la façon dont nous avons créé un ensemble de proportions prédites de ménages ayant uniquement un service sans fil, $\{w_{st}\}$, pour chaque État et vague au sein de nos données.

Ces proportions étaient à peu près nulles dans tous les États avant l'an 2000, mais ont régulièrement augmenté au fil du temps. Même si ces proportions ne sont pas disponibles pour chaque vague, les meilleures données disponibles sont les proportions estimées de juin et décembre du service sans fil seulement pour chaque État de 2007 à 2014 provenant du National Health Interview Survey, menée par le National Center for Health Statistics (Blumberg et Luke, 2013). Nous avons transformé ces proportions au moyen de fonctions logits empiriques et ajusté les valeurs transformées en fonctions linéaires continues par morceaux propres aux États enregistrant un changement de pente en 2010. Même si l'incertitude des covariables qui sont elles-mêmes des estimations d'enquête pourrait être traitée officiellement (par exemple Ybarra et Lohr (2008); Bell, Chung, Datta et Franco (2019)), les estimations des services sans fil sont précises et varient graduellement dans cette application, comme le reflète partiellement la valeur R^2 ajustée de 0,9948 pour le modèle ajusté. Nous ignorons donc l'incertitude d'échantillonnage dans les estimations du service sans fil dans ce qui suit. La rétrotransformation en proportions et l'extrapolation rétrospective donnent un ensemble de proportions prédites de ménages ayant uniquement un service sans fil, $\{w_{st}\}$, pour chaque État et chaque vague au sein de nos données.

B. Modèle de variance du plan de sondage

Nous devons estimer σ^2_{Tst} et σ^2_{Mst} dans les modèles de variance de l'erreur d'échantillonnage (2.3) et (2.4), en intégrant les estimations approximativement sans biais par rapport au plan, respectivement $\hat{V}_{Tst}$ et $\hat{V}_{Mst}$ de V_{Tst} et V_{Mst}. La modélisation ou le lissage des variances de plan de sondage avant l'intégration dans la méthodologie de Fay-Herriot est une pratique standard (voir, notamment You et Chapman (2006); You

(2021); You et Hidiroglou (2023)). Nous suivons une approche étroitement liée à l'estimation de fonction de variance généralisée (par exemple chapitre 7 de Wolter (2007)).

Supposons que n_{Tst} désigne la taille de l'échantillon de l'enquête par téléphone dans l'État s et la vague t, si elle n'est pas nulle et supposons que n_{Mst} désigne la taille de l'échantillon de l'enquête par la poste, si elle n'est pas nulle. Supposons qu'étant donné T_{st} et M_{st}, les coefficients de variation empiriques au carré sont répartis de façon log-normale, indépendamment des estimations d'effort $\hat{T}_{st}$ et $\hat{M}_{st}$:

$$\ln\left(\frac{\hat{V}_{Tst}}{\exp(2\hat{T}_{st})}\right) = \mathbf{d}_{Tst}^{\mathrm{T}}\boldsymbol{\delta}_{T0} + \delta_{T1}\ln(n_{Tst}) + \xi_{Tst}, \quad \xi_{Tst} \sim \mathcal{N}(0, \tau_T^2) \tag{B.1}$$

où $\mathbf{d}_{Tst}$ est un vecteur de covariables connues (État, vague et interaction d'État par vague) et $\boldsymbol{\delta}_{T0}$, δ_{T1} sont des coefficients de régression inconnus, et supposons que :

$$\ln\left(\frac{\hat{V}_{Mst}}{\exp(2\hat{M}_{st})}\right) = \mathbf{d}_{Mst}^{\mathrm{T}}\boldsymbol{\delta}_{M0} + \delta_{M1}\ln(n_{Mst}) + \xi_{Mst}, \quad \xi_{Mst} \sim \mathcal{N}(0, \tau_M^2), \tag{B.2}$$

où $\mathbf{d}_{Mst}$ est un vecteur de covariables connues (État, vague et interaction d'État par vague) et $\boldsymbol{\delta}_{M0}$, δ_{M1} sont des coefficients de régression inconnus. Ces modèles peuvent être reformulés comme des modèles de régression pour les estimations de la variance de la base de sondage, avec des compensations connues :

$$\ln\left(\hat{V}_{Tst}\right) = 2\hat{T}_{st} + \mathbf{d}_{Tst}^{\mathrm{T}}\boldsymbol{\delta}_{T0} + \delta_{T1}\ln(n_{Tst}) + \xi_{Tst}, \quad \xi_{Tst} \sim \mathcal{N}(0, \tau_T^2) \tag{B.3}$$

et

$$\ln\left(\hat{V}_{Mst}\right) = 2\hat{M}_{st} + \mathbf{d}_{Mst}^{\mathrm{T}}\boldsymbol{\delta}_{M0} + \delta_{M1}\ln(n_{Mst}) + \xi_{Mst}, \quad \xi_{Mst} \sim \mathcal{N}(0, \tau_M^2). \tag{B.4}$$

Empiriquement, chacun de ces modèles est très bien ajusté : 94,54 % de valeurs R^2 ajustées pour l'enquête par téléphone et 98,01 % de valeurs R^2 ajustées pour l'enquête par la poste.

Ces modèles empiriques peuvent être d'un intérêt indépendant comme fonctions de variance généralisées pour l'estimation de la variance sur l'échelle d'origine : en intégrant l'estimation ponctuelle, l'État, la vague et la taille de l'échantillon dans les versions ajustées de (B.3) et (B.4), on obtient d'excellentes estimations ponctuelles de la variance logarithmique de la base de sondage.

En supposant que $\hat{V}_{Tst}$ est exactement sans biais pour V_{Tst}, nous obtenons alors du modèle de CV log-normal (B.1) et de l'hypothèse d'indépendance conditionnelle de $\hat{V}_{Tst}$ et $\hat{T}_{st}$ étant donné T_{st} :

$$\begin{aligned}
\exp\left\{\mathbf{d}_{Tst}^{\mathrm{T}}\boldsymbol{\delta}_{T0} + \delta_{T1}\ln(n_{Tst}) + \frac{\tau_T^2}{2}\right\} &= \mathrm{E}\left[\frac{\hat{V}_{Tst}}{\exp\left(2\hat{T}_{st}\right)} \,\middle|\, T_{st}\right] \\
&= \mathrm{E}\left[\hat{V}_{Tst} \,\middle|\, T_{st}\right]\mathrm{E}\left[\exp\left(-2\hat{T}_{st}\right)\middle|\, T_{st}\right] \\
&= V_{Tst}\exp\left(-2T_{st} + 2\sigma_{Tst}^2\right),
\end{aligned} \tag{B.5}$$

et de façon similaire

$$\begin{aligned}\exp\left\{\mathbf{d}_{Mst}^{\mathrm{T}}\boldsymbol{\delta}_{M0}+\delta_{M1}\ln(n_{Mst})+\frac{\tau_M^2}{2}\right\} &= \mathrm{E}\left[\frac{\hat{V}_{Mst}}{\exp\left(2\hat{M}_{st}\right)}\;\middle|\; M_{st}\right]\\ &= \mathrm{E}\left[\hat{V}_{Mst}\middle| M_{st}\right]\mathrm{E}\left[\exp\left(-2\hat{M}_{st}\right)\middle| M_{st}\right] \qquad \text{(B.6)}\\ &= V_{Mst}\exp\left(-2M_{st}+2\sigma_{Mst}^2\right).\end{aligned}$$

Ainsi, nous obtenons de (2.3) et (B.5) :

$$\begin{aligned}\exp\left\{\mathbf{d}_{Tst}^{\mathrm{T}}\boldsymbol{\delta}_{T0}+\delta_{T1}\ln(n_{Tst})+\frac{\tau_T^2}{2}\right\} &= \left\{\exp(\sigma_{Tst}^2)-1\right\}\exp\left\{2T_{st}+\sigma_{Tst}^2\right\}\exp\left(-2T_{st}+2\sigma_{Tst}^2\right)\\ &= \exp(4\sigma_{Tst}^2)-\exp\left(3\sigma_{Tst}^2\right)\end{aligned} \qquad \text{(B.7)}$$

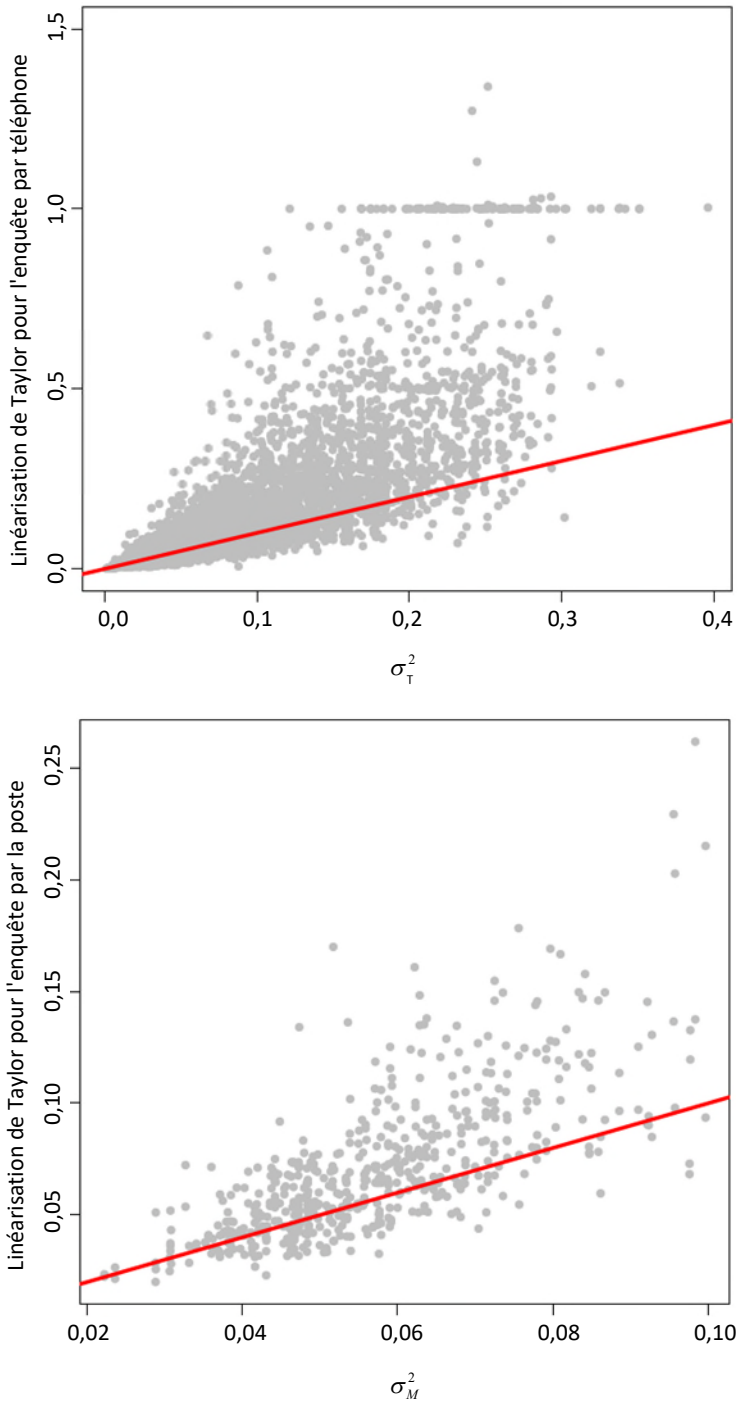
et de (2.4) et (B.6) :

$$\begin{aligned}\exp\left\{\mathbf{d}_{Mst}^{\mathrm{T}}\boldsymbol{\delta}_{M0}+\delta_{M1}\ln(n_{Mst})+\frac{\tau_M^2}{2}\right\} &= \left\{\exp(\sigma_{Mst}^2)-1\right\}\exp\left\{2M_{st}+\sigma_{Mst}^2\right\}\exp\left(-2M_{st}+2\sigma_{Mst}^2\right)\\ &= \exp(4\sigma_{Mst}^2)-\exp\left(3\sigma_{Mst}^2\right).\end{aligned} \qquad \text{(B.8)}$$

Les paramètres de gauche de (B.7) peuvent être estimés à partir de (B.1) et les paramètres de gauche de (B.8) peuvent être estimés à partir de (B.2). Les estimations résultantes de σ_{Tst}^2 et σ_{Mst}^2 peuvent alors être obtenues en résolvant les équations (B.7) et (B.8), qui sont des polynômes quartiques dans $\exp(\sigma_{Tst}^2)$ et $\exp(\sigma_{Mst}^2)$. À l'aide de la règle des signes de Descartes, on peut montrer que chacune de ces équations quartiques a une racine négative réelle, deux racines conjuguées complexes et une racine positive réelle. Les solutions pour σ_{Tst}^2 et σ_{Mst}^2 sont alors les logarithmes des racines positives réelles uniques, qui peuvent être obtenues par des procédures numériques standard. Même si ces solutions sont en fait des estimations, nous les traiterons comme étant fixes et connues dans ce qui suit, selon la norme des techniques d'estimation sur petits domaines que nous appliquons dans les sections suivantes.

Les variances de plan de sondage obtenues sur l'échelle logarithmique, σ_{Tst}^2 et σ_{Mst}^2, présentent des corrélations robustes (respectivement 0,798 et 0,803) avec les approximations de variance issues de la linéarisation de Taylor, $\hat{V}_{Tst}\exp\left(-2\hat{T}_{st}\right)$ et $\hat{V}_{Mst}\exp\left(-2\hat{M}_{st}\right)$. Elles ne sont cependant pas identiques (voir la figure B.1) et la méthode décrite force la cohérence analytique entre le modèle moyen et le modèle de variance et lisse un peu les variances de plan de sondage. De plus, cette méthode produit des estimations raisonnables pour certains cas pour lesquels un seuil de valeur maximale a été artificiellement appliqué aux variances de plan de sondage, comme le montre la partie gauche de la figure B.1.

Figure B.1 Variances de plan de sondage estimées pour l'effort (log de sorties) au moyen d'une linéarisation de Taylor et résolution des équations polynomiales quartiques (B.7) pour l'enquête par téléphone (partie du haut) et (B.8) pour l'enquête par la poste (partie du bas).



Bibliographie

Andrews, R., Brick, J.M. et Mathiowetz, N.A. (2014). Development and testing of recreational fishing effort surveys: Testing a mail survey design. Rapport technique, National Marine Fisheries Service. https://www.st.nmfs.noaa.gov/pims/main/public?method=DOWNLOAD_FR_PDF&record_id=1179.

Bell, W.R., Chung, H.C., Datta, G.S. et Franco, C. (2019). Erreur de mesure dans l'estimation sur petits domaines : comparaison de modèles fonctionnels, structurels et naïfs. *Techniques d'enquête*, 45, 1, 65-86. Article accessibla à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2019001/article/00005-fra.pdf.

Blumberg, S., et Luke, J. (2013). Wireless substitution: Early release of estimates from the National Health Interview Survey, Juillet-décembre 2012. Rapport technique, National Center for Health Statistics. http://www.cdc.gov/nchs/nhis.htm.

Boonstra, H.J. et van den Brakel, J. (2022). Multilevel time-series models for small area estimation at different frequencies and domain levels. *The Annals of Applied Statistics*, 16(4), 2314-2338.

Boonstra, H.J., van den Brakel, J. et Das, S. (2021). Multilevel time series modelling of mobility trends in the Netherlands for small domains. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(3), 985-1007.

Curtin, R., Presser, S. et Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly*, 69(1), 87.

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

Datta, G.S., Rao, J.N.K. et Smith, D.D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, 92(1), 183-196.

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

Erciulescu, A.L., Opsomer, J.D. et Breidt, J.F. (2021). A bridging model to reconcile statistics based on data from multiple surveys. *The Annals of Applied Statistics*, 15(2), 1068-1079.

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

Feder, M. (2001). Time series analysis of repeated surveys: The state-space approach. *Statistica Neerlandica*, 55(2), 182-199.

Jiang, J., et Lahiri, P. (2006). Mixed model prediction and small area estimation. *TEST*, 15(1), 1.

Ljung, G.M., et Box, G.E. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2), 297-303.

Lohr, S.L., et Brick, J.M. (2012). Blending domain estimates from two victimization surveys with possible bias. *Canadian Journal of Statistics*, 40(4), 679-696.

Manzi, G., Spiegelhalter, D.J., Turner, R.M., Flowers, J. et Thompson, S.G. (2011). Modelling bias in combining small area prevalence estimates from multiple surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(1), 31-50.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1), 27-48.

Molina, I., et Marhuenda, Y. (2015). `sae`: An `R` package for small area estimation. *The R Journal*, 7/1, 81-98.

National Research Council (2006). *Review of Recreational Fisheries Survey Methods*. Washington, DC: The National Academies Press.

Olson, K., Smyth, J.D., Horwitz, R., Keeter, S., Lesser, V., Marken, S., Mathiowetz, N.A., McCarthy, J.S., O'Brien, E., Opsomer, J.D., Steiger, D., Sterrett, D., Su, J., Suzer-Gurtekin, Z.T., Turakhia, C. et Wagner, J. (2020). Transitions from telephone surveys to self-administered and mixed-mode surveys: AAPOR task force report. *Journal of Survey Statistics and Methodology*.

Pfeffermann, D. (2002). Small area estimation: New developments and directions. *Revue Internationale de Statistique*, 70(1), 125-143.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28(1), 40-68.

Pfeffermann, D., et Tiller, R. (2006). Small-area estimation with state-space models subject to benchmark constraints. *Journal of the American Statistical Association*, 101(476), 1387-1397.

Raghunathan, T.E., Xie, D., Schenker, N., Parsons, V.L., Davis, W.W., Dodd, K.W. et Feuer, E.J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association*, 102(478), 474-486.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation*. Wiley Series in Survey Methodology. Wiley.

Rao, J.N.K., et Yu, M. (1994). Small-area estimation by combining time-series and cross-sectional data. *Canadian Journal of Statistics*, 22(4), 511-528.

Slud, E.V., et Maiti, T. (2006). Mean-squared error estimation in transformed Fay-Herriot models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(2), 239-257.

US Census Bureau (2016). *State Population Totals Datasets: 2010-2016*, 2016. https://www.census.gov/data/datasets/2016/demo/popest/state-total.html.

van den Brakel, J.A. (2008). Design-based analysis of embedded experiments with applications in the Dutch Labour Force Survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(3), 581-613.

van den Brakel, J.A. (2010). Sampling and estimation techniques for the implementation of new classification systems: The change-over from NACE Rev. 1.1 to NACE Rev. 2 in business surveys. *Survey Research Methods*, 4, 103-119.

van den Brakel, J.A. (2013). Analyse fondée sur le plan de sondage de plans d'expérience factoriels intégrés dans des échantillons probabilistes. *Techniques d'enquête*, 39, 2, 355-383. Article accessibla à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2013002/article/11870-fra.pdf.

van den Brakel, J., Zhang, X. et Tam, S.-M. (2020). Measuring discontinuities in time series obtained with repeated sample surveys. *Revue Internationale de Statistique*, 88(1), 155-175.

Wang, J.C., Holan, S.H., Nandram, B., Barboza, W., Toto, C. et Anderson, E. (2012). A Bayesian approach to estimating agricultural yield based on multiple repeated surveys. *Journal of Agricultural, Biological, and Environmental Statistics*, 17(1), 84-106.

Wolter, K.M. (2007). *Introduction to Variance Estimation (2nd Edition)*. New York: Springer-Verlag Inc.

Ybarra, L.M., et Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95(4), 919-931.

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

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

You, Y., et Hidiroglou, M.A. (2023). Application of sampling variance smoothing methods for small area proportion estimation. *Journal of Official Statistics*, 39(4), 571-590.

Techniques d'enquête, décembre 2024
Vol. 50, nº 2, pp. 375-405
Statistique Canada, nº 12-001-X au catalogue

Données entièrement synthétiques pour des enquêtes complexes

Shirley Mathur, Yajuan Si et Jerome P. Reiter[1]

Résumé

Lorsqu'ils souhaitent diffuser des fichiers à grande diffusion pour des données confidentielles, les organismes de statistique peuvent générer des données entièrement synthétiques. Nous proposons une méthode pour construire des données entièrement synthétiques à partir d'enquêtes dont les données sont recueillies selon des plans d'échantillonnage complexes. Notre méthode respecte la stratégie générale proposée par Rubin (1993). Plus précisément, nous générons des pseudo-populations en appliquant la méthode du bootstrap bayésien en population finie pondéré pour tenir compte des poids d'enquête, tirons des échantillons aléatoires simples de ces pseudo-populations, estimons des modèles de synthèse en utilisant ces échantillons aléatoires simples et diffusons des données simulées tirées des modèles sous la forme de fichiers à grande diffusion. Pour faciliter l'estimation de la variance, nous utilisons le cadre d'imputation multiple et deux stratégies de génération de données. Dans la première, nous générons plusieurs ensembles de données à partir de chaque échantillon aléatoire simple. Dans la seconde, nous générons un seul ensemble de données synthétiques à partir de chaque échantillon aléatoire simple. Nous présentons des règles de combinaison de l'imputation multiple pour chaque scénario. Nous illustrons les propriétés d'échantillonnage répété des règles de combinaison au moyen d'études par simulations, ce qui comprend des comparaisons avec la génération de données synthétiques en fonction de méthodes de pseudo-vraisemblance. Nous appliquons les méthodes proposées à un sous-ensemble de données tirées de l'American Community Survey.

Mots-clés : Bootstrap; confidentialité; divulgation; poids; protection des renseignements personnels.

1. Introduction

Un grand nombre d'organismes nationaux de statistique, d'organisations qui mènent des enquêtes et de chercheurs – ci-après tous regroupés sous le nom d'organismes – diffusent au public des microdonnées, c'est-à-dire des données sur des unités individuelles. La société profite grandement de la vaste diffusion de microdonnées, permettant à de larges sous-ensembles de la communauté de recherche d'accéder aux données recueillies et de les analyser (Reiter, 2009). Cependant, il arrive souvent que les organismes ne puissent pas diffuser les microdonnées telles que recueillies, car ce faisant, ils pourraient révéler les identités ou les valeurs des attributs de nature délicate des répondants aux enquêtes, et ainsi ne pas respecter les exigences éthiques ou juridiques de protéger la confidentialité des personnes concernées (Reiter et Raghunathan, 2007).

Pour gérer ces risques, de nombreux organismes ont mis en œuvre ou envisagent des méthodes de données synthétiques, comme Rubin (1993) l'a d'abord proposé. Dans cette méthode, l'organisme : i) échantillonne aléatoirement et indépendamment des unités tirées de la base de sondage pour former chaque ensemble de données synthétiques; ii) impute les valeurs de données inconnues pour les unités dans les échantillons synthétiques en utilisant des modèles qui correspondent aux données d'enquête initiales; iii)

1. Shirley Mathur, Department of Statistics, B-313 Padelford Hall, Université de Washington, Seattle, WA 98195-4322; Yajuan Si, Survey Research Center, Institute for Social Research, Université du Michigan, Rm 4014, 426 Thompson St., Ann Arbor, MI 48104. Courriel : yajuan@umich.edu; Jerome P. Reiter, Department of Statistical Science, 214a Old Chemistry Building, Université Duke, Durham, NC 27708-0251.

diffuse plusieurs versions de ces ensembles de données au public. On les appelle ensembles de données entièrement synthétiques (Drechsler, 2011; Raghunathan, 2021). La diffusion de données entièrement synthétiques peut préserver la confidentialité, puisque l'identification des unités et de leurs données de nature délicate peut être difficile lorsque les données diffusées ne sont pas des valeurs recueillies réelles (Reiter et Drechsler, 2010). Des méthodes d'inférence à partir de ces fichiers de données à imputation multiple ont été mises au point pour diverses tâches d'inférence statistique (Raghunathan, Reiter et Rubin, 2003; Reiter, 2002, 2005a,b; Drechsler et Reiter, 2010; Si et Reiter, 2011).

Bien qu'il existe des applications bien connues de données entièrement synthétiques pour les recensements ou les données administratives (par exemple Kinney, Reiter, Reznek, Miranda, Jarmin et Abowd, 2011), bon nombre d'ensembles de données de recherche sont fondés sur des enquêtes dont les données ont été recueillies au moyen de plans d'échantillonnage qui reposent sur des probabilités de sélection inégales. Des recherches antérieures sur l'imputation multiple pour les données manquantes laissent supposer que les modèles d'imputation devraient tenir compte des caractéristiques du plan de sondage, comme la stratification, la mise en grappes et les poids d'enquête (Reiter, Raghunathan et Kinney, 2006). Dans le même ordre d'idées, lors de l'utilisation de l'imputation multiple pour les données synthétiques, les modèles devraient aussi tenir compte du plan d'enquête (Mitra et Reiter, 2006; Fienberg, 2010; Kim, Drechsler et Thompson, 2021). Le défi principal consiste à incorporer adéquatement les poids dans les modèles de synthèse, ce qui nous ramène au débat de longue date à propos du rôle des poids d'enquête dans les inférences basées sur un modèle (Pfeffermann, 1993, 2011; Little, 2004).

Les chercheurs ont proposé diverses approches pour générer des données entièrement synthétiques dans les enquêtes complexes. Dans les premiers travaux, on suggérait (Rubin, 1993; Raghunathan et coll., 2003; Reiter, 2002) d'adopter une approche bayésienne d'inférence de population finie, dans le cadre de laquelle l'organisme : i) construit des modèles prédictifs pour les variables d'enquête conditionnels sur des caractéristiques du plan de sondage comme les indicateurs de strate ou de grappe ou les mesures de taille, que l'on présume connus par l'organisme pour chaque unité de la population; ii) impute les variables d'enquête manquantes pour les unités non échantillonnées de la population; iii) tire un échantillon aléatoire simple à partir de la population complète pour le diffuser en un seul ensemble de données synthétiques. Une approche connexe consiste à utiliser le bootstrap bayésien en population finie pondéré (BBPFP) (Dong, Elliott et Raghunathan, 2014), dans le cadre duquel l'organisme génère des populations complètes en reproduisant les personnes à partir des données confidentielles en proportion de leurs poids d'enquête et diffuse ensuite les populations complètes, abandonnant ainsi l'étape d'échantillonnage aléatoire simple. Plus récemment, il a été proposé de construire des modèles de données synthétiques qui tiennent compte directement du plan d'échantillonnage, de sorte qu'ils estiment la distribution conjointe des données de population. Par exemple, l'organisme peut utiliser une approche de pseudo-vraisemblance (Pfeffermann, 1993; Savitsky et Toth, 2016), dans laquelle la contribution de chaque personne à la fonction de

vraisemblance d'un modèle de synthèse est élevée à une puissance qui est une fonction des poids d'enquête (Kim et coll., 2021). Pour s'écarter de la proposition de Rubin (1993), une approche complètement différente consiste à créer de nouveaux poids et à les coupler à des enregistrements de données synthétiques simulés à partir de modèles qui sont indépendants des poids d'enquête (Commission économique des Nations Unies pour l'Europe, 2022). Dans ce cas-ci, le but est de permettre aux utilisateurs d'utiliser des estimations pondérées dont l'ampleur correspond à la population finie. Les nouveaux poids peuvent être créés en traitant les poids d'enquêtes comme une variable dans la synthèse, afin que l'organisme précise un modèle prédictif pour les poids. Les poids simulés peuvent être corrigés par ratissage ou calage avant d'être ajoutés dans le fichier diffusé.

Chacune de ces méthodes présente des inconvénients potentiels. La méthode bayésienne d'inférence de population finie, bien que théorique par principe, exige de créer des populations complètes, ce qui peut être encombrant, et la disponibilité de variables du plan de sondage pour tous les enregistrements dans la population, ce qui pourrait ne pas être le cas dans certaines enquêtes. La méthode du BBPFP diffuse (en multiples copies) les enregistrements de données réelles des personnes, ce qui crée des risques évidents de divulgation. Les méthodes de pseudo-vraisemblance peuvent ne pas estimer la variabilité d'échantillonnage correctement (Williams et Savitsky, 2021), et l'on ne sait pas avec certitude dans quelle mesure il est facile de les mettre en œuvre au moyen de synthétiseurs d'apprentissage automatique comme les arbres de classification et de régression (Reiter, 2005c), qui sont couramment utilisés dans les projets de données synthétiques pratiques (Raab, Nowok et Dibben, 2018). En ayant des poids synthétisés, les analystes secondaires devraient utiliser les poids simulés pour donner une approximation de l'inférence fondée sur le plan de sondage. Cette approximation ne repose sur aucune base théorique; à ce titre, on ne sait pas avec certitude si la méthode de poids synthétiques facilite l'obtention d'inférences exactes en général.

Dans le présent article, nous proposons une méthode pour générer des données entièrement synthétiques à partir d'échantillons complexes dans l'esprit de la proposition initiale de Rubin (1993), c'est-à-dire l'organisme diffuse des échantillons aléatoires simples qui n'exigent pas des utilisateurs qu'ils effectuent des analyses pondérées par les poids d'enquête à l'aide de données synthétiques. Pour ce faire, nous nous inspirons de la méthode du BBPFP de Dong et coll. (2014) en créant d'abord des pseudo-populations qui tiennent compte des poids d'enquête. Nous tirons ensuite des échantillons aléatoires simples (EAS) de chaque pseudo-population, estimons des modèles de synthèse à partir de chaque EAS et générons des tirages à partir de ces modèles pour créer des fichiers de données à grande diffusion entièrement synthétiques à imputation multiple. La dernière étape offre une protection de la confidentialité, car l'organisme ne diffuse aucun enregistrement réel. Nous tenons compte de deux processus pour la dernière étape de génération de données synthétiques. Dans *Synrep-R*, nous générons plusieurs ensembles de données synthétiques à partir de chaque EAS. Dans *SynRep-1*, nous générons un ensemble de données synthétiques à partir de chaque EAS. *SynRep-R* diffuse plus d'ensembles de données que *SynRep-1*, ce qui peut donner lieu à des variances réduites. Cependant, les ensembles de données supplémentaires peuvent accroître les coûts indirects pour

l'organisme et les analystes secondaires, et ils fournissent des renseignements supplémentaires pour les adversaires à la recherche de divulgations. Pour les deux approches, nous obtenons des règles de combinaison d'imputation multiple qui permettent d'estimer les variances. Au moyen d'études par simulations, nous illustrons les rendements d'échantillonnage répété des règles de combinaison et les comparons aux données entièrement synthétiques générées tout en ne tenant aucunement compte du plan d'échantillonnage. Nous les comparons aussi aux approches qui reposent sur des modèles de synthèse estimés au moyen de pseudo-vraisemblances pondérées (Kim et coll., 2021). Finalement, nous illustrons les méthodes proposées en utilisant un sous-ensemble de données de l'American Community Survey (ACS). Le code pour les études par simulations et l'exemple à partir de l'ACS est disponible à l'adresse https://github.com/yajuansi-sophie/SynRep.

Le reste du présent article est structuré de la façon suivante. La section 2 décrit les deux processus de génération de données synthétiques en détail et présente les nouvelles règles de combinaison. La section 3 porte sur les études par simulations. La section 4 présente l'exemple au moyen des données de l'ACS. Enfin, la section 5 propose des sujets pour les recherches futures.

2. Méthodes proposées pour générer des données d'enquête entièrement synthétiques

Supposons que $\mathcal{D}$ est un échantillon probabiliste de taille n tiré aléatoirement d'une population finie comportant N unités. Pour $i = 1, \ldots, N,$ supposons que π_i est la probabilité de sélection de l'unité i et supposons que $w_i = 1/\pi_i$ est le poids d'enquête de l'unité. Dans ce cas-ci, nous n'affichons aucune préférence quant à déterminer si w_i est potentiellement corrigé, par exemple aux fins de normalisation, de calage ou de non-réponse, bien que dans nos études par simulations, nous utilisions des poids de sondage purs. Pour $i = 1, \ldots, N,$ supposons que Y_i est le vecteur $p \times 1$ des variables de l'enquête. Donc, $\mathcal{D} = \{(w_i, Y_i) : i = 1, \ldots, n\}.$ Pour simplifier l'explication, nous supposons que $p = 1,$ afin que Y_i soit un scalaire. *SynRep-R* et *SynRep-1*, et leurs méthodes d'inférence correspondantes, peuvent aussi être utilisés à l'aide de données d'enquête multivariées.

À la section 2.1, nous décrivons les processus de génération de données synthétiques. À la section 2.2, nous décrivons les méthodes d'inférence. Comme on l'a mentionné à la section 1 et selon la proposition de Rubin (1993), nous fixons l'objectif de permettre aux utilisateurs secondaires d'analyser les ensembles de données diffusées comme s'il s'agissait d'échantillons aléatoires simples tirés de la population.

2.1 Processus de génération de données

La figure 2.1 et la figure 2.2 illustrent les processus de génération de données synthétiques pour *SynRep-R* et *SynRep-1*, respectivement. Nous décrivons maintenant les étapes en détail.

Figure 2.1 Processus de génération de données synthétiques au moyen de plusieurs ensembles de données par échantillon aléatoire simple, ce que nous appelons *SynRep-R*.

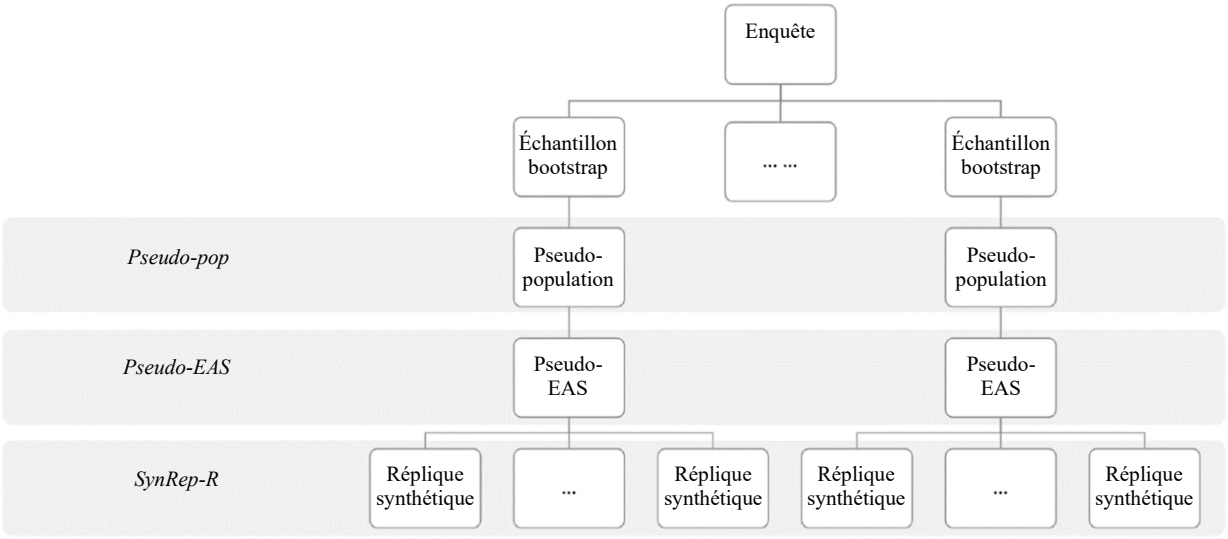
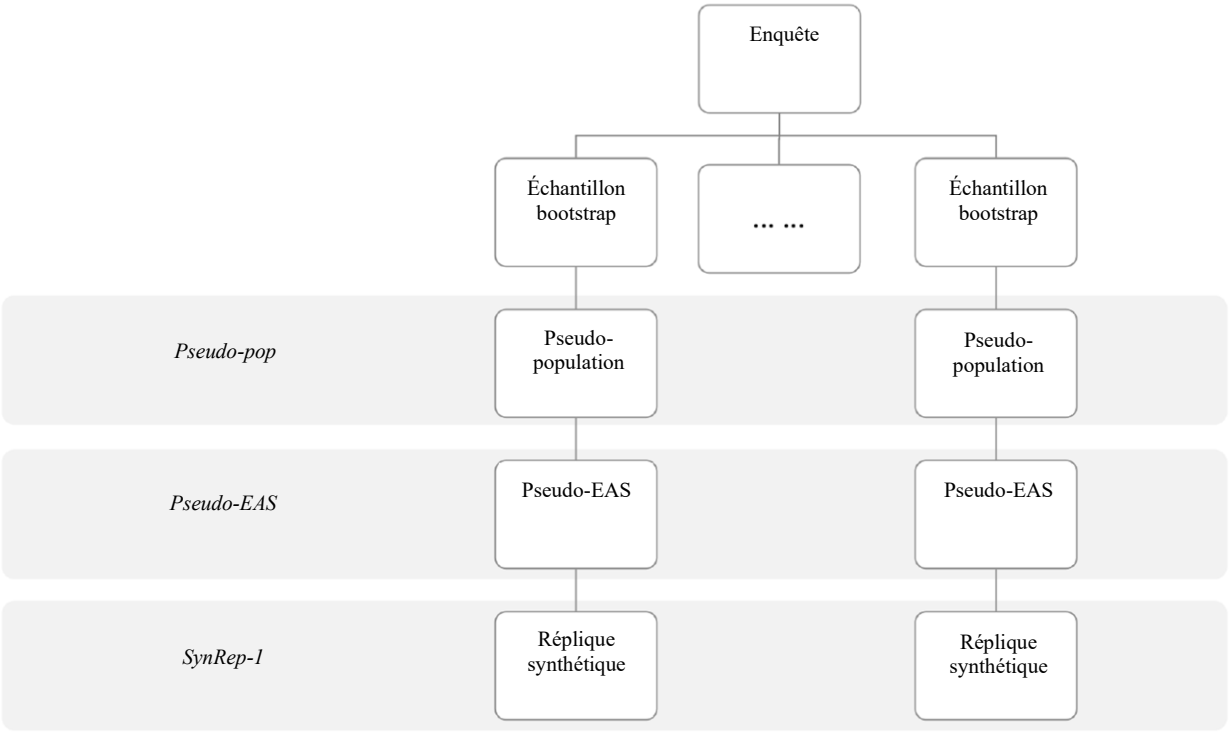


Figure 2.2 Processus de génération de données synthétiques au moyen d'un ensemble de données par échantillon aléatoire simple, ce que nous appelons *SynRep-1*.



Dans l'un ou l'autre des processus, la première étape consiste à générer des pseudo-populations en utilisant le BBPFP (Dong et coll., 2014). Le BBPFP génère des pseudo-populations en « défaisant » le plan d'échantillonnage complexe et en tenant compte des poids d'échantillonnage. L'idée est de s'appuyer sur la distribution prédictive *a posteriori* de données non observées (Y_{nob}) compte tenu des données observées (Y_{obs}) et des poids d'enquête, c'est-à-dire en s'appuyant sur $P(Y_{\text{nob}} \mid Y_{\text{obs}}, w_1, \ldots, w_n)$. Cette distribution laisse supposer que la population est formée des valeurs uniques de $Y_i \in \mathcal{D}$ et que les nombres correspondants pour chaque valeur dans la population suivent une distribution multinomiale. Au moyen d'une distribution *a priori* de Dirichlet non informative sur les probabilités multinomiales, la distribution de Pólya peut être utilisée pour tirer les échantillons prédictifs au lieu de la distribution multinomiale de Dirichlet.

En ayant cela en tête, le processus de génération de données synthétiques est décrit ci-dessous.

1. **Rééchantillonner par bootstrap bayésien :** Pour introduire une variabilité d'échantillonnage suffisante, en utilisant les données de l'échantillon « parent » $\mathcal{D}$, nous générons les échantillons M, $(\mathcal{S}^{(1)}, \ldots, \mathcal{S}^{(M)})$, chacun étant de taille n, en utilisant des bootstraps bayésiens indépendants (Rubin, 1981). Pour chaque $\mathcal{S}^{(m)}$ et pour $i = 1, \ldots, n$, supposons que $w_i^{(m)} = c w_i r_i^{(m)}$, où $r_i^{(m)}$ est le nombre de fois que l'élément i de $\mathcal{D}$ figure dans $\mathcal{S}^{(m)}$. Le c est une constante de normalisation permettant de garantir que la somme des nouveaux poids correspond à la taille de la population N. Donc, dans chaque $\mathcal{S}^{(m)}$, pour $i = 1, \ldots, n$, nous créons $w_i^{(m)} = \left(N w_i r_i^{(m)}\right) \Big/ \left(\sum_k w_k r_k^{(m)}\right)$.
2. **Utiliser le BBPFP pour produire des pseudo-populations :** Pour chaque $\mathcal{S}^{(m)}$, nous construisons une urne de Pólya initiale en utilisant l'ensemble de $\left\{Y_i, w_i^{(m)}\right\}$. Nous tirons ensuite $N - n$ unités en utilisant les probabilités $\left(p_1^{(m)}, \ldots, p_n^{(m)}\right)$ déterminées à partir de

$$p_i^{(m)} = \frac{w_i^{(m)} - 1 + l_{i,k-1}^{(m)} (N-n)/n}{N - n + (k-1)(N-n)/n}, \tag{2.1}$$

pour le k^{e} tirage, $k \in \{1, \ldots, N-n\}$, où $l_{i,k-1}^{(m)}$ est le nombre de sélections bootstrap de Y_i parmi les éléments présents dans l'urne au tirage $k-1$. Les $N-n$ tirages combinés avec les données dans $\mathcal{S}^{(m)}$ forment une pseudo-population, $\mathcal{P}^{(m)}$. Nous répétons ce segment pour $m = 1, \ldots, M$ pour créer $\mathcal{P}_{\text{pseudo}} = \{\mathcal{P}^{(m)} : m = 1, \ldots, M\}$. Quand N est très vaste, nous pouvons économiser des coûts de mémoire et de calcul en créant une pseudo-population qui est assez grande pour être pratiquement la même aux fins d'inférence qu'une population de taille N, que nous opérationnalisons en générant $50n$ plutôt que $N-n$ enregistrements.
3. **Tirer un EAS à partir de chaque pseudo-population :** Pour $m = 1, \ldots, M$, tirer un échantillon aléatoire simple $\mathcal{D}^{(m)}$ de taille n à partir de $\mathcal{P}^{(m)}$. Soit $\mathcal{D}_{\text{EAS}} = \{\mathcal{D}^{(m)} : m = 1, \ldots, M\}$.
4. **Générer des répliques de données synthétiques :** Pour $m = 1, \ldots, M$, estimer un modèle de synthèse en utilisant $\mathcal{D}^{(m)}$ et faire un tirage à partir des distributions prédictives pour former des répliques de données synthétiques en utilisant l'étape 4a ou l'étape 4b.

4a. *SynRep-R :* Pour $m = 1, \ldots, M,$ tirer $R > 1$ répliques synthétiques $\mathcal{D}_{\text{syn}}^{(m,r)}$ de taille $n,$ où $r = 1, \ldots, R,$ en utilisant chaque $\mathcal{D}^{(m)}$. Nous diffusons $\mathcal{D}_{\text{syn}} = \{\mathcal{D}_{\text{syn}}^{(m,r)} : m = 1, \ldots, M;\ r = 1, \ldots, R\},$ y compris les indicateurs pour lesquels m de chaque $\mathcal{D}_{\text{syn}}^{(m,r)}$ appartient.

4b. *SynRep-1 :* Pour $m = 1, \ldots, M,$ tirer un échantillon de données synthétiques $\mathcal{D}_{\text{syn}}^{(m)}$ de taille n de chaque $\mathcal{D}^{(m)}$. Diffuser $\mathcal{D}_{\text{syn}} = \{\mathcal{D}_{\text{syn}}^{(m)} : m = 1, \ldots, M\}$.

Le modèle de synthèse pour chaque $\mathcal{D}^{(m)}$ peut utiliser des valeurs par substitution des paramètres du modèle, par exemple leurs estimations du maximum de vraisemblance. Il n'est pas nécessaire d'utiliser des distributions *a posteriori* à cette étape du processus (Reiter et Kinney, 2012).

Comme ces deux processus de génération de données synthétiques diffèrent de ceux de Raghunathan et coll. (2003), ainsi que d'autres scénarios de données synthétiques tels que ceux de Reiter (2003, 2004), nous avons besoin de nouvelles méthodes pour les inférences, ce que nous allons maintenant examiner.

2.2 Inférences pour *SynRep-R* et *SynRep-1*

Pour obtenir les méthodes d'inférence, nous suivons la stratégie générale d'imputation multiple (Rubin, 1987) et utilisons une méthode d'inférence bayésienne. Pour toute quantité de population $Q,$ comme la moyenne de population $Q \equiv \bar{Y},$ nous cherchons la distribution *a posteriori* $P(Q \mid \mathcal{D}_{\text{syn}})$. Suivant la méthode de Raghunathan et coll. (2003), nous calculons l'intégrale suivante en fonction de chaque niveau du processus de synthèse de données de la figure 2.1 ou de la figure 2.2.

$$\begin{aligned} P(Q \mid \mathcal{D}_{\text{syn}}) = \iiint & P(Q \mid \mathcal{D}_{\text{syn}}, \mathcal{D}_{\text{EAS}}, \mathcal{P}_{\text{pseudo}}, \mathcal{D})\, P(\mathcal{D} \mid \mathcal{D}_{\text{syn}}, \mathcal{D}_{\text{EAS}}, \mathcal{P}_{\text{pseudo}}) \\ & P(\mathcal{P}_{\text{pseudo}} \mid \mathcal{D}_{\text{syn}}, \mathcal{D}_{\text{EAS}})\, P(\mathcal{D}_{\text{EAS}} \mid \mathcal{D}_{\text{syn}})\, d\mathcal{D} d\mathcal{P}_{\text{pseudo}}\, d\mathcal{D}_{\text{EAS}}. \end{aligned} \tag{2.2}$$

Lorsque nous conditionnons sur $\mathcal{D},$ les valeurs de $(\mathcal{D}_{\text{syn}}, \mathcal{D}_{\text{EAS}}, \mathcal{P}_{\text{pseudo}})$ ne fournissent aucun renseignement supplémentaire à propos de Q. Nous pouvons donc simplifier $P(Q \mid \mathcal{D}_{\text{syn}}, \mathcal{D}_{\text{EAS}}, \mathcal{P}_{\text{pseudo}}, \mathcal{D}) = P(Q \mid \mathcal{D})$. Lorsque nous conditionnons sur $\mathcal{P}_{\text{pseudo}},$ les valeurs de $(\mathcal{D}_{\text{rep}}, \mathcal{D}_{\text{syn}})$ ne fournissent aucun renseignement supplémentaire à propos de $\mathcal{D}$. Nous pouvons donc simplifier $P(\mathcal{D} \mid \mathcal{D}_{\text{syn}}, \mathcal{D}_{\text{EAS}}, \mathcal{P}_{\text{pseudo}}) = P(\mathcal{D} \mid \mathcal{P}_{\text{pseudo}})$. Lorsque nous conditionnons sur $\mathcal{D}_{\text{EAS}},$ la valeur de $\mathcal{D}_{\text{syn}}$ ne fournit aucun renseignement à propos de $\mathcal{P}_{\text{pseudo}}$. Donc, $P(\mathcal{P}_{\text{pseudo}} \mid \mathcal{D}_{\text{syn}}, \mathcal{D}_{\text{EAS}}) = P(\mathcal{P}_{\text{pseudo}} \mid \mathcal{D}_{\text{EAS}})$. Après avoir fait une réorganisation pour faciliter l'interprétation, nous exprimons de nouveau (2.2) comme suit :

$$\begin{aligned} P(Q \mid \mathcal{D}_{\text{syn}}) = \int \Big[\int \Big[\int P(Q \mid \mathcal{D})\, P(\mathcal{D} \mid \mathcal{P}_{\text{pseudo}})\, d\mathcal{D} \Big] P(\mathcal{P}_{\text{pseudo}} \mid \mathcal{D}_{\text{EAS}})\, d\mathcal{P}_{\text{pseudo}} \Big] \\ P(\mathcal{D}_{\text{EAS}} \mid \mathcal{D}_{\text{syn}})\, d\mathcal{D}_{\text{EAS}}. \end{aligned} \tag{2.3}$$

Nous commençons par $P(Q \mid \mathcal{P}_{\text{pseudo}}) = \int P(Q \mid \mathcal{D})\, P(\mathcal{D} \mid \mathcal{P}_{\text{pseudo}})\, d\mathcal{D}.$ Nous présumons que, pour un grand $M,$ il s'agit approximativement d'une distribution normale. Cela devrait être raisonnable dans les grands échantillons, qui sont habituels dans les cas où les organismes souhaitent diffuser des données pour le public. Nous avons uniquement besoin que la distribution *a posteriori* de Q soit normale, et non pas la

distribution des variables de l'enquête elles-mêmes; en effet, les données sous-jacentes peuvent être catégoriques. Nous notons que les méthodes d'inférence ne sont pas conçues pour des quantités comme les médianes ou d'autres quantiles; les méthodes d'inférence pour de telles quantités sont un sujet pour de plus amples recherches.

Nous n'avons besoin que de moyennes et de variances pour caractériser des distributions d'échantillonnage normales. Nous nous concentrons donc sur l'estimation des distributions des deux premiers moments. Pour $m = 1, \ldots, M$, supposons que $Q^{(m)}$ serait la valeur calculée de Q si nous avions accès à $\mathcal{P}^{(m)}$. Rubin (1987) montre que

$$(Q \mid \mathcal{P}_{\text{pseudo}}) \sim t_{M-1}\left(\bar{Q}, (1 + M^{-1})\, B\right), \tag{2.4}$$

où $\bar{Q} = \sum_m Q^{(m)}/M$ et $B = \sum_m (Q^{(m)} - \bar{Q})^2/(M-1)$. Dans ce cas-ci, $t_\nu(\mu, \sigma^2)$ désigne une distribution t à ν degrés de liberté, de position μ et de variance σ^2. Dans les calculs, par souci de commodité, nous évaluons approximativement la distribution t dans (2.4) comme une distribution normale, ce qui devrait être raisonnable pour un M plutôt grand.

Nous nous penchons ensuite sur $P(\mathcal{P}_{\text{pseudo}} \mid \mathcal{D}_{\text{EAS}})$. Dans ce cas-ci, nous avons uniquement besoin de $P(\bar{Q}, B \mid \mathcal{D}_{\text{EAS}})$. Pour $m = 1, \ldots, M$, supposons que $q^{(m)}$ est l'estimation de $Q^{(m)}$ et que $v^{(m)}$ est l'estimation de la variance échantillonnale associée à $q^{(m)}$; nous pourrions les calculer si nous avions accès à $\mathcal{D}^{(m)}$. Nous supposons que $\{q^{(m)}, v^{(m)} : m = 1, \ldots, M\}$ sont valides dans le sens suivant.

1) Pour chaque m, $q^{(m)}$ est approximativement sans biais pour $Q^{(m)}$ et suit une distribution asymptotique normale, en ce qui a trait à l'échantillonnage répété à partir de la pseudo-population $\mathcal{P}^{(m)}$ ayant la variance échantillonnale $V^{(m)}$. Autrement dit, nous avons $(q^{(m)} \mid \mathcal{P}^{(m)}) \sim N(Q^{(m)}, V^{(m)})$.
2) L'estimation de la variance échantillonnale $v^{(m)}$ est approximativement sans biais pour $V^{(m)}$ et la variabilité d'échantillonnage dans $v^{(m)}$ est négligeable. Autrement dit, $(v^{(m)} \mid \mathcal{P}^{(m)}) \approx V^{(m)}$.
3) La variation dans $V^{(m)}$ parmi les M pseudo-populations est négligeable; soit, $V^{(m)} \approx V \approx \bar{v}$, où $\bar{v} = \sum_m v^{(m)}/M$.

En utilisant les arguments bayésiens standards fondés sur ces distributions d'échantillonnage, il s'en suit que

$$(Q^{(m)} \mid q^{(m)}, \bar{v}) \sim N(q^{(m)}, \bar{v}) \tag{2.5}$$

$$(\bar{Q} \mid \bar{q}, \bar{v}) \sim N(\bar{q}, \bar{v}/M), \tag{2.6}$$

où $\bar{q} = \sum_m q^{(m)}/M$.

Pour obtenir la distribution de $(Q \mid \mathcal{D}_{\text{EAS}})$, nous intégrons la distribution dans (2.4), que nous évaluons approximativement comme une distribution normale, en ce qui a trait aux distributions de $\bar{Q}$ et de B. Nous

avons uniquement besoin des deux premiers moments, car la distribution en résultant est une distribution normale. Nous avons

$$E(Q \mid \mathcal{D}_{\text{EAS}}) = E(E(Q \mid \bar{Q}) \mid \mathcal{D}_{\text{EAS}}) = E(\bar{Q} \mid \mathcal{D}_{\text{EAS}}) = \bar{q}. \tag{2.7}$$

Nous avons également

$$\begin{aligned} \text{Var}(Q \mid \mathcal{D}_{\text{EAS}}) &= E(\text{Var}(Q \mid \mathcal{P}_{\text{pseudo}}) \mid \mathcal{D}_{\text{EAS}}) + \text{Var}(E(Q \mid \mathcal{P}_{\text{pseudo}}) \mid \mathcal{D}_{\text{EAS}}) \\ &= (1 + M^{-1}) E(B \mid \mathcal{D}_{\text{EAS}}) + \bar{v}/M. \end{aligned} \tag{2.8}$$

Il s'agit de l'estimateur de la variance dans Raghunathan et coll. (2003), que les analystes utiliseraient si l'organisme diffusait $\mathcal{D}_{\text{EAS}}$ sous la forme de fichiers de données à grande diffusion. Cependant, puisque nous prenons une mesure supplémentaire qui consiste à remplacer chaque $\mathcal{D}^{(m)}$ par des valeurs simulées, nous devons établir une moyenne sur les distributions de $(\bar{q}, \bar{v}, B)$. Le résultat dépend de notre décision d'utiliser *SynRep-R* ou *SynRep-1*, comme nous le décrivons maintenant.

2.2.1 Calcul pour *SynRep-R*

Pour chaque $\mathcal{D}_{\text{syn}}^{(m,r)}$, supposons que $q_{\text{syn}}^{(m,r)}$ est l'estimation ponctuelle de Q et supposons que $v_{\text{syn}}^{(m,r)}$ est l'estimation de la variance associée à $q_{\text{syn}}^{(m,r)}$. L'analyste calcule $q_{\text{syn}}^{(m,r)}$ et $v_{\text{syn}}^{(m,r)}$ en agissant comme si $\mathcal{D}_{\text{syn}}^{(m,r)}$ est les données recueillies obtenues à l'aide d'un échantillon aléatoire simple de taille n tiré dans la population. L'analyste doit calculer les quantités suivantes.

$$\bar{q}_{\text{syn}}^{(m)} = \sum_{r=1}^{R} q_{\text{syn}}^{(m,r)} / R \tag{2.9}$$

$$\bar{q}_{\text{syn}} = \sum_{m=1}^{M} \bar{q}_{\text{syn}}^{(m)} / M \tag{2.10}$$

$$b_{\text{syn}} = \sum_{m=1}^{M} (\bar{q}_{\text{syn}}^{(m)} - \bar{q}_{\text{syn}})^2 / (M-1) \tag{2.11}$$

$$w_{\text{syn}}^{(m)} = \sum_{r=1}^{R} (q_{\text{syn}}^{(m,r)} - \bar{q}_{\text{syn}}^{(m)})^2 / (R-1) \tag{2.12}$$

$$\bar{w}_{\text{syn}} = \sum_{m=1}^{M} w_{\text{syn}}^{(m)} / M \tag{2.13}$$

$$\bar{v}_{\text{syn}} = \sum_{m=1}^{M} \sum_{r=1}^{R} v_{\text{syn}}^{(m,r)} / MR. \tag{2.14}$$

Nous terminons maintenant le calcul de la distribution *a posteriori* pour $(Q \mid \mathcal{D}_{\text{syn}})$ dans la méthode *SynRep-R*. Pour ce faire, nous supposons des approximations normales de grands échantillons pour les

distributions d'échantillonnage des estimations ponctuelles. Plus précisément, pour tous les (m, r), nous présumons que

$$q_{\text{syn}}^{(m,r)} \sim N(q^{(m)}, W^{(m)}), \tag{2.15}$$

où $W^{(m)}$ est la variance échantillonnale pour $q_{\text{syn}}^{(m,r)}$ sur les tirages de données synthétiques de $\mathcal{D}^{(m)}$. La normalité devrait être raisonnable lorsque n est grand. En présumant des distributions *a priori* diffuses et un conditionnement sur $W^{(m)}$, nous avons

$$(q^{(m)} \mid \mathcal{D}_{\text{syn}}^{(m,1)}, \ldots, \mathcal{D}_{\text{syn}}^{(m,R)}, W^{(m)}) \sim N(\bar{q}_{\text{syn}}^{(m)}, W^{(m)}/R) \tag{2.16}$$

$$(\bar{q} \mid \mathcal{D}_{\text{syn}}, \bar{W}) \sim N(\bar{q}_{\text{syn}}, \bar{W}/MR), \tag{2.17}$$

où $\bar{W} = \sum_m W^{(m)}/M$.

Ayant maintenant déterminé les distributions pour les estimateurs ponctuels, nous rassemblons tous les éléments pour la distribution *a posteriori* de Q. Puisque tous les composantes sont des distributions normales, $P(Q \mid \mathcal{D}_{\text{syn}})$ est une distribution normale. Ainsi, pour l'espérance, nous utilisons (2.7) et (2.17) pour obtenir

$$E(Q \mid \mathcal{D}_{\text{syn}}) = (E(Q \mid \mathcal{D}_{\text{EAS}}) \mid \mathcal{D}_{\text{syn}}) = E(\bar{q} \mid \mathcal{D}_{\text{syn}}) = \bar{q}_{\text{syn}}. \tag{2.18}$$

Pour la variance, nous écrivons d'abord la variance en termes de $(B, \bar{v}, \bar{W})$ et substituons ensuite des estimations ponctuelles de ces termes. Pour insister sur l'utilisation de $(B, \bar{v}, \bar{W})$, nous écrivons

$$\begin{aligned} \text{Var}(Q \mid \mathcal{D}_{\text{syn}}, B, \bar{v}_M, \bar{W}) &= E(((1 + M^{-1})\, B + \bar{v}/M) \mid \mathcal{D}_{\text{syn}}, B, \bar{v}, \bar{W}) + \text{Var}(\bar{q} \mid \mathcal{D}_{\text{syn}}, B, \bar{v}, \bar{W}) \\ &= (1 + M^{-1})\, B + \bar{v}/M + \bar{W}/MR. \end{aligned} \tag{2.19}$$

Nous définissons maintenant les estimations pour $(B, \bar{v}, \bar{W})$, que nous substituons dans l'équation (2.19). Pour $\bar{v}$, nous supposons que $\bar{v}_{\text{syn}} \approx \bar{v}$. Cette hypothèse provient de la justification de Raghunathan et coll. (2003), qui affirment que c'est le cas lorsque les données synthétiques sont générées à partir de la même distribution sous-jacente que les données utilisées pour ajuster les modèles.

Pour $\bar{W}$, nous notons que (2.15) suppose que, pour $m = 1, \ldots, M$,

$$\frac{(R-1)\, w_{\text{syn}}^{(m)}}{W^{(m)}} \sim \chi_{R-1}^2. \tag{2.20}$$

Nous présumons ensuite que chaque $W^{(m)} \approx \bar{W}$. Cette hypothèse correspond à une hypothèse similaire fournie dans Reiter (2004) concernant la variabilité des variances *a posteriori*. Essentiellement, comme on l'indique dans Reiter (2004), cette hypothèse découle de l'observation selon laquelle la variabilité dans les variances *a posteriori* est habituellement inférieure à la variabilité dans les espérances *a posteriori*. Au moyen de cette hypothèse et de (2.20), nous avons

$$\sum_{m=1}^{M} \frac{(R-1)\, w_{\text{syn}}^{(m)}}{\bar{W}} \sim \chi^2_{M(R-1)}. \tag{2.21}$$

Nous avons ainsi

$$\mathrm{E}\left(\sum_{m=1}^{M} \frac{(R-1)\, w_{\text{syn}}^{(m)}}{\bar{W}} \right) = M(R-1). \tag{2.22}$$

En utilisant une approche de type méthode des moments pour évaluer approximativement $\bar{W}$, nous obtenons $\bar{W} \approx \bar{w}_{\text{syn}}$.

Pour évaluer approximativement B, nous notons que la distribution d'échantillonnage d'un $\bar{q}_{\text{syn}}^{(m)}$ généré aléatoirement sur toutes les étapes du processus de génération de données est $N(Q, B + \bar{v} + \bar{W}/R)$. En utilisant ce fait, nous avons

$$\frac{\sum_{m=1}^{M} (\bar{q}_{\text{syn}}^{(m)} - \bar{q}_{\text{syn}})^2}{B + \bar{v} + \bar{W}/R} \sim \chi^2_{M-1}, \tag{2.23}$$

de sorte que

$$\mathrm{E}\left(\frac{\sum_{m=1}^{M} (\bar{q}_{\text{syn}}^{(m)} - \bar{q}_{\text{syn}})^2}{B + \bar{v} + \bar{W}/R} \right) = M - 1. \tag{2.24}$$

En utilisant une approche de type méthode des moments et la définition de b_{syn} dans l'équation (2.11), et l'estimation de substitution $\bar{w}_{\text{syn}}$ pour $\bar{W}$, nous avons $b_{\text{syn}} \approx B + \bar{v}_{\text{syn}} + \bar{w}_{\text{syn}}/R$, de sorte que $B \approx b_{\text{syn}} - \bar{v}_{\text{syn}} - \bar{w}_{\text{syn}}/R$.

En rassemblant tous les éléments, nous pouvons évaluer approximativement $\mathrm{Var}(Q \mid \mathcal{D}_{\text{syn}})$ à l'aide de l'estimation T_r, où

$$\begin{aligned} T_r &= (1 + M^{-1})\left(b_{\text{syn}} - \bar{v}_{\text{syn}} - \bar{w}_{\text{syn}}/R\right) + \bar{v}_{\text{syn}}/M + \bar{w}_{\text{syn}}/MR \\ &= (1 + M^{-1})\, b_{\text{syn}} - \bar{v}_{\text{syn}} - \bar{w}_{\text{syn}}/R. \end{aligned} \tag{2.25}$$

Nous calculons des intervalles d'environ 95 % pour Q comme $\bar{q}_{\text{syn}} \pm t_{0{,}975;\, M-1} \sqrt{T_r}$. La distribution t est une approximation simple basée sur les degrés de liberté dans (2.4). À l'instar de l'estimateur de la variance dans Raghunathan et coll. (2003), l'estimation T_r peut être négative, particulièrement pour les petits M. Comme ajustement ponctuel lorsque $T_r < 0$, nous recommandons de remplacer B par $\bar{v}$ dans l'équation (2.19) et d'utiliser $T_r^* = (1 + 2/M)\, \bar{v}_{\text{syn}} + \bar{w}_{\text{syn}}/MR$.

2.2.2 Calcul pour *SynRep-1*

En présence de grands M et R, *SynRep-R* donne lieu à de nombreux ensembles de données synthétiques, ce qui pourrait être indésirable du point de vue de l'organisme et des analystes de données

secondaires. Les organismes voudront plutôt utiliser *SynRep-1*. Pour obtenir des inférences pour Q dans ce scénario, nous tirons avantage de la méthodologie de Raab et coll. (2018), qui ont observé que lorsque les données sources proviennent d'un échantillon aléatoire simple, comme c'est le cas pour chaque $\mathcal{D}^{(m)}$, nous pouvons obtenir des estimations de la variance valides ayant des répliques simples en ajustant les règles de combinaison. Nous décrivons maintenant le calcul.

Pour $m = 1, \ldots, M$, supposons que $q_{\text{syn}}^{(m)}$ est l'estimation ponctuelle de Q calculée en utilisant $\mathcal{D}_{\text{syn}}^{(m)}$ et supposons que $v_{\text{syn}}^{(m)}$ est la variance estimée associée à $q_{\text{syn}}^{(m)}$. L'analyste calcule chaque $(q_{\text{syn}}^{(m)}, v_{\text{syn}}^{(m)})$ en agissant comme si $\mathcal{D}_{\text{syn}}^{(m)}$ était un EAS de taille n tiré de la population. Nous avons besoin des quantités suivantes aux fins d'inférence. Pour économiser sur les notations, nous réutilisons certaines des notations présentées à la section 2.2.1.

$$\overline{q}_{\text{syn}} = \sum_{m=1}^{M} q_{\text{syn}}^{(m)} / M \tag{2.26}$$

$$b_{\text{syn}} = \sum_{m=1}^{M} (q_{\text{syn}}^{(m)} - \overline{q}_{\text{syn}})^2 / (M-1) \tag{2.27}$$

$$\overline{v}_{\text{syn}} = \sum_{m=1}^{M} v_{\text{syn}}^{(m)} / M. \tag{2.28}$$

Les paires d'équations (2.26) et (2.10), (2.27) et (2.11), et (2.28) et (2.14) peuvent être considérées comme équivalentes lorsque $R = 1$.

Afin d'effectuer le calcul pour *SynRep-1*, nous suivons la logique de Raab et coll. (2018) et supposons que $q_{\text{syn}}^{(m)} \sim N(q^{(m)}, V^{(m)})$. En supposant que $V^{(m)} \approx \overline{v}$ pour tout m, nous avons

$$(q^{(m)} \mid \mathcal{D}_{\text{syn}}^{(m)}, \overline{v}) \sim N(q_{\text{syn}}^{(m)}, \overline{v}) \tag{2.29}$$

$$(\overline{q} \mid \mathcal{D}_{\text{syn}}, \overline{v}) \sim N(\overline{q}_{\text{syn}}, \overline{v}/M). \tag{2.30}$$

Nous notons toutefois qu'il ne faut pas supposer que $B \approx \overline{v}$ aussi. Comme $\mathcal{D}$ est un échantillon complexe, il produit des variances échantillonnales qui pourraient être différentes des variances d'échantillonnage aléatoire simple associées à $\mathcal{D}_{\text{EAS}}$.

Puisque tous les composantes sont des distributions normales approximatives, $P(Q \mid \mathcal{D}_{\text{syn}})$ est aussi une distribution normale approximative. Pour son espérance, nous utilisons (2.7) et (2.30) pour obtenir

$$E(Q \mid \mathcal{D}_{\text{syn}}) = E(E(Q \mid \mathcal{D}_{\text{EAS}}) \mid \mathcal{D}_{\text{syn}}) = E(\overline{q} \mid \mathcal{D}_{\text{syn}}) = \overline{q}_{\text{syn}}. \tag{2.31}$$

Pour sa variance, à l'instar de *SynRep-R*, nous écrivons la variance en termes de $(B, \overline{v})$ et ajoutons ensuite par substitution des estimations ponctuelles de ces termes. Nous avons

$$\begin{aligned}\mathrm{Var}(Q \mid \mathcal{D}_{\mathrm{syn}}, B, \overline{v}) &= E((1+M^{-1})\,B + \overline{v}/M) \mid \mathcal{D}_{\mathrm{syn}}, B, \overline{v}) + \mathrm{Var}(\overline{q} \mid \mathcal{D}_{\mathrm{syn}}, B, \overline{v}) \\ &= (1+M^{-1})\,B + \overline{v}/M + \overline{v}/M = (1+M^{-1})\,B + 2\overline{v}/M.\end{aligned} \tag{2.32}$$

Nous définissons maintenant les estimations pour $(B, \overline{v})$ à ajouter par substitution dans l'équation (2.32). Pour $\overline{v}$, nous utilisons $\overline{v}_{\mathrm{syn}}$ défini dans l'équation (2.28). Cela devrait être raisonnable, puisque nous remplaçons l'ensemble complet de chaque $\mathcal{D}^{(m)}$ par des valeurs synthétiques. Pour évaluer approximativement B, nous notons que la distribution d'échantillonnage d'un $q_{\mathrm{syn}}^{(m)}$ généré aléatoirement sur toutes les étapes du processus de génération de données est $N(Q, B + 2\overline{v})$. En utilisant ce fait, nous avons

$$\frac{\sum_{m=1}^{M}(q_{\mathrm{syn}}^{(m)} - \overline{q}_{\mathrm{syn}})^2}{B + 2\overline{v}} \sim \chi^2_{M-1}, \tag{2.33}$$

de sorte que

$$\mathrm{E}\left(\frac{\sum_{m=1}^{M}(q_{\mathrm{syn}}^{(m)} - \overline{q}_{\mathrm{syn}})^2}{B + 2\overline{v}}\right) = M - 1. \tag{2.34}$$

En utilisant une approche de type méthode des moments et la définition de b_{syn} dans l'équation (2.27), nous avons $b_{\mathrm{syn}} \approx B + 2\overline{v}_{\mathrm{syn}}$, de sorte que $B \approx b_{\mathrm{syn}} - 2\overline{v}_{\mathrm{syn}}$.

Nous pouvons donc évaluer approximativement $\mathrm{Var}(Q \mid \mathcal{D}_{\mathrm{syn}})$ à l'aide de l'estimation T_m, où

$$T_m = \left(1 + M^{-1}\right) b_{\mathrm{syn}} - 2\overline{v}_{\mathrm{syn}}. \tag{2.35}$$

Nous calculons des intervalles d'environ 95 % pour Q comme $\overline{q}_{\mathrm{syn}} \pm t_{0{,}975;\,M-1}\sqrt{T_m}$. Lorsque $T_m < 0$, en tant qu'estimation de la variance ponctuelle, nous remplaçons B par $\overline{v}$ dans l'équation (2.32) et utilisons $T_m^* = \left(1 + 3/M\right)\overline{v}_{\mathrm{syn}}$.

3. Études par simulations

Dans la présente section, nous présentons les études par simulations pour illustrer les propriétés d'échantillonnage répété des méthodes d'inférence de la section 2.2 pour diverses quantités de population finie.

3.1 Plan de sondage en simulation

Nous construisons une population finie en fonction des données des échantillons de microdonnées à grande diffusion de l'American Community Survey de 2021 (Bureau du recensement des États-Unis, 2021). Le fichier comprend 3 252 599 personnes, que nous traitons comme une population de taille N. Le fichier comporte également des poids au niveau de la personne (nommés « PWGTP » dans le fichier de données). Nous ne les traitons pas comme des poids d'enquête en tant que tels; nous les traitons plutôt comme des

variables de taille x_i, où $i = 1, \ldots, N$, à des fins d'utilisation dans un échantillonnage avec probabilité proportionnelle à la taille (PPT). Nous utilisons aussi ces mesures de taille construites pour générer deux variables d'enquête, (y_{i1}, y_{i2}), où $i = 1, \ldots, N$. Plus précisément, nous supposons que chaque y_{i1} est une variable binaire échantillonnée à partir d'une distribution de Bernoulli ayant une probabilité $\Pr(y_{i1} = 1) = \exp(-7 + 2\log x_i) / (1 + \exp(-7 + 2\log x_i))$. Nous supposons que chaque y_{i2} est une variable continue échantillonnée à partir d'une distribution normale ayant une moyenne de $20 + 50y_{i1}$ et un écart-type de 50. Nous estimons la proportion de population finie $\overline{Y}_1 = \sum_{i=1}^{N} y_{i1} / N \approx 0{,}765$, la moyenne de population finie $\overline{Y}_2 = \sum_{i=1}^{N} y_{i2} / N \approx 58{,}2$ et le coefficient de régression de population finie de Y_1 dans la régression linéaire de Y_2 sur Y_1, qui est $\beta \approx 50$.

À partir de cette population, nous échantillonnons $\mathcal{D}$ en utilisant un échantillon avec PPT de taille $n = 500$ unités d'enquête, en établissant que $\pi_i = n x_i / \sum_{i=1}^{N} x_i$ et en utilisant la fonction « ppss » dans le paquet R « pps » (Gambino, 2021). Selon ce plan d'échantillonnage avec PPT, nous nous attendons à ce que les inférences non pondérées reposant sur $\mathcal{D}$ soient fortement biaisées pour $(\overline{Y}_1, \overline{Y}_2)$, mais peut-être pas autant pour β. Nous répétons le processus d'échantillonnage pour créer 1 000 réalisations indépendantes de $\mathcal{D}$.

Pour chaque $\mathcal{D}$, nous mettons en œuvre *SynRep-R* et *SynRep-1* selon divers (M, R). Plus précisément, nous examinons $(M = 4, R = 5)$, $(M = 10, R = 5)$, $(M = 50, R = 5)$, $(M = 10, R = 10)$, $(M = 10, R = 25)$ et $(M = 10, R = 50)$. Le choix de R ne touche que *SynRep-R*. Nous mettons en œuvre le BBPFP en utilisant le paquet « polyapost » dans R (Meeden, Lazar et Geyer, 2020), pour créer des pseudo-populations $(\mathcal{P}^{(1)}, \ldots, \mathcal{P}^{(M)})$ comportant chacune 25 000 personnes. À partir de chaque $\mathcal{P}^{(m)}$ où $m = 1, \ldots, M$, nous tirons un échantillon aléatoire simple de taille n pour produire un $\mathcal{D}^{(m)}$ correspondant. Pour que chaque réplique de données synthétiques découle de chaque $\mathcal{D}^{(m)}$, nous échantillonnons n valeurs synthétiques pour Y_1 en utilisant une distribution de Bernoulli pour laquelle la probabilité est fixée à la proportion empirique de Y_1 dans $\mathcal{D}^{(m)}$. Nous échantillonnons les valeurs synthétiques correspondantes de Y_2 à partir des distributions normales ayant des moyennes égales aux valeurs prédites à partir de la régression de Y_2 sur Y_1, calculée en utilisant les valeurs synthétiques de Y_1 et les estimations non biaisées des coefficients calculés au moyen de $\mathcal{D}^{(m)}$, et une variance égale à l'estimation non biaisée de la variance de régression calculée au moyen de $\mathcal{D}^{(m)}$.

Pour faciliter l'évaluation des rendements d'échantillonnage répété de *SynRep-1* et de *SynRep-R*, nous utilisons aussi les résultats calculés à l'aide de $\mathcal{P}_{\text{pseudo}}$ et de $\mathcal{D}_{\text{EAS}}$. Plus précisément, dans chacune des 1 000 exécutions de simulation, nous définissons *Pseudo-pop* en tant que procédure qui repose sur un estimateur ponctuel de $\overline{Q}$ et un estimateur de la variance de $(1 + 1/M) B$ calculé à l'aide des pseudo-populations générées par le BBPFP $(\mathcal{P}^{(1)}, \ldots, \mathcal{P}^{(M)})$. Nous définissons *Pseudo-EAS* en tant que procédure qui repose sur un estimateur ponctuel de $\overline{q}$ et un estimateur de la variance de Raghunathan et coll. (2003) calculé à l'aide de $(\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(M)})$. Comme comparaison avec ce qui se produit si nous ignorons le plan de sondage complètement, nous définissons *EASsyn* en tant que procédure qui génère des données synthétiques au moyen : i) de la proportion de l'échantillon non pondéré pour Y_1 en tant que probabilité de Bernoulli pour générer n valeurs synthétiques de Y_1; ii) des estimations non pondérées des paramètres dans la régression de Y_2 sur Y_1 en tant que paramètres de la distribution normale pour générer les n valeurs synthétiques correspondantes de Y_2.

Nous évaluons aussi les rendements d'échantillonnage répété des approches de pseudo-vraisemblance pour produire des données entièrement synthétiques. Pour chaque modèle de synthèse, c'est-à-dire les modèles de régression linéaire et de Bernoulli, nous commençons par une fonction de vraisemblance définie comme le produit des contributions de chaque personne dans $\mathcal{D}$. Nous créons la pseudo-vraisemblance en augmentant la contribution de chaque personne d'une puissance déterminée par le poids d'enquête de la personne. Nous utilisons ces pseudo-vraisemblances pondérées pour estimer les paramètres du modèle de synthèse. Nous mettons en œuvre cette approche en utilisant le logiciel *Stan* (Stan Development Team, 2024), qui peut générer des échantillons *a posteriori* des paramètres du modèle d'après les fonctions de vraisemblance précisées par l'utilisateur. Nous exécutons *Stan* pour créer quatre chaînes de 4 000 itérations et supprimons les 2 000 premières itérations dans le cadre du rodage. Nous échantillonnons aléatoirement un des tirages en résultant et utilisons ses valeurs de paramètre dans les modèles de régression linéaire et de Bernoulli pour générer les données synthétiques. Nous répétons ce processus M fois et appliquons les règles d'inférence de Raghunathan et coll. (2003). Nous appelons cette méthode *Wtreg*. Nous notons que Kim et coll. (2021) utilisent l'estimateur de la variance dans (2.8) de Raghunathan et coll. (2003) où $\overline{v} = 0$. Kim et coll. (2021) diffusent des populations synthétiques (où $\overline{v} = 0$) plutôt que des échantillons synthétiques (où $\overline{v} > 0$).

Nous tenons aussi compte d'une modification de *Wtreg* pour nous pencher sur une sous-estimation potentielle de la variabilité dans les tirages de paramètre. Nous appelons cette méthode *Wtreg-Boot*. Tout d'abord, nous tirons un échantillon bootstrap de taille n de $\mathcal{D}$. Nous construisons les fonctions de pseudo-vraisemblance en utilisant les données soumises à la méthode bootstrap et le poids d'enquête calibré pour chaque personne rééchantillonnée. En utilisant cette fonction de pseudo-vraisemblance, nous générons et analysons ensuite les données synthétiques en suivant les étapes décrites pour *Wtreg*.

Finalement, nous définissons *Direct* comme reposant sur la moyenne d'échantillon non pondéré et l'écart-type de $\mathcal{D}$, c'est-à-dire en ignorant les poids d'enquête, et *HT* comme reposant sur l'estimateur de Horvitz et Thompson (1952) et sa variance estimée reposant sur $\mathcal{D}$. Nous utilisons ces deux dernières procédures pour évaluer l'importance de tenir compte du plan de sondage dans les inférences avec $\mathcal{D}$.

Supposons que l'indice en exposant s constitue les résultats de l'exécution de simulation s, où $s = 1, \ldots, 1\,000$. Pour tout estimateur $\hat{q}$ pour toutes les méthodes que nous examinons, nous calculons le biais de pourcentage, $100 \sum_{s=1}^{1\,000} (\hat{q}^s - Q) \,/\, (1\,000\, Q)$. Nous calculons la proportion des 1 000 intervalles de confiance à 95 % en fonction de $\hat{q}$ et son estimation de la variance correspondante qui couvrent Q. Nous calculons aussi le ratio de la variance empirique des 1 000 valeurs de $\hat{q}$ à la variance empirique des 1 000 valeurs de l'estimateur ponctuel de *HT*. Pour évaluer l'exactitude des estimateurs de la variance, pour chaque méthode, nous calculons le ratio de la moyenne des 1 000 estimations de la variance sur sa variance empirique correspondante. Finalement, pour examiner la stabilité de l'estimateur de la variance pour chaque méthode, nous calculons l'écart-type des 1 000 estimations de la variance. Nous présentons les résultats pour les quatre premières quantités dans le texte principal et ceux pour la dernière quantité dans l'annexe.

3.2 Résultats

Nous examinons d'abord les propriétés de *SynRep-R* et de *SynRep-1* pour les divers scénarios de (M, R). La figure 3.1, la figure 3.2 et la figure 3.3 présentent les résultats pour $\overline{Y}_1$, $\overline{Y}_2$ et β, respectivement, pour ces

deux méthodes ainsi que pour *Pseudo-pop* et *Pseudo-EAS*. Les quatre méthodes offrent des estimations ponctuelles approximativement sans biais des trois quantités de population finie, et les biais en pourcentage simulés se situent généralement autour de 1 % ou moins. Ces petits biais proviennent principalement de l'étape qui consiste à compléter les populations, alors que les biais dans *Pseudo-pop* sont près des biais dans les trois autres méthodes. Comme prévu, comparativement à la variance pour *HT*, les variances simulées augmentent graduellement à mesure que M diminue. En maintenant $M = 10$ constant, le fait de diminuer R a tendance à augmenter les variances simulées, même si les effets sont moins prononcés que ceux découlant de la diminution de M. La variabilité des résultats dans *SynRep-1* selon un M fixe reflète l'erreur de Monte Carlo. Ensemble, ces résultats laissent supposer qu'il est préférable d'augmenter M plutôt que R lorsque MR demeure constant. Par exemple, lorsque nous comparons *SynRep-R* selon $(M = 10, R = 5)$ à *SynRep-1* selon $M = 50,$ ce dernier a tendance à donner lieu à une variance empirique plus petite et à des des taux de couverture plus près de la valeur nominale. Des avantages similaires apparaissent lorsque l'on compare *SynRep-R* selon $(M = 10, R = 25)$ à *SynRep-R* selon $(M = 50, R = 5).$ Cette conclusion correspond aux résultats de Reiter (2008), qui a considéré un compromis similaire pour l'imputation multiple à plusieurs critères de classification pour les données manquantes et partiellement synthétiques. Nous notons que l'utilisation de valeurs plus élevées de M offre également une variabilité plus petite des variances estimées, comme on le montre à l'annexe.

Figure 3.1 Propriétés d'échantillonnage répété de *SynRep-1* et de *SynRep-R* pour $\bar{Y}_1$ sous différents nombres d'échantillons synthétiques (M) et de répliques (R), selon un plan de sondage avec probabilité proportionnelle à la taille.

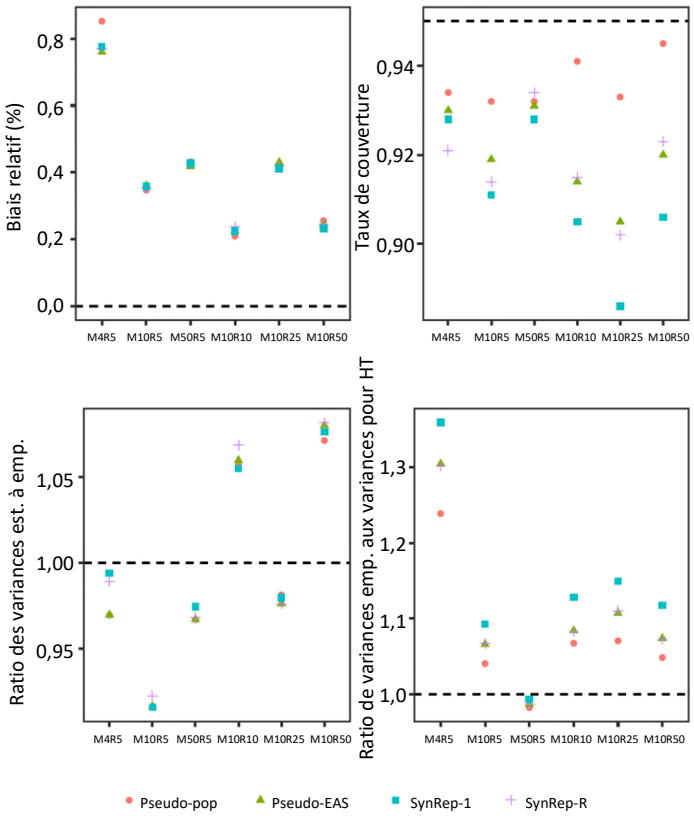
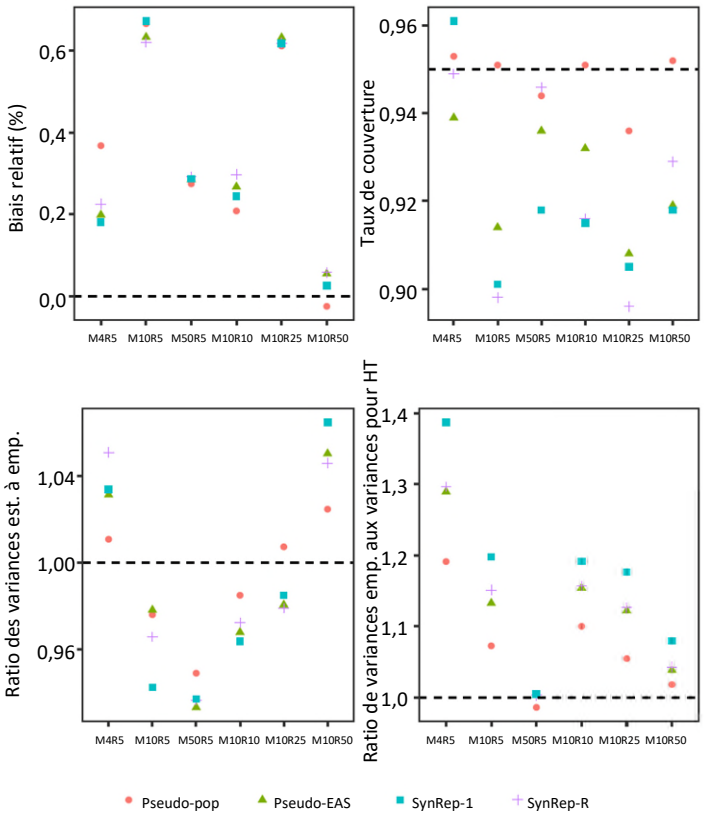


Figure 3.2 Propriétés d'échantillonnage répété de *SynRep-1* et de *SynRep-R* pour $\bar{Y}_2$ sous différents nombres d'échantillons synthétiques (M) et de répliques (R), selon un plan de sondage avec probabilité proportionnelle à la taille.



En comparant les ratios des variances empiriques aux variances pour *HT*, nous pouvons voir l'effet sur l'efficacité des étapes dans le processus de synthèse. Les variances augmentent en général à mesure que nous passons de *Pseudo-pop* à *Pseudo-EAS*, à *SynRep-R* ou à *SynRep-1*; autrement dit, elles augmentent à mesure que nous ajoutons plus d'étapes qui concernent le caractère aléatoire. Les variances pour *SynRep-R* sont en général légèrement plus petites que celles pour *SynRep-1*, ce qui indique l'avantage aux fins d'efficacité des renseignements supplémentaires provenant de MR plutôt que de M ensembles de données synthétiques. Nous notons que l'inflation de la variance occasionnée par l'utilisation des procédures de données synthétiques, comparativement à *HT*, disparaît en grande partie lorsque $M = 50$.

Pour les quatre méthodes de données synthétiques, les estimations de la variance moyenne sont raisonnablement semblables aux variances empiriques. De nouveau, les disparités par rapport aux ratios de 1 semblent découler principalement de l'étape qui consiste à compléter les populations. Les taux de couverture de l'intervalle de confiance varient d'un seuil inférieur s'établissant à 88 % à un seuil supérieur s'établissant à 96 %, et la plupart sont légèrement inférieurs à la valeur nominale. Les taux de couverture pour *SynRep-R* et *SynRep-1* ont tendance à être les plus élevés lorsque $M = 50$, ce qui indique encore une fois les avantages d'utiliser un plus grand M. Pour $M \geq 10$, les taux de couverture pour *SynRep-R* ont

tendance à être supérieurs à ceux pour *SynRep-1*, bien que la différence soit généralement de seulement un point ou deux.

Les règles de combinaison dans les équations (2.25) et (2.35) donnent lieu à des estimations de la variance négatives, comme il est démontré au tableau 3.1. Dans les simulations, nous utilisons T_r^* et T_m^* pour produire des intervalles de confiance, au besoin. À mesure que M augmente, le nombre d'estimations de la variance négatives diminue. En fait, quand $M = 50,$ toutes les estimations de la variance sont positives, ce qui offre un soutien supplémentaire pour rendre M élevé. Les estimations de b_{syn} deviennent moins variables à mesure que M augmente et elles contribuent à éviter les variances négatives. Les taux de variance négatifs ont tendance à être plus faibles pour *SynRep-R* que pour *SynRep-1*, ce qui montre les avantages des ensembles de données plus grands pour estimer les paramètres de variance. Bien qu'ils ne soient pas indiqués dans le tableau 3.1, les taux de variance négatifs quand $M = 10$ ne changent pas beaucoup alors que nous augmentons $R \geq 5$. Nous notons que les taux de variance négatifs pour *SynReg-R* sont semblables à ceux pour *Pseudo-EAS*. Évidemment, quand MR est élevé, les renseignements disponibles dans $\mathcal{D}_{\text{syn}}$ pour estimer b_{syn} sont équivalents aux renseignements disponibles dans $\mathcal{D}_{\text{EAS}}$.

Figure 3.3 Propriétés d'échantillonnage répété de *SynRep-1* et de *SynRep-R* pour β sous différents nombres d'échantillons synthétiques (M) et de répliques (R), selon un plan de sondage avec probabilité proportionnelle à la taille.

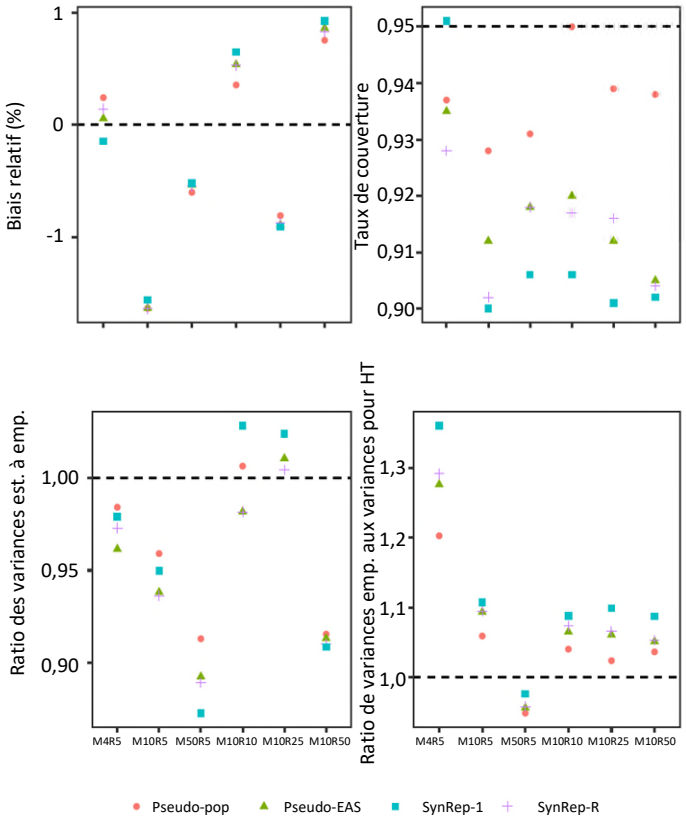
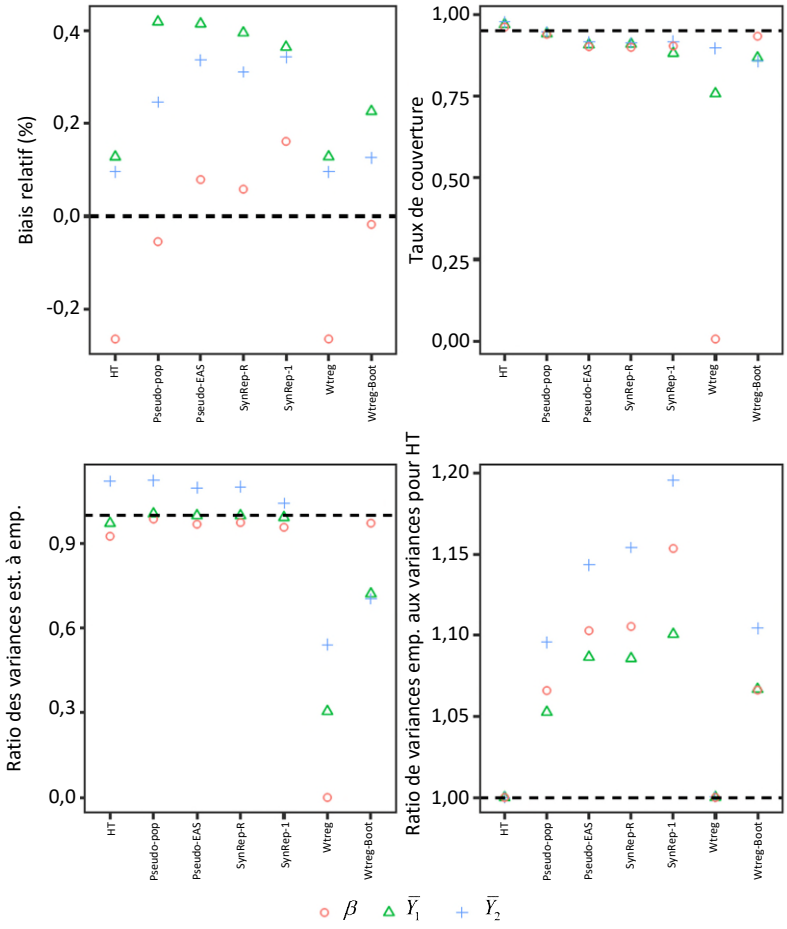


Tableau 3.1
Proportion des estimations de la variance négatives dans les études par simulations avec PPT. Quand $M = 50$, toutes les estimations de la variance sont positives.

(M, R)	Méthode	$\bar{Y}_1$	$\bar{Y}_2$	β
M4R5	*Pseudo-EAS*	0,09	0,15	0,13
M4R5	*SynRep-R*	0,11	0,17	0,13
M4R5	*SynRep-1*	0,17	0,26	0,22
M10R5	*Pseudo-EAS*	0,01	0,02	0,02
M10R5	*SynRep-R*	0,01	0,03	0,02
M10R5	*SynRep-1*	0,04	0,09	0,07

Nous nous penchons ensuite sur la comparaison de *SynRep-R* et de *SynRep-1* avec d'autres approches, en particulier *Wtreg*, *Wtreg-Boot* et *EASsyn*. Dans ce cas-ci, nous établissons que $M = 10$ et, lorsque cela est pertinent, que $R = 10$, puis nous tirons 500 échantillons répétés. La figure 3.4 résume les rendements de l'échantillonnage répété des méthodes qui tiennent compte des poids d'enquête. Pour toutes ces méthodes, les estimateurs ponctuels ont simulé des biais de pourcentage qui sont généralement négligeables. Pour *SynRep-R* et *SynRep-1*, les estimations de la variance moyennes sont proches de leurs variances empiriques correspondantes, et les taux de couverture sont proches de la valeur nominale. Pour *Wtreg* et *Wtreg-Boot*, les estimateurs de la variance peuvent sous-estimer grandement les variances empiriques correspondantes, plus particulièrement pour $\bar{Y}_1$ et $\bar{Y}_2$, ce qui donne lieu à des taux de couverture de l'intervalle de confiance qui peuvent être considérablement plus faibles que la valeur nominale de 95 %. L'étape bootstrap dans *Wtreg-Boot* donne lieu à des estimations de la variance plus fiables que dans *Wtreg*, mais *Wtreg-Boot* n'est pas aussi bien calé que *SynRep-R* et *SynRep-1*, qui ont des taux de couverture plus proches de la valeur nominale. Comme nous nous y attendions, *HT* donne lieu à des estimations exactes et des taux de couverture proches de la valeur nominale. Nous notons que la figure 3.4 n'affiche pas les résultats pour les méthodes *Direct* et *EASsyn*, car elles donnent de piètres résultats pour $\bar{Y}_1$ et $\bar{Y}_2$. Pour ces deux méthodes, les biais simulés pour $\bar{Y}_1$ et $\bar{Y}_2$ se situent à environ 16 % et 11 %, respectivement, et les taux de couverture se situent à environ 0 % et 30 %, respectivement. Les résultats accentuent l'importance de tenir compte des plans de sondage informatifs lors de la génération de données entièrement synthétiques qui peuvent être analysées en tant qu'échantillons aléatoires simples.

Figure 3.4 Propriétés d'échantillonnage répété de différentes quantités et procédures selon $M = 10$ échantillons synthétiques et $R = 10$ répliques, en fonction d'un plan de sondage avec probabilité proportionnelle à la taille.



Dans l'ensemble, les études par simulations laissent supposer que *SynRep-R* et *SynRep-1* peuvent fournir des inférences approximativement valides et que ces méthodes sont supérieures sur le plan inférentiel aux données entièrement synthétiques qui ignorent le plan de sondage complexe. L'annexe comprend les résultats des études par simulations où nous échantillonnons $\mathcal{D}$ à l'aide d'échantillons aléatoires simples. Ces résultats confirment que les règles de combinaison offrent un rendement raisonnable, même sans probabilités de sélection inégales.

4. Exemple au moyen des données de l'ACS

Nous illustrons *SynRep-R* et *SynRep-1* en supposant que $\mathcal{D}$ est un sous-ensemble de données tirées de l'échantillon de microdonnées à grande diffusion de l'ACS de 2021 pour $n =$ 84 128 personnes de l'État du Michigan. Les variables pour notre exemple comprennent le poids, l'âge et le revenu total de chaque

participant. Pour imiter les variables dans les simulations, nous créons, à partir de l'âge, un indicateur binaire Y_1 qui équivaut à 1 lorsqu'une personne a au moins 65 ans; nous appelons cet indicateur le statut de personne âgée. Aux fins de synthèse, nous transformons le revenu en prenant sa racine cubique. Les modèles de synthèse sont alors une distribution de Bernoulli pour Y_1 et une régression linéaire de la racine cubique du revenu total sur Y_1. Après avoir synthétisé les valeurs de la racine cubique du revenu, nous les élevons à la troisième puissance pour obtenir les revenus sur l'échelle initiale. Nous mettons en œuvre chaque méthode en suivant les procédures de la section 3. Pour *SynRep-R* et *SynRep-1*, nous établissons que $M = 10$ et que $R = 10$.

Pour les quantités de population, nous estimons la proportion de personnes dans la population ayant le statut de personne âgée $\overline{Y}_1$, la moyenne de population des valeurs de revenu, $\overline{Y}_2$, et le coefficient β de Y_1 dans le modèle de régression linéaire du revenu transformé à la racine cubique sur le statut de personne âgée.

La figure 4.1 présente les estimations ponctuelles et les intervalles de confiance à 95 % pour les trois quantités de population. Puisque les méthodes *Direct* et *EASsyn* ignorent le plan de sondage, elles donnent lieu à des résultats relativement imprécis, plus particulièrement pour $\overline{Y}_1$. En revanche, les estimations ponctuelles pour les méthodes de données synthétiques qui tiennent compte des poids d'enquête sont plus proches des estimations ponctuelles de *HT*. De plus, les intervalles de confiance à 95 % pour ces méthodes et les intervalles de confiance de la méthode de *HT* se chevauchent largement. Nous notons toutefois que *Wtreg* semble souffrir d'une sous-estimation de la variance, particulièrement pour β. Les intervalles de confiance pour les méthodes de pseudo-vraisemblance peuvent également être plus étroits que les intervalles de confiance pour *HT* ainsi que *SynRep-R* et *SynRep-1*.

Nous pouvons aussi examiner les risques de divulgation pour les méthodes de données synthétiques. Dans ce cas-ci, nous imitons un scénario d'attaque décrit par Kim et coll. (2021) dans lequel nous tenons compte d'un adversaire qui utilise les données synthétiques pour estimer la valeur de revenu la plus élevée dans $\mathcal{D}$. Plus précisément, nous examinons les différences entre le revenu synthétique maximum dans chaque ensemble de données synthétiques et le revenu maximum dans $\mathcal{D}$. Cette évaluation ne vise pas à illustrer un processus rigoureux et exhaustif d'évaluation des risques de divulgation. Nous utilisons plutôt ce scénario d'attaque pour comparer les différentes procédures de synthèse.

Le tableau 4.1 présente la répartition des différences pour les méthodes de synthèse qui tiennent compte du plan d'enquête. Dans l'ensemble, les résultats sont raisonnablement similaires d'une méthode à l'autre, ce qui laisse supposer qu'elles offrent des degrés de protection similaires dans ce scénario. Elles donnent toutes lieu à d'importantes différences entre les revenus synthétiques les plus élevés et les revenus observés les plus élevés. Les résultats laissent supposer qu'un adversaire qui choisit cette stratégie d'attaque n'est pas susceptible d'estimer le revenu le plus élevé avec exactitude.

Figure 4.1 Estimations ponctuelles et intervalles de confiance à 95 % pour $\bar{Y}_1$, $\bar{Y}_2$ et β dans l'exemple au moyens des données de l'ACS. Résultats basés sur $M = 10$ échantillons synthétiques et $R = 10$ répétitions.

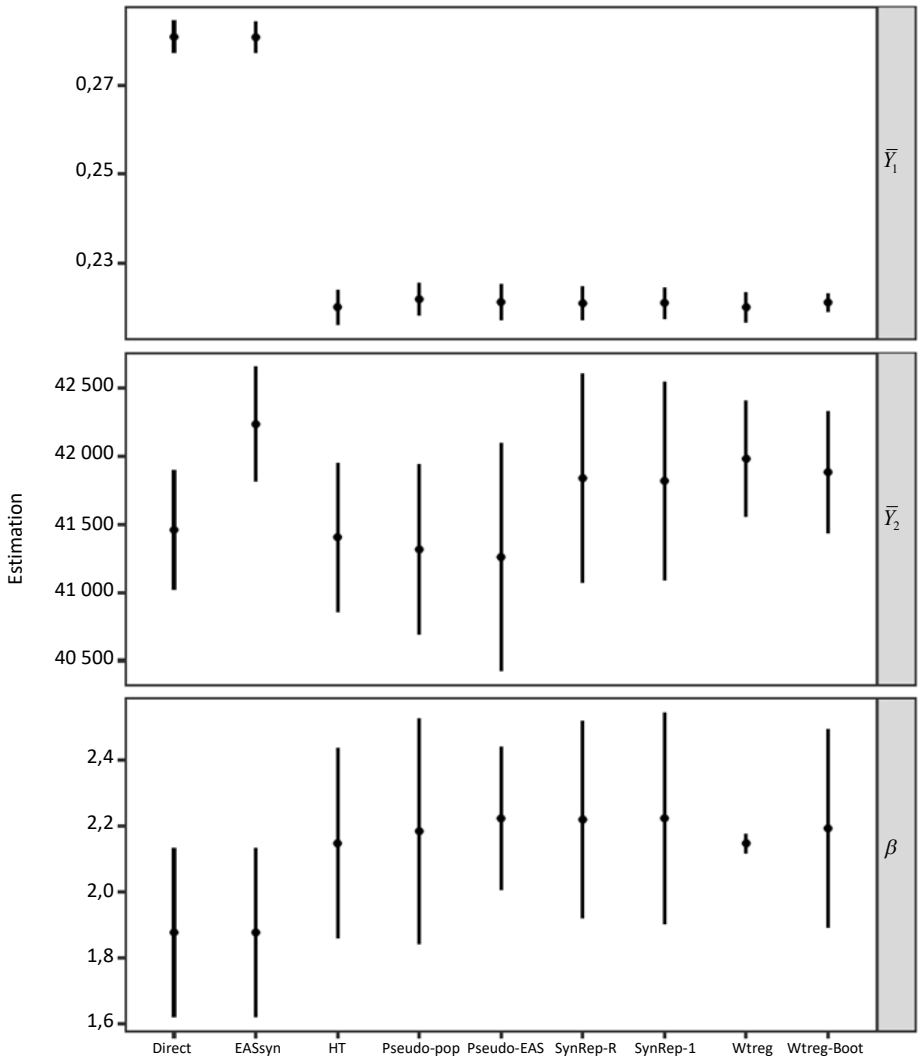


Tableau 4.1
Résumés des différences ($) en ce qui concerne la valeur de revenu la plus élevée dans les données synthétiques et les données de l'American Community Survey. La valeur réelle la plus élevée est 1 029 000 $.

Méthode	Min.	1er quartile	Médiane	Moyenne	3e quartile	Max.
SynRep-R	-424 323	-298 230	-252 984	-214 476	-139 874	465 380
SynRep-1	-371 466	-297 180	-287 199	-268 711	-267 428	-40 405
Wtreg	-440 253	-297 689	-242 095	-218 810	-159 766	707 411
Wtreg-Boot	-410 354	-275 398	-209 513	-174 444	-139 109	133 759

5. Discussion

SynRep-R et *SynRep-1* représentent une stratégie générale pour construire des données entièrement synthétiques qui tiennent compte des plans de sondage complexes : utiliser le BBPFP pour « déconstruire » le plan de sondage et remplacer ensuite les valeurs confidentielles par des valeurs simulées. La diffusion de multiples ensembles de données synthétiques, c'est-à-dire établir que $MR > 1$, peut accroître l'efficacité statistique et faciliter l'estimation de la variance. Cependant, les organismes peuvent aussi utiliser la méthode de BBPFP selon $MR = 1$. Bien que la diffusion d'un seul ensemble de données synthétiques ne puisse peut-être pas permettre d'obtenir une estimation de la variance approximativement valide pour des enquêtes complexes, elle peut tout de même être utile dans certaines situations, par exemple quand les données synthétiques servent à l'entraînement de codes ou à des analyses exploratoires où l'estimation de la variance n'est pas nécessaire.

Comme un évaluateur l'a noté, de nombreux organismes qui adoptent des approches de données synthétiques fournissent aussi des moyens permettant aux utilisateurs de vérifier la qualité de leurs inférences tirées de données synthétiques. Par exemple, les utilisateurs peuvent soumettre leur code à l'organisme qui a diffusé les données synthétiques qui peut par la suite exécuter le code et leur communiquer les résultats protégés contre la divulgation. Ce processus s'appelle la validation des résultats (Barrientos, Bolton, Balmat, Reiter, de Figueiredo, Machanavajjhala, Chen, Kneifel et DeLong, 2018). De plus, les utilisateurs peuvent soumettre des requêtes à un serveur qui analyse les données confidentielles et synthétiques, et fait ensuite un compte rendu des mesures de similarité des deux résultats d'analyse, par exemple le chevauchement dans les intervalles de confiance (Karr, Kohnen, Oganian, Reiter et Sanil, 2006). Ce processus s'appelle vérification des résultats (Barrientos et coll., 2018). En raison de la validation ou de la vérification, les utilisateurs de *SynRep-R* et de *SynRep-1* peuvent se buter à un obstacle supplémentaire. Si l'organisme exécute directement les codes d'analyse soumis par les utilisateurs, ces derniers seront peut-être obligés de préciser une version du code pondérée par les poids d'enquête aux fins de validation, même s'ils ont utilisé une analyse d'échantillon aléatoire simple pour les données synthétiques. Bien entendu, pour de nombreuses analyses, par exemple la modélisation par régression, certains utilisateurs se privent d'analyses pondérées, auquel cas la question est sans intérêt. L'organisme peut aussi automatiser la validation ou la vérification, auquel cas il pourrait être en mesure de transformer automatiquement les requêtes soumises par les utilisateurs en versions pondérées par les poids d'enquête en arrière-plan; il s'agit là d'un domaine pour de futures recherches.

Nous avons choisi de mettre au point des méthodes qui permettent aux organismes d'appliquer l'idée formulée dans Rubin (1993) : diffuser des données qui peuvent être analysées comme des échantillons aléatoires simples. Cela peut faciliter les analyses pour les utilisateurs, car ils n'ont pas à trouver une manière de traiter les poids dans le fichier, par exemple dans une estimation de la variance. La diffusion d'échantillons aléatoires simples pourrait aussi contribuer à atténuer les risques de divulgation qui pourraient

découler de la diffusion de poids d'enquête. Par exemple, si les poids diffusés dans les fichiers synthétiques sont échantillonnés directement à partir des valeurs de poids dans $\mathcal{D}$ sans altération, les poids pourraient révéler des renseignements à propos des personnes concernées, ce qui est considéré comme un risque de divulgation inacceptable (Fienberg, 2010). Finalement, la diffusion d'échantillons aléatoires simples évite de devoir estimer les relations entre les poids et les variables de résultat, une tâche pouvant être compliquée en pratique. Néanmoins, il serait intéressant de comparer les profils de risque et d'utilité de ces approches avec ceux des méthodes qui sont mises au point dans le présent article.

Il y a un grand nombre d'autres sujets liés à la stratégie générale sur lesquels il vaudrait la peine de s'attarder davantage. Premièrement, en pratique, les poids d'enquête peuvent être grandement variables et ne pas être fortement liés aux variables d'intérêt de l'enquête; cela pourrait faire en sorte que les estimations pondérées par les poids d'enquête aient des variances gonflées. Il est possible d'éviter cela en partie, par exemple en utilisant des approches basées sur des modèles pour lisser les poids (Beaumont, 2008; Xia et Elliott, 2016; Si, Trangucci, Gabry et Gelman, 2020). La génération de données synthétiques basée sur le BBPFP (ou sur toute autre approche) n'est pas immunisée contre ces problèmes de pondération. Il serait donc intéressant d'examiner si et comment le modèle de synthèse peut réduire les effets de l'inflation de la variance attribuable à des poids extrêmes.

Deuxièmement, nous nous concentrons sur la mise au point du cadre de données entièrement synthétiques et de ses règles de combinaison correspondantes, en utilisant des scénarios simples et des modèles de synthèse pour illustrer les méthodes. Conceptuellement, les organismes peuvent appliquer *SynRep-R* et *SynRep-1* à des données multivariées et pour divers paramètres d'intérêt, par exemple les moyennes de sous-domaine et les coefficients de régressions multiples. Dans de tels cas, il pourrait être avantageux d'utiliser des méthodes de modélisation souples, comme les modèles fondés sur des arbres ou d'autres algorithmes d'apprentissage automatique. De futurs travaux pourraient porter sur le rendement de ces synthétiseurs en combinaison avec les étapes de la génération de pseudo-populations et de pseudo-EAS.

Troisièmement, nous obtenons les règles de combinaison en supposant que les données d'enquête originales sont complètes. Les organismes pourraient imputer les données d'enquête manquantes et générer des répliques synthétiques simultanément, en tenant possiblement compte du plan d'échantillonnage complexe dans le modèle d'imputation et l'approche de synthèse. Cette stratégie pourrait nécessiter de nouvelles règles de combinaison semblables à celles de Reiter (2004).

Quatrièmement, nous présentons des ajustements ponctuels pour traiter les valeurs négatives des estimations de la variance. Nous pourrions être en mesure d'améliorer ces ajustements. Par exemple, nous pourrions être en mesure d'adapter la stratégie dans Si et Reiter (2011), qui élaborent des méthodes d'inférence pour les données entièrement synthétiques en fonction de l'échantillonnage provenant des distributions utilisées dans les calculs des règles de combinaison. De plus, comme un réviseur l'a souligné, il pourrait être avantageux d'utiliser l'interprétation de Raab et coll. (2018) pour les composantes

d'échantillonnage et de synthèse du calcul dans *SynRep-R*. Cela donne lieu à un autre estimateur de la variance, $(1+M^{-1})\,b_{\text{syn}} - (1+R^{-1})\,\overline{v}_{\text{syn}}$. De futurs travaux pourront permettre d'évaluer le rendement de ces méthodes d'inférence de rechange.

Cinquièmement, il serait intéressant de généraliser l'application de *SynRep-R* et de *Syn-Rep-1* à d'autres plans d'échantillonnage complexes, comme les plans d'échantillonnage en grappes stratifiés à plusieurs degrés qui sont de pratique courante. Zhou, Elliott et Raghunathan (2016) ont élargi le BBPFP pour tenir compte des strates, de la mise en grappes et des poids d'enquête dans la génération de populations synthétiques. Nous nous attendons à ce qu'il soit possible de tirer des échantillons aléatoires simples de ces pseudo-populations, de générer des répliques synthétiques, en utilisant possiblement des modèles de synthèse qui saisissent les renseignements du plan de sondage, comme on l'a suggéré dans Reiter (2002), et d'élargir les règles de combinaison présentées dans le présent article. Il s'agirait d'une extension naturelle pour évaluer de manière exhaustive les rendements de l'échantillonnage répété de *SynRep-R* et de *Syn-Rep-1* dans de tels échantillons complexes à plusieurs degrés.

Enfin, il serait utile de mettre au point des approches raisonnées pour mesurer les risques de divulgation pour ces méthodes. Pour *SynRep-R* et *SynRep-1*, il serait conceptuellement possible d'estimer la distribution *a posteriori* d'un adversaire pour les valeurs de données confidentielles compte tenu des valeurs synthétiques diffusées, par exemple comme cela est décrit pour les scénarios simples dans Reiter, Wang et Zhang (2014) et Hu, Reiter et Wang (2015). Cependant, il serait en pratique difficile de le faire sur le plan des calculs. Il faudrait tenir compte du processus de génération de données synthétiques au complet – y compris la méthode bootstrap, l'échantillonnage et la synthèse – lors du calcul de cette distribution *a posteriori*. En effet, pour autant que nous le sachions, les organismes qui diffusent des données synthétiques utilisent des méthodes ponctuelles pour évaluer les risques de divulgation, par exemple en comparant la similarité des valeurs aberrantes dans les données confidentielles et synthétiques, comme nous l'avons illustré dans le présent article (Kinney, Reiter et Miranda, 2014). La mise au point de méthodes d'évaluation des risques de divulgation est un domaine clé pour les recherches futures axées sur toutes les méthodes de génération de données entièrement synthétiques.

Remerciements

Les travaux ont été financés par une subvention de la U.S. National Science Foundation (SES 2217456) et un projet pilote du Michigan Center on the Demography of Aging financé par le National Institute on Aging (P30 AG012846).

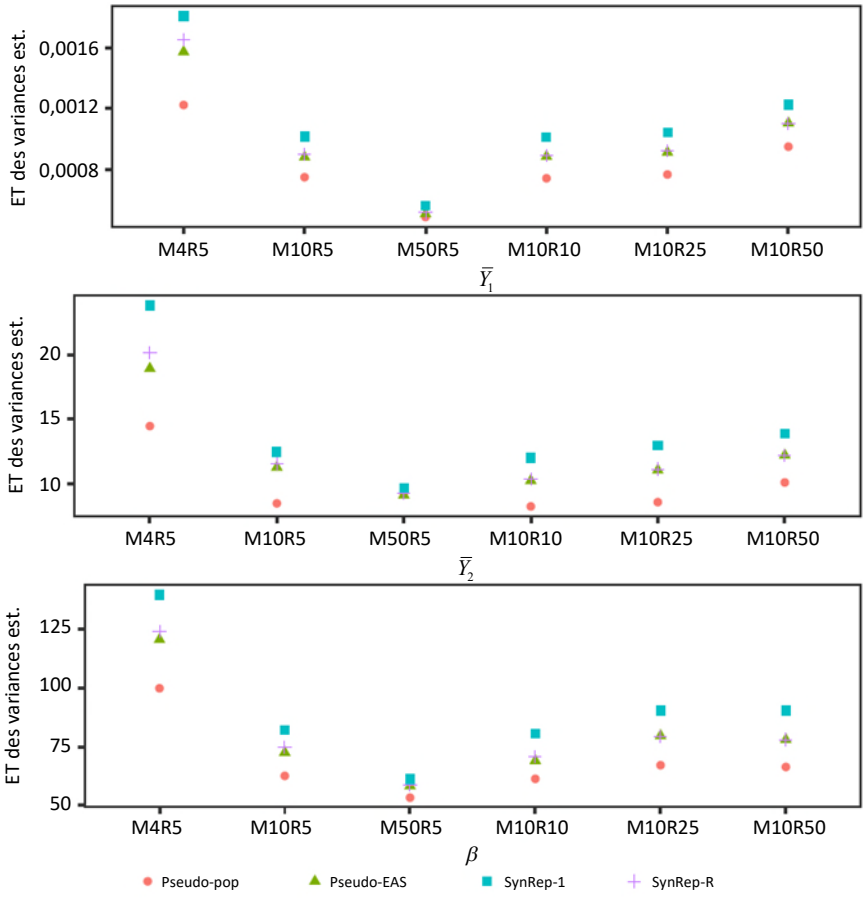
Annexe

A. Résultats des simulations supplémentaires

La figure A.1 affiche la variabilité des 1 000 valeurs de variances estimées des estimateurs ponctuels pour $\beta, \overline{Y}_1$ et $\overline{Y}_2$ pour la simulation selon le plan de sondage avec PPT. La variabilité a tendance à diminuer

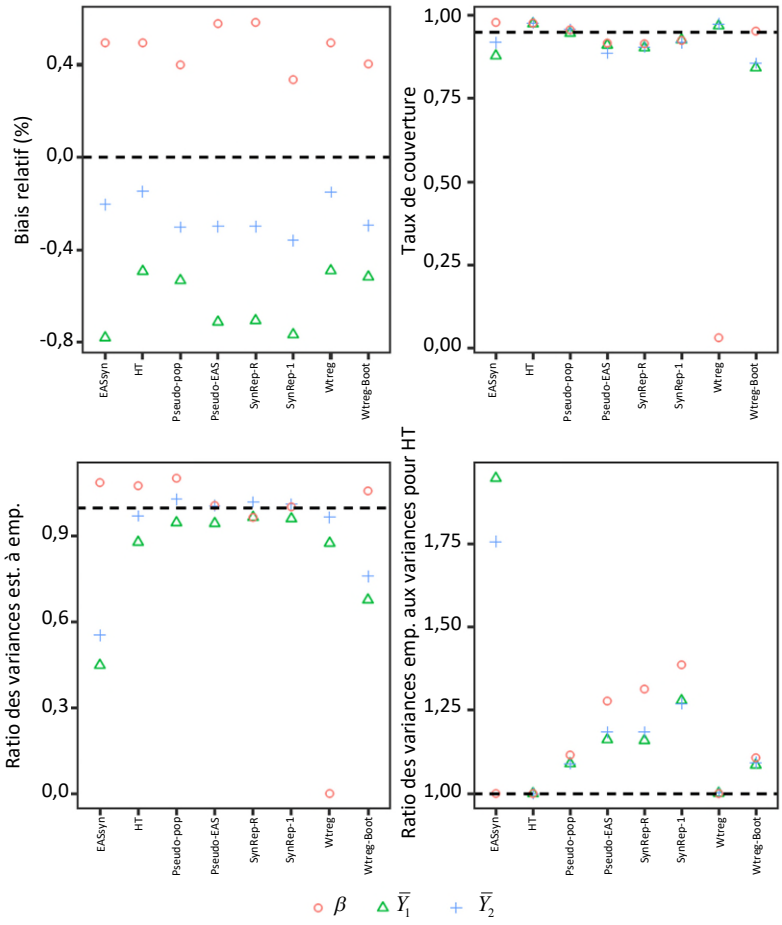
avec M. L'augmentation de R quand M demeure constant ne semble pas avoir beaucoup d'incidence sur la stabilité des résultats. Nous constatons une variabilité accrue à mesure que les procédures introduisent plus d'étapes faisant intervenir le caractère aléatoire; en d'autres mots, à mesure que nous passons de *Pseudo-pop* à *Pseudo-EAS,* à *SynRep-R* et à *SynRep-1*. La variabilité a tendance à être la plus importante pour *SynRep-1*.

Figure A.1 Écart-type (ET) des variances estimées (est.) de différentes quantités de population au moyen de différentes procédures pour différents nombres d'échantillons synthétiques (M) et de répliques (R) selon un plan de sondage avec probabilité proportionnelle à la taille.



Pour vérifier de nouveau la validité des règles de combinaison, nous répétons les simulations de la section 3 en utilisant un EAS au lieu d'un plan de sondage avec PPT. Plus précisément, nous utilisons la population décrite dans la section 3.1, mais nous utilisons un EAS de $n = 500$ enregistrements pour chaque $\mathcal{D}$. La figure A.2 présente les résultats. Dans l'ensemble, les rendements de *SynRep-R* et de *SynRep-1* sont le reflet des tendances observées pour le plan de sondage avec PPT dans la section 3.

Figure A.2 Propriétés d'échantillonnage répété de différentes quantités et procédures selon $M = 10$ échantillons synthétiques et $R = 10$ répliques, en fonction d'un plan d'échantillonnage aléatoire simple.



Bibliographie

Barrientos, A.F., Bolton, A., Balmat, T., Reiter, J.P., de Figueiredo, J.M., Machanavajjhala, A., Chen, Y., Kneifel, C. et DeLong, M. (2018). Providing access to confidential research data through synthesis and verification: An application to data on employees of the U.S. federal government. *Annals of Applied Statistics*, 12, 1124-1156.

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, 95, 3, 539-553.

Dong, Q., Elliott, M.R. et Raghunathan, T.E. (2014). Une méthode non paramétrique de production de populations synthétiques qui tient compte des caractéristiques des plans de sondage complexes. *Techniques d'enquête*, 40, 1, 33-52. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2014001/article/14003-fra.pdf.

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control*. New York: Springer.

Drechsler, J., et Reiter, J.P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105, 1347-1357.

Fienberg, S.E. (2010). The relevance or irrelevance of weights for confidentiality and statistical analyses. *Journal of Privacy and Confidentiality*, 1, 183-195.

Gambino, J.G. (2021). R package pps: PPS Sampling. https://cran.r-project.org/web/packages/pps/index.html.

Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663-685.

Hu, J., Reiter, J.P. et Wang, Q. (2015). Disclosure risk evaluation for fully synthetic data. Dans *Privacy in Statistical Databases*, (Éd., J. Domngo-Ferrer), 185-199. Heidelberg: Springer.

Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P. et Sanil, A.P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60, 224-232.

Kim, H.J., Drechsler, J. et Thompson, K.J. (2021). Synthetic microdata for establishment surveys under informative sampling. *Journal of Royal Statistical Society, Series A*, 184, 255-281.

Kinney, S.K., Reiter, J.P. et Miranda, J. (2014). Synlbd 2.0: Improving the Synthetic Longitudinal Business Database. *Statistical Journal of the International Association for Offcial Statistics*, 30, 129-135.

Kinney, S.K., Reiter, J.P., Reznek, A.P., Miranda, J., Jarmin, R.S. et Abowd, J.M. (2011). Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. *Revue Internationale de Statistique*, 79, 363-384.

Little, R.J. (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.

Meeden, G., Lazar, R. et Geyer, C.J. (2020). R package polyapost: Simulating from the Polya posterior. https://cran.r-project.org/web/packages/polyapost/index.html.

Mitra, R., et Reiter, J.P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. Dans *Privacy in Statistical Databases*, (Éds., J. Domingo-Ferrer et L. Franconi), 177-188. New York: Springer-Verlag.

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

Pfeffermann, D. (2011). Modélisation des données d'enquêtes complexes : Pourquoi les modéliser ? Pourquoi est-ce un problème ? Comment le résoudre ? *Techniques d'enquête*, 37, 2, 123-146. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2011002/article/11602-fra.pdf.

Raab, G.M., Nowok, B. et Dibben, C. (2018). Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, 7(3), 67-97.

Raghunathan, T.E. (2021). Synthetic data. *Annual Review of Statistics and Its Application*, 8, 129-140.

Raghunathan, T.E., Reiter, J.P. et Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Offcial Statistics*, 19, 1-16.

Reiter, J.P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Offcial Statistics*, 18, 531-544.

Reiter, J.P. (2003). Inférence pour les ensembles de microdonnées à grande diffusion partiellement synthétiques. *Techniques d'enquête*, 29, 2, 203-211. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2003002/article/6785-fra.pdf.

Reiter, J.P. (2004). Utilisation simultanée de l'imputation multiple pour les données manquantes et le contrôle de la divulgation. *Techniques d'enquête*, 30, 2, 263-271. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2004002/article/7755-fra.pdf.

Reiter, J.P. (2005a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*, 168, 185-205.

Reiter, J.P. (2005b). Significance tests for multi-component estimands from multiply imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131, 365-377.

Reiter, J.P. (2005c). Using CART to generate partially synthetic, public use microdata. *Journal of Offcial Statistics*, 21, 441-462.

Reiter, J.P. (2008). Selecting the number of imputed datasets when using multiple imputation for missing data and disclosure limitation. *Statistics and Probability Letters*, 78, 15-20.

Reiter, J.P. (2009). Using multiple imputation to integrate and disseminate confidential microdata. *Revue Internationale de Statistique*, 77, 179-195.

Reiter, J.P., et Drechsler, J. (2010). Releasing multiply-imputed, synthetic data generated in two stages to protect confidentiality. *Statistica Sinica*, 20, 405-422.

Reiter, J.P., et Kinney, S.K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Offcial Statistics*, 28, 583-590.

Reiter, J.P., et Raghunathan, T.E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102, 1462-1471.

Reiter, J.P., Raghunathan, T.E. et Kinney, S.K. (2006). L'importance de la modélisation du plan d'échantillonnage dans l'imputation multiple pour les données manquantes. *Techniques d'enquête*, 32, 2, 161-168. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006002/article/9548-fra.pdf.

Reiter, J.P., Wang, Q. et Zhang, B. (2014). Bayesian estimation of disclosure risks in multiply imputed, synthetic data. *Journal of Privacy and Confidentiality 6*, Article 2.

Rubin, D.B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9, 130-134.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Offcial Statistics*, 9, 462-468.

Savitsky, T.D., et Toth, D. (2016). Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10, 1677-1708.

Si, Y., et Reiter, J.P. (2011). A comparison of posterior simulation and inference by combining rules for multiple imputation. *Journal of Statistical Theory and Practice*, 5, 335-347.

Si, Y., Trangucci, R., Gabry, J.S. et Gelman, A. (2020). Ajustement de pondération hiérarchique bayésienne et inférence d'enquête. *Techniques d'enquête*, 46, 2, 193-228. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2020002/article/00003-fra.pdf.

Stan Development Team (2024). Stan: A C++ library for probability and sampling. http://mc-stan.org.

United Nations Economic Commission for Europe (2022). Synthetic Data for National Statistical Organizations. https://statswiki.unece.org/display/SDS/Synthetic+Data+Sets+public?preview=%2F282330193%2F330369384%2FHLG-MOS+Synthetic+Data+Guide.docx. Consulté : 2022-01-12.

United States Bureau of the Census (2021). Accessing American Community Survey PUMS data. https://www.census.gov/programs-surveys/acs/microdata/access.html.

Williams, M.R., et Savitsky, T.D. (2021). Uncertainty estimation for pseudo-Bayesian inference under complex sampling. *Revue Internationale de Statistique*, 89, 72-107.

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

Zhou, H., Elliott, M.R. et Raghunathan, T.E. (2016). Synthetic multiple-imputation procedure for multistage complex samples. *Journal of Offcial Statistics*, 32, 231-256.

Techniques d'enquête, décembre 2024
Vol. 50, nº 2, pp. 407-441
Statistique Canada, nº 12-001-X au catalogue

Modèles d'erreur de couplage pour l'estimation par capture-recapture sans vérifications manuelles

Abel Dasylva, Arthur Goussanou et Christian-Olivier Nambeu[1]

Résumé

Il est possible d'appliquer la méthode de capture-recapture pour mesurer la couverture des sources de données administratives et de mégadonnées dans les statistiques officielles. Dans sa forme de base, elle comporte le couplage de deux sources tout en supposant un couplage parfait et d'autres hypothèses types. En pratique, des erreurs de couplage surviennent et constituent une source potentielle de biais quand le couplage est fondé sur des quasi-identificateurs. Ces erreurs comprennent des faux positifs et des faux négatifs, où les premiers se produisent quand un lien est établi entre des enregistrements provenant de différentes unités, et les deuxièmes surviennent lorsqu'il n'y a pas de lien entre des enregistrements provenant de la même unité. Jusqu'à présent, les solutions trouvées ont reposé sur des vérifications manuelles coûteuses ou ont posé l'hypothèse restrictive de l'indépendance conditionnelle. Dans le présent article, on assouplit ces exigences en modélisant plutôt le nombre de liens à partir d'un enregistrement. Cette méthode peut aussi être adoptée pour estimer l'exactitude du couplage sans vérifications manuelles, quand on lie deux sources ayant chacune de la sous-couverture.

Mots-clés : Appariement de données; couplage d'enregistrements; estimation de système dual; intégration de données; mégadonnées; qualité.

1. Introduction

La méthode de capture-recapture est un outil important pour estimer la couverture des sources de données administratives et des mégadonnées qui sont de plus en plus employées dans les statistiques officielles (Zhang, 2015). Dans sa forme la plus simple, elle permet d'estimer la couverture de deux sources sur la même population finie, en déterminant les unités sélectionnées dans les deux sources, c'est-à-dire leur intersection, selon des hypothèses types qui comprennent un couplage parfait. Alors, la couverture estimée est fondée sur l'estimateur très connu de Petersen (1896) et Lincoln (1930). Il peut toutefois y avoir des erreurs de couplage parce que le couplage est souvent fondé sur des quasi-identificateurs comme les noms et les dates. Ces erreurs peuvent biaiser l'estimation de la couverture, qui doit être corrigée.

Pour ce qui est de l'exactitude, une erreur de couplage est définie comme un *faux négatif* ou un *faux positif*, où un faux négatif correspond à l'absence de lien entre des enregistrements d'une même unité, et un faux positif correspond à la présence d'un lien entre des enregistrements de différentes unités. En relation avec ces concepts, une paire d'enregistrements est dite *appariée* si ses enregistrements proviennent de la même unité (Fellegi et Sunter, 1969; Herzog, Scheuren et Winkler, 2007). Sinon, elle est dite *non appariée*. L'exactitude du couplage peut être mesurée au moyen de la vérification manuelle, d'un modèle statistique ou d'une combinaison des deux méthodes. Les vérifications manuelles consistent à inspecter visuellement un échantillon probabiliste de paires d'enregistrements afin de déterminer si elles sont appariées (Dasylva, Abeysundera, Akpoué, Haddou et Saïdi, 2016). Elles sont très souples et s'appliquent quelles que soient les détails du couplage. Elles sont cependant très coûteuses. La solution de rechange aux vérifications manuelles

1. Abel Dasylva, Arthur Goussanou et Christian-Olivier Nambeu, Statistique Canada, 150, promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6. Courriels : abel.dasylva@statcan.gc.ca, arthur.goussanou@statcan.gc.ca, christianolivier.nambeu@statcan.gc.ca.

consiste à ajuster un modèle statistique comme plusieurs études l'ont proposé, notamment par des mélanges log-linéaires (Fellegi et Sunter, 1969; Thibaudeau, 1993; Winkler, 1993; Daggy, Xu, Hui, Gamache et Grannis, 2013; Chipperfield, Hansen et Rossiter, 2018; Winglee, Valliant et Scheuren, 2005; Chipperfield et Chambers, 2015; Haque, Mengersen et Stern, 2021; Haque et Mengersen, 2022), des modèles du poids de couplage probabiliste d'une paire (Belin et Rubin, 1995; Sariyar, Borg et Pommerening, 2011), des modèles bayésiens (Fortini, Liseo, Nuccitelli et Scanu, 2001; Tancredi et Liseo, 2011; Sadinle, 2017; Steorts, Hall et Fienberg, 2016), et des modèles fondés sur le nombre de liens à partir d'un enregistrement donné (Blakely et Salmond, 2002; Dasylva et Goussanou, 2022). La dernière méthode de modélisation mentionnée présente un intérêt particulier dans le présent travail, car elle ne se limite pas aux couplages probabilistes et tient implicitement compte de toutes les interactions entre les variables de couplage. La méthode de modélisation n'est pas aussi coûteuse que les vérifications manuelles, mais elle est moins souple puisqu'elle repose sur des hypothèses concernant la procédure de couplage. Elle peut aussi constituer un défi quand le couplage est soumis à la contrainte d'avoir au plus un lien par enregistrement ou exactement un lien par enregistrement (Lahiri et Larsen, 2005, page 226). Chipperfield et Chambers (2015) et Sadinle (2017) ont traité cette question. Cependant, les méthodes proposées nécessitent d'importantes ressources informatiques et dépendent de l'hypothèse restrictive selon laquelle les variables de couplage sont conditionnellement indépendantes, c'est-à-dire qu'elles sont indépendantes, que la paire soit appariée ou non. En effet, cette hypothèse est une source potentielle de biais, selon Newcombe (1988, chapitre E.6, page 149), Belin et Rubin (1995) et Blakely et Salmond (2002). À la suite de Larsen et Rubin (2001), il est également possible de combiner les examens manuels et la modélisation statistique pour tirer parti de la souplesse des premiers et des faibles coûts de la seconde. Il reste que les coûts globaux dépassent le budget de nombreuses études. Pour ce qui est de la méthode par capture-recapture, les erreurs de couplage sont nuisibles parce qu'elles peuvent biaiser la couverture estimée. En effet, un faux négatif peut entraîner la sous-estimation de la couverture, tandis qu'un faux positif peut produire un biais dans la direction opposée. Bien entendu, il faut éliminer ce biais pour estimer la couverture avec exactitude.

De nombreuses méthodes de correction des erreurs ont été décrites, qui reposent sur les hypothèses types de capture-recapture sauf pour le couplage imparfait, c'est-à-dire une population fermée, des unités indépendantes sélectionnées indépendamment par chaque source, une probabilité de capture homogène par au moins une source et aucune unité en double ou hors du champ de l'enquête dans l'une des sources. Elles comprennent des solutions qui nécessitent des estimations manuelles de l'exactitude du couplage (Ding et Fienberg, 1994; Di Consiglio et Tuoto, 2015; de Wolf, van der Laan et Zult, 2019; Brown, Bycroft, Di Cecco, Elleouet, Powell, Račinskij, Smith, Tam, Tuoto et Zhang, 2020) et d'autres solutions qui reposent sur un modèle statistique selon l'hypothèse d'indépendance conditionnelle (Tancredi et Liseo, 2011; Račinskij, Smith et van der Heijden, 2019). Ding et Fienberg (1994), Di Consiglio et Tuoto (2015) et de Wolf et coll. (2019) présentent trois solutions étroitement apparentées du premier type, où ils restreignent le couplage à au plus un lien par enregistrement et ils supposent que la probabilité de faux positifs est négligeable pour les unités capturées par les deux sources. Cependant, ils estiment l'exactitude du couplage au moyen de vérifications manuelles, qui sont coûteuses, mais qui constituent la seule solution pratique, compte tenu des contraintes du couplage. Brown et coll. (2020) exposent une solution différente, qui repose

aussi sur des estimations manuelles de l'exactitude du couplage, et dans laquelle les deux sources doivent être liées deux fois au moyen de procédures de couplage différentes, en partant de l'hypothèse que les indicateurs de couplage associés sont indépendants dans chaque paire appariée. Tancredi et Liseo (2011) et Račinskij et coll. (2019) utilisent des modèles statistiques pour estimer conjointement l'exactitude du couplage et la couverture sans vérifications manuelles. Ils posent toutefois l'hypothèse restrictive d'indépendance conditionnelle.

Le présent article vise à estimer conjointement la couverture et l'exactitude du couplage sans vérification manuelle, tout en relâchant l'hypothèse selon laquelle les variables de couplage sont conditionnellement indépendantes. À cette fin, l'article présente une nouvelle méthodologie, qui constitue une extension d'un modèle antérieur d'erreur de couplage (Dasylva et Goussanou, 2022), selon les hypothèses types de capture-recapture, sauf pour l'hypothèse de couplage parfait. Dans cette méthode fondée sur un modèle, on estime la couverture en couplant les enregistrements avec un rappel suffisamment élevé, ou en précisant les interactions dans les paires appariées, tout en permettant des interactions arbitraires dans les paires non appariées. Les mêmes modèles peuvent servir à estimer le rappel et la précision lors du couplage de deux sources ayant chacune de la sous-couverture.

Les autres sections de l'article porteront, dans cet ordre, sur les notations et les hypothèses, le contexte, la méthodologie proposée, les simulations et la conclusion.

2. Notations et hypothèses

Dans la version de base du problème de capture-recapture, il faut estimer la couverture d'une liste tirée d'une population finie en exploitant une deuxième liste tirée de la même population finie, selon des hypothèses types, qui comprennent une population fermée, des unités indépendantes sélectionnées indépendamment par chaque liste, une capture homogène par au moins une liste, aucune unité en double ou hors du champ de l'enquête dans l'une des listes et un couplage parfait des deux listes. Généralement, ces listes correspondent à des échantillons probabilistes ayant des probabilités de sélection inconnues. Dans ce qui suit, on suppose que les hypothèses types de capture-recapture se vérifient, sauf pour l'hypothèse de couplage parfait qui est assouplie. Chaque liste est modélisée par un échantillon de Bernoulli, où chaque unité incluse est associée à un enregistrement susceptible de contenir des erreurs typographiques. Par exemple, pour ce qui est des listes de personnes, cet enregistrement peut contenir le nom de famille et la date de naissance. On suppose que les valeurs des enregistrements sont indépendantes entre unités, mais aucune hypothèse n'est faite au sujet de la dépendance de ces valeurs pour les enregistrements qui proviennent de la même unité, ou de la dépendance des variables dans ces enregistrements. Pour satisfaire l'hypothèse de capture homogène, on suppose qu'une unité est capturée dans la première liste indépendamment des valeurs d'enregistrement associées (par exemple le nom de famille et la date de naissance enregistrés). La capture dans la deuxième liste peut toutefois dépendre de ces valeurs. Dans ce qui suit, on désigne la cardinalité d'un ensemble s par $|s|$, et pour un uplet $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ on définit $|\mathbf{x}| = |x_1| + \ldots + |x_d|$.

2.1 Population finie et sources de données

Les unités proviennent d'une population finie U ayant N unités étiquetées de 1 à N. Les unités sont sélectionnées selon deux échantillons de Bernoulli qui sont désignés par S_A et S_B et définis par deux sous-ensembles de $\{1, \ldots, N\}$, où l'on suppose que la probabilité d'inclusion $P(i \in S_A)$ ne dépend pas de N. Par exemple, S_A peut être le recensement de la population et S_B peut être une enquête sur la couverture. Dans chaque liste où elle est incluse, l'unité i est associée à un enregistrement dont la valeur est sa caractéristique déterminante. Dans ce qui suit, le terme « enregistrement » signifiera aussi cette valeur quand le contexte permettra de clairement le comprendre. La valeur de l'enregistrement est censée se trouver dans l'espace d'enregistrement $\mathcal{V}_N$, qui est fini mais qui peut être de grande taille. Par exemple, l'espace d'enregistrement peut comprendre toutes les chaînes écrites avec au plus 32 caractères alphabétiques, si le couplage est établi avec le nom de famille. Dans S_B, cet enregistrement est désigné par V_i. Pour S_A, l'étiquetage des enregistrements dépend d'une permutation aléatoire uniforme $\Pi(.)$ de $\{1, \ldots, N\}$, pour modéliser l'absence totale d'informations sur les enregistrements associés à la même unité. Dans cette liste, l'unité i est associée à l'enregistrement $V'_{\Pi(i)}$. L'utilisation d'une permutation aléatoire inconnue est un procédé courant dans la littérature sur le couplage d'enregistrements (Lahiri et Larsen, 2005; Chambers, 2009). Il est commode sur le plan mathématique de visualiser les deux listes sous forme d'échantillons tirés de registres conceptuels A et B, où chaque unité est associée à un enregistrement. Ensuite, on suppose que les processus de création des enregistrements et d'inclusion dans les listes sont tels que les observations

$$\left[\left(I(i \in S_A), I(i \in S_B), V_i, V'_{\Pi(i)}\right)\right]_{1 \le i \le N}$$

sont indépendantes et identiquement distribuées et indépendantes de la permutation aléatoire $\Pi(.)$. Cela signifie que les deux listes sont étiquetées indépendamment et que l'on peut considérer seulement le cas où la permutation $\Pi(.)$ est l'identité (c'est-à-dire conditionner au fait que $\Pi(.)$ soit l'identité dans ce qui suit), sans perte de généralité. Ainsi, l'unité i est associée à V_i dans S_B et à V'_i dans S_A. Dans l'exposé qui suit, tous les arguments sont à la condition que $\Pi(.)$ soit égale à l'identité. On omet toutefois cette information pour simplifier la notation. Afin de satisfaire l'hypothèse de capture homogène, on suppose que la capture dans S_A est indépendante de V_i et V'_i. Toutefois, la capture dans S_B peut dépendre de ces enregistrements, c'est-à-dire que nous pourrions avoir

$$P\left(i \in S_B \mid V_i, V'_i\right) \neq P(i \in S_B).$$

Par exemple, la probabilité de capture peut varier selon des postrates qui sont basées sur V_i.

2.2 Couplage d'enregistrements et erreurs connexes

L'indicateur d'un lien entre V_i et V'_j est désigné par L_{ij} et appelé *décision de couplage* pour la paire (i, j). Soit n_i le nombre de liens issus de V_i dans S_B, c'est-à-dire

$$n_i = \sum_{j \in S_A} L_{ij}. \tag{2.1}$$

On suppose que le couplage est tel que L_{ij} est seulement une fonction de V_i et V'_j, c'est-à-dire que la décision de lier deux enregistrements ne dépend pas d'autres enregistrements. Dans la configuration actuelle, cette hypothèse signifie précisément que

$$E\left[L_{ij} \middle| \begin{matrix} \left[\left(I(k \in S_A), I(k \in S_B), V_k, V'_k \right) \right]_{k \in \{1, \ldots, N\} \setminus \{i, j\}}, \\ (i, j) \in S_A \times S_B, V_i, V'_j \end{matrix} \right] = E\left[L_{ij} \middle| (i, j) \in S_A \times S_B, V_i, V'_j \right].$$

Cette condition concerne une vaste gamme de stratégies de couplage pratiques qu'il est possible de mettre en œuvre au moyen de méthodes probabilistes, déterministes ou hiérarchiques. Toutefois, elle exclut les couplages qui limitent le nombre de liens par enregistrement (par exemple exactement un seul ou au plus un) même si de tels liens peuvent être établis à partir de couplages plus simples, qui répondent à la condition. Une paire d'enregistrements est désignée par un élément (i, j) de $S_A \times S_B$. Comme cela a été mentionné plus haut, une paire est dite *appariée* si les deux enregistrements proviennent de la même unité. Sinon, elle est dite *non appariée*. Dans l'exposé sur les erreurs de couplage, un *faux négatif* est une paire appariée qui n'est pas liée, un *faux positif* est une paire non appariée qui est liée, et un *vrai positif* est une paire appariée qui est liée. Par souci d'exhaustivité, nous définissions un *vrai négatif* comme une paire non appariée qui n'est pas liée. Par souci de commodité, supposons que FN, FP, VP et VN désignent respectivement le nombre total de faux négatifs, de faux positifs, de vrais positifs et de vrais négatifs. Il est courant de représenter les différents types de paires d'enregistrements dans un tableau 2×2 appelé matrice de confusion où les cellules hors de la diagonale représentent les erreurs, comme le montre le tableau 2.1. L'exactitude du couplage est habituellement mesurée par le rappel et la précision, le *rappel* étant la proportion de paires appariées qui sont liées (c'est-à-dire $\mathrm{VP}/(\mathrm{VP} + \mathrm{FN})$) et la *précision* étant la proportion de paires liées qui sont appariées (c'est-à-dire $\mathrm{VP}/(\mathrm{VP} + \mathrm{FP})$). Elle est également mesurée par le *taux de faux négatifs*, qui est la proportion de paires appariées qui ne sont pas liées (c'est-à-dire $\mathrm{FN}/(\mathrm{VP} + \mathrm{FN})$), et le *taux de faux positifs* (TFP), qui est la proportion de paires non appariées qui sont liées (c'est-à-dire $\mathrm{FP}/(\mathrm{VN} + \mathrm{FP})$). En cas de couplage parfait, la précision et le rappel sont égaux à 1,0 et le TFP est nul. Dans cette situation idéale, la taille de la population est estimée selon Petersen (1896) et Lincoln (1930) par

$$\hat{N} = \frac{|S_A||S_B|}{|S_A \cap S_B|}.$$

Par conséquent, la couverture estimée de S_A est donnée par $|S_A \cap S_B| / |S_B|$, tandis que celle de S_B est donnée par $|S_A \cap S_B| / |S_A|$. En cas d'erreurs de couplage, l'intersection des deux listes n'est pas observée directement. Il faut plutôt inférer la taille de cette intersection à partir des liens observés et de l'exactitude du couplage qui peut être estimée par la modélisation du nombre de liens à partir d'un enregistrement donné.

Tableau 2.1
Matrice de confusion.

	Lien	Non liée
Appariée	VP	FN
Non appariée	FP	VN

3. Contexte

La présente section fournit quelques éléments de contexte sur le modèle d'erreur (Dasylva et Goussanou, 2022), qui doit être adapté au problème qui nous concerne. Ce modèle s'applique quand S_A est un recensement (c'est-à-dire $S_A = U$) et que le couplage est tel que la décision de lier deux enregistrements donnés ne dépend d'aucun autre enregistrement. Dans ce cas, on peut estimer l'exactitude du couplage en modélisant la distribution du nombre de liens à partir d'un enregistrement donné, sans faire de suppositions sur la dépendance des variables de couplage.

3.1 Relation entre les erreurs et le nombre de liens d'un enregistrement

En général, il existe un lien étroit entre le nombre de liens d'un enregistrement donné et les erreurs de couplage qui s'y rapportent. Cette relation est décrite dans le tableau 3.1 quand S_A est un recensement (Dasylva et Goussanou, 2020). Quand $n_i = 0,$ on sait qu'il n'y a pas de faux positif mais un faux négatif parce que S_A est un recensement. Quand $n_i = 1, \ldots, N-1,$ il n'y a pas de faux négatif ou il y en a un et donc n_i ou $n_i - 1$ faux positifs, selon que l'enregistrement est lié à l'enregistrement de recensement apparié ou non, parce que S_A est un recensement et qu'il n'y a pas d'enregistrement en double. Quand $n_i = N,$ on sait qu'il n'y a pas de faux négatifs et qu'il y a $N-1$ faux positifs, pour les mêmes raisons. À titre d'illustration, examinons l'exemple présenté dans le tableau 3.2, où $N = 5,$ $S_A = U = \{1, 2, 3, 4, 5\},$ $S_B = \{2, 3\}$ et la nature de chaque paire d'enregistrements est également indiquée. Dans cet exemple, il y a quatre liens, à savoir $(2,1),$ $(2, 2),$ $(3, 2)$ et $(3, 4).$ Selon le tableau 3.2, il est possible de vérifier qu'il y a une relation entre n_i et les erreurs de couplage. Quand $n_i = 1, \ldots, N-1,$ les erreurs ne sont pas entièrement connues et peuvent être prédites au moyen d'un modèle.

Tableau 3.1
Relation entre n_i et les erreurs quand S_A est un recensement.

n_i	Faux négatifs	Faux positifs
0	1	0
$1 \le n_i \le N-1$	0 ou 1	$n_i - 1$ ou n_i
N	0	$N-1$

Tableau 3.2
Exemple, où $N = 5,$ $S_A = U = \{1, 2, 3, 4, 5\}$ et $S_B = \{2, 3\}$ et les liens sont indiqués par les coches.

	$j =$							
	1	2	3	4	5	n_i	N^{bre} FN	N^{bre} FP
$i = 2$	√ FP	√ VP	VN	VN	VN	2	0	1
3	VN	√ FP	FN	√ FP	VN	2	1	2

3.2 Modèle pour enregistrements homogènes

Blakely et Salmond (2002) modélisent n_i par la somme d'une variable de Bernoulli (pour les vrais positifs) avec une variable binomiale indépendante (pour les faux positifs) et ils estiment les paramètres qui s'y rapportent au moyen d'une équation quadratique. Cependant, la distribution de n_i doit être identique pour tous les enregistrements. Sinon, l'estimateur pourrait être biaisé ou ne pas exister (Dasylva et Goussanou, 2022) si l'équation quadratique n'a pas de solution. En pratique, cette question peut se poser quand on apparie avec des noms ou d'autres caractéristiques, qui se présentent à des fréquences différentes dans la population.

3.3 Modèle pour enregistrements hétérogènes

Pour régler le problème, Dasylva et Goussanou (2022) ont étendu le modèle (Blakely et Salmond, 2002) à un mélange fini, qui s'applique quand N devient grand dans des conditions de régularité. Pour décrire ces conditions, supposons que

$$\mathcal{V}_N^* = \left\{ v \in \mathcal{V}_N \text{ s.t. } P\left(V_i = v \middle| \, i \in S_B\right) > 0 \right\}. \tag{3.1}$$

En d'autres termes, $\mathcal{V}_N^*$ est le sous-ensemble des valeurs d'enregistrement qui peuvent être observées dans S_B avec une probabilité positive. À ce stade, il est utile d'examiner le sous-ensemble de toutes les valeurs d'enregistrement (issues de $\mathcal{V}_N$) qui sont liées avec une probabilité positive à une valeur d'enregistrement particulière, ainsi qu'un sur ensemble de cet ensemble, appelé « *voisinage* » et désigné par $\mathcal{B}_N(v)$ pour la valeur $v \in \mathcal{V}_N^*$. Ainsi,

$$\mathcal{B}_N(v) \supset \left\{ v' \in \mathcal{V}_N \text{ s.t. } E\left[L_{ij} \middle| \, i \in S_B, \left(V_i, V_j'\right) = (v, v') \right] > 0 \right\}. \tag{3.2}$$

De façon informelle, le voisinage d'une valeur d'enregistrement particulière est un sous-ensemble de valeurs d'enregistrement qui ressemblent à cette valeur selon un certain critère. Par exemple, on envisage de lier les enregistrements en fonction du nom de famille en lettres majuscules et on suppose que deux enregistrements sont liés s'ils correspondent exactement pour cette variable. Dans ce cas, la valeur d'enregistrement (v) « JARO » peut être associée au voisinage singleton $(\mathcal{B}_N(v) = \{v\})$ {« JARO »}. Pour affiner cet exemple, supposons maintenant que deux enregistrements sont liés si les noms de famille sont identiques, ou s'ils ont la même longueur et diffèrent seulement par une lettre. Dans ce cas, la valeur « JARO » peut être associée au voisinage

$$\begin{aligned} &\{\text{“AARO”, “BARO”, …, “ZARO”}\} \cup \{\text{“JARO”, “JBRO”, …, “JZRO”}\} \cup \\ &\quad \{\text{“JAAO”, “JABO”, …, “JAZO”}\} \cup \{\text{“JARA”, “JARB”, …, “JARZ”}\}. \end{aligned}$$

Le concept de voisinage est utile quand on caractérise le pouvoir discriminatif des variables de couplage et quand on définit des conditions de régularité pour l'estimation convergente du rappel et de la précision sans vérifications manuelles. Pour décrire ces conditions, définissons les fonctions $p_N(.)$, $\lambda_N(.)$ et $\lambda_N^{(0)}(.)$ qui

donnent la probabilité de vrais positifs, la probabilité de faux positifs et la probabilité qu'un enregistrement non apparié se trouve dans le voisinage, c'est-à-dire

$$p_N(v) = E\Big[L_{ii} \,\Big|\, i \in S_B, V_i = v \Big], \tag{3.3}$$

$$\lambda_N(v) = E\Big[L_{ij} \,\Big|\, i \in S_B, V_i = v \Big], \; j \neq i, \tag{3.4}$$

$$\lambda_N^{(0)}(v) = P\Big(V_j' \in \mathcal{B}_N\left(V_i\right) \Big|\, i \in S_B, V_i = v \Big). \tag{3.5}$$

Alors, les deux premières conditions de régularité sont données par les équations suivantes, où Λ est positif et fini, F est une distribution bivariée avec support contenu dans $[0,1] \times [0, \Lambda]$ et ni l'un ni l'autre ne dépend de N.

$$\sup_{v \in \mathcal{V}_N^*} (N-1)\, \lambda_N^{(0)}(v) \;\leq\; \Lambda, \tag{3.6}$$

$$\big(p_N(V_i), (N-1)\, \lambda_N(V_i) \big) \,\big|\, \{ i \in S_B \} \;\xrightarrow{d}\; F. \tag{3.7}$$

La première condition implique que le nombre attendu de faux positifs possède une borne supérieure pour chaque enregistrement. Quand la probabilité de faux positifs a comme borne inférieure δ (Dasylva et Goussanou, 2020, équation 6), cela signifie aussi que la précision n'est pas inférieure à $\delta / (\delta + \Lambda)$ en tout et pour chaque poststrate, qui est définie en fonction de V_i (Dasylva et Goussanou, 2020). La deuxième condition signifie que la distribution conjointe du nombre espéré de vrais positifs et du nombre espéré de faux positifs est approximativement donnée par F quand N est de grande taille. Quand F est discret et comporte G atomes, ces deux conditions supposent la convergence en distribution suivante (Dasylva et Goussanou, 2022, Lemme 1).

$$n_i \,\big|\, \{ i \in S_B \} \;\xrightarrow{d}\; \sum_{g=1}^{G} \alpha_g \text{ Bernoulli}(p_g) \,*\, \text{Poisson}(\lambda_g), \tag{3.8}$$

où $*$ désigne l'opérateur de convolution. Cela signifie que, à la limite, un enregistrement appartient à l'une des G classes latentes, où α_g est la probabilité de la classe g, et p_g et λ_g sont les nombres espérés de vrais positifs et de faux positifs pour les enregistrements de cette classe. On peut estimer les paramètres du modèle en maximisant la probabilité composite des n_i. Ceux-ci sont liés à l'exactitude du couplage au moyen des nombres espérés de vrais positifs et de faux positifs par enregistrement dans S_B, qui sont donnés par $\bar{p} = \sum_{g=1}^{G} \alpha_g p_g$ et $\bar{\lambda} = \sum_{g=1}^{G} \alpha_g \lambda_g$. En effet, le rappel et la précision convergent en probabilité respectivement vers $\bar{p}$ et $\bar{p} / (\bar{p} + \bar{\lambda})$, sous les deux conditions de régularité supplémentaires suivantes (Dasylva et Goussanou, 2022, Corollaire 1), où $i \neq i'$ et c est une constante finie positive ne dépendant pas de N.

$$NP\Big(\mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) \neq \varnothing \,\Big|\, \{i, i'\} \subset S_B \Big) \leq c, \tag{3.9}$$

$$NP\Big(V_i' \in \mathcal{B}_N(V_i) \,\Big|\, \{i, i'\} \subset S_B, \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) \neq \varnothing \Big) \leq c. \tag{3.10}$$

Ces deux conditions signifient que les enregistrements de différentes unités sont très susceptibles d'avoir des voisinages disjoints (3.9) et que les enregistrements appariés ne sont pas très éloignés (3.10). Étant donné les autres conditions de régularité (c'est-à-dire (3.6)-(3.7)), elles impliquent également la convergence des estimateurs composites par le maximum de vraisemblance (Dasylva et Goussanou, 2022, théorème 3).

Dans l'ensemble, cette méthodologie présente plusieurs avantages par rapport à d'autres solutions fondées sur un modèle, parce qu'elle en prend en compte à la fois les interactions entre les variables de couplage et l'hétérogénéité des enregistrements, de façon transparente. En outre, on peut tester l'adéquation du modèle au moyen de la procédure décrite par Dasylva et Goussanou (2024) pour tenir compte de la corrélation des n_i. Quand S_A est un recensement, la méthodologie peut aussi servir à estimer les faux négatifs générés par la procédure de création des pochettes (Dasylva et Goussanou, 2021). Cependant, il faut l'adapter quand S_A présente de la sous-couverture.

4. Méthodologie

La méthodologie proposée est fondée sur deux extensions du modèle décrit dans la section précédente, qui sera appelé par la suite *modèle de voisinage univarié* ou simplement modèle de voisinage. La première extension tient compte de la sous-couverture de S_A et modifie seulement l'interprétation de certains paramètres. Elle sert à estimer la couverture en cas de couplage lorsque le rappel est suffisamment élevé. La deuxième extension est plus importante, car elle remplace n_i par un vecteur de telles variables, chacune représentant le nombre de liens pour une règle de couplage distincte. Le modèle qui en résulte est appelé *modèle de voisinage multivarié*. On l'utilise pour estimer la couverture en précisant les interactions dans les paires appariées, tout en permettant des interactions arbitraires dans les paires non appariées. Les paragraphes suivants traitent de la stratégie de couplage avant de présenter en détail les différentes extensions et leur emploi aux fins d'estimation de la couverture.

4.1 Couplage des sources

Dans leurs solutions, Ding et Fienberg (1994), Di Consiglio et Tuoto (2015) et de Wolf et coll. (2019) limitent chaque enregistrement à un maximum d'un seul lien. Or, cette contrainte complique considérablement la modélisation des erreurs de couplage, comme cela a été indiqué précédemment. Nous proposons plutôt ici de coupler les enregistrements sans cette contrainte, en fonction d'une règle telle que la décision de coupler deux enregistrements ne dépend d'aucun autre enregistrement. Ainsi, un enregistrement peut avoir zéro lien, un lien ou plusieurs liens. Cette règle de couplage peut être au moyen des méthodes de couplage d'enregistrements déterministe, hiérarchique ou probabiliste.

4.2 Extension du modèle de voisinage univarié

Quand S_A n'est pas un recensement, la relation entre n_i et les erreurs est conforme au tableau 4.1, qui diffère du tableau 3.1 quand $n_i = 0$ et $n_i = |S_A|$. Quand $n_i = 0$, il n'y a aucune certitude quant à l'occurrence d'un faux négatif, car on ne sait pas si l'unité correspondante est dans S_A, contrairement à ce

qui se passe dans le tableau 3.1. Quand $n_i = |S_A|$, le nombre de faux positifs n'est pas connu avec certitude pour la même raison. Pour tenir compte de la sous-couverture de S_A, redéfinissons $\mathcal{B}_N(v)$ en tant que sous-ensemble de $\mathcal{V}_N$ de sorte que

$$\mathcal{B}_N(v) \supset \left\{v' \in \mathcal{V}_N \text{ s.t. } E\left[L_{ij} \,\middle|\, (i,j) \in S_B \times S_A, \left(V_i, V_j'\right) = (v, v')\right] > 0\right\}. \tag{4.1}$$

Tableau 4.1
Relation entre n_i et les erreurs quand S_A n'est pas un recensement.

n_i	Faux négatifs	Faux positifs
0	0 ou 1	0
$1 \le n_i \le \lvert S_A \rvert - 1$	0 ou 1	$n_i - 1$ ou n_i
$n_i = \lvert S_A \rvert$	0	$\lvert S_A \rvert - 1$ ou $\lvert S_A \rvert$

Redéfinissons aussi $p_N(.)$, $\lambda_N(.)$ et $\lambda_N^{(0)}(.)$ comme étant

$$p_N(v) = E\left[I\left(i \in S_A\right) L_{ii} \,\middle|\, i \in S_B, V_i = v\right], \tag{4.2}$$

$$\lambda_N(v) = E\left[I\left(j \in S_A\right) L_{ij} \,\middle|\, i \in S_B, V_i = v\right], \; j \neq i, \tag{4.3}$$

$$\lambda_N^{(0)}(v) = P\left(j \in S_A, V_j' \in \mathcal{B}_N\left(V_i\right) \,\middle|\, i \in S_B, V_i = v\right), \tag{4.4}$$

où $p_N(v)$ est la probabilité conjointe d'inclure i dans S_A et d'avoir un vrai positif, $\lambda_N(v)$ reste la probabilité de faux positifs et $\lambda_N^{(0)}(v)$ est la probabilité d'avoir un enregistrement non apparié dans le voisinage. Avec cette mise-à-jour de ces définitions, il est facile de démontrer que (3.8) s'applique quand $N \to \infty$ dans les conditions de régularité données par (3.6)-(3.7) et que F est discrète et comporte G atomes. En effet, la démonstration du cas de recensement continue de s'appliquer avec n_i selon (2.1). Voir Dasylva et Goussanou (2022, Lemme 1) pour en savoir plus. Les paramètres α_g et λ_g ont la même interprétation, mais p_g correspond maintenant au produit de la probabilité d'inclusion $P(i \in S_A)$ par la probabilité d'un vrai positif pour un enregistrement dans la classe g. Comme auparavant, soit $\overline{p} = \sum_{g=1}^{G} \alpha_g p_g$ et $\overline{\lambda} = \sum_{g=1}^{G} \alpha_g \lambda_g$, où $\overline{p}$ est le nombre attendu de vrais positifs par enregistrement, qui est aussi égal à $E\left[I(i \in S_A)\, L_{ii} \,\middle|\, i \in S_B\right]$, et où $\overline{\lambda}$ est le nombre attendu de faux positifs par enregistrement. Ainsi, $\overline{p}$ est une borne inférieure utile sur $P(i \in S_A)$. On peut estimer les paramètres du modèle en maximisant la vraisemblance composite des n_i quand G est donné, et en sélectionnant ce dernier paramètre par la minimisation du critère d'information d'Akaike comme dans le cas du recensement (Dasylva et Goussanou, 2022). Supposons que $\hat{\overline{p}}$ et $\hat{\overline{\lambda}}$ désignent les estimateurs par le maximum de vraisemblance qui en résultent. À l'annexe A, il est indiqué que le rappel et la précision (deux paramètres de population finie) convergent respectivement en probabilité vers $P(i \in S_A)^{-1} \overline{p} = E\left[L_{ii} \,\middle|\, i \in S_A \cap S_B\right]$ et $\overline{p} / (\overline{p} + \overline{\lambda})$, dans des conditions de

régularité. Dans les mêmes conditions, $\hat{\bar{p}}$ et $\hat{\bar{\lambda}}$ sont aussi des estimateurs convergents de $\bar{p}$ et $\bar{\lambda}$, de sorte que $P(i \in S_A)^{-1}\hat{\bar{p}}$ et $\hat{\bar{p}}/(\hat{\bar{p}}+\hat{\bar{\lambda}})$ sont respectivement des estimateurs convergents du rappel et de la précision.

4.3 Estimations de la couverture lorsque le rappel est élevé

À partir de l'exposé ci-dessus, on peut obtenir un estimateur convergent de la couverture $P(i \in S_A)$ comme suit :

$$\left(\frac{\text{VP}}{\text{VP + FN}}\right)^{-1}\hat{\bar{p}},$$

si le rappel (c'est-à-dire $\text{VP}/(\text{VP + FN})$) est connu. En particulier, $\hat{\bar{p}}$ est convergent, si l'on sait que le rappel est parfait; autrement dit, $\text{VP}/(\text{VP + FN}) = 1{,}0$, ce qui équivaut à ne pas avoir de faux négatifs. Cependant, il convient de mentionner que le modèle proposé est intéressant uniquement quand le couplage n'est pas parfait, c'est-à-dire si le rappel, la précision ou les deux sont inférieurs à 1,0. Autrement, l'estimateur standard par capture-recapture s'appliquerait, y compris dans la situation idéale où la clé de couplage est un identificateur unique sans erreur, c'est-à-dire une clé de couplage parfaite. En outre, le modèle de voisinage n'est pas conseillé avec une telle clé, car il se peut que certaines hypothèses du modèle ne se vérifient pas.

Pour appliquer la méthode ci-dessus, il faudrait concevoir la règle de couplage de sorte qu'elle génère très peu de faux négatifs, voire aucun, et idéalement en présentant un taux de faux négatifs inférieur à $\min\left(P(i \in S_A), 1-P(i \in S_A)\right)$ d'un ordre de grandeur. Cela peut s'inspirer des procédures de création des pochettes, qui sont utilisées dans les couplages probabilistes pour sélectionner un petit sous-ensemble du produit cartésien avec la majorité des paires appariées. Christen (2012) donne un bon aperçu de ces procédures, qui sont indispensables quand les sources sont de grande taille. Cependant, l'obtention d'un rappel suffisamment élevé peut se faire au détriment de la précision et nécessiter de tolérer une précision très faible, ce qui peut empêcher l'estimation de la couverture avec l'exactitude requise. Dans de tels cas, il est proposé d'estimer la couverture en précisant les interactions dans les paires appariées. Cela nécessite toutefois une extension multivariée du modèle de voisinage.

4.4 Modèle de voisinage multivarié

L'extension multivariée concerne les ensembles finis de règles de couplage simples qui sont également mutuellement exclusives, c'est-à-dire que pour chaque règle, la décision de lier deux enregistrements ne dépend d'aucun autre enregistrement, et chaque paire est liée par au plus une règle. La meilleure façon d'expliquer la nécessité de cette extension est de donner un exemple. Par souci de simplicité, supposons que S_A est connu pour être un recensement complet et que les deux sources doivent être liées au prénom, au nom de famille et à la date de naissance. Pour ce faire, il faut évaluer sept règles, identifiées dans le tableau 4.2, où γ se trouve dans l'ensemble fini $\Gamma = \{0,1\}^3 \setminus \{(0,0,0)\}$, et les composantes de γ indiquent

s'il y a une concordance exacte pour ce qui est du nom de famille, du prénom et de la date de naissance, dans cet ordre. Pour la règle γ, on désigne par $n_i^{(\gamma)}$ le nombre correspondant de liens pour l'enregistrement de l'échantillon i, par exemple $n_i^{(0,0,1)}$ est le nombre de liens quand le couplage est fondé sur le fait d'avoir les mêmes noms, mais une date de naissance différente. Une façon simple d'évaluer les différentes règles consiste à ajuster un modèle de la forme

$$n_i^{(\gamma)} \Big| \{i \in S_B\} \sim \sum_{g=1}^{G^{(\gamma)}} \alpha_g^{(\gamma)} \text{Bernoulli}\left(p_g^{(\gamma)}\right) * \text{Poisson}\left(\lambda_g^{(\gamma)}\right), \tag{4.5}$$

séparément pour chaque γ, où les nombres attendus de vrais positifs et de faux positifs par enregistrement sont respectivement $\overline{p}^{(\gamma)} = \sum_{g=1}^{G^{(\gamma)}} \alpha_g^{(\gamma)} p_g^{(\gamma)}$ et $\overline{\lambda}^{(\gamma)} = \sum_{g=1}^{G^{(\gamma)}} \alpha_g^{(\gamma)} \lambda_g^{(\gamma)}$. Notons que nous avons nécessairement la contrainte $\sum_{\gamma \in \Gamma} \overline{p}^{(\gamma)} \leq 1$, parce que les règles sont mutuellement exclusives. Cependant, il se peut que les estimateurs du rappel et de la précision qui en résultent soient inefficaces parce que des informations importantes ne sont pas pris en compte, comme la contrainte $\sum_{\gamma \in \Gamma} \overline{p}^{(\gamma)} \leq 1$ ou la corrélation entre les nombres de liens des différentes règles pour le même enregistrement de l'échantillon, qui n'est pas exploitée non plus. De plus, quand on choisit le nombre de classes $G^{(\gamma)}$ en fonction du critère d'information d'Akaike, le résultat $\hat{G}^{(\gamma)}$ peut varier selon les différentes règles, ce qui est contre-intuitif. Ajoutons que, même dans le meilleur des cas où $\hat{G}^{(\gamma)}$ est identique pour toutes les règles, les classes peuvent correspondre à différentes partitions latentes des enregistrements de l'échantillon selon les différentes règles, ce qui est également contre-intuitif et non souhaitable. De plus, les limites ci-dessus s'appliquent dans la situation plus générale où S_A n'est pas un recensement et où sa couverture est inconnue.

Tableau 4.2
Règles mutuellement exclusives basées sur le prénom, le nom de famille et la date de naissance.

Indice de la règle $\gamma = (\gamma_1, \gamma_2, \gamma_3)$	Nom de famille identique	Prénom identique	Date de naissance identique
(0,0,1)	x	x	√
(0,1,0)	x	√	x
(0,1,1)	x	√	√
(1,0,0)	√	x	x
(1,0,1)	√	x	√
(1,1,0)	√	√	x
(1,1,1)	√	√	√

La solution consiste à modéliser la distribution conjointe du vecteur de fréquences $\left[n_i^{(\gamma)}\right]_{\gamma \in \Gamma}$ au moyen d'une extension multivariée du modèle de voisinage, comme suit (des précisions sont données à l'annexe B). Pour décrire cette extension, il est commode de définir les distributions multivariées suivantes. La première distribution est la distribution conjointe de variables mutuellement indépendantes qui sont indexées sur un ensemble fini Γ, où la variable γ suit la distribution $\text{Poisson}\left(\lambda^{(\gamma)}\right)$. La distribution conjointe est alors simplement la distribution du produit. Pour faciliter la notation, nous définissons $\boldsymbol{\lambda} = \left[\lambda^{(\gamma)}\right]_{\gamma \in \Gamma}$ et désignons cette distribution par $\text{PPoisson}(\boldsymbol{\lambda})$, où le premier « P » signifie produit. La deuxième distribution correspond à la distribution conjointe du nombre de cellules dans une expérience multinomiale avec n essais, où la

dernière cellule est exclue, les autres cellules sont indexées sur Γ, et la probabilité d'observer la cellule γ est indiquée par $p^{(\gamma)}$, de sorte que $\sum_{\gamma\in\Gamma} p^{(\gamma)} \le 1$. Dans ce cas, nous définissons $\mathbf{p} = \left[p^{(\gamma)} \right]_{\gamma\in\Gamma}$ et nous désignons la distribution conjointe par $\text{IMultinomial}(n, \mathbf{p})$, où le « I » signifie « incomplète ». De façon générale, on peut envisager l'extension multivariée pour modéliser la distribution conjointe du nombre de liens, qui sont le résultat de l'application de règles de couplage simples mutuellement exclusives, indexées sur un ensemble fini Γ, où $n_i^{(\gamma)}$ désigne le nombre de liens de la règle γ pour l'enregistrement de l'échantillon i et $\mathbf{n}_i = \left[n_i^{(\gamma)} \right]_{\gamma\in\Gamma}$. Le modèle multivarié est un mélange fini de distri-butions discrètes à $|\Gamma|$ dimensions, où chaque composante est la convolution d'une distribution multi-nomiale incomplète avec un produit de distributions de Poisson indépendantes, à savoir

$$\mathbf{n}_i \big| \{i \in S_B\} \sim \sum_{g=1}^{G} \alpha_g \, \text{IMultinomial}(1, \mathbf{p}_g) * \text{PPoisson}\left(\boldsymbol{\lambda}_g\right), \tag{4.6}$$

où G est le nombre de classes d'enregistrements, α_g est la probabilité qu'un enregistrement d'échantillon provienne de la classe g, et $\mathbf{p}_g = \left[p_g^{(\gamma)} \right]_{\gamma\in\Gamma}$ et $\boldsymbol{\lambda}_g = \left[\lambda_g^{(\gamma)} \right]_{\gamma\in\Gamma}$ sont les vecteurs des nombres espérés de vrais positifs et de faux positifs pour un enregistrement dans la classe. De plus, $p_g^{(\gamma)}$ est le nombre espéré de vrais positifs et $\lambda_g^{(\gamma)}$ est le nombre espéré de faux positifs, selon la règle γ. Ensuite, étant donné G, le modèle est paramétré par $\left[\left(\alpha_g, \mathbf{p}_g, \boldsymbol{\lambda}_g \right) \right]_{1\le g\le G}$. Quand les enregistrements sont homogènes,

$$\mathbf{n}_i \,\big|\, \{i \in S_B\} \sim \text{IMultinomial}(1, \mathbf{p}) * \text{PPoisson}\left(\boldsymbol{\lambda}\right), \tag{4.7}$$

où $\mathbf{p} = \left[p^{(\gamma)} \right]_{\gamma\in\Gamma}$ et $\boldsymbol{\lambda} = \left[\lambda^{(\gamma)} \right]_{\gamma\in\Gamma}$. De plus, si $\min_{\gamma\in\Gamma} \lambda^{(\gamma)} > 0$ et $\mathbf{t} = \left[t^{(\gamma)} \right]_{\gamma\in\Gamma}$, nous avons

$$\begin{aligned} P\left(\mathbf{n}_i = \mathbf{t} \,\middle|\, i \in S_B\right) \;&= I\left(|\mathbf{t}| = 0\right)\left(1 - |\mathbf{p}|\right) e^{-|\boldsymbol{\lambda}|} \\ &\quad + I\left(|\mathbf{t}| > 1\right) \Bigg(\left(1 - |\mathbf{p}|\right) \prod_{\gamma\in\Gamma} \frac{e^{-\lambda^{(\gamma)}} \left(\lambda^{(\gamma)}\right)^{t^{(\gamma)}}}{t^{(\gamma)}!} \\ &\qquad + \sum_{\gamma\in\Gamma;\, t^{(\gamma)} > 0} p^{(\gamma)} \, \frac{e^{-\lambda^{(\gamma)}} \left(\lambda^{(\gamma)}\right)^{t^{(\gamma)}-1}}{\left(t^{(\gamma)} - 1\right)!} \prod_{\gamma'\in\Gamma\setminus\{\gamma\}} \frac{e^{-\lambda^{(\gamma')}} \left(\lambda^{(\gamma')}\right)^{t^{(\gamma')}}}{t^{(\gamma')}!} \Bigg). \end{aligned} \tag{4.8}$$

Comme auparavant, le modèle est motivé par la convergence de la distribution du vecteur des fréquences $\mathbf{n}_i = \left[n_i^{(\gamma)} \right]_{\gamma\in\Gamma}$, quand $N \to \infty$, comme l'indique le lemme 2 de l'annexe B. À partir du mélange multivarié, il s'ensuit que la distribution marginale de $n_i^{(\gamma)}$ est toujours donnée par (4.5), sauf que $G^{(\gamma)}$ et $\alpha_g^{(\gamma)}$ sont identiques pour toutes les règles, comme cela était souhaité. De plus, les catégories d'enregistrements correspondent maintenant pour toutes les règles. On obtient un modèle restreint quand $\mathbf{p}_g = \varrho\left(\boldsymbol{\beta}_g\right)$ pour chaque classe, où $\varrho(.)$ est une fonction injective connue et $\boldsymbol{\beta}_g$ est un vecteur de coefficients de régression de dimension inférieure à $|\Gamma|$. Cela signifie que les $|\Gamma|$ probabilités de vrais positifs pour les différentes règles ne sont pas libres, mais liées par moins de coefficients de régression, pour chaque classe d'enregistrements. Une telle restriction est utile quand on exploite l'information sur les interactions des variables dans les paires appariées, au moyen d'une spécification log-linéaire. Ensuite, un mélange multivarié ayant G composantes est paramétré par $\left[\left(\alpha_g, \boldsymbol{\beta}_g, \boldsymbol{\lambda}_g \right) \right]_{1\le g\le G}$. Selon le théorème 2 de l'annexe B, les paramètres du

modèle (pour chaque paramétrisation proposée) sont estimés de manière convergente en maximisant la probabilité composite des vecteurs de fréquences observés $\left[\mathbf{n}_i\right]_{i\in S_B}$, quand G est connu ou inconnu. Dans ce dernier cas, on peut sélectionner G selon le critère d'information minimum d'Akaike comme auparavant.

4.5 Estimer la couverture au moyen de la structure de corrélation

Quand les enregistrements sont liés avec un rappel parfait, la couverture peut être estimée au moyen du modèle de mélange univarié évoqué plus haut. Sinon, on peut estimer la couverture au moyen du modèle de voisinage multivarié, où les probabilités de vrais positifs sont contraintes selon la structure de corrélation des variables de couplage au moyen d'une spécification log-linéaire. Plus précisément, cela signifie que le modèle multivarié proposé est fondé sur un ensemble de règles de couplage simples (c'est-à-dire que chaque règle est telle que la décision de coupler deux enregistrements ne dépend d'aucun autre enregistrement), qui sont elles-mêmes fondées sur un premier ensemble de règles de couplage simples divisées en K groupes. Dans le groupe k, il y a H_k règles mutuellement exclusives (c'est-à-dire qu'une paire est liée par au plus une règle à partir du groupe) et $L_{ij}^{(k,h)}$ indique si la paire (i, j) est liée par la règle h. Par exemple, chaque groupe peut être fondé sur une seule variable, et les règles peuvent correspondre à différents niveaux de concordance pour cette variable. À titre d'exemple, pour le nom de famille, ces niveaux peuvent comprendre la concordance exacte, la concordance basée sur une erreur typographique (qui exclut la concordance exacte) et la concordance du code SOUNDEX (qui exclut les concordances exactes et celles basée sur une erreur typographique). Toutefois, un groupe peut comporter plusieurs variables. On obtient un deuxième ensemble de règles en combinant les règles du premier ensemble comme suit. Soit l'ensemble des indices $\Gamma = \{0,\ldots,H_1\}\times\ldots\times\{0,\ldots,H_K\} - \mathbf{0}_K$, et pour $\gamma = (\gamma_1,\ldots,\gamma_K)\in\Gamma$, désignons par $L_{ij}^{(\gamma)}$ l'indicateur que la règle γ lie la paire (i, j), où $L_{ij}^{(\gamma)} = 1$ seulement si $L_{ij}^{(k,\gamma_k)} = 1$ pour chaque k de sorte que $\gamma_k \geq 1$, et $\sum_{h=1}^{H_k} L_{ij}^{(k,h)} = 0$ pour chaque k de sorte que $\gamma_k = 0$. Le modèle proposé est le cas particulier du modèle de voisinage multivarié (4.6), où $p_g^{(\gamma)}$ a la forme suivante, pour un vecteur de covariables $\mathbf{z}^{(\gamma)}$ et des coefficients de régression $\mathbf{u}_g$.

$$p_g^{(\gamma)} = \frac{P(i\in S_A)\exp\left(\mathbf{z}^{(\gamma)\top}\mathbf{u}_g\right)}{1+\sum_{\gamma'\in\Gamma}\exp\left(\mathbf{z}^{(\gamma')\top}\mathbf{u}_g\right)}. \tag{4.9}$$

Pour un nombre donné de classes G, les paramètres du modèle incluent $\left[\left(\alpha_g,\mathbf{u}_g,\boldsymbol{\lambda}_\mathbf{g}\right)\right]_{1\leq g\leq G}$ et $P(i\in S_A)$. La forme particulière de $\mathbf{z}^{(\gamma)}$ et $\mathbf{u}_g$ dépend du modèle. L'exemple suivant en est une illustration, où le modèle comprend tous les termes principaux et aucune interaction. Cela signifie également que les composantes de γ sont indépendantes dans les paires appariées. Dans ce cas, le coefficient correspondant à l'événement $\{\gamma_k = l_k\}$ est indiqué par $u_{g,k(l_k)}$. Selon la convention de codage « dummy », le coefficient est fixé à zéro si $l_k = 0$. La covariable correspondant à ce coefficient est l'indicateur $I(\gamma_k = l_k)$ de sorte que

$$\mathbf{z}^{(\gamma)} = \left[I(\gamma_1=1)\ldots I(\gamma_1=H_1)\ldots I(\gamma_K=1)\ldots I(\gamma_K=H_K)\right]^\top, \tag{4.10}$$

$$\mathbf{u}_g = \left[u_{g,1(1)}\ldots u_{g,1(H_1)}\ldots u_{g,K(1)}\ldots u_{g,K(H_K)}\right]^\top. \tag{4.11}$$

Dans l'exemple suivant, le modèle comprend tous les termes principaux et les interactions du deuxième ordre, mais aucune interaction d'ordre supérieur. Pour $1 \le k_1 < k_2 \le K,$ le coefficient de l'interaction entre les événements $\{\gamma_{k_1} = l_{k_1}\}$ et $\{\gamma_{k_2} = l_{k_2}\}$ est désigné par $u_{g,k_1k_2(l_{k_1}l_{k_2})}$. Selon la même convention de codage, le coefficient est fixé à zéro si $l_{k_1} = 0$ ou $l_{k_2} = 0$. La covariable associée au coefficient est l'indicateur $I\left(\left(\gamma_{k_1}, \gamma_{k_2}\right) = \left(l_{k_1}, l_{k_2}\right)\right)$. Dans ce cas, $\mathbf{z}^{(\gamma)}$ et $\mathbf{u}_g$ ont des expressions plus complexes. Écrivons-les dans une formule claire en définissant

$$\mathbf{z}_{1k}^{(\gamma)} = \left[I\left(\gamma_k = 1\right) \ldots I\left(\gamma_k = H_k\right) \right]^\top.$$

De plus, désignons le deuxième membre de (4.10) par $\mathbf{z}_1^{(\gamma)},$ le deuxième membre de (4.11) par $\mathbf{u}_{g1}^{(\gamma)}$ et, pour $k = 1, \ldots, K-1,$ définissons

$$\begin{aligned}
\mathbf{z}_{2k}^{(\gamma)} &= \left(\left[\mathbf{z}_{1(k+1)}^{(\gamma)\top} \ldots \mathbf{z}_{1K}^{(\gamma)\top} \right] \otimes \mathbf{z}_{1k}^{(\gamma)\top} \right)^\top, \\
\mathbf{u}_{g2k} &= \Bigg[\overbrace{u_{g,k(k+1)(11)} \ldots u_{g,k(k+1)(H_k 1)}}^{\substack{\text{Termes d'interaction entre} \\ \text{le niveau 1 de } \gamma_{k+1} \text{ et} \\ \text{tous les niveaux de } \gamma_k}} \ldots \overbrace{u_{g,k(k+1)(1H_{k+1})} \ldots u_{g,k(k+1)(H_k H_{k+1})}}^{\substack{\text{Termes d'interaction entre} \\ \text{le niveau } H_{k+1} \text{ de } \gamma_{k+1} \text{ et} \\ \text{tous les niveaux de } \gamma_k}} \ldots \\
&\qquad \overbrace{u_{g,kK(11)} \ldots u_{g,kK(H_k 1)}}^{\substack{\text{Termes d'interaction entre} \\ \text{le niveau 1 de } \gamma_K \text{ et} \\ \text{tous les niveaux de } \gamma_k}} \ldots \overbrace{u_{g,kK(1H_K)} \ldots u_{g,kK(H_k H_K)}}^{\substack{\text{Termes d'interaction entre} \\ \text{le niveau } H_K \text{ de } \gamma_K \text{ et} \\ \text{tous les niveaux de } \gamma_k}} \Bigg]^\top,
\end{aligned}$$

où $\otimes$ est le produit de Kronecker et $\mathbf{u}_{g2k}$ sont les termes d'interaction entre γ_k et $\gamma_{k+1}, \ldots, \gamma_K$. Nous obtenons alors

$$\begin{aligned}
\mathbf{z}^{(\gamma)} &= \left[\mathbf{z}_1^{(\gamma)\top} \; \mathbf{z}_{21}^{(\gamma)\top} \ldots \mathbf{z}_{2(K-1)}^{(\gamma)\top} \right]^\top, \\
\mathbf{u}_g &= \left[\mathbf{u}_{g1}^\top \, \mathbf{u}_{g21}^\top \ldots \mathbf{u}_{g2(K-1)}^\top \right]^\top.
\end{aligned}$$

En général, pour $t \le K,$ le coefficient de l'interaction entre les événements $\{\gamma_{k_1} = l_{k_1}\}, \ldots, \{\gamma_{k_t} = l_{k_t}\}$ est désigné par $u_{g,k_1 \ldots k_t(l_{k_1} \ldots l_{k_t})}$. Comme auparavant, le coefficient est fixé à zéro par convention si $\min\left(l_1, \ldots, l_t\right) = 0.$ La covariable associée au coefficient est l'indicateur $I\left(\left(\gamma_{k_1}, \ldots, \gamma_{k_t}\right) = \left(l_{k_1}, \ldots, l_{k_t}\right)\right)$. Quand le modèle comprend tous les principaux termes et interactions d'ordre d ou d'un ordre inférieur, nous avons

$$\begin{aligned}
\mathbf{z}^{(\gamma)\top} \mathbf{u}_g &= \sum_{t=1}^{d} \sum_{1 \le k_1 < \ldots < k_t \le K} \sum_{l_{k_1}=1}^{H_{k_1}} \ldots \sum_{l_{k_t}=1}^{H_{k_t}} \\
&\qquad I\left(\left(\gamma_{k_1}, \ldots, \gamma_{k_t}\right) = \left(l_{k_1}, \ldots, l_{k_t}\right)\right) u_{g,k_1 \ldots k_t(l_{k_1} \ldots l_{k_t})}.
\end{aligned}$$

Selon le théorème 2 de l'annexe B, cela signifie que l'on peut estimer les paramètres du modèle de mélange asymptotique (dont la couverture $P(i \in S_A)$) de manière convergente en maximisant la probabilité des $\mathbf{n}_i,$ dans les conditions énoncées. À titre d'exemple, cette méthodologie peut être intéressante dans la configuration simple suivante, où le couplage est fondé sur des comparaisons exactes du nom de famille (premier groupe), du prénom (deuxième groupe) et de la date de naissance (troisième groupe), avec $K = 3,$ $H_1 = H_2 = H_3 = 1,$ $\Gamma = \{0,1\}^3 \setminus \{(0,0,0)\}$ et $|\Gamma| = 7.$ On peut estimer la couverture en maximisant la

probabilité des $\mathbf{n}_i$, si les différentes concordances n'ont pas d'interactions du troisième ordre dans les paires appariées pour chaque valeur possible d'un enregistrement dans S_B, c'est-à-dire que tous les termes principaux et les interactions du deuxième ordre peuvent être inclus (cela donne au total six paramètres en plus de la couverture $P(i \in S_A)$). Cela est particulièrement vrai si les concordances sont indépendantes dans les paires appariées. Dans ce cas, la solution est liée à celle décrite par Račinskij et coll. (2019). Elle est également liée à la solution décrite par Brown et coll. (2020), sauf qu'elle ne repose pas sur des examens manuels. Au-delà de ce cas particulier, la distribution des vrais positifs est associée à sept paramètres inconnus ($P(i \in S_A)$ et les six paramètres log-linéaires) et à sept équations (une pour $p^{(\gamma)}$ pour chaque γ), pour chaque composante du mélange.

Ici, quelques remarques sont nécessaires. Premièrement, le modèle proposé tient implicitement compte de toutes les interactions entre les variables de couplage dans les paires non appariées, tout en tenant compte de toutes les interactions d'ordre $K-1$ ou d'ordre inférieur dans les paires appariées, dans chaque classe d'enregistrement. Il offre donc une plus grande flexibilité de modélisation que les mélanges log-linéaires classiques, tout en conservant la propriété d'identification (voir les lemmes 3 et 4) et la capacité d'estimer de manière convergente les paramètres connexes (voir le théorème 2). Cela s'observe le mieux dans le cas plus simple où la distribution des vrais positifs est homogène entre les enregistrements et $H_1 = \ldots = H_K = 1$, c'est-à-dire K comparaisons dichotomiques. Dans ce cas, il est clair que l'on ne peut pas utiliser un mélange log-linéaire à K dimensions à deux composantes pour modéliser les paires d'enregistrements, tout en tenant compte de toutes les interactions d'ordre $K-1$ ou d'ordre inférieur dans les paires appariées. En effet, cela signifie qu'il y a au moins 2^K paramètres libres, dont $2^K - 2$ paramètres pour les paires appariées, un paramètre pour la proportion de mélange et au moins un paramètre pour les paires non appariées. Cependant, il n'y a que 2^K combinaisons observables et, par conséquent, seulement $2^K - 1$ équations pour déterminer les paramètres. Le même problème se produit quand $K = 2$, même en l'absence d'interactions dans la distribution des paires appariées. Deuxièmement, la souplesse de modélisation supplémentaire facilite grandement la conception des règles de couplage, qui répondent aux conditions de l'estimation convergente de la couverture (voir le théorème 2 et le lemme 4).

4.6 Capture hétérogène et enregistrements incomplets

La méthodologie est applicable quand la probabilité de capture varie selon la poststrate en fonction des covariables qui sont enregistrées sans erreur dans chaque échantillon, dans la mesure où les hypothèses posées (voir les sections 2.1, 3.3 et l'annexe) se vérifient dans chaque poststrate. Dans ce cas, S_A et S_B correspondent aux sous-ensembles d'enregistrements d'une poststrate, où la probabilité de capture peut être estimée au moyen d'un des modèles de voisinage. Bien entendu, la construction des poststrates est une question pratique importante, qui sera étudiée dans de futurs travaux.

Une autre préoccupation d'ordre pratique est la présence d'enregistrements incomplets dans l'un des échantillons. Pour traiter la question sans encombrer la notation, supposons que S_A et S_B désignent maintenant les deux échantillons dans une poststrate, que S'_A et S'_B désignent les sous-échantillons correspondants des enregistrements complets, et que $P(i \in S'_A)$ désigne la couverture de S'_A dans la poststrate,

où i est une unité qui s'y trouve. Little et Rubin (1987, chapitre 1.2) ont décrit différentes stratégies pour effectuer une analyse statistique en présence d'enregistrements incomplets. Une première solution consiste à utiliser uniquement les enregistrements complets, avec ou sans pondération visant à tenir compte des enregistrements incomplets. Les deux autres solutions sont l'imputation des valeurs manquantes et la méthode fondée sur un modèle, qui consiste à maximiser la probabilité que les données soient incomplètes. Dans ce cas, on peut envisager la première solution sans repondérer les enregistrements complets, quand les hypothèses énoncées s'appliquent dans chaque poststrate, notamment le fait que l'inclusion dans S'_A et celle dans S'_B sont indépendantes et que la probabilité d'inclusion dans S'_A est uniforme. Essentiellement, l'idée consiste à traiter les données manquantes comme une deuxième étape de la sélection et à estimer la couverture en deux étapes comme suit. Premièrement, on applique l'une des deux méthodes proposées dans la poststrate pour obtenir une estimation de $\hat{P}(i \in S'_A)$ de la couverture de S'_A. Deuxièmement, on estime la couverture de S_A (dans la poststrate) par $|S_A| / (|S'_A| / \hat{P}(i \in S'_A))$.

5. Simulations

La méthodologie proposée est évaluée au moyen de simulations Monte Carlo comprenant 100 répétitions. Dans une répétition, on génère une population finie comportant 100 000 personnes, où chaque personne se voit attribuer un nom de famille et une date de naissance. On tire deux échantillons de Bernoulli de cette population, dans lesquels le nom de famille et la date de naissance ont pu être enregistrés avec des erreurs typographiques. Ensuite, les échantillons sont liés et la couverture est estimée au moyen des modèles proposés et selon Ding et Fienberg (1994), Di Consiglio et Tuoto (2015) et Račinskij et coll. (2019), à des fins de comparaison. Différents scénarios sont envisagés en fonction de différentes règles de couplage, y compris certains où s'applique l'hypothèse d'indépendance conditionnelle et où le rappel est parfait. Des précisions sont données dans les paragraphes qui suivent.

5.1 Population finie et sources de données

Pour le nom de famille et la date de naissance, les fréquences sont fondées sur le croisement des répartitions du nom de famille et de l'âge du recensement de la population de 2010 aux États-Unis (US Census Bureau, 2020, 2016). Pour le nom de famille, la fréquence relative est calculée après exclusion de l'observation « tous les autres noms de famille ». Par souci de simplicité, le mois est tiré uniformément de $\{1, \ldots, 12\}$ et le jour est tirée indépendamment et uniformément de $\{1, \ldots, 30\}$. Par conséquent, les composantes du nom de famille et de la date sont mutuellement indépendantes dans la population. Deux registres complets sont créés et les variables sont perturbées dans le deuxième registre. Cette perturbation est décrite du point de vue de la concordance exacte pour le nom de famille, le jour de naissance ou le mois de naissance, et d'un critère de référence, qui consiste à avoir le même code SOUNDEX pour le nom de famille, la même année de naissance, ainsi qu'une différence absolue inférieure à 2 pour le jour et le mois. Pour être plus précis, supposons que γ_1, γ_2 et γ_3 désignent les variables indicatrices, qui correspondent à la

satisfaction du critère de référence en plus d'avoir le même nom de famille, le même jour de naissance ou le même mois de naissance, comme le montre le tableau 5.1. Par exemple, quand $\gamma_1 = 1,$ le critère de référence est satisfait et le nom de famille est identique. Quand $\gamma_1 = 0,$ le critère de référence n'est pas satisfait ou il est satisfait et le nom de famille est différent. Dans chaque cas, le jour et le mois de naissance peuvent être identiques ou différents. Pour une personne donnée, les enregistrements associés dans le premier registre et le deuxième registre sont appelés respectivement « premier enregistrement » et « second enregistrement ». On obtient le second enregistrement en tirant d'abord $\gamma = (\gamma_1, \gamma_2, \gamma_3),$ puis en choisissant la valeur de l'enregistrement en fonction de la valeur du premier enregistrement et $\gamma.$ Par exemple, quand $\gamma = (0,1,1),$ le second enregistrement est tel qu'il a la même date de naissance que le premier enregistrement, mais un nom de famille différent ayant le même code SOUNDEX. Au moyen de la convention de codage « dummy », nous pouvons écrire la distribution de γ sous une forme log-linéaire comme suit :

$$P(\gamma) = \exp\left(u + \sum_{k=1}^{3} \gamma_k u_{k(1)} + \sum_{1 \le k_1 < k_2 \le 3} \gamma_{k_1}\gamma_{k_2} u_{k_1 k_2 (11)} + \gamma_1\gamma_2\gamma_3 u_{123(111)} \right), \tag{5.1}$$

où le terme constant u est une fonction des termes principaux et des termes d'interaction parce que $\sum_{\gamma \in \{0,1\}^3} P(\gamma) = 1.$ Par souci de simplicité, nous choisissons les paramètres de sorte que $u_{1(1)} = u_{2(1)} = u_{3(1)}$ et $u_{12(11)} = u_{13(11)} = u_{23(11)}.$ Quand $\gamma_1 = 0,$ le nom de famille dans le second enregistrement est tiré des noms de famille de l'autre recensement ayant le même code SOUNDEX, selon leur fréquence. Quand $\gamma_2 = 0,$ on obtient le jour dans le second enregistrement en augmentant ou en diminuant aléatoirement le jour du premier enregistrement de 1, avec une probabilité de 1/2 pour chaque option, sauf quand le jour est 1 ou 30 dans le premier enregistrement, auquel cas le jour est respectivement augmenté de 1 ou réduit de 1. De même, quand $\gamma_3 = 0,$ on obtient le mois du second enregistrement en augmentant ou en diminuant aléatoirement le mois du premier enregistrement de 1, avec une probabilité de 1/2 pour chaque option, sauf quand le mois est 1 ou 12 dans le premier enregistrement, auquel cas le mois est augmenté de 1 ou réduit de 1, respectivement. On tire un échantillon Bernoulli indépendant de chaque registre, avec une probabilité d'inclusion de 0,9, qui est la couverture réelle.

Tableau 5.1
Indicateurs des perturbations dans un enregistrement.

Indicateur	Critère de référence	Nom de famille identique	Jour de naissance identique	Mois de naissance identique
γ_1	√	√	?	?
γ_2	√	?	√	?
γ_3	√	?	?	√

5.2 Couplage

On considère deux règles de couplage. La première règle permet de coupler les paires qui répondent au critère de référence, ou le sous-ensemble de ces paires dans lequel il y a au moins une concordance exacte pour le nom de famille, le jour de naissance ou le mois de naissance, selon le scénario. On utilise les liens qui en résultent pour estimer la couverture au moyen des modèles de voisinage univariés et multivariés et d'un modèle classique de mélange log-linéaire qui intègre l'hypothèse d'indépendance conditionnelle décrite par Račinskij et coll. (2019). Pour une paire donnée, le vecteur des résultats est basé sur les indicateurs de concordance exacte pour le nom de famille, le jour de naissance et le mois de naissance, par exemple $(1,1,1)$ pour une concordance parfaite sur le nom de famille et les deux composantes de la date. Afin d'estimer la couverture au moyen des méthodes proposées par Ding et Fienberg (1994) et Di Consiglio et Tuoto (2015), il faut une deuxième règle de couplage, avec au plus un lien par enregistrement ainsi que des estimations manuelles de l'exactitude du couplage qui en résulte. On dérive cette règle à partir de la première en supprimant un lien, si au moins un enregistrement impliqué comporte plusieurs liens. Les estimations manuelles de l'exactitude du couplage sont fondées sur un échantillon aléatoire simple de 1 000 paires d'enregistrements, qui satisfont au critère de référence, et l'utilisation de la table de vérité. Notons que cette procédure ne tient pas compte des faux négatifs générés par le critère de référence (semblable à un critère de création des pochettes), lorsqu'ils existent.

5.3 Scénarios

Cinq scénarios sont envisagés. Dans le premier scénario, l'hypothèse d'indépendance conditionnelle est satisfaite à partir de $u_{1(1)} = 1, u_{12(11)} = 0$ et $u_{123(111)} = 0,$ et la première règle de couplage est fondée sur le critère de référence et le rappel est parfait. Dans le deuxième scénario, il y a un écart par rapport à l'indépendance conditionnelle en raison des interactions du deuxième ordre, à partir de $u_{1(1)} = 1$ et $u_{12(11)} = 1$ et $u_{123(111)} = 0,$ mais la première règle de couplage reste fondée sur le critère de référence. Dans le troisième scénario, il y a aussi un écart par rapport à l'indépendance conditionnelle en raison des interactions des deuxième et troisième ordres, à partir de $u_{1(1)} = 1, u_{12(11)} = 1$ et $u_{123(111)} = 1/4,$ mais aucun changement n'est apporté à la première règle de couplage. Le quatrième scénario est identique au deuxième scénario, sauf que la première règle de couplage permet de lier maintenant les paires qui répondent au critère de référence et qui ont au moins une concordance exacte sur le nom de famille, le jour de naissance ou le mois de naissance. Enfin, le cinquième scénario est identique au quatrième, sauf que l'interaction du troisième ordre est ajoutée à partir de $u_{123(111)} = 1/4.$ Les caractéristiques des différents scénarios sont résumées dans les tableaux 5.2 et 5.3. Dans ce dernier tableau, les chiffres correspondent aux moyennes des répétitions.

Tableau 5.2
Scénarios de simulation.

Scénario	Paramètres log-linéaires			Indépendance conditionnelle
	$u_{k(1)}$	$u_{k_1k_2(11)}$	$u_{123(111)}$	
1	1	0	0	√
2 et 4	1	1	0	X
3 et 5	1	1	1-4	X

Tableau 5.3
Moyennes empiriques des taux d'erreur de couplage.

Scénario	Couplage	Rappel	Précision	Taux de faux positifs (TFP) $\times 10^{-9}$	Rappel parfait
1	1	1,000	0,952	498,92	√
	2	0,944	1,000	4,15	X
2 et 3	1	1,000	0,953	495,03	√
	2	0,950	1,000	4,17	X
4 et 5	1	0,996	0,964	371,84	X
	2	0,957	1,000	3,48	X

5.4 Estimateurs

Les modèles de voisinage sont appliqués selon une distribution homogène des vrais positifs, pour traduire la configuration actuelle et la situation en pratique, où l'on s'attend à ce que l'hétérogénéité de la distribution des faux positifs soit la source dominante d'hétérogénéité pour la distribution des n_i (Dasylva et Goussanou, 2022). Cela signifie que la probabilité p_g et le vecteur $\mathbf{p}_g = \left[p_g^{(\gamma)} \right]_{\gamma \in \Gamma}$ sont identiques pour toutes les classes. Cela signifie également que le paramètre $\boldsymbol{\beta}_g$ est identique dans toutes les classes, si $\mathbf{p}_g$ a la forme $\varrho\left(\boldsymbol{\beta}_g\right)$ pour une fonction connue $\varrho(.)$. Par souci de commodité, supposons que p, $\mathbf{p}$ et $\boldsymbol{\beta}$ désignent respectivement les valeurs communes de p_g, $\mathbf{p}_g$ et $\boldsymbol{\beta}_g$. Supposons aussi que

$$r^{(\gamma)} = \exp\left(u + \sum_{k=1}^{3} \gamma_k u_{k(1)} + \sum_{1 \le k_1 < k_2 \le 3} \gamma_{k_1} \gamma_{k_2} u_{k_1 k_2 (11)} \right), \gamma \in \Gamma = \{0,1\}^3, \tag{5.2}$$

où $\sum_{\gamma \in \{0,1\}^3} r^{(\gamma)} = 1$, le terme constant u est une fonction des autres paramètres, et le deuxième membre inclut seulement les interactions du deuxième ordre contrairement à celui de (5.1). Ensuite

$$\boldsymbol{\beta} = \left(u_{1(1)}, u_{2(1)}, u_{3(1)}, u_{12(11)}, u_{13(11)}, u_{23(11)} \right),$$

et la fonction $\varrho(.)$ est caractérisée par

$$p^{(\gamma)} = P(i \in S_A)\, r^{(\gamma)}, \gamma \in \Gamma = \{0,1\}^3 \setminus \{(0,0,0)\}.$$

On calcule les estimations en maximisant la probabilité numériquement dans R, où le nombre de classes est choisi en minimisant le critère d'information d'Akaike. Dans le modèle univarié, les estimations sont fondées sur le plafonnement des n_i par 10 (c'est-à-dire le remplacement de n_i par $\min(10, n_i)$) et la maximisation de la probabilité des observations qui en résultent, comme le décrivent Dasylva et Goussanou (2022). Lorsque l'on utilise le modèle multivarié, on estime la couverture quand seuls les termes principaux sont inclus et aussi quand les termes d'interaction du deuxième ordre sont inclus et qu'on ignore que les termes principaux sont égaux, et que les termes d'interaction du deuxième ordre sont égaux. En général, la maximisation numérique de la probabilité est plus difficile qu'avec le modèle univarié, et les estimations qui en résultent perdent en précision quand la précision du couplage diminue. C'est pourquoi il faut une bonne procédure d'initialisation, décrite à l'annexe C. À des fins de comparaison, nous calculons aussi les estimateurs proposés par Ding et Fienberg (1994), Di Consiglio et Tuoto (2015) et Račinskij et coll. (2019) ainsi que l'estimateur naïf de capture-recapture, qui ne tient pas compte des erreurs de couplage. Notons

que ce dernier estimateur est calculé comme étant le rapport du nombre de liens par la deuxième règle de couplage sur $|S_B|$.

5.5 Résultats

Les résultats de la simulation sont présentés dans le tableau 5.4. Dans le scénario 1, où l'hypothèse d'indépendance conditionnelle s'applique, la meilleure performance est obtenue par les estimateurs de Račinskij et coll. (2019) et les estimateurs de voisinage, tant pour ce qui est du biais relatif que de l'erreur quadratique moyenne, avec un avantage pour les estimateurs de voisinage quand on observe cette dernière mesure de performance. Parmi les estimateurs de voisinage, le modèle univarié affiche les meilleures performances pour ce qui est du biais, de la variance et de l'erreur quadratique moyenne. Sans surprise, l'estimateur naïf affiche les pires performances, tandis que les estimateurs de Ding et Fienberg (1994), et de Di Consiglio et Tuoto (2015) donnent de meilleurs résultats, mais une grande variance, parce qu'ils intègrent des estimations manuelles de l'exactitude du couplage. Notons que le mélange log-linéaire proposé par Račinskij et coll. (2019) présente un biais, une variance et une erreur quadratique moyenne plus grands que les estimateurs basés sur des modèles de voisinage multivariés, à une petite exception près. (Dans le tableau 5.4, le mélange log-linéaire comporte un biais relatif légèrement plus petit que celui du modèle de voisinage multivarié avec interactions.) En effet, tous ces estimateurs visent à estimer la couverture en exploitant la structure de corrélation des variables de couplage, ce que fait pleinement le mélange log-linéaire en intégrant l'indépendance des variables de couplage dans les paires appariées et dans les paires non appariées. Toutefois, les modèles de voisinage multivariés n'exploitent que l'information sur la structure de corrélation dans les paires appariées sans contrainte sur les paires non appariées. Pourtant, ils donnent des estimateurs qui sont nettement plus précis que le mélange log-linéaire pour ce qui est de l'erreur quadratique moyenne. Cela illustre la différence importante entre les mélanges log-linéaires classiques et les modèles de voisinage multivariés, quand ces derniers intègrent une spécification log-linéaire de la structure de corrélation dans les paires appariées.

Dans le scénario 2, les estimateurs proposés continuent de donner les plus petites erreurs quadratiques moyennes. Comme auparavant, le modèle univarié affiche les meilleures performances globales pour ce qui est du biais, de la variance et de l'erreur quadratique moyenne. Parmi les deux modèles multivariés, celui qui comprend les interactions donne de meilleurs résultats, comme on peut s'y attendre, étant donné un biais environ quarante fois plus petit et une erreur quadratique moyenne environ cinq fois plus petite. Sans surprise, l'estimateur de Račinskij et coll. (2019) donne de moins bons résultats que dans le scénario précédent, en raison de la violation de l'hypothèse d'indépendance conditionnelle dans ce scénario. Cette dégradation est toutefois telle que ses performances sont pires que celles de l'estimateur naïf pour chaque mesure de performance. De façon intéressante, nous observons que ses performances sont aussi nettement moins bonnes que celles de l'estimateur basé sur le modèle de voisinage multivarié sans interactions, étant donné qu'il comporte un biais relatif et une erreur quadratique moyenne plus grands de plus d'un ordre de grandeur, tandis que ce dernier estimateur donne de meilleurs résultats que l'estimateur naïf pour chaque mesure de performance. Entre autres explications possibles, cela est attribuable au fait que le modèle de

voisinage multivarié tient implicitement compte de toutes les interactions dans la distribution des paires non appariées, tandis que le modèle de Račinskij et coll (2019) ignore ces interactions. Cela est une illustration supplémentaire de la différence entre les mélanges log-linéaires classiques et les modèles de voisinage multivariés. Comme auparavant, les estimateurs de Ding et Fienberg (1994) et de Di Consiglio et Tuoto (2015) affichent de moins bons résultats que les modèles de voisinage, en raison de la variance de l'exactitude estimée du couplage. Cependant, pour ce qui est du biais et de l'erreur quadratique moyenne, ils donnent de meilleurs résultats que l'estimateur naïf et que celui de Račinskij et coll. (2019). Le scénario 3 diffère du scénario 2 par l'ajout d'une interaction du troisième ordre avec le coefficient 1/4. Ce changement a toutefois un effet négligeable sur les résultats obtenus et les tendances observées.

Dans le scénario 4, la première règle de couplage a un faible taux de faux négatifs d'environ 0,4 % (c'est-à-dire un rappel imparfait), car elle ne permet pas de lier les paires qui n'ont pas de concordance exacte pour le nom de famille, le jour de naissance ou le mois de naissance. Cela a un effet direct sur l'estimateur de voisinage univarié, qui a maintenant la troisième plus petite erreur quadratique moyenne, derrière les deux estimateurs de voisinage multivariés, celui qui présente les interactions ayant les meilleures performances. L'exclusion des paires sans concordance exacte dégrade davantage les performances de l'estimateur de mélange log-linaire (comparativement aux scénarios 3 et 4), qui présente toujours la plus grande erreur quadratique moyenne et de moins bonnes performances que l'estimateur naïf. Toutefois, ce changement a un effet limité sur les performances des estimateurs de Ding et Fienberg (1994) et de Di Consiglio et Tuoto (2015). Le scénario 5 diffère du scénario 4 par l'ajout d'une interaction du troisième ordre avec le coefficient 1/4. Ce changement a toutefois un effet négligeable sur les résultats.

En résumé, les résultats de simulation démontrent que les estimateurs proposés peuvent servir à estimer la couverture en ayant un petit biais relatif et une erreur quadratique moyenne plus petite que les autres estimateurs proposés par Ding et Fienberg (1994), Di Consiglio et Tuoto (2015) et Račinskij et coll. (2019) quand les faux négatifs sont négligeables ou que les probabilités de vrais positifs sont contraintes par une spécification log-linéaire.

Tableau 5.4
Résultats de simulation.

Scénario	Estimateur	Biais relatif (%)	Variance $\times 10^{-7}$	Erreur quadratique moyenne $\times 10^{-7}$
1	Naïf	-5,522	12,90	24 711,31
	R	-0,034	824,62	817,32
	DF	1,618	2 377,74	4 475,68
	DT	1,159	2 550,13	3 613,51
	VUNI	-0,003	8,43	8,35
	VMULTI sans interactions	-0,023	28,25	28,41
	VMULTI avec interactions du deuxième ordre	-0,119	27,06	38,25

DF : estimateur de Ding et Fienberg (1994)
DT : estimateur de Di Consiglio et Tuoto (2015)
VMULTI : estimateur basé sur le modèle de voisinage multivarié
Naïf : estimateur de Lincoln-Petersen qui ignore les erreurs de couplage
R : estimateur de Račinskij et coll. (2019)
VUNI : estimateur basé sur le modèle de voisinage univarié

Tableau 5.4(suite)
Résultats de simulation.

Scénario	Estimateur	Biais relatif (%)	Variance $\times 10^{-7}$	Erreur quadratique moyenne $\times 10^{-7}$
2 et 3	Naïf	-4,994	15,10	20 216,56
	R	-7,784	324,65	49 403,40
	DF	1,667	3 381,80	5 598,43
	DT	0,961	3 560,11	4 272,01
	VUNI	-0,004	8,31	8,24
	VMULTI sans interactions	-0,423	21,05	165,44
	VMULTI avec interactions du deuxième ordre	-0,091	23,70	30,12
4 et 5	Naïf	-4,292	15,46	14 934,57
	R	-3,497	280 414,54	287 515,73
	DF	1,760	2 255,34	4 740,96
	DT	1,159	2 550,13	3 613,51
	VUNI	-0,393	9,30	134,50
	VMULTI sans interactions	-0,423	21,05	165,44
	VMULTI avec interactions du deuxième ordre	-0,091	23,70	30,12

DF : estimateur de Ding et Fienberg (1994)
DT : estimateur de Di Consiglio et Tuoto (2015)
VMULTI : estimateur basé sur le modèle de voisinage multivarié
Naïf : estimateur de Lincoln-Petersen qui ignore les erreurs de couplage
R : estimateur de Račinskij et coll. (2019)
VUNI : estimateur basé sur le modèle de voisinage univarié

6. Conclusion

Nous avons décrit une nouvelle méthodologie aux fins d'estimation par la capture-recapture comportant des erreurs de couplage, qui est fondée sur la modélisation du nombre de liens à partir d'un enregistrement, sans examens manuels, notamment un modèle univarié et un modèle multivarié connexe. Dans le modèle univarié, on estime la couverture en liant les enregistrements dont le rappel est suffisamment élevé. Dans le modèle multivarié, on estime la couverture en contraignant les interactions dans les paires appariées au moyen d'une spécification log-linéaire, tout en permettant des interactions arbitraires dans les paires non appariées, ce qui tranche nettement avec les mélanges log-linéaires classiques. Dans ce dernier cas, il faut lier les enregistrements avec une grande précision pour obtenir une estimation fiable de la couverture. Les simulations à partir des données de recensement publiques démontrent les bonnes performances des estimateurs proposés, comparativement aux solutions précédentes. De futurs travaux porteront sur l'obtention des variances et d'intervalles de confiance, ainsi que sur la validation de la spécification log-linéaire quand le modèle multivarié est utilisé.

Remerciements

Le présent article expose les opinions des auteurs qui ne sont pas nécessairement celles de Statistique Canada.

Annexe

A. Extension pour la sous-couverture

La présente annexe vise à étendre les résultats de Dasylva et Goussanou (2022) pour démontrer que

$$\begin{aligned}
\frac{\text{VP}}{\text{VP + FN}} &\xrightarrow{p} P(i \in S_A)^{-1}\, \overline{p}, \\
\frac{\text{VP}}{\text{VP + FP}} &\xrightarrow{p} \frac{\overline{p}}{\overline{p} + \overline{\lambda}}, \\
\hat{\overline{p}} &\xrightarrow{p} \overline{p}, \\
\frac{\hat{\overline{p}}}{\hat{\overline{p}} + \hat{\overline{\lambda}}} &\xrightarrow{p} \frac{\overline{p}}{\overline{p} + \overline{\lambda}}.
\end{aligned}$$

Par conséquent, $P(i \in S_A)^{-1}\, \hat{\overline{p}}$ et $\hat{\overline{p}} \Big/ \left(\hat{\overline{p}} + \hat{\overline{\lambda}}\right)$ estiment le rappel et la précision de façon convergente, en ce sens que

$$\begin{aligned}
P(i \in S_A)^{-1}\, \hat{\overline{p}} - \frac{\text{VP}}{\text{VP + FN}} &\xrightarrow{p} 0, \\
\frac{\hat{\overline{p}}}{\hat{\overline{p}} + \hat{\overline{\lambda}}} - \frac{\text{VP}}{\text{VP + FP}} &\xrightarrow{p} 0.
\end{aligned}$$

L'extension consiste à tenir compte de la sous-couverture dans S_A. Pour continuer, il faut des notations supplémentaires. Appelons V_j' un *voisin* de V_i, si l'unité j est incluse dans S_A et V_j' est contenu dans le voisinage de V_i, c'est-à-dire $\mathcal{B}_N(V_i)$. Le voisin est considéré comme apparié si les deux enregistrements proviennent de la même unité. Sinon, il est considéré comme non apparié. De plus, définissons la notation supplémentaire suivante. Pour $i, i' \in S_B$ de sorte que $i \neq i'$, soit

$$n_i^{(0)} = \sum_{j \in S_A} I\left(V_j' \in \mathcal{B}_N(V_i)\right), \tag{A.1}$$

$$\left(n_{i|U}^{(0)}, n_{i|U}\right) = \left(n_i^{(0)}, n_i\right) - \left(I\left(\{i \in S_A\} \cap \left\{V_i' \in \mathcal{B}_N(V_i)\right\}\right), I(i \in S_A)\, L_{ii}\right), \tag{A.2}$$

$$\left(n_{ii'|U}^{(0)}\right) = \sum_{t \in S_A : t \neq i, i'} I\left(V_t' \in \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'})\right), \tag{A.3}$$

où $n_i^{(0)}$ est le nombre de voisins de V_i, $n_{i|U}^{(0)}$ est le nombre de voisins non appariés de cet enregistrement, $n_{i|U}$ est le nombre d'enregistrements non appariés, qui sont liés au même enregistrement, et $n_{ii'|U}^{(0)}$ est le nombre de voisins non appariés, qui sont communs à V_i et $V_{i'}$. En ce qui concerne $n_i^{(0)}$ et $n_{i|U}^{(0)}$, soit $S_{Ai} = \left\{t \in S_A \text{ s.t. } V_t' \in \mathcal{B}_N\left(V_i\right)\right\}$ le sous-ensemble d'unités, qui sont incluses dans S_A et ont leur enregistrement dans le voisinage, et soit $S_{Ai|U} = S_{Ai} - \{i\}$ le sous-ensemble de ces unités qui sont différentes de l'unité i. Ces dernières unités sont associées à des voisins non appariés. Enfin, pour les variables aléatoires (ou vecteurs) X, Y et Z, notons l'indépendance de X et Y par $X \perp\!\!\!\perp Y$, et leur indépendance conditionnelle étant donné Z par $X \perp\!\!\!\perp Y \mid Z$.

Le lemme suivant est une extension du lemme 2 présenté dans Dasylva et Goussanou (2022). Toutes les démonstrations se trouvent dans la version longue de l'article (Dasylva, Goussanou et Nambeu, 2024).

Lemme 1. *Supposons que* $\left[\left(I(i \in S_A), I(i \in S_B), V_i, V_i'\right)\right]_{1 \le i \le N}$ *sont indépendants et identiquement distribués et que* $Z_1, \ldots, Z_N$ *désignent des variables aléatoires identiquement distribuées, de sorte qu'elles soient conditionnellement indépendantes étant donné* $\left[\left(I(i \in S_A), I(i \in S_B), V_i, V_i'\right)\right]_{1 \le i \le N}$ *avec une distribution marginale conditionnelle de* Z_i *qui est seulement une fonction de* $I(i \in S_A)$, $I(i \in S_B)$, V_i, $S_{Ai|U}$, $\left[V_t'\right]_{t \in S_{Ai|U}}$ *et* V_i'. *Alors*

$$\left(I(i' \in S_A), V_{i'}', \left[V_t'\right]_{t \in S_{Ai'|U}}\right) \perp\!\!\!\perp \left(I(i \in S_A), V_i', \left[V_t'\right]_{t \in S_{Ai'|U}}\right) \left| \begin{array}{l} i \in S_B, V_i, S_{Ai|U}, \\ i' \in S_B, V_i', S_{Ai'|U}, \\ \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) = \varnothing, \\ V_i' \notin \mathcal{B}_N(V_{i'}), \\ V_{i'}' \notin \mathcal{B}_N(V_i), \end{array} \right. \tag{A.4}$$

$$S_{Ai'|U} \perp\!\!\!\perp \left(I(i \in S_A), V_i', \left[V_t'\right]_{t \in S_{Ai|U}}\right) \left| \begin{array}{l} i \in S_B, V_i, S_{Ai|U}, \\ i' \in S_B, V_{i'}, \\ \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) = \varnothing, \\ V_i' \notin \mathcal{B}_N(V_{i'}), \\ V_{i'}' \notin \mathcal{B}_N(V_i). \end{array} \right. \tag{A.5}$$

Aussi, pour tous $(v_i, v_{i'}) \in \mathcal{V}_N^* \times \mathcal{V}_N^*$, $a_i \in \{0,1\}$, $w_i \notin \mathcal{B}_N(v_{i'})$, $s_{Ai|U} \subset \{1, \ldots, N\} \setminus \{i, i'\}$ *et* $\left[w_t\right]_{t \in s_{Ai|U}}$ *fixes de sorte que* $\mathcal{B}_N(v_i) \cap \mathcal{B}_N(v_{i'}) = \varnothing$ *et* $w_t \in \mathcal{B}_N(v_i)$ *pour tous les* $t \in s_{Ai|U}$

$$P\left(\begin{array}{l|l} I(i \in S_A) = a_i, & i \in S_B, V_i = v_i, \\ V_i' = w_i, & S_{Ai|U} = s_{Ai|U}, \\ \left[V_t'\right]_{t \in s_{Ai|U}} = & i' \in S_B, V_{i'} = v_{i'} \\ \left[w_t\right]_{t \in s_{Ai|U}} & \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) = \varnothing, \\ & V_i' \notin \mathcal{B}_N(V_{i'}), \\ & V_{i'}' \notin \mathcal{B}_N(V_i) \end{array}\right) = P\left(\begin{array}{l|l} I(i \in S_A) = a_i, & \\ V_i' = w_i, & i \in S_B, V_i = v_i, \\ \left[V_t'\right]_{t \in s_{Ai|U}} = & S_{Ai|U} = s_{Ai|U}, \\ \left[w_t\right]_{t \in s_{Ai|U}} & V_i' \notin \mathcal{B}_N(v_{i'}) \end{array}\right). \tag{A.6}$$

Donc

$$E\left[Z_i Z_{i'} \left| \begin{array}{l} i \in S_B, V_i = v_i, n_{i|U}^{(0)} = k, \\ i' \in S_B, V_{i'} = v_{i'}, n_{i'|U}^{(0)} = \ell, \\ \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) = \varnothing, \\ V_i' \notin \mathcal{B}_N(V_{i'}), V_{i'}' \notin \mathcal{B}_N(V_i) \end{array} \right. \right] = g(v_i, k; v_{i'})\, g(v_{i'}, \ell; v_i), \tag{A.7}$$

où

$$g(v_i, k; v_{i'}) = E\left[Z_i \left| \begin{array}{l} i \in S_B, V_i = v_i, \\ n_{i|U}^{(0)} = k, V_i' \notin \mathcal{B}_N(v_{i'}) \end{array} \right. \right]. \tag{A.8}$$

Le lemme ci-dessus mène au théorème suivant, qui prolonge le théorème 1 présenté dans Dasylva et Goussanou (2022).

Théorème 1. *Considérons que* $V_N,\ \mathcal{B}_N(.),\ S_A,\ S_B$ *et* $[Z_i]_{1\le i\le N}$ *sont des variables aléatoires identiquement distribuées, de sorte que les conditions suivantes s'appliquent.*

(C.1) $\lim_{N\to\infty} P(i\in S_B)=\tau$ *pour un certain* τ *positif.*

(C.2) Pour $\Lambda\ge 0$ *ne dépendant pas de* N, $\sup_{v\in V_N^*}(N-1)\,\lambda_N^{(0)}(v)\le\Lambda$.

(C.3) Pour $c\ge 0$ *ne dépendant pas de* N

$$NP\left(\mathcal{B}_N(V_i)\cap\mathcal{B}_N(V_{i'})\ne\varnothing\mid\{i,i'\}\subset S_B\right)\le c,$$

$$NP\left(V_i'\in\mathcal{B}_N(V_{i'})\,\middle|\,\{i,i'\}\subset S_B,\ \mathcal{B}_N(V_i)\cap\mathcal{B}_N(V_{i'})=\varnothing\right)\le c.$$

(C.4) $Z_1,\ldots,Z_N$ *sont conditionnellement indépendants étant donné* $\left[\left(I(i\in S_A),I(i\in S_B),V_i,V_i'\right)\right]_{1\le i\le N}$ *de sorte que la distribution marginale conditionnelle de* Z_i *est seulement une fonction de* $I(i\in S_A)$, $I(i\in S_B)$, V_i, $S_{Ai|U}$, $[V_t']_{t\in S_{Ai|U}}$ *et* V_i'.

(C.5) $|Z_i|\le R_N\left(n_{i|U}^{(0)}\right)$ *où* $R_N(.)$ *est un polynôme de degré fini ne dépendant pas de* N *et des coefficients non négatifs de* $O\left((\log N)^d\right)$, *où* d *ne dépend pas non plus de* N.

(C.6) $\lim_{N\to\infty} E\left[Z_i\,\middle|\,i\in S_B\right]=\mu$, *où* $|\mu|<\infty$.

Alors

$$\frac{1}{|S_B|}\sum_{i\in S_B} Z_i\ \xrightarrow{p}\ \mu. \qquad (A.9)$$

Le résultat suivant montre la convergence du rappel et de la précision respectivement vers $P(i\in S_A)^{-1}\,\overline{p}$ et $\overline{p}/(\overline{p}+\overline{\lambda})$. Il prolonge le corollaire 1 dans Dasylva et Goussanou (2022) et est une conséquence directe du théorème 1.

Corollaire 1. *Supposons que les hypothèses C.1 à C.3 se vérifient et que le couplage respecte les conditions suivantes :*

(C.7) $[L_{1j}]_{1\le j\le N},\ldots,[L_{Nj}]_{1\le j\le N}$ *sont conditionnellement indépendants étant donné que* $[(I(i\in S_A),I(i\in S_B),V_i,V_i')]_{1\le i\le N}$, *quand la distribution conditionnelle de* $[L_{ij}]_{1\le j\le N}$ *est seulement une fonction de* $I(i\in S_A)$, $I(i\in S_B)$, V_i, $S_{Ai|U}$, $[V_t']_{t\in S_{Ai|U}}$ *et* V_i'.

(C.8)

$$\lim_{N\to\infty} E\left[\left(p_N(V_i),(N-1)\,\lambda_N(V_i)\right)\ \middle|\ i\in S_B\right]=\left(\overline{p},\overline{\lambda}\right). \qquad (A.10)$$

Alors

$$\left(\frac{\text{VP}}{\text{VP}+\text{FN}},\ \frac{\text{VP}}{\text{VP}+\text{FP}}\right) \xrightarrow{p} \left(\frac{\overline{p}}{P(i\in S_A)},\ \frac{\overline{p}}{\overline{p}+\overline{\lambda}}\right). \tag{A.11}$$

En particulier, (A.11) se vérifie si C.8 est remplacé par la condition

$$\left(p_N(V_i),\,(N-1)\,\lambda_N(V_i)\right)\ \Big|\ \{i\in S_B\} \xrightarrow{d} F(.,.)$$

avec $\overline{p}=\int p dF(p,\lambda)$ *et* $\overline{\lambda}=\int \lambda dF(p,\lambda).$

D'autres résultats présentés dans Dasylva et Goussanou (2022) demeurent valides. Ils sont fondés sur le théorème 1 et le corollaire 1, comme le lemme 2 (convergence dans la distribution de n_i vers un mélange comme dans le deuxième membre de (3.8)), le théorème 2 (convergence de l'estimateur de Blakely et Salmond (2002) dans le cas homogène) et le théorème 3 (convergence de l'estimateur par le maximum de vraisemblance). On peut facilement le constater en examinant les démonstrations connexes dans Dasylva et Goussanou (2022). Ainsi, les estimateurs $\hat{\overline{p}}$ et $\hat{\overline{\lambda}}$ sont convergents, et $P(i\in S_A)^{-1}\hat{\overline{p}}$ et $\hat{\overline{p}}\Big/\left(\hat{\overline{p}}+\hat{\overline{\lambda}}\right)$ sont des estimateurs convergents du rappel et de la précision, respectivement.

B. Extension multivariée

Pour décrire la version multivariée du modèle de voisinage, supposons que Γ désigne l'ensemble d'indices d'un ensemble de règles et que $L_{ij}^{(\gamma)}$ indique si la paire (i,j) est couplée par la règle $\gamma\in\Gamma$. (Pour éviter tout conflit avec la notation définie précédemment, on suppose que les règles sont indexées de sorte que Γ ne contienne pas 0.) L'ensemble Γ peut prendre diverses formes, comme un sous-ensemble de nombres entiers consécutifs à partir de 1, ou il peut être un sous-ensemble de $\{0,1\}^K$ si l'on lie les enregistrements avec les K variables de couplage et si l'on effectue une comparaison exacte pour chaque variable. Pour $\gamma\in\Gamma$, soit $n_i^{(\gamma)}=\sum_{j\in S_A} L_{ij}^{(\gamma)}$, $\mathbf{n}_i=\left[n_i^{(\gamma)}\right]_{\gamma\in\Gamma}$, et redéfinissons $\mathcal{B}_N(v)$ comme un sous-ensemble de $\mathcal{V}_N$, qui satisfait à la condition

$$\mathcal{B}_N(v)\ \supset\ \left\{v'\in\mathcal{V}_N\ \text{s.t.}\ E\left[\sum_{\gamma\in\Gamma} L_{ij}^{(\gamma)}\ \middle|\ (i,j)\in S_B\times S_A, \left(V_i,\,V_j'\right)=(v,v')\right]\ >\ 0\right\}.$$

En d'autres termes, $\mathcal{B}_N(v)$ est un surensemble de valeurs d'enregistrement, qui sont liées par au moins une règle de l'ensemble avec une probabilité positive, étant donné que $i\in S_B$ et $V_i=v$. La fonction $\lambda_N^{(0)}(.)$ reste définie par (4.4), tandis que $p_N(.)$ et $\lambda_N(.)$ sont remplacées par les vecteurs $\mathbf{p}_N(v)=\left[p_N^{(\gamma)}(v)\right]_{\gamma\in\Gamma}$ et $\boldsymbol{\lambda}_N(v)=\left[\lambda_N^{(\gamma)}(v)\right]_{\gamma\in\Gamma}$, avec

$$p_N^{(\gamma)}(v)\ =\ E\left[I\left(i\in S_A\right)L_{ii}^{(\gamma)}\ \middle|\ i\in S_B, V_i=v\right], \tag{B.1}$$

$$\lambda_N^{(\gamma)}(v)\ =\ E\left[I\left(j\in S_A\right)L_{ij}^{(\gamma)}\ \middle|\ i\in S_B, V_i=v\right],\ j\neq i, \tag{B.2}$$

et $\sum_{\gamma\in\Gamma} p_N^{(\gamma)}(v) \le 1$ pour tous les $v \in \mathcal{V}_N^*$. En d'autres termes, $p_N^{(\gamma)}(v)$ et $\lambda_N^{(\gamma)}(v)$ sont les nombres attendus de vrais positifs et de faux positifs pour la règle γ, étant donné que $i \in S_B$ et $V_i = v$. La condition de régularité de (3.7) est remplacée par la condition plus générale suivante :

$$\left(\mathbf{p}_N(V_i),\ (N-1)\,\boldsymbol{\lambda}_N(V_i)\right) \,\middle|\, \{i \in S_B\} \xrightarrow{d} F. \tag{B.3}$$

Nous avons un cas particulièrement intéressant quand $\mathbf{p}_N(v)$ a la forme $\varrho\left(\boldsymbol{\beta}_N(v)\right)$, pour une certaine fonction $\boldsymbol{\beta}_N : \mathcal{V}^* \to \mathbb{R}^m$, et une certaine *injection* $\varrho\colon \mathbb{R}^m \to [0,1]^{|\Gamma|}$ indépendante de N, où $m < |\Gamma|$. Dans ce cas, (3.7) est remplacée par la condition suivante :

$$\left(\boldsymbol{\beta}_N(V_i), (N-1)\,\boldsymbol{\lambda}_N(V_i)\right) \,\middle|\, \{i \in S_B\} \xrightarrow{d} H, \tag{B.4}$$

où H ne dépend pas de N.

Le lemme suivant indique la convergence de $\mathbf{n}_i$ vers un mélange multivarié, quand $N \to \infty$ dans les conditions données par (3.6) et (B.3) ou (B.4). La distribution latente est donnée par F ou H selon que (B.3) ou (B.4) s'applique. Dans les deux cas, les distributions des composantes proviennent de familles de distributions discrètes à $|\Gamma|$ dimensions, qui correspondent à la convolution d'une distribution multinomiale avec un produit de distributions de Poisson indépendantes. Pour décrire plus en détail les distributions limites, supposons que $\mathcal{F}$ désigne la famille des distributions des composantes selon (B.3), où chaque membre a la forme

$$\text{IMultinomial}(1, \mathbf{p}) \ * \ \text{PPoisson}(\boldsymbol{\lambda}),$$

pour $\mathbf{p} = \left[p^{(\gamma)}\right]_{\gamma\in\Gamma}$ et $\boldsymbol{\lambda} = \left[\lambda^{(\gamma)}\right]_{\gamma\in\Gamma}$. Quand (B.4) s'applique, les distributions des composantes proviennent du sous-ensemble $\mathcal{F}_\varrho$ de $\mathcal{F}$, où $\mathbf{p} = \varrho(\boldsymbol{\beta})$ pour un certain $\boldsymbol{\beta} \in \mathbb{R}^m$. Un membre de cette famille est une distribution paramétrique dont les paramètres sont $\boldsymbol{\beta} \in \mathbb{R}^m$ et $\boldsymbol{\lambda} \in (0, +\infty)^{|\Gamma|}$. Comme auparavant, toutes les démonstrations se trouvent dans la version longue de l'article (Dasylva et coll., 2024).

Lemme 2. *Supposons que (3.6) s'applique. Si (B.3) s'applique aussi,* $\mathbf{n}_i$ *converge en distribution vers le mélange des distributions provenant de* $\mathcal{F}$*, avec la distribution latente* F*. Si (B.4) s'applique aussi, alors* $\mathbf{n}_i$ *converge en distribution vers le mélange des distributions provenant de* $\mathcal{F}_\varrho$*, avec la distribution latente* H*.*

Le lemme suivant est une extension du lemme 4 présenté dans Dasylva et Goussanou (2022). Il fournit des conditions suffisantes pour l'identification de mélanges finis sur $\mathcal{F}$ (ou $\mathcal{F}_\varrho$); une propriété essentielle pour prouver la convergence des estimateurs par le maximum de vraisemblance. Le lemme nécessite un ordre lexicographique sur $(0,\infty)^{|\Gamma|}$. Pour ce faire, ordonnons les éléments de Γ en fonction d'une certaine application bijective provenant de $\{1,\ldots,|\Gamma|\}$, dans Γ, ce qui est désigné par $\gamma(.)$ avec un léger abus de la notation. Ensuite, désignons un uplet $\left[\lambda^{(\gamma)}\right]_{\gamma\in\Gamma}$ de manière équivalente par $\left[\lambda^{(\gamma(t))}\right]_{1\le t\le|\Gamma|}$, et écrivons $\boldsymbol{\lambda} \succ \boldsymbol{\lambda}'$ (c'est-à-dire $\boldsymbol{\lambda}$ est supérieur à $\boldsymbol{\lambda}'$), si $\lambda^{(\gamma(1))} > \lambda'^{(\gamma(1))}$ ou s'il existe un $t_0 = 2,\ldots,|\Gamma|$ tel que

$\lambda^{(\gamma(t))} = \lambda'^{(\gamma(t))}$ pour $t < t_0$ et $\lambda^{(\gamma(t_0))} > \lambda'^{(\gamma(t_0))}$. Aussi, convenons que $\max(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = \boldsymbol{\lambda}$ si $\boldsymbol{\lambda} \succ \boldsymbol{\lambda}'$ sinon $\max(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = \boldsymbol{\lambda}'$.

Lemme 3. *Pour les nombres entiers naturels G et G', soit $\boldsymbol{\lambda}_1 \succ \ldots \succ \boldsymbol{\lambda}_G$, et $\boldsymbol{\lambda}'_1 \succ \ldots \succ \boldsymbol{\lambda}'_G$, et désignons par h_g et h'_g les membres de $\mathcal{F}$ dont les paramètres sont $\left(\mathbf{p}_g, \boldsymbol{\lambda}_g\right)$ et $\left(\mathbf{p}'_g, \boldsymbol{\lambda}'_g\right)$, respectivement, et supposons que les mélanges $h = \sum_{g=1}^{G} \alpha_g h_g$ et $h' = \sum_{g=1}^{G'} \alpha'_g h'_g$ sont égaux, où α_g et α'_g sont positifs pour chaque g. Alors $G = G'$, $\alpha_g = \alpha'_g$ et $\left(\mathbf{p}_g, \boldsymbol{\lambda}_g\right) = \left(\mathbf{p}'_g, \boldsymbol{\lambda}'_g\right)$, pour $g = 1, \ldots, G$. De plus, s'il existe une injection $\varrho\colon \mathbb{R}^m \to [0,1]^{|\Gamma|}$ telle que $\mathbf{p}_g = \varrho(\boldsymbol{\beta}_g)$ et $\mathbf{p}'_g = \varrho(\boldsymbol{\beta}'_g)$ pour chaque g, nous avons aussi $\boldsymbol{\beta}_g = \boldsymbol{\beta}'_g$ pour chaque g.*

Le théorème suivant étend le théorème 3 de Dasylva et Goussanou (2022), qui concerne la convergence des estimateurs par le maximum de vraisemblance. Des notations supplémentaires sont nécessaires pour préciser cette extension. Pour $G \geq 1$, considérons le mélange fini de G distributions provenant de $\mathcal{F}$, où la g^{e} composante composant a une probabilité α_g et des paramètres $\mathbf{p}_g$ et $\boldsymbol{\lambda}_g$. Aussi, désignons par $\boldsymbol{\theta} = \left[\left(\alpha_g, \mathbf{p}_g, \boldsymbol{\lambda}_g\right)\right]_{1 \leq g \leq G}$ les paramètres de mélange associés et par $q(.;\boldsymbol{\theta})$, la fonction de masse de probabilité correspondante.

$$
\begin{aligned}
q(\mathbf{n};\boldsymbol{\theta}) \quad = \quad & \sum_{g=1}^{G} \alpha_g \left(I(|\,\mathbf{n}\,| = 0)\left(1 - |\mathbf{p}_g|\right) e^{-|\boldsymbol{\lambda}_g|} + I(|\,\mathbf{n}\,| > 1)\left(\left(1 - |\mathbf{p}_g|\right) \prod_{\gamma \in \Gamma} \frac{e^{-\lambda_g^{(\gamma)}} \left(\lambda_g^{(\gamma)}\right)^{n^{(\gamma)}}}{n^{(\gamma)}!} \right.\right. \\
& \left.\left. + \sum_{\gamma \in \Gamma:\, n^{(\gamma)} > 0} p_g^{(\gamma)} \frac{e^{-\lambda_g^{(\gamma)}} \left(\lambda_g^{(\gamma)}\right)^{n^{(\gamma)}-1}}{\left(n^{(\gamma)} - 1\right)!} \prod_{\gamma' \in \Gamma \setminus \{\gamma\}} \frac{e^{-\lambda_g^{(\gamma')}} \left(\lambda_g^{(\gamma')}\right)^{n^{(\gamma')}}}{n^{(\gamma')}!} \right)\right), \quad \mathbf{n} = \left[n^{(\gamma)}\right]_{\gamma \in \Gamma} \in \mathbb{N}^{|\Gamma|},
\end{aligned}
\tag{B.5}
$$

Définissons

$$
M_N(\boldsymbol{\theta}) = \frac{1}{|\,S_B\,|} \sum_{i \in S_B} \log q(\mathbf{n}_i;\boldsymbol{\theta}), \tag{B.6}
$$

et pour un nombre naturel $\tau_N > 0$, définissons

$$
\begin{aligned}
M_N(\boldsymbol{\theta};\tau_N) \quad = \quad & \frac{1}{|\,S_B\,|} \sum_{i \in S_B} \left(I\left(|\,\mathbf{n}_i\,| \leq \tau_N\right) \log q(\mathbf{n}_i;\boldsymbol{\theta}) \right. \\
& \left. + I\left(|\,\mathbf{n}_i\,| \geq \tau_N + 1\right) \log\left(\sum_{\mathbf{n} \in \mathbb{N}^{|\Gamma|}:\, |\mathbf{n}| \geq \tau_N + 1} q(\mathbf{n};\boldsymbol{\theta}) \right) \right).
\end{aligned}
\tag{B.7}
$$

Pour $\boldsymbol{\theta}_\varrho = \left[\left(\alpha_g, \boldsymbol{\beta}_g, \boldsymbol{\lambda}_g\right)\right]_{1 \leq g \leq G}$, soit $\boldsymbol{\theta}\left(\boldsymbol{\theta}_\varrho\right) = \left[\left(\alpha_g, \varrho\left(\boldsymbol{\beta}_g\right), \boldsymbol{\lambda}_g\right)\right]_{1 \leq g \leq G}$. De même pour un nombre naturel d, supposons que $\mathbf{0}_d$ désigne le d-uplet avec tous les zéros.

Comme c'est le cas précédemment, on peut estimer les paramètres de mélange en maximisant la log-vraisemblance des $\mathbf{n}_i$, c'est-à-dire $M_N(.)$ ou $M_N(;\tau_N)$. Le théorème suivant indique que les estimateurs

qui en résultent sont convergents dans des conditions appropriées, qui comprennent (B.3) ou (B.4). Dans ce dernier cas, on suppose que la fonction ϱ est une injection.

Théorème 2. *Pour $G^* \geq 2$ et $\nu \in (0,1)$, supposons que $\Theta_1, \ldots, \Theta_{G^*}$ désignent des sous-ensembles compacts de $\mathbb{R}^{(2|\Gamma|+1)G^*-1}$ tels que*

$$\begin{aligned}\Theta_G \subset \Big\{ & \Big[\big(\alpha_g, \mathbf{p}_g, \boldsymbol{\lambda}_g\big)\Big]_{1\leq g\leq G^*} \in \mathbb{R}^{(2|\Gamma|+1)G^*-1} \text{ s.t.} \\ & \big(\alpha_g, \mathbf{p}_g, \boldsymbol{\lambda}_g\big) \in (\nu,1]\times[0,1]^{|\Gamma|}\times[\nu,\Lambda]^{|\Gamma|} \text{ et } \big|\mathbf{p}_g\big| \leq 1-\nu \text{ et} \\ & \boldsymbol{\lambda}_{g+1} \succ \boldsymbol{\lambda}_g \text{ si } g \geq G^*-G+1, \\ & \big(\alpha_g, \mathbf{p}_g, \boldsymbol{\lambda}_g\big) = (0, \mathbf{0}_{|\Gamma|}, \mathbf{0}_{|\Gamma|}) \text{ si } g \leq G^*-G, \\ & \alpha_1+\ldots+\alpha_{G^*}=1\Big\},\ G=1,\ldots,G^*.\end{aligned}$$

et supposons que $\Theta = \bigcup_{G=1}^{G^*}\Theta_G$. Pour une application injective $\varrho\colon \mathbb{R}^m \to [0,1]^{|\Gamma|}$, supposons aussi que $\Theta_{\varrho 1}, \ldots, \Theta_{\varrho G^*}$ désigne des sous-ensembles compacts de $\mathbb{R}^{(|\Gamma|+m+1)G^*-1}$ tels que

$$\begin{aligned}\Theta_{\varrho G} \subset \Big\{ & \Big[\big(\alpha_g, \boldsymbol{\beta}_g, \boldsymbol{\lambda}_g\big)\Big]_{1\leq g\leq G^*} \in \mathbb{R}^{(|\Gamma|+m+1)G^*-1} \text{ s.t.} \\ & \big(\alpha_g, \boldsymbol{\beta}_g, \boldsymbol{\lambda}_g\big) \in (\nu,1]\times\mathbb{R}^m\times[\nu,\Lambda]^{|\Gamma|} \text{ et } \big|\varrho\big(\boldsymbol{\beta}_g\big)\big| \leq 1-\nu \text{ et} \\ & \boldsymbol{\lambda}_{g+1} \succ \boldsymbol{\lambda}_g \text{ si } g \geq G^*-G+1, \\ & \big(\alpha_g, \boldsymbol{\beta}_g, \boldsymbol{\lambda}_g\big) = (0, \mathbf{0}_m, \mathbf{0}_{|\Gamma|}) \text{ si } g \leq G^*-G, \\ & \alpha_1+\ldots+\alpha_{G^*}=1\Big\}, G=1,\ldots,G^*,\end{aligned}$$

et supposons que $\Theta_\varrho = \bigcup_{G=1}^{G^*}\Theta_{\varrho G}$. Supposons que toutes les règles de couplage sont simples (c'est-à-dire que chaque règle est telle que la décision de coupler deux enregistrements ne dépend d'aucun autre enregistrement), que C.1-C.3 (du théorème 1) s'appliquent et que (B.3) s'applique aussi avec

$$F(\mathbf{p},\boldsymbol{\lambda}) = \sum_{g=1}^{G^*}\alpha_{0g} I\Big((\mathbf{p},\boldsymbol{\lambda}) = \big(\mathbf{p}_{0g},\boldsymbol{\lambda}_{0g}\big)\Big),$$

$\boldsymbol{\theta}_0 = \Big[\big(\alpha_{0g},\mathbf{p}_{0g},\boldsymbol{\lambda}_{0g}\big)\Big]_{1\leq g\leq G^*} \in \Theta_{G_0}$ *et* $1 \leq G_0 \leq G^*$, *et supposons que* $\hat{\boldsymbol{\theta}}_{1N}$, $\hat{\boldsymbol{\theta}}_{2N}$ *et* $\hat{\boldsymbol{\theta}}_{3N}$ *désignent les estimateurs, qui maximisent respectivement $M_N(.)$ sur Θ_{G_0}, $M_N(.)$ sur Θ, et $M_N(.;\tau_N)$ sur Θ, où τ_N est un nombre entier naturel tel que $\tau_N \to \infty$ et $\tau_N = O\big(\log N\big)$. Alors $\hat{\boldsymbol{\theta}}_{1N}$, $\hat{\boldsymbol{\theta}}_{2N}$ et $\hat{\boldsymbol{\theta}}_{3N}$ convergent en probabilité vers $\boldsymbol{\theta}_0$. De plus, supposons que $\mathbf{p}_N(v)$ a aussi la forme $\varrho(\boldsymbol{\beta}_N(v))$ pour $\boldsymbol{\beta}_N\colon \mathcal{V}_N^* \to \mathbb{R}^m$, qui satisfait à (B.4) avec*

$$H(\boldsymbol{\beta},\boldsymbol{\lambda}) = \sum_{g=1}^{G^*}\alpha_{0g} I\Big((\boldsymbol{\beta},\boldsymbol{\lambda}) = \big(\boldsymbol{\beta}_{0g},\boldsymbol{\lambda}_{0g}\big)\Big),$$

$\varrho\big(\boldsymbol{\beta}_{0g}\big) = \mathbf{p}_{0g}$ *pour* $g = G^*-G_0+1,\ldots,G^*$ *et* $\boldsymbol{\theta}_{\varrho 0} = \Big[\big(\alpha_{0g},\boldsymbol{\beta}_{0g},\boldsymbol{\lambda}_{0g}\big)\Big]_{1\leq g\leq G^*} \in \Theta_{\varrho G_0}$. *Alors, on estime aussi $\boldsymbol{\theta}_{\varrho 0}$ de manière convergente en maximisant $M_N(.)$ sur $\Theta_{\varrho G_0}$, $M_N(.)$ sur Θ_ϱ, ou $M_N(.;\tau_N)$ sur Θ_ϱ.*

Dans le théorème ci-dessus, la fonction $\varrho(.)$ doit être une injection. Le lemme suivant montre que cette condition est remplie quand $\varrho(.)$ est basée sur une spécification log-linéaire non saturée des interactions dans les paires appariées.

Lemme 4. *Pour les nombres entiers naturels* $K, H_1, \ldots, H_K$, *supposons que* $\Gamma = \{0, \ldots, H_1\} \times \ldots \times \{0, \ldots, H_K\} - \mathbf{0}_K$ *et* $\mathbf{p} = \left[p^{(\gamma)} \right]_{\gamma \in \Gamma}$ *ont la forme* $p^{(\gamma)} = P(i \in S_A)\, r^{(\gamma)}$, *où* $\sum_{\gamma \in \Gamma \cup \{\mathbf{0}_K\}} r^{(\gamma)} = 1$ *et* $r^{(\gamma)}$ *a la forme log-linéaire suivante sans aucune interaction d'ordre supérieur à* $d < K$.

$$r^{(\gamma)} = \exp\left(u + \sum_{t=1}^{d} \sum_{1 \le k_1 < \ldots < k_t \le K} u_{k_1 \ldots k_t (\gamma_{k_1} \ldots \gamma_{k_t})} \right), \tag{B.8}$$

où le terme $u_{k_1 \ldots k_t (\gamma_{k_1} \ldots \gamma_{k_t})}$ *est mis à zéro si l'un des* $\gamma_{k_1}, \ldots, \gamma_{k_t}$ *est nul, selon la convention de codage « dummy ». Soit* $\boldsymbol{\beta}$ *le vecteur qui comprend* $P(i \in S_A)$ *et les paramètres de* $r^{(\gamma)}$, *qui ne sont pas mis à zéro par cette convention. Alors, la fonction* $\varrho : \boldsymbol{\beta} \mapsto \mathbf{p}$ *est injective.*

C. Procédure d'initialisation

La présente section brosse un tableau de la procédure d'initialisation permettant l'ajustement du modèle de voisinage multivarié dans les simulations. Les paramètres du modèle comprennent les proportions de mélange (les α_g), les paramètres de la distribution des faux positifs (les $\boldsymbol{\lambda}_g$) et ceux de la distribution des vrais positifs (la valeur commune des $\mathbf{p}_g$). Lorsqu'il y a G classes, la proportion de mélange α_g est mise à $1/G$ pour chaque classe. Les autres valeurs initiales sont choisies comme suit.

Pour la distribution des faux positifs, chaque $\boldsymbol{\lambda}_g$ est mis à une valeur commune désignée par $\hat{\boldsymbol{\lambda}} = \left[\hat{\lambda}^{(\gamma)} \right]_{\gamma \in \Gamma}$, où $\Gamma = \{0,1\}^3 \setminus \{(0,0,0)\}$. Pour $\gamma \in \Gamma$, $\hat{\lambda}^{(\gamma)}$ est mis à l'estimation de $\bar{\lambda}$, qui est obtenue par l'ajustement du modèle de voisinage univarié, quand les enregistrements sont liés selon γ. Par exemple, quand $\gamma = (1,0,1)$, cela signifie qu'il faut lier une paire si elle est liée par la première règle de couplage et s'il y a une concordance exacte concernant le nom de famille et le mois de naissance, mais pas de concordance sur le jour de naissance.

À partir de $\hat{\boldsymbol{\lambda}}$, les valeurs initiales sont choisies pour la probabilité de couverture $P(i \in S_A)$ et les paramètres log-linéaires (c'est-à-dire $u_{1(1)}$, $u_{2(1)}$, $u_{3(1)}$, $u_{12(11)}$, $u_{13(11)}$ et $u_{23(11)}$). Ces valeurs s'obtiennent en trois étapes, expliquées ci-dessous. Dans la première étape, on calcule une estimation $\hat{\mathbf{p}} = \left[\hat{p}^{(\gamma)} \right]_{\gamma \in \Gamma}$ du vecteur des probabilités de vrais positifs en maximisant la log-vraisemblance du modèle multivarié avec une seule classe, où $\hat{\boldsymbol{\lambda}}$ remplace $\boldsymbol{\lambda}$, et les probabilités de vrais positifs n'ont pas nécessairement une forme log-linéaire. Cette étape correspond à une optimisation convexe, car la log-vraisemblance du modèle multivarié est concave par rapport aux probabilités de vrais positifs, quand les autres paramètres sont donnés. Dans la deuxième étape, les valeurs initiales pour les paramètres log-linéaires de $r^{(\gamma)}$ (dans (5.2)) sont trouvées par une méthode des moments, basée sur $\hat{\mathbf{p}}$ comme suit. En cas d'ajustement du modèle sans interactions, supposons que $\hat{q}_k = \sum_{\gamma \in \Gamma : \gamma_k = 1} \hat{p}^{(\gamma)}$ (pour $k = 1, 2, 3$) et $\hat{q}_{k_1 k_2} = \sum_{\gamma \in \Gamma : \gamma_{k_1} = \gamma_{k_2} = 1} \hat{p}^{(\gamma)}$ (pour $1 \le k_1 < k_2 \le 3$) et choisissons les valeurs initiales comme étant

$$\hat{u}_{1(1)} = \text{logit}\left(\frac{1}{2}\left(\frac{\hat{q}_{12}}{\hat{q}_2} + \frac{\hat{q}_{13}}{\hat{q}_3}\right)\right),$$

$$\hat{u}_{2(1)} = \text{logit}\left(\frac{1}{2}\left(\frac{\hat{q}_{12}}{\hat{q}_1} + \frac{\hat{q}_{23}}{\hat{q}_3}\right)\right),$$

$$\hat{u}_{3(1)} = \text{logit}\left(\frac{1}{2}\left(\frac{\hat{q}_{13}}{\hat{q}_1} + \frac{\hat{q}_{23}}{\hat{q}_2}\right)\right),$$

avec $\hat{u}_{12(11)} = \hat{u}_{13(11)} = \hat{u}_{23(11)} = 0$. Lorsque les interactions sont incluses, choisissons les valeurs initiales comme suit :

$$\begin{aligned}
\hat{u}_{1(1)} &= \log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right) - \left(\log\left(\frac{\hat{p}^{(1,0,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,0)}}{\hat{p}^{(1,1,1)}}\right)\right), \\
\hat{u}_{2(1)} &= \log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right) - \left(\log\left(\frac{\hat{p}^{(1,0,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,0,1)}}{\hat{p}^{(1,1,1)}}\right)\right), \\
\hat{u}_{3(1)} &= \log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right) - \left(\log\left(\frac{\hat{p}^{(0,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,0,1)}}{\hat{p}^{(1,1,1)}}\right)\right), \\
\hat{u}_{12(11)} &= -\left(\log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right)\right) + \log\left(\frac{\hat{p}^{(1,0,0)}}{\hat{p}^{(1,1,1)}}\right), \\
\hat{u}_{13(11)} &= -\left(\log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right)\right) + \log\left(\frac{\hat{p}^{(0,1,0)}}{\hat{p}^{(1,1,1)}}\right), \\
\hat{u}_{23(11)} &= -\left(\log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right)\right) + \log\left(\frac{\hat{p}^{(0,0,1)}}{\hat{p}^{(1,1,1)}}\right).
\end{aligned}$$

Enfin, supposons que $\hat{\mathbf{r}} = \left[\hat{r}^{(\gamma)}\right]_{\gamma \in \{0,1\}^3}$ est le vecteur de probabilités qui correspond aux valeurs initiales ci-dessus des paramètres log-linéaires (c'est-à-dire $\hat{r}^{(\gamma)}$ est égal au deuxième membre de (5.2), où les valeurs initiales sont obtenues en remplaçant les vraies valeurs des paramètres par les valeurs estimées et choisissons la couverture initiale comme suit :

$$\hat{P}(i \in S_A) = \frac{\sum_{\gamma \in \Gamma} \hat{p}^{(\gamma)}}{\sum_{\gamma \in \Gamma} \hat{r}^{(\gamma)}}.$$

Bibliographie

Belin, T.R., et Rubin, D.B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90, 694-707.

Blakely, T., et Salmond, C. (2002). Probabilistic record linkage and a method to calculate the positive predicted value. *International Journal of Epidemiology*, 31, 1246-1252.

Brown, J., Bycroft, C., Di Cecco, D., Elleouet, J., Powell, G., Račinskij, V., Smith, P.A., Tam, S.-M., Tuoto, T. et Zhang, L.-C. (2020). Exploring developments in population size estimation. *Survey Statistician*, 82, 27-39.

Chambers, R. (2009). Regression analysis of probability-linked data. Dans *Research Series in Official Statistics*. Gouvernement de Nouvelle-Zélande.

Chipperfield, J.O., et Chambers, R.L. (2015). Using the bootstrap to analyse binary data obtained via probabilistic linkage. *Journal of Official Statistics*, 31, 397-414.

Chipperfield, J., Hansen, N. et Rossiter, P. (2018). Estimating precision and recall for deterministic and probabilistic record linkage. *Revue Internationale de Statistique*, 86, 219-236.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection*. New York: Springer.

Daggy, J.K., Xu, H. Hui, S.J., Gamache, R.E. et Grannis, S.J. (2013). A practical approach for incorporating dependence among fields in probabilistic record linkage. *BMC Medical Informatics and Decision Making*, 13, 1-8.

Dasylva, A., Abeysundera, M., Akpoué, B., Haddou, M. et Saïdi, A. (2016). Measuring the quality of a probabilistic linkage through clerical reviews. Recueil : *Symposium 2016, Croissance de l'information statistique : défis et bénéfices*, Statistique Canada.

Dasylva, A., et Goussanou, A. (2020). Estimating linkage errors under regularity conditions. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 687-692.

Dasylva, A., et Goussanou, A. (2021). Estimation des faux négatifs attribuables à la création des pochettes dans le couplage d'enregistrements. *Techniques d'enquête*, 47, 2, 325-338. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2021002/article/00002-fra.pdf.

Dasylva, A., et Goussanou, A. (2022). On the consistent estimation of linkage errors without training data. *Japanese Journal of Statistics and Data Science*, 5, 181-216. DOI: https ://doi.org/10.1007/s42081-022-00153-3.

Dasylva, A., et Goussanou, A. (2024). Making statistical inferences about linkage errors. *Japanese Journal of Statistics and Data Science*, 7, 17-56. DOI: https ://doi.org/10.1007/s42081-023-00228-9.

Dasylva, A., Goussanou, A. et Nambeu, C.O. (2024). *Models of Linkage Error for Capture-Recapture Estimation without Clerical Reviews*. https://arxiv.org/pdf/2403.11438.pdf.

De Wolf, P.-P., van der Laan, J. et Zult, D. (2019). Connection correction methods for linkage error in capture-recapture. *Journal of Official Statitsics*, 35, 577-597.

Di Consiglio, L., et Tuoto, T. (2015). Coverage evaluation on probabilistically linked data. *Journal of Official Statistics*, 31, 415-429.

Ding, Y., et Fienberg, S. (1994). Dual system estimation of census undercount in the presence of matching error. *Journal of the American Statistical Association*, 20, 149-158.

Fellegi, I.P., et Sunter, A.B. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Fortini, M., Liseo, B., Nuccitelli, A. et Scanu, M. (2001). On bayesian record linkage. *Research in Official Statistics*, 4, 185-198.

Haque, S., et Mengersen, K. (2022). Improved assessment of the accuracy of record linkage via an extended macsim approach. *Journal of Official Statistics*, 38, 429-451.

Haque, S., Mengersen, K. et Stern, S. (2021). Assessing the accuracy of record linkages with markov chain based monte carlo simulation approach. *Journal of Big Data*, 8, 1-26.

Herzog, T.N., Scheuren, F.J. et Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques*. New York: Springer.

Lahiri, P., et Larsen, D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 222-227.

Larsen, M., et Rubin, D. (2001). Iterated automated record linkage using mixture models. *Journal of the American Statistical Association*, 96, 32-41.

Lincoln, F. (1930). Calculating waterfowl abundance on the basis of banding returns. *United States Department of Agriculture Circular*, 118, 1-4.

Little, R., et Rubin, D. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

Newcombe, H. (1988). *Handbook of Record Linkage*. New York: Oxford University Press.

Petersen, C. (1896). The yearly immigration of young plaice into the limfjord from the german sea. *Report of the Danish Biological Station*, 6, 5-84.

Račinskij, V., Smith, P.A. et van der Heijden, P. (2019). *Linkage Free Dual System Estimation*. https://arxiv.org/abs/1903.10894, 2019.

Sadinle, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112, 600-612.

Sariyar, M., Borg, A. et Pommerening, K. (2011). Controlling false match rates in record linkage using extreme value theory. *Journal of Biomedical Informatics*, 44, 648-654.

Steorts, R., Hall, R. et Fienberg, S.E. (2016). A bayesian approach to graphical record linkage and de-duplication. *Journal of the American Statistical Association*, 111, 1660-1672.

Tancredi, A., et Liseo, B. (2011). A hierarchical bayesian approach to record linkage and population size problems. *Annals of Applied Statistics*, 5, 1553-1585.

Thibaudeau, Y. (1993). Le pouvoir discriminant des structures de dépendance dans le couplage d'enregistrements. *Techniques d'enquête*, 19, 1, 35-43. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1993001/article/14477-fra.pdf.

US Census Bureau (2016). *File b: Surnames Occurring 100 or More Times*. https://www2.census.gov/topics/genealogy/2010surnames/names.zip. (Consulté : 2020-10-17).

US Census Bureau (2020). *Annual State Resident Population Estimates for 6 Race Groups (5 Race Alone Groups and Two or More Races) by Age, Sex, and Hispanic Origin: April 1, 2010 to July 1, 2019*. https://www2.census.gov/programs-surveys/popest/tables/2010-2019/state/asrh/sc-est2019-alldata6.csv. (Consulté : 2020-10-17).

Winglee, M., Valliant, R. et Scheuren, F. (2005). Une étude de cas en couplage d'enregistrements. *Techniques d'enquête*, 31, 1, 3-13. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2005001/article/8085-fra.pdf.

Winkler, W.E. (1993). Improved decision rules in the fellegi-sunter model of record linkage. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 274-279.

Zhang, L.-C. (2015). On modelling register coverage errors. *Journal of Official Statistics*, 31, 381-396.

Techniques d'enquête, décembre 2024
Vol. 50, nº 2, pp. 443-472
Statistique Canada, nº 12-001-X au catalogue

Examiner les effets du mode d'enquête dans les variances de l'intervieweur grâce à deux enquêtes multimodales représentatives

Wenshan Yu, Michael R. Elliott et Trivellore E. Raghunathan[1]

Résumé

À mesure que les plans de sondage à mode mixte deviennent de plus en plus populaires, leurs effets sur la qualité des données ont attiré l'attention de plusieurs chercheurs. La plupart des études se sont concentrées sur les propriétés de biais des plans de sondage à mode mixte; peu d'entre elles ont cherché à savoir si les plans de sondage à mode mixte ont des structures de variance hétérogènes entre les modes. Bien que de nombreuses caractéristiques des plans de sondage à mode mixte, comme l'utilisation d'intervieweurs variés, les différences systématiques entre les répondants, les différents niveaux de biais dû à la désirabilité sociale, entre autres, peuvent conduire à des variances hétérogènes dans les estimations ponctuelles des moyennes de population propre à un mode, la présente étude permet d'examiner précisément si les variances de l'intervieweur demeurent cohérentes entre les différents modes dans les études multimodales. Pour répondre à cette question de recherche, nous utilisons les données recueillies grâce à deux modèles d'études distincts. Dans le premier modèle, lorsque les intervieweurs sont responsables soit du mode face à face, soit du mode téléphonique, nous examinons s'il y a des différences de mode dans les variances de l'intervieweur pour : 1) les questions politiques de nature délicate; 2) les éléments internationaux; 3) les indicateurs d'éléments manquants sur des éléments internationaux, grâce aux données sur la Jordanie de la vague 6 de l'Arab Barometer. Dans le deuxième modèle, nous nous appuyons sur les données de base de la Health and Retirement Study (HRS) de 2016 pour examiner la question sur trois sujets lorsque les intervieweurs sont responsables des deux modes. Les sujets traités comprennent : 1) l'échelle d'évaluation de la dépression du Center for Epidemiological Studies-Depression (CES-D); 2) les observations des intervieweurs; 3) l'échelle d'activité physique. Pour tenir compte du manque de plans de sondage interpénétrés dans les deux sources de données, nous incluons des covariables au niveau du répondant dans nos modèles. Nous constatons des différences importantes dans les variances de l'intervieweur sur un élément (12 éléments au total) dans l'enquête de l'Arab Barometer, alors que pour la HRS, les résultats sont de 3 sur 18. Dans l'ensemble, nous constatons que l'ampleur des variances de l'intervieweur est plus grande en personne que par téléphone pour les éléments de nature délicate. Nous effectuons des simulations pour comprendre le pouvoir de détecter les effets du mode d'enquête dans les tailles d'échantillons typiquement modestes de l'intervieweur.

Mots-clés : Effets de l'intervieweur; effets du mode d'enquête; étude multimodale; plan de sondage à mode mixte.

1. Introduction

Les intervieweurs jouent un rôle central dans la collecte des données d'enquête. Selon le mode et le plan d'échantillonnage de la collecte de données, il peut être nécessaire qu'ils répertorient des adresses pour générer des bases de sondage, qu'ils recrutent des répondants, qu'ils posent des questions d'enquête et qu'ils consignent les réponses des participants. Par conséquent, à partir du cadre de l'erreur d'enquête totale, les intervieweurs peuvent avoir une incidence sur la qualité des données d'enquête en générant ou en réduisant les erreurs de couverture, les erreurs de non-réponse, les erreurs de mesure et les erreurs de traitement (West et Blom, 2017). La plupart des recherches portant sur les effets des intervieweurs portent sur les erreurs de mesure (Schuman et Converse, 1971; Hanson et Marks, 1958; Ehrlich et Riesman, 1961), qui peuvent être

1. Wenshan Yu, Survey Research Center de l'Institute for Social Research, Université du Michigan. Courriel : yuwens@umich.edu; M. Michael R. Elliott, Survey Research Center de l'Institute for Social Research et département de biostatistique, Université du Michigan. Courriel : mrelliot@umich.edu; M. Trivellore E. Raghunathan, Survey Research Center de l'Institute for Social Research et département de biostatistique, Université du Michigan. Courriel : teraghu@umich.edu.

encore décomposées en une partie systématique, les biais attribuables aux intervieweurs (lorsque les répondants modifient les réponses soit en raison de la présence des intervieweurs ou de leurs caractéristiques observables), et une composante aléatoire, la variance de l'intervieweur. Cette variance de l'intervieweur augmente l'incertitude des estimations, parfois dans une bien plus grande mesure que la corrélation induite par le regroupement géographique (Schnell et Kreuter, 2003). La présente étude vise à déterminer l'effet des différents modes de collecte de données – plus précisément par téléphone par rapport à en personne – sur les variances de l'intervieweur dans les enquêtes multimodales.

Les variances de l'intervieweur ont été étudiées pour la première fois dans le cadre d'interviews sur place (Kish, 1962). Lorsque les enquêtes téléphoniques sont devenues une solution de rechange aux interviews sur place, les chercheurs ont évalué les variances de l'intervieweur dans les enquêtes téléphoniques et ont généralement constaté qu'elles étaient moins importantes que celles des enquêtes menées sur place (Groves et Magilavy, 1986; Tucker, 1983; Groves et Kahn, 1979). Plus précisément, la corrélation intraclasse ρ_{int}, une mesure courante utilisée pour évaluer les effets des intervieweurs et définie par le rapport des variances de l'intervieweur à la variance totale, variait de 0,005 à 0,102 dans les enquêtes en personne, alors que celles calculées dans les enquêtes téléphoniques centralisées variaient de 0,0018 à 0,0184 (Groves et Magilavy, 1986; Groves et Kahn, 1979). La constatation correspond aux attentes théoriques, puisque les intervieweurs dans l'environnement des enquêtes téléphoniques centralisées sont plus étroitement surveillés et supervisés que les intervieweurs sur le terrain (Schaeffer, Dykema et Maynard, 2010). Depuis lors, le domaine de la recherche a reçu peu d'attention scientifique. Cependant, à mesure que les plans de sondage à mode mixte deviennent de plus en plus utilisés, l'objet à l'étude nécessite davantage de recherches. Il y a un manque de preuve directe, car les résultats antérieurs reposent principalement sur des enquêtes différentes qui utilisent un mode, soit en personne ou par téléphone. En outre, les enquêtes multimodales offrent naturellement une occasion où le contexte de l'enquête et les questionnaires sont très comparables (sinon les mêmes) lorsqu'on compare les variances de l'intervieweur dans les deux modes. De plus, selon que les intervieweurs sont responsables des deux modes dans les enquêtes multimodales, ils seraient susceptibles d'exercer leur influence d'un mode à l'autre. Ces facteurs peuvent mener à des résultats différents lorsqu'on compare les variances de l'intervieweur entre les modes.

L'examen des effets du mode d'enquête dans les variances de l'intervieweur est également utile pour faciliter la conception des plans de sondage à mode mixte et servir d'indicateur de la qualité des données. Tout d'abord, la quantification de la variance de l'intervieweur en fonction du mode peut aider les chercheurs à déterminer et à choisir le mode ayant une faible variance de l'intervieweur dans un plan de sondage à mode mixte. La stratégie d'inférence multimodale de pointe actuelle met l'accent sur la propriété biaisée des modes (Elliott, Zaslavsky, Goldstein, Lehrman, Hambarsoomians, Beckett et Giordano, 2009; Kolenikov et Kennedy, 2014), mais peu a été fait pour intégrer la possible structure de variance hétérogène (Suzer-Gurtekin, Heeringa et Valliant, 2013). Une partie de la raison est que peu de littérature met en

lumière les propriétés de variance des plans de sondage à mode mixte (Vannieuwenhuyze, 2015), en particulier ce qui entre dans les variances. Deuxièmement, la détermination des questions associées aux importants effets du mode d'enquête de la variance de l'intervieweur peut renseigner sur la façon dont la variance de l'intervieweur est générée et peut donc être réduite. Par exemple, les chercheurs montrent que les questions relatives au comportement, de nature délicate, ambiguës, complexes et ouvertes sont généralement plus vulnérables aux effets de l'intervieweur (Schaeffer, Dykema et Maynard, 2010), car ces questions offrent plus de possibilités à l'intervieweur d'aider les répondants (West et Blom, 2017). Si les questions de nature délicate présentent un grand effet de l'intervieweur pour les enquêtes en personne, mais pas pour les enquêtes téléphoniques, cela peut laisser entendre que les questions constituent un fardeau pour les intervieweurs sur le terrain. À cette fin, les organismes d'enquête peuvent offrir une formation supplémentaire pour normaliser la façon de poser les questions ou d'utiliser d'autres approches [comme l'auto-interview assistée par ordinateur avec interface audio [audio-AIAO] ou la technique du dénombrement d'items (Holbrook et Krosnick, 2010)] pour recueillir des renseignements sur les éléments de nature délicate. Troisièmement, dans les plans de sondage à mode mixte où les intervieweurs sont responsables des deux modes, nous pouvons potentiellement trouver des intervieweurs qui ont un effet important sur les réponses dans les deux modes ou seulement dans un mode, qui fournissent la base de l'intervention en temps réel et de la formation des intervieweurs à un niveau plus granulaire.

Dans le présent document, nous examinons deux études multimodales représentatives : 1) la vague 6 de l'expérience en Jordanie dans l'enquête de l'Arab Barometer; 2) la Health and Retirement Study (HRS) de 2016. En nous appuyant sur les deux sources de données, nous examinons les effets du mode d'enquête dans les variances de l'intervieweur pour des intervieweurs dans différents pays, pour différentes populations cibles et pour une variété de variables de résultats. De plus, l'utilisation des deux études offre des perspectives distinctes pour l'examen de notre question de recherche. Le plan pour l'intervieweur de l'enquête de l'Arab Barometer est couramment utilisé dans les enquêtes où différents modes sont gérés par différents organismes de collecte de données, ce qui se traduit par des intervieweurs différents selon les modes. D'autre part, le plan pour l'intervieweur de la HRS, où les mêmes intervieweurs sont utilisés dans les deux modes, favorise une estimation plus précise des différences dans les variances de l'intervieweur uniquement en raison des modes, en éliminant la partie des variances de l'intervieweur qui résultent de l'utilisation de différents intervieweurs entre les divers modes.

Le reste du présent document est organisé comme suit. Dans la section 2, nous décrivons le plan de l'étude et la stratégie analytique, et présentons les résultats en utilisant notre première source de données – l'enquête de l'Arab Barometer. La section 3 présente la deuxième source de données – la HRS, ainsi que l'approche analytique correspondante et les résultats liés à la variance de l'intervieweur associée aux données de la HRS. Dans la section 4, nous menons une étude de simulation pour illustrer la puissance de détection des effets du mode d'enquête dans les variances de l'intervieweur en utilisant à la fois la

configuration de l'enquête de l'Arab Barometer et de la HRS. Enfin, à la section 5, nous discutons des répercussions de notre étude.

2. L'enquête de l'Arab Barometer

2.1 Description de l'enquête

L'Enquête de l'Arab Barometer est le plus grand entrepôt de données d'opinion publique dans la région du Moyen-Orient et de l'Afrique du Nord (MENA). Lors de la vague 6, une expérience sur le mode d'enquête en Jordanie a été intégrée entre mars et avril 2021, où les participants ont été répartis au hasard, soit pour participer à une enquête en personne ou à une enquête téléphonique de relance. Le Center for Strategic Studies de Jordanie a effectué les travaux sur le terrain en utilisant le recensement de la population et du logement de 2015 comme base de sondage. Ils ont mis en place un échantillon probabiliste de zone stratifié en fonction des gouvernorats et du clivage entre les régions urbaines et rurales. Des intervieweurs distincts ont été utilisés dans les interviews sur place et par téléphone. Les ménages sélectionnés pour l'interview par téléphone ont d'abord été mobilisés par l'entremise de l'interview en personne pour une courte enquête de 5 minutes, et la majorité des questions de l'enquête ont été posées environ une semaine plus tard lors d'un suivi téléphonique. Dans le mode d'interview en personne, 31 intervieweurs ont recueilli des données auprès de 1 193 répondants, tandis que 13 intervieweurs ont interviewé 1 212 participants par téléphone.

Nous nous concentrons sur trois types de variables de résultats (Y) : 1) questions d'ordre politique de nature délicate (6 éléments); 2) questions internationales de nature moins délicate (3 éléments); 3) question de savoir si les répondants ne connaissent pas ou refusent de répondre aux questions de relations internationales (3 éléments). À l'exception des indicateurs d'éléments manquants, les autres variables de résultat ont été initialement mesurées par quatre catégories ordinales; nous les avons agrégées en résultats binaires en fixant le point limite au milieu. Les catégories d'origine et les catégories agrégées sont disponibles à l'annexe A de (Yu, Elliott et Raghunathan, 2024).

Les variables de résultat (Y) peuvent être assujetties à deux types d'effets du mode d'enquête : 1) les effets du mode d'enquête qui entraînent un changement dans les moyennes des variables de résultat (appelés effets du mode d'enquête dans les moyennes); 2) les effets du mode d'enquête dans les variances de l'intervieweur. Au total, nous considérons que q intervieweurs ne recueillent des renseignements que dans l'un des deux modes (en personne et par téléphone) de n unités d'échantillonnage tirées d'une population finie. Les intervieweurs recueillent également des covariables au niveau des répondants (X) qui prédisent les variables de résultats (Y). Les covariables (X) sont censées être indépendantes de tous les effets du mode d'enquête. Nous considérons les covariables (X) qui comprennent l'âge, le genre, l'état matrimonial, la taille du ménage des répondants et les régions dans le présent document.

2.2 Stratégie analytique

Tout d'abord, pour illustrer les statistiques descriptives de la variation de l'intervieweur dans les réponses recueillies, nous calculons l'écart-type entre les intervieweurs et l'écart-type moyen de l'intervieweur lui-même. Plus précisément, nous calculons les proportions moyennes pour chaque variable et chaque intervieweur ($\overline{y}_{(m)j}$). Dans la configuration de l'enquête de l'Arab Barometer, où les intervieweurs sont imbriqués dans chaque mode, ces statistiques sont intrinsèquement propres au mode; par conséquent, nous écrivons m entre parenthèses pour souligner ce point. Nous calculons ensuite l'écart-type de ces proportions moyennes entre les intervieweurs, appelé l'écart-type entre les intervieweurs. L'écart-type de l'intervieweur lui-même (v_j^m) est tiré des réponses recueillies par chaque intervieweur. L'écart-type moyen de l'intervieweur lui-même est calculé comme la moyenne des écart-type de l'intervieweur lui-même pour tous les intervieweurs pour chaque mode. Nous présentons la formule pour calculer les statistiques pertinentes en (2.1), où i représente l'indice pour les répondants, j représente l'indice pour les intervieweurs, m représente l'indice pour les modes, $n_{(m)j}$ reflète le nombre d'interviews réalisées par l'intervieweur j en utilisant le mode m, n_m représente le nombre de répondants en mode m, n_j^m indique le nombre d'intervieweurs utilisant le mode m, et $y_{i(m)j}$ indique les réponses fournies par le répondant i interviewé par l'intervieweur j en utilisant le mode m. Pour les organismes de collecte de données d'enquête, un petit écart-type entre les intervieweurs et un grand écart-type moyen de l'intervieweur lui-même sont souhaitables, car cela peut indiquer une affectation d'intervieweur proche d'être aléatoire et des effets minimes des intervieweurs sur les réponses collectées. Nous présentons les statistiques pour les covariables et les résultats d'intérêt. Les statistiques des covariables peuvent indiquer des effets dus à la sélection des intervieweurs, ce qui souligne l'importance de considérer les covariables dans le modèle analytique final. Les statistiques relatives aux variables de résultat peuvent fournir une preuve initiale de la présence des effets de l'intervieweur et justifier une enquête plus approfondie.

$$
\begin{aligned}
\text{Proportion moyenne par intervieweur } \overline{y}_{(m)j} &= \frac{\sum_i^{n_{(m)j}} y_{i(m)j}}{n_{(m)j}} \\
\text{Proportion moyenne par mode } \overline{y}_m &= \frac{\sum_i^{n_m} y_{i(m)j}}{n_m} \\
\text{Écart-type entre les intervieweurs} &= \sqrt{\frac{\sum_j^{n_j^m} (\overline{y}_{(m)j} - \overline{y}_m)^2}{n_j^m}} \\
\text{Écart-type de l'intervieweur lui-même } v_j^m &= \sqrt{\frac{\sum_i^{n_{(m)j}} (\overline{y}_{i(m)j} - \overline{y}_{(m)j})^2}{n_{(m)j}}} \\
\text{Écart-type moyen de l'intervieweur lui-même} &= \frac{\sum_j^{n_j^m} v_j^m}{n_j^m}.
\end{aligned}
\tag{2.1}
$$

Pour tester si les variances de l'intervieweur sont égales entre les modes, puisque toutes les variables de résultat sont binaires, nous ajustons le modèle probit suivant à chacune des variables, où m représente l'indice pour les modes (f pour « en personne » et t pour « par téléphone »), M et $J_{j,\ j=1,\ldots,q-1}$ sont des variables nominales (longueur de n) pour indiquer les modes ($M=1$ pour le mode en personne et $M=0$ pour le mode par téléphone) et les intervieweurs :

$$\begin{aligned}
Y^*_{ij(m)} &= \beta_0 + \beta_1 M_i + b_{j(m)} + \epsilon_{ij(m)},\\
Y_{ij(m)} &= 1 \text{ si } Y^*_{ij(m)} > 0 \text{ et } Y_{ij(m)} = 0 \text{ si } Y^*_{ij(m)} \le 0,\\
b_{j(m)} &\sim N(0, \sigma^2_m),\\
\epsilon_{ij(m)} &\sim N(0,1),\\
\sigma_f, \sigma_t &\sim \text{ moitié } - T(3,1) \text{ (pour la modélisation bayésienne)},\\
\boldsymbol{\gamma}, \beta_0, \beta_1 &\sim N(0, 10^6) \text{ (pour la modélisation bayésienne)}.
\end{aligned} \tag{2.2}$$

Dans le modèle (2.2), les effets aléatoires de l'intervieweur sont représentés comme $b_{j(m)}$ puisque les intervieweurs sont imbriqués dans les modes. Notre question de recherche, « Les variances de l'intervieweur sont-elles égales entre les modes selon un plan de sondage à mode mixte randomisé ? » est traitée en évaluant si $\alpha = \log(\sigma_f) - \log(\sigma_t)$ est égal à zéro pour chaque variable du modèle (2.2). Pour déterminer cela, nous évaluons si les intervalles de confiance de 95 % ou les intervalles de plus haute densité *a posteriori* crédibles de α comprennent zéro. Si les intervalles ne comprennent pas zéro pour certaines variables, cela semble indiquer que les variances de l'intervieweur ne sont pas égales entre les modes pour ces variables.

En ajustant (2.2), nous pouvons également obtenir des estimations des effets du mode d'enquête (β_1) pour chaque variable en calculant et en testant si la quantité diffère de 0. Il convient de mentionner que les estimations peuvent inclure certains effets de sélection de mode; malgré l'attribution aléatoire du mode, la non-réponse différentielle peut se produire entre les modes (West, Kreuter et Jaenichen, 2013).

Supposons que d'après des données probantes $\alpha \neq 0$, nous envisageons alors que la variance de l'intervieweur propre à un mode est fallacieuse en raison de l'absence de plans de sondage interpénétrés en ajoutant des covariables au niveau du répondant (x_{si}, où s désigne la covariable s) au modèle (2.2) :

$$\begin{aligned}
Y^*_{ij(m)} &= \beta_0 + \beta_1 M_i + b_{j(m)} + \sum_s^S \gamma_s x_{si} + \epsilon_{ij(m)},\\
Y_{ij(m)} &= 1 \text{ si } Y^*_{ij(m)} > 0 \text{ et } Y_{ij(m)} = 0 \text{ si } Y^*_{ij(m)} \le 0,\\
b_{j(m)} &\sim N(0, \sigma^2_m),\\
\epsilon_{ij(m)} &\sim N(0,1),\\
\sigma_f, \sigma_t &\sim \text{ moitié } - T(3,1) \text{ (pour la modélisation bayésienne)},\\
\boldsymbol{\gamma}, \beta_0, \beta_1 &\sim N(0, 10^6) \text{ (pour la modélisation bayésienne)}.
\end{aligned} \tag{2.3}$$

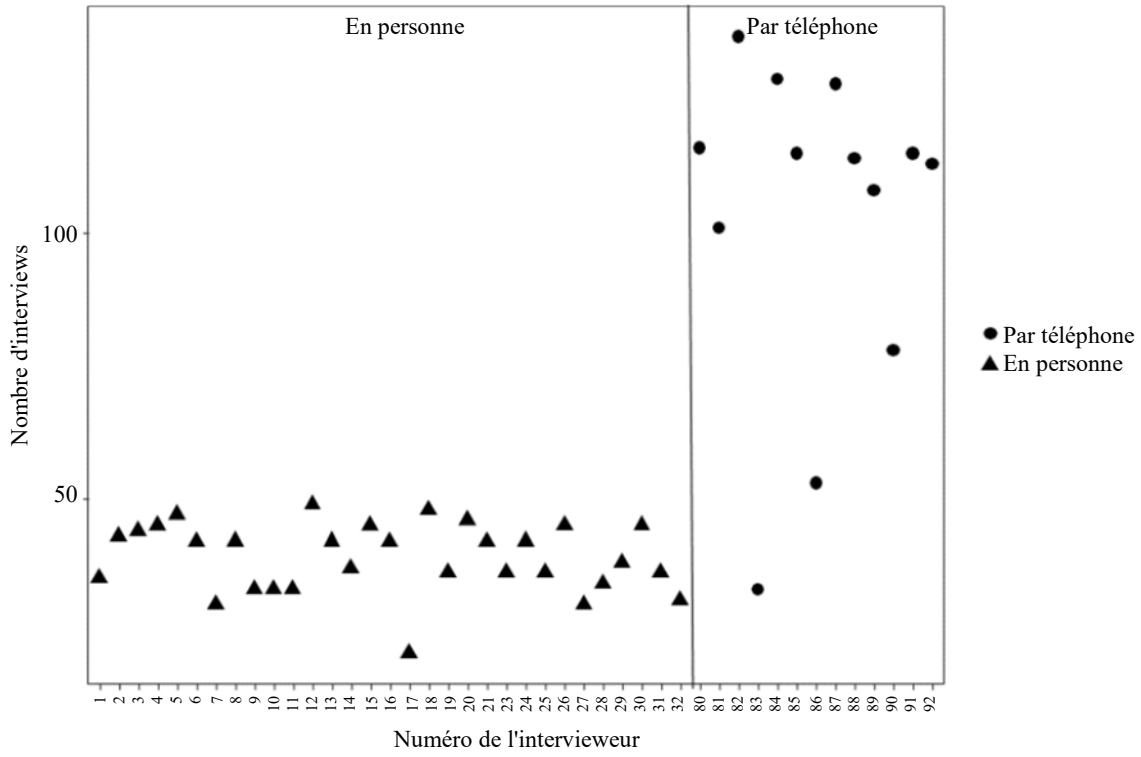
Nous mettons en œuvre les modèles en utilisant à la fois les approches de vraisemblance (Proc Nlmixed) et bayésienne (Proc MCMC) dans le langage de programmation SAS. Dans l'approche de vraisemblance, nous effectuons la transformation logarithmique sur σ_f^2 et σ_t^2 pour stabiliser la variance des paramètres et améliorer la propriété de couverture. Nous calculons la variance de la variable estimée α en utilisant la méthode delta, donnée par $\text{var}(\alpha) = \frac{1}{4}\text{var}(\log(\sigma_f^2)) + \frac{1}{4}\text{var}(\log(\sigma_t^2))$ (voir les dérivations à l'annexe A), puis nous utilisons une distribution normale pour estimer l'intervalle de confiance de 95 %. Dans l'approche bayésienne, nous utilisons une chaîne avec 200 000 à 300 000 tirages, selon l'autocorrélation et la taille effective de l'échantillon, et sélectionnons chaque 100^e valeur comme le taux d'amincissement. Pour faciliter l'illustration, nous ne présentons que les résultats du modèle avec des covariables ajoutées et estimées en utilisant la modélisation bayésienne (2.3) dans la section suivante.

2.3 Résultats

2.3.1 Statistiques descriptives

Nous supposons que les intervieweurs sont interchangeables dans le présent document. Pour évaluer en partie cette hypothèse, nous présentons les charges de travail de l'intervieweur dans les modes en personne et par téléphone dans l'étude de l'Arab Barometer à la figure 2.1. Comme l'indique la figure 2.1, nous constatons que dans le mode en personne, chaque intervieweur effectue un nombre semblable d'interviews. En revanche, la moyenne et la variation du nombre d'interviews par intervieweur sont plus grandes dans le mode par téléphone.

Figure 2.1 Charges de travail de l'intervieweur selon le mode dans l'étude de l'Arab Barometer.



Nous déclarons les moyennes d'échantillon propres au mode, les écarts-types entre les intervieweurs et les écarts-types moyens de l'intervieweur lui-même dans le tableau 2.1. D'après le tableau 2.1, nous observons d'une part que pour les questions politiques de nature délicate, les proportions moyennes déclarées par téléphone sont généralement plus élevées que celles rapportées lors d'entrevues en personne, ce qui donne à penser que l'interview par téléphone peut être associée à des déclarations plus positives. D'autre part, les écarts-types entre intervieweurs pour les interviews en personne sont généralement plus importants que ceux pour les interviews par téléphone pour la plupart des résultats, tandis que les écarts-types moyens de l'intervieweur lui-même sont plus grands par téléphone qu'en personne pour les questions politiques de nature délicate et les indicateurs manquants. Cela fournit des preuves initiales que les intervieweurs semblent avoir un effet plus grand en personne que par téléphone. Nous fournissons la répartition des variables de résultats par intervieweur à l'annexe C de (Yu, Elliott et Raghunathan, 2024).

Tableau 2.1
Répartition des variables de résultats dans l'étude de l'Arab Barometer entre les intervieweurs selon les modes.

Questions	Moyenne (EP)	Moyenne (TEL)	Écart-type entre les intervieweurs (EP)	Écart-type entre les intervieweurs (TEL)	Écart-type moyen de l'intervieweur lui-même (EP)	Écart-type moyen de l'intervieweur lui-même (TEL)
Questions politiques de nature délicate						
1. Liberté des médias	0,403	0,588	0,191	0,117	0,455	0,480
2. Confiance envers le gouvernement	0,356	0,533	0,165	0,122	0,455	0,487
3. Confiance envers les tribunaux	0,594	0,770	0,139	0,123	0,477	0,398
4. Satisfait des soins de santé	0,491	0,592	0,155	0,071	0,482	0,489
5. Rendement en matière d'inflation	0,140	0,243	0,146	0,142	0,291	0,406
6. Rendement pendant la COVID-19	0,402	0,576	0,171	0,161	0,464	0,470
Questions internationales						
7. Attitude positive à l'égard des États-Unis	0,394	0,415	0,187	0,189	0,467	0,459
8. Attitude positive à l'égard de l'Allemagne	0,488	0,560	0,224	0,186	0,464	0,464
9. Attitude positive à l'égard de la Chine	0,468	0,507	0,207	0,203	0,470	0,463
Réponses manquantes sur les questions internationales (construites)						
10. Réponse manquante sur l'attitude positive à l'égard des États-Unis	0,253	0,297	0,235	0,158	0,341	0,425
11. Réponse manquante sur l'attitude positive à l'égard de l'Allemagne	0,320	0,381	0,247	0,199	0,384	0,442
12. Réponse manquante sur l'attitude positive à l'égard de la Chine	0,283	0,329	0,252	0,180	0,359	0,431

Note EP = en personne; TEL = par téléphone.

Nous présentons les caractéristiques non pondérées des échantillons dans les modes en personne et par téléphone au tableau 2.2. Dans le cadre du plan de sondage à mode mixte randomisé, l'échantillon de la

Jordanie est à peu près équilibré en fonction des variables démographiques et socioéconomiques clés (âge, genre, éducation, état matrimonial, taille du ménage et région) entre les modes. Toutefois, il y a un peu plus de répondants de sexe masculin (0,55 contre 0,50) dans le mode par téléphone par rapport au mode en personne, peut-être en raison de la non-réponse différentielle. Nous constatons que pour ces covariables, l'écart-type entre intervieweurs pour l'interview en personne est généralement beaucoup plus grand que celui pour l'interview par téléphone, ce qui semble indiquer des effets de sélection potentiellement plus importants pour l'interview en personne, puisque nous supposons que les covariables ne sont pas susceptibles à l'erreur de mesure.

Tableau 2.2
Répartition des caractéristiques de l'échantillon de l'étude de l'Arab Barometer entre les intervieweurs selon le mode.

Variables du répondant	Moyenne (EP)	Moyenne (TEL)	Écart-type entre les intervieweurs (EP)	Écart-type entre les intervieweurs (TEL)	Écart-type moyen de l'intervieweur lui-même (EP)	Écart-type moyen de l'intervieweur lui-même (TEL)
18 à 24 ans	0,166	0,164	0,085	0,039	0,361	0,369
25 à 34 ans	0,226	0,203	0,072	0,038	0,415	0,402
35 à 44 ans	0,227	0,215	0,088	0,052	0,412	0,408
45 à 54 ans	0,199	0,219	0,069	0,031	0,394	0,414
55 ans et plus	0,183	0,198	0,071	0,032	0,381	0,399
Masculin	0,497	0,549	0,291	0,041	0,369	0,499
Pas de diplôme d'études secondaires	0,345	0,337	0,125	0,106	0,463	0,463
Diplôme d'études secondaires	0,365	0,357	0,098	0,082	0,477	0,474
Diplôme d'études postsecondaires	0,290	0,307	0,101	0,051	0,445	0,461
Non marié(e)s	0,238	0,264	0,106	0,063	0,412	0,438
Marié(e)s	0,693	0,684	0,082	0,062	0,459	0,463
Divorcé(e)s, veufs et veuves, séparé(e)s	0,069	0,053	0,044	0,024	0,230	0,219
Taille du ménage : Moins de 3	0,208	0,222	0,083	0,040	0,399	0,416
Taille du ménage : 4 à 5	0,345	0,349	0,081	0,061	0,475	0,475
Taille du ménage : 6 à 7	0,281	0,288	0,079	0,072	0,447	0,449
Taille du ménage : 8 et plus	0,165	0,141	0,089	0,056	0,353	0,330
Région : Central	0,523	0,509	0,154	0,255	0,482	0,429
Région : Nord	0,261	0,282	0,101	0,188	0,424	0,388
Région : Sud	0,216	0,209	0,175	0,119	0,333	0,367

Note EP = en personne; TEL = par téléphone.

2.3.2 Effets du mode d'enquête sur les moyennes et les variances de l'intervieweur

La présente section montre les résultats de modélisation qui incorporent le modèle d'information sur les répondants (2.3) à l'aide de l'estimation bayésienne au tableau 2.3. En ce qui concerne les effets du mode d'enquête dans les moyennes, nous observons des estimations négatives pour tous les éléments de nature délicate. Par exemple, la probabilité qu'un participant masculin célibataire âgé de 18 à 24 ans, ayant un

diplôme d'études postsecondaires, vivant dans un foyer de moins de trois personnes et résidant dans la région nord de la Jordanie, signalant que la liberté des médias est garantie dans une large ou moyenne mesure, diminue de 17,9 % si la méthode d'interviews en personne est utilisée comparativement aux interviews par téléphone. Ce pourcentage de 17,9 % est calculé à l'aide de $\phi(\beta_0 + \beta_1 + \sum_s^S \gamma_s x_{si})\,\beta_1$, où ϕ est la fonction de densité de probabilité d'une distribution normale standard et S est le nombre de covariables (x). Les estimations de γ_s ne sont pas indiquées dans le présent document, mais peuvent être fournies sur demande. Les effets négatifs du mode d'enquête dans les moyennes donnent à penser que les répondants ont exprimé de moins bonnes opinions sur le gouvernement lorsqu'elles ont répondu à des interviews en personne, ce qui pourrait représenter des réponses plus honnêtes compte tenu du régime autoritaire de la Jordanie. Le tableau 2.3 indique aussi que les taux manquants pour les questions internationales sont plus faibles dans les interviews en personne que dans les interviews par téléphone (bien que cela ne soit pas statistiquement significatif au seuil de 0,05). Nous n'avons pas intégré les poids de l'échantillon dans l'analyse, car notre objectif d'inférence est l'échantillonnage répété dans le même plan d'enquête.

Ensuite, nous tournons notre attention vers les variances de l'intervieweur. Premièrement, l'ampleur des variances de l'intervieweur est généralement importante dans l'étude de l'Arab Barometer. Pour les questions politiques de nature délicate, les variances de l'intervieweur varient de 0,03 à 0,393 (tableau 2.3). La documentation précédente sur les effets de l'intervieweur signalait habituellement une corrélation intraclasse de l'intervieweur (ρ_{int}) pour refléter la proportion de variances des intervieweurs. Pour calculer la variable propre au mode $\rho_{m,\text{int}}$, nous pouvons utiliser la formule $\rho_{m,\text{int}} = \frac{\text{var}_{m,\text{int}}}{1+\text{var}_{m,\text{int}}}$, puisque la variance résiduelle dans le modèle probit est 1. Par conséquent, les résultats mentionnés précédemment correspondent à ρ_{int} allant de 0,029 à 0,282. À titre de référence, d'après la littérature, une valeur ρ_{int} inférieure à 0,01 est considérée comme faible, alors qu'une valeur supérieure à 0,12 est considérée comme importante (West et Olson, 2010). Dans le tableau 2.3, nous observons que $\rho_{f,\text{int}}$ et $\rho_{t,\text{int}}$ peuvent varier considérablement pour le même résultat. Par exemple, pour la satisfaction à l'égard des soins de santé, $\rho_{f,\text{int}}$ est de 0,125, alors que $\rho_{t,\text{int}}$ est de 0,029. Il est important de tenir compte de ces différences lorsqu'on utilise les valeurs $\rho_{m,\text{int}}$ pour calculer les tailles d'échantillon effectives associées à un mode de collecte de données précis.

Pour un élément de nature délicate, le rendement dans le système de soins de santé, nous observons une différence marginalement importante dans les variances de l'intervieweur dans le tableau 2.3 en utilisant l'estimation bayésienne. Les résultats sont importants lorsqu'on utilise l'estimation par la méthode du maximum de vraisemblance, comme le montre l'annexe D de (Yu, Elliott et Raghunathan, 2024). Dans cet élément, les estimations des variances de l'intervieweur sont considérablement plus grandes dans le mode de l'interview en personne. Pour 5 éléments de nature délicate sur 6, les variances de l'intervieweur pour l'interview en personne sont un peu plus importantes que les variances de l'intervieweur pour l'interview par téléphone. Les différences ne sont pas statistiquement significatives, peut-être en raison de la puissance

limitée déterminée par le petit nombre d'intervieweurs dans la présente étude. Les plus grandes variances de l'intervieweur pour l'interview en personne correspondent aux attentes théoriques, car les intervieweurs peuvent démontrer une plus grande hétérogénéité dans l'administration des questions de nature délicate et établir des rapports avec les répondants au cours des interviews en personne.

Tableau 2.3
Variances de l'intervieweur par mode pour certains éléments de l'étude de l'Arab Barometer en tenant compte des covariables à l'aide de l'estimation bayésienne.

Questions	σ_f^2	σ_t^2	$\rho_{f,\text{int}}$	$\rho_{t,\text{int}}$	α	β_1
Questions politiques de nature délicate						
1. Liberté des médias	0,252	0,135	0,201	0,119	0,355	**-0,526**
	[0,122; 0,428]	[0,036; 0,284]	[0,109; 0,3]	[0,035; 0,221]	[-0,223; 0,898]	**[-0,795; -0,222]**
2. Confiance envers le gouvernement	0,188	0,127	0,158	0,113	0,239	**-0,504**
	[0,083; 0,322]	[0,029; 0,275]	[0,077; 0,244]	[0,028; 0,216]	[-0,382; 0,838]	**[-0,768; -0,238]**
3. Confiance envers les tribunaux	0,113	0,214	0,102	0,176	-0,29	**-0,555**
	[0,038; 0,201]	[0,05; 0,445]	[0,037; 0,167]	[0,048; 0,308]	[-0,94; 0,318]	**[-0,881; -0,273]**
4. Satisfait des soins de santé	0,143	0,03	0,125	0,029	0,906	**-0,278**
	[0,051; 0,251]	[0; 0,075]	[0,049; 0,201]	[0; 0,07]	[-0,054; 1,758]	**[-0,475; -0,085]**
5. Rendement en matière d'inflation	0,393	0,204	0,282	0,169	0,361	**-0,523**
	[0,153; 0,672]	[0,051; 0,435]	[0,133; 0,402]	[0,049; 0,303]	[-0,275; 0,927]	**[-0,861; -0,153]**
6. Rendement pendant la COVID-19	0,202	0,224	0,168	0,183	-0,026	**-0,51**
	[0,084; 0,34]	[0,07; 0,443]	[0,077; 0,254]	[0,065; 0,307]	[-0,602; 0,508]	**[-0,841; -0,205]**
Questions internationales						
7. Attitude positive à l'égard des États-Unis	0,198	0,362	0,165	0,266	-0,278	-0,057
	[0,074; 0,34]	[0,104; 0,719]	[0,069; 0,254]	[0,094; 0,418]	[-0,841; 0,282]	[-0,45; 0,318]
8. Attitude positive à l'égard de l'Allemagne	0,292	0,33	0,226	0,248	-0,037	-0,147
	[0,12; 0,514]	[0,092; 0,663]	[0,107; 0,339]	[0,084; 0,399]	[-0,603; 0,548]	[-0,551; 0,236]
9. Attitude positive à l'égard de la Chine	0,205	0,378	0,17	0,274	-0,282	-0,15
	[0,083; 0,361]	[0,116; 0,787]	[0,077; 0,265]	[0,104; 0,44]	[-0,869; 0,245]	[-0,549; 0,19]
Réponses manquantes sur les questions internationales (construites)						
10. Réponse manquante sur l'attitude positive à l'égard des États-Unis	0,995	0,343	0,499	0,255	**0,557**	-0,298
	[0,48; 1,71]	[0,104; 0,668]	[0,324; 0,631]	[0,094; 0,4]	**[0,014; 1,121]**	[-0,805; 0,172]
11. Réponse manquante sur l'attitude positive à l'égard de l'Allemagne	0,844	0,464	0,458	0,317	0,324	-0,287 (0,24)
	[0,404; 1,324]	[0,16; 0,857]	[0,288; 0,57]	[0,138; 0,461]	[-0,169; 0,839]	[-0,765; 0,149]
12. Réponse manquante sur l'attitude positive à l'égard de la Chine	0,936	0,452	0,483	0,311	0,398	-0,244
	[0,434; 1,552]	[0,118; 0,933]	[0,303; 0,608]	[0,106; 0,483]	[-0,134; 0,949]	[-0,73; 0,229]

Notes : Les résultats importants sont indiqués en gras. β_1 désigne les estimations de l'effet du mode d'enquête dans les moyennes. α_1 désigne les estimations de l'effet du mode d'enquête dans les variances de l'intervieweur. σ_f^2 est la variance de l'intervieweur pour l'interview en personne. σ_t^2 est la variance de l'intervieweur pour l'interview par téléphone. $\rho_{f,\text{int}}$ et $\rho_{t,\text{int}}$ représentent la corrélation intraclasse de l'intervieweur dans les modes d'interview en personne et d'interview par téléphone, respectivement.

De manière contre-intuitive, pour ce qui est des réponses substantielles aux questions d'attitude internationales qui ne sont pas de nature délicate (éléments 7 à 9), les estimations de la variance de l'intervieweur sont généralement plus grandes pour le mode par téléphone que pour le mode en personne (mais pas de

façon importante). Les variances de l'intervieweur liées au fait de déclarer « ne sait pas » ou « refuse de répondre » aux questions internationales qui ne sont pas de nature délicate sont plus importantes en mode en personne qu'en mode par téléphone (important pour le premier point). Cette constatation peut s'expliquer par le fait que les intervieweurs assignés au mode en personne ont tenté de persuader les répondants de donner des réponses substantielles, et que la persuasion se produise ou réussisse peut varier selon les intervieweurs.

3. Health and retirement study de 2016

3.1 Description de l'étude

La HRS est une étude longitudinale par panel qui vise les personnes de plus de 50 ans (et de leurs conjoints ou conjointes) aux États-Unis. Elle est menée tous les deux ans, a commencé en 1992, et a permis d'étudier plus de 43 000 personnes (Fisher et Ryan, 2018). La HRS est parrainée par le National Institute on Aging (numéro de subvention NIA U01AG009740) et est dirigé par l'Université du Michigan. L'échantillon de la HRS a été établi à l'aide d'une base de sondage probabiliste en plusieurs étapes et regroupé par région à l'échelle nationale (Heeringa et Connor, 1995). Depuis 2006, la HRS a initié la rotation des interviews en personne et par téléphone améliorés d'une vague à l'autre au niveau des ménages, sauf lorsque le ménage comprend des participants âgés de 80 ans ou plus, qui alternent entre l'interview en personne régulier et amélioré, ou des participants nouvellement recrutés, qui sont affectés à l'interview en personne amélioré lors de leur première vague. Dans la présente étude, nous souhaitons analyser les données de la HRS de 2016, lorsque la cohorte des derniers baby-boomers a été ajoutée pour réapprovisionner l'échantillon de la HRS. Bien que tous les intervieweurs ne recueillent pas de données dans les deux modes, selon le plan de sondage de la HRS, les intervieweurs sont responsables de la collecte de données dans les modes en personne et par téléphone. La HRS de 2016 a été menée d'avril 2016 à avril 2018, dont la taille de l'échantillon était de 20 912 [taux de réponse : 82,8 % (HRS, 2023)]. Dans notre échantillon analytique, nous avons exclu les répondants qui n'avaient pas de données sur les indicateurs de mode, pas de numéros d'intervieweur ou pour qui des covariables étaient manquantes, ce qui a donné une taille d'échantillon de 20 868.

Nous examinons quatre types de variables de résultat dans la HRS, notamment : 1) neuf éléments de l'échelle d'évaluation de la dépression du Center for Epidemiological Studies (CES-D); 2) six éléments des observations de l'intervieweur; 3) trois éléments de l'échelle de l'activité physique. La formulation de la question, les catégories de réponses originales et les catégories utilisées dans l'étude se trouvent à l'annexe E de (Yu, Elliott et Raghunathan, 2024). Nous considérons huit covariables au niveau du répondant (X), y compris l'âge, le genre, la race/l'origine ethnique, la langue de l'interview, l'éducation, si les répondants sont en couple et travaillent. Tous les participants sont compris dans notre échantillon, à moins qu'ils ne

manquent des données dans les variables de résultat ou les variables prédictives. Les taux manquants pour les variables prédictives sont mineurs, et ceux pour les variables de résultat sont inférieurs à 0,05.

3.2 Stratégie analytique

À l'instar des statistiques descriptives rapportées dans l'étude de l'Arab Barometer, nous rapportons l'écart-type entre intervieweurs et l'écart-type moyen de l'intervieweur lui-même afin d'obtenir une compréhension intuitive des effets de l'intervieweur dans la HRS.

Ensuite, nous ajustons les modèles à plusieurs niveaux à chacune des variables de résultat en utilisant la même notation que dans le modèle (2.2). Contrairement à l'étude de l'Arab Barometer, les intervieweurs ne sont pas imbriqués dans un modèle, donc un intervieweur peut faire des interviews dans les deux modes, ce qui permet de corréler les effets de l'intervieweur entre les modes. Par conséquent, nous présentons un modèle bivarié normal pour les effets de l'intervieweur :

$$
\begin{aligned}
Y^*_{ijm} &= \beta_0 + \beta_1 M_i + b_{jm} + \sum_s^S \gamma_s x_{si} + \epsilon_{ijm}, \\
&- \ Y_{ijm} = 1 \text{ si } Y^*_{ijm} > 0 \text{ et } Y_{ijm} = 0 \text{ si } Y^*_{ijm} \le 0, \\
&\begin{pmatrix} b_{jf} \\ b_{jt} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_f^2 & \rho\sigma_f\sigma_t \\ \rho\sigma_f\sigma_t & \sigma_t^2 \end{pmatrix} \right), \\
&\epsilon_{ijm} \sim N(0,1), \qquad (3.1) \\
&\sigma_f, \sigma_t \sim \text{ moitié } - T(3,1) \text{ (pour la modélisation bayésienne)}, \\
&\rho \sim U(-1,1) \text{ (pour la modélisation bayésienne)}, \\
&\gamma, \beta_0, \beta_1 \sim N(0, 10^6) \text{ (pour la modélisation bayésienne)}.
\end{aligned}
$$

De même, nous utilisons $\alpha = \log(\sigma_f) - \log(\sigma_t)$ comme mesure pour répondre à notre question de recherche. Pour vérifier si α est égal à zéro pour chaque variable, nous évaluons si les intervalles crédibles à 95 % ou les intervalles de confiance comprennent zéro. De plus, pour contrôler les effets de sélection des intervieweurs, nous incluons des covariables au niveau du répondant comme effets fixes dans le modèle.

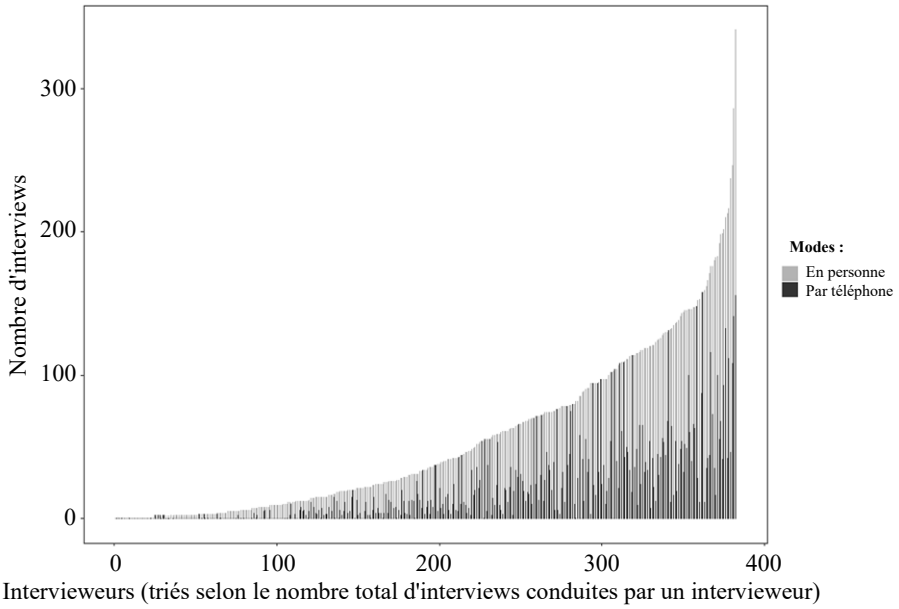
Nous appliquons la transformation Fisher Z $\left(z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)\right)$ lors de la construction de l'intervalle de confiance de 95 % pour ρ dans l'approche de vraisemblance. Nous calculons la variance de α grâce à l'utilisation de la méthode delta, donnée par $\text{var}(\alpha) = \frac{1}{4}\text{var}(\log(\sigma_f^2)) + \frac{1}{4}\text{var}(\log(\sigma_t^2)) - \frac{1}{2}\text{cov}(\log(\sigma_f^2), \log(\sigma_t^2))$, qui est légèrement différente de l'étude de l'Arab Barometer (voir les dérivations à l'annexe A).

3.3 Résultats

3.3.1 Statistiques descriptives

Premièrement, nous illustrons la charge de l'intervieweur à la figure 3.1. Dans la HRS de 2016, 382 intervieweurs ont été employés pour la collecte de données. Le nombre d'interviews réalisées en personne et par téléphone est très différent d'un intervieweur à l'autre. Au total, 82 (21,5 %) intervieweurs ont mené exclusivement des interviews par téléphone, tandis que 37 (9,7 %) ont uniquement mené des interviews en personne. Les 263 autres intervieweurs (68,9 %) ont effectué les deux types d'interviews. Toutes les interviews sont comprises dans l'analyse, bien que l'estimation des covariances entre les effets de l'interview en personne et de l'interview par téléphone pour l'intervieweur lui-même se limite au sous-échantillon des intervieweurs qui ont effectué les deux types d'interviews.

Figure 3.1 Charges de travail des intervieweurs par mode dans la Health and Retirement Study.



Deuxièmement, nous présentons les caractéristiques non pondérées des échantillons pour les modes en personne et par téléphone au tableau 3.1. Comparativement aux répondants par téléphone, une proportion plus élevée de répondants en personne avaient moins de 60 ans ou plus de 80 ans, appartenaient à des groupes minoritaires, n'étaient pas en couple, n'avaient pas terminé leurs études secondaires et étaient employés. Cette répartition déséquilibrée de l'échantillon souligne l'importance d'inclure des variables démographiques et du statut socioéconomique dans le modèle analytique lors de l'analyse des effets de l'intervieweur. En comparant les statistiques de la HRS à celles de l'étude de l'Arab Barometer, nous

constatons que dans la HRS, les écarts-types entre les intervieweurs sont généralement plus élevés et les écarts-types moyens de l'intervieweur lui-même sont généralement plus faibles. Cela donne à penser que les effets de sélection des intervieweurs sont potentiellement une menace plus grande lorsqu'on analyse la variance des intervieweurs dans la HRS. Cela est conforme à nos attentes, car l'attribution du mode se fait de façon aléatoire dans l'étude de l'Arab Barometer, mais pas dans la HRS.

Tableau 3.1
Répartition des caractéristiques de l'échantillon de la Health and Retirement Study 2016 entre les intervieweurs selon le mode.

Caractéristiques du répondant	Moyenne (EP)	Moyenne (TEL)	Écart-type entre intervieweurs (EP)	Écart-type entre intervieweurs (TEL)	Écart-type moyen de l'intervieweur lui-même (EP)	Écart-type moyen de l'intervieweur lui-même (TEL)
Âge : Moins de 60 ans	0,449	0,305	0,359	0,315	0,294	0,363
Âge : 60 à 69 ans	0,188	0,343	0,156	0,239	0,260	0,398
Âge : 70 à 79 ans	0,181	0,287	0,150	0,255	0,240	0,342
Âge : 80 ans et plus	0,182	0,066	0,185	0,151	0,232	0,163
Travaille actuellement	0,368	0,329	0,280	0,265	0,426	0,427
Masculin	0,414	0,415	0,177	0,198	0,503	0,503
Hispanique hispanophone	0,091	0,085	0,194	0,212	0,087	0,072
Hispanique anglophone	0,077	0,074	0,158	0,146	0,198	0,189
Noir	0,219	0,200	0,274	0,246	0,315	0,344
Blanc	0,613	0,641	0,315	0,300	0,376	0,397
En couple	0,601	0,632	0,260	0,275	0,431	0,428
Scolarité : moins de 12 ans	0,203	0,188	0,190	0,225	0,348	0,324
Scolarité : 12 ans	0,303	0,290	0,188	0,218	0,420	0,416
Scolarité : 13 à 15 ans	0,259	0,268	0,196	0,200	0,406	0,421
Scolarité : 16 ans et plus	0,249	0,262	0,214	0,220	0,391	0,374

Note : EP = en personne; TEL = par téléphone.

Ensuite, nous présentons les statistiques descriptives de la HRS, y compris les moyennes d'échantillons propres à un mode, les écart-types entre les intervieweurs et les écart-types moyens de l'intervieweur lui-même au tableau 3.2. Premièrement, selon l'échelle du CES-D, les taux de prévalence sont généralement plus élevés dans les interviews en personne que dans les interviews par téléphone, ce qui laisse entendre que le mode en personne peut être associé à des déclarations plus honnêtes. Deuxièmement, l'ampleur des écart-types entre les intervieweurs semble plus grande dans les observations de l'intervieweur et les éléments d'activité physique que dans les éléments du CES-D, ce qui indique des niveaux potentiellement différents d'effets de l'intervieweur dans différents résultats.

Tableau 3.2
Répartition des variables de résultat de la Health and Retirement Study 2016 entre les intervieweurs selon le mode.

Questions	Moyenne (EP)	Moyenne (TEL)	Écart-type entre les intervieweurs (EP)	Écart-type entre les intervieweurs (TEL)	Écart-type moyen de l'intervieweur lui-même (EP)	Écart-type moyen de l'intervieweur lui-même (TEL)
Questions du CES-D						
1. Vous vous sentiez déprimé.	0,156	0,117	0,177	0,148	0,306	0,268
2. Vous estimiez que tout ce que vous faisiez représentait un effort.	0,336	0,252	0,230	0,215	0,431	0,389
3. Votre sommeil était agité.	0,352	0,301	0,222	0,219	0,442	0,428
4. Vous étiez heureux (CODE INVERSÉ).	0,174	0,142	0,190	0,173	0,325	0,295
5. Vous vous sentiez seul.	0,207	0,152	0,191	0,158	0,365	0,321
6. Vous profitiez de la vie (CODE INVERSÉ).	0,113	0,077	0,151	0,115	0,261	0,214
7. Vous vous sentiez triste.	0,246	0,192	0,218	0,184	0,381	0,353
8. Vous ne pouviez pas vous motiver.	0,211	0,173	0,178	0,176	0,375	0,332
9. Déprimé (≥ 4 symptômes)	0,182	0,117	0,185	0,138	0,338	0,275
Observations de l'intervieweur						
10. Attentif aux questions	0,799	0,797	0,210	0,230	0,334	0,320
11. Compréhension des questions	0,463	0,474	0,272	0,304	0,437	0,405
12. Coopération	0,716	0,660	0,259	0,284	0,377	0,396
13. Difficulté à se souvenir des faits	0,539	0,588	0,288	0,316	0,422	0,383
14. Difficulté à vous entendre	0,803	0,741	0,202	0,254	0,325	0,359
15. Qualité de cette entrevue	0,591	0,623	0,323	0,326	0,378	0,366
Activité physique						
16. Activités ou sports vigoureux	0,352	0,347	0,218	0,221	0,443	0,454
17. Activités ou sports modérément exigeants	0,673	0,651	0,216	0,234	0,426	0,441
18. Activités ou sports peu exigeants	0,806	0,771	0,168	0,207	0,362	0,379

Note : EP = en personne; TEL = par téléphone; CES-D = Center for Epidemiological Studies-Depression.

3.3.2 Effets du mode d'enquête sur les moyennes et les variances de l'intervieweur

Finalement, nous discutons des résultats de modélisation présentés dans le tableau 3.3 à l'aide de l'estimation bayésienne. Les effets positifs du mode d'enquête dans les moyennes se retrouvent dans quatre des neuf éléments associés à la dépression. Ces éléments sont « vous vous sentiez déprimé », « tout représentait un effort », « votre sommeil était agité » et un indicateur global de la dépression. Par exemple, pour une femme de moins de 60 ans, qui est hispanique anglophone, qui n'est pas en couple, qui n'est pas actuellement employée, et qui n'a pas fait d'études secondaires, participer à une interview en personne augmente la probabilité d'être qualifiée de dépressive de 8,01 %, par rapport à une interview par téléphone. De même, nous calculons 8,01 % en utilisant $\phi(\beta_0 + \beta_1 + \sum_s^S \gamma_s x_{si})\,\beta_1$, où ϕ est la fonction de densité de probabilité d'une distribution normale standard et S représente le nombre de covariables (x). Étant donné que les symptômes dépressifs constituent des renseignements de nature délicate et que le fait de les avouer pourrait causer de l'embarras aux répondants, nous croyons qu'un niveau plus élevé de symptômes dépressifs déclarés est plus proche de la vérité. Pour les éléments d'observation de l'intervieweur, les effets positifs du mode d'enquête dans les moyennes sont présents dans trois des six éléments. Dans le mode en personne, les intervieweurs ont évalué les répondants comme étant plus coopératifs et ayant une meilleure

écoute et ont déclaré que l'interview était de meilleure qualité dans l'ensemble, comparativement au mode par téléphone (tableau 3.3). Enfin, dans les éléments de l'activité physique, les répondants ont tendance à signaler plus souvent qu'ils pratiquent des sports peu exigeants lorsqu'ils répondent par l'entremise d'une interview en personne comparativement à une interview par téléphone.

Nous observons de plus faibles variances de l'intervieweur entre les réponses substantielles dans la HRS (tableau 3.3) comparativement à l'étude de l'Arab Barometer. Pour les éléments associés à la dépression, les écarts de l'intervieweur pour l'interview en personne et l'interview par téléphone varient de 0,002 à 0,032, correspondant aux coefficients de corrélation interne (CCI) entre 0,002 et 0,032. Dans les éléments d'activité physique, les variances de l'intervieweur varient de 0,007 (CCI : 0,007) à 0,031 (CCI : 0,030). Lorsque nous comparons l'ampleur des variances de l'intervieweur entre les variables, nous constatons des variances plus importantes de l'intervieweur pour les éléments d'observation de l'intervieweur (allant de 0,271 [CCI : 0,213] à 0,881 [CCI : 0,468]).

En ce qui a trait aux effets du mode d'enquête sur les variances de l'intervieweur, nous constatons des différences importantes pour 3 des 18 questions examinées dans la HRS, plus particulièrement une dans l'échelle d'évaluation de la dépression et deux dans les questions d'observation de l'intervieweur (tableau 3.3). Lorsqu'on demande aux participants s'ils se sentaient tristes, les résultats révèlent que le mode en personne est associé à de plus grandes variances de l'intervieweur. De plus, la variance de l'intervieweur pour le mode en personne est légèrement plus grande que pour le mode par téléphone pour l'élément « tout représentait un effort ». En général, pour les éléments de l'échelle d'évaluation de la dépression, les variances de l'intervieweur pour le mode en personne sont plus importantes que celles pour le mode par téléphone pour sept éléments sur neuf, mais pas toujours de façon importante. Ce résultat s'harmonise avec les résultats de l'Arab Barometer et peut être attribuable au fait que les intervieweurs abordent les éléments de nature délicate de manière différente dans le mode en personne par rapport au mode par téléphone.

Pour déterminer si les répondants ont de la difficulté à se souvenir et à entendre des choses, les résultats laissent entendre que les variances de l'intervieweur pour le mode par téléphone sont plus importantes que les variances de l'intervieweur pour le mode en personne. Cette constatation peut être attribuée au fait que les intervieweurs ont moins d'indices pour évaluer la qualité de l'interview par téléphone, par opposition au mode en personne, où les intervieweurs peuvent se fier aux expressions faciales ou au langage corporel des répondants pour déduire la capacité des participants à entendre les questions. Cela pourrait mener à des réponses qui seraient principalement déterminées par les jugements subjectifs des intervieweurs et entraînerait donc de plus grandes variances. En ce qui concerne les éléments d'activité physique, il n'y a aucune preuve pour rejeter l'hypothèse nulle selon laquelle les variances de l'intervieweur sont égales entre les modes.

Il n'est pas surprenant de trouver des corrélations plus élevées ($\rho > 0{,}8$) entre les effets aléatoires de l'intervieweur selon les modes pour les variables d'observation de l'intervieweur, auxquelles les intervieweurs répondent directement. Par contraste, pour les deux autres échelles (échelle d'évaluation de

la dépression du CES-D et échelle de l'activité physique), les effets des intervieweurs sur les réponses sont induits par l'intermédiaire des répondants, ce qui entraîne une corrélation plus faible et moins stable entre les modes en personne et par téléphone.

Même si nous nous concentrons sur la communication des résultats bayésiens, nous fournissons des inférences à partir des procédures bayésiennes et de vraisemblance à l'annexe B. Nous constatons qu'en général, les estimations tirées des deux méthodes sont semblables, sauf pour l'estimation de la corrélation (ρ). Les corrélations sont associées à de larges intervalles dans les échelles du CES-D et les éléments de l'activité physique. En outre, les estimations ponctuelles de la corrélation sont parfois très différentes entre les deux procédures, en particulier pour les deux types d'éléments mentionnés ci-dessus. Sur deux éléments, soit « heureux » et « se sentait triste », la corrélation ne peut être estimée en utilisant l'approche de vraisemblance. Cela peut être attribuable aux faibles variances de l'intervieweur dans l'échelle, ce qui rend l'estimation de la covariance numériquement difficile et donc instable. En outre, cela peut être attribué au déséquilibre du fardeau des intervieweurs entre les modes. Environ 30 % des intervieweurs ne mènent des interviews qu'en un seul mode, et 51 % des intervieweurs effectuent moins de cinq interviews, que ce soit en personne ou par téléphone. Ce déséquilibre peut entraîner un manque de renseignements pour estimer ρ.

Tableau 3.3
Variances de l'intervieweur par mode pour certains éléments de la Health and Retirement Study en tenant compte des covariables à l'aide de l'estimation bayésienne.

Questions	σ_f^2	σ_t^2	$\rho_{f,\text{int}}$	$\rho_{t,\text{int}}$	α	β_1	ρ
Questions du CES-D							
Se sentait déprimé	0,011	0,013	0,011	0,013	0,044	**0,056**	0,07
	[0; 0,022]	[0; 0,03]	[0,000; 0,022]	[0,000; 0,029]	[-1,148; 1,533]	[0,005; 0,114]	[-0,551; 0,874]
Tout représentait un effort	0,025	0,007	0,024	0,007	0,746	**0,118**	-0,128
	[0,013; 0,037]	[0,001; 0,016]	[0,013; 0,036]	[0,001; 0,016]	[-0,002; 1,496]	[0,071; 0,175]	[-0,56; 0,254]
Sommeil agité	0,002	0,005	0,002	0,005	-0,486	**0,053**	0,337
	[0; 0,007]	[0; 0,012]	[0,000; 0,007]	[0,000; 0,012]	[-1,89; 0,925]	[0,011; 0,095]	[-0,162; 0,849]
Heureux	0,011	0,011	0,011	0,011	0,128	0,032	**-0,518**
	[0,003; 0,021]	[0; 0,022]	[0,003; 0,021]	[0,000; 0,022]	[-0,889; 1,333]]	[-0,024; 0,083]	[-0,989; -0,006]
Se sentait seul	0,006	0,006	0,006	0,006	0,178	0,048	0,055
	[0; 0,014]	[0; 0,016]	[0,000; 0,014]	[0,000; 0,016]	[-1,455; 1,846]]	[-0,005; 0,099]	[-0,108; 0,218]
Profitait de la vie	0,01	0,007	0,010	0,007	0,551	0,061	**0,56**
	[0,001; 0,021]	[0; 0,025]	[0,001; 0,021]	[0,000; 0,024]	[-1,096; 2,148]]	[-0,005; 0,134]	[0,223; 0,921]
Se sentait triste	0,032	0,003	0,031	0,003	**1,694**	0,046	**0,296**
	[0,018; 0,048]	[0; 0,009]	[0,018; 0,046]	[0,000; 0,009]	[0,463; 3,775]	[-0,01; 0,097]	[0,037; 0,577]
Ne pouvait pas se motiver	0,02	0,02	0,020	0,020	-0,051	0,051	0,274
	[0,007; 0,029]	[0,006; 0,035]	[0,007; 0,028]	[0,006; 0,034]	[-0,732; 0,515]	[-0,01; 0,109]	[-0,346; 0,797]
Indicateur global	0,016	0,012	0,016	0,012	0,093	**0,15**	0,244
	[0,002; 0,024]	[0,001; 0,027]	[0,002; 0,023]	[0,001; 0,026]	[-0,901; 1,075]	[0,102; 0,207]	[-0,18; 0,573]

Notes : β_1 est l'effet de mode dans les moyennes, calculé comme la moyenne de l'estimation en personne moins la moyenne de l'estimation par téléphone. σ_f^2 est la variance de l'intervieweur pour l'interview en personne. σ_t^2 est la variance de l'intervieweur pour l'interview par téléphone. $\rho_{f,\text{int}}$ est la corrélation interne de l'intervieweur associée au mode en personne. $\rho_{t,\text{int}}$ est la corrélation interne de l'intervieweur associée au mode par téléphone. α désigne les différences de logarithme entre les variations de l'intervieweur des modes en personne et par téléphone. ρ est la corrélation entre les effets aléatoires de l'intervieweur des modes en personne et par téléphone. CES-D = Center for Epidemiological Studies-Depression.

Tableau 3.3(suite)
Variances de l'intervieweur par mode pour certains éléments de la Health and Retirement Study en tenant compte des covariables à l'aide de l'estimation bayésienne.

Questions	σ_f^2	σ_t^2	$\rho_{f,\text{int}}$	$\rho_{t,\text{int}}$	α	β_1	ρ
				Observations de l'intervieweur			
Attentif	0,298	0,351	0,230	0,260	-0,081	0,018	**0,878**
	[0,233; 0,356]	[0,262; 0,431]	[0,189; 0,263]	[0,208; 0,301]	[-0,197; 0,038]	[-0,049; 0,088]	[0,803; 0,955]
Compréhension	0,413	0,465	0,292	0,317	-0,058	0	**0,91**
	[0,341; 0,493]	[0,366; 0,56]	[0,254; 0,330]	[0,268; 0,359]	[-0,149; 0,043]	[-0,064; 0,061]	[0,861; 0,958]
Coopération	0,459	0,41	0,315	0,291	0,057	**0,178**	**0,931**
	[0,378; 0,556]	[0,321; 0,51]	[0,274; 0,357]	[0,243; 0,338]	[-0,039; 0,138]	[0,108; 0,236]	[0,881; 0,971]
Se souvenir	0,483	0,605	0,326	0,377	**-0,112**	-0,062	**0,931**
	[0,392; 0,574]	[0,489; 0,721]	[0,282; 0,365]	[0,328; 0,419]	[-0,205; -0,028]]	[-0,124; 0,002]	[0,885; 0,972]
Entendre	0,271	0,375	0,213	0,273	**-0,161**	**0,151**	**0,87**
	[0,212; 0,335]	[0,274; 0,462]	[0,175; 0,251]	[0,215; 0,316]	[-0,284; -0,037]	[0,084; 0,229]	[0,795; 0,947]]
Qualité globale	0,881	0,788	0,468	0,441	0,057	**0,086**	**0,94**
	[0,749; 1,04]	[0,641; 0,949]	[0,428; 0,510]	[0,391; 0,487]	[-0,032; 0,14]	[0,014; 0,158]	[0,913; 0,983]
				Activité physique			
Sports vigoureux	0,017	0,007	0,017	0,007	0,523	-0,037	0,36
	[0,007; 0,026]	[0; 0,015]	[0,007; 0,025]	[0,000; 0,015]	[-0,209; 1,45]	[-0,081; 0,014]	[-0,446; 0,827]
Sport modéré	0,015	0,019	0,015	0,019	-0,086	0,031	0,233
	[0,006; 0,024]	[0,004; 0,033]	[0,006; 0,023]	[0,004; 0,032]	[-0,655; 0,464]	[-0,019; 0,078]	[-0,351; 0,698]
Sport léger	0,02	0,031	0,020	0,030	-0,355	**0,134**	0,144
	[0,002; 0,03]	[0,014; 0,052]	[0,002; 0,029]	[0,014; 0,049]	[-1,097; 0,324]	[0,073; 0,184]	[-0,264; 0,962]

Notes : β_1 est l'effet de mode dans les moyennes, calculé comme la moyenne de l'estimation en personne moins la moyenne de l'estimation par téléphone. σ_f^2 est la variance de l'intervieweur pour l'interview en personne. σ_t^2 est la variance de l'intervieweur pour l'interview par téléphone. $\rho_{f,\text{int}}$ est la corrélation interne de l'intervieweur associée au mode en personne. $\rho_{t,\text{int}}$ est la corrélation interne de l'intervieweur associée au mode par téléphone. α désigne les différences de logarithme entre les variations de l'intervieweur des modes en personne et par téléphone. ρ est la corrélation entre les effets aléatoires de l'intervieweur des modes en personne et par téléphone. CES-D = Center for Epidemiological Studies-Depression.

Pour relever les défis numériques et évaluer si l'estimation d'autres paramètres (par exemple, σ_f^2 σ_t^2 et α) est de nature délicate par rapport à ρ, nous avons fixé ρ à 0 et à la moyenne postérieure obtenue grâce à la procédure bayésienne, et nous avons réappliqué le modèle (3.1) pour les éléments du CES-D. Nous constatons que les estimations des variances de l'intervieweur demeurent pratiquement inchangées lorsqu'il s'agit de définir ρ à des valeurs différentes ou d'estimer ρ. Les résultats se trouvent à l'annexe G de (Yu, Elliott et Raghunathan, 2024). Ainsi, nous concluons qu'il y a peu de sensibilité dans les inférences fournies par l'estimation de vraisemblance à ρ.

4. Étude par simulation

Pour comprendre les propriétés d'échantillonnage répété de la méthode que nous proposons, y compris le pouvoir de détecter les effets du mode d'enquête dans les tailles d'échantillons typiquement modestes de l'intervieweur qui sont disponibles, nous avons effectué des études de simulation en utilisant la configuration de l'étude de l'Arab Barometer et de la HRS.

4.1 Étude de l'Arab Barometer

Cette étude de simulation est conçue de telle sorte que le nombre de répondants (n = 2 521) et d'intervieweurs (13 pour le mode par téléphone et 31 pour le mode en personne) soit le même que dans l'étude de

l'Arab Barometer, ainsi que la façon dont les répondants sont jumelés aux intervieweurs. Nous considérons quatre scénarios : 1) scénario sans différence où la variance de l'intervieweur pour le mode en personne est égale à la variance de l'intervieweur pour le mode par téléphone ($\sigma_f^2 = \sigma_t^2 = 0{,}14;\ \alpha_0 = -0{,}98$ et $\alpha = 0$); 2) de petites différences où $\sigma_f^2 = 0{,}20;\ \sigma_t^2 = 0{,}14;\ \alpha_0 = -0{,}98$ et $\alpha = 0{,}18$; 3) de moyennes différences où $\sigma_f^2 = 0{,}24;\ \sigma_t^2 = 0{,}14;\ \alpha_0 = -0{,}98$ et $\alpha = 0{,}27$; 4) de grandes différences où $\sigma_f^2 = 0{,}50;\ \sigma_t^2 = 0{,}14;\ \alpha_0 = -0{,}98$ et $\alpha = 0{,}64$. Nous considérons le vrai modèle de génération de données comme suit :

$$\begin{aligned} \eta_i &= \Phi(\beta_0 + \beta_1 M_{ij} + b_{j(m)}), \\ b_{j(m)} &\sim N(0, \sigma_m^2), \\ y_i &\sim \text{Bernoulli}(\eta_i), \end{aligned}$$

où i représente l'indice pour les répondants, j représente l'indice pour les intervieweurs, m indique les modes (f ou t) $\Phi()$ est la fonction de distribution cumulative de la distribution normale standard, et M est un vecteur $n \times 1$ du mode utilisé par chaque participant pour participer à l'enquête.

Nous adaptons le même modèle analytique (2.2) aux données simulées, mises en œuvre séparément en utilisant Proc Nlmixed et Proc MCMC dans le langage de programmation SAS. La simulation est répétée $K = 200$ fois, où pour chaque itération, les estimations ponctuelles, les erreurs-types et les intervalles de confiance de 95 % ou les intervalles crédibles de β_1, σ_f^2, σ_t^2 et α sont calculés et enregistrés. En fonction de ces statistiques, nous présentons le biais, le taux de couverture, le ratio d'erreur-type (ET) et la puissance dans chaque scénario pour les paramètres.

$$\begin{aligned} \text{Biais}(\hat{\delta}) &= \frac{1}{K}\sum_k^K \hat{\delta}_k - \delta, \\ \text{Taux de couverture}(\hat{\delta}) &= \frac{1}{K}\sum_k^K I(\hat{\delta}_{k,\text{lw}} < \delta \ \&\ \hat{\delta}_{k,\text{up}} > \delta), \\ \text{Ratio d'ET}(\hat{\delta}) &= \frac{1}{K}\sum_k^K \sqrt{\hat{\text{var}}(\hat{\delta}_k)} \Big/ \sqrt{\frac{1}{K-1}\sum_k^K (\hat{\delta}_k - \bar{\hat{\delta}}_k)^2}, \\ \text{Puissance}(\hat{\delta}) &= 1 - \frac{1}{K}\sum_k^K I(\hat{\delta}_{k,\text{lw}} < 0 \ \&\ \hat{\delta}_{k,\text{up}} > 0) \ \text{ quand } \delta \neq 0, \end{aligned}$$

où δ désigne les paramètres que nous souhaitons estimer (c'est-à-dire σ_f^2, σ_t^2, β_1 et α), $\hat{\delta}_k$ est l'estimation ponctuelle estimative de δ obtenue dans l'itération k, $\hat{\delta}_{k,\text{lw}}$ et $\hat{\delta}_{k,\text{up}}$ est la limite inférieure et la limite supérieure du paramètre estimé.

Le tableau 4.1 affiche les résultats de la simulation qui repose sur la configuration de l'étude de l'Arab Barometer. Lorsque $\alpha = 0,$ la puissance indiquée dans le tableau 4.1 représente le taux d'erreur de type 1. Nous observons que le pouvoir de rejeter l'hypothèse nulle selon laquelle les variances de l'intervieweur sont égales $(\alpha = 0)$ est limité dans tous les scénarios. Cependant, à mesure que les différences augmentent $(0{,}18 - 0{,}64),$ la puissance augmente de 0,075 à 0,520 dans la procédure bayésienne et de 0,110 à 0,633 dans l'approche « fréquentiste ». Il y a quelques différences dans le pouvoir fourni par les approches de

vraisemblance et bayésienne. En effet, les procédures de vraisemblance n'offrent pas de taux de couverture nominal dans les scénarios 1 à 3; par conséquent, la puissance obtenue des procédures de vraisemblance et bayésienne repose sur des niveaux de signification différents. La faible puissance de α est principalement attribuable au nombre très limité d'intervieweurs dans les modes en personne et par téléphone. Inversement, le pouvoir de rejeter l'hypothèse nulle selon laquelle il n'y a pas d'effets du mode d'enquête dans les moyennes (β_1) lorsque l'hypothèse de rechange est vraie est considérablement plus élevé (autour de 0,90). Cependant, comme α devient plus important et que les variances de l'intervieweur augmentent simultanément, nous observons une baisse de puissance de β_1, en raison de la diminution de la taille effective de l'échantillon du CCI accru.

Tableau 4.1
Étude de simulation qui repose sur la configuration de l'étude de l'Arab Barometer.

Paramètres	Résultats de vraisemblance				Résultats bayésiens			
	Biais	Taux de couverture	Ratio d'erreur-type	Puissance	Biais	Taux de couverture	Ratio d'erreur-type	Puissance
Scénario 1 : Aucune différence								
$\sigma_f^2 = 0{,}14$	-0,002	0,950	1,000	N/A	0,017	0,940	1,059	N/A
$\sigma_t^2 = 0{,}14$	-0,001	0,955	1,014	N/A	0,049	0,975	1,346	N/A
$\beta_1 = 0{,}5$	-0,003	0,965	1,023	0,935	0,006	0,955	1,121	0,930
$\alpha = 0$	0,028	0,930	0,888	0,070	-0,033	0,985	1,107	0,015
Scénario 2 : Petites différences								
$\sigma_f^2 = 0{,}20$	-0,012	0,960	0,948	N/A	0,028	0,975	1,105	N/A
$\sigma_t^2 = 0{,}14$	-0,007	0,935	0,974	N/A	0,059	0,955	1,161	N/A
$\beta_1 = 0{,}5$	-0,002	0,940	0,926	0,950	-0,001	0,950	1,078	0,900
$\alpha = 0{,}18$	0,042	0,920	0,928	0,110	-0,020	0,950	0,955	0,075
Scénario 3 : Différences moyennes								
$\sigma_f^2 = 0{,}24$	-0,002	0,920	0,947	N/A	0,039	0,920	0,980	N/A
$\sigma_t^2 = 0{,}14$	-0,013	0,955	1,009	N/A	0,061	0,980	1,311	N/A
$\beta_1 = 0{,}5$	0,004	0,935	0,940	0,920	-0,010	0,960	1,184	0,860
$\alpha = 0{,}27$	0,079	0,905	0,922	0,230	-0,042	0,960	1,075	0,085
Scénario 4 : Grandes différences								
$\sigma_f^2 = 0{,}50$	-0,007	0,970	1,058	N/A	0,078	0,950	1,093	N/A
$\sigma_t^2 = 0{,}14$	-0,009	0,960	1,055	N/A	0,054	0,935	1,231	N/A
$\beta_1 = 0{,}5$	0,022	0,935	0,965	0,824	-0,016	0,980	1,097	0,690
$\alpha = 0{,}64$	0,079	0,945	0,906	0,633	0,012	0,955	0,882	0,520

Notes : β_1 est l'effet du mode d'enquête dans les moyennes, calculé comme la moyenne de l'estimation en personne moins la moyenne de l'estimation par téléphone. σ_f^2 représente les variances de l'intervieweur pour le mode en personne. σ_t^2 est la variance de l'intervieweur pour le mode par téléphone. α désigne les différences de logarithme entre les variances de l'intervieweur pour les modes en personne et par téléphone.

4.2 Health and Retirement Study

Dans l'étude de simulation utilisant la configuration de la HRS, nous considérons le modèle de génération de données suivant en utilisant les mêmes notations que dans l'étude de simulation de l'Arab Barometer. Nous utilisons b_{jf} pour représenter les effets aléatoires de l'intervieweur dans le mode en personne et b_{jt} pour représenter les effets aléatoires de l'intervieweur dans le mode par téléphone :

$$\begin{aligned} \eta_i &= \Phi(\beta_0 + \beta_1 M_{ij} + b_{jf} M_{ij} + b_{jt}(1 - M_{ij})), \\ \begin{pmatrix} b_{jf} \\ b_{jt} \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_f^2 & \rho\sigma_f\sigma_t \\ \rho\sigma_f\sigma_t & \sigma_t^2 \end{pmatrix}\right), \\ y_i &\sim \text{Bernoulli}(\eta_i). \end{aligned}$$

Nous examinons quatre scénarios : 1) $\sigma_f^2 = \sigma_t^2 = 0{,}03$; $\alpha_0 = -1{,}75$ et $\alpha = 0$; 2) $\sigma_f^2 = 0{,}05$; $\sigma_t^2 = 0{,}03$; $\alpha_0 = -1{,}75$; et $\alpha = 0{,}26$; 3) $\sigma_f^2 = 0{,}06$; $\sigma_t^2 = 0{,}03$; $\alpha_0 = -1{,}75$; et $\alpha = 0{,}35$; 4) $\sigma_f^2 = 0{,}09$; $\sigma_t^2 = 0{,}03$; $\alpha_0 = -1{,}75$ et $\alpha = 0{,}55$. Dans tous les scénarios, $\beta_1 = 0{,}5$ et $\rho = 0{,}5$. Nous présentons le biais, le taux de couverture, le ratio d'ET, et la puissance pour ces paramètres et les différences logarithmiques des variances de l'intervieweur entre les modes en personne et par téléphone (α) au tableau 4.2.

Tableau 4.2
Étude de simulation qui repose sur la configuration de la Health and Retirement Study.

Paramètres	Résultats de vraisemblance				Résultats bayésiens			
	Biais	Taux de couverture	Ratio d'erreur-type	Puissance	Biais	Taux de couverture	Ratio d'erreur-type	Puissance
Scénario 1 : Aucune différence								
$\sigma_f^2 = 0{,}03$	-0,000	0,980	1,085	N/A	0,003	0,965	1,704	N/A
$\sigma_t^2 = 0{,}03$	-0,001	0,975	1,049	N/A	0,002	0,935	1,469	N/A
$\beta_1 = 0{,}5$	-0,002	0,940	1,029	1,000	-0,000	0,960	1,128	1,000
$\rho = 0{,}5$	0,012	0,965	1,009	0,470	-0,020	0,925	1,061	0,690
$\alpha = 0$	0,022	0,965	1,019	0,035	0,047	0,965	0,928	0,035
Scénario 2 : Petites différences								
$\sigma_f^2 = 0{,}05$	0,000	0,940	0,999	N/A	0,001	0,955	1,507	N/A
$\sigma_t^2 = 0{,}03$	-0,000	0,975	1,125	N/A	0,002	0,945	1,249	N/A
$\beta_1 = 0{,}5$	0,003	0,960	0,996	1,000	0,001	0,950	0,983	1,000
$\rho = 0{,}5$	0,020	0,980	1,084	0,695	-0,021	0,925	1,032	0,755
$\alpha = 0{,}26$	0,018	0,940	0,978	0,270	0,008	0,940	0,934	0,295
Scénario 3 : Différences moyennes								
$\sigma_f^2 = 0{,}06$	-0,001	0,945	0,999	N/A	0,001	0,950	1,268	N/A
$\sigma_t^2 = 0{,}03$	-0,001	0,975	1,045	N/A	0,002	0,940	1,103	N/A
$\beta_1 = 0{,}5$	-0,001	0,920	0,993	1,000	0,001	0,965	1,007	1,000
$\rho = 0{,}5$	0,011	0,970	1,030	0,665	-0,009	0,930	1,014	0,815
$\alpha = 0{,}35$	0,024	0,910	0,919	0,510	0,008	0,945	0,949	0,530
Scénario 4 : Grandes différences								
$\sigma_f^2 = 0{,}09$	0,000	0,930	0,983	N/A	0,002	0,950	1,201	N/A
$\sigma_t^2 = 0{,}03$	-0,001	0,955	1,054	N/A	-0,001	0,915	1,089	N/A
$\beta_1 = 0{,}5$	0,004	0,950	1,009	1,000	-0,002	0,955	1,031	1,000
$\rho = 0{,}5$	0,009	0,985	1,085	0,750	0,004	0,970	1,121	0,860
$\alpha = 0{,}55$	0,029	0,915	0,977	0,935	0,070	0,950	0,955	0,990

Notes : β_1 est l'effet du mode d'enquête dans les moyennes, calculé comme la moyenne de l'estimation en personne moins la moyenne de l'estimation par téléphone. σ_f^2 représente les variances de l'intervieweur pour le mode en personne. σ_t^2 est la variance de l'intervieweur pour le mode par téléphone. α désigne les différences de logarithme entre les variances de l'intervieweur pour les modes en personne et par téléphone. ρ est la corrélation entre les effets aléatoires de l'intervieweur pour le mode en personne et le mode par téléphone.

Le tableau 4.2 montre qu'à mesure que α augmente de 0 à 0,55, la puissance augmente proportionnellement de 0,035 à 0,990 à l'aide de la procédure bayésienne, et de 0,035 à 0,935 en utilisant l'approche de vraisemblance. Les résultats donnent à penser que lorsque α est assez grand, nous pouvons atteindre une puissance raisonnablement élevée grâce à la configuration de la HRS. En comparant le tableau 4.1 et le tableau 4.2, nous constatons que le pouvoir de rejeter l'hypothèse nulle affirmant des variances égales des

intervieweurs, lorsque l'hypothèse de rechange est vraie, dépasse celui de la simulation de l'étude de l'Arab Barometer. Ce résultat est conforme aux attentes, étant donné le grand nombre d'intervieweurs qui participent dans la HRS. En outre, nous constatons que l'approche de vraisemblance ne peut pas toujours atteindre les taux de couverture nominale de 95 % (dans les scénarios 3 et 4), de sorte que la puissance calculée en utilisant les procédures de vraisemblance et bayésienne repose sur des niveaux de signification différents.

5. Analyse

Le présent document permet d'examiner la présence d'effets du mode d'enquête dans les variances de l'intervieweur sur plusieurs éléments dans deux enquêtes nationales. Dans l'étude de l'Arab Barometer, nous trouvons des preuves statistiques de divergences dans les effets des intervieweurs entre les modes en personne et par téléphone dans un des six éléments de nature délicate (légèrement) et un des trois indicateurs manquants. En outre, pour les éléments de nature délicate et les indicateurs manquants dans l'étude de l'Arab Barometer, les variances de l'intervieweur du mode en personne sont généralement plus importantes que celles du mode par téléphone. Parallèlement, nous devrions interpréter les résultats de l'étude de l'Arab Barometer avec prudence. En raison du petit nombre d'intervieweurs qui participent à l'étude, les résultats nuls ne peuvent être traduits en effets faibles ou nuls, ce qui entrave quelque peu la force de la preuve. À l'aide des données de la HRS de 2016, nous observons des effets importants du mode d'enquête dans les variances des intervieweurs sur deux éléments de dépression (un légèrement) et deux éléments d'observation des intervieweurs. Pour les éléments de la dépression de nature sensible, un modèle semblable se dégage, avec de plus grandes variances de l'intervieweur pour le mode en personne comparativement au mode par téléphone. Ces résultats semblent indiquer que les questions de nature délicate et les éléments manquants sont des défis cruciaux pour régulariser les variances de l'intervieweur entre les modes. De plus, l'ampleur des variances de l'intervieweur est beaucoup plus grande sur les éléments d'observation de l'intervieweur que sur les réponses substantielles. De plus, des données probantes montrent que les variances de l'intervieweur pour le mode par téléphone sont plus importantes que les variances de l'intervieweur pour le mode en personne pour ces éléments. Cela pourrait s'expliquer par le fait que ces questions comportent des évaluations plus subjectives et peuvent offrir de plus grandes possibilités de réduire les variances de l'intervieweur en normalisant les protocoles d'intervieweurs pour ces éléments, en particulier du mode par téléphone.

Les études de simulation donnent à penser qu'il est possible d'obtenir une puissance raisonnable selon la configuration de l'étude de l'Arab Barometer ou de la HRS s'il y a d'importants effets du mode d'enquête dans les variances de l'intervieweur. Cependant, avec de petits effets du mode d'enquête, la puissance est limitée, en particulier dans la configuration de l'étude de l'Arab Barometer. L'observation d'effets

importants du mode d'enquête dans les variances des intervieweurs dans les données de l'étude de l'Arab Barometer et de la HRS souligne l'importance de tenir compte du rôle des modes sur les effets de l'intervieweur, particulièrement lorsqu'il s'agit de sujets de nature délicate et de la non-réponse d'un élément. Étant donné le nombre généralement limité d'intervieweurs employés dans la plupart des enquêtes, une constatation nulle n'indique pas nécessairement une variance de l'intervieweur qui est égale. Toutefois, il est encore utile pour les organismes d'enquête de considérer ce genre d'enquête comme une conclusion positive, et cela devrait capter l'attention des chercheurs. En outre, en présence de multiples études « insuffisantes » qui emploient peu d'intervieweurs, une méta-analyse peut être effectuée pour combiner les inférences tirées de ces études et mieux explorer les effets du mode d'enquête dans les variances de l'intervieweur.

La documentation a abondamment documenté la question de savoir si les modes ont une incidence sur les erreurs de mesure au niveau du répondant (Tourangeau et Smith, 1996; Kreuter, Presser et Tourangeau, 2008). Toutefois, peu d'études ont examiné si et comment les modes influent sur les erreurs de mesure liées à l'intervieweur, particulièrement après l'adoption généralisée de plans de sondage à mode mixte. Le présent document vise à combler cet écart en analysant deux enquêtes nationales comportant des caractéristiques distinctes de plan de sondage à mode mixte, comme le nombre d'intervieweurs et la question de savoir si les intervieweurs sont imbriqués dans des modes. Lorsque les intervieweurs sont imbriqués dans des modes, il est difficile de déterminer si les différences observées sont attribuables aux modes ou aux intervieweurs. L'approche actuelle de modélisation présuppose que toutes les différences systématiques entre les réponses recueillies par téléphone et en personne sont la conséquence des modes, et non des intervieweurs. Si les organismes d'enquête disposent de renseignements sur les caractéristiques des intervieweurs, ils peuvent évaluer cette hypothèse en comparant les caractéristiques des intervieweurs entre les modes. Une telle analyse aiderait à démêler les effets des modes d'enquête de ceux des intervieweurs, fournissant des indications précieuses sur la qualité des données de l'enquête.

Pour les plans de sondage qui permettent aux intervieweurs de recueillir des données selon les deux modes, les modèles présentés dans le présent document permettent d'estimer les effets par intervieweur dans chaque mode. Cela est utile pour détecter les intervieweurs qui ont une incidence importante sur les réponses dans un mode ou les deux. En utilisant ces effets estimés de l'intervieweur, nous pouvons déterminer si certains intervieweurs ont constamment des effets importants entre les variables, ce qui peut indiquer le besoin d'une intervention de la part des superviseurs des intervieweurs. Si des variables particulières sont associées à des variances importantes de l'intervieweur dans un certain mode, cela peut justifier de fournir une meilleure formation aux intervieweurs pour ces éléments. Par exemple, en fonction des résultats de la présente étude, on pourrait envisager un protocole d'interview plus standardisé pour les éléments de nature délicate et lorsque les répondants répondent « je ne sais pas » aux questions du mode en personne. À ce titre, nous recommandons que les organismes d'enquête intègrent ces analyses dans leurs évaluations

courantes de la qualité des données. Des recherches futures pourraient chercher à déterminer si les caractéristiques de l'intervieweur peuvent expliquer les effets différentiels de l'intervieweur observés entre les modes, ce qui pourrait faire la lumière sur les mécanismes sous-jacents en jeu.

Lorsqu'on détermine le mode à utiliser pour produire des estimations de population dans des études à mode mixte, il est souhaitable d'avoir un biais plus petit et des variances de l'intervieweur plus faibles, ce qui pourrait entraîner une erreur quadratique moyenne plus faible. Toutefois, en réalité, le mode ayant un biais plus petit et une variance de l'intervieweur plus faible peut ne pas toujours être le même, comme le montre le présent document. Par exemple, les interviews en personne peuvent être liées à moins de biais, mais à une plus grande variance de l'intervieweur. La façon d'équilibrer les compromis entre le biais et la variance dans une méthode formelle sera un sujet de recherche future. Cette étude présente deux exemples d'enquêtes pour évaluer les effets des modes, tant dans les moyennes que dans les variances de l'intervieweur. Si de telles analyses sont couramment adoptées par les chercheurs qui conçoivent et mettent en œuvre des études à mode mixte, on peut accumuler plus de preuves sur la question de savoir si et comment les intervieweurs auraient pu effectuer différemment la collecte selon les différents modes de collecte de données. Cela peut devenir la base du développement de futurs protocoles à mode mixte. Lorsque nous présentons les résultats de l'analyse, nous recommandons que les organismes d'enquête expliquent de quelle façon leurs intervieweurs sont affectés à différents modes ou sélectionnent eux-mêmes différents modes, et précisent si les effets du mode d'enquête observés sur les variances de l'intervieweur correspondent à leurs attentes.

Dans le présent document, nous observons des différences entre les résultats obtenus de la procédure de maximum de vraisemblance et la procédure bayésienne mise en œuvre dans le langage de programmation SAS. Lorsque les variances de l'intervieweur sont faibles, il peut être difficile d'adapter le modèle analytique aux effets aléatoires des intervieweurs corrélés selon les modes en utilisant l'approche de vraisemblance. Dans cette situation, l'approche bayésienne peut être particulièrement utile, car l'emploi de distributions appropriées et informatives nous permet de nous assurer que nous tirons des inférences de distributions postérieures appropriées.

La présente étude comporte trois principales limites. Tout d'abord, comme d'autres études semblables (West, Ong, Conrad, Schober, Larsen et Hupp, 2022; Groves et Magilavy, 1986), elle est confrontée à la question de la puissance statistique limitée, comme le démontre l'étude de simulation. Deuxièmement, nous considérons les résultats dichotomisés dans cette étude en raison de contraintes de calcul; cependant, il se peut que ce ne soit pas une approche optimale pour étudier la variance des intervieweurs, car agréger des catégories peut réduire les variances. Les études futures peuvent explorer cette question de recherche en utilisant différents types de résultats et des tailles d'échantillons plus grandes. Enfin, les deux enquêtes n'ont pas utilisé la randomisation dans le système d'affectation des intervieweurs. Idéalement, lors de l'estimation des variances de l'intervieweur, on devrait utiliser des plans de sondage interpénétrés pour s'assurer que la

variabilité est uniquement due au processus de mesure de l'intervieweur, plutôt qu'aux différences entre les répondants. Pour contourner l'absence de randomisation, nous avons inclus les caractéristiques des répondants dans le modèle d'analyse. Toutefois, les variances de l'intervieweur pourraient encore être surestimées en raison de covariables non observées qui n'ont pas été prises en compte dans les modèles.

Remerciements

Ce projet a été soutenu par la Daniel Katz Dissertation Fellowship in Psychology and Survey Methodology à l'Institute for Social Research de l'Université du Michigan. Les auteurs remercient les chercheurs, le personnel et les participants de l'étude de l'Arab Barometer et de la Health and Retirement Study pour leur précieuse contribution. Nous remercions le Dr Brady West pour son aide précieuse dans le cadre de cette étude. Nous reconnaissons avoir utilisé le modèle de langage d'intelligence artificielle, ChatGPT, développé par OpenAI, pour aider à corriger la grammaire et à perfectionner la rédaction de ce document.

Annexe A

Dérivations de la variance de α à l'aide de la méthode Delta

$$
\begin{aligned}
\mathrm{var}(\alpha) &= \mathrm{var}\left(\log(\sigma_f) - \log(\sigma_t)\right) \\
&= \mathrm{var}\left(\log(\sigma_f)\right) + \mathrm{var}\left(\log(\sigma_t)\right) - 2\mathrm{cov}\left(\log(\sigma_f), \log(\sigma_t)\right) \\
&= \frac{1}{4}\mathrm{var}\left(2\log(\sigma_f)\right) + \frac{1}{4}\mathrm{var}\left(2\log(\sigma_t)\right) - 2\mathrm{cov}\left(\log(\sigma_f), \log(\sigma_t)\right) \\
&= \frac{1}{4}\mathrm{var}\left(\log(\sigma_f^2)\right) + \frac{1}{4}\mathrm{var}\left(\log(\sigma_t^2)\right) - \frac{1}{4} \times 2\mathrm{cov}\left(\log(\sigma_f^2), \log(\sigma_t^2)\right) \\
&= \frac{1}{4}\mathrm{var}\left(\log(\sigma_f^2)\right) + \frac{1}{4}\mathrm{var}\left(\log(\sigma_t^2)\right) - \frac{1}{2}\mathrm{cov}\left(\log(\sigma_f^2), \log(\sigma_t^2)\right).
\end{aligned}
$$

Nous exprimons $\mathrm{var}(\alpha)$ comme fonction de $\mathrm{var}\left(\log(\sigma_f^2)\right)$ $\mathrm{var}\left(\log(\sigma_t^2)\right)$, et $\mathrm{cov}\left(\log(\sigma_f^2), \log(\sigma_t^2)\right)$, en appliquant une transformation logarithmique pour σ_t^2 et σ_f^2 afin de stabiliser leurs variances. On peut supposer que la covariance entre $\log(\sigma_f^2)$ et $\log(\sigma_t^2)$ est 0 lorsque les effets aléatoires de l'intervieweur pour le mode en personne et le mode par téléphone ne sont pas corrélés, comme c'est le cas dans l'étude de l'Arab Barometer. En revanche, dans la HRS, lorsque les effets aléatoires de l'intervieweur sont corrélés entre les modes, la covariance entre les deux estimations doit être prise en compte lors du calcul de $\mathrm{var}(\alpha)$.

Annexe B

Résultats complets des variances de l'intervieweur dans la Health and Retirement Study

Tableau B.1
Variances de l'intervieweur par mode pour certains éléments de la Health and Retirement Study en tenant compte des covariables.

Questions	Vraisemblance					Bayésienne				
	σ_f^2	σ_t^2	α	β_1	ρ	σ_f^2	σ_t^2	α	β_1	ρ
Questions du CES-D										
Se sentait déprimé	0,011	0,014	-0,095	0,056	0,222	0,011	0,013	0,044	**0,056**	0,07
	(0,006)	(0,008)	(0,389)	(0,029)	(0,448)	(0,006)	(0,009)	(0,643)	(0,029)	(0,391)
	[0,004;	[0,004;	[-0,857;	[-0,001;	[-0,603;	[0;	[0;	[-1,148;	[0,005;	[-0,551;
	0,031]	0,045]	0,667]	0,113]	0,818]	0,022]	0,03]	1,533]	0,114]	0,874]
Tout représentait un effort	0,022	0,004	0,893	**0,116**	-0,099	0,025	0,007	0,746	**0,118**	-0,128
	(0,006)	(0,005)	(0,682)	(0,025)	(0,618)	(0,014)	(0,005)	(0,38)	(0,029)	(0,264)
	[0,013;	[0;	[-0,445;	[0,066;	[-0,867;	[0,013;	[0,001;	[-0,002;	[0,071;	[-0,56;
	0,037]	0,052]	2,23]	0,165]	0,809]	0,037]	0,016]	1,496]	0,175]	0,254]
Sommeil agité	0,003	0,004	-0,13	**0,057**	-0,698	0,002	0,005	-0,486	**0,053**	0,337
	(0,003)	(0,004)	(0,719)	(0,022)	(1,018)	(0,002)	(0,004)	(0,754)	(0,021)	(0,312)
	[0;	[0;	[-1,54;	[0,013;	[-1;	[0;	[0;	[-1,89;	[0,011;	[-0,162;
	0,02]	0,032]	1,279]	0,1]	0,995]	0,007]	0,012]	0,925]	0,095]	0,849]
Heureux	0,007	0,010	-0,253	0,033	NA	0,011	0,011	0,128	0,032	-0,518
	(0,015)	(0,014)	(1,519)	(0,032)	(NA)	(0,005)	(0,007)	(0,539)	(0,027)	(0,314)
	[0;	[0;	[-2,348;	[-0,019;	[NA;	[0,003;	[0;	[-0,889;	[-0,024;	[-0,989;
	0,019]	0,023]	2,915]	0,085]	NA]	0,021]	0,022]	1,333]	0,083]	-0,006]
Se sentait seul	0,006	0,004	0,223	0,046	-0,208	0,006	0,006	0,178	0,048	0,055
	(0,004)	(0,006)	(0,973)	(0,026)	(1,053)	(0,004)	(0,005)	(0,878)	(0,028)	(0,084)
	[0,001;	[0;	[-1,685;	[-0,004;	[-0,983;	[0;	[0;	[-1,455;	[-0,005;	[-0,108;
	0,025]	0,12]	2,131]	0,097]	0,96]	0,014]	0,016]	1,846]	0,099]	0,218]
Profitait de la vie	0,009	0,011	-0,124	**0,07**	-0,823	0,01	0,007	0,551	0,061	**0,56**
	(0,006)	(0,009)	(0,528)	(0,033)	(0,718)	(0,006)	(0,009)	(0,944)	(0,036)	(0,187)
	[0,002;	[0,002;	[-1,16;	[0,006;	[-1;	[0,001;	[0;	[-1,096;	[-0,005;	[0,223;
	0,037]	0,052]	0,911]	0,134]	0,997]	0,021]	0,025]	2,148]	0,134]	0,921]
Se sentait triste	0,03	0	2,475	0,047	NA	0,032	0,003	**1,694**	0,046	**0,296**
	(0,007)	(0,001)	(2,213)	(0,025)	(NA)	(0,008)	(0,003)	(0,951)	(0,027)	(0,137)
	[0,018;	[0;	[-1,863;	[-0,003;	[NA;	[0,018;	[0;	[0,463;	[-0,01;	[0,037;
	0,048]	1,263]	6,813]	0,097]	NA]	0,048]	0,009]	3,775]	0,097]	0,577]
Ne pouvait pas se motiver	0,016	0,019	-0,098	0,05	0,314	0,02	0,02	-0,051	0,051	0,274
	(0,005)	(0,008)	(0,274)	(0,027)	(0,318)	(0,032)	(0,008)	(0,339)	(0,033)	(0,307)
	[0,008;	[0,008;	[-0,635;	[-0,003;	[-0,351;	[0,007;	[0,006;	[-0,732;	[-0,01;	[-0,346;
	0,03]	0,044]	0,439]	0,103]	0,768]	0,029]	0,035]	0,515]	0,109]	0,797]
Indicateur global	0,012	0,012	0,006	**0,152**	-0,264	0,016	0,012	0,093	**0,15**	0,244
	(0,005)	(0,007)	(0,37)	(0,028)	(0,428)	(0,028)	(0,008)	(0,487)	(0,027)	(0,211)
	[0,005;	[0,004;	[-0,718;	[0,096;	[-0,825;	[0,002;	[0,001;	[-0,901;	[0,102;	[-0,18;
	0,029]	0,04]	0,731]	0,207]	0,558]	0,024]	0,027]	1,075]	0,207]	0,573]

Notes : β_1 représente les effets du mode d'enquête dans les moyennes, calculés comme la moyenne de l'estimation en personne moins la moyenne de l'estimation par téléphone. σ_f^2 représente les variances de l'intervieweur pour le mode en personne. σ_t^2 est la variance de l'intervieweur pour le mode par téléphone. α désigne les différences de logarithme entre les variations de l'intervieweur pour le mode en personne et le mode par téléphone. ρ est la corrélation entre les effets aléatoires de l'intervieweur pour le mode en personne et pour le mode par téléphone. Nous utilisons S.O. pour masquer des estimations qui ne peuvent pas être estimées en raison de difficultés numériques. CES-D = Center for Epidemiological Studies-Depression.

Tableau B.1(suite)
Variances de l'intervieweur par mode pour certains éléments de la Health and Retirement Study en tenant compte des covariables.

Questions	Vraisemblance					Bayésienne				
	σ_f^2	σ_t^2	α	β_1	ρ	σ_f^2	σ_t^2	α	β_1	ρ
Observations de l'intervieweur										
Attentif	0,29	0,342	-0,082	0,013	**0,893**	0,298	0,351	-0,081	0,018	**0,878**
	(0,032)	(0,043)	(0,063)	(0,035)	(0,036)	(0,032)	(0,044)	(0,062)	(0,035)	(0,039)
	[0,233;	[0,268;	[-0,205;	[-0,056;	[0,795;	[0,233;	[0,262;	[-0,197;	[-0,049;	[0,803;
	0,361]	0,438]	0,041]	0,081]	0,946]	0,356]	0,431]	0,038]	0,088]	0,955]
Compréhension	0,408	0,461	-0,062	-0,003	**0,921**	0,413	0,465	-0,058	0	**0,91**
	(0,04)	(0,05)	(0,049)	(0,032)	(0,023)	(0,039)	(0,051)	(0,049)	(0,032)	(0,041)
	[0,336;	[0,373;	[-0,158;	[-0,066;	[0,86;	[0,341;	[0,366;	[-0,149;	[-0,064;	[0,861;
	0,494]	0,571]	0,033]	0,059]	0,956]	0,493]	0,56]	0,043]	0,061]	0,958]
Coopération	0,45	0,404	0,053	0,174	**0,941**	0,459	0,41	0,057	0,178	**0,931**
	(0,043)	(0,044)	(0,047)	(0,031)	(0,021)	(0,047)	(0,048)	(0,047)	(0,032)	(0,025)
	[0,373;	[0,327;	[-0,039;	[0,113;	[0,884;	[0,378;	[0,321;	[-0,039;	[0,108;	[0,881;
	0,542]	0,5]	0,145]	0,234]	0,971]	0,556]	0,51]	0,138]	0,236]	0,971]
Se souvenir	0,482	0,593	**-0,103**	**-0,065**	**0,941**	0,483	0,605	**-0,112**	**-0,062**	**0,931**
	(0,047)	(0,065)	(0,047)	(0,032)	(0,019)	(0,047)	(0,059)	(0,047)	(0,033)	(0,029)
	[0,398;	[0,478;	[-0,195;	[-0,128;	[0,89;	[0,392;	[0,489;	[-0,205;	[-0,124;	[0,885;
	0,584]	0,735]	-0,011]	-0,001]	0,969]	0,574]	0,721]	-0,028]	0,002]	0,972]
Entendre	0,27	0,372	**-0,161**	**0,152**	**0,888**	0,271	0,375	**-0,161**	**0,151**	**0,87**
	(0,03)	(0,046)	(0,063)	(0,034)	(0,035)	(0,032)	(0,048)	(0,065)	(0,038)	(0,064)
	[0,217;	[0,291;	[-0,285;	[0,084;	[0,796;	[0,212;	[0,274;	[-0,284;	[0,084;	[0,795;
	0,336]	0,476]	-0,038]	0,219]	0,94]	0,335]	0,462]	-0,037]	0,229]	0,947]
Qualité globale	0,879	0,782	0,058	**0,09**	**0,96**	0,881	0,788	0,057	**0,086**	**0,94**
	(0,08)	(0,079)	(0,04)	(0,034)	(0,014)	(0,077)	(0,08)	(0,046)	(0,038)	(0,088)
	[0,736;	[0,642;	[-0,019;	[0,023;	[0,923;	[0,749;	[0,641;	[-0,032;	[0,014;	[0,913;
	1,051]	0,953]	0,136]	0,156]	0,98]	1,04]	0,949]	0,14]	0,158]	0,983]
Activité physique										
Sports vigoureux	0,015	0,008	0,298	-0,036	0,642	0,017	0,007	0,523	-0,037	0,36
	(0,004)	(0,006)	(0,381)	(0,022)	(0,41)	(0,011)	(0,004)	(0,406)	(0,026)	(0,372)
	[0,008;	[0,002;	[-0,45;	[-0,079;	[-0,541;	[0,007;	[0;	[-0,209;	[-0,081;	[-0,446;
	0,027]	0,032]	1,045]	0,007]	0,972]	0,026]	0,015]	1,45]	0,014]	0,827]
Sports modérément exigeants	0,013	0,015	-0,043	0,028	0,478	0,015	0,019	-0,086	0,031	0,233
	(0,004)	(0,006)	(0,274)	(0,023)	(0,326)	(0,008)	(0,008)	(0,28)	(0,025)	(0,248)
	[0,007;	[0,006;	[-0,581;	[-0,017;	[-0,297;	[0,006;	[0,004;	[-0,655;	[-0,019;	[-0,351;
	0,026]	0,035]	0,494]	0,073]	0,873]	0,024]	0,033]	0,464]	0,078]	0,698]
Sports légèrement exigeants	0,015	0,03	-0,353	**0,135**	0,107	0,02	0,031	-0,355	**0,134**	0,144
	(0,005)	(0,009)	(0,238)	(0,028)	(0,278)	(0,042)	(0,01)	(0,361)	(0,028)	(0,29)
	[0,007;	[0,016;	[-0,819;	[0,08;	[-0,416;	[0,002;	[0,014;	[-1,097;	[0,073;	[-0,264;
	0,03]	0,056]	0,114]	0,19]	0,577]	0,03]	0,052]	0,324]	0,184]	0,962]

Notes : β_1 représente les effets du mode d'enquête dans les moyennes, calculés comme la moyenne de l'estimation en personne moins la moyenne de l'estimation par téléphone. σ_f^2 représente les variances de l'intervieweur pour le mode en personne. σ_t^2 est la variance de l'intervieweur pour le mode par téléphone. α désigne les différences de logarithme entre les variations de l'intervieweur pour le mode en personne et le mode par téléphone. ρ est la corrélation entre les effets aléatoires de l'intervieweur pour le mode en personne et pour le mode par téléphone. Nous utilisons S.O. pour masquer des estimations qui ne peuvent pas être estimées en raison de difficultés numériques. CES-D = Center for Epidemiological Studies-Depression.

Bibliographie

Ehrlich, J.S., et Riesman, D. (1961). Age and authority in the interview. *Public Opinion Quarterly*, 39-56.

Elliott, M.N., Zaslavsky, A.M. Goldstein, E., Lehrman, W., Hambarsoomians, K., Beckett, M.K. et Giordano, L. (2009). Effects of survey mode, patient mix, and nonresponse on CAHPS hospital survey scores. *Health Services Research*, 44, 501-518.

Fisher, G.G., et Ryan, L.H. (2018). Overview of the health and retirement study and introduction to the special issue. *Work, Aging and Retirement*, 4, 1-9.

Groves, R.M., et Kahn, R.L. (1979). *Surveys by Telephone: A National Comparison with Personal Interviews*.

Groves, R.M., et Magilavy, L.J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly*, 50, 251-266.

Hanson, R.H., et Marks, E.S. (1958). Influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association*, 53, 635-655.

Heeringa, S.G., et Connor, J.H. (1995). Technical description of the health and retirement survey sample design. *Ann Arbor: University of Michigan*.

Holbrook, A.L., et Krosnick, J.A. (2010). Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly*, 74, 37-67.

Health and Retirement Study (HRS), Staff (2023). HRS core interview sample sizes and response rates. Rapport technique, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI. Disponible en ligne.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.

Kolenikov, S., et Kennedy, C. (2014). Evaluating three approaches to statistically adjust for mode effects. *Journal of Survey Statistics and Methodology*, 2, 126-158.

Kreuter, F., Presser, S. et Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and WEB surveys the effects of mode and question sensitivity. *Public Opinion Quarterly*, 72, 847-865.

Schaeffer, N.C., Dykema, J. et Maynard, D.W. (2010). Interviewers and interviewing. *Handbook of Survey Research*, 2, 437-471.

Schnell, R., et Kreuter, F. (2003). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 21, 3, 389-410.

Schuman, H., et Converse, J.M. (1971). The effects of black and white interviewers on black responses in 1968. *Public Opinion Quarterly*, 35, 44-68.

Suzer-Gurtekin, Z.T., Heeringa, S.G. et Valliant, R. (2013). Investigating the bias of alternative statistical inference methods in mixed-mode surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3393-3407.

Tourangeau, R., et Smith, T.W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, 60, 275-304.

Tucker, C. (1983). Interviewer effects in telephone surveys. *Public Opinion Quarterly*, 47, 84-95.

Vannieuwenhuyze, J.T.A. (2015). Mode effects on variances, covariances, standard deviations, and correlations. *Journal of Survey Statistics and Methodology*, 3, 3, 296-316.

West, B.T., et Blom, A.G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, 5, 175-211.

West, B.T., Kreuter, F. et Jaenichen, U. (2013). "Interviewer" effects in face-to-face surveys: A function of sampling, measurement error, or nonresponse? *Journal of Official Statistics*, 29, 277-297.

West, B.T., et Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74, 1004-1026.

West, B.T., Ong, A.R., Conrad, F.G., Schober, M.F., Larsen, K.M. et Hupp, A.L. (2022). Interviewer effects in live video and prerecorded video interviewing. *Journal of Survey Statistics and Methodology*, 10, 317-336.

Yu, W., Elliott, M.R. et Raghunathan, T.E. (2024). Investigating mode effects in interviewer variances using two representative multi-mode surveys. *arXiv preprint arXiv:2408.11874*.

Techniques d'enquête, décembre 2024
Vol. 50, nº 2, pp. 473-505
Statistique Canada, nº 12-001-X au catalogue

Plan d'enquête adaptatif robuste pour les variations temporelles des propensions à répondre dans un contexte multimodal

Shiya Wu, Harm-Jan Boonstra, Mirjam Moerbeek et Barry Schouten[1]

Résumé

Les plans d'enquête adaptatifs (PEA) permettent d'adapter les protocoles de recrutement aux sous-groupes de population qui présentent un intérêt pour une enquête. Ces dernières années, l'optimisation efficace d'un PEA a fait l'objet de recherches et de plusieurs applications. Toutefois, le rendement d'un PEA optimisé au fil du temps réagit aux variations temporelles des propensions à répondre. On ne comprend pas encore entièrement la façon dont les stratégies d'adaptation peuvent s'adapter à une telle variation au fil du temps. Dans la présente étude, nous proposons une approche d'optimisation robuste dans le contexte d'enquêtes séquentielles multimodales à l'aide d'une analyse bayésienne. Cette approche s'exprime sous la forme d'un problème de programmation mathématique qui tient explicitement compte de l'incertitude attribuable à la variation temporelle. Des décisions en matière de PEA peuvent alors être prises en tenant compte de la variation chronologique de la propension à répondre conditionnelle au mode et des corrélations de la propension à répondre entre les modes. La démonstration de cette approche fait appel à une étude de cas : l'Enquête sur la santé aux Pays-Bas de 2014 à 2017. Nous évaluons la sensibilité du rendement des PEA 1) au niveau budgétaire et 2) à la durée des données de série chronologique historiques applicables. Nous concluons que la dépendance au niveau budgétaire n'est que modérée et que la dépendance aux données historiques est tempérée par le degré de saisonnalité au cours de l'année.

Mots-clés : Analyse de séries chronologiques; approche bayésienne; modèle de propension à répondre; optimisation; plans d'enquête adaptatifs.

1. Introduction

Les plans d'enquête adaptatifs (PEA) (Wagner, 2008; Schouten, Peytchev et Wagner, 2017) sont progressivement devenus un choix viable dans les enquêtes contemporaines; on ne propose plus un seul protocole d'enquête pour toutes les personnes et tous les sous-groupes, mais on l'adapte afin d'obtenir des réponses des personnes de manière efficace, en fonction de caractéristiques de population connues et de caractéristiques observées au cours du travail sur le terrain. Cette transition a été accélérée par la baisse persistante des taux de réponse, les budgets limités, une plus grande variété de sources de données, l'émergence de toutes sortes d'appareils mobiles et la migration progressive vers des enquêtes multimodales. Cette évolution suppose une plus grande urgence et davantage d'options en ce qui concerne les plans d'enquête.

Un élément clé du PEA est la stratégie d'optimisation, c'est-à-dire l'ensemble de règles de décision. De telles stratégies reposent sur les intrants de propensions à répondre et d'autres paramètres du plan d'enquête. Parmi les principales approches en matière d'optimisation figurent la priorisation des cas (Peytchev, Riley, Rosen, Murphy et Lindblad, 2010; Wagner, 2013; Wagner et Hubbard, 2013), l'approche par essais et erreurs, ainsi que l'optimisation mathématique et statistique (van Berkel, van der Doef et Schouten, 2020;

1. Shiya Wu, Utrecht University, Department of Methodology and Statistics; Harm-Jan Boonstra, Statistics Netherlands, Department of Statistical Methods; Mirjam Moerbeek, Utrecht University, Department of Methodology and Statistics; Barry Schouten, Statistics Netherlands, Department of Statistical Methods and Utrecht University, Department of Methodology and Statistics. Courriel : jg.schouten@cbs.nl.

Calinescu, Bhulai et Schouten, 2013; Schouten, Calinescu et Luiten, 2013); voir Schouten et coll. pour connaître les avantages et les inconvénients de chaque approche. Cependant, dans le cadre de ces contributions, l'inexactitude des propensions à répondre estimées à partir des données historiques a été la plupart du temps ignorée. Dans la programmation mathématique, les objectifs peuvent être paramétrés sous forme de fonctions de propension à répondre agissant comme l'un des principaux intrants de l'optimisation. Une erreur serait introduite dans la prise de décisions en cas de variation des réelles propensions à répondre au fil du temps, lorsque cette variation n'est pas prise en compte dans l'estimation. Par conséquent, cette inexactitude rend tout PEA sous-optimal, ou, pire, inefficace. Placer l'optimisation du PEA dans un contexte bayésien est naturel pour répondre à cet enjeu; pourtant, les études pertinentes sur cette méthodologie d'enquête en sont toujours à leurs débuts. Récemment, Ma (2021) a élaboré une méthodologie pour optimiser efficacement une stratification en tenant compte de l'exactitude des estimations des propensions à répondre de façon bayésienne, en fonction des données historiques les plus récentes.

La variation temporelle des propensions à répondre et des estimations inexactes de données historiques compromet la robustesse de l'optimisation des PEA; voir Schouten et coll. (2017) ainsi que Chun, Heeringa et Schouten (2018) pour une analyse plus approfondie. Récemment, les travaux de recherche en matière de PEA ont commencé à mettre l'accent sur l'élaboration de modèles de propension à répondre et l'amélioration de l'exactitude des prédictions; Schouten, Mushkudiani, Shlomo, Durrant, Lundquist et Wagner (2018) ont été les premiers à élaborer des méthodes de mise à jour bayésiennes afin de contrer ce biais en utilisant statistiquement les données d'enquête accumulées et les données historiques générées à partir de mises en œuvre passées d'une même enquête. En étant les plus instructives, les croyances *a priori* découlant de données d'enquête antérieures peuvent améliorer les données actuelles à des fins de prédiction. Manifestement, traduire des sources de données externes en croyances *a priori* est nécessaire pour élaborer des modèles de propension à répondre. Pour ce faire, l'analyse documentaire (West, Wagner, Coffey et Elliott, 2023) et l'obtention de connaissances spécialisées (Coffey, West, Wagner et Elliott, 2020; Wu, Schouten, Meijers et Schouten, 2022) sont des approches récentes permettant d'obtenir des renseignements *a priori*. Les spécialistes de la recherche sur les enquêtes traitent le sujet de l'actualité des données historiques de façon incomplète et considèrent globalement les propensions à répondre à différentes étapes de l'enquête, alors que des faits, notamment des taux de réponse constamment en baisse au fil des ans, indiquent que l'exactitude des estimations des propensions à répondre dépend du temps et que les propensions à répondre à des enquêtes séquentielles sont probablement corrélées.

Les travaux les plus pertinents menés par Wu, Boonstra, Moerbeek et Schouten (2023) ont porté sur la décomposition de la variation temporelle des propensions à répondre à de multiples niveaux afin d'étudier l'influence de la durée des données d'enquête historiques applicables sur l'exactitude des prédictions de la propension à répondre. Dans ce cas précis, seule l'étape de collecte de données des interviews Web assistées par ordinateur (IWAO) dans le cadre de l'Enquête sur la santé aux Pays-Bas (GEZO) a été prise en compte et non l'étape d'interview sur place assistée par ordinateur (IPAO) applicable aux non-répondants aux IWAO. Dans la présente étude, nous généralisons l'élaboration du modèle de prédiction de la propension à

répondre aux multiples étapes de la collecte de données, en particulier au cas des étapes de collecte de données par IWAO et IPAO de l'enquête GEZO. Cela permet, par exemple, d'évaluer l'exactitude conditionnelle des prédictions des propensions à répondre par IPAO étant donné la réalisation de la réponse par IWAO sur une certaine période. Comme dans Wu et coll. (2023), nous adoptons une approche bayésienne permettant une quantification complète de l'incertitude des propensions à répondre et des quantités dérivées. La deuxième extension de la présente étude par rapport à Wu et coll. (2023) est l'analyse du rendement des PEA selon diverses contraintes externes, en tenant compte de l'incertitude des prédictions des propensions à répondre.

Dans l'ensemble, la présente étude vise deux contributions aux modèles séquentiels multimodaux (MM) : prédire aussi exactement que possible la propension à répondre de chaque mode d'enquête et prendre des décisions adaptatives de manière aussi optimale que possible. Pour satisfaire cette ambition en utilisant des données de séries chronologiques historiques dans l'évaluation, nous posons trois questions de recherche :

- Comment construire des modèles de série chronologique pour améliorer l'exactitude de la prédiction de la propension à répondre dans un plan séquentiel multimodal ?
- Dans quelle mesure le rendement des PEA est-il sensible au niveau budgétaire donné ?
- Dans quelle mesure le rendement des PEA dépend-il de la durée des données historiques applicables ?

En réponse à la première question, nous étendons les modèles hiérarchiques binomiaux de série chronologique pour un seul mode proposés par Wu et coll. (2023) à des modèles de série chronologique hiérarchiques multinomiaux pour de multiples modes, et nous illustrons cela au moyen d'une application à l'enquête GEZO, pour laquelle nous utilisons des données relatives à la période de 2014 à 2017. Cette extension tient également compte de l'intégration de paramètres de corrélation entre les modes à la modélisation des propensions à répondre des modes d'IWAO et d'IPAO.

Pour ce qui est des deuxième et troisième questions, une stratégie tenant compte de l'incertitude quant aux intrants de l'optimisation lors de l'optimisation de répartitions probabilistes est très recherchée. Le rendement du plan d'enquête est suivi en évaluant la représentation de caractéristiques contextuelles pertinentes couplées au moyen de données administratives. Cette représentation est opérationnalisée par le coefficient de variation (c.v.) des propensions à répondre (Schouten, Cobben et Bethlehem, 2009). Nous étalonnons le rendement des PEA par rapport au rendement des plans par IWAO uniquement et des plans non adaptatifs pour veiller à ce que les répartitions déterminées améliorent le rendement des PEA. Pour déterminer la sensibilité du rendement des PEA, nous menons deux expériences. Dans la première expérience, nous sélectionnons un trimestre de l'année et diminuons progressivement le niveau budgétaire. Cette expérience nous permet d'étudier la sensibilité du rendement des PEA aux contraintes budgétaires. Dans la deuxième expérience, le niveau budgétaire est fixe et la fenêtre de la série chronologique des données historiques se déplace vers le nouveau trimestre suivant de collecte de données. Des données

historiques sont d'abord utilisées pour établir une probabilité *a priori*. Ensuite, au fil des nouveaux trimestres, la probabilité *a priori* est mise à jour de manière répétée en une probabilité *a posteriori* servant de probabilité *a priori* pour le trimestre suivant. Nous évaluons la façon dont la probabilité *a priori* varie ainsi que les effets que cela a sur les répartitions des unités d'échantillonnage.

La présente étude s'articule comme suit : nous commençons par construire le modèle de série chronologique pour les plans séquentiels multimodaux dans la deuxième section. Dans la troisième section, nous décrivons le problème d'optimisation. Nous introduisons l'étude de cas dans la quatrième section et répondons aux questions de recherche. Dans la dernière section, nous traitons des avantages et des inconvénients de notre méthode et concluons avec des idées de travaux de recherche futurs.

2. Méthodes

Dans la présente section, un modèle de série chronologique multivarié est élaboré pour les propensions à répondre de plans séquentiels multimodaux. Nous élargissons le modèle hiérarchique de série chronologique suggéré par Wu et coll. (2023) en introduisant des propensions conditionnelles à répondre de modes de suivi.

2.1 Modélisation des propensions à répondre dans des plans d'enquête séquentiels multimodaux

Le modèle de série chronologique de Wu et coll. (2023) produit des estimations précises des propensions à répondre pour des plans d'enquête à un seul mode, ou pour le premier mode des enquêtes multimodales au cours du travail sur le terrain. Ici, l'objectif évolue pour permettre des prédictions fiables pour chaque mode d'enquêtes multimodales afin d'étendre l'attrait du modèle. On considère notamment des données de série chronologique à valeurs discrètes, y compris la taille d'un échantillon et le nombre de répondants pour chaque mode, d'une répartition multinomiale, alors que Wu et coll. (2023) considéraient une répartition binomiale des données.

La propension à répondre (PR) est la propension théorique d'un sujet échantillonné à répondre pour un mode d'interview particulier selon un ensemble de caractéristiques connues. Ces caractéristiques peuvent comprendre des paradonnées recueillies à une étape particulière ou un mode particulier de l'enquête. Ce sujet peut être soit une personne, soit un groupe de personnes bien défini. Il convient de mentionner qu'un groupe est établi en recoupant plusieurs variables auxiliaires considérées comme étant des prédicteurs puissants des variables d'enquête. Au sein de ce groupe, les unités présentent des attributs démographiques homogènes, comme l'âge. Un tel ensemble de groupes peut varier au fil du temps ou des variations de conception (voir Schouten et coll., 2017 pour en savoir plus sur la stratification), mais nous supposons ici que la stratification est constante et précisée avant l'ajustement du modèle.

Pour modéliser la propension à répondre au niveau du mode, nous supprimons, dans la présente section, les indices indiquant un groupe particulier dans le paramètre de propension ainsi qu'un moment particulier; toutefois, dans la section suivante, nous devons préciser cet indice pour décomposer une série chronologique en effets fixes ou aléatoires au niveau de la strate, du temps et/ou du mode.

Supposons qu'une enquête multimodale comprend $M-1$ modes de collecte de données. Nous ajoutons un M^{e} « mode » correspondant à la non-réponse, c'est-à-dire la catégorie représentant une non-réponse aux $M-1$ modes. Soit un échantillon aléatoire de taille n connu avant le début de la collecte de données et r_j, le nombre observé de répondants pour le j^{e} mode, où $j \in \{1, \ldots, M\}$. Supposons une répartition multinomiale en M modes avec une propension à répondre ρ_j pour le j^{e} mode, où $\rho_j \in [0,1]$. ρ_M est la propension de non-réponse; toutefois, celle-ci n'est plus explicitement modélisée plus tard.

Le vecteur $\mathbf{r} = (r_1, \ldots, r_M)$ suit une répartition multinomiale dont la taille d'échantillon est n et la propension à répondre est $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_M)$, c'est-à-dire la répartition conjointe de $\mathbf{r}$ est une généralisation multivariée d'une répartition binomiale,

$$\text{mult}(\mathbf{r} \mid n, \boldsymbol{\rho}) = \frac{n!}{\prod_{j=1}^{M} r_j!} \prod_{j=1}^{M} \rho_j^{r_j}. \tag{2.1}$$

Linderman, Johnson et Adams (2015) ont eu recours à une transformation par rupture de bâton pour reformuler la répartition multinomiale en un produit de répartitions binomiales, au sein de laquelle les paramètres créés sont dépendants. Cela a donné l'occasion de reformuler l'équation (2.1) à m dimensions de façon récursive en termes de $M-1$ binômes. Dans la représentation par rupture de bâton, le vecteur de propension $\boldsymbol{\rho}$ sert de bâton divisé de manière récursive en deux morceaux pour créer une variable binomiale $\tilde{\boldsymbol{\rho}} = (\tilde{\rho}_1, \ldots, \tilde{\rho}_{M-1})$. Pour fournir une dérivation, posons que la variable r_j de réponse pour le j^{e} mode suit une densité binomiale avec les paramètres n_j et $\tilde{\rho}_j$, c'est-à-dire $\text{bin}(r_j \mid n_j, \tilde{\rho}_j)$, où n_j et $\tilde{\rho}_j$ représentent la taille restante de l'échantillon et la fraction de la probabilité restante approchée par le mode j^{e},

$$n_j = n - \sum_{k<j} r_k, \tag{2.2}$$

$$\tilde{\rho}_j = \frac{\rho_j}{1 - \sum_{k<j} \rho_k}, \tag{2.3}$$

où $j \in \{2, \ldots, M\}$. Lorsque $j = 1$, le paramètre $n = n_1 = \sum_{j \in \{1, \ldots, M\}} r_j$ et le paramètre $\tilde{\rho}_j = \rho_1$. À l'aide de (2.3), nous obtenons

$$\tilde{\rho}_j^{r_j} (1 - \tilde{\rho}_j)^{n_j - r_j} = \rho_j^{r_j} \frac{1}{\left(1 - \sum_{k<j} \rho_k\right)^{n_j}} \left(1 - \sum_{k \le j} \rho_k\right)^{n_j - r_j}. \tag{2.4}$$

Notons que la somme de r_j est n sur j et $n_j = n_{j-1} - r_{j-1}$ pour tout $j \in \{2, \ldots, M\}$. Cela signifie que les paires de termes $\left(1 - \sum_{k<j} \rho_k\right)^{n_j}$ s'annuleront dans le produit de (2.4) sur j, menant au même format que le terme exponentiel multinomial dans (2.1),

$$\prod_{j=1}^{M-1} \tilde{\rho}_j^{r_j} \, (1-\tilde{\rho}_j)^{n_j - r_j} = \prod_{j=1}^{M} \rho_j^{r_j}. \tag{2.5}$$

Les constantes de normalisation suivent le même raisonnement. Combinée aux termes exponentiels dans (2.5), l'équation (2.1) peut être réécrite sous la forme

$$\text{mult}(\mathbf{r} \mid n, \boldsymbol{\rho}) = \prod_{j=1}^{M-1} \text{bin}(r_j \mid n_j, \tilde{\rho}_j). \tag{2.6}$$

Nous utilisons la représentation par rupture de bâton du modèle multinomial à des fins pratiques : pour cette représentation, il existe un échantillonneur de Gibbs simple et efficace pour le modèle (hiérarchique) multinomial. Comme l'expliquent Linderman et coll. (2015), ils emploient la même méthodologie d'augmentation de données Polya-Gamma (Polson, Scott et Windle, 2013) utilisée pour les modèles binomiaux dans Wu et coll. (2023). La représentation par rupture de bâton permet d'échantillonner les coefficients du modèle pour toutes les $M-1$ catégories modélisées dans un bloc, améliorant ainsi la convergence de l'échantillonneur de Gibbs. Cette représentation présente également un inconvénient : la définition de $\tilde{\rho}_j$ rend l'interprétation des coefficients de corrélation sous-jacents plus difficile, en particulier l'interprétation des coefficients de corrélation dans les modèles décrits en détail dans la section 2.2.

Dans la section 2.2, un modèle de série chronologique structurel est employé pour décomposer une série chronologique observée en des composantes temporelles sous-jacentes.

2.2 Modèle hiérarchique multinomial de série chronologique

Pour mesurer la dépendance des propensions à répondre parmi les modes, nous élargissons les modèles de Wu et coll. (2023) en introduisant un nouveau paramètre hiérarchique qui désigne les coefficients de corrélation. Une telle dépendance répartie sur les composantes d'intérêt de la série chronologique est similaire à ce qu'ont adopté Wu et coll. (2023); il est suggéré de consulter cette étude pour obtenir de plus amples précisions sur chaque définition de composante ainsi que des renseignements techniques.

Pour décrire chaque composante du modèle au niveau le plus détaillé, posons que le vecteur de paramètre de propension dépendante des modes séquentiels est associé à une strate de PEA particulière et à un moment particulier, c'est-à-dire $\tilde{\boldsymbol{\rho}}_{g,t} = \{\tilde{\rho}_{g,t,j} \mid j \in \{1, \ldots, M-1\}\}$, où la j^{e} entrée désigne le paramètre de propension du j^{e} mode dans la strate de PEA g au moment t, comme le définit (2.3). Les nombres de strates de PEA, de moments et de modes d'enquête sont respectivement G, T et $M-1$. Il convient de mentionner que dans la présente étude, nous utilisons le terme « strate de PEA » pour désigner un groupe de population pouvant recevoir un traitement différent. Cela ne doit pas être confondu avec des strates d'échantillon recevant des probabilités d'inclusion différentes. Nous omettons la référence au PEA dans les strates de PEA dans la majeure partie de ce qui suit.

Dans la présente étude, nous supposons que le choix des strates de PEA a déjà été effectué et que la sélection de variables elle-même ne fait pas partie de la stratégie de modélisation. Les strates seront fondées

sur des renseignements auxiliaires prédictifs des variables d'intérêt de l'enquête ou de la non-réponse à l'enquête. Toutefois, en cas de volume important de renseignements auxiliaires, une stratégie de sélection de variables devrait être intégrée à l'ajustement du modèle et à la stratégie d'optimisation du PEA. Nous revisitons cet enjeu dans la section sur l'analyse.

Soit $\theta_{g,t,j} = \text{logit}(\tilde{\rho}_{g,t,j})$. Une fonction de lien logit est utilisée pour convertir l'échelle limitée d'une probabilité à l'échelle non limitée d'un prédicteur linéaire, c'est-à-dire une combinaison linéaire de composantes de modèle. Soit $\theta_{g,t,j} = \text{logit}(\tilde{\rho}_{g,t,j})$, le prédicteur linéaire. Par conséquent, la fonction de probabilité multinomiale dans l'équation (2.1) peut être reformulée en substituant la transformation inverse à $\tilde{\rho}_{g,t,j}$,

$$\text{mult}(\mathbf{r} \mid n, \boldsymbol{\rho}) \propto \prod_{j=1}^{M-1} \left(\frac{e^{\theta_{g,t,j}}}{1+e^{\theta_{g,t,j}}} \right)^{r_j}. \tag{2.7}$$

Les modèles hiérarchiques considérés pour modéliser le prédicteur linéaire $\theta_{g,t,j}$ prennent la forme générale d'une décomposition additive, qui fait référence à une fonction de la somme des composantes de la série chronologique. Donc,

$$\theta_{g,t,j} = \beta_j + \beta'_{xj} x_g + \delta'_s s_t + u_{t,j} + v_{g,j} + z_{g,t,j} + e_{g,t,j}, \tag{2.8}$$

dont les trois premiers et les quatre derniers termes sont respectivement modélisés en tant qu'effets fixes et effets aléatoires.

Les premiers effets fixes de régression β_j sont des ordonnées à l'origine propres au mode, mesurant l'effet principal sur $\theta_{g,t,j}$. Les deuxièmes effets fixes β_{xj} sont des coefficients de régression propres au mode associés à la covariable x_g de vecteur p, correspondant aux caractéristiques démographiques particulières associées à la strate g. Les troisièmes effets fixes δ_s sont des coefficients de régression propres à la saison associés à la variable d'indicateur de saison s_t de vecteur q. Voir la définition des strates et des saisons à l'annexe A. Actuellement, les effets de la saison et du mode sont communs à toutes les strates. Plus globalement, ces effets fixes peuvent être propres à la strate. Dans l'application présente tout au long de cette étude, x_g et s_t sont des vecteurs binaires correspondant à des variables catégoriques, mais des variables ordinales ou numériques, voire variant dans le temps, peuvent également être prises en compte, le cas échéant.

Chaque terme d'effet aléatoire dans l'équation (2.8) permet implicitement la corrélation entre les modes d'enquête. Veuillez consulter l'étude de Wu et coll. (2023) pour obtenir une description des composantes d'effet aléatoire. Comme nous l'avons souligné, ces termes sont désormais croisés avec le mode, c'est-à-dire que des paramètres de variance distincts pour chaque mode et des paramètres de corrélation entre les modes sont introduits. La tendance temporelle globale $\mathbf{u}$, l'ordonnée à l'origine aléatoire pour la strate $\mathbf{v}_g$ et la tendance propre à la strate $\mathbf{z}_g$ respectent cette règle. Des effets aléatoires de bruit blanc $e_{g,t,j}$ sont également croisés avec le mode, mais nous utilisons un seul paramètre de variance commun pour tous les modes et aucune corrélation.

Nous adoptons une approche bayésienne pour estimer le modèle dans l'équation (2.8) et obtenir des prédictions fiables des propensions à répondre au niveau du mode, de la strate et du temps. Comme nous l'avons mentionné, les probabilités *a priori* sont les mêmes pour les coefficients correspondant aux différents modes. Pour simplifier la notation, nous supprimons les indices g, t et j dans chaque terme de composante de modèle. Les effets fixes β et δ se voient attribuer des probabilités *a priori* faiblement informatives à répartition normale de moyenne nulle et de matrice de variance diagonale, où l'erreur-type prend une valeur relativement grande de 10. On suppose que chaque vecteur d'effets aléatoires respecte la probabilité *a priori* normale multivariée de moyenne $\mathbf{0}$ et de matrice de covariance définie comme le produit de Kronecker de deux matrices de covariance $\mathbf{A}$ et $\mathbf{V}$. Ici, $\mathbf{V}$ est une matrice de covariance entièrement paramétrée, à laquelle est attribuée une loi *a priori* de Wishart inversée et mise à l'échelle (Gelman et Hill, 2007; O'Malley et Zaslavsky, 2008) et $\mathbf{A}$ est une matrice fixe, qui peut être une matrice diagonale simple pour des effets non structurés ou une matrice structurée correspondant à des marches aléatoires au fil du temps. De plus amples renseignements techniques relatifs à la spécification de la probabilité *a priori* et de la stratégie d'estimation, notamment l'échantillonneur de Gibbs et les répartitions conditionnelles complètes nécessaires, figurent dans les études de Boonstra et van den Brakel (2019) et de Wu et coll. (2023).

Un modèle plus parcimonieux peut être obtenu en omettant l'interaction axée sur le mode et en remplaçant la matrice de covariance entièrement paramétrée $\mathbf{V}$ par une matrice diagonale, si les effets entre les modes sur les prédictions de propension présentent peu d'intérêt. On appelle ce modèle le modèle sans corrélation; il fournit un modèle de base par rapport auquel on peut évaluer le rendement du modèle de corrélation complet.

L'estimation au moyen du modèle est effectuée à l'aide du paquet R *mcmcsae* (Boonstra, 2022). La convergence des résultats de la simulation de la méthode de Monte Carlo par chaîne de Markov (MCMC) est traitée en utilisant le tracé de trajectoire et la réduction scalaire potentielle ou diagnostic R-hat (Gelman et Rubin, 1992). Ces diagnostics montrent que l'échantillonneur de Gibbs converge rapidement, à la fois pour les modèles incluant et excluant des corrélations entre les modes. Dans une large mesure, cela est attribuable à l'échantillonnage de tous les effets fixes et aléatoires dans un seul bloc, ce qui est possible en raison de la représentation par rupture de bâton de la répartition multinomiale en combinaison avec l'augmentation de données Polya-Gamma, comme nous l'avons mentionné précédemment.

2.3 Extensions à des plans d'enquête complexes plus généraux

Le modèle présenté à la section 2.2 prend en charge des plans d'enquête d'échantillonnage aléatoire stratifié, mais pas d'autres caractéristiques d'échantillonnage complexe. Cela est justifiable pour l'application de l'enquête GEZO sur laquelle nous nous concentrons. Même si l'enquête GEZO repose sur un plan de sondage à deux degrés (les municipalités étant les unités d'échantillonnage de premier degré), les effets de catégorisation des résultats obtenus sont très mineurs, car la plupart des municipalités sont sélectionnées.

De plus, les probabilités d'inclusion aux premier et deuxième degrés sont telles que les probabilités d'inclusion générale sont égales pour toutes les personnes. La stratification utilisée dans la présente étude est choisie de sorte que les probabilités de réponse soient raisonnablement homogènes au sein d'une strate.

Nous allons maintenant décrire brièvement dans quelle mesure notre méthode peut être étendue pour prendre en charge les enquêtes à plans d'échantillonnage complexe plus généraux. Si les probabilités d'échantillonnage sont inégales, l'une peut, dans certains cas, tout de même définir une stratification de telle sorte que les probabilités soient (approximativement) égales au sein d'une strate. Si cela n'est pas possible, une analyse fondée sur un modèle au niveau de l'unité serait plus appropriée. Le modèle multinomial au niveau de la strate devient alors un modèle catégorique ou Multinoulli au niveau de la personne, c'est-à-dire le cas spécial de la répartition multinomiale avec $n_i = 1$ pour chaque personne i. Il convient de souligner que les dérivations relatives à la représentation par rupture de bâton de la section 2.2 sont toujours valides dans ce cas. Pour l'essentiel, les indices de strate g dans l'équation (2.8) deviendraient des indices de personne i.

Des covariables au niveau de l'unité expliquant la variation à la fois de l'échantillonnage et des probabilités de réponse devraient être incluses dans le modèle pour atténuer le biais autant que possible. En particulier, les probabilités d'inclusion elles-mêmes ou les variables sous-jacentes dont elles dépendent sont des covariables précieuses. Il peut également être utile de modéliser la dépendance envers les probabilités d'inclusion de manière souple; voir par exemple Chen, Elliott et Little (2010). Une stratification peut toujours être définie expressément aux fins du PEA, de sorte que les probabilités de réponse soient raisonnablement homogènes au sein d'une strate. Il n'est pas nécessaire qu'une telle stratification coïncide avec une stratification possible utilisée pour l'échantillonnage. Cela peut être géré par un modèle au niveau de l'unité en incluant des indicateurs de strate dans le modèle. De la même manière, pour un plan par grappe, le modèle au niveau de l'unité peut tenir compte des effets de grappe en incluant des indicateurs de grappe. Toutefois, puisque généralement seul un sous-échantillon de grappes est inclus dans l'échantillon et que le nombre d'observations par grappe observée peut être petit, les coefficients de grappe correspondants devraient être modélisés en tant qu'effets aléatoires (Scott et Smith, 1969). Enfin, nous soulignons que les extensions de modèle décrites pour prendre en charge des plans d'échantillonnage plus complexes peuvent être traitées à l'aide du même cadre d'échantillonneur de Gibbs employé dans l'application de l'enquête GEZO. En particulier, la même approche d'augmentation des données peut être utilisée pour la famille catégorique ou Multinoulli, puisqu'il s'agit d'un cas particulier de la répartition multinomiale. La seule différence est que l'on peut devoir intégrer d'autres covariables et leurs effets fixes correspondants, ainsi que des effets aléatoires supplémentaires pour les grappes, possiblement à diverses étapes. De tels modèles au niveau de l'unité peuvent être ajustés de la même façon, par exemple à l'aide du paquet R *mcmcsae*, même si les temps de calcul augmenteront du fait de la taille supérieure des données au niveau de l'unité et de la taille du modèle.

3. Optimisation de la répartition de modes dans le modèle de série chronologique hiérarchique bayésien

La présente section vise à étudier un problème de répartition rendant compte de l'incertitude en matière d'actualité et de mise en œuvre de PEA. La section 3.1 souligne les principaux éléments de construction et de fonctionnement de ce problème dans un cadre bayésien. Une stratégie est proposée à la section 3.2 pour évaluer l'intérêt de répartitions adaptatives par rapport à des répartitions non adaptatives en matière de réduction du risque de biais de non-réponse en suivant une mesure du risque de biais.

3.1 Principaux éléments

En général, une optimisation mathématique fait intervenir la sélection des « meilleures valeurs disponibles » d'une fonction objectif quelconque par rapport à un nombre de contraintes en choisissant des valeurs d'intrants parmi un ensemble autorisé. Établir des modèles d'optimisation fait intervenir trois éléments principaux : des variables de décision visant à optimiser l'objectif, des objectifs à minimiser ou à maximiser et des contraintes sur les variables de décision. Du fait de l'optimisation relativement au paramètre bayésien, nous insistons sur le fait que tous les paramètres statistiques mentionnés sont considérés comme des variables aléatoires ayant des valeurs variant dans le temps. Par conséquent, des fonctions objectifs et des fonctions de contraintes sont également des variables aléatoires. Dans ce qui suit, les principaux éléments sont d'abord présentés pour un cadre non bayésien et sont ensuite développés pour le cadre bayésien.

Les *variables de décision* sont des représentations symboliques d'une intervention décidée par le décideur. Elles représentent des parties inconnues d'une fonction objectif pouvant être manipulées et peuvent prendre toute valeur possible au sein d'un ensemble autorisé s'il est précisé. Dans la présente étude, une intervention est censée attribuer des modes d'interview à des strates lorsque les modes précédents ne permettent pas d'obtenir les données. Les variables de décision font par conséquent référence à des probabilités de répartition indiquant dans quelles mesures les non-répondants sont susceptibles d'être approchés par un mode de suivi. La probabilité de répartition $s_{g,t} \in [0,1]$ prend une décision sur la taille des candidats au suivi dans la strate g au moment t, où $s_{g,t} = 0$ signifie que la collecte de données pour la strate g cesse, et $s_{g,t} = 1$ signifie que tous les non-répondants de la strate g des modes précédents sont affectés au mode suivant.

La *fonction objectif* définit le critère permettant d'évaluer les valeurs des variables de décision candidates. Mis à part les variables de décision, elle dépend uniquement de quantités observées et estimées. Notre objectif d'optimisation est de minimiser le risque attendu de biais de non-réponse grâce à une répartition optimale. Puisque la non-réponse ne peut pas être observée directement, la présente étude tient compte d'un indicateur substitut du biais de non-réponse qui est une fonction des propensions à répondre. Nous employons le coefficient de variation (c.v.) des propensions à répondre (Schouten et coll., 2009). Le c.v. réel de la population limite le biais normalisé absolu des moyennes des répondants. Toutefois, le c.v.

estimé sur un ensemble particulier de variables auxiliaires respectera uniquement une partie de ce biais global. Même si d'autres indicateurs existent (voir Moore, Durrant et Smith (2018) ainsi que Nishimura, Wagner et Elliott (2016)), ce sont principalement les renseignements auxiliaires disponibles qui jouent un rôle décisif. Dans des enquêtes polyvalentes, la disponibilité peut ne pas être un réel enjeu, puisque l'on peut se concentrer sur une représentation générale et que toute amélioration sera utile. Dans des enquêtes ne comprenant que quelques statistiques clés, la disponibilité de variables auxiliaires pertinentes est essentielle.

Le c.v. est l'écart-type pondéré divisé par le taux de réponse pondéré :

$$\text{c.v.}(s,t) \;=\; \frac{\sqrt{\sum_g d_{g,t}\left(\rho_{g,t}-\overline{\rho}_t\right)^2}}{\overline{\rho}_t}. \tag{3.1}$$

Le poids $d_{g,t}$ est la proportion de l'échantillon de la taille de strate g au moment t par rapport à la taille globale au moment t, c'est-à-dire $d_{g,t} = n_{g,t} \Big/ \sum_g n_{g,t}$. Cette notation suppose implicitement que le plan de sondage donne lieu à des poids d'inclusion égaux; dans le cas contraire, les poids de sondage devraient alors être également intégrés. Cet ajout est simple, mais rend la notation difficile. La propension à répondre mixte $\rho_{g,t}$ désigne la propension générale pour les modes, qui est la somme de la propension marginale à répondre du mode de départ et des propensions conjointes à répondre du mode $j \geq 2$, en supposant que la strate g n'a pas répondu aux derniers $j-1$ modes. (Ici, nous supposons implicitement que tous les non-répondants d'un mode sont admissibles au suivi. En pratique, certains types de non-réponse, comme celle attribuable à une maladie physique ou mentale, peuvent ne pas être admissibles.)

$$\rho_{g,t} = \rho_{g,t,1} + \sum_{j\in\{2,\ldots,m-1\}} \rho_{g,t,j} \prod_{i\leq j-1} (1-\rho_{g,t,k}). \tag{3.2}$$

Les propensions individuelles (conditionnelles) $\rho_{g,t,j}$ pour un mode j sont estimées par des modèles multinomiaux dans la section 2. L'équation (3.2) suppose que tous les non-répondants aux modes précédents seront recrutés par le mode j pour une enquête non adaptative; cependant, cela peut être modifié en une enquête adaptative en réduisant la propension conjointe à la variable de décision $s_{g,t,j} \in [0,1]$; l'équation mise à jour devient donc

$$\rho_{g,t} = \rho_{g,t,1} + \sum_{j\in\{2,..,m-1\}} s_{g,t,j}\, \rho_{g,t,j} \prod_{i\leq j-1} (1-\rho_{g,t,k}). \tag{3.3}$$

Il est clair que l'équation (3.3) est équivalente à l'équation (3.2) lorsque tous les $s_{g,t,j} = 1$.

Les dénominateurs de l'équation (3.1), appelés les taux de réponse pondérés sur les strates, indiquent le niveau estimé de propensions inconnues, définies comme la somme pondérée des propensions mixtes de l'équation (3.2) et de l'équation (3.3). Nous disons que le c.v. est non adaptatif lorsque tous les $s_{g,t,j} = 1$ et adaptatif, lorsqu'au moins un $s_{g,t,j} \neq 1$.

Nous utilisons un seul indicateur motivé d'enquêtes polyvalentes, c'est-à-dire présentant un éventail important et diversifié de variables d'enquête cibles. Toutefois, dans les enquêtes comptant une ou quelques

variables d'enquête cibles, des indicateurs plus ciblés peuvent être utilisés. Un bon exemple est l'indicateur $H1$ de Särndal et Lundström (2010). Faire cela modifierait l'utilisation de données d'enquête historiques et inclurait également des associations avec la ou les variables d'enquête cibles.

Les *contraintes* sont des inégalités ou équations fonctionnelles représentant des restrictions logiques quant aux variables de décision autorisées. Des contraintes peuvent garantir, par exemple, une recherche approfondie de solutions possibles à partir d'un espace de solutions finies. Dans le contexte des plans d'enquête, une contrainte peut être une limite imposée sur la qualité de l'enquête, comme des solutions entraînant un taux de réponse global supérieur à 0,5, ou sur le coût de l'enquête, comme le coût global des intervieweurs s'adressant aux non-répondants demeurant inférieur à un montant donné. Dans la présente étude, nous nous concentrons sur des contraintes de coût quant à la charge de travail relative à la prise de contact avec les candidats non répondants au moyen d'un mode de suivi. Dans les ententes sur les niveaux de service avec des commanditaires d'enquête, une probabilité maximale de dépassement budgétaire est souvent spécifiée, disons de 10 %. α désigne cette proportion de dépassement de budget. Si le budget est strictement limité, alors $\alpha = 0.$ Soit h le niveau budgétaire précisé. La contrainte de coût est désormais :

$$p\left(\sum_g s_{g,t}\,(n_{g,t} - r_{g,t}) \geq \sum_g h(n_{g,t} - r_{g,t})\right) \leq \alpha, \tag{3.4}$$

où p est la probabilité que la charge de travail adaptative dépasse la charge de travail non adaptative. Lorsque les valeurs de $s_{g,t}$ respectent la contrainte (3.4), la solution correspondante de la variable de décision est dite acceptable; sinon, la solution contredit la règle. Il est naturel de préciser des limites inférieure et supérieure de la variable de décision $\mathbf{s} = \{s_{g,t} \mid \forall g, t\},$ appelées « contraintes de boîtes »,

$$0 \leq \mathbf{s} \leq 1. \tag{3.5}$$

Ainsi, le problème d'optimisation dans une application de répartition d'échantillon dans le cadre non bayésien est formulé afin de détecter un vecteur $\mathbf{s}$ minimisant l'objectif (3.1) sujet aux contraintes (3.4) et (3.5) en fonction des paramètres $(n, r).$ Comme nous l'avons énoncé auparavant, (3.1) et la charge de travail figurant dans (3.4) sont des variables aléatoires dans le cadre de l'approche bayésienne, de sorte que nous tirons des espérances des répartitions *a posteriori*.

Puisqu'aucune expression explicite n'existe pour les répartitions *a posteriori*, nous les estimons empiriquement. Nous utilisons les répliques de l'échantillonneur de Gibbs des répartitions *a posteriori* et calculons, par réplique, le c.v. et les charges de travail. Nous obtenons :

$$\hat{E}\left(\text{c.v.}(s,t)\right) = \frac{1}{K}\sum_k \frac{\sqrt{\sum_g d_{g,t}\left(\rho_{g,t}^{(k)} - \bar{\rho}_t^{(k)}\right)^2}}{\bar{\rho}_t^{(k)}}, \tag{3.6}$$

où $\hat{E}\left(\text{c.v.}(s,t)\right)$ est l'espérance estimée *a posteriori* au moment $t,$ $\rho_{g,t}^{(k)}$ est la k^{e} estimation itérée de la fonction prédictive *a posteriori* de $\rho_{g,t},$ et l'indice k désigne les passages sur les tirages MCMC. La probabilité d'un dépassement de budget est estimée empiriquement par la fréquence à laquelle le budget nécessaire pour assurer le plan adaptatif dépasse le budget établi :

$$\frac{\sum_k \mathbf{1}_{\sum_g s_{g,t}\left(n_{g,t}-r_{g,t}^{(k)}\right) \geq \sum_g h\left(n_{g,t}-r_{g,t}^{(k)}\right)}}{K} \leq \alpha. \tag{3.7}$$

$\mathbf{1}_{\sum_g s_{g,t}(n_{g,t}-r_{g,t}^{(k)}) \geq \sum_g h(n_{g,t}-r_{g,t}^{(k)})}$ est une fonction indicatrice qui prend la valeur de un lorsque l'inégalité dans son indice est satisfaite pour la k^{e} itération et de zéro autrement. Ainsi, l'optimisation bayésienne vise à réduire au minimum l'objectif de l'équation (3.6) en fonction des contraintes des équations (3.5) et (3.7).

Étalonnage : dans le problème d'optimisation bayésien, nous établissons un étalonnage afin d'évaluer le rendement de PEA de deux points de vue : améliorer la qualité et réaliser des économies. En particulier, promouvoir la représentativité de l'échantillon par recrutement peut améliorer la qualité de la collecte de données, alors que répartir les ressources coûteuses là où elles sont les plus nécessaires peut permettre de réaliser des économies. Cet objectif peut être atteint en passant, par exemple, d'un mode simple à des modes mixtes, mais optimalement répartis de nouveau ou en passant de modes mixtes complets à des modes mixtes partiels. En définissant les variables de décision en $\mathbf{s}=0$ ou $\mathbf{s}=1$, le problème d'optimisation proposé ci-dessus peut établir ces nouvelles répartitions. Pour ce faire, les rendements du plan d'enquête à mode unique et du plan multimodal complet sont des normes de rendement des PEA et peuvent être comparés.

3.2 Optimisation de plans d'enquête adaptatifs statiques

Par analogie avec les méthodes de traitement adaptatif, nous disons que les PEA sont *statiques* lorsqu'ils reposent uniquement sur les renseignements disponibles dans les données de registre et de la base de sondage avant le début de la collecte de données et qu'ils sont *dynamiques* lorsqu'ils reposent (également) sur des paradonnées (données recueillies au cours de la collecte de données). Les PEA *dynamiques* reflètent la nature dynamique de l'optimisation, puisque celle-ci a lieu à chaque étape de la collecte de données, c'est-à-dire à l'issue de chaque mode.

Pour ce qui est des PEA *dynamiques* dans le présent contexte, on prend des décisions d'affectation d'intervieweurs aux strates en fonction des résultats d'enquête intermédiaires provenant des modes précédents. La corrélation entre les propensions à répondre pour le mode sans intervieweur et celles pour le mode avec intervieweur mène à une mise à jour intermédiaire des répartitions *a priori* du mode avec intervieweur. Théoriquement, l'évaluation peut déterminer les priorités des personnes qui refusent de répondre dans les strates contenant des répondants à interviewer et organiser la charge de travail des intervieweurs. En réalité, le temps peut manquer pour calculer à temps la charge de travail des intervieweurs faisant l'objet d'une nouvelle affectation du fait du regroupement géographique. De plus, la réaffectation nécessite une logistique complexe de gestion des cas; nous reviendrons sur ce point dans la section sur l'analyse.

Par conséquent, dans la présente étude, nous nous concentrons sur l'environnement des PEA statiques; autrement dit, la décision d'effectuer un suivi est prise au début de la collecte de données. Nous nous limitons à cette intervention, mais la méthodologie permettrait de multiples interventions, par exemple après plusieurs appels ou visites.

Dans la section 3.3, nous élaborons une stratégie pour tenir compte de l'incertitude de la prise de décision et préciser la routine d'optimisation afin de déterminer les affectations optimales du problème d'optimisation bayésien présenté à la section 3.1.

3.3 Stratégie d'optimisation

Pour résoudre le problème d'optimisation formulé dans la section 3.1, nous proposons une stratégie à deux étapes au moment t :

1. *Élaboration de la répartition a posteriori des nombres de réponses* $r_{g,t}$. Posons que les ensembles de données de série chronologique historiques jusqu'au moment $t-1$ sont des données utilisées pour l'entraînement du modèle et que les ensembles de données au moment t sont des données d'essai pour la prédiction. Tous les coefficients et hyperparamètres de modèle de l'équation (2.8) peuvent être estimés par la taille d'un échantillon $\mathbf{n}_{1:t-1} = \{\mathbf{n}_{g,1:t-1} \mid \forall g\}$ et les nombres de réponses pour tous les modes $\mathbf{r}_{1:t-1} = \{\mathbf{r}_{g,1:t-1,j} \mid \forall g, j\}$. Selon le modèle estimé, des prédictions peuvent être obtenues sur des propensions dépendantes $\tilde{\boldsymbol{\rho}}_t = \{\tilde{\rho}_{g,t,j} \mid \forall g, j\}$ et $\mathbf{r}_t = \{\mathbf{r}_{g,tj} \mid \forall g, j\}$, en fonction des données $\mathbf{n}_t = \{n_{g,t} \mid \forall g\}$.
2. *Détermination des répartitions optimales*. Précisons le niveau budgétaire h et le seuil de dépassement α. Définissons plusieurs vecteurs de départ de répartitions de strates $\mathbf{s}$, chaque vecteur étant considéré comme un état initial et ayant chacun un nombre fini d'états successifs bien définis. Pour toute strate g, supposons K itérations des estimations de $\mathbf{r}_{g,t} = \{r_{g,t,m} \mid \forall j\}$ et $\tilde{\boldsymbol{\rho}}_{g,t} = \left\{\tilde{\rho}_{g,t,j} \mid \forall j\right\}$ générées à partir des répartitions *a posteriori* en 1. Ces estimations *a posteriori* et paramètres donnés h et α sont substitués séparément dans les équations (3.6) et (3.7) pour calculer l'espérance *a posteriori* $\hat{E}\left(\text{c.v.}(s,t)\right)$ et la probabilité *a posteriori* de l'excès de charge de travail. Pour détecter les optima, en commençant à chaque état initial, un tel calcul se déroule en passant par ses états successifs, produit un extrant et finalement atteint un état final. Éliminons les états ne respectant pas les contraintes, et leurs extrants, et conservons les états respectant les contraintes et leurs extrants. Dans ces résultats, additionnons le minimum de $\hat{E}\left(\text{c.v.}(s,t)\right)$ et ses optima de répartition correspondants.

Résoudre ce programme mathématique constitue une tâche qui nécessite beaucoup de calculs. Par conséquent, les méthodes de la première étape sont mises en œuvre dans R à l'aide du paquet *mcmcsae* (Boonstra, 2022), alors que les méthodes de la deuxième étape sont mises en œuvre dans R à l'aide de la fonction *auglag* (algorithme de minimisation lagrangien augmenté) du paquet *Alabama* (Varadhan, 2022) pour une optimisation non linéaire limitée.

3.4 Évaluation du rendement

La présente section introduit un critère d'évaluation afin d'évaluer l'exactitude de la prédiction. Ce critère peut mettre en lumière l'amélioration en matière de réduction du risque de non-réponse des différents

modèles ou des différents plans d'enquête. Cette amélioration est quantifiée par la racine carrée de l'erreur quadratique moyenne (REQM) de la répartition *a posteriori* d'un paramètre τ (par exemple la propension à répondre ou le c.v.), par rapport à la « réalité », cette dernière étant estimée au moyen d'observations.

Nous considérons un rendement tout au long d'une fenêtre dynamique de trois mois. Ce choix est motivé par la durée de trois mois du travail sur le terrain de l'application de la présente étude, mais peut être modifié pour refléter n'importe quelle durée. Dans la fenêtre temporelle $q = \{t, t+1, t+2 \mid \forall t\}$, la REQM de la g^{e} strate est alors exprimée sous la forme

$$\text{REQM}(\tau, q) = \sqrt{B^2(\tau, q) + \text{SD}^2(\tau q)}, \tag{3.8}$$

où le premier terme s'appelle le terme de biais, représenté comme la différence quadratique entre la moyenne *a posteriori* du paramètre τ et le τ observé,

$$B^2(q) = \sum_g d_{g,q} \left(E_{\pi_q} \text{c.v.}(s_q, q) - \widehat{\text{c.v.}}(s_q, q) \right)^2 \tag{3.9}$$

et le second terme est la variance *a posteriori* du c.v., qui est la forme quadratique de l'écart-type (é.-t.),

$$\text{é.-t.}^2(q) = \sum_g d_{g,q} \, \text{Var}_{\pi_q} \text{c.v.}(s_q, q). \tag{3.10}$$

Le poids $d_{g,q} = n_{g,q} \Big/ \sum_g n_{g,q}$ est le ratio de la taille de la strate g sur la taille de l'échantillon dans la fenêtre q. Les répartitions *a posteriori* π_q du c.v. et la répartition $\mathbf{s}_q$ sont dérivées de la stratégie de calcul présentée à la section 3.3.

Ces critères dépendent fortement de la variation d'échantillonnage de la taille de l'échantillon, en particulier pour les enquêtes présentant une petite taille d'échantillon. Des données empiriques assujetties à la variation d'échantillonnage sont utilisées pour évaluer le rendement. Alors que les enquêtes présentant d'importantes tailles d'échantillon fournissent des renseignements riches et, donc que leur rendement peut être évalué avec précision, pour les petites enquêtes, une contradiction par rapport à la variation temporelle devient forte, c'est-à-dire qu'une évaluation précise prend davantage de temps. Il est plus difficile de tirer une conclusion solide sur la mise en pratique de l'adaptation en cas de rendement basé sur des critères instables.

4. Étude de cas de l'Enquête sur la santé aux Pays-Bas

Dans la présente section, nous étudions et exploitons l'application des modèles de série chronologique multinomiaux exposés dans la deuxième section et l'approche d'optimisation dans un cadre bayésien présentée dans la troisième section à l'Enquête sur la santé aux Pays-Bas (enquête GEZO). La section 4.1 présente brièvement le contexte de l'enquête GEZO. Nous montrons séparément dans les sections 4.2 à 4.4 la façon dont les variations temporelles dans des propensions séquentielles peuvent être modélisées, dans quelle mesure le rendement de répartitions optimales dépend du niveau budgétaire et dans quelle mesure des décisions optimales dépendent de la durée des données historiques applicables.

4.1 Enquête sur la santé aux Pays-Bas

L'enquête GEZO est menée chaque année par le bureau central de la statistique des Pays-Bas, pour fournir un aperçu méthodique de l'évolution des contacts médicaux, du mode de vie, de la santé et des comportements préventifs de la population néerlandaise, notamment toutes les personnes vivant dans des ménages privés. Un plan de sondage autopondéré à deux degrés est utilisé; il produit d'abord un échantillon de municipalités, puis un échantillon de personnes vivant dans certaines municipalités. Cette enquête est passée à un plan de sondage multimodal après 2014. La méthode d'observation fait intervenir des interviews en ligne et sur place. Tout d'abord, des IWAO permettent de demander la participation d'unités d'échantillonnage de la population. Ensuite, des non-répondants sont recrutés pour participer à des IPAO. À compter de 2018, cependant, une adaptation a été mise en œuvre pour stabiliser la charge de travail des intervieweurs. Seule une portion des non-répondants aux IWAO est abordée de nouveau pour participer à une IPAO, afin de réduire les coûts de l'enquête et d'améliorer la représentativité. Les taux de réponse plus élevés des unités d'échantillonnage des IWAO signifient une possibilité moindre d'être abordés de nouveau.

Dans la présente étude, nous nous concentrons sur une série chronologique de données recueillies de 2014 à 2017, comprenant 48 mois. Il convient de mentionner que les données recueillies au début de 2017 étaient « anormales », du fait de problèmes techniques relatifs au serveur Web, ce qui a entraîné une interruption de la collecte de données. Cela se justifie par des raisons pratiques : le bureau central de la statistique des Pays-Bas a mis en œuvre des PEA statiques depuis 2018; cette adaptation peut gâcher la valeur potentielle des données historiques utilisées pour améliorer l'exactitude de la prédiction (Wu et coll., 2023). De plus, les unités d'échantillonnage sont stratifiées en 13 strates disjointes par deux variables auxiliaires provenant de la base de sondage ou de registres administratifs : l'âge et l'ethnicité. Voir la stratification à l'annexe A. Il convient de souligner que cette stratification est fixe tout au long de la présente étude et que nos strates de série chronologique sont différentes des strates de PEA (Van Berkel et coll., 2020). L'ensemble de variables auxiliaires disponibles avant le début du travail sur le terrain est bien plus grand. Il comprend plusieurs données démographiques, comme le genre, le pays de naissance, la composition du ménage, des caractéristiques socioéconomiques, comme le revenu personnel et le revenu du ménage enregistrés, le niveau de scolarité, le type de profession, ainsi que des caractéristiques relatives au logement et à la région, comme le type de logement, la valeur du logement et l'urbanisation. Une étude portant sur la sélection efficiente et efficace des strates est une étape importante à venir.

4.2 Comment construire un modèle de série chronologique sous un plan séquentiel multimodal

La présente section fournit des précisions sur l'approche permettant de créer des modèles de série chronologique hiérarchiques. Nous observons deux décisions potentiellement influentes. La première décision consiste à inclure une composante saisonnière. L'ajout d'une saison améliorera probablement l'exactitude, mais comporte un coût. Inclure une saison signifie qu'une série plus longue de données historiques est nécessaire; il faut observer au moins deux années de données pour évaluer une composante

saisonnière, mais préférablement davantage. La deuxième décision est l'inclusion d'associations ou de corrélations explicites entre les paramètres du modèle. Encore une fois, l'exactitude sera probablement légèrement plus élevée, mais davantage de données sont nécessaires. Nous pouvons encore différencier les décisions à trois autres niveaux : la strate, le temps et le mode. Nous pouvons ainsi rendre les paramètres des composantes saisonnières et les corrélations dépendants du mode et de la strate, et même du temps. Nous explorons quatre scénarios : composante saisonnière présente ou absente multipliée par corrélation présente ou absente.

Notre objectif est de trouver le scénario le plus avantageux parmi ces quatre scénarios. Puisqu'il existe de nombreux modèles possibles, par analogie avec Wu et coll. (2023), nous adoptons une stratégie pas à pas pour passer de modèles simples à des modèles plus avancés. En évaluant le rendement, nous tenons compte de deux critères d'information : le critère d'information de déviance (DIC, voir Spiegelhalter, Best, Carlin et van der Linde, 2002) et le critère d'information de Watanabe-Akaike (WAIC, voir Watanabe, 2010 et 2013). En offrant un compromis raisonnable entre la qualité de l'ajustement du modèle et la complexité du modèle, ces mesures devraient permettre de sélectionner des modèles adéquats pour la tâche de prédiction qui nous intéresse. Nous pouvons les interpréter comme étant des approximations de mesures de validation croisée avec retrait d'un élément, relativement faciles et peu onéreuses à calculer à partir de l'extrant de la simulation de la méthode MCMC (Vehtari, Gelman et Gabry, 2017). Lors de la construction des modèles, nous pouvons choisir parmi divers effets aléatoires : le bruit blanc, la tendance temporelle globale, les ordonnées à l'origine aléatoires pour chaque strate et les tendances temporelles propres à la strate. De plus, nous pouvons faire cela avec ou sans saisonnalité et disposons de paramètres de modèles dépendants ou indépendants du mode. Nous suivons les étapes ci-dessous :

1. Modèle de base : définir deux modèles de base; l'un sans composantes saisonnières et l'autre ayant des composantes saisonnières. Les deux modèles présentent des effets de mode fixes et des effets fixes de variables auxiliaires propres au mode β. La différence entre les deux modèles est l'inclusion ou non de la saisonnalité δ.
2. Tendance de base : ajouter un seul effet aléatoire, c'est-à-dire {tendance temporelle globale u_t}, {ordonnées à l'origine aléatoires pour les strates v_g} ou {bruit blanc e_{gt}} aux modèles de 1. Chaque effet aléatoire est corrélé avec des modes ou rendu indépendant des modes. Examiner si le modèle avec ou sans saisonnalité de 1 est amélioré par l'un des trois effets aléatoires.
3. Certaine dépendance à la strate : ajouter une combinaison d'effets aléatoires, c'est-à-dire, {tendance temporelle globale u_t, ordonnées à l'origine aléatoires pour les strates v_g} ou {ordonnées à l'origine aléatoires pour les strates v_g, tendance temporelle propre à la strate z_{gt}}, aux modèles de 1. Chaque effet aléatoire est corrélé avec des modes ou est indépendant des modes. Examiner si les modèles mis à jour fournissent de meilleurs résultats que les modèles exposés au point 2.
4. Dépendance modérée à la strate : ajouter une combinaison de trois effets aléatoires, c'est-à-dire $\{u_t, v_g, e_{gt}\}$ ou $\{u_t, v_g, z_{gt}\}$, aux modèles de 1. Chaque effet aléatoire est corrélé avec des modes ou est indépendant des modes. Examiner si les modèles mis à jour fournissent de meilleurs résultats que les modèles exposés au point 3.

5. Dépendance complète à la strate : ajouter tous les effets aléatoires aux modèles de 1. Chaque effet aléatoire est corrélé avec des modes ou est indépendant des modes. Examiner si la combinaison complète améliore le rendement du modèle.

Le tableau 4.1 présente les résultats de ces cinq étapes. Comme l'indique chaque ligne du tableau 4.1, les modèles avec et sans saisonnalité obtiennent une correspondance égale en matière d'ajustement et de prédiction. Les résultats des critères d'information (CI) des deux modèles de base (modèle 1) indiquent que le rendement du modèle avec saisonnalité est légèrement supérieur à celui du modèle sans saisonnalité. Cet avantage se maintient lorsque certains effets aléatoires s'ajoutent (voir les modèles 1, 3 et 6), alors que l'inclusion de la saisonnalité δ produit des CI inférieurs. D'autre part, les modèles avec saisonnalité présentent un rendement légèrement plus faible, comme on le voit pour les modèles 2, 4, 7, 8 et 9. En particulier, les résultats du modèle 5 présentent des résultats mitigés. Les résultats du DIC appuient la modélisation des effets saisonniers pour des prédictions de propension exactes, mais le WAIC ne peut pas s'y conformer.

En matière d'équilibre entre la complexité et la qualité de l'ajustement du modèle, le rendement des modèles indépendants des modes est pratiquement équivalent aux modèles corrélés aux modes, malgré un rendement légèrement plus élevé (modèle 5 pour le modèle sans saison et le modèle 6).

Tableau 4.1
Critères d'information et chiffres réels lors de l'évaluation de la qualité de l'ajustement et de la complexité des modèles.

Modèle	Effet fixe	Effet aléatoire	DIC		p_{DIC}		WAIC		p_{WAIC}	
			IND	COR	IND	COR	IND	COR	IND	COR
1	β	-	6 808		18		6 817		28	
	β, δ		6 800		25		6 813		38	
2	β	u_t	6 504	6 504	67	67	6 509	6 509	72	72
	β, δ		6 506	6 505	70	70	6 510	6 510	74	74
3	β	v_g	6 775	6 774	26	26	6 786	6 785	36	36
	β, δ		6 768	6 767	32	32	6 781	6 780	46	45
4	β	e_{gt}	6 602	6 601	322	321	6 479	6 479	200	200
	β, δ		6 609	6 608	322	321	6 488	6 487	201	200
5	β	u_t, v_g	6 488	6 490	147	146	6 472	6 475	131	130
	β, δ		6 474	6 472	77	77	6 477	6 476	80	81
6	β, δ	v_g, z_{gt}	6 494	6 495	143	144	6 481	6 482	130	129
	β, δ		6 488	6 490	147	146	6 472	6 475	131	130
7	β	u_t, v_g, e_{gt}	6 449	6 448	195	194	6 398	6 396	143	142
	β, δ		6 453	6 452	191	191	6 404	6 402	142	142
8	β	u_t, v_g, z_{gt}	6 396	6 395	109	108	6 381	6 380	94	94
	β, δ		6 398	6 397	112	111	6 382	6 381	96	95
9	β	u_t, v_g, z_{gt}, e_{gt}	6 398	6 398	137	137	6 371	6 371	110	110
	β, δ		6 400	6 397	132	134	6 375	6 372	108	108

Note : La série chronologique de 2014 à 2017 est ajustée à un modèle indépendant du mode (« IND ») et un modèle est corrélé au mode (« COR »). Chaque modèle est simplifié en utilisant uniquement les composantes d'effets fixes, puis en ajustant la corrélation au fil du temps ou entre les modes en tenant compte de plusieurs effets aléatoires.
DIC = Deviance Information Criterion;
WAIC = Widely Applicable Information Criterion.

En tenant compte des effets aléatoires, les modèles mixtes sont meilleurs, car ils entraînent une diminution des CI, contrairement aux modèles qui comportent uniquement des effets fixes (modèle 1). Comparer les modèles 2 à 4 au modèle 1 signifie que le modèle avec ou sans saisonnalité est amélioré en introduisant un seul effet aléatoire, où la tendance temporelle globale u_t entraîne la plus grande diminution des CI, suivie d'un bruit blanc e_{gt} et d'une ordonnée à l'origine aléatoire pour la strate. Une telle amélioration persiste pour les CI en appliquant la combinaison de deux effets aléatoires, comme le montrent les comparaisons du modèle 5 avec les modèles 2 et 3, et du modèle 6 avec le modèle 3. Apparemment, le modèle 5 présente la diminution la plus forte des CI jusqu'à présent. Les modèles 7 et 8 indiquent que les modèles peuvent être encore améliorés en ajoutant le bruit blanc e_{gt} et une tendance temporelle propre à la strate z_{gt} au modèle 5, et dans le modèle 8, les CI diminuent davantage que dans le modèle 7. Inclure le bruit blanc e_{gt} est utile pour obtenir un meilleur rendement, car cela ajoute peu à la diminution du WAIC du modèle 9 malgré la faible contribution apportée au DIC.

Comme le montre le modèle 9, les modèles corrélés au mode et indépendants du mode (respectivement les colonnes COR et IND) présentent un rendement semblable en matière de CI en ne tenant pas compte de la saisonnalité; toutefois, pour les modèles avec saisonnalité, la modélisation des corrélations (colonne COR) est meilleure en ce qui a trait aux résultats de CI par rapport à la colonne IND. Il est cependant difficile de conclure que le modèle avec saisonnalité présente un avantage absolu par rapport au modèle sans saisonnalité en termes de qualité de l'ajustement et de complexité du modèle. Pour déterminer si les effets saisonniers jouent un rôle essentiel dans des répartitions adaptatives, nous considérons les modèles sans saisonnalité et avec saisonnalité (indiqués en rouge) dans la section 4.4.

4.3 Dans quelle mesure le rendement des plans d'enquête adaptatifs est-il sensible au niveau budgétaire donné ?

Cette question de recherche porte sur la façon d'adapter les répartitions des non-réponses aux IWAO pour toutes les strates, afin de réduire au maximum le risque de non-réponse, en fonction d'un niveau budgétaire donné. Elle soulève également la question de savoir si une telle réduction peut se maintenir pour différents niveaux budgétaires.

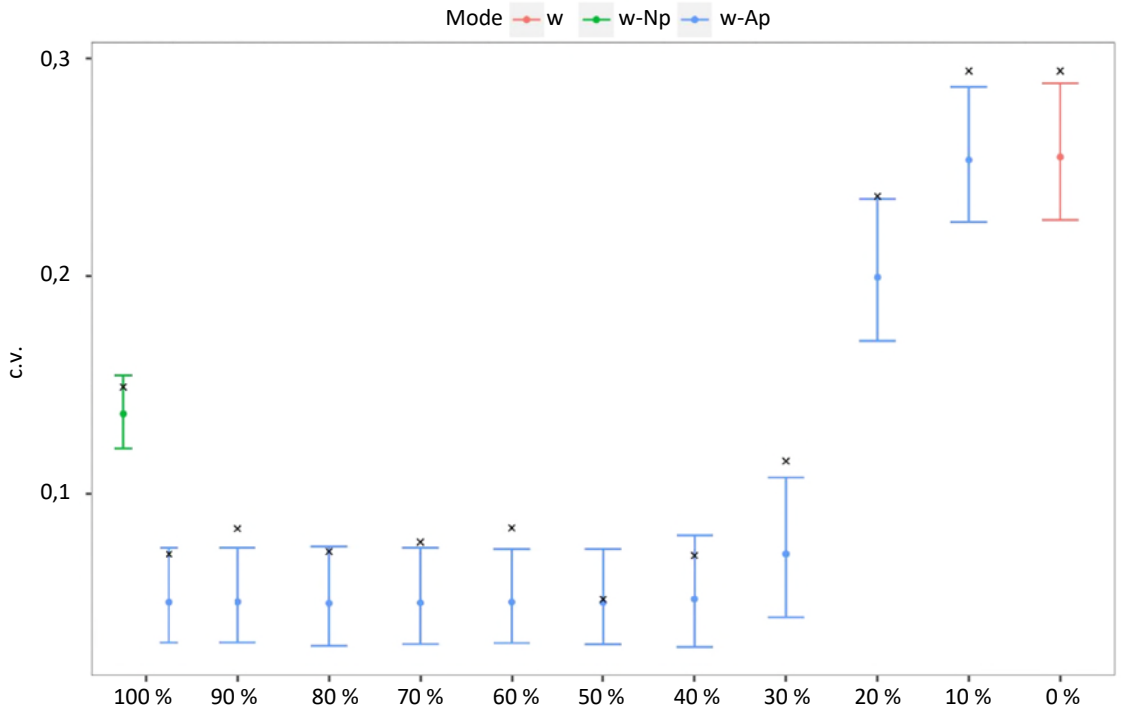
Nous répondons à cette question en réduisant d'abord au minimum (3.6) en fonction de (3.5) et (3.7) pour le trimestre de collecte de données suivant, lorsque le niveau budgétaire est précisé, puis en comparant la valeur optimale de (3.5) pour le c.v. réalisé pour le même niveau budgétaire, et enfin, en comparant la valeur optimale de (3.5) pour différents niveaux budgétaires. Nous nous concentrons sur le trimestre suivant, car, dans le cas statique, le nombre de répondants aux IWAO est inconnu tant que les données ne sont pas recueillies et car l'échantillon suffisant d'un trimestre peut assurer la précision de la prédiction. En se reportant à la stratégie d'optimisation de la section 3.3, la procédure d'évaluation du trimestre q est la suivante :

1. Soit le niveau budgétaire h commençant à 100 % puis diminuant successivement par paliers de 10 %, c'est-à-dire $h \in \{1, 0{,}9, 0{,}8, 0{,}7, 0{,}6, 0{,}5, 0{,}4, 0{,}3, 0{,}2, 0{,}1\}$.
2. Considérons le trimestre à venir q et l'ensemble de données dans q comme l'ensemble de données d'essai.
3. Posons les données de la série chronologique jusqu'au trimestre $q-1$ comme l'ensemble de données d'essai pour estimer les modèles sélectionnés. Les modèles comprennent les composantes du modèle 9 considérées comme la « meilleure » représentation.
4. Utilisons la taille d'échantillon dans q pour simuler les réponses aux IWAO. Pour chaque strate et chaque mois dans q, 3 000 tirages sont générés à partir des répartitions prédictives *a posteriori*.
5. Selon le modèle simulé à l'étape 3, des prédictions individuelles *a posteriori* des IWAO et des IPAO conditionnelles sont générées séparément 3 000 fois pour chaque strate et chaque mois dans q.
6. Substituons le niveau budgétaire précisé h et les réponses aux IWAO simulées à l'étape 4 dans la contrainte de coût (3.7).
7. Calculons les propensions mixtes en substituant le niveau h et les prédictions individuelles de l'étape 5 dans (3.3).
8. Initialisons trois solutions de départ de probabilités de répartition, $s \in \{0, 0.5, 1\}$, chacune s'appliquant à 13 strates simultanément.
9. Commençons à chaque point initial de l'étape 8 pour trouver les solutions optimales pour chaque strate par le résolveur *auglag* en fonction des étapes 6 et 7.
10. Faisons le lien entre les solutions relevées et l'échantillon réel pour le calcul des prédictions de c.v. *a posteriori* et les réalisations de c.v.
11. Effectuons des comparaisons en répétant les étapes 2 à 10 pour chaque niveau budgétaire exposé à l'étape 1.

$s = 0$ indique aucun suivi par IPAO, $s =$ 0,5 signifie que la moitié des non-réponses aux IWAO a été affectée aux IPAO, et $s = 1$ représente un suivi complet par IPAO. Pour distinguer les stratégies des différents modes et simplifier la notation, les plans IWAO uniquement, les plans non adaptatifs et les plans adaptatifs sont désignés par w, w-Ap et w-Np dans l'ensemble du document.

Dans la figure 4.1, les prédictions de c.v. *a posteriori* pour le premier trimestre (T1) 2017 sont résumées pour chaque niveau budgétaire; voir le tableau B.1 de l'annexe B pour connaître les résultats de c.v. tenant compte du biais. Nous étalonnons le rendement w-Ap en tant que fonction du niveau budgétaire h par rapport au rendement de w et de w-Np. Par souci de concision, les c.v. pour les plans par IWAO uniquement, les plans non adaptatifs et les plans adaptatifs, sont simplifiés sous la forme c.v.(w), c.v.(w-Np) et c.v.(w-Ap).

Figure 4.1 Comparaison des coefficients de variation (c.v.) des prédictions de propension à répondre fondées sur un modèle comportant des observations de c.v. et tenant compte du biais au premier trimestre 2017.



Note : Les estimations de c.v. sont effectuées séparément pour les plans par IWAO uniquement (« w »), les plans non adaptatifs (« w-Np ») et les plans adaptatifs (« w-Ap »). Les prédictions de c.v. *a posteriori* sont résumées par la région crédible à 95 %, alors que les observations sont indiquées par un nuage de points (« x »).

La comparaison du c.v.(w-Np) avec le c.v.(w) indique que le recrutement des non-répondants aux IWAO par IPAO peut donner lieu à une diminution de la non-réponse, puisque la région crédible à 95 % (intervalle de confiance [IC]) du c.v.(w-Np) *a posteriori* est bien plus étroite que celle du c.v.(w) et le quantile à 97,5 % du c.v.(w-Np) *a posteriori* est bien inférieur au quantile à 2,5 % du c.v.(w). Lorsque $h = 100$ %, une autre diminution de la variation globale est obtenue par les répartitions optimisées de l'enquête adaptative. Les prédictions *a posteriori* et les observations du c.v.(w-Ap) divergent de 0,1 et se rapprochent de 0 par rapport au quantile à 2,5 % du c.v.(w-Np); cependant, l'IC de l'approche adaptative indique que l'exactitude de la prédiction est relativement compromise. Puisque les IC sont à peine modifiés lorsque le budget est réduit pour passer de 90 % à 50 %, la réduction de l'incertitude associée au c.v.(w-Ap) n'augmenterait probablement pas de plus de 100 %. Cela signifie que dans l'intervalle des niveaux de 100 % à 50 %, l'effet du faible budget sur le risque de non-réponse estimée est équivalent à celui du budget élevé. Les limites supérieures des IC semblent s'approcher des c.v. observés, voire les dépasser; par exemple, au niveau de 50 %, l'observation chevauche la moyenne *a posteriori*.

Le risque de non-réponse augmente avec la réduction continue du budget, puisque les estimations du c.v.(w-Ap) augmentent et indiquent un risque accru de biais de non-réponse. Pour les niveaux budgétaires inférieurs à 50 %, la méthode de répartition indiquée accroît l'incertitude des estimations *a posteriori* de la variation globale. De plus, les limites inférieures se rapprochent de 0,1, et même le dépassent largement, lorsque le niveau est de 20 % ou de 10 %, auxquels cas le résolveur finit par obtenir un « optimum » local

faux, du fait du non-respect du critère de convergence. Pour le niveau de 10 %, l'intérêt de la méthode de répartition est particulièrement réduit et elle perd son avantage, comme le montre le c.v.(w-Ap) exactement identique au c.v.(w). Pour déterminer le niveau budgétaire préférable, nous adoptons un critère : le coût relatif défini comme le coût global de la taille adaptative pour les IPAO par rapport à la taille non adaptative limitée par le niveau budgétaire. Voir le tableau B.3 à l'annexe B pour connaître les résultats du coût relatif pour différents niveaux.

Des répartitions optimisées rendent le rendement de l'adaptation constant pour des niveaux budgétaires relativement élevés (100 % à 50 %). De plus, même si cela contribue légèrement à des pertes de précision, l'adaptation est néanmoins efficace pour le risque de non-réponse estimé plus faible par rapport aux plans w et w-Np (barres d'erreur rouges et vertes). Jusqu'à un niveau budgétaire de 40 %, le rendement s'inverse et évolue dans la direction opposée, ce qui signifie que le risque de non-réponse augmente nettement. On s'attend à ce que ce comportement général soit relativement robuste pour des variations raisonnables dans le choix de modèle, car les deux extrêmes de budget ne laissent aucune marge dans la répartition des PEA pour compenser des différences de propensions à répondre.

4.4 Dans quelle mesure le rendement des plans adaptatifs dépend-il des données historiques disponibles ?

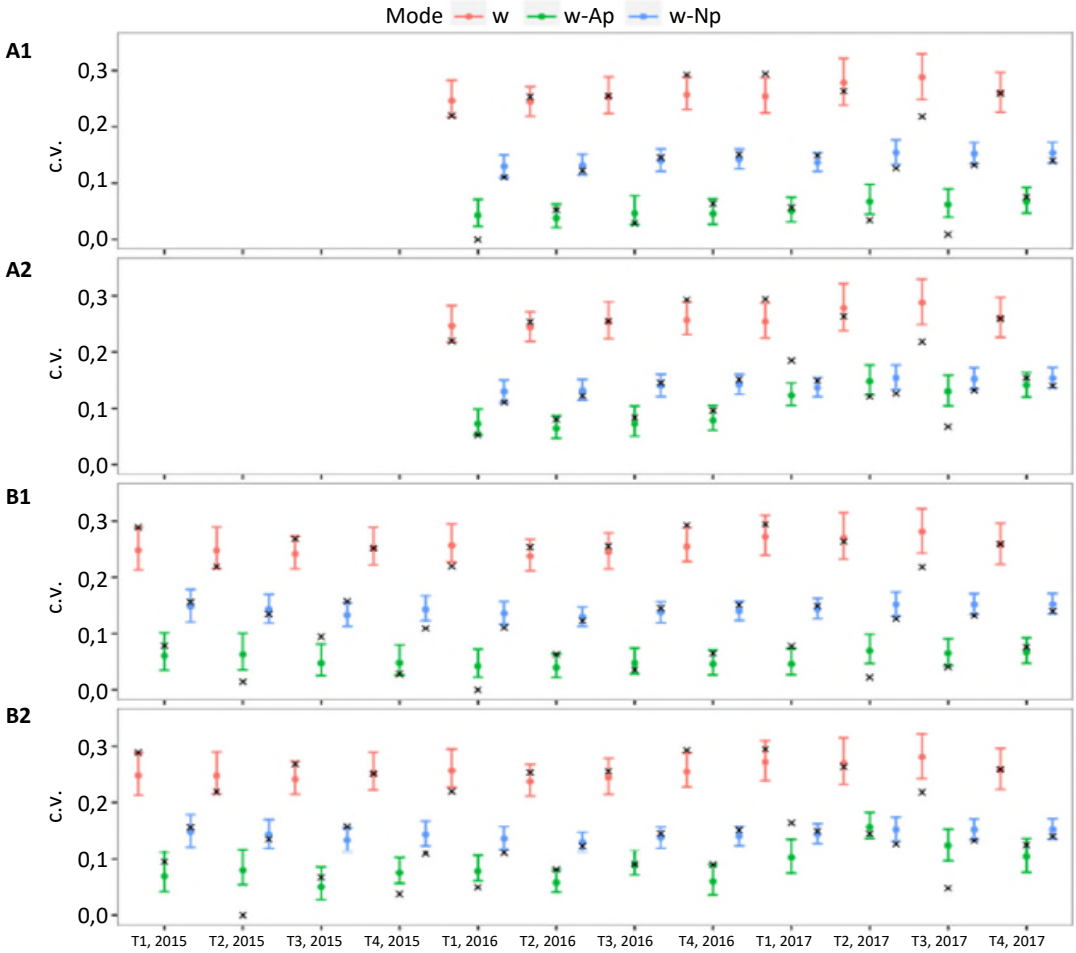
Cette question nécessite d'examiner dans quelle mesure la série chronologique historique accumulée influence le rendement des plans adaptatifs, c'est-à-dire le risque de non-réponse mesuré par les c.v. et l'équilibre entre le biais et la variance mesuré par la REQM. Pour répondre à cette question, nous étudions le rendement des plans w, w-Np et w-Ap au niveau du trimestre civil. De plus, nous étalonnons le rendement des plans adaptatifs par rapport au rendement des plans w et w-Np.

Nous comparons et évaluons les modèles avec et sans l'inclusion de la saisonnalité, et ce, pour des niveaux budgétaires de 50 % et de 30 %. La section 4.2 indique que la saisonnalité est un facteur pouvant être ignoré, puisque les modèles avec et sans ce résultat présentent une qualité de l'ajustement et une complexité semblables. La section 4.3 donne à penser que pour une fenêtre temporelle particulière, le niveau budgétaire de 50 % favorise le rendement de PEA le plus rentable et que le PEA perd son avantage absolu pour des valeurs plus petites. La sensibilité du rendement à la durée de la série chronologique est moins évidente lorsque les modèles tiennent compte de la saisonnalité ou lorsque le niveau budgétaire est inférieur à 50 %; il est donc prématuré de les ignorer dans l'analyse. En croisant les deux conditions, il est possible de procéder à des comparaisons simultanées des quatre scénarios de modèles : 1) l'inclusion de la saisonnalité et un niveau budgétaire de 50 %; 2) l'inclusion de la saisonnalité et un niveau budgétaire de 30 %; 3) la non-inclusion de la saisonnalité et un niveau budgétaire de 50 %; 4) la non-inclusion de la saisonnalité et un niveau budgétaire de 30 %.

Pour étudier la sensibilité à la durée de la série chronologique historique, l'analyse est effectuée sur une base dynamique en ajoutant un mois à chaque fois. Il convient de se rappeler que la durée de la série chronologique historique initiale devrait être d'au moins un an pour les modèles n'incluant pas la saisonnalité (scénarios 3 et 4), mais d'au moins deux ans pour les modèles comprenant la saisonnalité (scénarios 1 et 2). Pour chacun, le processus d'entraînement se termine au troisième trimestre de 2017), car un trimestre doit demeurer pour la prédiction.

Dans la figure 4.2, l'incertitude relative aux c.v. estimés évaluée par la région crédible à 95 % et les moyennes *a posteriori* sont, ensemble, comparées avec les observations des c.v. tout au long des trimestres et selon différents plans. Dans les plans w et w-Np, les c.v. observés se situent dans les intervalles ou sont très proches des limites des intervalles pour la plupart des trimestres, à l'exception du c.v.(w) du troisième trimestre de 2017. Les résultats des PEA des sections A1, A2 et B2 soutiennent cette constatation. De plus, des observations très éloignées des IC apparaissent au deuxième trimestre de 2015 dans la section B2 et au premier trimestre de 2017 dans les sections A2 et B2. Cette exception laisse entendre que pour des trimestres correspondants, on est moins convaincu que le rendement adaptatif évalué reproduit le rendement en pratique.

Figure 4.2 Pour un niveau budgétaire donné, coefficient de variation (c.v.) *a posteriori* pour les plans adaptatifs, les plans non adaptatifs et les plans par IWAO uniquement, par rapport aux observations tout au long des trimestres.



Note : Les régions crédibles à 95 % avec les espérances *a posteriori* sont résumées pour les plans par IWAO uniquement (« w »), les plans non adaptatifs (« w- p), et les plans adaptatifs (« w-Ap »). On désigne les observations par des croix « x » noires. Les sections « A » présentent les résultats des modèles avec saisons et les sections « B », les résultats des modèles sans saisons. Les sections « A1 » et « B1 » correspondent au niveau budgétaire de 50 %, alors que les sections « A2 » et « B2 » correspondent au niveau budgétaire de 30 %. Le trimestre sur l'axe des x désigne le trimestre existant à des fins de prédiction.

Comme cela a été mentionné à la section 4.3, le rendement des plans d'enquête adaptatifs au niveau budgétaire de 50 % est constamment supérieur au rendement des plans au niveau budgétaire de 30 % pour tous les trimestres, comme l'indiquent les comparaisons de A1 à A2 ou de B1 à B2. Au niveau de 50 %, le c.v.(w-Ap) estimé est plus précis du fait de régions crédibles plus étroites, ce qui est implicitement visible. De plus, nous pouvons observer l'avantage absolu de plans adaptatifs dont le rendement est supérieur à celui de plans non adaptatifs, puisque les limites supérieures du c.v.(w-Ap) s'écartent de façon substantielle des limites inférieures du c.v.(w-Np); cependant, au niveau de 30 %, ils sont concurrentiels, par exemple pour les premier et deuxième trimestres de 2017 (voir les sections A2 et B2).

À mesure que les données historiques s'accumulent, on suppose que les modèles peuvent être encore optimisés; l'exactitude de la prédiction des c.v. peut présenter une augmentation systématique et le rendement obtenu peut être amélioré. Il est clair que cela est le cas dans la section B1 jusqu'au quatrième trimestre 2016; toutefois, à partir de cette date, il semble ne plus être possible d'améliorer le rendement, et les estimations *a posteriori* du c.v.(w-Ap) se stabilisent pour se situer à environ 0,1. Les résultats pour les sections A1 et B1 sont similaires; on doit donc conclure que la modélisation de la saisonnalité contribue peu à l'exactitude de la prédiction.

Tirer des conclusions sur la robustesse du rendement de PEA est dogmatique pour deux raisons. Tout d'abord, l'avantage peut être plus grand pour certaines strates. Leurs estimations individuelles de c.v. peuvent être moins biaisées et varier pour tendre vers l'option qui inclut la saisonnalité malgré peu de différence dans la variation globale. Deuxièmement, les tailles d'échantillons de certaines strates sont relativement petites. Tôt au cours des étapes de collecte de données, elles peuvent présenter un comportement volatile quant à l'équilibre entre l'erreur et la variation. La durée de la série chronologique historique déterminée en fonction de ces résultats n'est pas une garantie de robustesse.

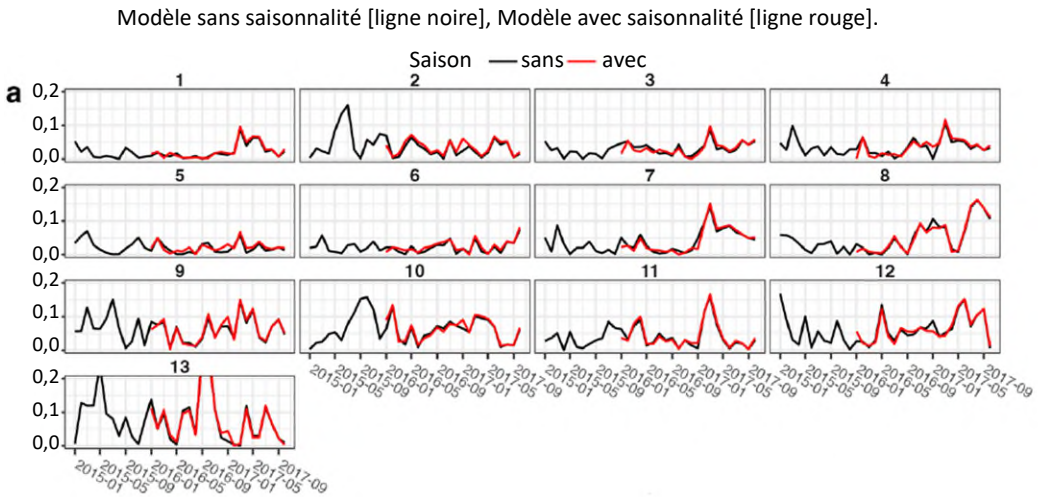
Par conséquent, nous évaluons le rendement de strates individuelles mesuré selon le critère mentionné à la section 3.4. Comme plus haut, nous appliquons l'approche de la fenêtre dynamique tout au long d'une série chronologique. À titre d'illustration, cela est utilisé dans les plans non adaptatifs. Dans le cadre d'une application aux PEA, les répartitions doivent être optimisées de nouveau pour la fenêtre temporelle suivante à l'aide de la stratégique de la section 3.3.

Cette fenêtre temporelle se déplace à mesure que la largeur augmente pour inclure la nouvelle période temporelle suivante. Au trimestre q, nous pouvons évaluer le rendement de la prédiction pour chaque strate en substituant des estimations individuelles *a posteriori* des propensions à répondre et des réalisations individuelles dans (3.8) à (3.10), c'est-à-dire $\text{REQM}(g,q)$, $B(g,q)$ et é.-t.(g,q). Puisque cette analyse est itérée de manière continue, une série chronologique suffisamment longue permet de bien comprendre la variation de chaque rendement de prédiction de strate au fil du temps.

La figure 4.3 montre que lors d'une comparaison des modèles sans saisonnalité (courbes rouges) et des modèles avec saisons (courbes noires), l'introduction de la saisonnalité ne déclenche probablement pas de

réduction efficace du biais et de la variance. Cela s'observe clairement pour pratiquement tous les trimestres, à l'exception du trimestre comprenant les mois de janvier à mars 2017 dans certaines strates (comme la strate 8) pour les estimations des biais et de la REQM. Selon les observations des sections a et b, la variation estimée de la propension à répondre diminue globalement en douceur, ce qui contraste vivement avec le niveau estimé de propension à répondre présentant un comportement volatile. Cette volatilité diffère selon les strates. Les résultats estimés des biais de certaines strates (strates 1 à 8) fluctuent d'environ 0,05 pour tous les trimestres jusqu'à celui commençant en janvier 2017. À partir de cette date, ils présentent une hausse temporaire en raison d'un problème technique à ce moment (voir Wu et coll., 2023 pour une analyse et une éventuelle solution). Lorsque les données d'entraînement sont étendues pour inclure les données « normales », les biais peuvent rapidement diminuer pour atteindre 0,05. Il convient de mentionner que la strate 8 se comporte de manière opposée. Les strates 9 à 13, qui portent sur des tailles d'échantillons relativement petites, obtiennent quant à elles des estimations de l'espérance de la propension à répondre relativement plus biaisées pour la plupart des trimestres.

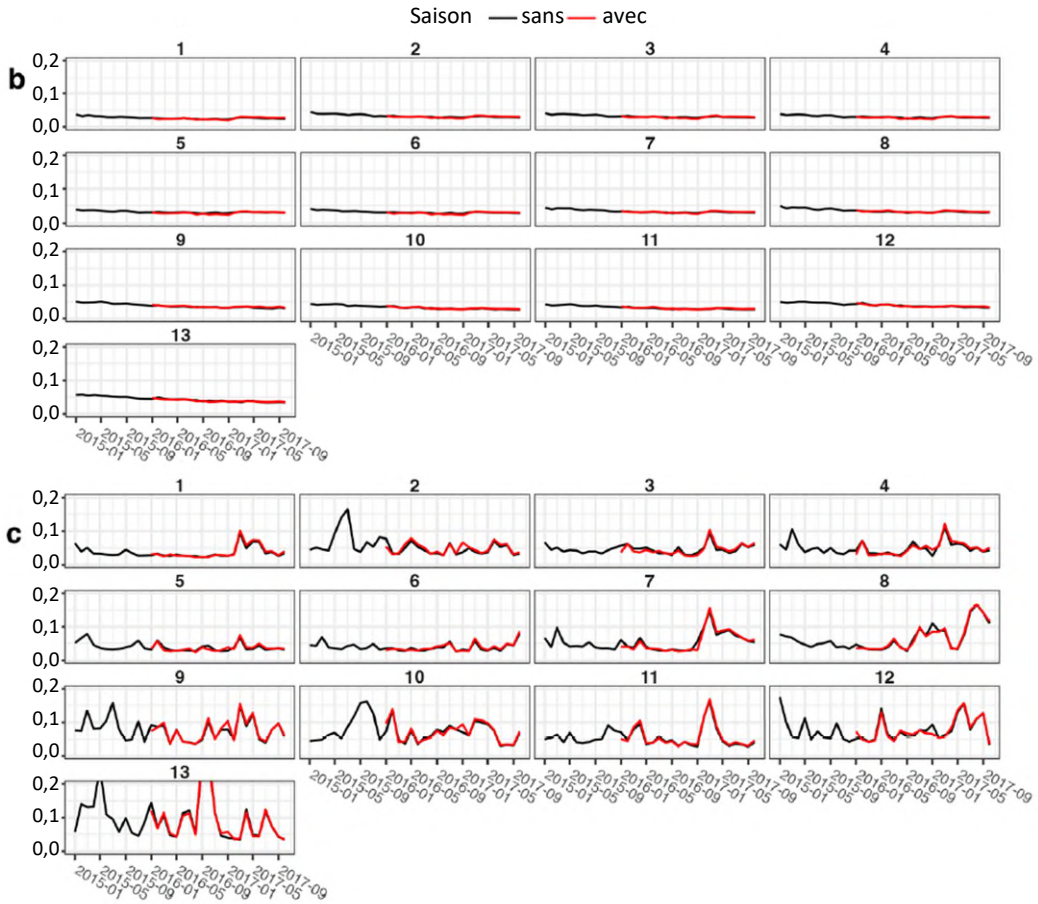
Figure 4.3 Moyennes mobiles prévisionnelles à une étape du biais estimé (section a), de l'écart-type (section b) et de la racine carrée de l'erreur quadratique moyenne (section c) de la propension à répondre au niveau de la strate.



Note : La courbe « noire » représente le modèle sans saisonnalité, alors que la courbe « rouge » fait référence au modèle avec saison. Les deux modèles incluent les corrélations entre les IWAO et les IPAO concernant les prédictions de propensions. L'axe horizontal représente le moment où une décision de PEA est prise.

Figure 4.3(suite) Moyennes mobiles prévisionnelles à une étape du biais estimé (section a), de l'écart-type (section b) et de la racine carrée de l'erreur quadratique moyenne (section c) de la propension à répondre au niveau de la strate.

Modèle sans saisonnalité [ligne noire], Modèle avec saisonnalité [ligne rouge].



Note : La courbe « noire » représente le modèle sans saisonnalité, alors que la courbe « rouge » fait référence au modèle avec saison. Les deux modèles incluent les corrélations entre les IWAO et les IPAO concernant les prédictions de propensions. L'axe horizontal représente le moment où une décision de PEA est prise.

Finalement, les résultats de l'analyse indiquent que lors de la modélisation d'une série chronologique courte, les effets saisonniers, lorsque l'on suppose qu'ils sont semblables pour différents modes, peuvent être moins importants pour l'amélioration du rendement d'un PEA. Lorsque davantage de données d'entraînement sont disponibles, le rendement du PEA peut être constamment amélioré jusqu'à un moment donné; cela sous-entend qu'une règle pour arrêter la collecte de données peut être mise en place et qu'une stratégie fondée sur l'effort peut être adoptée pour les strates présentant de petites tailles d'échantillon.

5. Analyse

En fonction du budget d'une enquête, un PEA cherche la correspondance optimale entre le comportement des répondants et les caractéristiques du plan, c'est-à-dire un ensemble de règles de décision qui peut être

déterminé par des approches d'optimisation (voir Schouten et coll., 2017 pour connaître les avantages et les inconvénients de diverses démarches). Servant d'intrants principaux pour l'optimisation de PEA, des estimations exactes de paramètres de plans d'enquête, comme des propensions à répondre, sont nécessaires pour toute stratégie fiable. À proprement parler, l'inexactitude compromet le rendement et la conception d'un PEA du fait des décisions sous-optimales et inefficaces prises dans le cadre de la démarche d'optimisation. Une incidence négative est évidente lorsque les propensions à répondre varient progressivement au fil du temps.

Dans la présente étude, nous avons discuté d'une méthodologie permettant d'évaluer l'incidence de facteurs temporels (par exemple la saisonnalité) sur l'exactitude de prédictions de propensions séquentielles à répondre dans une enquête multimodale à répliques, et exploré la manière d'appliquer des méthodes de répartition optimale à des strates de population et l'échéancier correspondant. Nous avons introduit un modèle chronologique multinomial bayésien pour les propensions séquentielles à répondre et un modèle d'optimisation de PEA. La forme générale du modèle relatif à la propension décrivait de multiples facteurs temporels et liés aux strates et tenait compte de la dépendance des propensions à répondre du mode actuel envers les comportements de réponse des modes précédents. Le modèle d'optimisation, en revanche, permettait l'inclusion de l'incertitude relative à la charge de travail de suivi et décrivait la façon d'affecter des réviseurs à chaque strate pour favoriser la plus forte diminution du risque de non-réponse. Ce cadre correspond à la plupart des enquêtes transversales multimodales ou à mode unique menées sur de nombreuses années. De plus, nous avons élaboré une analyse pour l'enquête GEZO, afin d'examiner le rendement le plus élevé du modèle de propension. Du fait de diverses compositions de modèle, des critères d'information mesurant la qualité de l'ajustement et la complexité du modèle de propension ont été adoptés pour comparer le rendement de différents modèles. Nous avons ainsi pu atteindre le premier objectif de cette étude visant à sélectionner et à élaborer le modèle de série chronologique « favori » (le modèle 9 enregistrant les résultats de critères d'information les plus bas) contribuant le plus à l'exactitude des prédictions pour une enquête séquentielle multimodale.

Les deuxième et troisième objectifs ont été d'examiner la sensibilité du rendement de PEA au niveau budgétaire donné et au volume de données historiques incluses, respectivement. Dans l'évaluation, le rendement des PEA a dû être optimisé à nouveau, lors de la mise à jour du niveau budgétaire ou de la durée de la série chronologique de données historiques applicables. Ainsi, nous avons étalonné le rendement des PEA par rapport au rendement des plans par IWAO uniquement et des plans non adaptatifs. Cette analyse est essentiellement une comparaison de la réduction du risque de non-réponse, lors de l'affectation à des intervieweurs d'une fraction de non-réponses aux IWAO (sans aucun suivi et avec suivi complet comme cas particulier). Pour que cela soit comparable pour un éventail de scénarios, nous avons utilisé les propriétés des répartitions *a posteriori*, c'est-à-dire la région crédible et l'espérance. L'évaluation a permis d'examiner, pour une fenêtre de temps précise, l'amélioration du rendement pour différents niveaux budgétaires. De

plus, elle a permis d'examiner, pour un niveau budgétaire particulier, l'amélioration par rapport à des fenêtres temporelles dynamiques. L'évaluation a révélé que le rendement d'un PEA était relativement robuste pour des niveaux budgétaires supérieurs à 50 %, mais était moins bon pour des niveaux budgétaires inférieurs à 50 %. Cette évaluation a en outre montré que lorsque nous ne tenions pas compte de la saisonnalité, le rendement d'un PEA était clairement amélioré au début de l'accumulation des données. Ensuite, cette tendance ralentissait, voire disparaissait, malgré l'augmentation des données; le comportement étant pratiquement semblable au modèle avec saison, et laissait entendre, par conséquent, que la saisonnalité n'améliorait pas davantage l'exactitude des prédictions et le rendement du PEA.

Notre étude comporte des limites exigeant de plus amples recherches et la reproduction dans d'autres enquêtes multimodales.

Ignorer les effets saisonniers propres au mode a été notre première simplification. Cela a simplifié la complexité de la spécification des modèles, mais a mené à apparemment compenser les effets saisonniers sur les prédictions de propensions. Nous pensons, toutefois, qu'il est possible d'adapter de manière pratique au modèle ajusté les effets saisonniers propres à chaque modèle, si la saisonnalité s'avère un prédicteur puissant des prédictions de propensions.

Pour assurer l'exactitude et la fiabilité des prédictions, nous avons uniquement envisagé deux stratégies de collecte de données candidates (uniquement en ligne et en ligne suivi d'interviews en personne) comme deuxième simplification. Le nombre de visites d'IPAO aux unités d'échantillonnage peut être encore personnalisé et l'optimisation peut comprendre le nombre réel de visites. Les propensions à répondre après chaque visite peuvent être modélisées et estimées simultanément, et l'on pourrait supposer que les prédictions d'un mode de suivi sont corrélées uniquement à son prédécesseur immédiat. Une telle application est simple, mais fait intervenir une vérification attentive de la fiabilité des prédictions.

La troisième limite a été la sensibilité de notre modèle de propension aux variations structurelles du plan (voir Wu et coll., 2023 pour une analyse plus approfondie). Par conséquent, dans l'étude de cas de la présente étude, nous avons noté une erreur de spécification temporaire dans les répartitions *a priori* des propensions à répondre. Dans un effort de caractérisation d'une variation imprévue, il est important de déterminer les paramètres (et strates) touchés. La robustesse peut alors être améliorée par des paramètres de modèle hiérarchiques supplémentaires. Nous reportons cette extension à des études ultérieures.

Notre dernière limite était que, dans nos modèles hiérarchiques de série chronologique, nous avons supposé que les strates du PEA étaient précisées d'avance. Nous n'avons pas intégré d'étape de sélection de variables dans l'ajustement et l'optimisation du modèle. Lorsque le nombre de variables auxiliaires est important, un compromis intervient entre le temps d'entraînement de l'analyse bayésienne des propensions à répondre et l'utilité de l'optimisation; il s'agit d'un sujet important pour des études ultérieures.

L'inclusion dans les modèles hiérarchiques de série chronologique de paradonnées et d'autres covariables variant dans le temps serait une autre extension pertinente. Cela permettrait l'optimisation dans un

cadre dynamique, c'est-à-dire au cours du travail sur le terrain. Dans le cadre d'une interview sur place, une telle approche dynamique n'est pas simple du point de vue opérationnel. Les charges de travail des intervieweurs sont uniquement connues à un moment proche du travail sur le terrain et à des moments définis, souvent mensuels. Ce serait également le cas, dans une moindre mesure, pour un suivi téléphonique. Une solution pratique, appliquée par le bureau central de la statistique des Pays-Bas, est d'ajouter un sous-échantillonnage aléatoire de non-répondants. Les probabilités de sous-échantillonnage dépendent des charges de travail prédéterminées et fixes dans les régions des intervieweurs. On peut également envisager une étape supplémentaire permettant à un plan dynamique d'intervenir lors du suivi en personne. Afin de laisser une certaine liberté aux intervieweurs, cela peut être mis en œuvre sous la forme d'un ou de plusieurs points de liaisons prédéfinis d'arrêt ou non des unités d'échantillonnage. En considérant en plus les étapes de collecte de données générales (séquentielles), il serait utile d'étendre la méthodologie de la présente étude. Une telle extension fait face à trois défis statistiques. Le premier est l'inclusion de nouvelles données auxiliaires d'intrants (par exemple les paradonnées) dans le modèle et l'optimisation. La décision et la façon d'inclure ces données seraient un compromis entre l'efficacité et l'efficience. Le deuxième est qu'il est nécessaire de mettre davantage l'accent sur la spécification correcte des covariances des propensions à répondre des différentes étapes de collecte de données. Les erreurs de spécifications prolifèrent dans les étapes suivantes. Le troisième est que l'optimisation doit être effectuée pendant la collecte de données, ce qui exige de limiter ses ambitions.

Annexe A

Tableau A.1
Les variables auxiliaires forment 13 strates et la saisonnalité est jugée être un facteur influant sur la prédiction des propensions à répondre.

Strate	Âge	Ethnicité
1	0 à 17 ans	Occidentaux
2	18 à 24 ans	Occidentaux
3	25 à 34 ans	Occidentaux
4	35 à 54 ans	Occidentaux
5	55 à 64 ans	Occidentaux
6	65 à 74 ans	Occidentaux
7	75 ans et plus	Occidentaux
8	0 à 17 ans	Non occidentaux
9	18 à 24 ans	Non occidentaux
10	25 à 34 ans	Non occidentaux
11	35 à 54 ans	Non occidentaux
12	55 à 64 ans	Non occidentaux
13	65 ans et plus	Non occidentaux

Annexe B

Tableau B.1
Observations des coefficients de variation tenant compte du biais (corrigés) ou non (non corrigés) et des erreurs-types (ET) lors de la prise en compte du biais au premier trimestre 2017 pour différents niveaux budgétaires.

	W			w-Np			w-Ap		
	Non corrigés	Corrigés	ET	Non corrigés	Corrigés	ET	Non corrigés	Corrigés	ET
0 %*	0,305	0,294	0,023	-	-	-	-	-	-
100 %	-	-	-	0,156	0,149	0,013	0,102	0,072	0,021
90 %	-	-	-	-	-	-	0,093	0,084	0,021
80 %	-	-	-	-	-	-	0,104	0,074	0,020
70 %	-	-	-	-	-	-	0,092	0,078	0,021
60 %	-	-	-	-	-	-	0,093	0,084	0,021
50 %	-	-	-	-	-	-	0,106	0,052	0,020
40 %	-	-	-	-	-	-	0,114	0,072	0,021
30 %	-	-	-	-	-	-	0,136	0,115	0,022
20 %	-	-	-	-	-	-	0,259	0,237	0,022
10 %	-	-	-	-	-	-	0,305	0,294	0,023

**Aucun budget indique seulement que le mode IWAO est utilisé. « - » désigne aucun résultat*. w désigne les plans par IWAO uniquement, w-Ap désigne les plans non adaptatifs, w-Np désigne les plans adaptatifs.

Tableau B.2
Observations de c.v. tenant compte du biais (corrigés) ou non (non corrigés) et erreurs-types (ET) lors de la prise en compte du biais pour le niveau budgétaire de 100 % pour chaque trimestre.

Année	Trimestre	w			w-Np			w-Ap		
		Non corrigés	Corrigés	ET	Non corrigés	Corrigés	ET	Non corrigés	Corrigés	ET
2016	Q1				-	-	-	-	-	-
	Q2	-	-	-						
	Q3	-	-	-	-	-	-	0,093	0,084	0,021
	Q4	-	-	-	-	-	-	0,104	0,074	0,020
2017	Q1	-	-	-	-	-	-	0,092	0,078	0,021
	Q2	-	-	-	-	-	-	0,093	0,084	0,021
	Q3	-	-	-	-	-	-	0,106	0,052	0,020
	Q4	-	-	-	-	-	-	0,114	0,072	0,021

Note : w désigne les plans par IWAO uniquement, w-Ap désigne les plans non adaptatifs, w-Np désigne les plans adaptatifs.

Tableau B.3
Coût relatif (c) du premier trimestre 2017 pour différents niveaux budgétaires pour des enquêtes adaptatives.

	100 %	90 %	80 %	70 %	60 %	50 %	40 %	30 %	20 %	10 %
c	0,425	0,473	0,532	0,608	0,709	0,851	0,983	1,311	1,966	0,977
C	VRAI	VRAI	VRAI	VRAI	VRAI	VRAI	VRAI	FAUX	FAUX	FAUX

Note : Les répartitions dans chaque cas sont déterminées par le résolveur d'optimisation « auglag » en commençant au point initial 0 défini pour les 13 strates. Si la convergence (C) indique VRAI, il est possible de trouver l'optimum local et de produire les répartitions correspondantes; autrement, le processus mène à un faux « optimum » local.

Bibliographie

Boonstra, H.-J. (2022). *mcmcsae*: Markov Chain Monte Carlo small area estimation. R package version 0.7.2.

Boonstra, H.-J., et van den Brakel, J.A. (2019). Estimation du niveau et de la variation du chômage au moyen de modèles de séries chronologiques structurels. *Techniques d'enquête*, 45, 3, 421-454. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2019003/article/00005-fra.pdf.

Calinescu, M., Bhulai, S. et Schouten, B. (2013). Optimal resource allocation in survey designs. *European Journal of Operational Research*, 226(1), 115-121. DOI: https://doi.org/10.1016/j.ejor.2012.10.046.

Chen, Q., Elliott, M.R. et Little, R.J.A. (2010). Inférence basée sur un modèle bayésien avec splines pénalisées pour les proportions de population finie dans l'échantillonnage avec probabilités inégales. *Techniques d'enquête*, 36, 1, 25-37. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2010001/article/11250-fra.pdf.

Chun, A.Y., Heeringa, S.G. et Schouten, B. (2018). Responsive and adaptive design for survey optimization. *Journal of Official Statistics*, 34(3), 581-597. DOI: https://doi.org/10.2478/jos-2018-0028.

Coffey, S., West, B.T., Wagner, J. et Elliott, M.R. (2020). What do you think? Using expert opinion to improve predictions of response propensity under a Bayesian framework. *Methods, Data, Analyses*. DOI: https://doi.org/10.12758/mda.2020.05.

Gelman, A., et Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

Gelman, A., et Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457-472.

Linderman, S., Johnson, M.J. et Adams, R.P. (2015). Dependent multinomial models made easy: Stick-breaking with the polya-gamma augmentation. *Advances in Neural Information Processing Systems*, 28. DOI: https://doi.org/10.48550/arXiv.1506.05843.

Ma, Y. (2021). *Optimal Stratification in Bayesian Adaptive Survey Designs*. Thèse de doctorat, University Utrecht, Les Pays-Bas.

Moore, J.C., Durrant, G.B. et Smith, P.W.F. (2018). Data set representativeness during data collection in three UK social surveys: Generalizability and the effects of auxiliary covariate choice. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(1), 229-248. DOI: https://doi.org/10.1111/rssa.12256.

Nishimura, R., Wagner, J. et Elliott, M. (2016). Alternative indicators for the risk of non-response bias: A simulation study. *Revue Internationale de Statistique*, 84(1), 43-62. DOI: https://doi.org/10.1111/INSR.12100.

O'Malley, A.J., et Zaslavsky, A.M. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association*, 103(484), 1405-1418. DOI: https://doi.org/10.1198/016214508000000724.

Polson, N.G., Scott, J.G. et Windle, J. (2013). Bayesian inference for logistic models using Pólya-gamma latent variables. *Journal of the American Statistical Association*, 108, 1339-1349.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. et Lindblad, M. (2010). Reduction of nonresponse bias through case prioritization. *Survey Research Methods*, 4(1), 21-29. DOI: https://doi.org/10.18148/SRM/2010.V4I1.3037.

Särndal, C.-E., et Lundström, S. (2010). Plan d'estimation : détermination de vecteurs auxiliaires en vue de réduire le biais de non-réponse. *Techniques d'enquête*, 36, 2, 141-156. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2010002/article/11376-fra.pdf.

Schouten, B., Calinescu, M. et Luiten, A. (2013). Optimiser la qualité de la réponse au moyen de plans de collecte adaptatifs. *Techniques d'enquête*, 39, 1, 33-66. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2013001/article/11824-fra.pdf.

Schouten, B., Cobben, F. et Bethlehem, J. (2009). Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête*, 35, 1, 107-121. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2009001/article/10887-fra.pdf.

Schouten, B., Mushkudiani, N., Shlomo, N., Durrant, G., Lundquist, P. et Wagner, J. (2018). A Bayesian analysis of design parameters in survey data collection. *Journal of Survey Statistics and Methodology*. DOI: https://doi.org/10.1093/jssam/smy012.

Schouten, B., Peytchev, A. et Wagner, J. (2017). *Adaptive Survey Design*. Chapman and Hall/CRC.

Scott, A., et Smith, T.M. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association*, 64(327), 830-840.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. et van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583-639. DOI: https://doi.org/10.1111/1467-9868.00353.

van Berkel, K., van der Doef, S. et Schouten, B. (2020). Implementing adaptive survey design with an application to the Dutch Health Survey. *Journal of Official Statistics*, 36(3), 609-629. DOI: https://doi.org/10.2478/jos-2020-0031.

Varadhan, R. (2022). *alabama: Constrained Nonlinear Optimization*. R package version 2022.4-1.

Vehtari, A., Gelman, A. et Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27, 1413-1432. DOI: https://doi.org/10.1007/s11222-016-9696-4.

Wagner, J. (2013). Adaptive contact strategies in telephone and face-to-face surveys. *Survey Research Methods*, 7(1), 45-55. DOI: https://doi.org/10.18148/SRM/2013.V7I1.5037.

Wagner, J., et Hubbard, F. (2013). Using propensity models during data collection for responsive designs: Issues with estimation. Lors de la 68e conference de l'AAPOR, mai (pp. 16-19).

Wagner, J.R. (2008). *Adaptive Survey Design to Reduce Nonresponse Bias*. (Thèse de doctorat, University of Michigan).

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571-3594. https://arxiv.org/abs/1004.2316.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14, 867-897. https://arxiv.org/abs/1208.6338.

West, B.T., Wagner, J., Coffey, S. et Elliott, M.R. (2023). Deriving priors for Bayesian prediction of daily response propensity in responsive survey design: Historical data analysis vs. literature review. *Journal of Survey Statistics and Methodology,* 11(2), 367-392.

Wu, S., Boonstra, H.-J., Moerbeek, M. et Schouten, B. (2023). Modélisation de la variation temporelle des taux de réponse aux enquêtes : approche bayésienne s'appliquant à l'enquête sur la santé réalisée aux Pays-Bas. *Techniques d'enquête*, 49, 1, 177-208. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2023001/article/00010-fra.pdf.

Wu, S., Schouten, B., Meijers, R. et Moerbeek, M. (2022). Data collection expert prior elicitation in survey design: Two case studies. *Journal of Official Statistics*, 38(2), 637-662. DOI: https://doi.org/10.2478/JOS-2022-0028.

Techniques d'enquête, décembre 2024
Vol. 50, nº 2, pp. 507-530
Statistique Canada, nº 12-001-X au catalogue

Inférence prédictive bayésienne d'une moyenne de population finie sans préciser la relation entre la variable d'étude et les covariables

Ashley Lockwood et Balgobin Nandram[1]

Résumé

Tout en évitant de préciser la relation paramétrique entre la variable d'étude et les covariables, nous illustrons l'avantage d'inclure une composante spatiale pour mieux tenir compte des covariables dans nos modèles en vue de faire une inférence prédictive bayésienne. Nous traitons chaque combinaison unique de covariables comme une strate individuelle, puis nous utilisons les techniques d'estimation sur petits domaines pour faire une inférence à propos de la population moyenne finie de la variable de réponse continue. Les deux modèles spatiaux utilisés sont le modèle d'autorégression conditionnel et le modèle d'autorégression conditionnelle simple. Nous incluons les effets spatiaux en créant la matrice d'adjacence à partir de la distance de Mahalanobis entre les covariables. Nous démontrons également la façon d'incorporer les poids d'enquête dans les modèles spatiaux en cas de données d'enquête probabiliste. Nous comparons les résultats des deux modèles non spatiaux, à savoir le modèle Scott-Smith et le modèle Battese, Harter et Fuller, aux modèles spatiaux. Nous illustrons la comparaison entre les deux modèles mentionnés et une application au moyen des données sur l'indice de masse corporelle de huit comtés en Californie. Notre but est d'obtenir des strates voisines donnant des prédictions similaires et d'augmenter la différence entre les strates qui ne sont pas voisines. Finalement, l'utilisation des modèles spatiaux montre un regroupement global moindre par rapport aux modèles non spatiaux, ce qui correspond au résultat souhaité.

Mots-clés : Modèle d'autorégression conditionnelle; modèle d'autorégression conditionnelle simple; modèle hiérarchique bayésien; modélisation spatiale.

1. Introduction

Dans le présent article, lorsque nous faisons l'inférence de la moyenne de la population finie, nous nous gardons de présumer une relation entre la variable réponse et les covariables. Nous évitons de formuler les fortes hypothèses des modèles de régression et, par conséquent, nous augmentons le nombre de situations auxquelles nos modèles peuvent s'appliquer. Les méthodes que nous présentons évitent de définir cette relation en considérant chaque combinaison unique de covariables dans la population comme une strate individuelle. Nous adaptons les covariables en utilisant le modèle spatial plutôt qu'un modèle de régression. Puis, nous utilisons les techniques d'estimation sur petits domaines pour produire une inférence à propos de chaque strate de la population en fonction de ses covariables sous-jacentes. Nous pouvons alors comprendre la population globale en regroupant les prédictions des strates (Rao et Molina, 2015).

Nous présentons deux versions des modèles spatiaux, un modèle d'autorégression conditionnelle (CAR pour « conditional autoregressive ») et un modèle d'autorégression conditionnelle simple (SCAR pour « simple conditional autoregressive ») (Chung et Datta, 2022). Pour les modèles spatiaux, nous incluons les effets spatiaux en créant la matrice d'adjacence à partir de la distance de Mahalanobis entre les covariables de chaque strate. Nous utilisons ces modèles spatiaux pour créer une relation de voisinage entre les strates

1. Ashley Lockwood, Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA. Courriel : anlockwood@wpi.edu; Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA. Courriel : balnan@wpi.edu.

similaires pour réduire le regroupement global de la moyenne d'échantillon globale. En permettant aux strates d'avoir des voisins, nous pensons que les voisinages se regrouperont sans regrouper des strates éloignées. Le recours au modèle spatial plutôt qu'au modèle non spatial devrait fournir des prédictions *a posteriori* ayant une plus grande variation des moyennes de strate prédites.

Nous présentons également deux modèles non spatiaux aux fins de comparaison, soit une forme du modèle Scott-Smith (Scott et Smith, 1969) et une forme du modèle Battese, Harter et Fuller (BHF) (Battese, Harter et Fuller, 1988). Le modèle BHF est une version plus générale du modèle Scott-Smith qui comprend les covariables du modèle, alors que le modèle Scott-Smith n'en comprend pas. Nous utilisons les deux modèles non spatiaux comme base de comparaison pour constater l'effet de l'inclusion d'une relation spatiale dans les modèles. L'annexe A contient les renseignements techniques complets du modèle Scott-Smith et l'annexe B, les renseignements techniques du modèle BHF.

En outre, Datta et Ghosh (1991) poursuivent la recherche menée par Battese, Harter et Fuller (1988) en offrant une analyse complète du modèle hiérarchique bayésien de régression à erreurs emboîtées qui cible en particulier l'estimation sur petits domaines. Ce travail fait progresser de manière importante la généralisation de formules de calcul permettant de dériver les prédicteurs bayésiens et les erreurs-types qui y sont associées. Toutefois, il faut souligner que l'approche de Datta et Ghosh (1991) continue de s'appuyer sur des modèles linéaires mixtes, qui établissent explicitement la relation entre les covariables et la variable de réponse. Par contre, la méthodologie que nous proposons dans le présent article évite intentionnellement de telles définitions explicites de relation, ce qui donne lieu à une différence méthodologique importante.

Il y a plusieurs modèles de régression traditionnels qui font des inférences à propos d'une caractéristique d'une population, notamment la régression logistique, les modèles linéaires généraux, les modèles normaux multivariés généraux et l'arbre de classification et de régression (CART pour « classification and regression tree »). Voir Lindley et Smith (1972), Ghosh, Natarajan, Stroud et Carlin (1998), Albert et Chib (1993), Box et Tiao (1973), ainsi que Chipman, George et McCulloch (1998) pour obtenir des précisions sur chaque modèle. Bien que ces modèles aient été largement utilisés au fil du temps, les fortes hypothèses de distribution émises pour l'utilisation appropriée de ces modèles limitent les types de données et de situations permettant leur application.

Il existe aussi d'autres modèles ne comportant pas de coefficients de régression qui répondent à une question similaire, notamment les modèles de Dirichlet, le modèle de l'urne de Polya et les arbres de régression additive bayésienne (BART pour « Bayesian additive regression trees »). Voir Blackwell et MacQueen (1973), Antoniak (1974), Yin et Nandram (2020), Teh, Jordan, Beal et Blei (2006) ainsi que Chipman, George et McCulloch (2010) pour obtenir des renseignements sur ces autres modèles. Les modèles de Dirichlet et les modèles de l'urne de Polya sont populaires dans la modélisation bayésienne; toutefois, ces calculs complexes peuvent mener à un mauvais mélange dans l'algorithme de la méthode de Monte-Carlo par chaîne de Markov (MCMC pour « Markov chain Monte Carlo »). La méthode BART est plus nouvelle, mais elle va à l'encontre de la logique bayésienne traditionnelle en faisant un double usage des données. Les données sont utilisées dans la probabilité du modèle BART, puis de nouveau dans les

données préalables de deux hyperparamètres (Hill, Linero et Murray, 2020). Nous cherchons à améliorer le calcul des modèles sans coefficient de régression tout en maintenant la cohérence du paradigme bayésien.

Le reste du présent article portera sur la méthodologie des deux modèles spatiaux et sur l'ajout de poids d'enquête dans les modèles de la section 2. Puis, dans la section 3, nous traiterons de l'application des données sur l'indice de masse corporelle (IMC) à chacun des modèles. La conclusion est présentée à la section 4. L'annexe A contient les renseignements techniques relatifs au modèle Scott-Smith et la façon d'inclure les poids d'enquête dans ce modèle. De même, l'annexe B présente les renseignements techniques relatifs au modèle BHF et la façon d'inclure les poids d'enquête.

2. Méthodologie

Dans la section qui suit, nous proposons deux modèles spatiaux et la façon d'y inclure les poids d'enquête. D'abord, dans la section 2.1, nous présentons les modèles spatiaux, soit le modèle CAR dans la section 2.1.1 et le modèle SCAR dans la section 2.1.2. Puis, dans la section 2.2, nous illustrons la façon d'inclure les poids d'enquête dans les modèles spatiaux présentés dans la section 2.1. Les méthodes propres au modèle Scott-Smith et au modèle BHF sont décrites dans l'annexe A et l'annexe B, respectivement.

Dans les quatre modèles, nous observons une variable de réponse continue y_{ij} pour l'unité d'échantillonnage $j = 1,\ldots,n_i,$ appartenant à la strate $i = 1,\ldots,\ell,$ et ces réponses sont regroupées en fonction de la valeur de leurs covariables. Chaque combinaison possible de covariables est examinée, et chaque combinaison unique est considérée comme une strate. Par conséquent, chaque $\mathbf{y}_i$ a une combinaison de variables, unique et correspondante, désignée par $\mathbf{x}_i,$ où $\mathbf{y}_i$ est le vecteur regroupé des réponses de la longueur n_i de chaque strate. La matrice de covariables $\mathbf{X} = \left(\mathbf{x}_i'\right)$ a la dimension $\ell \times p,$ où p est le nombre de covariables dans les données. La matrice $\mathbf{X}$ ne comprend pas la colonne de l'ordonnée à l'origine et $\mathbf{x}_i$ correspond aux lignes uniques de $\mathbf{X}$. Nous faisons l'inférence de la moyenne de la population finie, $\bar{Y}_i = \sum_{j=1}^{N_i} y_{ij} / N_i,$ en fonction des données observées de $\mathbf{y}_i$. Nous désignons la fraction d'échantillonnage par $f_i = n_i / N_i,$ où n_i représente la taille de l'échantillon et N_i représente la taille de la population d'une strate donnée. Dans notre application, les N_i sont inconnues. Nous aborderons ultérieurement la façon de les estimer à l'aide de la pondération de probabilité inverse. Dans la mesure du possible, l'utilisation des tailles réelles de population, $N_i,$ est celle optimale et privilégiée. Lorsque le nombre de strates est exceptionnellement élevé, les estimations de N_i peuvent devenir de plus en plus sensibles au bruit.

2.1 Modèles spatiaux

Pour les modèles spatiaux, nous incluons les effets spatiaux en créant la matrice d'adjacence symétrique, $\mathbf{W}$ de dimension $\ell \times \ell,$ au moyen de la distance de Mahalanobis entre $\mathbf{x}_i$ et $\mathbf{x}_{i'}$ pour $i = 1,\ldots,\ell,$ $i' = 1,\ldots,\ell,$ et $i \neq i'.$ La distance de Mahalanobis est définie comme suit :

$$d_{ii'} = \sqrt{\left(\mathbf{x}_i - \mathbf{x}_{i'}\right)' \mathbf{S}^{-1} \left(\mathbf{x}_i - \mathbf{x}_{i'}\right)}, \tag{2.1}$$

où $\mathbf{S}$ est la matrice de covariance de $\mathbf{X}$ et $d_{ii} = 0$. Nous définissons $\mathbf{W}$ en établissant $w_{ii'} = 1$ si $d_{ii'} \le d_0$ et $w_{ii'} = 0$ si $d_{ii'} > d_0$, les zéros étant sur la diagonale (soit, $w_{ii} = 0$). Une recherche par quadrillage est menée pour déterminer la valeur de d_0 qui donne une matrice $\mathbf{W}$ qui maximise I de Moran, défini comme suit :

$$I = \frac{\ell}{w_{\cdot\cdot}} \frac{\sum_i \sum_{i'} w_{ii'} (\bar{y}_i - \bar{y})(\bar{y}_{i'} - \bar{y})}{\sum_i (\bar{y}_i - \bar{y})^2}, \tag{2.2}$$

où $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} / n_i$ est la variable réponse; $\bar{y} = \sum_{i=1}^{\ell} \bar{y}_i / \ell$ est la réponse moyenne de l'échantillon global, $w_{ii'}$ correspond aux éléments de $\mathbf{W}$ et $w_{\cdot\cdot} = \sum_i \sum_{i'} w_{ii'}$.

Dans la section 2.1.1, nous décrivons le modèle CAR et dans la section 2.1.2, nous établissons la différence entre ce modèle et le modèle SCAR.

2.1.1 Modèle d'autorégression conditionnelle

Le modèle d'autorégression conditionnelle (CAR pour « conditional autoregressive ») hiérarchique bayésien est :

$$\begin{aligned} y_{ij} \mid \boldsymbol{\mu}, \sigma^2 &\sim \text{Normal}\left(\mu_i, \sigma^2\right), \\ \boldsymbol{\mu} \mid \theta, \rho, \sigma^2, \gamma &\sim \text{Normal}\left(\theta \mathbf{1}, \frac{\rho}{1-\rho} \sigma^2 \left(\mathbf{R} - \gamma \mathbf{W}\right)^{-1}\right), \\ \pi\left(\theta, \sigma^2, \rho\right) \propto \frac{1}{\sigma^2}, \gamma &\sim \text{Uniform}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_\ell}\right), \frac{1}{\lambda_1} \le \gamma \le \frac{1}{\lambda_\ell}, -\infty < \theta < \infty, 0 < \rho < 1, \sigma^2 > 0, \\ & j = 1, \ldots, n_i, \; i = 1, \ldots, \ell, \end{aligned} \tag{2.3}$$

où ℓ, dans ce cas, représente le nombre total de combinaisons possibles de covariables qui sont considérées comme des strates individuelles. Nous discrétisons les variables continues de sorte qu'il y a un nombre fini de combinaisons possibles de covariables. Puis, nous gardons les réponses continues, $\mathbf{y}_i$, de sorte qu'il y a n_i réponses pour chaque strate $i = 1, \ldots, \ell$. Ici, $\mathbf{R}$ est une matrice de précision diagonale $\ell \times \ell$ définie par $\mathbf{R} = \text{diag}\left\{w_{i\cdot}\right\}_{i=1}^{\ell}$, où $w_{i\cdot} = \sum_{j=1}^{n_i} w_{ij}$ est la somme de la i^{e} ligne de $\mathbf{W}$. De plus, λ_1 est la valeur propre minimale de $\mathbf{R}^{-1}\mathbf{W}$ et λ_ℓ est la valeur propre maximale de $\mathbf{R}^{-1}\mathbf{W}$, et puisque $\sum_{i=1}^{\ell} w_{ii} = 0$, on obtient $\lambda_1 < 0 < \lambda_\ell$ (Chung et Datta, 2022). Ici, $\left(\mathbf{R} - \gamma \mathbf{W}\right)$ sera définie positivement pour autant que γ est dans l'intervalle $\frac{1}{\lambda_1} \le \gamma \le \frac{1}{\lambda_\ell}$. Pour obtenir des échantillons de la densité *a posteriori* conjointe de ce modèle, nous pouvons éliminer par intégration $\boldsymbol{\mu}$, θ et σ^2; il nous suffira ensuite simplement d'établir γ et ρ à l'aide d'un échantillonneur de Gibbs (Ritter et Tanner, 1992).

Nous pouvons vectoriser la variable de réponse continue, y_{ij}, pour qu'elle soit $\mathbf{y}$ ayant une dimension $n \times 1$, où $n = \sum_{i=1}^{\ell} n_i$ de sorte que

$$\mathbf{y}_{(n\times 1)} \mid \boldsymbol{\mu}, \sigma^2 \sim \text{Normal}\left(A\boldsymbol{\mu}, \sigma^2 \mathbb{I}_{n\times n}\right), \tag{2.4}$$

où $\mathbb{I}_{n\times n}$ est la matrice d'identité et A a la dimension $n \times \ell$ et peut être définie par

$$A = \begin{pmatrix} \mathbf{1_1} & 0 & \cdots & 0 \\ 0 & \mathbf{1_2} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbf{1_\ell} \end{pmatrix}, \tag{2.5}$$

et $\mathbf{1_1}$ à $\mathbf{1_\ell}$ sont des vecteurs des valeurs 1 ayant des longueurs correspondant au nombre d'observations dans cette strate. Par conséquent, $\mathbf{1_1}$ est un vecteur des valeurs 1 ayant une longueur n_1, $\mathbf{1_2}$ est un vecteur des valeurs 1 ayant une longueur n_2, et ainsi de suite jusqu'à $\mathbf{1_\ell}$. Nous écrivons ce modèle de cette façon pour pouvoir utiliser le lemme de la section 2 de l'article de Lindley et Smith (1972) pour obtenir la distribution *a posteriori* de $\boldsymbol{\mu}$ dont nous tirons les échantillons de $\boldsymbol{\mu}$:

$$\begin{aligned} \boldsymbol{\mu} \,|\, \Omega, \mathbf{y} \;\sim\; \text{Normal}\Bigg[& \left(\text{diag}\left(n_1,\ldots,n_\ell\right) + \frac{1-\rho}{\rho}\left(\mathbf{R}-\gamma\mathbf{W}\right) \right)^{-1} \left(A'\mathbf{y} + \left(\frac{1-\rho}{\rho}\left(\mathbf{R}-\gamma\mathbf{W}\right) \right) \theta\mathbf{1} \right), \\ & \sigma^2 \left(\text{diag}\left(n_1,\ldots,n_\ell\right) + \frac{1-\rho}{\rho}\left(\mathbf{R}-\gamma\mathbf{W}\right) \right)^{-1} \Bigg]. \end{aligned} \tag{2.6}$$

Désignons $\Omega = \left(\theta, \rho, \sigma^2, \gamma\right)$ pour simplifier la notation.

Une autre façon d'écrire ce modèle spatial pour le simplifier en vue d'éliminer par intégration $\boldsymbol{\mu}$ serait

$$\begin{aligned} \overline{\mathbf{y}} \,|\, \boldsymbol{\mu}, \sigma^2 \;&\sim\; \text{Normal}\left(\boldsymbol{\mu}, \sigma^2 \text{diag}\left(\frac{1}{n_1}, \ldots, \frac{1}{n_\ell} \right) \right), \\ \boldsymbol{\mu} \,|\, \theta, \rho, \sigma^2, \gamma \;&\sim\; \text{Normal}\left(\theta\mathbf{1}, \frac{\rho}{1-\rho}\sigma^2 \left(\mathbf{R}-\gamma\mathbf{W}\right)^{-1} \right). \end{aligned} \tag{2.7}$$

Maintenant, si nous éliminons par intégration $\boldsymbol{\mu}$ à partir de ce modèle, nous nous retrouvons avec la densité *a posteriori*

$$\begin{aligned} \pi\left(\theta,\rho,\sigma^2,\gamma \,|\, \mathbf{y}\right) \;\propto\; & \det\left[\sigma^2 \left(\text{diag}\left(\frac{1}{n_1},\ldots,\frac{1}{n_\ell} \right) + \frac{\rho}{1-\rho}\left(\mathbf{R}-\gamma\mathbf{W}\right)^{-1} \right) \right]^{-1/2} \\ & \times \exp\left\{ \frac{-1}{2\sigma^2} \left(\overline{\mathbf{y}}-\theta\mathbf{1}\right)' \left(\text{diag}\left(\frac{1}{n_1},\ldots,\frac{1}{n_\ell} \right) + \frac{\rho}{1-\rho}\left(\mathbf{R}-\gamma\mathbf{W}\right)^{-1} \right)^{-1} \left(\overline{\mathbf{y}}-\theta\mathbf{1}\right) \right\} \\ & \times \left(\frac{1}{\sigma^2} \right)^{(n-\ell)/2} \exp\left\{ \frac{-1}{2\sigma^2} \sum_{i=1}^{\ell} \left(n_i - 1\right) s_i^2 \right\} \times \frac{1}{\sigma^2} \end{aligned} \tag{2.8}$$

où $s_i^2 = \frac{\sum_{j=1}^{n_i} (y_{ij}-\overline{y}_i)^2}{(n_i-1)}$. À partir de cette valeur de densité, nous pouvons constater que θ suit une distribution normale

$$\theta \,|\, \sigma^2, \rho, \gamma, \overline{\mathbf{y}} \;\sim\; \text{Normal}\left(\hat{\theta}, \frac{\sigma^2}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}} \right), \tag{2.9}$$

où $\hat{\theta} = \mathbf{1}'\boldsymbol{\Sigma}\bar{\mathbf{y}}/\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}$ et $\boldsymbol{\Sigma} = \left[\text{diag}\left(\frac{1}{n_1},\ldots,\frac{1}{n_\ell}\right) + \frac{\rho}{1-\rho}\left(\mathbf{R}-\gamma\mathbf{W}\right)^{-1}\right]^{-1}$. Nous pouvons utiliser ce fait pour éliminer par intégration θ, et ainsi obtenir

$$\begin{aligned}\pi\left(\rho,\sigma^2,\gamma \mid \mathbf{y}\right) \propto\ & \det\left[\sigma^2\boldsymbol{\Sigma}^{-1}\right]^{-1/2} \\ & \times \exp\left\{\frac{-1}{2\sigma^2}\left(\hat{\theta}\mathbf{1}-\bar{\mathbf{y}}\right)'\boldsymbol{\Sigma}\left(\hat{\theta}\mathbf{1}-\bar{\mathbf{y}}\right)\right\} \times \sqrt{2\pi\sigma^2/\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}} \\ & \times \left(\frac{1}{\sigma^2}\right)^{(n-\ell)/2} \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{\ell}(n_i-1)\,s_i^2\right\} \times \frac{1}{\sigma^2}.\end{aligned} \tag{2.10}$$

À partir de cette valeur de densité, nous pouvons obtenir la distribution gamma inverse

$$\sigma^2 \mid \rho,\gamma,\bar{\mathbf{y}} \sim \text{InvGam}\left(\frac{n-1}{2},\left[\left(\hat{\theta}\mathbf{1}-\bar{\mathbf{y}}\right)'\boldsymbol{\Sigma}\left(\hat{\theta}\mathbf{1}-\bar{\mathbf{y}}\right) + \sum_{i=1}^{\ell}(n_i-1)\,s_i^2\right]\Big/2\right). \tag{2.11}$$

Finalement, après l'élimination par intégration de σ^2, nous obtenons la densité *a posteriori* conjointe non normalisée

$$\pi\left(\rho,\gamma \mid \mathbf{y}\right) \propto \det\left[\boldsymbol{\Sigma}^{-1}\right]^{-1/2}\left[\left(\hat{\theta}\mathbf{1}-\bar{\mathbf{y}}\right)'\boldsymbol{\Sigma}\left(\hat{\theta}\mathbf{1}-\bar{\mathbf{y}}\right)+\sum_{i=1}^{\ell}(n_i-1)\,s_i^2\right]^{\frac{-n+1}{2}}\left(\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}\right)^{-1/2}. \tag{2.12}$$

En utilisant l'échantillonneur de Gibbs à grille, nous pouvons tirer les échantillons de ρ et de γ par (2.12) (Ritter et Tanner, 1992). Nous utilisons la même densité *a posteriori* conditionnelle pour générer les deux paramètres à partir d'une méthode de grille; toutefois, les grilles de ρ et de γ diffèrent puisque leurs plages de valeurs possibles ne sont pas équivalentes. Cette méthode fait un bon mélange et converge rapidement. Soulignons que l'échantillonneur de Gibbs à grille converge vers une approximation de la distribution d'intérêt *a posteriori*, plutôt que vers sa forme exacte. Puis, poursuivant dans l'ordre inverse, nous pouvons entrer nos échantillons de ρ et de γ pour obtenir directement des échantillons de σ^2 à partir de (2.11), puis de θ à partir de (2.9), et finalement de $\boldsymbol{\mu}$ à partir de (2.6). L'obtention des échantillons de σ^2, θ et $\boldsymbol{\mu}$ est directe puisque leurs distributions sont connues. D'après nos échantillons de $\boldsymbol{\mu}$, σ^2 et les valeurs observées de $\mathbf{y}_i$, nous pouvons faire l'inférence de la population moyenne finie $\bar{Y}_i$ en utilisant le modèle

$$\bar{Y}_i \mid \mu_i,\sigma^2,\mathbf{y}_i \overset{\text{ind}}{\sim} \text{Normal}\left(f_i\bar{y}_i + (1-f_i)\,\mu_i, (1-f_i)\,\frac{\sigma^2}{N_i}\right). \tag{2.13}$$

Nous examinons le rendement de ce modèle dans la section 3.1 à l'aide d'une application reposant sur les données sur l'IMC.

2.1.2 Modèle d'autorégression conditionnelle simple

Nous décrivons maintenant le modèle d'autorégression conditionnelle simple (SCAR pour « simple conditional autoregressive ») et les différences entre ce modèle et le modèle CAR présenté précédemment. Sur le plan des calculs, les deux modèles se différencient principalement du fait que, dans le modèle CAR,

la matrice $\mathbf{R}$ est utilisée dans la variance sur $\boldsymbol{\mu}$ et que, dans le modèle SCAR, $\mathbf{R}$ est remplacé par la matrice d'identité, $\mathbb{I}$, ce qui simplifie le modèle. Par conséquent, le modèle SCAR hiérarchique bayésien est défini par (2.3) où $\mathbf{R} = \mathbb{I}$.

Mis à part les petits changements mentionnés sur le plan des calculs, les distributions et les méthodes que nous utilisons pour obtenir un échantillon de la densité *a posteriori* et la méthode que nous utilisons pour faire l'inférence à propos de la moyenne de la population finie, $\overline{Y}_i$ a la valeur décrite à la section 2.1.1. Nous substituons simplement $\mathbf{R}$ pour la matrice $\mathbb{I}$ et utilisons les valeurs actualisées de λ_1 et λ_ℓ en conséquence. Nous illustrons le rendement de ce modèle dans la section 3.1 à l'aide d'une application reposant sur les données sur l'IMC.

Dans le modèle SCAR, la matrice de précision est définie comme étant la matrice d'identité $\mathbb{I}$. Bien que les éléments diagonaux d'une matrice de précision sont tous égaux, les éléments diagonaux de l'inverse peuvent ne pas tous être égaux, ce qui permet l'hétéroscédasticité des effets aléatoires. Dans le modèle CAR, les entrées diagonales de la matrice de précision, $\mathbf{R}$, correspondent au nombre de voisins associés à chaque strate. Par conséquent, la matrice $\mathbf{R}$ pondère chaque ligne en fonction du nombre de voisins qu'elle comporte et agit comme matrice de normalisation. Tant le modèle SCAR que le modèle CAR présument que μ_i dépend seulement de la moyenne des strates voisines et non des strates éloignées (Chung and Datta, 2022). Soit $\mathbf{Q} = \mathbf{R} - \gamma\mathbf{W}$, alors μ_i et μ_j pour $i \neq j$ sont conditionnellement indépendantes, pour autant que les μ_k pour $k \neq i \neq j$, chaque fois que $Q_{ij} = 0$. Soulignons que, dans les modèles CAR et SCAR, il est important que les valeurs de ρ et de γ ne soient pas trop petites parce que nous voulons mettre en évidence la structure spatiale pour prendre en compte les covariables.

2.2 Inclusion des poids d'enquête

Dans la présente section, nous illustrons la façon d'inclure les poids d'enquête dans les deux modèles spatiaux que nous proposons. Nous utilisons, ici, les poids originaux, désignés par v_{ij} pour $j = 1,\ldots,n_i$ et $i = 1,\ldots \ell$, pour calculer la taille réelle de l'échantillon et les poids ajustés a_{ij}. D'abord, nous calculons la taille réelle de l'échantillon, $\hat{n}$:

$$\hat{n} = \frac{\left(\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij}\right)^2}{\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij}^2}. \tag{2.14}$$

La taille réelle de l'échantillon, $\hat{n}$, illustre l'importance de l'augmentation de la variance lorsque le poids est inégal (Nandram et Rao, 2021). Puis, nous calculons les poids ajustés, a_{ij}, qui sont utilisés pour éliminer le biais présent dans les poids originaux :

$$a_{ij} = \hat{n}\frac{v_{ij}}{\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij}}, \tag{2.15}$$

où $\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij} = \hat{N}$ est l'estimateur de la taille de la population Horvitz-Thompson, et $\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} a_{ij} = \hat{n}$, la taille réelle de l'échantillon.

Ces poids ajustés a_{ij} peuvent être utilisés dans un modèle lorsque les données ne comportent pas de valeurs aberrantes. Toutefois, dans l'exemple de l'IMC de la section 3, nos données comportent des valeurs aberrantes; alors nous utilisons la méthode d'estimation de Winsor, qui est une méthode efficace pour traiter les valeurs aberrantes en élaguant les poids d'enquête (Yang, Nandram et Choi, 2023). Les valeurs aberrantes sont définies ici comme les poids d'enquête observés qui sont supérieurs à $v_0 = Q_3 + 1{,}5(Q_3 - Q_1)$, où Q_1 est le premier quartile et Q_3 est le troisième quartile. Soit v^*, désignant les poids après élagage :

$$v_{ij}^* = \begin{cases} v_0, & v_{ij} \geq v_0 \\ rv_{ij}, & v_{ij} < v_0 \end{cases}, \tag{2.16}$$

où r est un paramètre de rééchelonnage de sorte que $\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij}^* = \sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij} = \hat{N}$. Alors, nous obtenons les poids ajustés et élagués a_{ij}^* :

$$\begin{aligned} \hat{n}^* &= \frac{\left(\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij}^*\right)^2}{\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij}^{*2}}, \\ a_{ij}^* &= \hat{n}^* \frac{v_{ij}^*}{\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} v_{ij}^*}. \end{aligned} \tag{2.17}$$

Ces poids ajustés et élagués a_{ij}^* sont utilisés dans les deux modèles, CAR et SCAR.

2.2.1 Inclusion des poids d'enquête dans le modèle d'autorégression conditionnelle

Le modèle CAR, dont les poids sont ajustés, peut être exprimé en remplaçant la variance de réponse dans la première ligne de (2.3), soit σ^2 par $\frac{\sigma^2}{a_{ij}^*}$, où a_{ij}^* provient de (2.17). Nous utilisons la même logique pour obtenir un échantillon de ce modèle comportant les poids ajustés, que celle que nous avons utilisée dans la section 2.1.1. La différence tient à la façon de faire les prédictions de la population; en incluant les poids d'enquête, nous avons maintenant besoin d'utiliser d'autres méthodes d'échantillonnage. Nous obtenons les prédictions de la population comme suit :

$$\overline{Y}_i \,|\, \boldsymbol{\mu}, \sigma^2 \;\sim\; \text{Normal}\left(\mu_i, \frac{\sigma^2}{\hat{N}_i}\right) \qquad i = 1, \ldots, \ell, \tag{2.18}$$

où $\hat{N}_i = \sum_{j=1}^{n_i} v_{ij}$ représente l'estimateur Horovitz-Thompson de la taille de la population de chaque strate $i = 1, \ldots, \ell$. Nous n'avons plus besoin d'inclure la moyenne de l'échantillon, $\overline{y}_i$, ou la fraction d'échantillonnage, f_i, dans cette prédiction de la population.

Auparavant, pour les prédictions de la population, nous combinions la partie échantillonnée et la partie non échantillonnée de la population pour obtenir un échantillon de $\overline{Y}_i$. Toutefois, puisque nous utilisons désormais les poids d'enquête provenant de l'échantillonnage probabiliste, nous pouvons utiliser d'autres méthodes d'échantillonnage et n'avons plus besoin d'inclure la partie échantillonnée. Les poids ajustés et élagués, $\mathbf{a}^*$, sont utilisés pour simuler un échantillon non biaisé de la population. Les poids ajustés et élagués étant inclus dans le modèle, nous devons échantillonner la population entière en utilisant d'autres

méthodes d'échantillonnage parce que les poids d'enquête, tant de l'échantillon que hors de l'échantillon, sont biaisés (Nandram, 2007; Nandram et Rao, 2021). Puis, lorsque nous faisons des prédictions de la population dans (2.18), nous utilisons des poids non ajustés pour la prédiction puisque ces poids représentent fidèlement la population entière. Nous examinons le rendement de ce modèle CAR incluant les poids d'enquête dans la section 3.1 à l'aide d'une application reposant sur les données sur l'IMC.

2.2.2 Inclusion des poids d'enquête dans le modèle d'autorégression conditionnelle simple

De même, le modèle SCAR dont les poids sont ajustés peut être exprimé en remplaçant la variance de réponse dans la première ligne de (2.3), soit σ^2 par $\frac{\sigma^2}{a_{ij}^*}$, où a_{ij}^* provient de (2.17) et en établissant $\mathbf{R} = \mathbb{I}$.

Nous utilisons la même logique pour obtenir un échantillon de ce modèle dont les poids sont ajustés que celle que nous avons utilisée dans la section 2.1.1. À l'instar du modèle CAR comportant les poids d'enquête, nous avons désormais besoin d'utiliser d'autres méthodes d'échantillonnage dans le modèle SCAR comportant les poids d'enquête pour faire des prédictions de la population. Nous utilisons (2.18) pour faire nos prédictions de la population dans le modèle SCAR comportant les poids d'enquête. Nous examinons le rendement de ce modèle SCAR comportant les poids d'enquête dans la section 3.1, à l'aide d'une application reposant sur des données sur l'IMC.

3. Application reposant sur des données sur l'indice de masse corporelle

Pour quiconque s'intéresse à la santé d'une population, les niveaux de l'indice de masse corporelle (IMC) des personnes peuvent être un important indicateur. Nous illustrons nos divers modèles non spatiaux et spatiaux en utilisant un échantillon probabiliste des données sur l'IMC de 1 867 personnes de huit comtés de la Californie enregistrés dans le National Health and Nutrition Examination Survey (NHANES III), (Nandram et Choi, 2005). Les poids d'enquête totalisent 12 232 099, ce qui signifie que notre échantillon correspond à 0,015 % de la population. En examinant les poids d'enquête originaux, nous constatons que la taille réelle de l'échantillon tirée de (2.14) est de 498, soit bien moindre que la taille de l'échantillon observé. Puisque les poids d'enquête sont asymétriques vers la droite et comportent quelques grandes valeurs aberrantes, nous appliquons la méthode d'estimation de Winsor pour réduire les poids. Cet ajustement donne une taille réelle de l'échantillon rééchelonné de 1 300, d'après (2.17), qui est plus près de la taille observée de l'échantillon. Les poids ajustés et élagués dans (2.17) sont distribués plus également, et ce sont ces poids qui seront utilisés dans cette application.

Ces données comportent quatre variables d'intérêt pour nous, soit l'âge, la race, le sexe et une mesure continue de l'IMC en $\mathrm{kg/m^2}$. Pour la race, les valeurs sont « blanc » ou « non blanc », et pour le sexe, les valeurs sont « homme » ou « femme ». La variable de l'âge est une variable continue comprise dans les données, et l'âge varie de 20 ans à 90 ans. Nous groupons par classe la variable de l'âge en groupes de deux

ans, soit les classes de 20 et 21 ans, 22 et 23 ans et ainsi de suite, pour obtenir un nombre fini de combinaisons possibles de covariables. Cette idée de grouper les variables par classe est une pratique courante qui permet de réduire le besoin de compter sur l'exactitude des données et de faire une inférence à propos d'un groupe d'âge élargi. Pour bien saisir toutes les combinaisons possibles de covariables présentes dans la population lorsque nous employons ce modèle, nous choisissons de grouper par classe toutes les covariables continues. Dans cet exemple de l'IMC, où les variables d'âge continues ont des bornes supérieures et inférieures inhérentes, ces bornes persisteront dans la population. Toutefois, si de telles bornes n'émergent pas naturellement dans les données, elles devraient être appliquées en utilisant les groupements par classe pour éviter que des groupes de la population ne soient pas représentés dans les données de l'échantillon. La valeur de la variable réponse continue de l'IMC s'étend de 15,8 kg/m^2 à 58,4 kg/m^2.

Après l'agrégation de toutes les combinaisons possibles d'âge, de race et de sexe, on obtient 144 strates ayant chacune son propre ensemble de valeurs de covariables. Toutefois, dans notre échantillon de données sur l'IMC, nous avons 12 strates sans observations et nous présumons que ce sont des zéros structuraux dans les données. Nous présumons donc que ces 12 groupes de personnes n'existent pas dans notre population. Par conséquent, le nombre total de strates, ℓ, dans ce cas, représente le nombre total de combinaisons possibles de covariables disponibles dans notre échantillon, et $\ell = 132$ après avoir retiré les 12 zéros structuraux des données. Si nous voulons éviter de faire cette hypothèse, nous pouvons atténuer les zéros structuraux en utilisant des groupes de variables plus grossiers. Si nous ne faisions pas de groupes plus grossiers, il serait nécessaire d'avoir la population totale connue, N_i, de chaque strate que nous incluons dans le modèle. Il nous faudrait alors nous fier aux poids d'enquête pour estimer N_i au moyen de l'estimateur de la taille de la population Horvitz-Thompson pour chaque strate $i = 1, \ldots, \ell$.

Si nous augmentons le nombre de covariables utilisées dans le modèle, la nature éparse de la matrice d'adjacence, $\mathbf{W}$, facilite la gestion de plus grands ensembles de covariables. Au besoin, des groupes plus grossiers des covariables peuvent également être employés pour diminuer le nombre de combinaisons de covariables et ainsi réduire les coûts de calcul. De plus, au lieu de miser uniquement sur l'optimisation du I de Moran pour la construction de la matrice adjacente, nous avons la possibilité de privilégier le recours à une plus grande parcimonie dans la matrice.

Dans la section 3.1, nous illustrons et comparons les résultats des deux modèles, spatiaux et non spatiaux, aux résultats comportant et excluant les poids d'enquête. Puis, dans la section 3.2 nous montrons dans quelle mesure l'inclusion de la composante spatiale dans nos modèles permet de réduire le nombre de regroupements globaux, par rapport aux modèles non spatiaux.

3.1 Comparaison des modèles d'application reposant sur l'indice de masse corporelle

Avant l'échantillonnage de l'un ou l'autre des modèles spatiaux, nous créons d'abord la matrice d'adjacence symétrique, $\mathbf{W}$ de dimension 132×132, en utilisant la distance de Mahalanobis décrite dans (2.1). Pour empêcher l'inclusion d'une covariable catégorique ordinale dans le calcul de la distance de

Mahalanobis, nous utilisons la valeur moyenne de chaque regroupement d'âges comme covariable d'âge. Rappelons que nous définissons $\mathbf{W}$ en établissant $w_{ii'} = 1$ si $d_{ii'} \leq d_0$ et $w_{ii'} = 0$ si $d_{ii'} > d_0$, les zéros étant sur la diagonale (soit, $w_{ii} = 0$), où d_0 est la valeur donnant la matrice $\mathbf{W}$ qui maximise le I de Moran, tiré de (2.2). Après avoir mené une recherche par quadrillage pour déterminer la valeur optimale de d_0, nous avons atteint la valeur maximale du I de Moran, soit $I = 0{,}212$, lorsque $d_0 = \text{moyenne}(d_{ij})/38 \approx 0{,}157$. En général, nous constatons que la diminution de d_0 augmente le I de Moran jusqu'à un certain point. Dans le présent cas, si nous continuons à réduire d_0 à moins de 0,157, nous n'obtenons pas d'augmentation du I de Moran. Toutefois, si nous augmentons d_0 à une valeur supérieure à 0,157, alors le I de Moran diminuera.

Comme nous avons désormais $\mathbf{W}$, nous pouvons maintenant tirer des échantillons des modèles CAR et SCAR décrits dans la section 2.1. L'échantillonnage de ces modèles est extrêmement similaire puisqu'ils commencent tous les deux par l'utilisation de l'échantillonneur de Gibbs à grille pour obtenir des échantillons de ρ et de γ simultanément (Ritter et Tanner, 1992). Pour le modèle CAR, l'intervalle de grille de γ est (-1,94; 1) et, pour le modèle SCAR, l'intervalle de grille de γ est (-0,390; 0,169). Les intervalles de grille de γ dans ces modèles diffèrent puisque l'intervalle de γ est fondée sur les valeurs propres de $\mathbf{R}^{-1}\mathbf{W}$, pour le modèle CAR, et les valeurs propres de $\mathbf{W}$ pour le modèle SCAR. L'intervalle de grille du modèle CAR est meilleur dans la mesure où il ramène γ plus proche de l'unité. Nous présentons également les résultats en incluant les poids dans les modèles CAR et SCAR.

Nous exécutons 10 000 itérations de l'échantillonneur, puis laissons tomber les 1 000 premières valeurs échantillonnées et choisissons la 9e valeur échantillonnée de chacune pour obtenir une taille d'échantillon définitive de 1 000 pour les deux paramètres. Dans les deux modèles, CAR et SCAR, l'échantillonneur de Gibbs à grille montre un bon rendement, comme en témoignent les représentations graphiques, les auto-corrélations, le test de stationnarité de Geweke et les tailles réelles d'échantillon. Pour ρ et γ, les valeurs de p du test de Geweke sont de 0,237 et de 0,286 respectivement dans le modèle CAR, et de 0,938 et de 0,833 respectivement dans le modèle SCAR, ce qui signifie que les deux paramètres respectent les exigences de stationnarité de chaque modèle. Comme l'indique le tableau 3.1, dans le modèle CAR excluant les poids d'enquête, les moyennes *a posteriori* de ρ et de γ sont de 0,188 et de 0,937, respectivement et, dans le modèle SCAR excluant les poids d'enquête, les moyennes *a posteriori* de ρ et de γ sont de 0,0444 et de 0,0165, respectivement. Puisqu'il est important que les valeurs de ρ et de γ ne soient pas trop petites étant donné que nous voulons insister sur la structure spatiale qui prend en compte les covariables, nous préférons le modèle CAR qui a de plus grandes valeurs de ρ et de γ. Dans le modèle CAR, γ est près de l'unité, ce qui indique bien que notre composante spatiale aura une incidence sur le modèle, par rapport à la valeur de γ beaucoup plus basse dans le modèle SCAR. Alors que ρ et γ diminuent, notre erreur-type *a posteriori* des prédictions de la population diminue également.

Dans le tableau 3.2, nous observons que $\overline{Y}$, dans le modèle SCAR, a une erreur-type *a posteriori* légèrement inférieure par rapport au modèle CAR et que ce résultat s'explique par les valeurs inférieures de ρ et de γ dans le modèle SCAR, présentées dans le tableau 3.1. Après avoir exécuté avec succès

l'échantillonneur de Gibbs à grille pour obtenir les valeurs de ρ et de γ, nous continuons à échantillonner les autres paramètres σ^2, θ et $\boldsymbol{\mu}$ directement de leurs densités *a posteriori* connues, tant pour le modèle CAR que pour le modèle SCAR. Chaque modèle comporte 136 paramètres totaux que nous pouvons ensuite utiliser pour prédire l'IMC de la population au moyen de (2.13). La méthode qui permet d'obtenir les échantillons des modèles CAR et SCAR comportant les poids est la même que celle décrite dans les modèles où les poids sont exclus, et l'échantillonneur de Gibbs offre un très bon rendement similaire (Ritter et Tanner, 1992). La façon de faire des prédictions de la population est différente dans les modèles CAR et SCAR lorsque nous incluons les poids d'enquête, comme illustré dans (2.18).

Tableau 3.1
Estimations *a posteriori* de ρ et de γ.

	MP	ETP	CV	95 % IHDP
Modèles excluant les poids d'enquête				
ρ (CAR)	0,188	0,045	0,244	(0,108; 0,289)
γ (CAR)	0,937	0,050	0,054	(0,848; 1,000)
ρ (SCAR)	0,044	0,013	0,292	(0,021; 0,069)
γ (SCAR)	0,165	0,004	0,026	(0,159; 0,169)
Modèles comportant les poids d'enquête				
ρ (CAR)	0,185	0,048	0,262	(0,098; 0,282)
γ (CAR)	0,940	0,052	0,055	(0,851; 1,000)
ρ (SCAR)	0,042	0,013	0,316	(0,018; 0,068)
γ (SCAR)	0,166	0,004	0,023	(0,159; 0,169)

Notes : MP = moyenne *a posteriori*; ETP = erreur-type *a posteriori*; CV = coefficient de variation; IHDP = intervalle à plus haute densité *a posteriori*; CAR = autorégression conditionnelle; SCAR = autorégression conditionnelle simple.

Tableau 3.2
Comparaison du modèle de prédiction de la population, selon l'indice de masse corporelle.

	Prédiction de $\bar{Y}$	ET de $\bar{Y}$	CV de $\bar{Y}$	95 % IHDP de $\bar{Y}$	DIC
Modèles excluant les poids d'enquête					
CAR	27,402	0,091	0,003	(27,233; 27,584)	-73,2
SCAR	27,418	0,088	0,003	(27,237; 27,579)	-70,5
Scott-Smith	27,375	0,129	0,005	(27,117; 27,634)	-61,0
BHF	27,347	0,132	0,005	(27,109; 27,608)	-61,4
Modèles comportant les poids d'enquête					
CAR	27,070	0,100	0,004	(26,879; 27,263)	-119,9
SCAR	27,090	0,100	0,004	(26,898; 27,268)	-113,0
Scott-Smith	27,380	0,147	0,005	(27,070; 27,656)	-55,5
BHF	27,294	0,161	0,006	(27,007; 27,614)	-60,4

Notes : ET = erreur-type; CV = coefficient de variation; IHDP = intervalle à plus haute densité *a posteriori*; DIC = critère d'information de déviance; CAR = autorégression conditionnelle; SCAR = autorégression conditionnelle simple; BHF = Battese, Harter et Fuller.

Un échantillon du modèle Scott-Smith, présenté dans l'annexe A, peut être obtenu directement sans passer par l'algorithme de la méthode MCMC. Après avoir échantillonné ρ à l'aide de la méthode de la grille, nous pouvons ensuite tirer des échantillons de σ^2, θ et $\boldsymbol{\mu}$ dans l'ordre, à partir de leurs densités

a posteriori conditionnelles connues. Puisqu'il y a 132 strates, notre modèle contient $\mu_1,\ldots,\mu_{132}$, qui produisent les 135 paramètres totaux échantillonnés dans ce modèle. Une taille d'échantillon de 1 000 a été utilisée pour prédire chaque paramètre du modèle Scott-Smith. Après l'obtention de ces paramètres, nous pouvons faire des prédictions de l'IMC de la population en utilisant (A.9).

À l'instar du modèle Scott-Smith, un échantillon du modèle BHF, présenté dans l'annexe B, peut également être obtenu sans passer par l'échantillonneur de la méthode MCMC. Pour ce modèle, nous échantillonnons ρ à l'aide de la méthode de la grille, nous pouvons ensuite tirer des échantillons de σ^2, $\boldsymbol{\beta}$ et $\boldsymbol{\nu}$ dans l'ordre, à partir de leurs densités *a posteriori* conditionnelles connues. Ce modèle est le seul qui contient les covariables directement; nous avons donc $\beta_0,\ldots,\beta_3$ où β_0 représente le coefficient de l'ordonnée à l'origine. Par conséquent, il y a 138 paramètres totaux qui, une fois échantillonnés, nous permettent de prédire l'IMC de la population en utilisant (B.12). Une taille d'échantillon de 1 000 a été utilisée pour prédire chaque paramètre du modèle BHF. Les modèles Scott-Smith et BHF comportant les poids sont adaptés en échantillonnant selon la même densité *a posteriori*, comme dans le cas des modèles excluant les poids. Seules les prédictions de la population sont différentes. Les prédictions de la population peuvent être faites en utilisant (A.11) et (B.14) pour les modèles Scott-Smith et BHF comportant les poids, respectivement. Les modèles Scott-Smith et BHF sont utilisés comme référence pour comparer le rendement du modèle spatial.

Le tableau 3.2 présente les résultats de la prédiction de l'IMC de la population pour les quatre modèles excluant et incluant les poids. Dans notre application, la variable de réponse est l'IMC; par conséquent, $\overline{Y}$ représente la moyenne générale de l'IMC de la population de huit comtés en Californie. Les résultats des deux modèles non spatiaux, soit les modèles Scott-Smith et BHF, sont très similaires pour les quatre mesures de la moyenne *a posteriori* (MP), de l'erreur-type *a posteriori* (ETP), du coefficient de variation (CV) et de l'intervalle à plus haute densité *a posteriori* (IHDP). Les deux modèles spatiaux, CAR et SCAR, ont un rendement similaire. Les modèles CAR et SCAR ont produit une prédiction légèrement plus élevée de l'IMC moyen de la population finie *a posteriori* de la population. Le rendement des modèles CAR et SCAR dépasse celui des modèles non spatiaux quant au critère d'information de déviance (DIC pour « Deviance Information Criterion ») : le modèle CAR donnant la plus faible valeur de DIC, suivi de près par le modèle SCAR. Les modèles spatiaux ont également une erreur-type *a posteriori* plus petite et un CV qui donne un IHDP plus serré par rapport aux modèles non spatiaux. Puisque les strates sont renforcées par les strates voisines dans les modèles spatiaux, nous constatons que les erreurs-types *a posteriori* diminuent alors que les moyennes *a posteriori* sont plus adaptées à chaque voisinage. Les strates dont la taille de l'échantillon est très petite ne reposent plus seulement sur leur nombre limité d'observations dans les modèles spatiaux, puisqu'elles sont désormais incluses dans les voisinages qui comptent, collectivement, un plus grand nombre d'observations. Dans les modèles sans composante spatiale, les prédictions des strates mènent plus généralement à des prédictions plus vagues centrées autour de la moyenne d'échantillon globale ayant une erreur-type *a posteriori* plus élevée.

Le tableau 3.2 présente également les résultats de la prédiction de l'IMC de la population lorsque les poids d'enquête ajustés et élagués sont inclus dans les modèles. Dans le tableau, nous pouvons observer que la prédiction de la population générale quant à l'IMC de la moyenne de la population finie est similaire à celle des modèles excluant les poids. Toutefois, l'erreur-type *a posteriori* et le CV augmentent dans les modèles incluant les poids par rapport aux modèles excluant les poids. Par contre, l'erreur-type plus grande entraîne aussi des IHDP plus larges dans les modèles incluant les poids. En incluant les poids ajustés et élagués dans le modèle, nous nous attendons à une augmentation de l'erreur-type *a posteriori* puisque l'inclusion des poids fait diminuer la taille de l'échantillon à la taille réelle de l'échantillon. Naturellement, si la taille de l'échantillon est plus petite, l'erreur-type *a posteriori* sera plus grande. Le DIC des deux modèles spatiaux est significativement plus petit lorsque les poids ajustés et élagués sont inclus, ce qui indique une préférence pour les modèles spatiaux incluant les poids.

Le tableau 3.3 présente les résultats de la prédiction de l'IMC pour les quatre modèles, en fonction des huit comtés inclus dans nos données d'enquête de l'IMC. Les modèles n'ont pas été adaptés à chaque comté; les résultats ont simplement été séparés par comté. Les résultats obtenus pour chaque comté étaient similaires aux résultats globaux décrits à partir du tableau 3.2; toutefois, puisque la taille des échantillons est réduite lorsque nous faisons un groupement par comté, les erreurs-types *a posteriori* augmenteront en raison de la plus petite taille de l'échantillon. La taille de l'échantillon est environ la même pour tous les comtés, à l'exception du comté 3 dont l'échantillon est exceptionnellement grand et compte 795 observations, et correspond à environ 43 % de la taille de l'échantillon des données d'IMC. Pour les autres comtés, les échantillons comptaient de 125 à 176 observations. La taille de l'échantillon comparativement plus grande du comté 3 produit des erreurs-types *a posteriori* plus petites par rapport aux autres comtés.

Tableau 3.3
Comparaison du modèle de prédiction de la population, selon l'indice de masse corporelle, par comté.

	Prédiction de $\bar{Y}$	ET de $\bar{Y}$	CV de $\bar{Y}$	95 % IHDP de $\bar{Y}$
Comté 1				
CAR	27,242	0,101	0,004	(27,061; 27,457)
SCAR	27,232	0,097	0,004	(27,044; 27,414)
Scott-Smith	27,194	0,149	0,005	(26,913; 27,481)
BHF	27,198	0,151	0,006	(26,891; 27,480)
Comté 2				
CAR	27,549	0,112	0,004	(27,308; 27,759)
SCAR	27,554	0,112	0,004	(27,336; 27,768)
Scott-Smith	27,491	0,178	0,006	(27,154; 27,852)
BHF	27,487	0,174	0,006	(27,138; 27,801)

Notes : ET = erreur-type; CV = coefficient de variation; IHDP = intervalle à plus haute densité *a posteriori*; CAR = autorégression conditionnelle; SCAR = autorégression conditionnelle simple; BHF = Battese, Harter et Fuller.

Tableau 3.3(suite)
Comparaison du modèle de prédiction de la population, selon l'indice de masse corporelle, par comté.

	Prédiction de $\bar{Y}$	ET de $\bar{Y}$	CV de $\bar{Y}$	95 % IHDP de $\bar{Y}$
Comté 3				
CAR	27,460	0,094	0,003	(27,268; 27,635)
SCAR	27,470	0,093	0,003	(27,292; 27,656)
Scott-Smith	27,424	0,140	0,005	(27,142; 27,701)
BHF	27,389	0,140	0,005	(27,100; 27,646)
Comté 4				
CAR	27,446	0,139	0,005	(27,178; 27,700)
SCAR	27,467	0,144	0,005	(27,164; 27,729)
Scott-Smith	27,462	0,212	0,008	(27,029; 27,842)
BHF	27,428	0,219	0,008	(27,008; 27,829)
Comté 5				
CAR	27,557	0,123	0,004	(27,310; 27,794)
SCAR	27,563	0,124	0,004	(27,329; 27,801)
Scott-Smith	27,551	0,188	0,007	(27,188; 27,939)
BHF	27,497	0,194	0,007	(27,125; 27,899)
Comté 6				
CAR	27,481	0,129	0,005	(27,236; 27,725)
SCAR	27,502	0,131	0,005	(27,260; 27,781)
Scott-Smith	27,408	0,205	0,007	(26,984; 27,766)
BHF	27,366	0,206	0,008	(26,967; 27,753)
Comté 7				
CAR	27,083	0,114	0,004	(26,856; 27,287)
SCAR	27,109	0,110	0,004	(26,892; 27,322)
Scott-Smith	27,113	0,163	0,006	(26,827; 27,450)
BHF	27,087	0,166	0,006	(26,769; 27,405)
Comté 8				
CAR	27,218	0,120	0,004	(26,983; 27,445)
SCAR	27,273	0,113	0,004	(27,050; 27,490)
Scott-Smith	27,205	0,163	0,006	(26,895; 27,503)
BHF	27,184	0,170	0,006	(26,885; 27,581)

Notes : ET = erreur-type; CV = coefficient de variation; IHDP = intervalle à plus haute densité *a posteriori*; CAR = autorégression conditionnelle; SCAR = autorégression conditionnelle simple; BHF = Battese, Harter et Fuller.

Le tableau 3.4 présente les résultats de la prédiction de l'IMC pour les quatre modèles comportant des poids, en fonction des huit comtés inclus dans nos données d'enquête sur l'IMC. De nouveau, les modèles n'ont pas été adaptés à chaque comté; les résultats ont simplement été séparés en fonction des comtés. Les résultats de chaque comté sont similaires aux résultats globaux décrits à partir du tableau 3.2. Dans le tableau 3.4, la taille des échantillons est réduite par l'inclusion des poids d'enquête, en plus de la répartition des tailles d'échantillons par comté. Les erreurs-types *a posteriori* continuent d'augmenter parce que les deux facteurs réduisent la taille des échantillons. Il faut rappeler que tous les comtés ont environ la même taille d'échantillon, à l'exception du comté 3 dont la taille de l'échantillon est exceptionnellement grande. La taille de l'échantillon plus grande du comté 3 produit des erreurs-types *a posteriori* plus petites par rapport aux autres comtés.

Tableau 3.4
Inclusion des poids d'enquête dans la prédiction de l'indice de masse corporelle de la population, par comté.

	Prédiction de $\bar{Y}$	ET de $\bar{Y}$	CV de $\bar{Y}$	95 % IHDP de $\bar{Y}$
Comté 1				
CAR	27,158	0,114	0,004	(26,949; 27,411)
SCAR	27,185	0,111	0,004	(26,970; 27,401)
Scott-Smith	27,347	0,169	0,006	(27,011; 27,666)
BHF	27,202	0,175	0,006	(26,888; 27,575)
Comté 2				
CAR	26,945	0,125	0,005	(26,709; 27,202)
SCAR	27,021	0,117	0,004	(26,806; 27,259)
Scott-Smith	27,275	0,164	0,006	(26,950; 27,575)
BHF	27,416	0,200	0,007	(27,048; 27,814)
Comté 3				
CAR	27,196	0,102	0,004	(26,999; 27,383)
SCAR	27,199	0,104	0,004	(27,014; 27,417)
Scott-Smith	27,362	0,145	0,005	(27,070; 27,631)
BHF	27,329	0,171	0,006	(26,987; 27,634)
Comté 4				
CAR	26,983	0,122	0,005	(26,749; 27,229)
SCAR	26,990	0,123	0,005	(26,734; 27,206)
Scott-Smith	27,268	0,165	0,006	(26,979; 27,607)
BHF	27,352	0,253	0,009	(26,852; 27,816)
Comté 5				
CAR	27,162	0,118	0,004	(26,932; 27,391)
SCAR	27,171	0,117	0,004	(26,943; 27,400)
Scott-Smith	27,284	0,174	0,006	(26,953; 27,633)
BHF	27,393	0,224	0,008	(26,967; 27,859)
Comté 6				
CAR	27,142	0,115	0,004	(26,924; 27,379)
SCAR	27,135	0,118	0,004	(26,887; 27,344)
Scott-Smith	27,320	0,179	0,007	(26,972; 27,659)
BHF	27,298	0,233	0,009	(26,871; 27,761)
Comté 7				
CAR	26,935	0,121	0,004	(26,706; 27,180)
SCAR	26,923	0,120	0,004	(26,694; 27,158)
Scott-Smith	27,107	0,174	0,006	(26,794; 27,477)
BHF	27,085	0,194	0,007	(26,725; 27,459)
Comté 8				
CAR	26,840	0,127	0,005	(26,590; 27,090)
SCAR	26,903	0,126	0,005	(26,669; 27,150)
Scott-Smith	27,054	0,171	0,006	(26,742; 27,403)
BHF	27,156	0,198	0,007	(26,774; 27,585)

Notes : ET = erreur-type; CV = coefficient de variation; IHDP = intervalle à plus haute densité *a posteriori*; CAR = autorégression conditionnelle; SCAR = autorégression conditionnelle simple; BHF = Battese, Harter et Fuller.

3.2 Réduction du regroupement global par modélisation spatiale

Notre principal objectif est de faire une inférence à propos de la moyenne de la population finie sans inclure directement les covariables dans nos modèles, ce que nous avons présenté. Nous avons choisi

d'introduire une composante spatiale dans nos modèles pour avoir des moyennes *a posteriori* du résultat des strates individuelles dans un regroupement moins global. Nous ne voulons pas la moyenne *a posteriori* de chaque strate individuelle pour simplement approcher la moyenne *a posteriori* de la population globale. Nous souhaiterions plutôt avoir des strates ayant des attributs de covariables similaires (par exemple voisins dans la matrice $\mathbf{W}$) se renforçant mutuellement pour produire une moyenne *a posteriori* de strate orientée vers la moyenne *a posteriori* du voisinage. Par conséquent, nous étudions la valeur $\boldsymbol{\mu}$ du modèle Scott-Smith et du modèle CAR pour illustrer qu'en incluant la composante spatiale, comme nous l'avons vu dans le modèle CAR, nous pouvons augmenter la variabilité des prédictions *a posteriori* de $\boldsymbol{\mu}$. Puisque les résultats du modèle Scott-Smith et du modèle BHF sont similaires, nous utilisons seulement le modèle BHF pour faire la comparaison. En général, nous souhaitons éviter d'utiliser les covariables dans nos modèles (comme on l'a vu dans le modèle BHF), si possible, et les résultats similaires des modèles Scott-Smith et BHF prouvent que l'inclusion des covariables dans le modèle n'améliore pas les résultats de prédiction. De même, puisque les modèles CAR et SCAR donnent des résultats similaires et que le modèle SCAR est une version simplifiée du modèle CAR, nous ferons la comparaison en utilisant le modèle CAR.

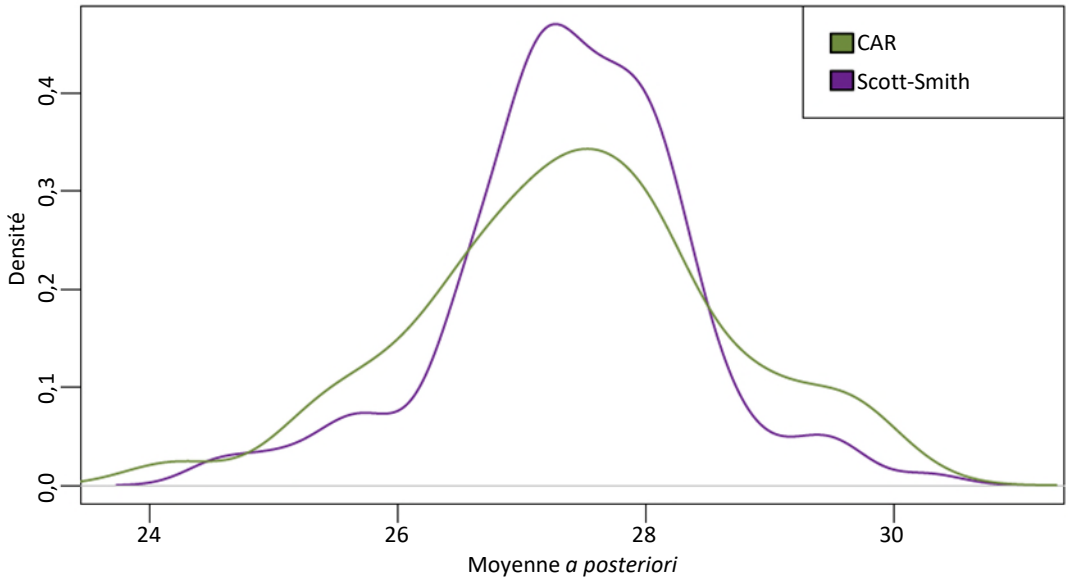
Nous analysons la valeur de $\boldsymbol{\mu}$ de chaque modèle; par conséquent, nous avons un échantillon de la densité *a posteriori* de μ_i pour chaque $i = 1,\ldots,\ell$. Trois principales indications illustrent que l'inclusion de la composante spatiale dans le modèle CAR réduit le regroupement global, lorsque nous faisons une comparaison au modèle Scott-Smith. Premièrement, les estimations de $\mu_1,\ldots,\mu_{132}$ obtenues à partir du modèle CAR ont un écart-type de 1,237, par rapport aux estimations du modèle Scott-Smith qui ont un écart-type de 0,967. Nous pouvons donc déjà constater une augmentation de la variation dans les estimations de $\boldsymbol{\mu}$ du modèle CAR, et cette augmentation de la variation est un signe de moindre regroupement global dans le modèle CAR puisque $\boldsymbol{\mu}$ comporte des valeurs plus dispersées. Deuxièmement, en observant la figure 3.1, qui présente les deux courbes de densité du noyau de $\boldsymbol{\mu}$ de chaque modèle, nous pouvons constater que $\boldsymbol{\mu}$ du modèle CAR a un pic moins élevé et des queues plus importantes. Les valeurs de $\boldsymbol{\mu}$ dans le modèle CAR ne convergent pas vers la moyenne de population globale aussi fortement que dans le modèle Scott-Smith.

Troisièmement, nous observons les paramètres de rétrécissement dans la moyenne *a posteriori* de $\boldsymbol{\mu}$ de chaque modèle. Il faut rappeler que la moyenne *a posteriori* de μ_i selon le modèle Scott-Smith dans (A.3) est : $\left(\lambda_i \bar{y}_i + (1-\lambda_i)\,\theta\right)$ où $\lambda_i = n_i\rho \big/ \left(\left(n_i - 1\right)\rho + 1\right)$ pour $i = 1\ldots,\ell$. Nous pouvons récrire la moyenne *a posteriori* de $\boldsymbol{\mu}$ du modèle CAR dans (2.6) comme suit : $\left(\Lambda\bar{\mathbf{y}} + (\mathbb{I} - \Lambda)\,\theta\mathbf{1}\right)$ où

$$\Lambda = \left(\operatorname{diag}\left(\frac{\sigma^2}{n_1},\ldots,\frac{\sigma^2}{n_\ell}\right)^{-1} + \left(\frac{\rho}{1-\rho}\sigma^2\left(\mathbf{R} - \gamma\mathbf{W}\right)^{-1}\right)^{-1}\right)^{-1} \operatorname{diag}\left(\frac{\sigma^2}{n_1},\ldots,\frac{\sigma^2}{n_\ell}\right)^{-1}.$$

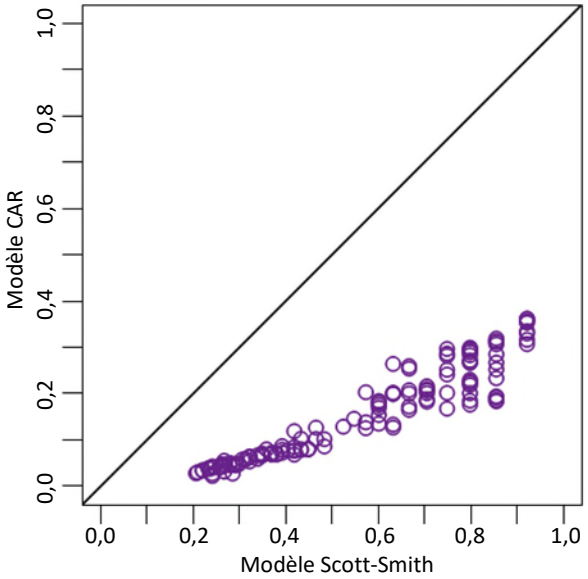
Lorsque les paramètres de rétrécissement du modèle Scott-Smith, $(1-\lambda_i)$, sont supérieurs à la somme des valeurs de la ligne des paramètres de rétrécissement du modèle CAR, $(\mathbb{I} - \Lambda)$, le modèle Scott-Smith non spatial tend plutôt vers le paramètre de regroupement global, θ, au lieu de maintenir les caractéristiques des strates individuelles. C'est exactement ce que montre la figure 3.2.

Figure 3.1 Comparaison de la distribution *a posteriori* de μ.



Note : CAR = autorégression conditionnelle.

Figure 3.2 Comparaison des paramètres de rétrécissement.



Note : CAR = autorégression conditionnelle.

La figure 3.2 montre que le modèle Scott-Smith attribue beaucoup plus de poids au paramètre de regroupement global, θ, par rapport au modèle CAR. Ce constat maintient notre objectif d'établir que le modèle CAR produit un regroupement global moindre en incluant les relations de voisinage. Il est aussi

important de préciser que toutes les valeurs de paramètre de rétrécissement du modèle Scott-Smith et du modèle CAR sont dans l'intervalle de $[0, 1]$.

4. Conclusion

En introduisant la composante spatiale, notre principal objectif est de prendre en compte les covariables sans utiliser un modèle de régression. Ce faisant, nous réduisons aussi l'ampleur du regroupement global et permettons plutôt des prédictions plus rapprochées pour les voisins ayant des attributs similaires. Nous avons démontré, dans notre comparaison de $\boldsymbol{\mu}$ pour les modèles CAR et Scott-Smith, que l'inclusion de cette relation spatiale des strates n'a pas pour effet de limiter le regroupement global. Nous y sommes arrivés en observant les paramètres de rétrécissement, les densités *a posteriori* et la variation *a posteriori* de $\boldsymbol{\mu}$ dans les deux modèles. Les modèles CAR et SCAR fonctionnent bien, tous les deux, comme modèles d'estimation sur petits domaines qui réduiront les effets du regroupement global sans définir la relation entre la réponse et les covariables. Toutefois, puisqu'il est important que ρ et γ ne soient pas trop petits pour pouvoir mettre l'accent sur la structure spatiale qui prend les covariables, nous préférons le modèle CAR qui a de plus grandes valeurs de ρ et de γ.

Dans le modèle CAR, γ est près de l'unité, ce qui indique bien que notre composante spatiale aura une incidence sur le modèle, par rapport à la valeur γ beaucoup plus basse dans le modèle SCAR. Alors que ρ et γ diminuent, notre erreur-type *a posteriori* des prédictions de la population diminue également. Nous avons également présenté la façon d'utiliser ces modèles spatiaux que nous prônons, avec et sans poids, et la façon de faire des prédictions de la population dans les deux cas. Enfin, nous ne cherchons pas à définir une relation entre la variable réponse $\mathbf{y}$ et $\mathbf{X}$ en ayant $\boldsymbol{\beta}$ dans le modèle, comme dans le cas du modèle BHF. Nous évitons d'émettre de fortes hypothèses à propos de cette relation et nous maximisons le nombre d'applications possibles que nos modèles peuvent accepter.

La recherche pourrait se poursuivre pour inclure ce genre de problème en adaptant les modèles pour couvrir la situation d'une variable réponse binaire au lieu d'une variable réponse continue. Bien que le cas de la variable réponse binaire exige plus de calculs, les applications utiles sont nombreuses. Par exemple, pour l'application à l'IMC, nous pourrions être davantage intéressés par la proportion de personnes au sein de la population qui sont obèses (soit, IMC ≥ 30), au lieu de prédire l'IMC global de la population finie. Il y a des cas où la proportion d'une caractéristique que possède une population donne plus de renseignements que la connaissance de la valeur moyenne de cette caractéristique dans la population.

Annexe A

Modèle Scott-Smith

Dans l'annexe A, nous présentons les précisions techniques du modèle Scott-Smith (Scott et Smith, 1969). Le modèle Scott-Smith adapté que nous utilisons peut être formulé comme suit :

$$\begin{aligned} y_{ij} \mid \mu_i, \sigma^2 &\sim \text{Normal}\left(\mu_i, \sigma^2\right), \\ \mu_i \mid \theta, \rho, \sigma^2 &\sim \text{Normal}\left(\theta, \frac{\rho}{1-\rho}\sigma^2\right), \pi\left(\theta, \sigma^2, \rho\right) \propto \frac{1}{\sigma^2}, -\infty < \theta < \infty, \quad 0 < \rho < 1, \quad \sigma^2 > 0, \\ & j = 1, \ldots, n_i, \quad i = 1, \ldots, \ell. \end{aligned} \tag{A.1}$$

Bien que les covariables soient absentes du modèle, les réponses restent regroupées en utilisant les valeurs des covariables, de sorte que chaque $\mathbf{y}_i$ a la même combinaison unique de covariables pour chaque strate $i = 1, \ldots \ell.$ Nandram, Toto et Choi (2011) ont démontré que ρ est une corrélation intra-classe courante. Ici μ_i suit une distribution normale :

$$\mu_i \mid \theta, \sigma^2, \rho, \mathbf{y} \sim \text{Normal}\left(\lambda_i \bar{y}_i + (1-\lambda_i)\,\theta, (1-\lambda_i)\,\rho\sigma^2 / (1-\rho)\right) \tag{A.2}$$

où $\lambda_i = n_i\rho / \left(\left(n_i - 1\right)\rho + 1\right)$ pour $i = 1 \ldots, \ell.$

La densité conditionnelle *a posteriori* de θ est :

$$\theta \mid \sigma^2, \rho, \mathbf{y} \sim \text{Normal}\left(\tilde{y}, \frac{\sigma^2\rho}{(1-\rho)\sum_{i=1}^{\ell}\lambda_i}\right). \tag{A.3}$$

Notons que

$$\tilde{y} = \left(\sum_{i=1}^{\ell}\left(n_i / \left(\left(n_i - 1\right)\rho + 1\right)\right)\bar{y}_i\right) \Big/ \left(\sum_{i=1}^{\ell}\left(n_i / \left(\left(n_i - 1\right)\rho + 1\right)\right)\right)$$

est bien défini pour tous les $0 \le \rho \le 1$ et $l \ge 2.$

La densité conditionnelle *a posteriori* de σ^2 est :

$$\sigma^2 \mid \rho, \mathbf{y} \sim \text{InvGamma}\left(\frac{n-1}{2}, \left\{\sum_{i=1}^{\ell}(n_i - 1)\, s_i^2 + \frac{1-\rho}{\rho}\left(\sum_{i=1}^{\ell}\lambda_i\left(\bar{y}_i - \tilde{y}\right)^2\right)\right\} \Big/ 2\right). \tag{A.4}$$

Finalement, après avoir éliminé par intégration $\sigma^2,$ il nous reste la densité *a posteriori* non normalisée :

$$\begin{aligned} \pi_4\left(\rho \mid \mathbf{y}\right) \propto\ & (1-\rho)^{(l-2)/2}\sqrt{\frac{\prod_{i=1}^{\ell} n_i / \left(\left(n_i - 1\right)\rho + 1\right)}{\sum_{i=1}^{\ell} n_i / \left(\left(n_i - 1\right)\rho + 1\right)}} \\ & \times \frac{1}{\left\{1 + (1-\rho)\left(\sum_{i=1}^{\ell}\left(n_i / \left(\left(n_i - 1\right)\rho + 1\right)\right)\left(\bar{y}_i - \tilde{y}\right)^2\right) \Big/ \left(\sum_{i=1}^{\ell}(n_i - 1)\, s_i^2\right)\right\}^{(n-1)/2}}. \end{aligned} \tag{A.5}$$

Cela prouve que la densité *a posteriori* conjointe est adéquate et que nous pouvons obtenir un échantillon de la densité *a posteriori* conjointe en échantillonnant à partir de $\pi_4\left(\rho \mid \mathbf{y}\right)$ d'abord, et en continuant à tirer des échantillons de leurs distributions connues en ordre inverse (Nandram, Toto and Choi, 2011). Par conséquent, nous commençons en tirant des échantillons de ρ à partir de (A.5) en utilisant la méthode de la grille. Ensuite, nous utilisons l'échantillon de ρ que nous avons obtenu pour tirer un échantillon de σ^2 directement à partir de (A.4). Puis, nous utilisons l'échantillon de ρ et de σ^2 pour tirer un échantillon de θ de sa distribution normale (A.3). Finalement, nous utilisons les échantillons de ρ, σ^2 et θ pour tirer un

échantillon de μ_i pour $i=1,\ldots,\ell$ à partir de (A.2). D'après nos échantillons de la densité *a posteriori* et les valeurs observées de $\mathbf{y}_i$, nous pouvons faire l'inférence de la population moyenne finie $\overline{Y}_i$ en utilisant le modèle suivant :

$$\overline{Y}_i \,|\, \mathbf{y}_i \overset{\text{ind}}{\sim} \text{Normal}\left(f_i\, \overline{y}_i + (1-f_i)\, \mu_i, (1-f_i)\, \frac{\sigma^2}{N_i} \right). \tag{A.6}$$

Les résultats de ce modèle sont présentés à la section 3.1, à l'aide d'une application aux données de l'IMC.

Inclusion des poids dans le modèle Scott-Smith

Nous pouvons également inclure les poids dans le modèle Scott-Smith et nous utilisons les mêmes poids ajustés et élagués, décrits dans la section 2.2. Le modèle Scott-Smith comportant les poids peut être exprimé en remplaçant la variance de réponse dans la première ligne de (A.1) soit σ^2 par $\frac{\sigma^2}{a_{ij}^*}$, où a_{ij}^* est tiré de (2.17).

Nous utilisons la même logique pour obtenir un échantillon de ce modèle comportant les poids ajustés que celle que nous avons utilisée précédemment dans le modèle sans poids. L'inférence des prédictions de la population diffère parce que nous obtenons les prédictions de la population comme suit :

$$\overline{Y}_i \,|\, \boldsymbol{\mu}, \sigma^2 \;\sim\; \text{Normal}\left(\mu_i, \frac{\sigma^2}{\hat{N}_i} \right) \qquad i=1,\ldots,\ell, \tag{A.7}$$

où $\hat{N}_i = \sum_{j=1}^{n_i} v_{ij}$ représente l'estimateur de taille de population Horovitz-Thompson de chaque strate $i=1,\ldots,\ell$.

Annexe B

Modèle Battese, Harter et Fuller

Dans l'annexe B, nous présentons les précisions techniques du modèle Battese, Harter et Fuller (BHF) (Battese, Harter et Fuller, 1988). Ce modèle non spatial introduit les covariables et comprend les effets aléatoires pour chaque strate. Le modèle BHF est le seul modèle utilisé dans le présent article qui précise la relation entre la réponse et les covariables. En général, nous voulons éviter de définir cette relation entre $\mathbf{y}_i$ et $\mathbf{x}_i$. Le modèle BHF est :

$$\begin{aligned}
y_{ij} \,|\, \mathbf{v}, \boldsymbol{\beta}, \sigma^2 &\overset{\text{ind}}{\sim} \text{Normal}\left(\mathbf{x}_i'\boldsymbol{\beta} + v_i, \sigma^2 \right), \\
\mathbf{v} \,|\, \rho, \sigma^2 &\overset{\text{ind}}{\sim} \text{Normal}\left(0, \frac{\rho}{1-\rho}\sigma^2 \right), \\
\pi\left(\boldsymbol{\beta}, \sigma^2, \rho\right) &\propto \frac{1}{\sigma^2}; \quad \sigma^2 > 0, \quad 0<\rho<1, \quad \boldsymbol{\beta} \in \mathbb{R}^p, \quad j=1,\ldots,n_i, \quad i=1,\ldots\ell.
\end{aligned} \tag{B.1}$$

Établissons $\lambda_i = \rho n_i / ((1-\rho)+\rho n_i)$, alors v_i suit une distribution normale :

$$v_i \mid \boldsymbol{\beta}, \sigma^2, \rho, \mathbf{y} \overset{\text{ind}}{\sim} \text{Normal}\left(\lambda_i \left(\overline{y}_i - \mathbf{x}_i' \boldsymbol{\beta} \right), \frac{(1-\lambda_i)\, \rho \sigma^2}{(1-\rho)} \right). \tag{B.2}$$

La densité conditionnelle *a posteriori* de $\boldsymbol{\beta}$ est :

$$\boldsymbol{\beta} \mid \sigma^2, \rho, \mathbf{y} \sim \text{Normal}\left(\hat{\boldsymbol{\beta}}, \sigma^2 \hat{\boldsymbol{\Sigma}} \right). \tag{B.3}$$

où

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}} \sum_{i=1}^{\ell} n_i \left(1-\lambda_i\right) \overline{y}_i \mathbf{x}_i' \quad \text{et} \quad \hat{\boldsymbol{\Sigma}} = \left(\sum_{i=1}^{\ell} n_i \left(1-\lambda_i\right) \mathbf{x}_i \mathbf{x}_i' \right)^{-1}. \tag{B.4}$$

Notons que $\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{\Sigma}}^{-1}$ sont bien définies pour toutes les ρ, pour autant que la matrice de plan $\mathbf{X} = \left(\mathbf{x}_i' \right)$ est de plein rang, où $\mathbf{x}_i'$ correspond aux lignes de $\mathbf{X}$.

La densité conditionnelle *a posteriori* de σ^2 est :

$$\sigma^2 \mid \rho, \mathbf{y} \sim \text{InvGamma}\left(\frac{n-p}{2}, \frac{\sum_{i=1}^{\ell} \left[n_i \left(1-\lambda_i\right) \left(\overline{y}_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}} \right)^2 + \sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_i \right)^2 \right]}{2} \right) \tag{B.5}$$

et $n = \sum_{i=1}^{\ell} n_i$. Par conséquent, après élimination par intégration de σ^2, il nous reste finalement que la densité *a posteriori* non normalisée de ρ :

$$\begin{aligned} \pi\left(\rho \mid \mathbf{y}\right) \propto \det & \left[\left(\sum_{i=1}^{\ell} n_i \left(1-\lambda_i\right) \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \right]^{1/2} \prod_{i=1}^{\ell} \left(1-\lambda_i\right)^{1/2} \\ & \times \left[\sum_{i=1}^{\ell} n_i \left(1-\lambda_i\right) \left(\overline{y}_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}} \right)^2 + \sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_i \right)^2 \right]^{\frac{-n+p}{2}}. \end{aligned} \tag{B.6}$$

Nous pouvons désormais obtenir directement un échantillon de la densité *a posteriori* conjointe en commençant par l'utilisation de la méthode de la grille pour échantillonner ρ. Par la suite, l'échantillonnage de σ^2, $\boldsymbol{\beta}$ et $\mathbf{v}$ est direct puisque chacun de ces paramètres a une forme normalisée. D'après nos échantillons de la densité *a posteriori* et les valeurs observées de $\mathbf{y}_i$, nous pouvons faire l'inférence de la population moyenne finie $\overline{Y}_i$ en utilisant le modèle suivant :

$$\overline{Y}_i \mid \mathbf{y}_i \overset{\text{ind}}{\sim} \text{Normal}\left(f_i \overline{y}_i + (1-f_i) \left[\mathbf{x}_i' \boldsymbol{\beta} + v_i \right], (1-f_i) \frac{\sigma^2}{N_i} \right). \tag{B.7}$$

Nous étudions le rendement de ce modèle dans la section 3.1, à l'aide d'une application reposant sur les données de l'IMC.

Inclusion des poids d'enquête dans le modèle Battese, Harter et Fuller

Nous pouvons également inclure les poids dans le modèle BHF en utilisant les mêmes poids ajustés et élagués, décrits dans la section 2.2. Le modèle BHF comportant les poids ajustés peut être exprimé en remplaçant la variance de réponse dans la première ligne de (B.1), soit σ^2 par $\frac{\sigma^2}{a_{ij}^*}$, où a_{ij}^* est tiré de (2.17).

Nous utilisons la même logique pour obtenir un échantillon de ce modèle comportant les poids ajustés, que celle que nous avons utilisée précédemment dans le modèle sans poids. L'inférence des prédictions de la population diffère parce que nous obtenons les prédictions de la population comme suit :

$$\overline{Y}_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \nu_i, \sigma^2 \sim \text{Normal}\left(\mathbf{x}_i'\boldsymbol{\beta} + \nu_i, \frac{\sigma^2}{\hat{N}_i}\right) \qquad i = 1, \ldots, \ell, \tag{B.8}$$

où $\hat{N}_i = \sum_{j=1}^{n_i} v_{ij}$ représente l'estimateur Horovitz-Thompson de la taille de la population de chaque strate $i = 1, \ldots, \ell$.

Bibliographie

Albert, J.H., et Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422), 669-679. https://doi.org/10.2307/2290350.

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6), 1152-1174. http://www.jstor.org/stable/2958336.

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Blackwell, D., et MacQueen, J.B. (1973). Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1(2), 353-355. http://www.jstor.org/stable/2958020.

Box, G.E.P., et Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Longman Higher Education. ISBN 10: 0201006227/ISBN 13: 9780201006223.

Chipman, H.A., George, E.I. et McCulloch, R.E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, 93(443), 935-948. https://doi.org/10.2307/2669832.

Chipman, H.A., George, E.I. et McCulloch, R.E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 266-298. http://www.jstor.org/stable/27801587.

Chung, H.C., et Datta, G.S. (2022). Modèles spatiaux bayésiens pour l'estimation des moyennes pour petites régions échantillonnées et non échantillonnées. *Techniques d'enquête*, 48, 2, 507-535. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2022002/article/00012-fra.pdf.

Datta, G., et Ghosh, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. *Annals of Statistics*, 19(4), 1748-1770. https://doi.org/10.1214/aos/1176348369.

Ghosh, M., Natarajan, K., Stroud, T.W.F. et Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93(441), 273-282. https://doi.org/10.2307/2669623.

Hill, J., Linero, A. et Murray, J. (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and its Application*. 7, 251-278. https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-031219-041110.

Lindley, D.V., et Smith, A.F.M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society,* Series B (Methodological), 34(1), 1-41. http://www.jstor.org/stable/2985048.

Nandram, B. (2007). Bayesian predictive inference under informative sampling via surrogate samples. *Bayesian Statistics and its Applications*, (Éds., S.K. Upadhyay, U. Singh et D. Dey), Anamaya, New Delhi, Chapter 25, 356-374.

Nandram, B., et Choi, J.W. (2005). Modèles de régression hiérarchiques bayésiens sous non-réponse non-ignorable pour petits domaines : Une application aux données de la NHANES. *Techniques d'enquête*, 31, 1, 79-92. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2005001/article/8089-fra.pdf.

Nandram, B., et Rao, J.N.K. (2021). A Bayesian approach for integrating a small probability sample with a non-probability sample. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1568-1603. http://www.asasrms.org/Proceedings/y2021/files/1912256.pdf.

Nandram, B., Toto, M.C. et Choi, J.W. (2011). A Bayesian benchmarking of the Scott-Smith model for small areas. *Journal of Statistical Computation and Simulation*, 81, 1593-1608. https://www.semanticscholar.org/paper/A-Bayesian-benchmarking-of-the-Scott%E2%80%93Smith-model-Nandram-Toto/a18cd37adaea51d06e81b2b525f61526d028fd73.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation*, Wiley Series in Survey Methodology.

Ritter, C., et Tanner, M.A. (1992). Facilitating the Gibbs Sampler: The Gibbs stopper and the Griddy-Gibbs sampler. *Journal of the American Statistical Association*, 87(419), 861-868. https://doi.org/10.2307/2290225.

Scott, A., et Smith, T.M.F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association*, 101, 1387-1397.

Teh, Y.W., Jordan, M.I., Beal, M.J. et Blei, D.M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 1566-1581. http://www.jstor.org/stable/27639773.

Yang, L., Nandram, B. et Choi, J.W. (2023). Bayesian predictive inference under nine methods for incorporating survey weights. *International Journal of Statistics and Probability*, l2, 1. https://ccsenet.org/journal/index.php/ijsp/article/view/0/48223.

Yin, J., et Nandram, B. (2020). A Bayesian small area model with Dirichlet processes on the responses. *Statistics in Transition New Series*, ISSN 2450-0291, Exeley, New York, 21, 3, 1-19, https://doi.org/10.21307/stattrans-2020-041.

Techniques d'enquête, décembre 2024
Vol. 50, nº 2, pp. 531-557
Statistique Canada, nº 12-001-X au catalogue

Algorithme récursif de Neyman pour la répartition optimale d'échantillons sous contraintes de boîtes sur les tailles d'échantillons dans les strates

Jacek Wesołowski, Robert Wieczorkowski et Wojciech Wójciak[1]

Résumé

La répartition optimale de l'échantillon dans le cadre d'un échantillonnage stratifié est l'une des questions fondamentales des techniques d'enquête. Il s'agit d'une procédure consistant à diviser la taille globale de l'échantillon en strates de telle sorte que, pour des plans d'échantillonnage donnés dans les strates, la variance de l'*estimateur stratifié* π du total (ou de la moyenne) de la population pour une variable à l'étude donnée atteigne son minimum. Dans le présent travail, nous considérons la répartition optimale d'un échantillon, sous des bornes inférieures et supérieures imposées conjointement sur les tailles d'échantillon dans les strates. Nous nous intéressons à la fonction de variance d'une forme générique qui, en particulier, couvre le cas de l'*échantillonnage aléatoire simple sans remise* dans les strates. L'objectif du présent document est double. Tout d'abord, nous établissons (à l'aide des conditions de Karush-Kuhn-Tucker) une forme générique de la solution optimale, appelée « conditions d'optimalité ». Ensuite, sur la base des conditions d'optimalité établies, nous dérivons un algorithme récursif efficace, appelé « *RNABOX* », qui résout le problème de répartition étudié. Le *RNABOX* peut être considéré comme une généralisation de l'algorithme récursif classique de répartition de Neyman, un outil populaire pour la répartition optimale lorsque seules des bornes supérieures sont imposées à la taille des strates d'échantillonnage. Nous mettons en œuvre le *RNABOX* dans `R` dans le cadre de notre paquet `stratallo` qui est disponible dans le dépôt Comprehensive R Archive Network (CRAN).

Mots-clés : Algorithme récursif de Neyman; échantillonnage stratifié; répartition de Neyman; répartition optimale de l'échantillon; répartition optimale sous contraintes de boîtes.

1. Introduction

Considérons une population finie U de taille N. Supposons que le paramètre d'intérêt soit le total de la population t d'une variable y dans U, c'est-à-dire $t = \sum_{k \in U} y_k$, où y_k représente la valeur de y pour l'élément de population $k \in U$. Pour estimer t, nous considérons l'*échantillonnage stratifié* avec l'*estimateur* π. Selon cette technique d'échantillonnage bien connue, la population U est stratifiée, c'est-à-dire $U = \bigcup_{h \in \mathcal{H}} U_h$, où $U_h, h \in \mathcal{H}$, appelées strates, sont disjointes et non vides, et où $\mathcal{H}$ désigne un ensemble fini d'étiquettes de strates. La taille de la strate U_h est désignée par $N_h, h \in \mathcal{H}$ et clairement $\sum_{h \in \mathcal{H}} N_h = N$. Les échantillons probabilistes $s_h \subseteq U_h$ de taille $n_h \le N_h, h \in \mathcal{H}$, sont sélectionnés indépendamment dans chaque strate selon des plans d'échantillonnage choisis qui sont souvent du même type dans toutes les strates. L'échantillon total qui en résulte est de taille $n = \sum_{h \in \mathcal{H}} n_h \le N$. Il est bien connu que l'*estimateur stratifié* π $\hat{t}_\pi$ de t et sa variance sont exprimés en termes de probabilités d'inclusion de premier et de second ordre (pour obtenir un exemple, voir Särndal, Swensson et Wretman, 1992, Résultat 3.7.1, page 102). En particulier, pour plusieurs plans d'échantillonnage importants

1. Jacek Wesołowski, Programmation, Coordination of Statistical Surveys and Registers Department, Statistics Poland, Aleja Niepodległości, 208, 00-925 Varsovie, Pologne, et Faculty of Mathematics and Information Science, Warsaw University of Technology, ul. Koszykowa 75, 00-662 Varsovie, Pologne. Courriel : jacek.wesolowski@pw.edu.pl; Robert Wieczorkowski, Programmation, Coordination of Statistical Surveys and Registers Department, Statistics Poland, Aleja Niepodległości 208, 00-925 Varsovie, Pologne. Courriel : R.Wieczorkowski@stat.gov.pl; Wojciech Wójciak, Faculty of Mathematics and Information Science, Warsaw University of Technology, ul. Koszykowa 75, 00-662 Varsovie, Pologne. Courriel : wojciech.wojciak.dokt@pw.edu.pl.

$$\mathrm{Var}\left(\hat{t}_\pi\right) = \sum_{h\in\mathcal{H}} \tfrac{A_h^2}{n_h} - B, \tag{1.1}$$

où $A_h > 0, B$ ne dépend pas de $n_h, h \in \mathcal{H}$. Parmi les plans d'échantillonnage les plus fondamentaux et les plus courants qui donnent lieu à la variance de la forme (1.1) figure l'*échantillonnage aléatoire simple sans remise* dans les strates. Dans ce cas, l'estimateur *stratifié* π de t prend la forme

$$\hat{t}_\pi = \sum_{h\in\mathcal{H}} \tfrac{N_h}{n_h} \sum_{k\in s_h} y_k, \tag{1.2}$$

ce qui donne en (1.1) : $A_h = N_h S_h$, où S_h représente l'écart-type de la strate de la variable à l'étude y, $h \in \mathcal{H}$, et $B = \sum_{h\in\mathcal{H}} N_h S_h^2$ (voir, par exemple, Särndal et coll., 1992, résultat 3.7.2, page 103).

Le problème classique de la répartition optimale des échantillons est formulé comme la détermination du vecteur de répartition $\mathbf{n} = \left(n_h, h \in \mathcal{H}\right)$ qui minimise la variance (1.1), sous la contrainte $\sum_{h\in\mathcal{H}} n_h = n$, pour un $n \le N$ donné (voir, par exemple, Särndal et coll., 1992, section 3.7.3, page 104). Dans le présent article, nous nous intéressons au problème classique de la répartition optimale des échantillons avec des contraintes bilatérales supplémentaires imposées sur la taille des échantillons dans les strates. Dans le langage de l'optimisation mathématique, nous formulons ce problème sous la forme du problème 1.1.

Problème 1.1. Étant donné un ensemble fini $\mathcal{H} \neq \varnothing$ et des nombres $A_h > 0, m_h, M_h, n,$ de sorte que $0 < m_h < M_h \le N_h, h \in \mathcal{H}$ et $\sum_{h\in\mathcal{H}} m_h \le n \le \sum_{h\in\mathcal{H}} M_h$,

$$\underset{\mathbf{x}=(x_h,\, h\in\mathcal{H})\in\mathbb{R}_+^{|\mathcal{H}|}}{\text{minimiser}} \quad \sum_{h\in\mathcal{H}} \tfrac{A_h^2}{x_h} \tag{1.3}$$

$$\text{sous la contrainte} \quad \sum_{h\in\mathcal{H}} x_h = n \tag{1.4}$$

$$m_h \le x_h \le M_h, \quad h \in \mathcal{H}. \tag{1.5}$$

Pour souligner le fait que la solution optimale au problème 1.1 peut ne pas être un nombre entier, nous désignons la variable d'optimisation par $\mathbf{x}$, et non par $\mathbf{n}$. Les hypothèses concernant $n, m_h, M_h, h \in \mathcal{H}$, garantissent que le problème 1.1 est réalisable.

Les bornes supérieures imposées à $x_h, h \in \mathcal{H}$, sont naturelles puisque, par exemple, la solution avec $x_h > N_h$ pour une certaine $h \in \mathcal{H}$ est impossible. Les bornes inférieures sont nécessaires, par exemple, pour l'estimation des variances des strates de population $S_h^2, h \in \mathcal{H}$. Elles apparaissent également lorsque nous traitons les strates comme des domaines et que nous attribuons des bornes supérieures aux variances des estimateurs des totaux dans les domaines. Une telle approche a été envisagée par exemple dans Choudhry, Rao et Hidiroglou (2012), où, à l'exception des contraintes de bornes supérieures $x_h \le N_h, h \in \mathcal{H}$, les contraintes supplémentaires $\left(\frac{1}{x_h} - \frac{1}{N_h}\right) N_h^2 S_h^2 \le R_h, h \in \mathcal{H}$, où $R_h, h \in \mathcal{H}$, sont des constantes données, ont été imposées. Évidemment, ce dernier système d'inégalités peut être réécrit comme des contraintes de bornes inférieures de la forme $x_h \ge m_h = \frac{N_h^2 S_h^2}{R_h + N_h S_h^2}, h \in \mathcal{H}$. La solution donnée dans Choudhry et coll. (2012) a été obtenue par la procédure basée sur l'algorithme de Newton-Raphson, une méthode numérique générale de

recherche de racines. Voir également un article connexe de Wright, Noble et Bailer (2007), dans lequel le problème de la répartition optimale sous la contrainte d'une précision égale pour l'estimation des moyennes des strates a été examiné.

Il est utile d'introduire la définition suivante pour les solutions réalisables du problème 1.1.

Définition 1.1. *Tout vecteur* $\mathbf{x} = (x_h, h \in \mathcal{H})$ *satisfaisant à (1.4) et (1.5) sera appelé «* répartition *».*

Une répartition $\mathbf{x} = (x_h, h \in \mathcal{H})$ *est appelée de «* sommet *» si et seulement si*

$$x_h = \begin{cases} m_h, & h \in \mathcal{L} \\ M_h, & h \in \mathcal{U}, \end{cases}$$

où $\mathcal{L}, \mathcal{U} \subseteq \mathcal{H}$ *sont tels que* $\mathcal{L} \cup \mathcal{U} = \mathcal{H}$ *et* $\mathcal{L} \cap \mathcal{U} = \varnothing$.

Une répartition qui n'est pas une répartition de sommet *sera appelée « répartition* régulière *».*

La solution au problème 1.1 sera appelée « répartition optimale *».*

Il convient de mentionner qu'une répartition optimale peut avoir une forme de type de *sommet* ou une forme *régulière*. Le nom de répartition de *sommet* fait référence au fait que dans ce cas, $\mathbf{x}$ est un sommet de l'hyperrectangle $\times_{h \in \mathcal{H}} [m_h, M_h]$. Nous notons que le problème 1.1 devient trivial si $n = \sum_{h \in \mathcal{H}} m_h$ ou $n = \sum_{h \in \mathcal{H}} M_h$. Dans le premier cas, la solution est $\mathbf{x}^* = (m_h, h \in \mathcal{H})$, et dans le second, $\mathbf{x}^* = (M_h, h \in \mathcal{H})$. Ces deux cas sont des cas limites de la répartition de *sommet*. Dans les enquêtes réelles comportant de nombreuses strates, on ne s'attendrait pas à ce qu'une allocation de *sommet* permette une répartition optimale. Néanmoins, par souci d'exhaustivité, nous considérons également ce cas dans le théorème 3.1, qui décrit la forme du vecteur de répartition optimale. Nous notons également qu'une répartition optimale *régulière* $\mathbf{x}^* \in \times_{h \in \mathcal{H}} (m_h, M_h)$ si et seulement si elle est la répartition classique de Tschuprow-Neyman

$$\mathbf{x}^* = \left(A_h \frac{n}{\sum_{v \in \mathcal{H}} A_v}, h \in \mathcal{H} \right)$$

(voir Neyman, 1934; Tschuprow, 1923).

Le reste du document est organisé de la façon suivante. La section 2 présente les motivations de cette étude ainsi qu'une brève analyse documentaire. Dans la section 3, nous déterminons que le problème 1.1 est un problème d'optimisation convexe et utilisons ensuite les conditions de Karush-Kuhn-Tucker pour établir des conditions nécessaires et suffisantes à une solution au problème d'optimisation 1.1. Ces conditions, appelées « conditions d'optimalité », sont présentées dans le théorème 3.1. Dans la section 4, sur la base de ces conditions d'optimalité, nous introduisons un nouvel algorithme, le *RNABOX*, et prouvons qu'il résout le problème 1.1 (voir le théorème 4.1). Le nom *RNABOX* fait référence au fait que cet algorithme généralise l'algorithme récursif de Neyman, désigné par le sigle *RNA* dans le présent document. Le *RNA* est une procédure d'attribution bien établie, couramment utilisée dans la pratique quotidienne des enquêtes. Il trouve une solution au problème de répartition 2.1 (voir ci-dessous), qui est une version assouplie du

problème 1.1. Dans la section 5, nous discutons des expériences numériques liées à l'efficacité de calcul de l'algorithme *RNABOX* et de l'algorithme de l'itération du point fixe (*fixed-point iteration, désigné par le sigle* FPIA*)* de Münnich, Sachs et Wagner (2012). Un résumé concis des résultats est présenté à la section 6, où nous commentons également brièvement certains aspects clés de l'arrondissement des répartitions optimales non entières. Les remarques et lemmes auxiliaires ainsi que les preuves des deux théorèmes figurent dans l'annexe.

Enfin, notons que la mise en œuvre de l'algorithme *RNABOX* est disponible dans notre paquet `R stratallo` (Wójciak, 2023b), qui est publiée dans le dépôt Comprehensive R Archive Network (CRAN) (R Core Team, 2023).

2. Motivation et analyse documentaire

Une abondante littérature est consacrée au problème de la répartition optimale de l'échantillon, qui remonte à la solution classique de Tschuprow (1923) et Neyman (1934), consacrée à l'échantillonnage aléatoire simple sans remise sans tenir compte des contraintes d'inégalité (1.5). Malgré cela, une analyse approfondie de la littérature montre que le problème 1.1 n'a pas encore été complètement compris et qu'il souffre de l'absence d'algorithmes pleinement satisfaisants.

Ci-dessous, nous passons brièvement en revue les méthodes existantes pour résoudre le problème 1.1, y compris les méthodes qui fournissent des solutions à valeurs entières.

2.1 Répartition non nécessairement basée sur des valeurs entières

Une solution approximative au problème 1.1 peut être obtenue par des méthodes génériques de programmation non linéaire (pour obtenir un exemple voir, la monographie Valliant, Dever et Kreuter, 2018 et les références qui y figurent). Ces méthodes ont été utilisées pour résoudre le problème de la répartition optimale des échantillons, car la résolution du problème de la répartition équivaut à trouver les points extrêmes (c'est-à-dire les points stationnaires) d'une certaine fonction objective sur un ensemble réalisable. Connaissant l'extrême (approximatif) de la fonction objective, nous pouvons déterminer les tailles (approximatives, mais généralement suffisamment précises) des échantillons alloués aux strates individuelles.

Suivant une approche similaire mais différente adoptée par exemple dans Münnich et coll. (2012), le problème 1.1 est transformé en problèmes de recherche de racines ou de points fixes (d'une fonction correctement définie) dont la solution est obtenue par des algorithmes à usage général comme *bisection* ou *regula falsi*.

Les algorithmes utilisés dans ces deux approches comportent en principe un nombre infini d'étapes et sont arrêtés par une décision arbitraire, généralement liée à la précision des itérations. Ce mode de fonctionnement présente deux faiblesses principales : l'absence de convergence de la méthode ou la lenteur de la convergence vers la solution optimale pour certains points de départ médiocres. En d'autres termes, les performances de ces algorithmes peuvent dépendre fortement du choix initial d'un point de départ, et ce choix est presque toujours quelque peu hasardeux. Prenons par exemple l'algorithme *FPIA*, de Münnich

et coll. (2012). Pour une population à quatre strates, de sorte que $A_1 = 380$, $A_2 = 140$, $A_3 = 230$, $A_4 = 1\,360$, les bornes $m_h = 10$, $M_h = 50$, $h \in \{1, 2, 3, 4\}$, la taille totale de l'échantillon $n = 80$, et pour le point de départ $\lambda_0 = 695{,}64$ (choisi de la manière suggérée dans le présent article), l'algorithme *FPIA* ne converge pas en raison d'oscillations autour de la solution optimale. Un autre inconvénient des algorithmes de ce type est leur sensibilité aux problèmes d'arithmétique de précision finie qui peuvent survenir lorsque le critère d'arrêt n'est pas exprimé directement en termes d'itérations du vecteur de répartition (ce qui est souvent le cas).

Contrairement à cela, dans les algorithmes récursifs (qui nous intéressent), la solution optimale est toujours trouvée par une recherche récursive de candidats réalisables pour la répartition optimale parmi les sous-ensembles de $\mathcal{H}$. Par conséquent, ils s'arrêtent toujours à la solution exacte et après un nombre fini d'itérations (ne dépassant pas le nombre de strates + 1, comme nous le verrons pour le cas du *RNABOX* dans la preuve du théorème 4.1). Un exemple important d'un tel algorithme est l'algorithme récursif de Neyman, *RNA*, dédié au problème 2.1, une version assouplie du problème 1.1.

Problème 2.1. Étant donné un ensemble fini $\mathcal{H} \neq \varnothing$ et les nombres $A_h > 0, M_h, n > 0,$ de sorte que $0 < M_h \leq N_h, h \in \mathcal{H}$ et $n \leq \sum_{h \in \mathcal{H}} M_h$,

$$\begin{aligned} &\underset{\mathbf{x}=(x_h,\, h\in\mathcal{H})\in\mathbb{R}_+^{|\mathcal{H}|}}{\text{minimiser}} \ \sum_{h\in\mathcal{H}} \frac{A_h^2}{x_h} \\ &\text{sous la contrainte} \ \sum_{h\in\mathcal{H}} x_h = n \\ &\qquad\qquad\qquad\quad x_h \leq M_h, \quad h \in \mathcal{H}. \end{aligned}$$

Bien que le *RNA* soit populaire parmi les praticiens, une preuve formelle du fait qu'il donne la solution optimale au problème 2.1 n'a été donnée que récemment dans Wesołowski, Wieczorkowski et Wójciak (2022). Pour examiner d'autres approches récursives du problème 2.1, voir aussi Stenger et Gabler (2005), Kadane (2005).

À notre connaissance, le seul algorithme récursif de répartition optimale à valeur non entières décrit dans la littérature et destiné à résoudre le problème 1.1 est la procédure *noptcond* proposée par Gabler, Ganninger et Münnich (2012). Contrairement à *RNABOX*, cette méthode permet notamment d'effectuer un tri par strates. Malheureusement, la répartition calculée par *noptcond* peut ne pas donner le minimum de la fonction objective (1.3). Ce fait peut être illustré par un court exemple numérique donné dans le tableau 2.1, qui suit Wójciak (2019, exemple 3.9).

Tableau 2.1
Deux répartitions pour une population à deux strates : répartition non optimale $\mathbf{x}^{\text{noptcond}}$ et répartition optimale $\mathbf{x}^*$.

h	A_h	m_h	M_h	$\mathbf{x}^{\text{noptcond}}$	$\mathbf{x}^*$
1	2 000	30	50	30	50
2	3 000	40	200	130	110
Taille totale de l'échantillon $n = 160$.					

2.2 Répartition à valeur entière

Des algorithmes à valeurs entières dédiés au problème 1.1 sont proposés dans Friedrich, Münnich, de Vries et Wagner (2015), Wright (2017, 2020). La version multivariée du problème de répartition optimale d'échantillons sous contrainte de boîtes dans laquelle $m_h = m, h \in \mathcal{H},$ pour une constante donnée $m,$ est étudiée dans l'article de Moura Brito, do Nascimento Silva, Silva Semaan et Maculan (2015). La procédure proposée pour résoudre ce problème repose sur un algorithme de programmation binaire en nombres entiers et peut être appliquée au cas univarié. Voir également Brito, Silva et Veiga (2017) pour obtenir des renseignements sur la mise en œuvre de cette approche dans le langage `R`.

Les méthodes de répartition à valeurs entières proposées dans ces documents sont précises (et non approximatives) et théoriquement solides. Cependant, elles sont relativement lentes par rapport aux algorithmes à valeurs non nécessairement entières. Par exemple, au moins pour les contraintes unilatérales, l'algorithme à valeurs entières *capacity scaling* de Friedrich et coll. (2015) peut être des milliers de fois plus lent que le *RNA* (voir Wesołowski et coll., 2022, section 4). Cela semble être un inconvénient majeur de ces méthodes, car les différences de variances des estimateurs basés sur la répartition optimale à valeurs non entières arrondies à nombres entiers et la répartition optimale à valeur entières sont négligeables, comme il est expliqué dans la section 6. L'efficacité de calcul est particulièrement importante lorsque le nombre de strates est élevé, voir, par exemple, l'application au recensement allemand dans Burgard et Münnich (2012), et elle est encore plus prononcée dans les solutions itératives aux problèmes de stratification, lorsque le nombre d'itérations peut se compter en millions (pour consulter un exemple, voir Khan, Nand et Ahmad, 2008; Baillargeon et Rivest, 2011; Barcaroli, 2014; Gunning et Horgan, 2004; Lednicki et Wieczorkowski, 2003).

Cela étant dit, la recherche de nouveaux algorithmes récursifs universels, théoriquement solides et efficaces en termes de calcul pour la répartition optimale des échantillons sous des contraintes bilatérales sur les tailles des strates de l'échantillon est cruciale à la fois pour la théorie et la pratique de l'échantillonnage des sondages.

3. Conditions d'optimalité

Dans la section qui suit, nous établissons des conditions d'optimalité, c'est-à-dire une forme générale de la solution au problème 1.1. Comme nous le verrons à la section 4, ces conditions d'optimalité sont cruciales pour la construction de l'algorithme *RNABOX*.

Avant d'établir les conditions d'optimalité nécessaires et suffisantes pour une solution au problème d'optimisation 1.1, nous définissons d'abord une fonction d'ensemble $s,$ ce qui simplifie considérablement la notation et les calculs.

Définition 3.1. Soit $\mathcal{H}, n, A_h > 0, m_h, M_h, h \in \mathcal{H}$ comme dans le problème 1.1 et soit $\mathcal{L}, \mathcal{U} \subseteq \mathcal{H}$ de sorte que $\mathcal{L} \cap \mathcal{U} = \varnothing, \mathcal{L} \cup \mathcal{U} \subsetneq \mathcal{H}.$ La fonction d'ensemble s est définie comme suit :

$$s(\mathcal{L}, \mathcal{U}) = \frac{n - \sum_{h \in \mathcal{L}} m_h - \sum_{h \in \mathcal{U}} M_h}{\sum_{h \in \mathcal{H} \setminus (\mathcal{L} \cup \mathcal{U})} A_h}. \tag{3.1}$$

Ci-dessous, nous introduirons le vecteur $\mathbf{x}^{(\mathcal{L},\mathcal{U})}$ pour $\mathcal{L},\mathcal{U}\subseteq\mathcal{H}$ disjoint. Il apparaît que la solution du problème 1.1 est nécessairement de la forme (3.2), les ensembles $\mathcal{L}$ et $\mathcal{U}$ étant définis implicitement par des systèmes d'équations ou d'inégalités établis dans le théorème 3.1.

Définition 3.2. Soit $\mathcal{H}, n, A_h>0, m_h, M_h, h\in\mathcal{H}$ comme dans le problème 1.1, et soit $\mathcal{L},\mathcal{U}\subseteq\mathcal{H}$ de sorte que $\mathcal{L}\cap\mathcal{U}=\varnothing$. Nous définissons le vecteur $\mathbf{x}^{(\mathcal{L},\mathcal{U})}=\left(x_h^{(\mathcal{L},\mathcal{U})}, h\in\mathcal{H}\right)$ comme suit :

$$x_h^{(\mathcal{L},\mathcal{U})}=\begin{cases} m_h, & h\in\mathcal{L} \\ M_h, & h\in\mathcal{U} \\ A_h\, s\left(\mathcal{L},\mathcal{U}\right), & h\in\mathcal{H}\setminus\left(\mathcal{L}\cup\mathcal{U}\right). \end{cases} \tag{3.2}$$

Le théorème 3.1 suivant caractérise la forme de la solution optimale au problème 1.1 et constitue donc l'un des principaux résultats du présent document.

Théorème 3.1 (conditions d'optimalité)**.** *Le problème d'optimisation 1.1 a une solution optimale unique. Le point* $\mathbf{x}^*\in\mathbb{R}_+^{|\mathcal{H}|}$ *est une solution au problème d'optimisation 1.1 si et seulement si* $\mathbf{x}^*=\mathbf{x}^{(\mathcal{L}^*,\mathcal{U}^*)}$, *avec* $\mathcal{L}^*,\mathcal{U}^*\subseteq\mathcal{H}$ *disjoint, de sorte que l'un des deux cas suivants se vérifie :*

CAS I : $\mathcal{L}^*\cup\mathcal{U}^*\subsetneq\mathcal{H}$ *et*

$$\begin{aligned} \mathcal{L}^* &= \left\{h\in\mathcal{H}:\, s\left(\mathcal{L}^*,\mathcal{U}^*\right)\le\tfrac{m_h}{A_h}\right\}, \\ \mathcal{U}^* &= \left\{h\in\mathcal{H}:\, s\left(\mathcal{L}^*,\mathcal{U}^*\right)\ge\tfrac{M_h}{A_h}\right\}. \end{aligned} \tag{3.3}$$

CAS II : $\mathcal{L}^*\cup\mathcal{U}^*=\mathcal{H}$ *et*

$$\max_{h\in\mathcal{U}^*}\tfrac{M_h}{A_h}\le\min_{h\in\mathcal{L}^*}\tfrac{m_h}{A_h}, \qquad \textit{si}\ \ \mathcal{U}^*\neq\varnothing\ \ \textit{et}\ \ \mathcal{L}^*\neq\varnothing, \tag{3.4}$$

$$\sum_{h\in\mathcal{L}^*}m_h+\sum_{h\in\mathcal{U}^*}M_h=n. \tag{3.5}$$

Remarque 3.1. *La répartition optimale* $\mathbf{x}^*$ *est une répartition* régulière *dans le CAS I et une répartition de* sommet *dans le CAS II.*

La preuve du théorème 3.1 est donnée dans l'annexe A. Notons que le théorème 3.1 décrit la forme générale de la répartition optimale jusqu'à la spécification des ensembles de strates *take-min* et *take-max* $\mathcal{L}^*$ et $\mathcal{U}^*$. La façon de définir les ensembles $\mathcal{L}^*$ et $\mathcal{U}^*$ qui déterminent la solution optimale $\mathbf{x}^*=\mathbf{x}^{(\mathcal{L}^*,\mathcal{U}^*)}$ fait l'objet de la section 4.

4. Algorithme récursif de Neyman sous contraintes de boîtes

4.1 L'algorithme *RNABOX*

Dans la section qui suit, nous présentons un algorithme permettant de résoudre le problème 1.1. Compte tenu du théorème 3.1, sa tâche essentielle consiste à diviser l'ensemble de toutes les étiquettes de strates $\mathcal{H}$

en trois sous-ensembles : *take-min* $\left(\mathcal{L}^*\right)$, *take-max* $\left(\mathcal{U}^*\right)$, et *take-Neyman* $\left(\mathcal{H}\setminus\left(\mathcal{L}^*\cup\mathcal{U}^*\right)\right)$. Nous appelons ce nouvel algorithme le *RNABOX,* car il généralise l'algorithme *RNA* existant dans le sens où le *RNABOX* résout le problème de la répartition optimale avec des bornes inférieures et supérieures simultanées, alors que le *RNA* est dédié au problème avec des bornes supérieures uniquement, c'est-à-dire pour le problème 2.1. De plus, *RNABOX* utilise le *RNA* dans l'une de ses étapes intermédiaires. Nous rappelons tout d'abord l'algorithme *RNA* et présentons ensuite l'algorithme *RNABOX*. Pour obtenir plus de renseignements sur l'algorithme *RNA*, voir Wesołowski et coll. (2022, section 2) ou Särndal et coll. (1992, remarque 12.7.1, page 466).

Algorithme *RNA*

Entrée :	$\mathcal{H}, \left(A_h\right)_{h\in\mathcal{H}}, \left(M_h\right)_{h\in\mathcal{H}}, n.$
Obligatoire :	$A_h > 0, M_h > 0, h\in\mathcal{H}, 0 < n \le \sum_{h\in\mathcal{H}} M_h.$
Étape 1 :	Attribuer une valeur à $\mathcal{U} = \varnothing.$
Étape 2 :	Déterminer $\tilde{\mathcal{U}} = \left\{h\in\mathcal{H}\setminus\mathcal{U}:\ A_h\, s\left(\varnothing,\mathcal{U}\right) \ge M_h\right\},$ où la fonction d'ensemble s est définie dans (3.1).
Étape 3 :	Si $\tilde{\mathcal{U}} = \varnothing,$ passer à l'étape 4. Sinon, mettre à jour $\mathcal{U}\leftarrow\mathcal{U}\cup\tilde{\mathcal{U}}$ et passer à l'étape 2.
Étape 4 :	Retourner $\mathbf{x}^* = \left(x_h^*, h\in\mathcal{H}\right)$ avec $x_h^* = \begin{cases} M_h, & h\in\mathcal{U} \\ A_h s\left(\varnothing,\mathcal{U}\right), & h\in\mathcal{H}\setminus\mathcal{U}. \end{cases}$

Algorithme *RNABOX*

Entrée :	$\mathcal{H}, \left(A_h\right)_{h\in\mathcal{H}}, \left(m_h\right)_{h\in\mathcal{H}}, \left(M_h\right)_{h\in\mathcal{H}}, n.$
Obligatoire :	$A_h > 0, 0 < m_h < M_h, h\in\mathcal{H}, \sum_{h\in\mathcal{H}} m_h \le n \le \sum_{h\in\mathcal{H}} M_h.$
Étape 1 :	Attribuer une valeur à $\mathcal{L} = \varnothing.$
Étape 2 :	Exécuter le *RNA* $\left[\mathcal{H}, \left(A_h\right)_{h\in\mathcal{H}}, \left(M_h\right)_{h\in\mathcal{H}}, n\right]$ pour obtenir $\left(x_h^{**}, h\in\mathcal{H}\right).$ Soit $\mathcal{U} = \left\{h\in\mathcal{H}:\ x_h^{**} = M_h\right\}.$
Étape 3 :	Déterminer $\tilde{\mathcal{L}} = \left\{h\in\mathcal{H}\setminus\mathcal{U}:\ x_h^{**} \le m_h\right\}.$
Étape 4 :	Si $\tilde{\mathcal{L}} = \varnothing,$ passer à l'étape 5. Sinon, mettre à jour $n\leftarrow n - \sum_{h\in\tilde{\mathcal{L}}} m_h, \mathcal{H}\leftarrow\mathcal{H}\setminus\tilde{\mathcal{L}},$ $\mathcal{L}\leftarrow\mathcal{L}\cup\tilde{\mathcal{L}}$ et passer à l'étape 2.
Étape 5 :	Retourner $\mathbf{x}^* = \left(x_h^*, h\in\mathcal{L}\cup\mathcal{H}\right)$ avec $x_h^* = \begin{cases} m_h, & h\in\mathcal{L} \\ x_h^{**}, & h\in\mathcal{H}. \end{cases}$

Nous notons que dans les applications réelles, les nombres $\left(A_h\right)_{h\in\mathcal{H}}$ sont généralement inconnus et que leurs estimations $\left(\hat{A}_h\right)_{h\in\mathcal{H}}$ sont donc utilisées à la place dans l'entrée des algorithmes.

Le théorème 4.1 est le principal résultat théorique de ce document et sa preuve est donnée à l'annexe B.

Théorème 4.1. L'algorithme *RNABOX* fournit la solution optimale au problème 1.1.

4.2 Exemple de rendement de l'algorithme *RNABOX*

Nous démontrons le comportement opérationnel de l'algorithme *RNABOX* pour une population artificielle comportant 10 strates et pour une taille d'échantillon totale $n = 5\,110$, comme le montre le tableau 4.1.

Tableau 4.1
Un exemple du rendement de l'algorithme *RNABOX* pour une population comportant 10 strates et pour une taille d'échantillon totale $n = 5\,110$.

h	A_h	m_h	M_h	$\mathcal{L}_1 / \mathcal{U}_1$	$\mathcal{L}_2 / \mathcal{U}_2$	$\mathcal{L}_3 / \mathcal{U}_3$	$\mathcal{L}_4 / \mathcal{U}_4$	$\mathcal{L}_5 / \mathcal{U}_5$	$\mathcal{L}_6 / \mathcal{U}_6$	x^*
1	2 700	750	900			□	□	□	□	750
2	2 000	450	500	■		□	□	□	□	450
3	4 200	250	300	■	■	■	■	■		261,08
4	4 400	350	400	■	■	■			□	350
5	3 200	150	200	■	■	■	■	■		198,92
6	6 000	550	600	■	■	■		□	□	550
7	8 400	650	700	■	■	■			□	650
8	1 900	50	100	■	■	■	■	■	■	100
9	5 400	850	900	■	■		□	□	□	850
10	2 000	950	1,000		□	□	□	□	□	950
SOMME		5 000	5 600	0/8	1/7	3/6	4/3	5/3	7/1	5 110

Notes : Les colonnes $\mathcal{L}_r / \mathcal{U}_r$, $r = 1, \ldots, 6$, représentent le contenu de ensembles $\mathcal{L}, \mathcal{U}$, respectivement dans la r^{e} itération de l'algorithme *RNABOX* (entre l'étape 3 et l'étape 4) : les symboles □ ou ■ indiquent que la strate avec l'étiquette h est dans $\mathcal{L}_r$ ou $\mathcal{U}_r$, respectivement.

Pour cet exemple, *RNABOX* s'arrête après six itérations avec un ensemble de strates *take-min* $\mathcal{L}^* = \{1, 2, 4, 6, 7, 9, 10\}$, un ensemble de strates *take-max* $\mathcal{U}^* = \{8\}$ et un ensemble de strates *take-Neyman* $\mathcal{H} \setminus \left(\mathcal{L}^* \cup \mathcal{U}^*\right) = \{3, 5\}$ (voir la colonne $\mathcal{L}_6 / \mathcal{U}_6$). La répartition optimale en est une *régulière* et elle est indiquée dans la colonne x^* du tableau 4.1. La valeur correspondante de la fonction objective (1.3) est 441 591,5. Les détails des attributions provisoires de strates aux ensembles $\mathcal{L}, \mathcal{U}$ à chacune des six itérations de l'algorithme sont indiqués dans les colonnes $\mathcal{L}_1 / \mathcal{U}_1 - \mathcal{L}_6 / \mathcal{U}_6$.

4.3 Modifications et améliorations possibles

Solutions de rechange pour le *RNA* à l'étape 2

L'algorithme *RNABOX* utilise le *RNA* à son étape 2. Cependant, il n'est pas difficile de voir que tout algorithme dédié au problème 2.1 (par exemple le *SGA* par Stenger et Gabler, 2005 ou le *COMA* par Wesołowski et coll., 2022) pourrait être utilisé à la place. Nous avons choisi le *RNA,* car cela permet de garder *RNABOX* exempt de tout tri de strates.

Une version jumelle du *RNABOX*

Observons que l'ordre dans lequel $\mathcal{L}$ et $\mathcal{U}$ sont calculés dans l'algorithme peut être interchangé. Un tel changement suppose que le *RNA* utilisé à l'étape 2 du *RNABOX*, soit remplacé par sa version jumelle, le *LRNA*, qui résout le problème de la répartition optimale en fonction de bornes inférieures unilatérales. Le *LRNA* est décrit en détail dans Wójciak (2023a).

Algorithme *LRNA*

Entrée : $\mathcal{H}, \left(A_h\right)_{h\in\mathcal{H}}, \left(m_h\right)_{h\in\mathcal{H}}, n.$

Obligatoire : $A_h > 0, m_h > 0, h \in \mathcal{H}, n \geq \sum_{h\in\mathcal{H}} m_h.$

Étape 1 : Attribuer une valeur à $\mathcal{L} = \varnothing.$

Étape 2 : Déterminer $\tilde{\mathcal{L}} = \left\{h \in \mathcal{H}\setminus\mathcal{L}\colon A_h\, s\left(\mathcal{L}, \varnothing\right) \leq m_h\right\},$ où la fonction d'ensemble s est définie dans (3.1).

Étape 3 : Si $\tilde{\mathcal{L}} = \varnothing,$ passer à l'étape 4. Sinon, mettre à jour $\mathcal{L} \leftarrow \mathcal{L} \cup \tilde{\mathcal{L}}$ et passer à l'étape 2.

Étape 4 : Retourner $\mathbf{x}^* = \left(x_h^*, h \in \mathcal{H}\right)$ avec $x_h^* = \begin{cases} m_h, & h \in \mathcal{L} \\ A_h\, s\left(\mathcal{L}, \varnothing\right), & h \in \mathcal{H}\setminus\mathcal{L}. \end{cases}$

Compte tenu de l'observation ci-dessus, les étapes 2 et 3 du *RNABOX* se lisent comme suit :

Étape 2 : Exécuter $LRNA\left[\mathcal{H}, \left(A_h\right)_{h\in\mathcal{H}}, \left(m_h\right)_{h\in\mathcal{H}}, n\right]$ pour obtenir $\left(x_h^{**}, h \in \mathcal{H}\right).$
Soit $\mathcal{L} = \left\{h \in \mathcal{H}\colon x_h^{**} = m_h\right\}.$

Étape 3 : Déterminer $\tilde{\mathcal{U}} = \left\{h \in \mathcal{H}\setminus\mathcal{L}\colon x_h^{**} \geq M_h\right\}.$

Les autres étapes doivent être adaptées en conséquence.

Utilisation de renseignements *a priori* dans le logarithme *RNA* à l'étape 2

Compte tenu du lemme B.2, en utilisant la notation introduite dans l'annexe B.1, à l'étape 2 de l'algorithme *RNABOX*, pour $r^* \geq 2$ nous avons

$$\mathcal{U}_r = \left\{h \in \mathcal{H}\setminus\mathcal{L}_r\colon x_h^{**} = M_h\right\} \subseteq \mathcal{U}_{r-1}, \qquad r = 2, \ldots, r^*.$$

Cela indique que le domaine de discours de $\mathcal{U}_r$ pourrait être réduit de $\mathcal{H}\setminus\mathcal{L}_r$ à $\mathcal{U}_{r-1} \subseteq \mathcal{H}\setminus\mathcal{L}_r,$ de telle sorte que

$$\mathcal{U}_r = \left\{h \in \mathcal{U}_{r-1}\colon x_h^{**} = M_h\right\}, \qquad r = 2, \ldots, r^*. \tag{4.1}$$

Compte tenu de l'observation ci-dessus et du fait que, du point de vue de la mise en œuvre, l'ensemble $\mathcal{U}_r$ est déterminé en interne par le *RNA*, il est tentant d'envisager la modification du *RNA* de manière à ce qu'il utilise le domaine de discours $\mathcal{U}_{r-1}$ pour l'ensemble $\mathcal{U}_r.$ Ce domaine pourrait être spécifié en tant que paramètre d'entrée supplémentaire, par exemple $\mathcal{J} \subseteq \mathcal{H},$ et l'étape 2 de l'algorithme *RNA* serait alors la suivante :

Étape 2 : Déterminer $\tilde{\mathcal{U}} = \left\{h \in \mathcal{J}\setminus\mathcal{U}\colon A_h\, s\left(\varnothing, \mathcal{U}\right) \geq M_h\right\}.$

Du point de vue de l'algorithme *RNABOX*, ce nouveau paramètre d'entrée du *RNA* doit être fixé à $\mathcal{J} = \mathcal{H}$ pour la première itération, puis à $\mathcal{J} = \mathcal{U}_{r-1}$ pour les itérations suivantes $r = 2, \ldots, r^* \geq 2$ (le cas échéant).

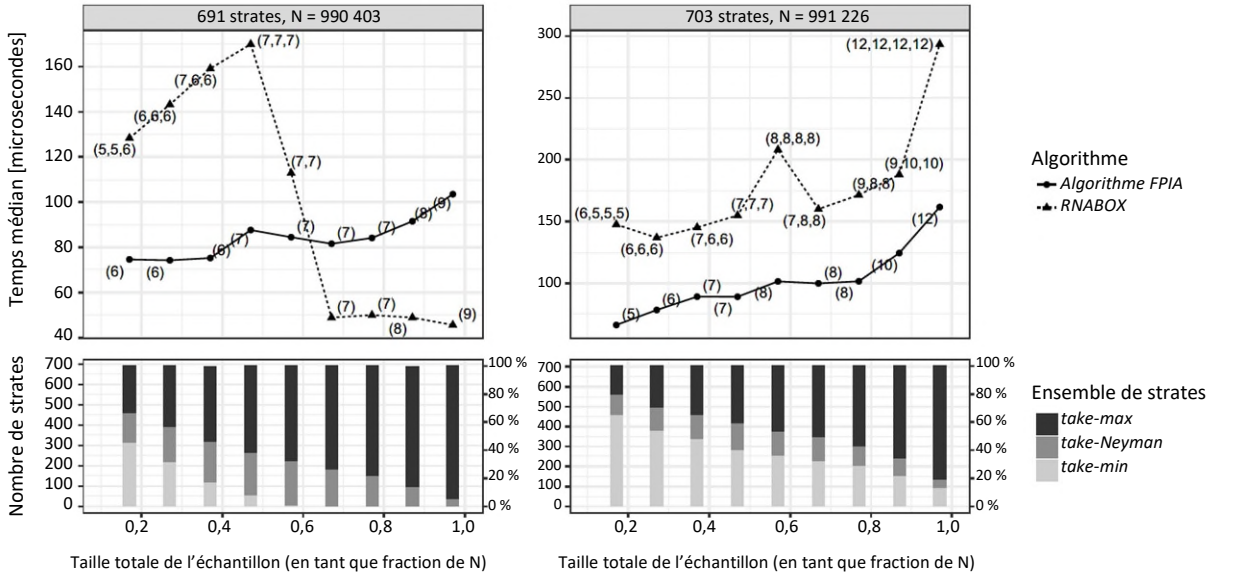
5. Résultats numériques

Dans les simulations, en utilisant le *Statistical Software* `R` (R Core Team, 2023) et le paquet `microbenchmark R` (Mersmann, 2021), nous avons comparé l'efficacité de calcul de l'algorithme *RNABOX* avec l'efficacité de l'algorithme *FPIA* de Münnich et coll. (2012). Ce dernier est connu pour être un algorithme efficace dédié au problème 1.1 et nous l'avons donc utilisé comme référence. La comparaison ne visait pas à vérifier la complexité théorique du calcul, mais plutôt à obtenir des résultats quantitatifs concernant l'efficacité du calcul pour les mises en œuvre particulières des deux algorithmes.

Pour comparer les performances des algorithmes, nous avons utilisé l'*échantillonnage aléatoire simple sans remise* dans les strates pour plusieurs populations créées artificiellement. Ici, nous avons choisi de présenter une simulation pour deux de ces populations comportant 691 et 703 strates, dont les résultats sont représentatifs des autres. Ces deux populations ont été construites en liant itérativement K ensembles de nombres, où K est égal à 100 (pour la première population) et à 200 (pour la deuxième population). Chaque ensemble, désigné par $i = 1, \ldots, K,$ contient 10 000 nombres aléatoires générés indépendamment à partir d'une distribution log-normale ayant les paramètres $\mu = 0$ et $\sigma = \log(1+i).$ Pour chaque ensemble $i = 1, \ldots, K,$ les bornes des strates ont été déterminées par la méthode de stratification géométrique de Gunning et Horgan (2004), le paramètre 10 étant le nombre de strates et le coefficient de variation ciblé étant égal à 0,05. Cette méthode de stratification est mise en œuvre dans le paquet `R stratification`, développée par Rivest et Baillargeon (2022) et décrite dans Baillargeon et Rivest (2011). Pour obtenir de plus amples précisions, voir le code `R` avec les essais, qui est placé dans notre dépôt GitHub (voir Wieczorkowski, Wesołowski et Wójciak, 2023).

Les résultats de ces simulations sont illustrés à la figure 5.1. Cette figure montre que, bien que dans la majorité des cas, l'algorithme *FPIA* soit légèrement plus rapide que l'algorithme *RNABOX*, les durées d'exécution de ces deux algorithmes sont généralement comparables. Le gain en temps d'exécution du *FPIA* résulte du fait qu'il parcourt généralement un plus petit nombre d'ensembles $\mathcal{L}, \mathcal{U} \subset \mathcal{H},$ que le *RNABOX* afin de trouver les ensembles optimaux $\mathcal{L}^*$ et $\mathcal{U}^*.$ Bien que cette approche donne généralement des résultats corrects (comme dans les simulations présentées dans la présente section), il peut arriver que le *FPIA* manque les ensembles optimaux $\mathcal{L}^*, \mathcal{U}^* \subsetneq \mathcal{H}$ et qu'il n'aboutisse donc pas à la répartition optimale correcte. Ce cas rare a été illustré par un exemple numérique donné à la section 2.1. Par ailleurs, nous soulignons que le *FPIA* n'est pas bien défini lorsque la répartition optimale est de type de *sommet*, c'est-à-dire lorsque $\mathcal{L}^* \cup \mathcal{U}^* = \mathcal{H}.$

Figure 5.1 Temps d'exécution de l'algorithme *FPIA* et de l'algorithme *RNABOX* pour deux populations artificielles.



Notes : Les graphiques du haut montrent la médiane empirique des temps d'exécution (calculée à partir de 100 répétitions) pour différentes tailles d'échantillons. Les chiffres entre parenthèses correspondent au nombre d'itérations d'un algorithme donné. Dans le cas de *RNABOX*, il s'agit d'un vecteur indiquant le nombre d'itérations de *RNA* (voir l'étape 2 du *RNABOX*) pour chaque itération de *RNABOX*. Ainsi, la longueur de ce vecteur est égale au nombre d'itérations du *RNABOX*. Les nombres de strates *take-min*, *take-Neyman*, et *take-max* sont indiqués dans les graphiques du bas.

6. Conclusions

Dans le présent document, nous avons examiné le problème 1.1 de la répartition optimale d'échantillons sous contraintes de boîtes. Le principal résultat de ce travail est la formulation mathématiquement précise des conditions nécessaires et suffisantes pour la solution du problème 1.1, données dans le théorème 3.1, ainsi que le développement du nouvel algorithme récursif, appelé « *RNABOX* », qui résout le problème 1.1. Les conditions d'optimalité sont fondamentales pour l'analyse du problème d'optimisation. Elles constituent une base fiable pour le développement d'algorithmes efficaces et peuvent être utilisées comme référence pour toute étude future de nouveaux algorithmes permettant de résoudre le problème 1.1. Les propriétés essentielles de l'algorithme *RNABOX*, qui le distinguent des autres algorithmes et approches existants du problème 1.1, sont les suivantes :

1. Universalité : L'algorithme *RNABOX* fournit une solution optimale à tous les cas de faisabilité du problème 1.1 (y compris le cas d'une répartition optimale de *sommet*).
2. Aucun problème d'initialisation : *RNABOX* ne nécessite aucune initialisation, aucun prétest ou autre qui pourrait avoir une incidence sur les résultats définitifs de l'algorithme. C'est ce qui se passe, par exemple, dans le cas des méthodes génériques de programmation non linéaire.
3. Pas de tri : Le *RNABOX* n'effectue aucun tri des strates.

4. Efficacité de calcul : Le temps d'exécution du *RNABOX* est comparable à celui du *FPIA* (qui est probablement l'algorithme de répartition optimale le plus rapide connu jusqu'à présent pour le problème considéré).
5. Caractère direct : Le *RNABOX* calcule des quantités importantes (y compris les éléments internes du *RNA*) au moyen de formules qui sont exprimées directement en termes de vecteur de répartition $\mathbf{x}^{(\mathcal{L},\mathcal{U})}$ (voir la définition 3.2). Cela réduit le risque de problèmes arithmétiques de précision finie, par rapport aux algorithmes qui basent leurs opérations clés sur certaines variables intermédiaires dont dépend la répartition optimale, comme c'est le cas, par exemple, des méthodes basées sur les méthodes génériques de programmation non linéaire.
6. Nature récursive : Le *RNABOX* applique de manière répétée les étapes 2 et 3 de la répartition à un ensemble de strates réduites par étapes, c'est-à-dire à des versions « plus petites » du même problème. Cela se traduit par une clarté des routines et une façon naturelle de penser le problème de la répartition.
7. Généralisation : Le *RNABOX*, du point de vue de sa construction, est une généralisation de l'algorithme populaire *RNA* qui résout le problème 2.1 de la répartition optimale de l'échantillon sous des limites unilatérales de la taille des strates de l'échantillon.

Enfin, nous tenons à souligner que le problème 1.1 examiné dans le présent document n'est pas un problème de répartition à valeurs entières, alors que les tailles d'échantillon dans les strates doivent bien entendu être des nombres naturels. D'autre part, les algorithmes de répartition optimale à valeurs entières sont relativement lents et peuvent donc être inefficaces dans certaines applications, comme nous l'avons déjà mentionné à la section 2. Si la vitesse d'un algorithme est une préoccupation et qu'un algorithme de répartition à valeur non nécessairement entière est choisi (par exemple le *RNABOX*), la solution naturelle est d'arrondir la répartition optimale non entière fournie par cet algorithme. Dans l'ensemble, cette procédure est toujours beaucoup plus rapide que les algorithmes de répartition à valeurs entières. Cependant, un simple arrondi de la solution non entière ne donne généralement pas le minimum de la fonction objective et peut même conduire à une solution infaisable, comme il est indiqué dans Friedrich et coll. (2015, section 1, page 3). Puisque l'infaisabilité ne peut en fait résulter que de la violation de la contrainte (1.4), elle peut être facilement évitée en utilisant la méthode d'arrondi de Cont et Heidari (2014) qui préserve la somme entière des nombres positifs. En outre, toutes les expériences numériques que nous avons réalisées montrent que les valeurs de la fonction objective obtenues pour la répartition optimale non entière avant et après l'arrondi et pour la répartition optimale entière sont pratiquement indiscernables. Par exemple, pour les deux populations utilisées dans la section 5, les ratios $V/V_{\text{int}} \in [0{,}999759;\ 1]$, tandis que $V_{\text{ronde}}/V_{\text{int}} = 1$ (jusqu'à 6 décimales), pour différentes tailles totales d'échantillon $n = 0{,}1\,N, \ldots, 0{,}9\,N$. Ici, V, V_{int} et V_{ronde} désignent les variances (1.1) calculées pour les répartitions optimales non entières, les répartitions optimales entières et les répartitions optimales non entières arrondies (avec la méthode d'arrondi de Cont et Heidari, 2014), respectivement.

Les observations ci-dessus laissent entendre que les algorithmes de répartition rapides, pas nécessairement à valeur entière, ayant des résultats correctement arrondis, peuvent constituer une solution de rechange bonne et raisonnable aux algorithmes à valeur entière plus lents lorsque la vitesse d'un algorithme est cruciale.

Remerciements

Nous remercions le rédacteur en chef adjoint et les examinateurs d'avoir consacré leur temps et leurs efforts à la révision de la première et de la deuxième version du présent document. Leurs commentaires nous ont beaucoup aidés à préparer la présente version du document. En particulier, nous apprécions beaucoup toutes les remarques et suggestions relatives à l'algorithme *fixed-point iteration* (FPIA).

Annexe

A. Preuve du théorème 3.1

Remarque A.1. *Le problème 1.1 est un problème d'optimisation convexe, car sa fonction objective est* $f\colon \mathbb{R}_+^{|\mathcal{H}|} \to \mathbb{R}_+$,

$$f(\mathbf{x}) = \sum_{h\in\mathcal{H}} \frac{A_h^2}{x_h}, \tag{A.1}$$

Et les fonctions de contrainte d'inégalité $g_h^m\colon \mathbb{R}_+^{|\mathcal{H}|} \to \mathbb{R}$, $g_h^M\colon \mathbb{R}_+^{|\mathcal{H}|} \to \mathbb{R}$,

$$g_h^m(\mathbf{x}) = m_h - x_h, \qquad h \in \mathcal{H}, \tag{A.2}$$

$$g_h^M(\mathbf{x}) = x_h - M_h, \qquad h \in \mathcal{H}, \tag{A.3}$$

sont des fonctions convexes, tandis que la fonction de contrainte d'égalité $w\colon \mathbb{R}_+^{|\mathcal{H}|} \to \mathbb{R}$,

$$w(\mathbf{x}) = \sum_{h\in\mathcal{H}} x_h - n$$

est affine. Plus précisément, le problème 1.1 est un problème d'optimisation convexe d'un type particulier dans lequel les fonctions de contrainte d'inégalité (A.2)-(A.3) sont affines. Voir l'annexe D pour la définition du problème d'optimisation convexe.

Preuve du théorème 3.1. Nous commençons par prouver que le problème 1.1 a une solution unique. Le problème d'optimisation 1.1 est réalisable puisque les exigences $m_h < M_h, h \in \mathcal{H}$ et $\sum_{h\in\mathcal{H}} m_h \le n \le \sum_{h\in\mathcal{H}} M_h$ garantissent que l'ensemble réalisable $F := \left\{\mathbf{x} \in \mathbb{R}_+^{|\mathcal{H}|}\colon (1.4) - (1.5) \text{ sont tous satisfaits}\right\}$ est non vide. La fonction objective (1.3) atteint son minimum sur F puisqu'il s'agit d'une fonction continue et que F est fermé et borné. Enfin, l'unicité de la solution est attribuable à la convexité stricte de la fonction objective sur F.

Comme il est expliqué dans la remarque A.1, le problème 1.1 est un problème d'optimisation convexe dans lequel les fonctions de contrainte d'inégalité $g_h^m, g_h^M, h \in \mathcal{H}$ sont affines. La solution optimale d'un tel problème peut être déterrminée grâce aux conditions de Karush-Kuhn-Tucker (KKT), auquel cas elles sont non seulement nécessaires mais aussi suffisantes; pour consulter plus de références, voir l'annexe D.

Les gradients de la fonction objective (A.1) et des fonctions de contrainte (A.2)-(A.3) sont les suivants :

$$\nabla f(\mathbf{x}) = \left(-\tfrac{A_h^2}{x_h^2}, h \in \mathcal{H}\right), \quad \nabla w(\mathbf{x}) = \underline{1}, \quad \nabla g_h^m(\mathbf{x}) = -\nabla g_h^M(\mathbf{x}) = -\underline{1}_h, \quad h \in \mathcal{H},$$

où $\underline{1}$ est un vecteur dont toutes les entrées sont à 1 et $\underline{1}_h$ est un vecteur dont toutes les entrées sont à 0, à l'exception de l'entrée portant l'étiquette h, qui est à 1. Par conséquent, les conditions de KKT (D.2) pour le problème 1.1 prennent la forme suivante :

$$-\tfrac{A_h^2}{\left(x_h^*\right)^2} + \lambda - \mu_h^m + \mu_h^M = 0, \qquad h \in \mathcal{H}, \tag{A.4}$$

$$\sum_{h \in \mathcal{H}} x_h^* - n = 0, \tag{A.5}$$

$$m_h \le x_h^* \le M_h, \qquad h \in \mathcal{H}, \tag{A.6}$$

$$\mu_h^m \left(m_h - x_h^*\right) = 0, \qquad h \in \mathcal{H}, \tag{A.7}$$

$$\mu_h^M \left(x_h^* - M_h\right) = 0, \qquad h \in \mathcal{H}. \tag{A.8}$$

Pour prouver le théorème 3.1, il suffit de montrer que pour $\mathbf{x}^* = \mathbf{x}^{(\mathcal{L}^*, \mathcal{U}^*)}$ ayant $\mathcal{L}^*, \mathcal{U}^*$ satisfaisant aux conditions du CAS I ou de CAS II, il existe $\lambda \in \mathbb{R}$ et $\mu_h^m, \mu_h^M \ge 0, h \in \mathcal{H},$ de telle sorte que (A.4)-(A.8) sont valables. Il convient également de mentionner que l'exigence $m_h < M_h, h \in \mathcal{H},$ garantit que $\mathcal{L}^*$ et $\mathcal{U}^*$ définis dans (3.3) et (3.4) sont disjoints. Par conséquent, $\mathbf{x}^{(\mathcal{L}^*, \mathcal{U}^*)}$ est bien défini selon la définition 3.2.

CAS I : Prenons $\mathbf{x}^* = \mathbf{x}^{(\mathcal{L}^*, \mathcal{U}^*)}$ avec $\mathcal{L}^*$ et $\mathcal{U}^*$ comme dans (3.3). Alors, (A.5) est clairement satisfait après référence à (3.2) et (3.1), tandis que (A.6) découle directement de (3.2) et (3.3), puisque (3.3) pour $h \in \mathcal{H} \setminus \left(\mathcal{L}^* \cup \mathcal{U}^*\right)$ suppose en particulier $m_h < A_h s\left(\mathcal{L}^*, \mathcal{U}^*\right) < M_h$. Prenons $\lambda = \frac{1}{s^2(\mathcal{L}^*, \mathcal{U}^*)}$ et

$$\mu_h^m = \begin{cases} \lambda - \frac{A_h^2}{m_h^2}, & h \in \mathcal{L}^* \\ 0, & h \in \mathcal{H} \setminus \mathcal{L}^*, \end{cases} \qquad \mu_h^M = \begin{cases} \frac{A_h^2}{M_h^2} - \lambda, & h \in \mathcal{U}^* \\ 0, & h \in \mathcal{H} \setminus \mathcal{U}^*. \end{cases} \tag{A.9}$$

Notons que (3.3) ainsi que l'exigence $n \ge \sum_{h \in \mathcal{H}} m_h$ $\left(\text{cette dernière étant nécessaire si } \mathcal{U}^* = \varnothing\right)$ garantissent $s\left(\mathcal{L}^*, \mathcal{U}^*\right) > 0,$ tandis que (3.3) seul suppose $\mu_h^m, \mu_h^M \ge 0, h \in \mathcal{H}.$ Après avoir fait référence à (3.2), c'est une simple question d'algèbre de vérifier (A.4), (A.7) et (A.8) pour $\lambda, \mu_h^m, \mu_h^M, h \in \mathcal{H}$ défini ci-dessus.

CAS II : Prenons $\mathbf{x}^* = \mathbf{x}^{(\mathcal{L}^*, \mathcal{U}^*)}$ avec $\mathcal{L}^*, \mathcal{U}^*$ satisfaisant a (3.4) et à (3.5). La condition (A.5) devient alors (3.5), tandis que (A.6) est satisfaite trivialement en raison de (3.2). Supposons que $\mathcal{L}^* \ne \varnothing$

et $\mathcal{U}^* \neq \varnothing$ (pour $\mathcal{L}^*$ ou $\mathcal{U}^*$ vides, les conditions (A.4), (A.7) et (A.8) sont t satisfaites trivialement). Prenons $\tilde{s} > 0$ de façon arbitraire de telle sorte que

$$\tilde{s} \in \left[\max_{h \in \mathcal{U}^*} \tfrac{M_h}{A_h}, \min_{h \in \mathcal{L}^*} \tfrac{m_h}{A_h} \right]. \tag{A.10}$$

Notons que (3.4) garantit que l'intervalle ci-dessus est bien défini. Soit $\lambda = \frac{1}{\tilde{s}^2}$ et $\mu_h^m, \mu_h^M, h \in \mathcal{H}$ comme dans (A.9). Notons que (A.10) garantit que $\mu_h^m, \mu_h^M \geq 0$ pour tout $h \in \mathcal{H}$. Il est alors facile de vérifier, de la même manière que dans le CAS I, que les conditions (A.4), (A.7) et (A.8) sont satisfaits.

B. Lemmes auxiliaires et preuve du théorème 4.1

B.1 Notation

Dans l'ensemble de l'annexe B, nous désignons par $\mathcal{U}_r, \mathcal{L}_r, \tilde{\mathcal{L}}_r,$ les ensembles $\mathcal{U}, \mathcal{L}, \tilde{\mathcal{L}},$ respectivement, tels qu'ils se présentent dans l'itération r^{e} de l'algorithme *RNABOX* après l'étape 3 et avant l'étape 4. L'indice d'itération r prend des valeurs dans l'ensemble $\{1, \ldots, r^*\},$ où $r^* \geq 1$ indique l'itération finale de l'algorithme. Sous cette notation, nous avons $\mathcal{L}_1 = \varnothing$ et en général, pour les itérations suivantes, s'il y en a $\left(\text{c'est-à-dire si } r^* \geq 2\right),$ nous obtenons

$$\mathcal{L}_r = \mathcal{L}_{r-1} \cup \tilde{\mathcal{L}}_{r-1} = \bigcup_{i=1}^{r-1} \tilde{\mathcal{L}}_i, \qquad r = 2, \ldots, r^*. \tag{B.1}$$

Au cours de l'itération de *RNABOX*, les objets désignés par les symboles n et $\mathcal{H}$ sont modifiés. Toutefois, dans la présente annexe B, chaque fois que nous nous référons à n et $\mathcal{H},$ nous indiquons toujours la taille totale de l'échantillon non modifié et l'ensemble des étiquettes de strates comme dans l'entrée du *RNABOX*. En particulier, cette fonction est également liée à la fonction d'ensemble s (définie dans (3.1)) qui dépend de n et $\mathcal{H}$.

Pour faciliter la notation, pour tout $\mathcal{A} \subseteq \mathcal{H}$ et tout ensemble de nombres réels $z_h, h \in \mathcal{A},$ nous notons

$$z_{\mathcal{A}} = \sum_{h \in \mathcal{A}} z_h.$$

B.2 Remarques auxiliaires et lemmes

Nous commençons par un lemme décrivant d'importantes propriétés de monotonicité des fonctions s.

Lemme B.1. *Soit $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{H}$ et $\mathcal{C} \subseteq \mathcal{D} \subseteq \mathcal{H}$.*

1. *Si $\mathcal{B} \cup \mathcal{D} \subsetneq \mathcal{H}$ et $\mathcal{B} \cap \mathcal{D} = \varnothing,$ alors*

$$s(\mathcal{A}, \mathcal{C}) \geq s(\mathcal{B}, \mathcal{D}) \quad \Leftrightarrow \quad s(\mathcal{A}, \mathcal{C})\left(A_{\mathcal{B}\setminus\mathcal{A}} + A_{\mathcal{D}\setminus\mathcal{C}}\right) \leq m_{\mathcal{B}\setminus\mathcal{A}} + M_{\mathcal{D}\setminus\mathcal{C}}. \tag{B.2}$$

2. *Si* $\mathcal{A} \cup \mathcal{D} \subsetneq \mathcal{H}$, $\mathcal{A} \cap \mathcal{D} = \varnothing$, $\mathcal{B} \cup \mathcal{C} \subsetneq \mathcal{H}$, $\mathcal{B} \cap \mathcal{C} = \varnothing$, *alors*

$$s(\mathcal{A}, \mathcal{D}) \geq s(\mathcal{B}, \mathcal{C}) \quad \Leftrightarrow \quad s(\mathcal{A}, \mathcal{D})\left(A_{\mathcal{B}\setminus\mathcal{A}} - A_{\mathcal{D}\setminus\mathcal{C}}\right) \leq m_{\mathcal{B}\setminus\mathcal{A}} - M_{\mathcal{D}\setminus\mathcal{C}}. \tag{B.3}$$

Preuve. Il est clair que pour tout $\alpha \in \mathbb{R}, \beta \in \mathbb{R}, \delta \in \mathbb{R}, \gamma > 0, \gamma + \delta > 0,$ nous avons

$$\frac{\alpha+\beta}{\gamma+\delta} \geq \frac{\alpha}{\gamma} \quad \Leftrightarrow \quad \frac{\alpha+\beta}{\gamma+\delta}\delta \leq \beta. \tag{B.4}$$

Pour prouver (B.2), prenons

$$\alpha = n - m_{\mathcal{B}} - M_{\mathcal{D}} \qquad \beta = m_{\mathcal{B}\setminus\mathcal{A}} + M_{\mathcal{D}\setminus\mathcal{C}}$$
$$\gamma = A_{\mathcal{H}} - A_{\mathcal{B}\cup\mathcal{D}} \qquad \delta = A_{\mathcal{B}\setminus\mathcal{A}} + A_{\mathcal{D}\setminus\mathcal{C}}.$$

Alors, $\frac{\alpha}{\gamma} = s(\mathcal{B}, \mathcal{D}), \frac{\alpha+\beta}{\gamma+\delta} = s(\mathcal{A}, \mathcal{C}),$ et donc (B.2) est une conséquence immédiate de (B.4).
De même, pour (B.3), prenons

$$\alpha = n - m_{\mathcal{B}} - M_{\mathcal{C}} \qquad \beta = m_{\mathcal{B}\setminus\mathcal{A}} - M_{\mathcal{D}\setminus\mathcal{C}}$$
$$\gamma = A_{\mathcal{H}} - A_{\mathcal{B}\cup\mathcal{C}} \qquad \delta = A_{\mathcal{B}\setminus\mathcal{A}} - A_{\mathcal{D}\setminus\mathcal{C}},$$

et notons que $\gamma + \delta = A_{\mathcal{H}} - A_{\mathcal{B}\cup\mathcal{C}} + A_{\mathcal{B}\setminus\mathcal{A}} - A_{\mathcal{D}\setminus\mathcal{C}} = A_{\mathcal{H}} - A_{\mathcal{B}} - A_{\mathcal{C}} + A_{\mathcal{B}} - A_{\mathcal{A}} - A_{\mathcal{D}} + A_{\mathcal{C}} = A_{\mathcal{H}} - A_{\mathcal{A}\cup\mathcal{D}} > 0$ en raison des hypothèses faites pour $\mathcal{A}, \mathcal{D}, \mathcal{B}, \mathcal{C},$ et $A_h > 0, h \in \mathcal{H}.$ Alors, $\frac{\alpha}{\gamma} = s(\mathcal{B}, \mathcal{C}), \frac{\alpha+\beta}{\gamma+\delta} = s(\mathcal{A}, \mathcal{D}),$ et donc (B.3) est une conséquence immédiate de (B.4).

La remarque ci-dessous décrit certaines relations entre les ensembles $\mathcal{L}_r$ et $\mathcal{U}_r, r = 1, \ldots, r^* \geq 1,$ apparaissant dans l'algorithme *RNABOX*. Ces relations sont particulièrement importantes pour comprendre les calculs faisant intervenir la fonction d'ensemble s (rappelons qu'elle n'est définie que pour deux ensembles disjoints dont l'union est un sous-ensemble propre de $\mathcal{H}$).

Remarque B.1. *Pour* $r^* \geq 1,$

$$\mathcal{L}_r \cap \mathcal{U}_r = \varnothing, \qquad r = 1, \ldots, r^*, \tag{B.5}$$

et pour $r^* \geq 2,$

$$\mathcal{L}_r \cup \mathcal{U}_r \subsetneq \mathcal{H}, \qquad r = 1, \ldots, r^* - 1. \tag{B.6}$$

En outre, la valeur de $\mathbf{x}^*$ *correspond à l'étape 5 de l'algorithme* RNABOX. *Ensuite, pour* $r^* \geq 1,$

$$\mathcal{L}_{r^*} \cup \mathcal{U}_{r^*} \subsetneq \mathcal{H} \quad \Leftrightarrow \quad \mathbf{x}^* \textit{ est une répartition} \text{ régulière}, \tag{B.7}$$

et

$$\mathcal{L}_{r^*} \cup \mathcal{U}_{r^*} = \mathcal{H} \quad \Leftrightarrow \quad \mathbf{x}^* \textit{ est une répartition de} \text{ sommet}. \tag{B.8}$$

Preuve. D'après la définition de l'ensemble $\mathcal{U}$ à l'étape 2 de *RNABOX*, pour $r^* \geq 1,$

$$\mathcal{U}_r \subseteq \mathcal{H} \setminus \mathcal{L}_r, \qquad r = 1, \ldots, r^*, \tag{B.9}$$

ce qui prouve (B.5). Suivant (B.1), pour $r^* \geq 2$,

$$\mathcal{L}_r = \bigcup_{i=1}^{r-1} \tilde{\mathcal{L}}_i \subseteq \mathcal{H}, \qquad r = 2, \ldots, r^*, \tag{B.10}$$

où l'inclusion est attribuable à la définition de l'ensemble $\tilde{\mathcal{L}}$ à l'étape 3 de *RNABOX*, c'est-à-dire $\tilde{\mathcal{L}}_r \subseteq \mathcal{H} \setminus (\mathcal{L}_r \cup \mathcal{U}_r)$ pour $r = 1, \ldots, r^*$. Les inclusions (B.9), (B.10) avec $\mathcal{L}_1 = \varnothing$ supposent

$$\mathcal{L}_r \cup \mathcal{U}_r \subseteq \mathcal{H}, \qquad r = 1, \ldots, r^* \geq 1. \tag{B.11}$$

Étant donné que $r^* \geq 2$, l'étape 4 de l'algorithme garantit que l'ensemble $\tilde{\mathcal{L}}_r \subseteq \mathcal{H} \setminus (\mathcal{L}_r \cup \mathcal{U}_r)$ est non vide pour $r = 1, \ldots, r^* - 1$, ce qui suppose $\mathcal{L}_r \cup \mathcal{U}_r \neq \mathcal{H}$. Ce fait combiné à (B.11) donne (B.6). Les équivalences (B.7) et (B.8) sont triviales nous nous référons à la définition 1.1 des répartitions *régulières* et des répartitions de *sommet*.

Les deux remarques suivantes résument certains faits importants découlant de l'étape 2 de l'algorithme *RNABOX*. Ces faits serviront de point de départ à la plupart des preuves présentées dans cette section.

Remarque B.2. *A chaque itération* $r = 1, \ldots, r^* \geq 1$, *de l'algorithme* RNABOX, *un vecteur* $\left(x_h^{**}, h \in \mathcal{H} \setminus \mathcal{L}_r\right)$ *obtenu à l'étape 2, possède des éléments de la forme*

$$x_h^{**} = \begin{cases} M_h, & h \in \mathcal{U}_r \subseteq \mathcal{H} \setminus \mathcal{L}_r \\ A_h s(\mathcal{L}_r, \mathcal{U}_r) < M_h, & h \in \mathcal{H} \setminus (\mathcal{L}_r \cup \mathcal{U}_r), \end{cases} \tag{B.12}$$

où la fonction d'ensemble s *est définie dans (3.1). L'équation (B.12) est une conséquence directe du théorème C.1.*

Remarque B.3. *La remarque B.2 ainsi que le théorème C.1, pour* $r^* \geq 2$, *donnent*

$$\mathcal{U}_r = \left\{h \in \mathcal{H} \setminus \mathcal{L}_r : \; A_h s(\mathcal{L}_r, \mathcal{U}_r) \geq M_h\right\}, \quad r = 1, \ldots, r^* - 1, \tag{B.13}$$

tandis que pour $r^* \geq 1$,

$$\mathcal{U}_{r^*} = \left\{h \in \mathcal{H} \setminus \mathcal{L}_{r^*} : \; A_h s\left(\mathcal{L}_{r^*}, \mathcal{U}_{r^*}\right) \geq M_h\right\}, \tag{B.14}$$

si et seulement si $\mathbf{x}^*$ *(calculé à l'étape 5 de l'algorithme* RNABOX*) est une répartition* régulière *ou une répartition de* sommet *avec* $\mathcal{L}_{r^*} = \mathcal{H}$.

De plus, pour $r^* \geq 1$,

$$\tilde{\mathcal{L}}_r = \left\{h \in \mathcal{H} \setminus (\mathcal{L}_r \cup \mathcal{U}_r) : \; A_h s(\mathcal{L}_r, \mathcal{U}_r) \leq m_h\right\}, \quad r = 1, \ldots, r^*. \tag{B.15}$$

Notons que dans la remarque B.3, la fonction s est bien définie grâce à la remarque B.1. La nécessité de limiter le champ d'application de (B.14) aux seules répartitions *régulières* est dictée par le fait que dans

le cas d'une répartition de *sommet*, nous avons $\mathcal{L}_{r^*} \cup \mathcal{U}_{r^*} = \mathcal{H}$ (voir (B.8)) et que, par conséquent, $s\left(\mathcal{L}_{r^*}, \mathcal{U}_{r^*}\right)$ n'est pas bien défini.

Le lemme B.2 et le lemme B.3 révèlent certaines propriétés de monotonicité de la séquence $\left(\mathcal{U}_r\right)_{r=1}^{r^*}$ et de la séquence $\left(s\left(\mathcal{L}_r, \mathcal{U}_r\right)\right)_{r=1}^{r^*}$, respectivement. Ces propriétés joueront un rôle crucial dans la démonstration du théorème 4.1.

Lemme B.2. *La séquence* $\left(\mathcal{U}_r\right)_{r=1}^{r^*}$ *est non croissante, c'est-à-dire que pour* $r^* \geq 2$,

$$\mathcal{U}_r \supseteq \mathcal{U}_{r+1}, \qquad r = 1, \ldots, r^* - 1. \tag{B.16}$$

Preuve. Soit $r^* \geq 2$ et $r = 1, \ldots, r^* - 1$. Alors, par (B.6), $\mathcal{L}_r \cup \mathcal{U}_r \subsetneq \mathcal{H}$. Suivant (B.13), le domaine de discours de $\mathcal{U}_r$ est $\mathcal{H} \setminus \mathcal{L}_r$, et en fait c'est $\mathcal{H} \setminus \left(\mathcal{L}_r \cup \tilde{\mathcal{L}}_r\right) = \mathcal{H} \setminus \mathcal{L}_{r+1}$, puisque $\mathcal{U}_r \not\subset \tilde{\mathcal{L}}_r$ est assuré par l'étape 3 de *RNABOX*. En d'autres termes, $\mathcal{U}_r$ et $\mathcal{U}_{r+1}$ ont essentiellement le même domaine de discours, qui est le suivant $\mathcal{H} \setminus \mathcal{L}_{r+1}$. Compte tenu de ce fait et de la forme du prédicat de construction d'ensemble dans (B.13)-(B.14) ainsi que de l'égalité $\mathcal{U}_{r^*} = \mathcal{H} \setminus \mathcal{L}_{r^*}$ pour le cas où $\mathbf{x}^*$ est une répartition de *sommet* (pour laquelle (B.14) ne s'applique pas), nous concluons que seul l'un des deux cas distincts suivants est possible : $\mathcal{U}_r \supseteq \mathcal{U}_{r+1}$ ou $\mathcal{U}_r \subsetneq \mathcal{U}_{r+1}$.

La preuve se fait par contradiction, c'est-à-dire qu'il faut supposer que l'équation (B.16) n'est pas vérifiée. Par conséquent, compte tenu de l'observation ci-dessus, il existe $r \in \left\{1, \ldots, r^* - 1\right\}$ de telle sorte que $\mathcal{U}_r \subsetneq \mathcal{U}_{r+1}$. Alors,

$$\varnothing \neq \left(\mathcal{U}_{r+1} \setminus \mathcal{U}_r\right) \subsetneq \mathcal{H} \setminus \left(\mathcal{L}_r \cup \mathcal{U}_r\right), \tag{B.17}$$

et donc, en raison de (B.12),

$$A_h s\left(\mathcal{L}_r, \mathcal{U}_r\right) < M_h, \qquad h \in \mathcal{U}_{r+1} \setminus \mathcal{U}_r. \tag{B.18}$$

D'autre part, d'après (B.15),

$$A_h s\left(\mathcal{L}_r, \mathcal{U}_r\right) \leq m_h, \qquad h \in \tilde{\mathcal{L}}_r. \tag{B.19}$$

En additionnant latéralement : (B.18) pour $h \in \mathcal{U}_{r+1} \setminus \mathcal{U}_r$, (B.19) pour $h \in \tilde{\mathcal{L}}_r$, puis tous ensemble, nous obtenons

$$s\left(\mathcal{L}_r, \mathcal{U}_r\right)\left(A_{\tilde{\mathcal{L}}_r} + A_{\mathcal{U}_{r+1} \setminus \mathcal{U}_r}\right) < m_{\tilde{\mathcal{L}}_r} + M_{\mathcal{U}_{r+1} \setminus \mathcal{U}_r}. \tag{B.20}$$

Le vecteur $\mathbf{x}^*$ *est une répartition* régulière : dans ce cas, en suivant la remarque B.1, nous voyons que l'inégalité (B.20) est le côté droit de l'équivalence (B.2) avec

$$\mathcal{A} = \mathcal{L}_r \subseteq \left(\mathcal{L}_r \cup \tilde{\mathcal{L}}_r\right) = \mathcal{L}_{r+1} = \mathcal{B} \subsetneq \mathcal{H}, \tag{B.21}$$

$$\mathcal{C} = \mathcal{U}_r \subsetneq \mathcal{U}_{r+1} = \mathcal{D} \subsetneq \mathcal{H}. \tag{B.22}$$

Alors, suivant le lemme B.1, l'inégalité (B.20) est équivalente à

$$s\left(\mathcal{L}_r, \mathcal{U}_r\right) > s\left(\mathcal{L}_{r+1}, \mathcal{U}_{r+1}\right). \tag{B.23}$$

Si nous combinons

$$s\left(\mathcal{L}_{r+1}, \mathcal{U}_{r+1}\right) \geq \tfrac{M_h}{A_h}, \qquad h \in \mathcal{U}_{r+1}, \tag{B.24}$$

(cela découle de (B.13)-(B.14)) avec les inégalités (B.23) et (B.18), nous obtenons la contradiction suivante :

$$\tfrac{M_h}{A_h} > s\left(\mathcal{L}_r, \mathcal{U}_r\right) > s\left(\mathcal{L}_{r+1}, \mathcal{U}_{r+1}\right) \geq \tfrac{M_h}{A_h}, \quad h \in \mathcal{U}_{r+1} \setminus \mathcal{U}_r. \tag{B.25}$$

Par conséquent, (B.16) est vrai, étant donné que $\mathcal{L}_{r^*} \cup \mathcal{U}_{r^*} \subsetneq \mathcal{H}$.

Le vecteur $\mathbf{x}^*$ *est une* répartition de sommet : Puisque $\mathcal{L}_{r+1} \cup \mathcal{U}_{r+1} \subsetneq \mathcal{H}$ pour $r = 1, \ldots, r^* - 2,$ la preuve de (B.16) pour un tel r est identique à la preuve pour le cas d'une répartition *régulière*. Il suffit donc de démontrer que (B.16) est valable pour $r = r^* - 1.$ Pour ce faire, nous exploiterons l'inégalité (B.20) qui, compte tenu de la définition 3.1 de la fonction d'ensemble $s,$ prend la forme suivante pour $r = r^* - 1,$

$$\tfrac{n - m_{\mathcal{L}_{r^*-1}} - M_{\mathcal{U}_{r^*-1}}}{A_{\mathcal{H}} - A_{\mathcal{L}_{r^*-1} \cup \mathcal{U}_{r^*-1}}} \left(A_{\tilde{\mathcal{L}}_{r^*-1}} + A_{\mathcal{U}_{r^*} \setminus \mathcal{U}_{r^*-1}} \right) < m_{\tilde{\mathcal{L}}_{r^*-1}} + M_{\mathcal{U}_{r^*} \setminus \mathcal{U}_{r^*-1}}. \tag{B.26}$$

Puisque $A_{\mathcal{H}} - A_{\mathcal{L}_{r^*-1} \cup \mathcal{U}_{r^*-1}} = A_{\tilde{\mathcal{L}}_{r^*-1}} + A_{\mathcal{U}_{r^*} \setminus \mathcal{U}_{r^*-1}},$ pour $\mathcal{L}_{r^*} \cup \mathcal{U}_{r^*} = \mathcal{H},$ l'inégalité (B.26) se simplifie en

$$n < m_{\tilde{\mathcal{L}}_{r^*-1}} + m_{\mathcal{L}_{r^*-1}} + M_{\mathcal{U}_{r^*} \setminus \mathcal{U}_{r^*-1}} + M_{\mathcal{U}_{r^*-1}} = m_{\mathcal{L}_{r^*}} + M_{\mathcal{U}_{r^*}} = n, \tag{B.27}$$

ce qui est une contradiction. Notons que la dernière égalité découle de l'étape 2 du *RNABOX* après référence à (C.3) et en s'appuyant sur le fait que $\mathcal{U}_{r^*} = \mathcal{H} \setminus \mathcal{L}_{r^*}$ pour une répartition de *sommet*. Par conséquent, (B.16) est également vrai pour $\mathcal{L}_{r^*} \cup \mathcal{U}_{r^*} = \mathcal{H}.$

Lemme B.3. *Soit* $r^* \geq 3.$ *Alors*

$$s(\mathcal{L}_r, \mathcal{U}_r) \geq s\left(\mathcal{L}_{r+1}, \mathcal{U}_{r+1}\right), \quad r = 1, \ldots, r^* - 2. \tag{B.28}$$

De plus, si $\mathbf{x}^*$ *(calculé à l'étape 5 de l'algorithme* RNABOX*) est une répartition* régulière *et* $r^* \geq 2,$ *alors*

$$s\left(\mathcal{L}_{r^*-1}, \mathcal{U}_{r^*-1}\right) \geq s\left(\mathcal{L}_{r^*}, \mathcal{U}_{r^*}\right). \tag{B.29}$$

Preuve. Nous commençons par prouver (B.28). Soit $r^* \geq 3$ et $r = 1, \ldots, r^* - 2.$ En suivant le lemme B.2 et en utilisant (B.13),

$$A_h s(\mathcal{L}_r, \mathcal{U}_r) \geq M_h, \qquad h \in \mathcal{U}_r \setminus \mathcal{U}_{r+1}. \tag{B.30}$$

D'autre part, d'après (B.15),

$$A_h s(\mathcal{L}_r, \mathcal{U}_r) \leq m_h, \qquad h \in \tilde{\mathcal{L}}_r. \tag{B.31}$$

En multipliant les deux côtés de l'inégalité (B.30) par $-1,$ en l'additionnant latéralement pour $h \in \mathcal{U}_r \setminus \mathcal{U}_{r+1}$ et en l'ajoutant à (B.31), dont la somme est antérieurement calculé latéralement pour $h \in \tilde{\mathcal{L}}_r,$ nous obtenons

$$s(\mathcal{L}_r, \mathcal{U}_r)\left(A_{\tilde{\mathcal{L}}_r} - A_{\mathcal{U}_r \setminus \mathcal{U}_{r+1}}\right) \leq m_{\tilde{\mathcal{L}}_r} - M_{\mathcal{U}_r \setminus \mathcal{U}_{r+1}}. \tag{B.32}$$

La relation (B.32) est la deuxième inégalité de (B.3) ayant

$$\mathcal{A} = \mathcal{L}_r \subsetneq \left(\mathcal{L}_r \cup \tilde{\mathcal{L}}_r\right) = \mathcal{L}_{r+1} = \mathcal{B} \subsetneq \mathcal{H}, \tag{B.33}$$

$$\mathcal{C} = \mathcal{U}_{r+1} \subseteq \mathcal{U}_r = \mathcal{D} \subsetneq \mathcal{H}. \tag{B.34}$$

Sur la base de la remarque B.1, nous voyons que $\mathcal{A} \cup \mathcal{D} \subsetneq \mathcal{H},$ $\mathcal{A} \cap \mathcal{D} = \varnothing,$ et $\mathcal{B} \cup \mathcal{C} \subsetneq \mathcal{H},$ $\mathcal{B} \cap \mathcal{C} = \varnothing,$ et donc la première inégalité dans (B.3) s'ensuit, c'est-à-dire

$$s(\mathcal{L}_r, \mathcal{U}_r) \geq s(\mathcal{L}_{r+1}, \mathcal{U}_{r+1}). \tag{B.35}$$

Par conséquent, (B.28) est prouvé.

Si $\mathbf{x}^*$ est une répartition *régulière*, compte tenu de la remarque B.1, le même raisonnement conduisant à l'inégalité (B.35) reste clairement valable pour $r = r^* - 1,\ r^* \geq 2.$

B.3 Preuve du théorème 4.1

Pour prouver le théorème 4.1, il faut démontrer que :

I) l'algorithme se termine en un nombre fini d'itérations, c'est-à-dire $r^* < \infty,$

II) la solution calculée à r^* est optimale.

La preuve de la partie (I) est relativement simple. À chaque itération $r = 1, \ldots, r^* - 1,\ r^* \geq 2,$ l'ensemble des étiquettes de strates $\mathcal{H}$ est réduit en soustrayant $\tilde{\mathcal{L}}_r.$ Par conséquent, $r^* \leq |\mathcal{H}| + 1 < \infty,$ où $r^* = |\mathcal{H}| + 1$ si et seulement si $\left|\tilde{\mathcal{L}}_r\right| = 1, r = 1, \ldots, r^* - 1.$ En d'autres termes, l'algorithme s'arrête après un maximum de $|\mathcal{H}| + 1$ itérations.

Afin de prouver la partie (II), suite au théorème 3.1 et à la remarque 3.1, il suffit de démontrer que lorsque $\mathbf{x}^*$ (calculé à l'étape 5 de l'algorithme *RNABOX*) est une répartition *régulière*, pour tout $h \in \mathcal{H},$

$$h \in \mathcal{L}_{r^*} \quad \Leftrightarrow \quad s(\mathcal{L}_{r^*}, \mathcal{U}_{r^*}) \leq \tfrac{m_h}{A_h}, \tag{B.36}$$

$$h \in \mathcal{U}_{r^*} \quad \Leftrightarrow \quad s(\mathcal{L}_{r^*}, \mathcal{U}_{r^*}) \geq \tfrac{M_h}{A_h}, \tag{B.37}$$

Et lorsque $\mathbf{x}^*$ est une répartition de *sommet*

$$\max_{h\in\mathcal{U}_{r^*}} \tfrac{M_h}{A_h} \le \min_{h\in\mathcal{L}_{r^*}} \tfrac{m_h}{A_h}, \qquad \text{lorsque } \mathcal{U}_{r^*}\neq\varnothing \text{ et } \mathcal{L}_{r^*}\neq\varnothing, \tag{B.38}$$

$$m_{\mathcal{L}_{r^*}} + M_{\mathcal{U}_{r^*}} = n. \tag{B.39}$$

Le vecteur $\mathbf{x}^*$ *est une répartition* régulière : Notons que la remarque B.1 suppose que $s(\mathcal{L}_r, \mathcal{U}_r)$ est bien défini. Nous commençons par l'équivalence (B.36).

Nécessité : Pour $r^*=1,$ nous avons $\mathcal{L}_{r^*}=\varnothing$ et par conséquent, le côté droit de l'équivalence (B.36) est satisfait trivialement. Soit $r^*\ge 2,$ et $h\in\mathcal{L}_{r^*}=\bigcup_{r=1}^{r^*-1}\tilde{\mathcal{L}}_r.$ Ainsi, $h\in\tilde{\mathcal{L}}_r$ pour une certaine $r\in\{1,\ldots,$ $r^*-1\}$ et ensuite, en raison de (B.15), nous avons $s(\mathcal{L}_r,\mathcal{U}_r)\le\frac{m_h}{A_h}.$ Par conséquent, (B.28) et (B.29) donnent $s(\mathcal{L}_{r^*},\mathcal{U}_{r^*})\le\frac{m_h}{A_h}.$

Suffisance : Puisque $\tilde{\mathcal{L}}_{r^*}=\varnothing,$ (B.15) suppose

$$\left\{h\in\mathcal{H}\setminus(\mathcal{L}_{r^*}\cup\mathcal{U}_{r^*})\colon s(\mathcal{L}_{r^*},\mathcal{U}_{r^*})\le\tfrac{m_h}{A_h}\right\}=\varnothing, \quad r^*\ge 1. \tag{B.40}$$

D'autre part, (B.14) ainsi que $m_h<M_h, h\in\mathcal{H},$ donnent

$$s(\mathcal{L}_{r^*},\mathcal{U}_{r^*})\ge\tfrac{M_h}{A_h}>\tfrac{m_h}{A_h}, \qquad h\in\mathcal{U}_{r^*}, \tag{B.41}$$

Et donc (B.40) se lit

$$\left\{h\in\mathcal{H}\setminus\mathcal{L}_{r^*}\colon s(\mathcal{L}_{r^*},\mathcal{U}_{r^*})\le\tfrac{m_h}{A_h}\right\}=\varnothing, \qquad r^*\ge 1. \tag{B.42}$$

La preuve de la nécessité en (B.37) est immédiate au vu de (B.14), tandis que la suffisance s'ensuit par contradiction. En effet, prenons $r^*\ge 1.$ Supposons que le côté droit de l'équivalence (B.37) est vérifié et que $h\notin\mathcal{U}_{r^*}.$ Alors, compte tenu de la remarque B.1, soit $h\in\mathcal{H}\setminus(\mathcal{L}_{r^*}\cup\mathcal{U}_{r^*})$ et donc à partir de (B.12)

$$s(\mathcal{L}_{r^*},\mathcal{U}_{r^*})<\tfrac{M_h}{A_h}, \tag{B.43}$$

cela constitue une contradiction, ou soit $h\in\mathcal{L}_{r^*}$ puis à partir de (B.36), compte tenu de $m_h<M_h, h\in\mathcal{H},$

$$s(\mathcal{L}_{r^*},\mathcal{U}_{r^*})\le\tfrac{m_h}{A_h}<\tfrac{M_h}{A_h}, \tag{B.44}$$

cela constitue une contradiction.

Le vecteur $\mathbf{x}^*$ *est une répartition de* sommet : Pour $r^*=1,$ la seule possibilité est que $\mathcal{U}_{r^*}=\mathcal{H}, \mathcal{L}_{r^*}=\varnothing.$ Dans ce cas, (B.38) est clairement respecté, tandis que (B.39) découle de l'étape 2 de *RNABOX* après référence à (C.3). Soit $r^*\ge 2.$ Alors, par (B.13) nous avons

$$s(\mathcal{L}_{r^*-1},\mathcal{U}_{r^*-1})\ge\tfrac{M_h}{A_h}, \quad h\in\mathcal{U}_{r^*-1}\supseteq\mathcal{U}_{r^*}, \tag{B.45}$$

où l'inclusion de l'ensemble est attribuable au lemme B.2. D'autre part, d'après (B.15), nous obtenons

$$s\left(\mathcal{L}_{r^*-1}, \mathcal{U}_{r^*-1}\right) \leq \frac{m_h}{A_h}, \quad h \in \mathcal{L}_{r^*-1} \cup \tilde{\mathcal{L}}_{r^*-1} = \mathcal{L}_{r^*}, \tag{B.46}$$

où le fait que l'inégalité ci-dessus est satisfaite pour $h \in \mathcal{L}_{r^*-1}$ découle de (B.28). En comparant (B.45) et (B.46), nous voyons clairement que (B.38) est satisfait. Enfin, l'équation (B.39) est satisfaite du fait de

$$n - m_{\tilde{\mathcal{L}}_1} - \ldots - m_{\tilde{\mathcal{L}}_{r^*-1}} = n - m_{\mathcal{L}_{r^*}} = M_{\mathcal{U}_{r^*}}, \tag{B.47}$$

où la première égalité découle de (B.1) tandis que la seconde découle de l'étape 2 du *RNABOX* après référence à (C.3) et en considérant le fait que $\mathcal{U}_{r^*} = \mathcal{H} \setminus \mathcal{L}_{r^*}$ pour une répartition de *sommet*.

C. Conditions d'optimalité pour le problème 2.1

Le théorème C.1 suivant fournit des conditions nécessaires et suffisantes pour la solution optimale du problème 2.1. Il a été donné à l'origine comme le théorème 1.1 dans Wesołowski et coll. (2022) et il est crucial pour la preuve du théorème 4.1. Nous le citerons ici sous une forme légèrement élargie afin qu'il couvre également le cas de $\mathcal{U}^* = \mathcal{H}$. Comme d'habitude, la fonction d'ensemble s est définie comme dans la définition 3.1. L'algorithme qui résout le problème 2.1 est le *RNA* et il est présenté dans la section 4 du présent document.

Théorème C.1. *Le problème d'optimisation 2.1 a une solution optimale unique. Le point* $\mathbf{x}^* = (x_h^*, h \in \mathcal{H}) \in \mathbb{R}_+^{|\mathcal{H}|}$ *est une solution au problème d'optimisation 2.1 si et seulement si* $\mathbf{x}^*$ *a des entrées de la forme*

$$x_h^* = \begin{cases} M_h, & h \in \mathcal{U}^* \\ A_h s\left(\varnothing, \mathcal{U}^*\right), & h \in \mathcal{H} \setminus \mathcal{U}^*, \end{cases} \tag{C.1}$$

avec $\mathcal{U}^* \subseteq \mathcal{H}$, *de sorte que l'un des deux cas suivants se vérifie :*

CAS I : $\mathcal{U}^* \subsetneq \mathcal{H}$ *et*

$$\mathcal{U}^* = \left\{h \in \mathcal{H}: \; A_h s\left(\varnothing, \mathcal{U}^*\right) \geq M_h\right\}. \tag{C.2}$$

CAS II : $\mathcal{U}^* = \mathcal{H}$ *et*

$$n = \sum_{h \in \mathcal{H}} M_h. \tag{C.3}$$

D. Schéma d'optimisation convexe et conditions de Karush-Kuhn-Tucker

Un problème d'optimisation convexe est un problème d'optimisation dans lequel la fonction objective est une fonction convexe et l'ensemble réalisable est un ensemble convexe. Sous sa forme standard, elle s'écrit

$$\begin{aligned} \underset{\mathbf{x}\in\mathcal{D}}{\text{minimise}} \quad & f(\mathbf{x}) \\ \text{sous la contrainte} \quad & w_i(\mathbf{x}) = 0, \quad i = 1, \ldots, k \\ & g_j(\mathbf{x}) \leq 0, \quad j = 1, \ldots, \ell, \end{aligned} \tag{D.1}$$

où $\mathbf{x}$ est la variable d'optimisation $\mathcal{D} \subseteq \mathbb{R}^p$, $p \in \mathbb{N}_+$, la fonction objective $f\colon \mathcal{D}_f \subseteq \mathbb{R}^p \to \mathbb{R}$ et les fonctions de contrainte d'inégalité $g_j\colon \mathcal{D}_{g_j} \subseteq \mathbb{R}^p \to \mathbb{R}, j = 1, \ldots, \ell,$ sont convexes, tandis que les fonctions de contrainte d'égalité $w_i\colon \mathcal{D}_{w_i} \subseteq \mathbb{R}^p \to \mathbb{R}, i = 1, \ldots, k,$ sont affines. Ici, $\mathcal{D} = \mathcal{D}_f \cap \bigcap_{i=1}^{k} \mathcal{D}_{w_i} \cap \bigcap_{j=1}^{\ell} \mathcal{D}_{g_j}$ désigne un domaine commun à toutes les fonctions. Le point $\mathbf{x} \in \mathcal{D}$ est appelé *faisable* s'il satisfait à toutes les contraintes, sinon le point est appelé *infaisable*. Un problème d'optimisation est appelé *faisable* s'il existe $\mathbf{x} \in \mathcal{D}$ qui est *faisable*, sinon le problème est appelé *infaisable*.

Dans le contexte du problème de répartition optimale 1.1 examiné dans le présent document, nous nous intéressons à un type particulier de problème convexe, à savoir (D.1) dans lequel toutes les fonctions de contrainte d'inégalité $g_j, j = 1, \ldots, \ell,$ sont affines. Il est bien connu, voir par exemple la monographie de Boyd et Vandenberghe (2004), que la solution d'un tel problème d'optimisation peut être dégagée au moyen d'un ensemble d'équations et d'inégalités connues sous le nom de conditions de Karush-Kuhn-Tucker (KKT), qui, dans ce cas, sont non seulement nécessaires mais aussi suffisantes.

Théorème D.1 (Conditions de KKT pour un problème d'optimisation convexe avec des contraintes d'inégalité affines)**.** *Un point* $\mathbf{x}^* \in \mathcal{D} \subseteq \mathbb{R}^p$, $p \in \mathbb{N}_+$, *est une solution au problème d'optimisation convexe (D.1) dans lequel les fonctions* $g_j, j = 1, \ldots, \ell,$ *sont affines si et seulement s'il existe des nombres* $\lambda_i \in \mathbb{R}$, $i = 1, \ldots, k,$ *et* $\mu_j \geq 0$, $j = 1, \ldots, \ell,$ *appelés multiplicateurs de KKT, de telle sorte que*

$$\begin{aligned} & \nabla f(\mathbf{x}^*) + \sum_{i=1}^{k} \lambda_i \nabla w_i(\mathbf{x}^*) + \sum_{j=1}^{\ell} \mu_j \nabla g_j(\mathbf{x}^*) = \mathbf{0} \\ & w_i(\mathbf{x}^*) = 0, \quad i = 1, \ldots, k \\ & g_j(\mathbf{x}^*) \leq 0, \quad j = 1, \ldots, \ell \\ & \mu_j g_j(\mathbf{x}^*) = 0, \quad j = 1, \ldots, \ell. \end{aligned} \tag{D.2}$$

Bibliographie

Baillargeon, S., et Rivest, L.-P. (2011). Élaboration de plans stratifiés en R à l'aide du programme *stratification*. *Techniques d'enquête*, 37, 1, 59-72. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2011001/article/11447-fra.pdf.

Barcaroli, G. (2014). `SamplingStrata`: An R Package for the Optimization of Stratified Sampling. *Journal of Statistical Software*, 61(4), 1-24. https://www.jstatsoft.org/index.php/jss/article/view/v061i04.

Boyd, S., et Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

Brito, J., Silva, P. et Veiga, T. (2017). `stratbr`: *Optimal Stratification in Stratified Sampling*. R package version 1.2. https://CRAN.R-project.org/package=stratbr.

Burgard, J.P., Münnich, R.T. (2012). Modelling over and undercounts for design-based Monte Carlo studies in small area estimation: An application to the German register-assisted census. *Computational Statistics & Data Analysis*, 56, 2856-2863. https://www.sciencedirect.com/science/article/pii/S0167947310004305.

Choudhry, G.H., Rao, J.N.K. et Hidiroglou, M.A. (2012). À propos de la répartition de l'échantillon pour une estimation sur domaine efficace. *Techniques d'enquête*, 38, 1, 25-32. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2012001/article/11682-fra.pdf.

Cont, R., et Heidari, M. (2014). Optimal rounding under integer constraints. https://arxiv.org/abs/1501.00014.

de Moura Brito, J.A., do Nascimento Silva, P.L., Silva Semaan, G. et Maculan, N. (2015). Application des formulations de la programmation en nombres entiers à la répartition optimale dans l'échantillonnage stratifié. *Techniques d'enquête*, 41, 2, 451-467. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2015002/article/14249-fra.pdf.

Friedrich, U., Münnich, R., de Vries, S. et Wagner, M. (2015). Fast integer-valued algorithms for optimal allocations under constraints in stratified sampling. *Computational Statistics & Data Analysis*, 92, 1-12. https://www.sciencedirect.com/science/article/pii/S0167947315001413.

Gabler, S., Ganninger, M., Münnich, R. (2012). Optimal allocation of the sample size to strata under box constraints. *Metrika*, 75(2), 151-161. DOI: https://doi.org/10.1007/s00184-010-0319-3.

Gunning, P., et Horgan, J.M. (2004). Un nouvel algorithme pour la construction de bornes de stratification dans les populations asymétriques. *Techniques d'enquête*, 30, 2, 177-185. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2004002/article/7749-fra.pdf.

Kadane, J.B. (2005). Optimal dynamic sample allocation among strata. *Journal of Official Statistics*, 21(4), 531-541. http://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/optimal-dynamic-sample-allocation-among-strata.pdf.

Khan, M.G., Nand, N. et Ahmad, N. (2008). Détermination des bornes optimales de strate au moyen de la programmation dynamique. *Techniques d'enquête*, 34, 2, 227-236. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2008002/article/10761-fra.pdf.

Lednicki, B., et Wieczorkowski, R. (2003). Optimal stratification and sample allocation between subpopulations and strata. *Statistics in Transition*, 6(2), 287-305. https://stat.gov.pl/download/gfx/portalinformacyjny/en/defaultstronaopisowa/3432/1/1/sit_volume_4-7.zip.

Mersmann, O. (2021). `microbenchmark`: *Accurate Timing Functions*. R package version 1.4.9. https://CRAN.R-project.org/package=microbenchmark.

Münnich, R.T., Sachs, E.W. et Wagner, M. (2012). Numerical solution of optimal allocation problems in stratified sampling under box constraints. *AStA Advances in Statistical Analysis*, 96(3), 435-450. DOI: https://doi.org/10.1007/s10182-011-0176-z.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4), 558-625.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienne, Autriche. https://www.R-project.org/.

Rivest, L.-P., et Baillargeon, S. (2022). `stratification`: *Univariate Stratification of Survey Populations*. R package version 2.2-7. https://CRAN.R-project.org/package=stratification.

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

Stenger, H., et Gabler, S. (2005). Combining random sampling and census strategies − Justification of inclusion probabilities equal to 1. *Metrika*, 61(2), 137-156. DOI: https://doi.org/10.1007/s001840400328.

Tschuprow, A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observation (Chapters 4-6). *Metron*, 2(4), 636-680.

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

Wesołowski, J., Wieczorkowski, R. et Wójciak, W. (2022). Optimality of the Recursive Neyman Allocation. *Journal of Survey Statistics and Methodology*, 10(5), 1263-1275. https://academic.oup.com/jssam/article-pdf/10/5/1263/46878255/smab018.pdf.

Wieczorkowski, R., Wesołowski, J. et Wójciak, W. (2023). Numerical Performance of the *RNABOX* Algorithm. https://github.com/rwieczor/recursive_Neyman_rnabox.

Wright, T. (2017). Exact optimal sample allocation: More efficient than Neyman. *Statistics & Probability Letters*, 129, 50-57. https://www.sciencedirect.com/science/article/pii/S0167715217301657.

Wright, T. (2020). A general exact optimal sample allocation algorithm: With bounded cost and bounded sample sizes. *Statistics & Probability Letters*, 165, 108829. https://www.sciencedirect.com/science/article/pii/S0167715220301322.

Wright, S.E., Noble, R. et Bailer, A.J. (2007). Equal-precision allocations and other constraints in stratified random sampling. *Journal of Statistical Computation and Simulation*, 77(12), 1081-1089. DOI: https://doi.org/10.1080/10629360600897191.

Wójciak, W. (2019). *Optimal Allocation in Stratified Sampling Schemes*. Thèse de doctorat, Warsaw University of Technology, Varsovie, Pologne. http://home.elka.pw.edu.pl/~wwojciak/msc_optimal_allocation.pdf.

Wójciak, W. (2023a). Another solution for some optimum allocation problem. *Statistics in Transition New Series*, 24(5), 203-219. DOI: https://doi.org/10.59170/stattrans-2023-071.

Wójciak, W. (2023b). `stratallo`: *Optimum Sample Allocation in Stratified Sampling*. R package version 2.2.1. https://CRAN.R-project.org/package=stratallo.

Techniques d'enquête, décembre 2024
Vol. 50, nº 2, pp. 559-590
Statistique Canada, nº 12-001-X au catalogue

Rythme quotidien de la qualité des données : résultats d'une enquête menée auprès des chômeurs au New Jersey

Jorge González Chapela[1]

Résumé

Le présent article porte sur la possibilité que la qualité des données d'enquête fluctue au fil de la journée. Après avoir exposé l'argument sur le plan théorique, les données recueillies au moyen d'un panel du *Survey of Unemployed Workers in New Jersey* (enquête sur les chômeurs au New Jersey) sont analysées. Plusieurs indicateurs indirects de l'erreur de réponse sont examinés, notamment la non-réponse partielle, la durée des interviews, l'arrondissement et les mesures de la qualité des données du journal sur l'emploi du temps. Les données probantes que nous avons rassemblées pour l'effet de l'heure de la journée sur les interviews sont fiables ou inexistantes. La non-réponse partielle et la probabilité que la durée des interviews fasse partie des 5 % des durées les plus courtes semblent augmenter en soirée, mais une évaluation plus détaillée demande de recourir à des variables instrumentales.

Mots-clés : Données recueillies au moyen d'un panel; heure de la journée; qualité des données d'enquête; Survey of Unemployed Workers in New Jersey.

1. Introduction

Le fait que les enquêtes constituent un outil essentiel à la recherche empirique semble aussi indéniable que le fait qu'une erreur de mesure peut compromettre la qualité de données d'enquête. Parmi les préceptes qui sous-tendent apparemment les ouvrages sur l'erreur de mesure se loge le principe voulant que le répondant à une enquête doive effectuer un ensemble d'opérations cognitives avant de répondre à une question (par exemple Tourangeau, Rips et Rasinski, 2000, chapitre 1). Chacune des opérations peut se révéler plutôt complexe et demander un travail cognitif très exigeant (Krosnick, 1999). Des recherches approfondies (résumées, entre autres, par Schmidt, Collette, Cajochen et Peigneux, 2007) ont montré que le rendement humain fluctue au fil de la journée pour une vaste gamme de tâches cognitives. Pourtant, les répercussions que ces fluctuations pourraient avoir sur la qualité des données d'enquête demeurent en grande partie inconnues.

Dans le présent article, on tente de déceler les heures de la journée qui pourraient poser problème à la qualité des données d'enquête en exploitant des microdonnées longitudinales à haute fréquence tirées de l'enquête américaine Survey of Unemployed Workers in New Jersey (SUWNJ). Dans l'enquête SUWNJ, on a mené des interviews en ligne chaque semaine, sur une période allant jusqu'à 24 semaines, auprès de quelque 6 000 travailleurs qui étaient sans emploi au début de l'enquête, en octobre 2009. Même si les répondants de l'enquête SUWNJ décidaient eux-mêmes du moment qui leur convenait le mieux pour répondre au questionnaire, la disponibilité d'observations répétées sur chaque répondant permet de soustraire les nombreux facteurs non observés qui sont restés constants au cours de la période relativement courte de l'enquête SUWNJ (comparativement à d'autres enquêtes longitudinales de grande envergure).

1. Jorge González Chapela, Centro Universitario de la Defensa de Zaragoza, Academia General Militar, Ctra. de Huesca s/n, 50090 Zaragoza, Espagne. Courriel : jorgegc@unizar.es.

L'article est structuré de la manière suivante. La section 2 présente des renseignements de base et le contexte de cette recherche. La section 3 décrit les données, la construction des variables principales et la sélection de l'échantillon. La section 4 traite de la méthodologie. Les résultats sont présentés à la section 5. Enfin, la section 6 résume les constatations et propose des orientations pour les recherches futures.

2. Renseignements de base et contexte

2.1 Renseignements de base

Des psychologues et des méthodologistes d'enquête ont caractérisé une série d'étapes cognitives en lien avec le fait de répondre à des questions d'enquête. Tourangeau et coll. (2000, page 8) distinguent quatre étapes (compréhension de la question, récupération de renseignements pertinents, utilisation de ces renseignements pour porter les jugements nécessaires, et sélection et énonciation d'une réponse) et présentent une liste indicative des processus mentaux qui interviennent possiblement dans le processus de réponse. L'attention et la mémoire font partie de cette liste, et l'on a démontré que toutes deux fluctuaient au fil de la journée.

La recherche des fluctuations du rendement cognitif humain selon l'heure de la journée repose de plus en plus sur le modèle dit à deux processus de régulation du sommeil et de l'éveil (Blatter et Cajochen, 2007; Schmidt et coll., 2007). Ce modèle postule que l'influence de l'heure de la journée sur le rendement cognitif est induite par la somnolence, qui est elle-même déterminée par les influences qui interagissent de deux propensions. La propension homéostatique au sommeil s'accumule de façon continue durant le temps d'éveil et diminue progressivement durant le sommeil. La propension à l'état d'éveil oscillatoire de près de 24 heures (ou circadien) équilibre la tendance au sommeil homéostatique accumulée pendant l'état d'éveil.

La propension circadienne à l'éveil, qui résulte d'une horloge interne synchronisée par des signaux créés par la rotation de la Terre (lumière, température, etc.) (Roenneberg, Kuehnle, Juda, Kantermann, Allebrandt, Gordijn et Merrow, 2007), atteint son maximum durant la soirée et son minimum tôt le matin. Ainsi, chez une personne qui dort habituellement de 23 h à 7 h, le rendement cognitif serait à un niveau plus faible durant la nuit et tôt le matin, atteindrait un meilleur niveau autour de midi, accuserait une baisse après le repas du midi (par exemple Bes, Jobert et Schulz, 2009) et serait à un niveau élevé en après-midi et en soirée (Valdez, 2019). Cependant, cette chronologie peut être modulée par le genre de tâche et les différences entre les personnes quant au rendement lié à la tâche (Blatter et Cajochen, 2007).

La phase de la propension circadienne à l'éveil et celle des signaux diffère d'une personne à l'autre, ce qui crée un lien entre le temps interne et externe appelé « phase d'entraînement ». Les personnes dont la phase d'entraînement diffère sont désignées comme ayant des chronotypes différents. L'harmonisation du chronotype avec l'heure de la journée améliore un certain nombre de fonctions cognitives, ce qui engendre l'effet dit de synchronie (par exemple Hasher, Goldstein et May, 2005; Hornik et Tal, 2010; Salehinejad, Wischnewski, Ghanavati, Mosayebi-Samani, Kuo et Nitsche, 2021; Guarana, Stevenson, Gish, Ryu et Crawley, 2022). Par conséquent, si les personnes répondaient à des questionnaires durant les heures

concordant avec leur chronotype (comme le laissent croire les données probantes dans Fordsham, Moss, Krumholtz, Roggina, Robinson et Litman, 2019), l'effet de l'heure de la journée serait modéré positivement par le tri des répondants aux heures optimales.

2.2 Contexte

Une exécution minutieuse et complète de chacune des quatre étapes du processus de réponse de l'enquête peut nécessiter une quantité importante d'efforts mentaux. Ainsi, selon le principe du seuil de satisfaction de l'individu de Krosnick (1991), les répondants de l'enquête pourraient simplement fournir une réponse satisfaisante, et la probabilité de le faire diminue en fonction de la capacité du répondant. Cette observation a favorisé la réalisation d'études portant sur le lien entre la capacité cognitive, considérée comme un trait stable ou changeant lentement, et la qualité des données. Consultez par exemple Kaminska, McCutcheon et Billiet (2010), Kroh, Lüdtke, Düzel et Winter (2016), Gideon, Helppie-McFall et Hsu (2017), Olson, Smyth et Ganshert (2019), Truebner (2021), Angrisani et Couper (2022), Bais, Schouten et Toepoel (2022), et Phillips et Stenger (2022). Comme l'avait prédit Krosnick (1991), la capacité cognitive et le principe du seuil de satisfaction de l'individu semblent habituellement présenter un lien inversé.

Les fluctuations du rendement cognitif selon l'heure de la journée pourraient constituer un autre volet de la capacité du répondant lié au principe du seuil de satisfaction de l'individu. Ce lien potentiel a toutefois suscité peu d'études. Ziniel (2008, chapitre 4) a examiné si la proportion de réponses « Je ne sais pas » fournies par les répondants de l'étude américaine *Health and Retirement Study* (étude sur la santé et la retraite) est sensible à l'heure de la journée, pour en arriver à une conclusion négative. Binder (2022) a recruté des participants d'Amazon Mechanical Turk (MTurk) afin d'examiner si les attentes en ce qui concerne l'inflation et les réponses objectivement correctes à des questions différaient selon l'heure de la journée, pour ne trouver que peu de différences. En revanche, une enquête menée auprès de fournisseurs en concurrence pour l'obtention de contrats publics en Irlande (Flynn, 2018) révèle que l'heure de la journée à laquelle les répondants avaient commencé à répondre au questionnaire permettait de savoir s'il serait dûment rempli.

Ces études antérieures étaient limitées du fait que les répondants décidaient eux-mêmes de répondre au questionnaire au moment qui leur convenait le mieux. Comme l'a reconnu Ziniel (2008, chapitre 4), les différences entre les personnes en matière de capacité cognitive ou de chronotype pourraient donc interférer avec d'éventuelles fluctuations du rendement cognitif selon l'heure de la journée. Pour être certains que des facteurs comme ceux-là n'interfèrent pas avec l'heure de la journée, Dickinson et McElroy (2010) ont attribué au hasard la fenêtre de réponse au questionnaire et constaté que l'heure de la journée (représentée par une variable binaire égale à une unité pour les heures de réponse allant de 1 h à 5 h et à zéro pour les heures de réponse allant de midi à 19 h) n'avait aucun effet sur le raisonnement itératif.

Il est pertinent de déterminer les heures de la journée problématiques pour la qualité des données d'enquête, tout d'abord dans le cadre des enquêtes, puisqu'il serait possible de mettre en place des mesures supplémentaires pour réduire l'étendue de l'erreur de mesure. Il serait par exemple possible de programmer l'envoi par courriel d'invitations ou de rappels pour remplir des questionnaires ou même organiser la

collecte des données à des heures de la journée qui conviendraient mieux pour augmenter la qualité des données. Cependant, forcer des répondants à remplir des questionnaires à des heures précises de la journée pourrait faire augmenter l'erreur de non-réponse (par exemple Weeks, Kulka et Pierson, 1987; Durrant, D'Arrigo et Steele, 2011). Alors dans le cadre de l'erreur d'enquête totale (par exemple Lyberg et Stukel, 2017), il faudrait étudier le compromis entre l'erreur de mesure et l'erreur de non-réponse.

Outre les articles déjà mentionnés, nos travaux reposent sur des ouvrages appartenant à d'autres courants. Certaines études ont porté sur les caractéristiques et les comportements des participants à des enquêtes en ligne en fonction de l'heure de la participation au cours de la journée (par exemple Arechar, Kraft-Todd et Rand, 2017; Casey, Chandler, Levine, Proctor et Strolovitch, 2017; Binder, 2022). Malgré une possible corrélation entre certaines caractéristiques des répondants et la qualité des données, nous nous concentrons sur la qualité des données et créons des effets nets liés à des facteurs individuels non observés et aux heures optimales de participation. Les fluctuations du rendement cognitif selon l'heure de la journée ont été tenues responsables de la variation au fil de la journée d'un vaste éventail de décisions et de capacités économiques; consultez par exemple Carrell, Maghakian et West (2011), Dickinson et McElroy (2017), Williams et Shapiro (2018), Collinson, Mathmann et Chylinski (2020), Dickinson, Chaudhuri et Greenaway-McGrevy (2020), et Guarana et coll. (2022). Mais la possibilité que la qualité des données d'enquête soit modulée selon l'heure de la journée demeure en grande partie inconnue. Enfin et surtout, en tirant parti de l'heure de début et de fin de chaque interview, nous établissons des liens avec la documentation en utilisant des paradonnées pour étudier l'erreur de mesure (examinée par Yan et Olson, 2013).

3. Données, mesures et sélection de l'échantillon

3.1 Enquête SUWNJ

Les données de la présente étude sont tirées de l'enquête SUWNJ, une enquête longitudinale menée en ligne d'octobre 2009 à avril 2010 auprès de chômeurs par le Survey Research Center de l'Université Princeton. Nous décrivons dans la présente les principaux volets de cette enquête, en nous reportant à Krueger et Mueller (2010 et 2011) pour le questionnaire de l'enquête, l'ensemble de données et une description plus détaillée de l'enquête SUWNJ. Le code Stata nécessaire pour passer des données brutes aux résultats est disponible sur demande auprès de l'auteur.

3.1.1 Échantillonnage et invitation

Les personnes sélectionnées pour faire partie de l'échantillon proviennent de l'univers des bénéficiaires de prestation d'assurance-emploi (AE) de l'État du New Jersey en date du 28 septembre 2009. Durant les années 2009 et 2010, le taux de chômage du New Jersey reflétait fidèlement la moyenne des États-Unis, mais sa population de bénéficiaires de l'AE se composait davantage de femmes et de personnes plus âgées et plus scolarisées que dans l'ensemble du pays. L'échantillon, sélectionné par échantillonnage aléatoire stratifié, se composait de strates définies par la durée initiale de chômage et la disponibilité d'une adresse

courriel. Les personnes sans emploi depuis au moins 60 semaines et celles possédant une adresse courriel étaient surreprésentées dans l'échantillon.

Les personnes sélectionnées ont été invitées à participer à l'enquête pour une durée de 12 semaines consécutives, alors que les chômeurs de longue date ont été invités à participer à une étude prolongée pendant 12 semaines supplémentaires. L'invitation initiale a fait l'objet d'un envoi par courriel ou (aux personnes sans adresse courriel) par lettre. Le courriel (ou la lettre) contenait un hyperlien vers le questionnaire en ligne. Les personnes ayant reçu une lettre devaient saisir une adresse courriel valide afin de recevoir des invitations par courriel concernant les interviews de suivi hebdomadaires. Un répondant sans adresse courriel pouvait tout de même prendre part aux interviews hebdomadaires en ouvrant une session sur la même page d'accès Web. Selon la Current Population Survey (enquête sur l'état de la population) d'octobre 2009, 15 % des chômeurs du New Jersey vivaient au sein de ménages dans lesquels personne n'utilisait Internet, mais aucune autre disposition n'a été prise pour faire en sorte que les non-utilisateurs d'Internet participent à l'enquête. Dans les courriels d'invitation (envoyés le matin), on demandait aux personnes de remplir le questionnaire dans les deux jours suivants, et ce, même s'ils avaient déjà trouvé un emploi.

3.1.2 Participation et pondération

Le RR6 (taux de réponse des interviews complètes et partielles) de l'American Association for Public Opinion Research (2023) pour la première interview s'élevait à 9,7 % (6 025 personnes). Ces répondants ont pris part à une moyenne de 4,1 interviews de suivi sur un maximum de 11 (sauf dans le suivi à long terme) et à 24 638 (37,2 %) des interviews de suivi potentielles. Seules 302 personnes ont pris part à 12 interviews, ce qui correspond à une érosion de 95,0 % de l'échantillon de l'étude initiale. Le RR6 pour la première interview de l'étude prolongée était plus important, s'établissant à 56,8 % (1 148 personnes). Ces répondants ont pris part à une moyenne de 6,4 interviews de suivi sur un maximum de 11 et à 7 390 (58,5 %) des interviews de suivi potentielles. Au total, 115 personnes ont pris part à 12 interviews, ce qui correspond à une érosion de 90,0 % de l'échantillon de l'étude prolongée. Le tout a mené à 39 201 interviews.

Les faibles taux de réponse ont occasionné des écarts considérables entre l'univers des bénéficiaires d'AE du New Jersey et les répondants. Krueger et Mueller (2011) ont créé des poids de propension inverse d'après des données administratives tirées du système d'AE. Les poids, appelés « poids de la semaine actuelle », prennent en compte les probabilités d'échantillonnage différentielles et les taux de réponse sur les 12 semaines de l'enquête (ou 24 semaines, pour les personnes ayant participé à l'étude prolongée). Les variables explicatives utilisées pour créer les « poids de la semaine actuelle » étaient des variables indicatrices des strates et des données démographiques sans variation dans le temps.

3.1.3 Instrument d'enquête

Le questionnaire de l'enquête SUWNJ compte deux parties : une enquête de début menée la première semaine et comprenant des questions sur les caractéristiques démographiques, le revenu et le patrimoine,

ainsi qu'une enquête hebdomadaire menée la première semaine et chaque semaine subséquente et comprenant des questions sur la satisfaction à l'égard de la vie, les dépenses pour la nourriture, les activités de recherche d'emploi et l'emploi du temps. Les renseignements sur l'emploi du temps, qui concernent la journée précédant l'interview, sont recueillis au moyen d'un journal sur l'emploi du temps à remplir soi-même de 7 h à 23 h et comportent deux questions sur l'heure du réveil et l'heure du coucher. Pour remplir le journal, les répondants pouvaient sélectionner jusqu'à deux activités pour chaque heure à partir d'une liste prédéterminée de 21 activités.

Après avoir commencé une interview, les répondants pouvaient aller plus loin dans le questionnaire et retourner en arrière, ainsi qu'interrompre l'interview et y revenir plus tard. Même s'il était possible de remplir le questionnaire à l'aide du navigateur d'un téléphone, le questionnaire n'était pas optimisé pour les appareils mobiles. L'ensemble de données comprend la date et l'heure (enregistrées à la seconde près) du début et de l'achèvement de chaque interview, ainsi que le temps de fin de la section sur l'emploi du temps (la troisième des cinq sections de l'enquête hebdomadaire).

3.2 Mesures

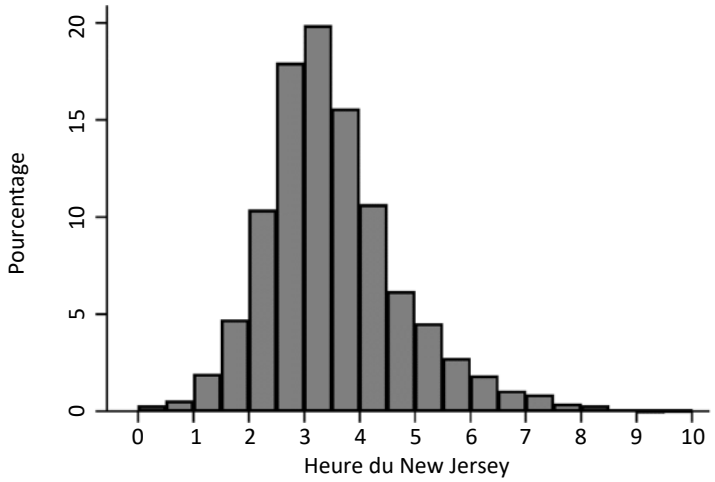
3.2.1 Heure de la journée

Les heures correspondent au fuseau horaire du New Jersey, mesurées en continu à partir de minuit et exprimées en heures (par exemple 9,5 pour 9 h 30). L'heure de la journée de l'interview (représentée par D) est approximée par l'heure médiane entre les heures de début et de fin de l'interview. Dans une vérification de la robustesse, elle sera approximée à l'aide de points sélectionnés au hasard entre les heures de début et de fin de l'interview (Ahn, Peng, Park et Jeon, 2012).

3.2.2 Chronotype

Roenneberg et coll. (2007) font appel au questionnaire chronotype de Munich pour déterminer le chronotype, mesuré comme étant le point à mi-chemin entre le début du sommeil et la fin du sommeil (ou mi-sommeil) durant les jours libres, qui sont corrigés en cas de durée plus longue du sommeil (MSF_{sc}). Une mesure indirecte du chronotype peut être bâtie de façon semblable à l'aide des renseignements sur l'emploi du temps tirés de l'enquête SUWNJ. La durée du sommeil est estimée à partir du temps écoulé entre le coucher et le réveil, et son point à mi-chemin correspond à une moyenne au cours des jours libres. Chez les personnes qui dorment davantage les jours libres que les jours de travail, la différence entre la durée du sommeil les jours libres et sa moyenne hebdomadaire (en supposant que la semaine de travail compte 5 jours) est soustraite du mi-sommeil les jours libres. La mesure résultante est désignée par MSF^{e}_{sc}. Le moment et la durée du sommeil sont des traits essentiellement indépendants (Roenneberg et coll., 2007). La corrélation entre MSF^{e}_{sc} et la durée moyenne du sommeil est de 0,06 (bien qu'elle soit statistiquement différente de zéro à un niveau de 5 %). La figure 3.1 montre la répartition de MSF^{e}_{sc} dans l'échantillon.

Figure 3.1 Chronotype (MSF_{sc}^{e}).



Source : Enquête SUWNJ.

3.2.3 Qualité des données

Nous analysons quatre ensembles de mesures de la qualité des données (Juster, 1986; Malhotra, 2008; Fricker et Tourangeau, 2010) : i) le pourcentage de non-réponse partielle; ii) les mesures de la qualité des données du journal sur l'emploi du temps (le nombre et la variété d'activités, et le nombre d'heures non codées dans le journal); iii) la durée de l'interview; iv) les valeurs arrondies de l'humeur à la maison, des dépenses liées à la nourriture à la maison et des dépenses liées aux sorties au restaurant. Le questionnaire de l'enquête SUWNJ semble ne pas contenir suffisamment d'éléments pour étudier les erreurs de réponse causées par la désirabilité sociale ou les réponses extrêmes, médianes ou indifférenciées (consultez par exemple Baumgartner et Steenkamp, 2001; Chang et Krosnick, 2009). Le fichier de données contient des interviews achevées, ce qui empêche de procéder à une analyse de l'interruption de l'enquête (par exemple Peytchev, 2009).

Nous définissons le pourcentage de non-réponse partielle (P_{INR}) comme étant le pourcentage de valeurs manquantes à des questions soumises à tous les répondants, dans le cadre d'une certaine interview. Cela exclut les questions de suivi et les questions qu'il est possible de reporter à la prochaine interview de la personne.

Pour faire le compte du nombre d'activités et de la variété des activités enregistrées dans le journal sur l'emploi du temps, nous suivons la convention voulant que si une activité intervient au milieu d'une certaine autre activité (par exemple magasinage sur le chemin du retour à la maison après une entrevue d'emploi), le nombre d'activités augmente de deux unités et la variété des activités, d'une unité (Juster, 1986). Nous présentons les résultats pour le nombre d'activités (désigné par NbrAct), puisque ceux relatifs à la variété obéissent aux mêmes schémas. En l'absence d'activité enregistrée à une heure donnée, l'heure est considérée comme étant non codée. La variable indiquant le nombre d'heures non codées est désignée par

HManquante. Juster (1986) a remarqué que seuls les journaux des jours de semaine (lundi à jeudi) présentent une importante détérioration de la qualité dans la mesure où ils nécessitent plus qu'un rappel de 24 heures, une constatation qui se révélera utile pour interpréter certains de nos résultats.

La relation entre le temps d'achèvement de l'interview et la qualité des données dans le cas d'enquêtes en ligne est complexe, puisque des temps d'achèvement courts et longs pourraient tous deux être attribuables à l'inattention du répondant (Malhotra, 2008; Read, Wolters et Berinsky, 2021). Ainsi, en plus d'une mesure continue du temps d'achèvement (désigné par IvDur), nous analysons des variables nominales pour les 5 % des temps d'achèvement les plus courts et les 5 % des plus longs, désignés par $P_{IVDUR5L}$ et $P_{IVDUR5H}$, respectivement. Ces variables nominales sont créées en calculant séparément les centiles correspondant à la première interview et aux interviews subséquentes, après avoir retiré les valeurs aberrantes (consultez la section 3.3).

Les renseignements sur l'humeur à la maison sont recueillis à l'aide de la question suivante : « Nous voudrions maintenant savoir comment vous vous sentez et quelle est votre humeur lorsque vous êtes à la maison. Lorsque vous êtes à la maison, quelle est la proportion de temps où vous êtes : de mauvaise humeur, un peu déprimé ou irritable, d'humeur plutôt agréable, de très bonne humeur ? » [*traduction*] Les répondants devaient indiquer la proportion de temps qu'ils passaient dans chacune des catégories d'humeur. Nous avons créé des variables nominales indiquant les répondants pour qui les pourcentages *des quatre* catégories sont des multiples de 50 (menant à des réponses de 0, de 50 ou de 100), de 25 ou de 10. Les trois variables binaires sont désignées par P_{MOOD50}, P_{MOOD25} et P_{MOOD10}, respectivement.

Deux questions ont permis de recueillir des renseignements sur les dépenses liées à la nourriture : i) « Au cours des sept derniers jours, combien vous et tout autre membre de votre famille avez dépensé pour de la nourriture utilisée à la maison ? Veuillez inclure les aliments achetés à l'aide de bons alimentaires. » ii) « Au cours des sept derniers jours, combien vous et tout autre membre de votre famille avez dépensé pour des sorties au restaurant ? » Nous avons créé des variables nominales indiquant les répondants pour qui certaines dépenses constituaient un multiple de 100 ou de 50, et ces variables sont désignées par $P_{FOODAH100}$, $P_{FOODAH50}$, $P_{EATING\text{-}OUT100}$ et $P_{EATING\text{-}OUT50}$. Une dépense nulle pourrait correspondre à un arrondissement, à une solution d'angle ou à la rareté des achats. Nous présentons les résultats en présumant que les dépenses nulles représentent un arrondissement et analysons leur robustesse pour présumer qu'elles n'en représentent pas un.

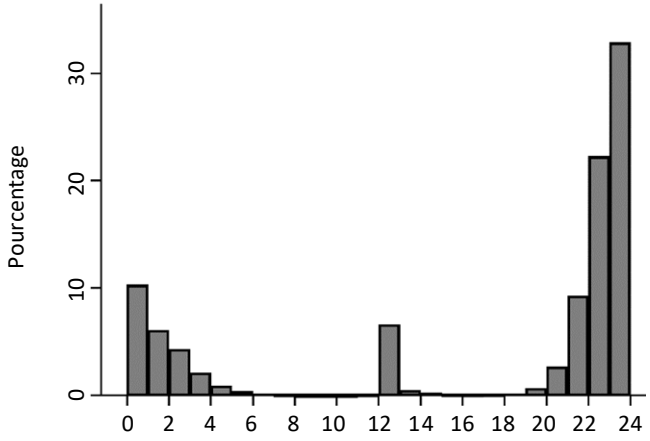
3.3 Sélection de l'échantillon

La répartition du temps d'achèvement des interviews présente une forte asymétrie à droite, le temps d'achèvement médian (moyen) étant de 13,2 (144,3) minutes pour les premières interviews et de 11,8 (75,6) minutes pour les interviews subséquentes. Pour éviter l'introduction de plus d'erreurs dans notre mesure de D, les interviews dont le temps d'achèvement est de plus de 60 minutes sont retirées, ce qui représente 5,7 % et 4,4 % des premières interviews et des interviews subséquentes. Nous avons également retiré les

premières interviews ou les interviews subséquentes dont le temps d'achèvement était inférieur au premier centile correspondant (4,8 minutes et 3,6 minutes, respectivement). D'après notre propre temps de lecture, cette borne inférieure supprime les interviews dans lesquelles le répondant ne peut pas avoir lu le questionnaire. (Sans compter que ces interviews ne changent rien aux conclusions.)

Nous avons aussi retiré les interviews comportant des données manquantes ou incohérentes concernant certaines variables utilisées dans la présente étude. Dans ce cas, une question demande une certaine analyse. Trois menus déroulants comprenant les heures, les minutes et la période de la journée (avant-midi ou après-midi) servaient à indiquer l'heure du coucher. Le menu de la période de la journée était réglé par défaut sur l'après-midi, et la figure 3.2 indique que cela aurait peut-être induit une erreur. Si l'heure du coucher déclarée se trouve entre 11 h et 11 h 59 dans 0,04 % des interviews, 6,6 % des répondants ont indiqué une heure de coucher entre 12 h et 12 h 59. Nous avons examiné les journaux sur l'emploi du temps pour trouver des heures de coucher incohérentes déclarées entre 12 h 00 et 14 h 59. En cas d'incohérence, l'interview était retirée. En tout, 2 578 interviews ont été supprimées pour cette raison. Lorsque nous vérifions la robustesse, nous devons les inclure dans l'échantillon en présumant que la période de l'avant-midi s'applique, et une variable nominale indiquant ces cas (désignée par $P_{PM\text{-}AM}$) sera analysée pour les effets de l'heure de la journée sur l'interview.

Figure 3.2 Heure du coucher.



Source : Enquête SUWNJ.

Enfin, la dernière interview d'une personne ayant achevé 25 interviews est retirée parce que l'on ne sait pas clairement si la personne a fini par abandonner l'enquête. Tout ce procédé nous laisse 5 531 personnes et un total de 33 000 interviews. La figure 3.3 présente un histogramme du nombre d'interviews réalisées par chaque personne. Le nombre moyen (médian) d'interviews s'élève à 6,0 (4). Le tableau 3.1 fournit des statistiques descriptives pour les variables principales utilisées dans la présente étude.

Figure 3.3 Nombre d'interviews effectuées pour l'échantillon.

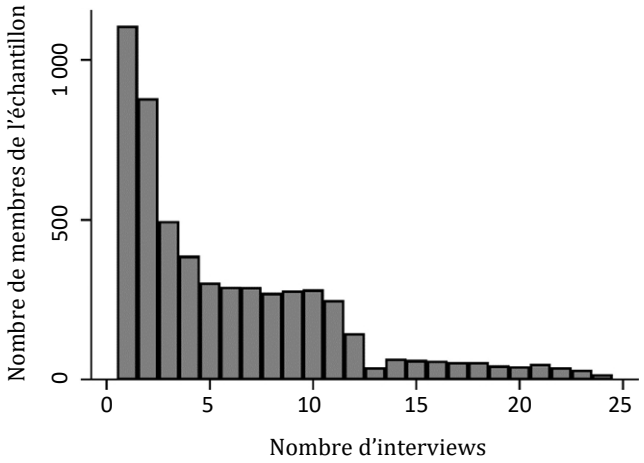


Tableau 3.1
Statistiques descriptives.

	Observations	Moyenne	Écart-type	Min.	Max.
P_{INR}[a]	33 000	2,64	6,58	0	60,87
NbrAct	33 000	16,77	7,14	1	32
HManquante	33 000	0,53	2,14	0	15
IvDur (minutes)	33 000	13,95	8,19	3,57	59,95
$P_{IVDUR5L}$[a]	33 000	4,99			
$P_{IVDUR5H}$[a]	33 000	4,99			
P_{MOOD10}[a]	32 877	50,40			
P_{MOOD25}[a]	32 877	15,84			
P_{MOOD50}[a]	32 877	10,79			
$P_{FOODAH50}$[a]	31 949	51,64			
$P_{FOODAH100}$[a]	31 949	30,63			
$P_{EATING\text{-}OUT50}$[a]	29 084	45,74			
$P_{EATING\text{-}OUT100}$[a]	29 084	34,67			
Heure de la journée de l'interview	33 000	12,94	4,80 (3,89) [3,45]	0	23,99
MSF^{c}_{sc}	5 531	3,56	1,65	0	23,99
Jour de l'interview	33 000				
Lundi[b]		8,56			
Mardi[b]		23,39			
Mercredi[b]		16,18			
Jeudi[b]		14,51			
Vendredi[b]		17,93			
Samedi[b]		12,76			
Dimanche[b]		6,67			
A travaillé[b]	33 000	14,00			
Durée du sommeil (heures)	33 000	8,35	2,10	0,50	23,58
Nbre d'interviews précédentes	33 000	5,34	5,00	0	23
Semaines entre $t-2$ et $t-1$	33 000	1,43	1,40	0	16

Notes : Les données portent sur 5 531 personnes. La variation dans l'échantillon de l'heure de la journée de l'interview se compose d'une intravariation (ou une variation chronologique) [montrée entre parenthèses] et d'une intervariation (section transversale) [montrée entre crochets]. « A travaillé » et « durée du sommeil » s'appliquent à un jour du journal. [a] : indicateur binaire pour le résultat inscrit dans l'indice du nom, exprimé sous forme de pourcentage. [b] : indicateur binaire exprimé sous forme de pourcentage.

4. Méthodes

4.1 Spécifications de référence

Puisque les répondants décident eux-mêmes de répondre aux questionnaires au moment qui leur convient le mieux, il est improbable qu'une simple comparaison des résultats de la qualité des données en fonction de l'heure de la journée de l'interview permette d'établir un effet de causalité. La disponibilité d'observations répétées sur chaque répondant de l'enquête SUWNJ nous donne les moyens de prendre en compte des facteurs non observés sans variation dans le temps, comme la capacité cognitive ou le chronotype. L'erreur de mesure (selon la définition de Biemer, Groves, Lyberg, Mathiowetz et Sudman, 2004, page xvii) découle aussi de la méthode de collecte des données et du questionnaire, mais puisque ces traits sont fixes d'une interview à l'autre, ils ne peuvent pas interférer avec nos estimations.

Le modèle suivant de données recueillies au moyen d'un panel d'effets non observés (Wooldridge, 2010, chapitre 10) est estimé ainsi :

$$y_{it} = \alpha\left(D_{it}\right) + \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it} \tag{4.1}$$

où y_{it} désigne une certaine mesure de la qualité des données pour chaque i $\left(i = 1, 2, \ldots, N\right)$ à un nombre d'interviews t $\left(t = 1, 2, \ldots, T_i\right)$, $\alpha\left(D_{it}\right)$ est une fonction scalaire de l'heure de la journée de l'interview, $\mathbf{x}_{it}$ est un vecteur des variables de contrôle observables variant selon l'interview, $\boldsymbol{\beta}$ est un vecteur des paramètres inconnus, c_i est un effet individuel non observé arbitrairement corrélé avec D_{it} et $\mathbf{x}_{it}$, et u_{it} est un terme d'erreur idiosyncrasique.

Sauf pour une ordonnée à l'origine, et selon Binder (2022) et Juster (1986), des variables nominales pour le jour de la semaine de l'interview sont comprises dans $\mathbf{x}_{it}$, de même qu'une variable nominale pour le fait que le répondant ait ou non travaillé le jour réservé au journal (ce renseignement n'est pas disponible pour le jour de l'interview). Un manque cumulatif de sommeil (par exemple Lowe, Safati et Hall, 2017) et des effets de synchronie pourraient également influer sur y_{it}. Par conséquent, la durée du sommeil le jour réservé au journal et l'interaction entre $\mathrm{MSF}^{\mathrm{e}}_{\mathrm{sc}}$ et les variables nominales pour une seule heure relatives à D_{it} sont comprises dans $\mathbf{x}_{it}$. Les variables nominales pour une seule heure sont construites à partir de l'arrondissement de D_{it} à l'heure entière la plus près, ce qui donne lieu à 24 variables nominales qui interagiront avec $\mathrm{MSF}^{\mathrm{e}}_{\mathrm{sc}}$. Là encore, une variable nominale est exclue en raison de sa colinéarité avec c_i. La médiane $\mathrm{MSF}^{\mathrm{e}}_{\mathrm{sc}}$ est soustraite de $\mathrm{MSF}^{\mathrm{e}}_{\mathrm{sc}}$, de sorte que $\alpha\left(D_{it}\right)$ représente le chronotype médian.

Les effets de conditionnement du panel peuvent agir dans une enquête longitudinale, ce qui entraînerait possiblement des conséquences positives ou négatives sur la qualité des données (par exemple Bach, 2021). Les répondants pourraient acquérir une meilleure compréhension du sens des questions grâce à la soumission répétée du questionnaire, ce qui rehausserait la fiabilité de leurs réponses (Kroh, Winter et Schupp, 2016). Par ailleurs, les répondants pourraient apprendre à donner de fausses réponses à certaines questions afin d'éviter les questions de suivi, ce qui réduirait la qualité des données (par exemple Davis, 2011). Pour tenir compte des effets de conditionnement du panel, un ensemble complet de variables nominales est inséré dans $\mathbf{x}_{it}$ pour le nombre d'interviews précédentes. Ce nombre peut être 0, 1, 2, …, 23,

ce qui produit 24 variables nominales. Encore une fois, une variable nominale est exclue en raison de sa colinéarité avec l'ordonnée à l'origine.

Soit $\mathbf{z}_{it}\boldsymbol{\theta} \equiv \alpha(D_{it}) + \mathbf{x}_{it}\boldsymbol{\beta}$ et $K = \dim(\boldsymbol{\theta})$. Selon l'hypothèse d'exogénéité stricte $E(u_{it} \mid \mathbf{z}_{i1}, \mathbf{z}_{i2}, \ldots, \mathbf{z}_{iT_i}, c_i) = 0$, il est possible d'estimer $\boldsymbol{\theta}$ à partir des moindres carrés ordinaires de

$$\Delta y_{it} = \Delta \mathbf{z}_{it}\boldsymbol{\theta} + e_{it}, \qquad t = 2, 3, \ldots, T_i, \tag{4.2}$$

où $\Delta y_{it} = y_{it} - y_{i,t-1}$, $\Delta \mathbf{z}_{it} = \mathbf{z}_{it} - \mathbf{z}_{i,t-1}$ et $e_{it} = u_{it} - u_{i,t-1}$. Les e_{it} sont présumés être répartis de façon indépendante entre les personnes, mais aucune restriction n'est imposée quant à la forme des auto-covariances pour une personne donnée. Des erreurs-types cohérentes en termes d'hétéroscédasticité et de corrélation sérielle sont obtenues à partir de l'estimateur suivant de la matrice de variance (Wooldridge, 2010, pages 172 et 318) :

$$\hat{\mathrm{V}}\left(\hat{\boldsymbol{\theta}}\right) = \left(\sum_{i=1}^{N} \Delta \mathbf{Z}_i' \Delta \mathbf{Z}_i\right)^{-1} \left(\sum_{i=1}^{N} \Delta \mathbf{Z}_i' \hat{\mathbf{e}}_i \hat{\mathbf{e}}_i' \Delta \mathbf{Z}_i\right) \left(\sum_{i=1}^{N} \Delta \mathbf{Z}_i' \Delta \mathbf{Z}_i\right)^{-1} \tag{4.3}$$

où $\Delta \mathbf{Z}_i$ est la matrice $(T_i - 1) \times K$ obtenue en regroupant $\Delta \mathbf{z}_{it}$ à partir de $t = 2, 3, \ldots, T_i$ et $\hat{\mathbf{e}}_i$ est le vecteur $(T_i - 1) \times 1$ des résidus des moindres carrés ordinaires $\hat{e}_{it}$, $t = 2, 3, \ldots, T_i$. Autrement, il est possible de concevoir une matrice de corrélation de travail pour la modélisation de corrélations parmi les personnes et d'estimer le modèle résultant à l'aide de méthodes pour la moyenne des populations, appelées estimateurs des moindres carrés généralisés réalisables (MCGR) en économétrie. Nous fournissons les résultats de deux estimateurs des MCGR dans un supplément distinct (González Chapela, 2024). Ils révèlent essentiellement les mêmes schémas que ceux indiqués dans la présente étude.

Pour évaluer l'exogénéité stricte de $\{D_{it} : t = 1, \ldots, T_i\}$, $\alpha(D_{it})$ sera ajouté à l'équation (4.2), puis sa signification sur le plan statistique sera mesurée à l'aide du test de Wald (Wooldridge, 2010, page 325).

4.2 Types de modèles et sélection du modèle

Notre objectif consiste à obtenir une représentation raisonnable et parcimonieuse de $\alpha(D_{it})$. Par conséquent, un critère d'information sert à sélectionner un modèle pour $\alpha(D_{it})$ parmi trois types de modèles ayant des paramètres linéaires : les fonctions constantes par morceaux (plus particulièrement celles d'Arechar et coll., 2017; de Binder, 2022; de Durrant et coll., 2011; de Flynn, 2018; de Valdez, 2019; de Weeks et coll., 1987), la fonction polynomiale du troisième degré et le modèle cosinor

$$\alpha(D_{it}) = \alpha_1 \sin(D_{it} \times 2\pi/24) + \alpha_2 \cos(D_{it} \times 2\pi/24), \tag{4.4}$$

où α_1 et α_2 sont des paramètres inconnus.

Le modèle cosinor est un type de représentation de la série de Fournier dans lequel des sinus et des cosinus servent à approximer des formes d'ondes mathématiques complexes (Brown et Czeisler, 1992; Cornelissen, 2014). Compte tenu du caractère ondulatoire des propensions homéostatique et circadienne au sommeil, le modèle cosinor pourrait fournir une représentation convenable de $\alpha(D_{it})$. Le modèle cosinor comporte un pic et un creux qui sont séparés par un intervalle de 12 heures et sont égaux en amplitude et en

largeur; leur emplacement est déterminé par α_1 et α_2. Le double de l'amplitude de l'onde du modèle cosinor, ou $2 \times \sqrt{\alpha_1^2 + \alpha_2^2}$, donne une mesure de l'étendue d'un changement prévisible dans la journée.

La fonction polynomiale du troisième degré est moins restrictive que le modèle cosinor parce qu'il est possible que le pic et le creux ne soient pas séparés par un intervalle de 12 heures; de même, l'amplitude et la largeur du pic pourraient différer de celles du creux. Par contre, la fonction polynomiale pourrait ne pas être périodique, c'est-à-dire que ses valeurs pourraient ne pas se répéter toutes les 24 heures. Pour assurer une périodicité, la restriction $\alpha(0) = \alpha(24)$ est imposée, pour donner

$$\alpha(D_{it}) = \alpha_1 D_{it}\left(1 - \left(D_{it}/24\right)^2\right) + \alpha_2 D_{it}^2\left(1 - D_{it}/24\right). \tag{4.5}$$

Pour faire un choix entre les modèles, le critère d'information bayésien (BIC) de Schwarz (1978)

$$\text{BIC} = \ln \text{SSR} + \frac{K \ln\left(\sum_{i=1}^{N}(T_i - 1)\right)}{\left(\sum_{i=1}^{N}(T_i - 1)\right)} \tag{4.6}$$

est utilisé, où SSR désigne la somme des résidus carrés d'un modèle. Le BIC est préféré à d'autres critères populaires quand certaines solutions de modélisationde rechange sont emboîtées (Nishii, 1988). La spécification de $\mathbf{x}_{it}$ reste la même tout au long du processus de sélection. Schwarz (1978) a établi la validité du BIC pour les observations indépendantes et réparties de façon identique. Pour se protéger contre de possibles biais créés par la corrélation de e_{it}, les valeurs du BIC ont été recalculées à l'aide de N au lieu de $\sum_{i=1}^{N}(T_i - 1)$ (StataCorp, 2019, page 104), ce qui produit la même sélection de modèles.

4.3 Attrition

Si l'attrition découle de facteurs non observés qui ne changent pas durant la période d'enquête, retirer c_i corrigerait alors le biais d'attrition. Il serait néanmoins possible de se soucier d'une attrition étant la conséquence de facteurs non observés variant selon l'interview. Nous utilisons une variante de la procédure proposée par Wooldridge (2010, page 837) pour tester et corriger le biais d'attrition, mais constatons que cette procédure ne permet pas de prendre en compte les personnes sélectionnées pour participer à l'enquête SUWNJ qui n'y ont jamais répondu. Puisque les données de chaque personne sont organisées en fonction du nombre d'interviews, l'attrition constitue un état sans retour.

Soit s_{it}, l'indicateur d'achèvement d'une interview, où $s_{it} = 1$ si une personne i a achevé les t interviews et $s_{it} = 0$ si i a délaissé l'enquête juste après l'interview $t-1$. L'équation pour l'achèvement de l'interview t conditionnel à $s_{i,t-1} = 1$ est

$$s_{it} = 1\left[\mathbf{w}_{it}\boldsymbol{\delta} + v_{it} > 0\right], \qquad t = 2, 3, \ldots, T_i, \tag{4.7}$$

où $1[\cdot]$ est la fonction indicatrice, $\mathbf{w}_{it}$ est un ensemble de variables qui sont observées que la personne fasse ou non l'objet d'une attrition, $\boldsymbol{\delta}$ est un vecteur de paramètres inconnus et v_{it} est un terme d'erreur normale standard présumé indépendant de $\left(\Delta\mathbf{z}_{it}, \mathbf{w}_{it}, s_{i,t-1} = 1\right)$. Une attrition non aléatoire survient en cas de corrélation entre v_{it} et e_{it}.

En supposant que e_{it} est indépendant de $(\Delta \mathbf{z}_{it}, \mathbf{w}_{it})$ et que $\mathrm{E}(e_{it} \mid v_{it}, s_{i,t-1}=1) = \rho_t v_{it}$, ρ_t étant un paramètre inconnu, il est possible d'estimer les paramètres inconnus de l'équation (4.1) à l'aide des moindres carrés ordinaires de

$$\Delta y_{it} = \Delta \mathbf{z}_{it}\boldsymbol{\theta} + \rho_2 d2_t \hat{\lambda}_{it} + \ldots + \rho_{24} d24_t \hat{\lambda}_{it} + \varepsilon_{it}, \qquad t = 2, 3, \ldots, T_i. \tag{4.8}$$

Dans cette expression, $d2_t, \ldots, d24_t$ sont des variables nominales d'interview, de sorte que $dj_t = 1$ si $t = j$ et $dj_t = 0$ si $t \neq j$, $\hat{\lambda}_{it} \equiv \lambda(\mathbf{w}_{it}\hat{\boldsymbol{\delta}}) = \phi(\mathbf{w}_{it}\hat{\boldsymbol{\delta}}) \,/\, \Phi(\mathbf{w}_{it}\hat{\boldsymbol{\delta}})$, où $\phi(\cdot)$ et $\Phi(\cdot)$ désignent la fonction de densité et la fonction de distribution cumulative de la distribution normale standard, est l'inverse estimé du rapport de Mills, et ε_{it} est un terme d'erreur.

Un estimateur de $\boldsymbol{\delta}$ est disponible à partir de l'estimation par probit groupé de l'équation relative à l'achèvement d'interviews :

$$\mathrm{P}(s_{it} = 1 \mid \mathbf{w}_{it}, s_{i,t-1} = 1) = \Phi(\mathbf{w}_{it}\boldsymbol{\delta}), \qquad t = 2, 3, \ldots, T_i. \tag{4.9}$$

Nous utilisons un probit groupé parce que l'on suppose que $\boldsymbol{\delta}$ est constant dans les interviews. Si $\boldsymbol{\delta}$ pouvait changer (comme dans la formulation originale de Wooldridge), un probit serait estimé pour chaque t. Cette méthode s'avère pourtant problématique parce qu'à de nombreuses occasions, les variables comprises dans $\mathbf{w}_{it}$ prévoient avec précision un des résultats. Le vecteur $\mathbf{w}_{it}$ comprend des variables nominales pour une seule heure pour $D_{i,t-1}$, $\mathbf{x}_{i,t-1}$ et le nombre de semaines écoulées entre $t-2$ et $t-1$. (Pour $t = 2$, nous comptons le nombre de semaines entre la semaine de l'envoi des invitations initiales à participer à l'enquête et la semaine de la première interview.)

Il est possible de tester le biais d'attrition à l'aide d'un test conjoint de $\mathrm{H}_0 : \rho_t = 0,\ t \geq 2,$ dans l'équation (4.8). Si H_0 est rejeté, les erreurs-types sont corrigées pour la présence de paramètres estimés dans $\hat{\lambda}_{it}$ en s'inspirant d'Arellano et Meghir (1992).

4.4 Pondération

Puisque les variables explicatives utilisées pour créer des « poids de la semaine actuelle » sont absorbées dans c_i, le modèle (4.1) comprend toutes les variables du plan de sondage et, par conséquent, le plan de sondage peut être considéré comme pouvant être ignoré (Pfeffermann, 1993). Ainsi, l'analyse principale est menée sans poids de sondage. Il est cependant utile de faire état des estimations pondérées pour la vérification des spécifications erronées, puisque le fait de ne pas modéliser les effets hétérogènes peut générer des contrastes significatifs entre les estimations pondérées et celles ne l'étant pas (par exemple Solon, Haider et Wooldridge, 2015). L'équation (4.2) sera donc estimée de nouveau à l'aide des moindres carrés pondérés.

4.5 Inférence multiple

Presque tous nos groupements de mesures de la qualité des données contiennent plus d'une mesure. Il pourrait par conséquent émerger par hasard des effets marqués pour certaines mesures, même en l'absence d'un effet sur le groupement. Pour tenir compte de cette éventualité, des corrections de Bonferroni sont

appliquées et la signification statistique est fixée au niveau $0{,}05/M$, M étant le nombre de mesures dans le groupement.

5. Résultats

5.1 Sélection du modèle

Le tableau 5.1 présente la liste des modèles de $\alpha(D_{it})$ convenant le mieux. Le modèle cosinor est l'option privilégiée pour l'analyse de la majorité des mesures de la qualité des données. Toutefois, la fonction constante par morceaux de Binder (2022) [indicateurs pour les périodes de 6 h à 11 h 59, de 12 h à 18 h 59 et de 19 h à 5 h 59] est la solution de rechange qui convient le mieux pour le nombre d'heures non codées dans le journal (Hmanquante), la probabilité de déclarer toutes les catégories d'humeur à la maison en multiples de 50 (P_{MOOD50}) et la probabilité de déclarer des dépenses en sorties au restaurant en multiples de 50 ($P_{EATING\text{-}OUT50}$). La fonction polynomiale du troisième degré est privilégiée pour la probabilité de se trouver dans les 5 % des temps d'achèvement les plus élevés ($P_{IVDUR5H}$) et la probabilité de déclarer des dépenses en sorties au restaurant en multiples de 100 ($P_{EATING\text{-}OUT100}$). Pour ce qui est de la probabilité de déclarer toutes les catégories d'humeur à la maison en multiples de 25 (P_{MOOD25}), la fonction constante par morceaux de Durrant et coll. (2011) [indicateurs pour les périodes de 0 h à 11 h 59, de 12 h à 16 h 59 et de 17 h à 23 h 59] est privilégiée.

Tableau 5.1
Modèle sélectionné pour $\alpha(D_{it})$.

Variable dépendante	Modèle	Valeur du BIC
P_{INR}	Cosinor	13,259
NbrAct	Cosinor	13,585
Hmanquante	Fonction constante par morceaux (Binder, 2022)	11,081
IvDur	Cosinor	14,327
$P_{IVDUR5L}$	Cosinor	16,476
$P_{IVDUR5H}$	Fonction polynomiale du troisième degré	16,716
P_{MOOD10}	Cosinor	18,031
P_{MOOD25}	Fonction constante par morceaux (Durrant et coll., 2011)	17,063
P_{MOOD50}	Fonction constante par morceaux (Binder, 2022)	16,670
$P_{FOODAH50}$	Cosinor	18,096
$P_{FOODAH100}$	Cosinor	18,018
$P_{EATING\text{-}OUT50}$	Fonction constante par morceaux (Binder, 2022)	17,956
$P_{EATING\text{-}OUT100}$	Fonction polynomiale du troisième degré	17,759

5.2 Résultats de référence

Les résultats de l'équation d'estimation (4.2) comportant les formes fonctionnelles énumérées au tableau 5.1 sont présentés aux tableaux 5.2 et 5.3. Le tableau 5.2 montre les résultats pour le pourcentage de non-réponse partielle (P_{INR}), les mesures du journal sur l'emploi du temps et le temps d'achèvement des interviews. Le tableau 5.3 regroupe les résultats pour les indicateurs d'arrondissement. Les dernières lignes des deux tableaux renferment la liste des valeurs de p pour les tests de signification de $\alpha(D_{it})$ et l'exogénéité stricte de $\{D_{it} : t = 1, \ldots, T_i\}$.

Une valeur de $\alpha(D_{it})$ significative sur le plan statistique est décelée dans quelques-unes des régressions, ce qui indique l'existence de certains effets sur la qualité des données de D_{it}. Dans le sens d'une valeur de p, les données probantes les plus fiables se trouvent dans les régressions relatives au nombre d'activités (NbrAct) et la probabilité de se trouver dans les 5 % des temps d'achèvement les plus bas ($P_{IVDUR5L}$). L'hypothèse nulle de l'absence d'effet est aussi rejetée à 5 % dans les régressions pour P_{INR} et $P_{EATING\text{-}OUT100}$. Aucun effet significatif sur le plan statistique n'a été décelé dans les cas restants.

Dans le cas de $P_{EATING\text{-}OUT100}$, le rejet de l'hypothèse nulle ne tient qu'en présumant que l'absence de dépenses (déclarée dans 28 % des interviews) ne correspond pas à un arrondissement (valeur de p de 0,55). De plus, l'effet sur $P_{EATING\text{-}OUT100}$ ne survit pas à une correction de Bonferroni dans deux tests simultanés menés dans le groupe de mesures évaluant les dépenses en sorties au restaurant, qui nécessiterait une valeur de $p < 0{,}025$.

Tableau 5.2
Effets de l'heure de la journée de l'interview sur la qualité des données.

	(1)		(2)		(3)		(4)		(5)		(6)	
	P_{INR}		NbrAct		Hmanquante		IvDur (min)		$P_{IVDUR5L}$		$P_{IVDUR5H}$	
Variables explicatives	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**
12 h à 18 h 59					0,049	0,029						
19 h à 5 h 59					0,048	0,039						
$D_{it}(1-(D_{it}/24)^2)$											-0,169	0,238
$D_{it}^2(1-D_{it}/24)$											0,039	0,024
$\sin(D_{it}\times 2\pi/24)$	-0,147*	0,070	0,259*	0,087			-0,020	0,135	-0,405	0,361		
$\cos(D_{it}\times 2\pi/24)$	0,040	0,064	-0,191*	0,086			-0,282*	0,138	0,814*	0,373		
Mardi	-0,190	0,108	1,480*	0,137	-0,103*	0,039	0,733*	0,192	-0,251	0,491	1,013	0,624
Mercredi	-0,157	0,118	1,207*	0,150	-0,079	0,041	0,430*	0,204	-0,960	0,547	-0,046	0,656
Jeudi	0,015	0,131	1,233*	0,159	-0,048	0,043	0,653*	0,226	-0,380	0,573	1,146	0,762
Vendredi	-0,008	0,118	1,132*	0,152	-0,075	0,042	0,272	0,206	0,188	0,505	-0,138	0,656
Samedi	0,012	0,124	0,755*	0,153	-0,027	0,041	0,387	0,246	0,514	0,605	0,140	0,781
Dimanche	0,289*	0,140	0,054	0,171	0,001	0,045	-0,380	0,242	0,151	0,660	-0,882	0,759
A travaillé	0,408*	0,137	-1,711*	0,158	-0,120*	0,037	-0,174	0,173	1,378*	0,609	0,623	0,555
Durée du sommeil	-0,024	0,018	-0,080*	0,025	-0,016*	0,007	-0,131*	0,028	0,400*	0,089	-0,178	0,095
Signification de $\alpha(D_{it})$	[0,04]		[0,00]		[0,22]		[0,10]		[0,01]		[0,14]	
Exogénéité stricte de $\{D_{it}\}$	[0,01]		[0,73]		[0,10]		[0,03]		[0,03]		[0,79]	
Observations	25 184		25 184		25 184		25 184		25 184		25 184	

Notes : Les estimations sont réalisées à l'aide de la première différenciation et comprennent des ensembles complets de variables nominales des différences premières pour un certain nombre d'interviews précédentes et de variables nominales pour une seule heure des différences premières pour D_{it} en interaction avec MSF_{sc}^{e}. Les variables dépendantes dont le nom commence par P sont des indicateurs binaires pour le résultat inscrit dans l'indice du nom, exprimé sous forme de pourcentage. Les erreurs-types prennent en compte l'hétéroscédasticité et la mise en grappes au niveau individuel. Les valeurs de probabilité sont présentées entre crochets. * : significatif à 5 %.

Les effets estimés sur P_{INR}, NbrAct et $P_{IVDUR5L}$, calculés en mettant à zéro toutes les variables de contrôle et en faisant varier D_{it}, sont illustrés à la figure 5.1. Les trois graphiques racontent une histoire plutôt cohérente : la qualité des données atteint un sommet tôt le matin et un creux en soirée. Le changement estimé au sein de la journée est de l'ordre de 0,30 point de pourcentage pour P_{INR}, de 0,64 activité pour NbrAct et de 1,82 point de pourcentage pour $P_{IVDUR5L}$, ce qui représente 11 %, 4 % et 36 % de la moyenne correspondante.

Tableau 5.3
Effets de l'heure de la journée de l'interview sur la qualité des données.

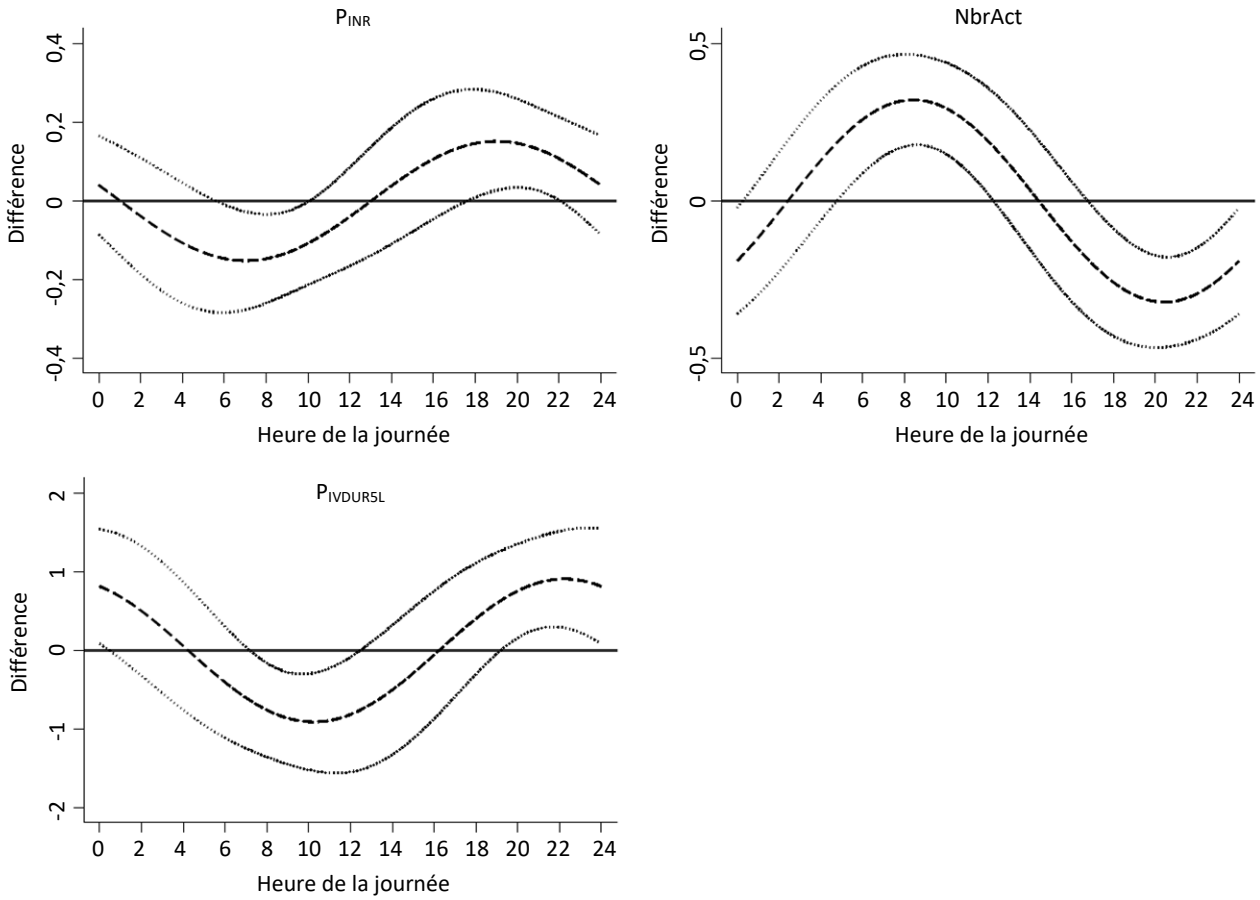
	(1)		(2)		(3)		(4)		(5)		(6)		(7)	
	P_{MOOD10}		P_{MOOD25}		P_{MOOD50}		$P_{FOODAH50}$		$P_{FOODAH100}$		$P_{EATING-OUT50}$		$P_{EATING-OUT100}$	
Variables explicatives	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**
12 h à 18 h 59					-0,109	0,500					-1,594	1,108		
19 h à 5 h 59					-0,023	0,675					-1,740	1,446		
12 h à 16 h 59			0,589	0,589										
17 h à 23 h 59			-0,502	0,733										
$D_{it}(1-(D_{it}/24)^2)$													0,839	0,449
$D_{it}^2(1-D_{it}/24)$													-0,032	0,045
$\sin(D_{it}\times 2\pi/24)$	-0,714	0,828					-0,277	0,898	-0,280	0,822				
$\cos(D_{it}\times 2\pi/24)$	0,820	0,820					-1,598	0,837	-0,883	0,799				
Mardi	-0,531	1,210	0,176	0,770	-0,706	0,662	-1,212	1,276	0,833	1,239	-0,501	1,378	0,273	1,267
Mercredi	-1,326	1,361	-0,768	0,842	-1,661*	0,722	-1,028	1,433	0,720	1,389	1,332	1,552	3,122*	1,417
Jeudi	-0,220	1,358	0,374	0,867	-0,873	0,737	-1,294	1,413	-0,873	1,388	0,669	1,575	1,909	1,416
Vendredi	-0,444	1,277	0,243	0,826	-0,455	0,663	-1,570	1,374	0,163	1,338	1,603	1,485	2,055	1,353
Samedi	-0,162	1,394	1,094	0,885	-0,068	0,721	-0,116	1,476	0,282	1,447	0,380	1,554	0,971	1,424
Dimanche	-0,907	1,544	-0,408	0,920	-1,711*	0,776	-3,351*	1,569	-2,001	1,539	0,264	1,692	2,627	1,529
A travaillé	-2,307*	1,103	0,721	0,678	-0,160	0,573	-1,491	1,272	-2,048	1,238	-2,277	1,360	-0,664	1,201
Durée du sommeil	0,215	0,182	0,223	0,117	0,166	0,097	-0,378	0,194	-0,326	0,192	0,193	0,216	0,244	0,196
Signification de $\alpha(D_{it})$	[0,28]		[0,28]		[0,97]		[0,16]		[0,54]		[0,30]		[0,04]	
Exogénéité stricte de $\{D_{it}\}$	[0,87]		[0,91]		[0,42]		[0,25]		[0,55]		[0,27]		[0,55]	
Observations	25 083		25 083		25 083		23 957		23 957		20 874		20 874	

Notes : Voir les notes du tableau 5.2.

Le nombre d'activités pourrait être moindre lorsque le journal est rempli en soirée en raison de la plus longue période de remémoration. Pour dissocier l'effet de D_{it} de celui de la période de remémoration, l'échantillon est divisé en journaux des jours de la semaine (lundi au jeudi) et des jours de la fin de semaine (vendredi au dimanche). Les résultats d'une nouvelle estimation de l'équation pour NbrAct dans chacun des deux sous-échantillons de journaux sont présentés au tableau 5.4. (N'oubliez pas que le jour indiqué dans les tableaux correspond au jour de l'interview.) $\alpha(D_{it})$ devient non significatif dans le sous-échantillon des journaux des jours de la fin de semaine, même si cette conclusion découle en partie de l'imprécision des estimations. De plus, l'étendue du changement dans un jour de fin de semaine se fait bien moins imposante que celle dans un jour de semaine : 0,43 activité par rapport à 1,01 activité, ce qui représente 2,7 % et 5,9 % de la moyenne correspondante. Par conséquent, une grande étendue du rythme quotidien de NbrAct est principalement attribuable à la période de remémoration.

Comme pour les effets sur les variables de contrôle, le nombre d'activités est plus élevé dans les journaux du lundi au jeudi, et les interviews semblent plus longues les mardis, les mercredis et les jeudis. Le fait de travailler et de dormir plus longtemps le jour réservé au journal a des effets contradictoires sur la qualité des données du journal sur l'emploi du temps, car cela a tendance à réduire le nombre d'activités et le nombre d'heures non codées. Ces effets s'expliquent probablement par le fait que travailler et dormir plus longtemps réduit le temps disponible pour d'autres activités, et la réduction des activités facilite leur remémoration. Travailler le jour réservé au journal accroît de 1,4 point de pourcentage (ou 28 %) la probabilité que l'interview se situe dans les 5 % des interviews les plus courtes.

Figure 5.1 Effets de l'heure de la journée de l'interview sur la qualité des données.



Notes : Les effets (lignes à tirets) sont calculés à partir des estimations correspondantes du tableau 5.2. Les lignes pointillées délimitent l'intervalle de confiance à 95 %.

Tableau 5.4
Effets de l'heure de la journée de l'interview sur NbrAct selon le jour du journal.

	(1) Journaux du lundi au jeudi		(2) Journaux du vendredi au dimanche	
Variables explicatives	**Coef.**	**ET**	**Coef.**	**ET**
$\sin(D_{it} \times 2\pi/24)$	0,476*	0,110	0,213	0,218
$\cos(D_{it} \times 2\pi/24)$	-0,171	0,112	-0,028	0,209
Lundi			Ref.	
Mardi	0,236	0,165		
Mercredi	0,099	0,172		
Jeudi	0,088	0,165		
Vendredi	Ref.			
Samedi			0,238	0,263
Dimanche			-0,415	0,304
A travaillé	-1,626*	0,213	-1,374*	0,345
Durée du sommeil	-0,093*	0,030	-0,021	0,057
Signification de $\alpha(D_{it})$	[0,00]		[0,59]	
Observations	14 904		3 073	

Notes : Les estimations sont réalisées à l'aide de la première différenciation et comprennent des ensembles complets de variables nominales des différences premières pour un certain nombre d'interviews précédentes et de variables nominales pour une seule heure des différences premières pour D_{it} en interaction avec MSF^{e}_{sc}. Les erreurs-types prennent en compte l'hétéroscédasticité et la mise en grappes au niveau individuel. Les valeurs de probabilité sont présentées entre crochets. * : significatif à 5 %.

5.3 Analyses complémentaires

5.3.1 Exogénéité stricte

Nous avons présumé que la variation dans D_{it} chez les répondants est strictement exogène. Cette hypothèse serait remise en question si, par exemple, des répondants s'empressaient de remplir le questionnaire ou devenaient distraits aux heures de la journée où le coût de renonciation pour achever l'interview serait le plus élevé. La valeur de p pour le test d'exogénéité stricte de $\{D_{it} : t = 1, \ldots, T_i\}$ est présentée à l'avant-dernière ligne des tableaux 5.2 et 5.3. Au niveau de 5 %, l'exogénéité est remise en question dans les régressions pour P_{INR}, le temps d'achèvement (IvDur) et $P_{IVDUR5L}$. Puisque l'estimateur à effets fixes tend à être plus résistant au non-respect de l'exogénéité stricte, nous révisons l'estimation de l'équation (4.1) à l'aide de l'estimateur des moindres carrés ordinaires tirés de la régression

$$y_{it} - \bar{y}_i = \left(\alpha\left(D_{it}\right) - \overline{\alpha\left(D_i\right)}\right) + \left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i\right)\boldsymbol{\beta} + u_{it} - \bar{u}_i \tag{5.1}$$

où $\bar{y}_i = T_i^{-1}\sum_{t=1}^{T_i} y_{it}$, $\overline{\alpha\left(D_i\right)} = T_i^{-1}\sum_{t=1}^{T_i}\alpha\left(D_{it}\right)$ et ainsi de suite. L'hypothèse nulle $H_0 : \alpha\left(D_{it}\right) = 0$ est rejetée dans la régression pour P_{INR} (valeur de p de 0,01), mais pas dans les régressions pour IvDur et $P_{IVDUR5L}$ (valeur de p de 0,39 dans les deux cas). Prenez toutefois note que l'estimateur par différence première et l'estimateur à effets fixes pourraient présenter des biais en cas d'échec de l'exogénéité stricte.

5.3.2 Robustesse

Les estimations changent peu lorsque la durée du sommeil est exclue de $\mathbf{x}_{it}$, ou que D_{it} est approximé selon l'heure de fin de la section sur l'emploi du temps du questionnaire ou selon des points sélectionnés aléatoirement dans les heures de début et de fin de l'interview (résultats non présentés). Quand les 2 578 interviews présentant des heures de coucher incohérentes sont comprises dans l'échantillon, le modèle privilégié pour $\alpha\left(D_{it}\right)$ change dans certains cas (tableau A.1 de l'annexe). Une valeur de $\alpha\left(D_{it}\right)$ significative sur le plan statistique est décelée dans les régressions pour P_{INR}, NbrAct, HManquante et IvDur, tandis que $\alpha\left(D_{it}\right)$ perd toute signification dans la régression pour $P_{IVDUR5L}$ (tableaux A.2 et A.3 de l'annexe). Lorsqu'un effet est décelé, cela indique que la qualité des données atteint un sommet tôt le matin.

5.3.3 Attrition

Le tableau 5.5 présente les données d'une estimation par probit ayant trait à la décision d'achever une interview. Il montre des coefficients $\boldsymbol{\delta}$ sélectionnés et les effets marginaux moyens (EMM) calculés en faisant la moyenne des effets marginaux dans les différentes observations. L'achèvement de l'interview $t-1$ entre le mardi et le samedi accroît la probabilité d'achever l'interview t. Travailler le jour réservé au journal accroît de 1,8 point de pourcentage cette probabilité, tandis que dormir une heure de plus la réduit de 0,6 point de pourcentage. Le nombre de semaines écoulées entre $t-2$ et $t-1$ est un prédicteur important de l'achèvement de l'interview t, dont la probabilité est réduite de 3,0 points de pourcentage chaque semaine écoulée. Aucune des variables nominales pour une seule heure liées à $D_{i,t-1}$ n'atteint le seuil de signification de 5 % (non illustré).

Après correction pour prendre en compte l'attrition non aléatoire, le modèle cosinor devient l'option privilégiée pour analyser P_{MOOD50}, alors que la fonction constante par morceaux de Binder (2022) se révèle la solution de rechange convenant le mieux pour la probabilité de déclarer des dépenses en nourriture à la maison en multiples de 100 ($P_{FOODAH100}$). L'hypothèse nulle de l'absence de biais d'attrition est remise en question dans les régressions pour IvDur, $P_{IVDUR5L}$ et $P_{FOODAH100}$. Toutefois, les estimations corrigées pour tenir compte de l'attrition (présentées aux tableaux A.4 et A.5 de l'annexe) révèlent essentiellement les mêmes schémas que celles qui n'ont pas été corrigées pour tenir compte de l'attrition. La correction pour tenir compte de l'attrition non aléatoire rend moins contestable l'hypothèse d'exogénéité stricte de $\alpha(D_{it})$ dans les régressions pour IvDur et $P_{IVDUR5L}$ (valeur de p de 0,12 dans les deux cas).

Tableau 5.5
Probit pour l'achèvement d'une interview.

	Variable dépendante : s_{it}, $t \geq 2$			
Variables explicatives $(t-1)$	**Coef.**	**ET**	**EMM**	**ET**
Mardi	0,154*	0,034	0,034*	0,008
Mercredi	0,175*	0,038	0,039*	0,008
Jeudi	0,184*	0,038	0,040*	0,008
Vendredi	0,215*	0,037	0,047*	0,008
Samedi	0,174*	0,039	0,038*	0,009
Dimanche	0,058	0,042	0,014	0,010
A travaillé	0,089*	0,028	0,018*	0,005
Durée du sommeil	-0,027*	0,004	-0,006*	0,001
Semaines entre $t-2$ et $t-1$	-0,145*	0,006	-0,030*	0,001
Ordonnée à l'origine	1,121*	0,105		
R au carré		0,070		
Observations		32 779		
Moyenne de s_{it}		0,859		

Notes : Les observations relatives à la dernière interview sont exclues parce que les personnes n'ont certainement pas poursuivi l'enquête. Comprend des variables nominales pour une seule heure relatives à $D_{i,t-1}$, des variables nominales pour le nombre d'interviews précédentes et l'interaction de MSF_{sc}^{e} avec les variables nominales pour une seule heure relatives à $D_{i,t-1}$. Les erreurs-types prennent en compte l'hétéroscédasticité et la mise en grappes au niveau individuel. *R* au carré est égal à un moins le logarithme du rapport de vraisemblance de la fonction ajustée par rapport au logarithme du rapport de vraisemblance d'une fonction ayant seulement une ordonnée à l'origine. * : significatif à 5 %.

5.3.4 Poids

Les tableaux 5.6 et 5.7 présentent les estimations des moindres carrés pondérés. Aucune valeur de $\alpha(D_{it})$ significative sur le plan statistique n'est décelée dans la majorité des régressions illustrées. Même si dans certains cas (par exemple la régression pour NbrAct), les coefficients estimés des moindres carrés pondérés se révèlent plus petits que ceux des moindres carrés ordinaires, l'inférence découle d'erreurs-types plus importantes dans la majorité des cas. Une valeur de $\alpha(D_{it})$ significative sur le plan statistique est décelée dans la régression pour $P_{IVDUR5H}$ (valeur de p de 0,03), mais cet effet ne survit pas à une correction de Bonferroni pour des tests simultanés dans le groupe de mesures qui évaluent le temps d'achèvement. L'hypothèse nulle de l'absence d'effet est aussi rejetée à 5 % dans les régressions pour $P_{EATING-OUT50}$ et $P_{EATING-OUT100}$ (valeur de p de 0,01 dans les deux cas), mais le rejet de l'hypothèse nulle ne tient dans aucun des deux cas si une absence de dépenses est censée ne pas correspondre à un arrondissement (valeurs de p de 0,43 et de 0,35, respectivement).

5.3.5 Sous-populations

Pour terminer, nous avons divisé l'échantillon par niveau de scolarité (au plus des études collégiales partielles par rapport à un diplôme d'études collégiales) afin d'examiner les effets de l'heure de la journée sur l'interview de certains types de personnes. Bien que les capacités cognitives soient d'importants prédicteurs du niveau de scolarité, nous ne nous attendons pas à percevoir d'énormes différences entre les groupes démographiques, puisque nos estimations sont nettes d'effets de synchronie et d'effets relatifs aux capacités cognitives. En fait, même si le modèle de $\alpha(D_{it})$ convenant le mieux change pour la majorité des variables dépendantes dans les deux sous-populations, les principales constatations demeurent (résultats non présentés).

Tableau 5.6
Effets de l'heure de la journée de l'interview sur la qualité des données. Estimations pondérées.

	(1) PINR		(2) NbrAct		(3) HManquante		(4) IvDur (min)		(5) PIVDUR5L		(6) PIVDUR5H	
Variables explicatives	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**
12 h à 18 h 59					0,224	0,124						
19 h à 5 h 59					0,063	0,113						
$D_{it}(1-(D_{it}/24)^2)$											-0,993*	0,474
$D_{it}^2(1-D_{it}/24)$											0,153*	0,058
$\sin(D_{it}\times 2\pi/24)$	-0,298	0,249	0,061	0,207			-0,581*	0,295	0,363	0,662		
$\cos(D_{it}\times 2\pi/24)$	0,133	0,139	-0,111	0,157			-0,314	0,293	0,082	1,023		
Mardi	-0,241	0,242	1,222*	0,260	-0,133	0,093	1,341*	0,374	-0,299	1,306	1,439	1,417
Mercredi	-0,124	0,334	1,248*	0,359	-0,185	0,138	1,028*	0,438	0,077	2,241	0,812	1,507
Jeudi	0,483	0,331	0,881*	0,312	0,020	0,098	1,025*	0,409	0,543	1,653	1,879	1,619
Vendredi	0,512	0,336	0,703*	0,291	0,104	0,127	0,621	0,431	1,067	1,422	0,262	1,533
Samedi	0,354	0,291	-0,009	0,411	0,038	0,128	-0,243	0,483	1,911	2,031	-1,101	1,531
Dimanche	-0,034	0,358	0,109	0,339	-0,075	0,130	-0,959*	0,485	0,381	1,556	-3,991*	1,663
A travaillé	0,200	0,250	-2,582*	0,260	-0,158	0,084	-0,706*	0,320	3,005*	1,494	-1,619	1,025
Durée du sommeil	-0,057	0,043	-0,077	0,046	-0,030	0,017	-0,097	0,062	0,656*	0,209	0,040	0,276
Signification de $\alpha(D_{it})$	[0,38]		[0,69]		[0,14]		[0,14]		[0,84]		[0,03]	
Exogénéité stricte de $\{D_{it}\}$	[0,26]		[0,92]		[0,36]		[0,02]		[0,25]		[0,23]	
Observations	25 184		25 184		25 184		25 184		25 184		25 184	

Notes : Les estimations sont réalisées à l'aide de la première différenciation et comprennent des ensembles complets de variables nominales des différences premières pour un certain nombre d'interviews précédentes et de variables nominales pour une seule heure des différences premières pour D_{it} en interaction avec MSF_{sc}^{e}. Les variables dépendantes dont le nom commence par P sont des indicateurs binaires pour le résultat inscrit dans l'indice du nom, exprimé sous forme de pourcentage. Les erreurs-types prennent en compte l'hétéroscédasticité et la mise en grappes au niveau individuel. Les valeurs de probabilité sont présentées entre crochets. * : significatif à 5 %.

Tableau 5.7
Effets de l'heure de la journée de l'interview sur la qualité des données. Estimations pondérées.

	(1)		(2)		(3)		(4)		(5)		(6)		(7)	
	P_{MOOD10}		P_{MOOD25}		P_{MOOD50}		$P_{FOODAH50}$		$P_{FOODAH100}$		$P_{EATING-OUT50}$		$P_{EATING-OUT100}$	
Variables explicatives	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**
12 h à 18 h 59					-1,514	1,575					-1,255	2,295		
19 h à 5 h 59					0,127	1,500					-8,426*	3,040		
12 h à 16 h 59			-1,515	1,582										
17 h à 23 h 59			-3,035*	1,436										
$D_{it}(1-(D_{it}/24)^2)$													1,074	0,889
$D_{it}^2(1-D_{it}/24)$													0,037	0,096
$\sin(D_{it}\times 2\pi/24)$	-1,143	1,758					-2,348	2,000	-2,029	2,399				
$\cos(D_{it}\times 2\pi/24)$	0,236	2,192					-4,367	2,250	-4,183	2,698				
Mardi	4,451	2,827	0,293	1,508	-0,836	1,199	2,645	2,530	3,310	2,620	-5,478	2,900	-2,843	2,732
Mercredi	1,520	2,894	-0,776	1,728	-0,807	1,502	1,089	3,120	0,796	3,130	-0,336	3,387	3,425	3,026
Jeudi	5,568	3,929	0,631	1,738	-0,993	1,292	3,577	2,871	2,227	3,014	0,251	4,177	1,842	3,670
Vendredi	7,726*	3,824	-0,968	1,819	-1,234	1,152	2,891	3,447	3,313	3,682	4,003	3,041	2,739	2,822
Samedi	3,580	3,349	0,399	1,659	-0,234	1,321	3,190	3,127	3,793	3,692	-0,065	3,192	1,163	3,009
Dimanche	5,240	3,578	-0,992	1,842	-2,572	1,450	2,211	3,327	0,165	3,645	-1,928	3,554	0,044	3,236
A travaillé	-3,964	2,072	0,514	1,111	-1,931*	0,918	-0,547	2,809	-0,775	2,856	-3,397	2,839	-2,032	2,383
Durée du sommeil	0,282	0,361	0,174	0,219	0,149	0,178	-0,855*	0,368	-0,576	0,440	0,330	0,441	0,280	0,432
Signification de $\alpha(D_{it})$	[0,77]		[0,11]		[0,35]		[0,09]		[0,28]		[0,01]		[0,01]	
Exogénéité stricte de $\{D_{it}\}$	[0,60]		[0,68]		[0,97]		[0,13]		[0,34]		[0,41]		[0,92]	
Observations	25 083		25 083		25 083		23 957		23 957		20 874		20 874	

Notes : Voir les notes du tableau 5.6.

6. Sommaire et discussion

L'analyse des microdonnées longitudinales à haute fréquence tirées de l'enquête SUWNJ ne révèle aucune donnée probante indiquant un effet de l'heure de la journée de l'interview sur la qualité des données du journal sur l'emploi du temps (au-delà de l'effet exercé par la longueur de la période de remémoration) ou sur la tendance à déclarer des valeurs arrondies de probabilités subjectives ou de dépenses pour la nourriture. En ce qui concerne la période de remémoration, nous avons constaté que le fait de remplir le journal de la veille en soirée réduisait le nombre d'activités déclarées, tandis que la quantité d'heures non codées ne subit aucune fluctuation quotidienne significative. Par conséquent, il semble que certaines activités sont sous-déclarées et que la durée de certaines autres est surestimée, ce qui introduit une erreur dans la mesure de l'emploi du temps. Toutes ces constatations tiennent compte des différences inter-personnelles en matière de capacité cognitive et d'effets de synchronie, ce qui pourrait expliquer la raison pour laquelle elles persistent entre les groupes de niveau de scolarité. Elles semblent également robustes pour une variété d'autres spécifications évaluant les répercussions d'une attrition non aléatoire, les effets hétérogènes non modélisés et les mesures différentes de l'heure de la journée de l'interview. Malgré l'existence de données probantes indiquant que la non-réponse partielle et la probabilité que la durée de l'interview fasse partie des 5 % des interviews les plus courtes augmentent lorsque le questionnaire est rempli en soirée, une évaluation plus détaillée nécessite le recours à des variables instrumentales.

Nos résultats les plus fiables étayent la conclusion des recherches antérieures voulant que la qualité des données d'enquête ne soit pas sensible à l'heure de la journée de l'interview (Ziniel, 2008; Dickinson et McElroy, 2010; Binder, 2022), mais contredisent celle de Flynn (2018), qui a constaté que les répondants qui commencent à remplir un questionnaire en soirée répondent à nettement plus de questions que ceux qui commencent à le remplir en matinée ou en après-midi. Cependant, l'échantillon de Flynn (2018) se compose de représentants de sociétés pour qui remplir un questionnaire en dehors des heures normales de travail pourrait permettre de réduire les contraintes de temps. Puisque les chômeurs (comparativement aux personnes occupées) n'ont pas à respecter de limites quant à des heures de travail, l'heure de la journée de l'interview peut se répartir de manière plus homogène sur les 24 heures, ce qui permet de déceler plus facilement les effets relatifs au moment de la journée. Il convient également de souligner que, contrairement aux échantillons de participants de MTurk (par exemple Binder, 2022), les interviews semblent plus longues les jeudis (ainsi que les mardis et les mercredis) et que le nombre d'activités déclarées est plus élevé dans les journaux du lundi au jeudi, comme dans Juster (1986).

Dans l'ensemble, il semble donc qu'outre l'effet exercé par la longueur de la période de remémoration, inciter des répondants à remplir des questionnaires à des heures précises de la journée pourrait avoir des répercussions limitées sur l'erreur de mesure. Par conséquent, les spécialistes des enquêtes ne devraient pas trop se soucier des conséquences qu'aurait sur l'erreur de mesure le fait de communiquer avec les sujets de l'interview au moment de la journée où ils sont les plus susceptibles d'être joignables.

Cela étant dit, nous reconnaissons que la présente étude comporte certaines limites. En ce qui concerne la question de savoir si nous constatons des effets de causalité pour la population à l'étude, il faut souligner que nous ne disposons pas de suffisamment de données sur le contexte situationnel dans lequel les interviews ont été réalisées (par exemple où se trouvait le répondant et ce qu'il faisait), et comme l'avancent Bison et Zhao (2023), les contextes situationnel et temporel pourraient être corrélés. Il est cependant difficile de proposer des variables instrumentales suffisamment corrélées avec l'heure de la journée de l'interview sans être corrélées avec des erreurs idiosyncrasiques, puisque la majorité des variables de l'enquête SUWNJ font référence à des jours autres que celui de l'interview. De plus, même si le pourcentage d'interviews de l'enquête SUWNJ réalisées à l'aide d'un appareil mobile était sans doute faible (Callegaro, 2010, rapporte par exemple que parmi tous les répondants ayant tenté de remplir un questionnaire en ligne sur la satisfaction de la clientèle, sondage réalisé en Amérique du Nord en juin 2010, 2,6 % ont effectivement utilisé un appareil mobile), si le fait d'utiliser un appareil mobile pour répondre à une interview nuit à la qualité des données (comme le laissent croire les données probantes examinées dans Toninelli et Revilla, 2020) et dépend du moment de la journée, nos résultats pourraient renfermer un biais. En ce qui a trait à la valeur prédictive de nos constatations dans un contexte différent, il convient de souligner que les résultats obtenus pour les chômeurs pourraient ne pas se révéler représentatifs des populations en général si, par exemple, les activités menées avant de prendre part à l'enquête interagissent avec le sommeil ou la fatigue.

De plus, en raison d'un manque de données, nous n'avons pas pu nous pencher sur la question de l'existence d'effets de l'heure de la journée de l'interview sur des mesures de rechange de la qualité des données, comme l'interruption de l'enquête et les erreurs de réponse causées par la désirabilité sociale ou les réponses extrêmes, médianes ou non différenciées. Pour ce qui est des effets de la longueur de la période de remémoration, il semble valoir la peine d'examiner si la soumission d'un journal de la veille par un intervieweur (qui pourrait stimuler l'attention et la motivation des répondants) ou la déclaration dans « leurs propres mots » des activités des répondants (ce qui évite de devoir mettre la réponse en correspondance avec l'option de réponse appropriée) pourrait rehausser la qualité des données de journaux sur l'emploi du temps.

Remerciements

Le présent article a grandement profité des commentaires et des suggestions de Jean-François Beaumont, de Kristen Olson et, spécialement, d'un rédacteur associé anonyme et de plusieurs réviseurs anonymes. Je remercie également Andreas Mueller pour son aide avec les données de l'enquête SUWNJ. Cette étude a été financée par le Gouvernement d'Aragon, subvention S32-23R.

Annexe

Tableau A.1
Modèle sélectionné pour $\alpha(D_{it})$. Comprend des observations présentant des heures de coucher incohérentes.

Variable dépendante	Modèle	Valeur du BIC
P_{INR}	Fonction constante par morceaux (Binder, 2022)	13,405
NbrAct	Cosinor	13,731
HManquante	Fonction constante par morceaux (Binder, 2022)	11,271
IvDur	Fonction constante par morceaux (Binder, 2022)	14,460
$P_{IVDUR5L}$	Cosinor	16,590
$P_{IVDUR5H}$	Fonction polynomiale du troisième degré	16,849
P_{MOOD10}	Fonction constante par morceaux (Binder, 2022)	18,162
P_{MOOD25}	Fonction constante par morceaux (Durrant et coll., 2011)	17,211
P_{MOOD50}	Cosinor	16,820
$P_{FOODAH50}$	Fonction polynomiale du troisième degré	18,226
$P_{FOODAH100}$	Fonction constante par morceaux (Binder, 2022)	18,145
$P_{EATING\text{-}OUT50}$	Fonction constante par morceaux (Binder, 2022)	18,080
$P_{EATING\text{-}OUT100}$	Fonction polynomiale du troisième degré	17,887

Tableau A.2
Effets de l'heure de la journée de l'interview sur la qualité des données. Comprend des observations présentant des heures de coucher incohérentes.

	(1)		(2)		(3)		(4)		(5)		(6)		(7)	
	P_{INR}		NbrAct		HManquante		IvDur (min)		$P_{IVDUR5L}$		$P_{IVDUR5H}$		P_{PM-AM}	
Variables explicatives	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**
12 h à 18 h 59	0,203*	0,073			0,070*	0,027	-0,368*	0,140						
19 h à 5 h 59	0,221*	0,101			0,054	0,036	-0,461*	0,189						
$D_{it}(1-(D_{it}/24)^2)$											-0,106	0,219		
$D_{it}^2(1-D_{it}/24)$											0,027	0,022		
$\sin(D_{it}\times 2\pi/24)$			0,264*	0,084					-0,099	0,355			0,611	0,499
$\cos(D_{it}\times 2\pi/24)$			-0,184*	0,080					0,427	0,333			-1,069*	0,495
Mardi	-0,177	0,101	1,420*	0,131	-0,086*	0,037	0,772*	0,184	-0,160	0,447	1,114	0,602	-2,738*	0,696
Mercredi	-0,145	0,109	1,180*	0,141	-0,056	0,039	0,434*	0,196	-0,672	0,502	0,075	0,635	-2,417*	0,763
Jeudi	0,024	0,121	1,247*	0,150	-0,011	0,041	0,780*	0,213	-0,469	0,519	1,255	0,724	-1,503	0,788
Vendredi	-0,009	0,113	1,091*	0,145	-0,057	0,042	0,355	0,196	0,095	0,476	0,241	0,626	-2,161*	0,748
Samedi	0,072	0,121	0,755*	0,147	-0,002	0,041	0,507*	0,233	0,838	0,561	0,482	0,755	-1,156	0,815
Dimanche	0,270*	0,136	-0,025	0,163	0,029	0,046	-0,283	0,228	-0,199	0,613	-0,538	0,728	-0,759	0,879
A travaillé	0,369*	0,129	-1,687*	0,152	-0,123*	0,035	-0,206	0,164	1,401*	0,565	0,454	0,531	-2,210*	0,659
Durée du sommeil	-0,028	0,018	-0,056*	0,023	-0,015*	0,007	-0,105*	0,026	0,361*	0,082	-0,131	0,090	-1,803*	0,124
Signification de $\alpha(D_{it})$	[0,01]		[0,00]		[0,03]		[0,01]		[0,35]		[0,30]		[0,02]	
Exogénéité stricte de $\{D_{it}\}$	[0,02]		[0,61]		[0,17]		[0,57]		[0,01]		[0,64]		[0,65]	
Observations	28 576		28 576		28 576		28 576		28 576		28 576		28 576	

Notes : Les estimations sont réalisées à l'aide de la première différenciation et comprennent des ensembles complets de variables nominales des différences premières pour un certain nombre d'interviews précédentes et de variables nominales pour une seule heure des différences premières pour D_{it} en interaction avec MSF_{sc}^{e}. Les variables dépendantes dont le nom commence par P sont des indicateurs binaires pour le résultat inscrit dans l'indice du nom, exprimé sous forme de pourcentage. Les erreurs-types prennent en compte l'hétéroscédasticité et la mise en grappes au niveau individuel. Les valeurs de probabilité sont présentées entre crochets. * : significatif à 5 %.

Tableau A.3
Effets de l'heure de la journée de l'interview sur la qualité des données. Comprend des observations présentant des heures de coucher incohérentes.

	(1)		(2)		(3)		(4)		(5)		(6)		(7)	
	P_{MOOD10}		P_{MOOD25}		P_{MOOD50}		$P_{FOODAH50}$		$P_{FOODAH100}$		$P_{EATING-OUT50}$		$P_{EATING-OUT100}$	
Variables explicatives	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**
12 h à 18 h 59	1,065	0,895							-0,600	0,911	-1,238	1,043		
19 h à 5 h 59	2,380*	1,197							-0,187	1,204	-0,950	1,373		
12 h à 16 h 59			0,525	0,573										
17 h à 23 h 59			0,206	0,686										
$D_{it}(1-(D_{it}/24)^2)$							0,283	0,451					0,578	0,411
$D_{it}^2(1-D_{it}/24)$							0,022	0,047					-0,022	0,041
$\sin(D_{it}\times 2\pi/24)$					0,111	0,440								
$\cos(D_{it}\times 2\pi/24)$					0,511	0,374								
Mardi	0,208	1,142	-0,282	0,723	-0,864	0,619	-0,697	1,205	0,729	1,168	-0,468	1,312	0,398	1,194
Mercredi	-0,735	1,277	-0,951	0,791	-1,703*	0,672	-0,453	1,363	0,947	1,320	1,463	1,471	3,295*	1,339
Jeudi	-0,176	1,270	0,191	0,814	-0,966	0,692	-1,183	1,351	-1,337	1,305	-0,120	1,473	1,251	1,322
Vendredi	0,034	1,207	-0,163	0,782	-0,636	0,623	-1,346	1,313	0,104	1,266	1,540	1,405	2,252	1,279
Samedi	0,078	1,300	0,353	0,840	-0,369	0,677	-0,448	1,392	-0,298	1,352	0,382	1,471	1,028	1,342
Dimanche	-0,489	1,440	-0,529	0,876	-1,590*	0,727	-2,606	1,462	-1,390	1,451	0,611	1,607	2,820	1,441
A travaillé	-2,167*	1,067	0,377	0,659	-0,363	0,542	-1,344	1,200	-2,024	1,162	-2,238	1,277	-0,372	1,147
Durée du sommeil	0,152	0,171	0,200	0,110	0,115	0,095	-0,401*	0,179	-0,336	0,178	0,139	0,204	0,190	0,187
Signification de $\alpha(D_{it})$	[0,13]		[0,65]		[0,39]		[0,10]		[0,79]		[0,49]		[0,17]	
Exogénéité stricte de $\{D_{it}\}$	[0,64]		[0,99]		[0,79]		[0,25]		[0,25]		[0,42]		[0,93]	
Observations	28 461		28 461		28 461		27 187		27 187		23 612		23 612	

Notes : Voir les notes du tableau A.2.

Tableau A.4
Effets de l'heure de la journée de l'interview sur la qualité des données. Estimations corrigées pour prendre en compte l'attrition.

	(1) PINR		(2) NbrAct		(3) HManquante		(4) IvDur (min)		(5) PIVDUR5L		(6) PIVDUR5H	
Variables explicatives	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**
12 h à 18 h 59					0,048	0,029						
19 h à 5 h 59					0,046	0,039						
$D_{it}(1-(D_{it}/24)^2)$											-0,157	0,238
$D_{it}^2(1-D_{it}/24)$											0,038	0,024
$\sin(D_{it}\times 2\pi/24)$	-0,143*	0,070	0,256*	0,087			0,026	0,135	-0,457	0,364		
$\cos(D_{it}\times 2\pi/24)$	0,041	0,064	-0,191*	0,086			-0,293*	0,138	0,827*	0,374		
Mardi	-0,175	0,113	1,475*	0,142	-0,096*	0,040	0,944*	0,202	-0,505	0,501	1,100	0,658
Mercredi	-0,147	0,121	1,209*	0,154	-0,074	0,043	0,645*	0,211	-1,195*	0,559	0,048	0,680
Jeudi	0,033	0,136	1,231*	0,163	-0,040	0,044	0,858*	0,231	-0,681	0,576	1,243	0,774
Vendredi	0,003	0,123	1,130*	0,155	-0,070	0,044	0,470*	0,212	-0,052	0,512	-0,055	0,673
Samedi	0,023	0,127	0,755*	0,155	-0,022	0,042	0,542*	0,249	0,319	0,610	0,187	0,795
Dimanche	0,289*	0,140	0,058	0,172	0,003	0,045	-0,336	0,242	0,121	0,661	-0,896	0,762
A travaillé	0,414*	0,137	-1,720*	0,158	-0,118*	0,037	-0,117	0,174	1,282*	0,608	0,717	0,556
Durée du sommeil	-0,025	0,018	-0,079*	0,025	-0,016*	0,007	-0,142*	0,028	0,416*	0,088	-0,188*	0,095
Biais d'attrition	[0,91]		[0,20]		[0,29]		[0,00]		[0,01]		[0,38]	
Signification de $\alpha(D_{it})$	[0,04]		[0,00]		[0,24]		[0,07]		[0,01]		[0,16]	
Exogénéité stricte de $\{D_{it}\}$	[0,00]		[0,54]		[0,08]		[0,12]		[0,12]		[0,95]	
Observations	25 184		25 184		25 184		25 184		25 184		25 184	

Notes : Les estimations sont réalisées à l'aide de la première différenciation et comprennent un ensemble complet de variables nominales des différences premières pour un certain nombre d'interviews précédentes et de variables nominales pour une seule heure des différences premières pour D_{it} en interaction avec MSF_{sc}^{e}, ainsi que l'inverse du rapport de Mills en interaction avec des valeurs nominales pour le nombre d'interviews. Les variables dépendantes dont le nom commence par P sont des indicateurs binaires pour le résultat inscrit dans l'indice du nom, exprimé sous forme de pourcentage. Les erreurs-types prennent en compte l'hétéroscédasticité et la mise en grappes au niveau individuel, en plus d'apporter une correction pour prendre en compte les variables explicatives générées. Les valeurs de probabilité sont présentées entre crochets. * : significatif à 5 %.

Tableau A.5
Effets de l'heure de la journée de l'interview sur la qualité des données. Estimations corrigées pour prendre en compte l'attrition.

	(1) PMOOD10		(2) PMOOD25		(3) PMOOD50		(4) PFOODAH50		(5) PFOODAH100		(6) PEATING-OUT50		(7) PEATING-OUT100	
Variables explicatives	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**	**Coef.**	**ET**
12 h à 18 h 59									-1,027	0,994	-1,642	1,113		
19 h à 5 h 59									-1,107	1,281	-1,824	1,446		
12 h à 16 h 59			0,621	0,589										
17 h à 23 h 59			-0,467	0,734										
$D_{it}(1-(D_{it}/24)^2)$													0,836	0,448
$D_{it}^2(1-D_{it}/24)$													-0,031	0,045
$\sin(D_{it}\times 2\pi/24)$	-0,741	0,831			-0,039	0,458	-0,316	0,899						
$\cos(D_{it}\times 2\pi/24)$	0,827	0,820			0,084	0,400	-1,588	0,836						
Mardi	-0,758	1,250	0,140	0,803	-0,844	0,685	-1,400	1,328	0,792	1,290	-0,348	1,429	0,201	1,304
Mercredi	-1,539	1,394	-0,823	0,874	-1,789*	0,744	-1,175	1,486	0,745	1,443	1,505	1,609	3,053*	1,459
Jeudi	-0,437	1,387	0,324	0,888	-1,011	0,749	-1,447	1,445	-0,919	1,429	0,822	1,620	1,867	1,451
Vendredi	-0,621	1,305	0,183	0,850	-0,570	0,679	-1,740	1,416	0,180	1,379	1,759	1,534	2,022	1,389
Samedi	-0,342	1,409	1,065	0,903	-0,162	0,733	-0,280	1,501	0,240	1,475	0,494	1,589	0,907	1,446
Dimanche	-0,976	1,544	-0,396	0,926	-1,741*	0,779	-3,374*	1,575	-2,103	1,543	0,286	1,702	2,611	1,533
A travaillé	-2,281*	1,106	0,653	0,684	-0,214	0,576	-1,511	1,277	-1,992	1,241	-2,132	1,369	-0,491	1,203
Durée du sommeil	0,221	0,184	0,232*	0,117	0,176	0,097	-0,366	0,195	-0,333	0,193	0,176	0,216	0,247	0,195
Biais d'attrition	[0,79]		[0,30]		[0,26]		[0,67]		[0,04]		[0,96]		[0,08]	
Signification de $\alpha(D_{it})$	[0,27]		[0,28]		[0,97]		[0,16]		[0,54]		[0,28]		[0,04]	
Exogénéité stricte de $\{D_{it}\}$	[0,85]		[0,98]		[0,68]		[0,25]		[0,61]		[0,33]		[0,50]	
Observations	25 083		25 083		25 083		23 957		23 957		20 874		20 874	

Notes : Voir les notes du tableau A.4.

Bibliographie

Ahn, J., Peng, M., Park, C. et Jeon, Y. (2012). A resampling approach for interval-valued data regression. *Statistical Analysis and Data Mining*, 5, 336-348.

American Association for Public Opinion Research (AAPOR) (2023). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys, 10th Edition.* AAPOR.

Angrisani, M., et Couper, M. (2022). A simple question goes a long way: A wording experiment on bank account ownership. *Journal of Survey Statistics and Methodology*, 10, 1172-1182.

Arechar, A., Kraft-Todd, G. et Rand, D. (2017). Turking overtime: how participant characteristics and behavior vary over time and day on Amazon Mechanical Turk. *Journal of the Economic Science Association*, 3, 1-11.

Arellano, M., et Meghir, C. (1992). Female labour supply and on-the-job search: An empirical model estimated using complementary data sets. *Review of Economic Studies*, 59, 537-557.

Bach, R. (2021). A methodological framework for the analysis of panel conditioning effects. Dans *Measurement Error in Longitudinal Data*, (Éds., Alexandru Cernat et Joseph Sakshaug), 19-42. Oxford: OUP.

Bais, F., Schouten, B. et Toepoel, V. (2022). Les comportements de réponse indésirables sont-ils constants d'une enquête à l'autre ? Un examen approfondi des caractéristiques de répondants. *Techniques d'enquête*, 48, 1, 209-243. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2022001/article/00001-fra.pdf.

Baumgartner, H., et Steenkamp, J.-B. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38(2), 143-156.

Bes, F., Jobert, M. et Schulz, H. (2009). Modeling napping, post-lunch dip, and other variations in human sleep propensity. *Sleep*, 32(3), 392-398.

Biemer, P., Groves, R., Lyberg, L. Mathiowetz, N. et Sudman, S. (éds.) (2004). *Measurement Errors in Surveys*. Hoboken, NJ: John Wiley & Sons, Inc.

Binder, C. (2022). Time-of-day and day-of-week variations in Amazon Mechanical Turk survey responses. *Journal of Macroeconomics*, 71, Article 103378.

Bison, I., et Zhao, H. (2023). Factors impacting the quality of user answers on smartphones. *CEUR Workshop Proceedings*, 3456, 208-213.

Blatter, K., et Cajochen, C. (2007). Circadian rhythms in cognitive performance: Methodological constraints, protocols, theoretical underpinnings. *Physiology and Behavior*, 90, 196-208.

Brown, E., et Czeisler, C. (1992). The statistical analysis of circadian phase and amplitude in constant-routine core-temperature data. *Journal of Biological Rhythms*, 7(3), 177-202.

Callegaro, M. (2010). Do you know which device your respondent has used to take your online survey? *Survey Practice*, 3(6). https://doi.org/10.29115/SP-2010-0028.

Carrell, S., Maghakian, T. et West, J. (2011). A's from zzzz's? The causal effect of school start time on the academic achievement of adolescents. *American Economic Journal: Economic Policy*, 3, 62-81.

Casey, L., Chandler, J., Levine, A., Proctor, A. et Strolovitch, D. (2017). Intertemporal differences among MTurk workers: Time-based sample variations and implications for online data collection. *SAGE Open*, 7(2), 1-15.

Chang, L., et Krosnick, J. (2009). National surveys via RDD telephone interviewing versus the Internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, 73(4), 641-678.

Collinson, J., Mathmann, F. et Chylinski, M. (2020). Time is money: Field evidence for the effect of time of day and product name on product purchase. *Journal of Retailing and Consumer Services*, 54, Article 102064.

Cornelissen, G. (2014). Cosinor-based rhythmometry. *Theoretical Biology and Medical Modelling*, 11, Article 16.

Davis, S. (2011). Commentaires sur : Krueger, A., et A. Mueller. Job search, emotional well-being and job finding in a period of mass unemployment: Evidence from high-frequency longitudinal data. *Brookings Papers on Economic Activity*, 42(1), 58-70.

Dickinson, D., et McElroy, T. (2010). Rationality around the clock: Sleep and time-of-day effects on guessing game responses. *Economics Letters*, 108, 245-248.

Dickinson, D., et McElroy, T. (2017). Sleep restriction and circadian effects on social decisions. *European Economic Review*, 97, 57-71.

Dickinson, D., Chaudhuri, A.·et Greenaway-McGrevy, R. (2020). Trading while sleepy? Circadian mismatch and mispricing in a global experimental asset market. *Experimental Economics*, 23, 526-553.

Durrant, G., D'Arrigo, J. et Steele, F. (2011). Using paradata to predict best times of contact, conditioning on household and interviewer influences. *Journal of the Royal Statistical Society. Series A*, 174(4), 1029-1049.

Flynn, A. (2018). e-Surveying and respondent behaviour: Insights from the public procurement field. *Electronic Journal of Business Research Methods*, 16(1), 38-53.

Fordsham, N., Moss, A., Krumholtz, S., Roggina, T., Robinson, J. et Litman, L. (2019). Variation among Mechanical Turk workers across time of day presents an opportunity and a challenge for research. PsyArXiv. doi:10.31234/osf.io/p8bns.

Fricker, S., et Tourangeau, R. (2010). Examining the relationship between nonresponse propensity and data quality in two national household surveys. *Public Opinion Quarterly*, 74(5), 934-955.

Gideon, M., Helppie-McFall, B. et Hsu, J. (2017). Heaping at round numbers on financial questions: The role of satisficing. *Survey Research Methods*, 11(2), 189-214.

González Chapela, J. (2024). Supplement to "Daily rhythm of data quality: Evidence from the survey of unemployed workers in New Jersey". Disponible à https://drive.google.com/file/d/14YPt9BmXlxFfuURatCQY0ak-OW9z1c7B/view?usp=sharing.

Guarana, C., Stevenson, R. Gish, J. Ryu, J.W. et Crawley, R. (2022). Owls, larks, or investment sharks? The role of circadian process in early-stage investment decisions. *Journal of Business Venturing*, 37, Article 106165.

Hasher, L., Goldstein, D. et May, C. (2005). It's about time: Circadian rhythms, memory, and aging. Dans *Human Learning and Memory: Advances in Theory and Applications*, (Éds., Chizuko Izawa et Nobuo Ohta), 199-217. New York: Lawrence Erlbaum Associates Publishers.

Hornik, J., et Tal, A. (2010). The effect of synchronizing consumers' diurnal preferences with time of response on data reliability. *Marketing Letters*, 21, 1-15.

Juster, T. (1986). Response errors in the measurement of time use. *Journal of the American Statistical Association*, 81(394), 390-402.

Kaminska, O., McCutcheon, A. et Billiet, J. (2010). Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly*, 74(5), 956-984.

Kroh, M., Winter, F. et Schupp, J. (2016). Using person-fit measures to assess the impact of panel conditioning on reliability. *Public Opinion Quarterly*, 80(4), 914-942.

Kroh, M., Lüdtke, D., Düzel, S. et Winter, F. (2016). Response error in a web survey and a mailed questionnaire: The role of cognitive functioning. SOEPpapers on Multidisciplinary Panel Data Research, No. 888. DIW, Berlin.

Krosnick, J. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213-236.

Krosnick, J. (1999). Survey research. *Annual Review of Psychology*, 50, 537-567.

Krueger, A., et Mueller, A. (2010). *Survey of Unemployed Workers in New Jersey* (version 12 novembre 2013) [Base de données]. Data Archive at the Office of Population Research, Princeton University. https://oprdata.princeton.edu/archive/njui/.

Krueger, A., et Mueller, A. (2011). Job search, emotional well-being and job finding in a period of mass unemployment: Evidence from high-frequency longitudinal data. *Brookings Papers on Economic Activity*, 42(1), 1-57.

Lowe, C., Safati, A., et Hall, P. (2017). The neurocognitive consequences of sleep restriction: A meta-analytic review. *Neuroscience and Biobehavioral Reviews*, 80, 586-604.

Lyberg, L., et Stukel, D. (2017). The roots and evolution of the total survey error concept. Dans *Total Survey Error in Practice*, (Éds., Paul Biemer, Edith de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars Lyberg, Clyde Tucker et Brady West), 1-22. Hoboken, NJ: John Wiley & Sons, Inc.

Malhotra, N. (2008). Completion time and response order effects in web surveys. *Public Opinion Quarterly*, 72(5), 914-934.

Nishii, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivariate Analysis*, 27, 392-403.

Olson, K., Smyth, J. et Ganshert, A. (2019). The effects of respondent and question characteristics on respondent answering behaviors in telephone interviews. *Journal of Survey Statistics and Methodology*, 7, 275-308.

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, 73(1), 74-97.

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

Phillips, A., et Stenger, R. (2022). The effect of burdensome survey questions on data quality in an omnibus survey. *Journal of Official Statistics*, 38(4), 1019-1050.

Read, B., Wolters, L. et Berinsky, A. (2021). Racing the clock: Using response time as a proxy for attentiveness on self-administered surveys. *Political Analysis*. Disponible à https://doi.org/10.1017/pan.2021.32.

Roenneberg, T., Kuehnle, T., Juda, M., Kantermann, T., Allebrandt, K., Gordijn, M. et Merrow, M. (2007). Epidemiology of the human circadian clock. *Sleep Medicine Reviews*, 11, 429-438.

Salehinejad, M., Wischnewski, M., Ghanavati, E., Mosayebi-Samani, M., Kuo, M.-F. et Nitsche, M. (2021). Cognitive functions and underlying parameters of human brain physiology are associated with chronotype. *Nature Communications*, 12, Article 4672.

Schmidt, C., Collette, F., Cajochen, C. et Peigneux, P. (2007). A time to think: Circadian rhythms in human cognition. *Cognitive Neuropsychology*, 24(7), 755-789.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461-464.

Solon, G., Haider, S. et Wooldridge, J. (2015). What are we weighting for? *Journal of Human Resources*, 50(2), 301-316.

StataCorp (2019). *Stata Base Reference Manual. Release 16*. College Station, TX: Stata Press.

Toninelli, D., et Revilla, M. (2020). How mobile device screen size affects data collected in web surveys. Dans *Advances in Questionnaire Design, Development, Evaluation and Testing*, (Éds., Paul Beatty, Debbie Collins, Lyn Kaye, Jose Luis Padilla, Gordon Willis et Amanda Wilmot), 349-373. Hoboken, NJ: John Wiley & Sons, Inc.

Tourangeau, R., Rips, L. et Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge, Royaume-Uni: Cambridge University Press.

Truebner, M. (2021). The dynamics of "neither agree nor disagree" answers in attitudinal questions. *Journal of Survey Statistics and Methodology*, 9, 51-72.

Valdez, P. (2019). Homeostatic and circadian regulation of cognitive performance. *Biological Rhythm Research*, 50, 85-93.

Weeks, M., Kulka, R. et Pierson, S. (1987). Optimal call scheduling for a telephone survey. *Public Opinion Quarterly*, 51, 540-549.

Williams, K., et Shapiro, T.M. (2018). Academic achievement across the day: Evidence from randomized class schedules. *Economics of Education Review*, 67, 158-170.

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data*. Second edition. Cambridge, MA: MIT Press.

Yan, T., et Olson, K. (2013). Analyzing paradata to investigate measurement error. Dans *Improving Surveys with Paradata: Analytic Uses of Process Information*, (Éd., Frauke Kreuter), 73-95. Hoboken, NJ: John Wiley & Sons, Inc.

Ziniel, S. (2008). *Cognitive Aging and Survey Measurement*. Thèse de doctorat, University of Michigan.

Techniques d'enquête, décembre 2024
Vol. 50, nº 2, pp. 591-600
Statistique Canada, nº 12-001-X au catalogue

Exploration d'une conjecture sur l'asymétrie : élargissement de la règle de Cochran à une proportion estimée à partir d'un échantillon complexe

Phillip S. Kott et Burton Levine[1]

Résumé

La règle de Cochran stipule qu'un intervalle de confiance (de Wald) à 95 % bilatéral standard autour de la moyenne d'un échantillon tiré d'une population présentant une asymétrie positive est raisonnable lorsque la taille de l'échantillon est supérieure à 25 fois le carré du coefficient d'asymétrie de la population. Nous examinons si une variante de cette règle brute s'applique à une proportion estimée à partir d'un échantillon aléatoire simple stratifié.

Mots-clés : Coefficient d'asymétrie; règle de suppression; taille d'échantillon effective; troisième moment central.

1. Introduction

Dans son célèbre manuel de statistiques d'enquête, William Cochran (1977) suggère que l'intervalle de confiance (de Wald) à 95 % bilatéral standard pour une proportion estimée de la population p (ayant une propriété) fondée sur un échantillon aléatoire simple fonctionne raisonnablement bien lorsque la valeur absolue du coefficient d'asymétrie (de Fisher) de l'estimation est inférieure à 0,2. Plus particulièrement, Cochran a suggéré que la proportion réelle de la population P se situera dans l'intervalle de confiance standard au moins 94 % du temps sur des échantillons répétés, *bien que la fraction d'omissions de chaque côté de l'intervalle ne soit pas nécessairement égale*.

La règle de Cochran pour l'échantillonnage aléatoire simple (appelée ainsi, pour autant que nous le sachions, par Sugden, Smith et Jones (2000)) est que la taille de l'échantillon n devrait être supérieure à $25\mathrm{G}_1{}^2$, où G_1 est le coefficient d'asymétrie de la répartition à partir de laquelle l'échantillon est tiré. Nous reprenons la définition du coefficient d'asymétrie utilisée par Cochran, que l'on trouve dans Evans, Hastings et Peacock (2000, page 15) : le rapport entre le troisième moment central de la répartition au numérateur et le deuxième moment central de la répartition élevé à la puissance $3/2$ au dénominateur.

Le coefficient d'asymétrie de la proportion de l'échantillon p est alors $\mathrm{G}(p) = \mathrm{G}_1/n^{1/2}$, en ignorant la correction de la population finie, de sorte que la règle se traduit par $|\mathrm{G}(p)| < 0{,}2$. La règle brute de Cochran (Cochran a qualifié sa proposition initiale de « brute ») s'applique à toute moyenne d'échantillon présentant une asymétrie positive. Nous limitons ici principalement l'analyse à une proportion estimée de p ayant une asymétrie positive. Il convient de noter que la proportion estimée $1-p$ est symétrique à p et présente une asymétrie négative. Par conséquent, nous présumons que l'intervalle de confiance bilatéral standard est

1. Phillip S. Kott, RTI International, retraité (États-Unis). Courriel : philkott1@gmail.com; Burton Levine, RTI International (États-Unis).

raisonnable lorsque $|G(p)| < 0{,}2$ pour toute proportion estimée presque sans biais (par rapport au plan) p calculée à partir d'un échantillon complexe. Nous étudions cette conjecture de manière empirique pour les estimations sans biais fondées sur des échantillons aléatoires simples stratifiés virtuels dans la section 3. Dans la section 4, nous analysons les répercussions pratiques de notre conjecture puisque, d'un point de vue opérationnel, $G(p)$ devra être remplacé par une estimation. Dans la section suivante, nous présentons quelques données statistiques issues de la théorie de l'échantillonnage probabiliste (souvent appelée « théorie de l'échantillonnage fondé sur le plan d'enquête »).

2. Quelques éléments de contexte

Pour un échantillon aléatoire simple stratifié, $h = 1, \ldots, H$ représente les strates; $n_h > 1$, la taille de l'échantillon dans la strate h; N_h, la taille de la population de la strate h; $n = \sum_{h=1}^{H} n_h$; et $N = \sum_{h=1}^{H} N_h$. Soit P_h la proportion de la population dans la strate h, et p_h la proportion de l'échantillon dans la strate h.

Les équations suivantes sont toutes bien connues. La proportion de la population P est égale à $P = \sum_{h=1}^{H} \frac{N_h}{N} P_h$, tandis que $p = \sum_{h=1}^{H} \frac{N_h}{N} p_h$ est son estimateur. En supposant (comme nous le ferons à partir de maintenant) que N est si grand que la correction de la population finie peut être ignorée, la variance de p est la suivante :

$$\mathrm{Var}(p) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \frac{P_h(1-P_h)}{n_h}, \tag{2.1}$$

et qu'un estimateur sans biais de cette variance est :

$$\mathrm{var}(p) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \frac{p_h(1-p_h)}{n_h - 1}. \tag{2.2}$$

Le troisième moment central de p est :

$$M_3(p) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^3 \frac{P_h(1-P_h)(1-2P_h)}{n_h^2}, \tag{2.3}$$

et un estimateur sans biais pour ce paramètre lorsque tous les $n_h > 2$ est :

$$m_3(p) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^3 \frac{p_h(1-p_h)(1-2p_h)}{(n_h-1)(n_h-2)}. \tag{2.4}$$

Le coefficient d'asymétrie de p est :

$$G(p) = M_3(p) \Big/ \left[\mathrm{Var}(p) \right]^{3/2}, \tag{2.5}$$

et un estimateur presque sans biais de ce paramètre (s'il existe) est :

$$g(p) = m_3(p) \Big/ \left[\mathrm{var}(p) \right]^{3/2}. \tag{2.6}$$

La mesure *ad hoc* de l'asymétrie, très répandue, évite de mesurer le troisième moment central de p :

$$\mathrm{G}^*(p) = \frac{(1-2P) \,/\, (n^+)^{1/2}}{\left[P(1-P)\right]^{1/2}} = \frac{(1-2P)}{P(1-P)}\left[\mathrm{Var}(p)\right]^{1/2}, \tag{2.7}$$

où $n^+ = \frac{P(1-P)}{\mathrm{Var}(p)}$ est la *taille d'échantillon effective* du plan d'échantillonnage et de l'estimateur.

Un estimateur presque sans biais pour $\mathrm{G}^*(p)$ est :

$$\mathrm{g}^*(p) = \frac{(1-2p) \,/\, (n^*)^{1/2}}{\left[p(1-p)\right]^{1/2}} = \frac{(1-2p)}{p(1-p)}\left[\mathrm{var}(p)\right]^{1/2}, \tag{2.8}$$

où $n^* = \frac{p(1-p)}{\mathrm{var}(p)}$ est la *taille d'échantillon effective estimée* du plan d'échantillonnage et de l'estimateur. Contrairement à $\mathrm{g}(p)$, $\mathrm{g}^*(p)$ peut être calculé lorsqu'un ou plusieurs $n_h = 2,$ ce qui est souvent le cas dans la pratique.

Dans le cas d'un échantillonnage aléatoire simple et d'un n de taille importante, $n = n^+ \approx n^*,$ de sorte que $\mathrm{G}^*(p) = \mathrm{G}(p)$ et $\mathrm{g}^*(p) \approx \mathrm{g}(p).$

3. Quelques expériences simulées

Dans le cadre d'un échantillonnage aléatoire simple, une taille d'échantillon d'environ 180 est nécessaire pour que $\mathrm{G}(p)$ soit inférieur à 0,2 lorsque le P ciblé est de 0,1. Les expériences simulées dans la présente section ont été conçues dans cette optique.

Notre objectif était d'évaluer des intervalles de confiance bilatéraux à 95 % pour différentes estimations fondées sur des échantillons stratifiés de 180 unités. Nous avons envisagé des plans d'échantillonnage à trois strates avec 60 unités d'échantillonnage chacune et des plans d'échantillonnage à 90 strates avec deux unités d'échantillonnage chacune. Nous avons envisagé la possibilité que la taille relative de la population dans chaque strate soit la même ou que la moitié de la population soit représentée par un tiers des unités d'échantillonnage, et ce, de deux manières différentes (comme nous l'expliquerons). Enfin, nous avons envisagé une méthode homoscédastique (variance unitaire égale) d'affectation des variables d'enquête dans laquelle chaque unité de population avait la même probabilité d'avoir une valeur d'enquête binaire de 1 (plutôt que 0) et que cette probabilité variait entre les 20 valeurs numériques (0,01; 0,02; [...]; 0,19; 0,20). Nous avons également envisagé une méthode d'affectation hétéroscédastique dans laquelle un tiers de la population n'avait aucune chance d'avoir une valeur d'enquête binaire de 1, un tiers avait la même chance d'avoir une valeur d'enquête de 1 que dans la méthode d'affectation homoscédastique, et un tiers avait deux fois plus de chances d'avoir une valeur d'enquête de 1 que dans la méthode homoscédastique.

Plutôt que de simuler 100 000 fois 180 unités d'échantillonnage tirées séparément pour 20 estimations dans chacun des 12 scénarios différents (2 méthodes d'attribution de variables × 2 ensembles de formations de strates × 3 ensembles de tailles de population relatives), nous avons fait l'équivalent pour alléger la charge

de calcul. Nous avons tiré 100 000 échantillons originaux ou primitifs. Chaque échantillon original ou primitif contenait 180 unités d'échantillonnage original ou primitif ordonnées. Dans chacun des 12 scénarios, chaque unité d'échantillonnage original ou primitif a été affectée à 20 valeurs d'enquête distinctes, et donc à 20 unités d'échantillonnage virtuelles. Nous appelons ce que nous avons trouvé sur les 20 intervalles de confiance estimés dans chaque scénario le résultat d'une « expérience simulée » parce que nous n'avons pas réellement prélevé d'échantillons dans chaque scénario. Néanmoins, nous appelons chacune des 100 000 sélections d'un échantillon original ou primitif de 180 unités et ses répercussions une « simulation ».

Les détails de ce que nous avons fait suivent. Chaque unité d'échantillonnage original ou primitif j a été associée à un tirage aléatoire indépendant d_j de la répartition uniforme sur l'intervalle mi-fermé, mi-ouvert [0, 1) (c'est-à-dire que 0 est inclus dans l'intervalle, mais que 1 ne l'est pas). En laissant P_v prendre les 20 valeurs (0,01; 0,02; [...]; 0,19; 0,20), chaque unité d'échantillonnage original ou primitif j s'est vu attribuer 20 variables d'enquête binaires de l'une des deux manières suivantes. Dans la méthode d'affectation homoscédastique des variables, $y_{jv} = 1$ lorsque $d_j < P_v$ et 0 sinon. Dans la méthode d'affectation hétéroscédastique des variables, $y_{jv} = 1$ lorsque $d_j < P_v a_j$ et 0 sinon, où $a_j = 0$ lorsque $j < 61$, $a_j = 1$ lorsque $60 < j < 121$, et $a_j = 2$ lorsque $j > 120$.

Les unités d'échantillonnage original ou primitif ordonnées j ont été affectées à des strates de la façon qui suit. Dans la méthode d'affectation à trois strates, l'unité d'échantillonnage original ou primitif a été affectée à la strate 1 lorsque $j < 61$; l'unité a été affectée à la strate 2 lorsque $60 < j < 121$; et l'unité a été affectée à la strate 3 lorsque $j > 120$. Dans la méthode à 90 strates, lorsque $j =$ équivaut à 1 ou 2, l'unité est affectée à la strate 1; lorsque $j =$ équivaut à 3 ou 4, l'unité est affectée à la strate 2, et ainsi de suite.

Enfin, les strates se sont vu attribuer des tailles de population relatives (c'est-à-dire $f_h = N_h / N$) de trois façons différentes. Pour la méthode d'affectation à trois strates, les trois méthodes étaient $f_1 = f_2 = f_3 = 1/3$; $f_1 = 1/2$ $f_2 = f_3 = 1/4$; et $f_1 = f_2 = 1/4$ $f_3 = 1/2$. Pour la méthode d'affectation à 90 strates, les trois méthodes étaient $f_h = 1/90$ pour toutes les valeurs h; $f_h = 1/60$ pour $h = 1$ à 30, sinon $f_h = 1/120$; et $f_h = 1/120$ pour $h = 1$ à 60, sinon $f_h = 1/60$.

Pour un grand nombre des 12 scénarios, la proportion cible des estimateurs P (définie dans la section 2) était la même que P_v (la simple moyenne des valeurs d'enquête dans l'échantillon virtuel). Les exceptions se sont produites avec la méthode d'affectation hétéroscédastique des variables lorsque les parts des strates (c'est-à-dire les tailles relatives des populations) n'étaient pas toutes égales. Lorsque les unités d'échantillonnage virtuelles de la strate la plus basse $(h = 1)$ selon la méthode d'assignation à trois strates ou des 30 strates les plus basses $(h = 1, \ldots, 30)$ selon la méthode d'assignation à 90 strates représentaient la moitié des estimations, le nombre d'unités d'échantillonnage virtuelles était inférieur à celui des unités d'échantillonnage virtuelles $P = (3/4) P_v$. Lorsque les échantillons de la strate la plus élevée $(h = 3)$ ou des 30 strates les plus élevées $(h = 61, \ldots, 90)$ représentaient la moitié des estimations $P = (5/4) P_v$.

Il faut souligner que P_v (et non P) figure sur l'axe des x dans les six graphiques des figures 3.1, 3.2 et 3.3. Ces graphiques montrent les couvertures moyennes sur 100 000 simulations des intervalles de confiance

bilatéraux à 95 % traditionnels dans le cadre des scénarios décrits ci-dessus pour P_v établi à 0,01; 0,02; [...] et 0,20.

Les intervalles de confiance ont été calculés à l'aide des estimations traditionnelles de $\mathrm{Var}(p)$ par échantillonnage probabiliste sans modèle en ignorant la correction de la population finie décrite à la section 2. Le fait que les données utilisées dans les expériences aient été générées à partir d'un modèle ne remet pas en cause l'utilité des méthodes d'inférence sans modèle (en particulier lorsque le modèle générant les données est inconnu, ce qui n'était pas le cas ici).

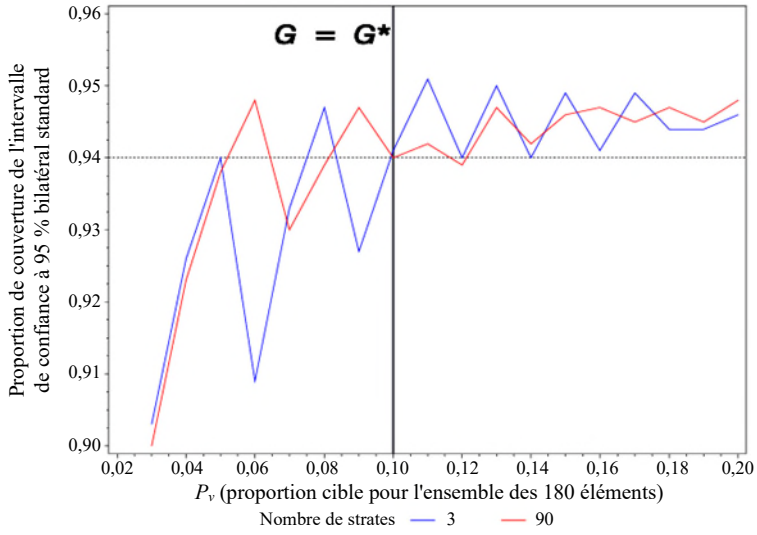
Nous nous intéressons à la relation entre les couvertures des intervalles de confiance bilatéraux à 95 % traditionnels et les valeurs de $\mathrm{G}(p)$ qui (comme $\mathrm{G}^*(p)$) diminuent toujours à mesure que P_v augmente, du moins dans la fourchette que nous avons étudiée, c'est-à-dire pour $P_v < 0{,}2$ (non illustré).

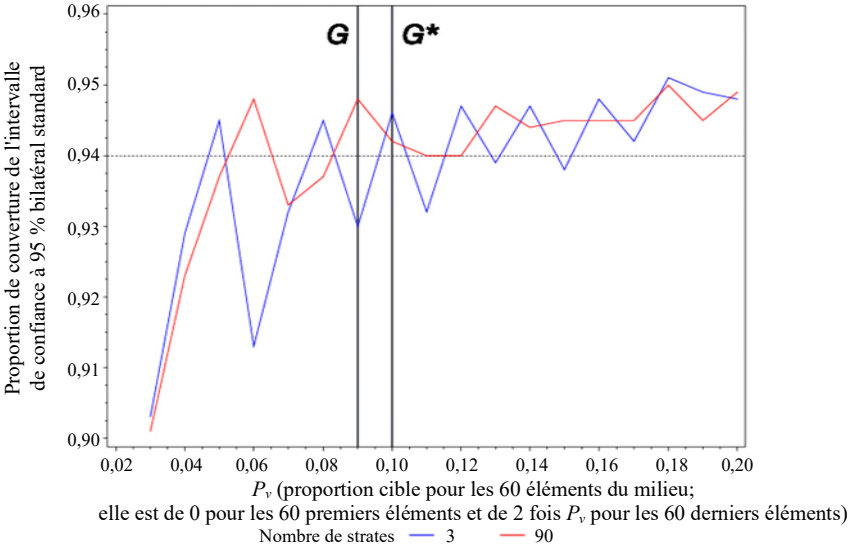
Il y a des lignes verticales pour la première valeur de P_v, pour laquelle $\mathrm{G}(p)$ est inférieur à 0,2. De telles lignes sont également visibles pour $\mathrm{G}^*(p)$ (les deux lignes sont identiques lorsque les variables de l'enquête sont homoscédastiques et que les parts des strates sont égales). Les lignes verticales nous permettent d'évaluer la version robuste de notre conjecture, à savoir que les couvertures à droite de la ligne $\mathrm{G}(p)$ ou $\mathrm{G}^*(p)$ devraient *toujours* être d'au moins 94 %. Cette version échoue à la fois pour $\mathrm{G}(p)$ et $\mathrm{G}^*(p)$, mais une version brute plus faible remplaçant « toujours » par « habituellement » n'échoue pas.

Bien que les échantillons virtuels et les estimations de p soient identiques lorsque l'affectation des variables et les parts relatives des strates sont identiques, les couvertures ne le sont pas, car les variances estimées $\mathrm{var}(p)$ (dans l'équation 2.2), contrairement aux variances réelles $\mathrm{Var}(p)$ (dans l'équation 2.1), diffèrent entre les méthodes d'affectation à trois strates et à 90 strates. C'est pour cette raison que les lignes rouges et bleues diffèrent dans chacun des six graphiques présentés dans les trois figures, tandis que les lignes $\mathrm{G}(p)$ et $\mathrm{G}^*(p)$ (des équations 2.5 et 2.7), qui sont des fonctions de l'échantillon, $\mathrm{Var}(p)$, $\mathrm{M}_3(p)$ (de l'équation 2.3), et P, sont les mêmes pour les deux méthodes d'affectation des strates. Cela nous a permis de présenter les résultats de 12 scénarios dans six graphiques.

Pour ces six graphiques, nous avons utilisé les intervalles de confiance à 95 % standard produits par PROC SURVEYMEANS (2020) du SAS, c'est-à-dire $\mathrm{IC}_d(P) = p \pm t_{0,05}(d)\sqrt{\mathrm{var}(p)}$, où $t_{0,95}(d) = 1{,}9754$ pour l'affectation à trois strates ($d = 180 - 30 = 150$ étant les degrés nominaux de liberté de $\mathrm{var}(p)$) et $t_{0,95}(d) = 1{,}9872$ pour l'affectation à 90 strates ($d = 180 - 90 = 90$). Comme indiqué ci-dessus, p est le même pour les deux méthodes d'affectation des strates, mais $\mathrm{var}(p)$ ne l'est pas. L'irrégularité des couvertures a été notée dans des travaux empiriques antérieurs (par exemple Brown, Cai et Dasgupta (2001) et Dean et Pagano (2015)) et est attribuée à la nature discrète de la détermination de la couverture (soit l'intervalle couvre P, soit il ne le couvre pas). Il est un peu surprenant que les lignes à 90 strates (rouges) soient moins irrégulières que les lignes à trois strates (bleues), même si ces dernières ont plus de degrés de liberté. Une enquête sur les raisons expliquant cette situation devra être menée ultérieurement.

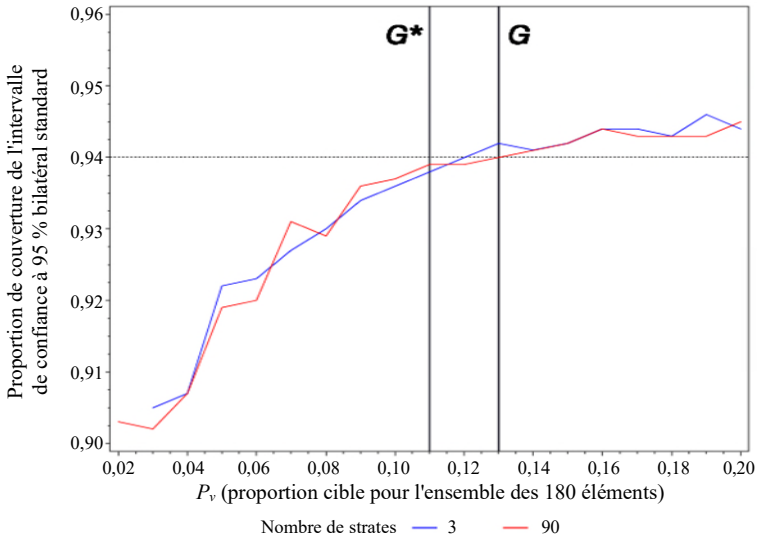
Figure 3.1 Couvertures lorsque toutes les parts de strates (f_h) sont égales.

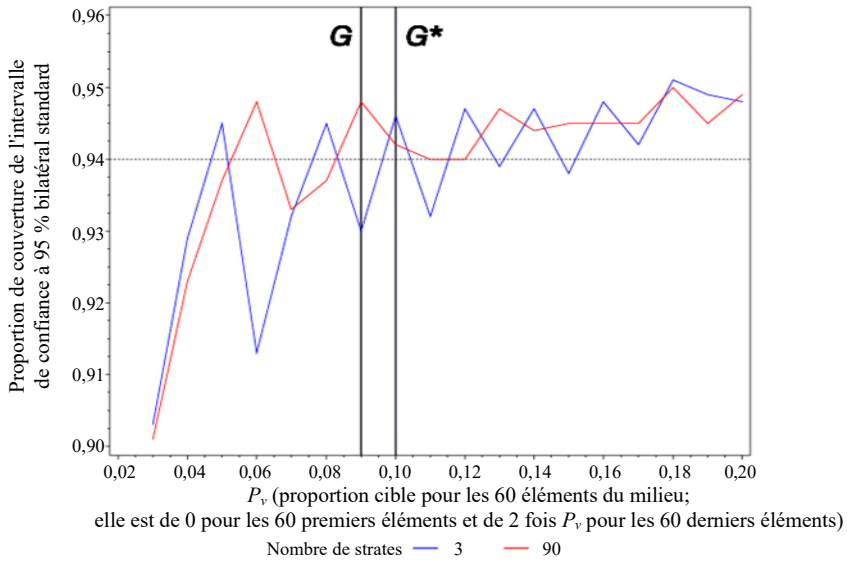




Note : G(G*) est d'abord inférieur à 0,2 au niveau de la ligne verticale, puis diminue à mesure que P_v augmente.

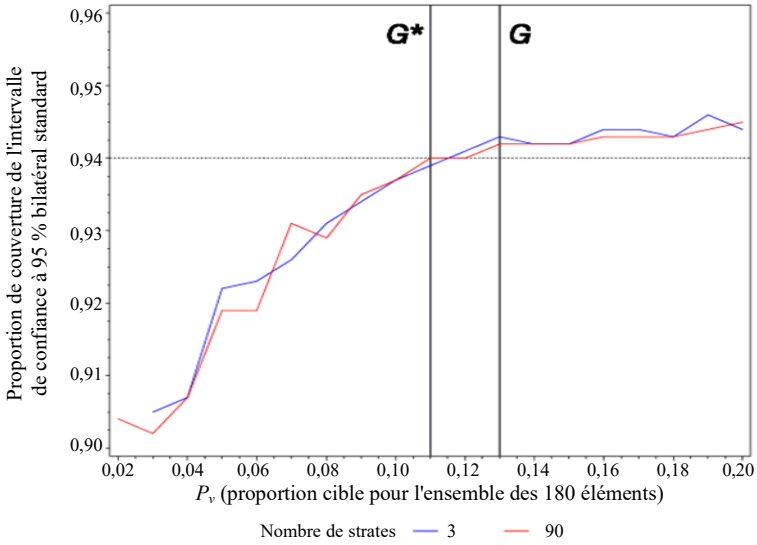
Figure 3.2 Couvertures lorsque le premier tiers de la strate a la moitié des parts.

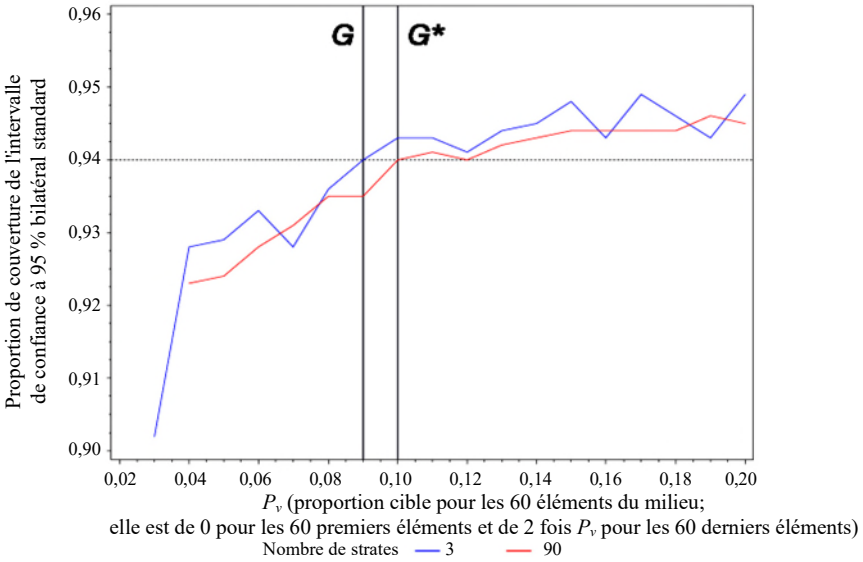




Note : G(G*) est d'abord inférieur à 0,2 au niveau de la ligne verticale, puis diminue à mesure que P_v augmente.

Figure 3.3 Couvertures lorsque le dernier tiers de la strate a la moitié des parts.





Note : G(G*) est d'abord inférieur à 0,2 au niveau de la ligne verticale, puis diminue à mesure que P_v augmente.

En réfléchissant un peu, on peut comprendre pourquoi les graphiques inférieurs des figures 3.1 et 3.2 coïncident exactement, alors que les graphiques supérieurs des figures 3.2 et 3.3 sont presque identiques (c'est seulement la nature finie des simulations qui les fait différer légèrement).

4. Analyse

L'utilisation d'un intervalle de confiance à 95 % bilatéral standard pour une proportion estimée repose sur le fait que l'estimation est asymptotiquement normale. Notamment, le coefficient d'asymétrie d'un estimateur normalement distribué est égal à 0. Une proportion estimée en fonction d'un échantillon fini a un coefficient d'asymétrie non nul. Le coefficient d'asymétrie tend à diminuer à mesure que la taille de l'échantillon sur lequel les estimations sont basées augmente. Nous avons élaboré notre version de la règle de Cochran afin de pouvoir déterminer quand la taille d'un échantillon complexe est suffisante pour permettre l'utilisation raisonnable d'un intervalle de confiance à 95 % bilatéral standard.

Notre version brute de la règle de Cochran, à savoir que l'intervalle de confiance à 95 % bilatéral standard pour une proportion estimée p fondée sur un échantillon complexe est raisonnable lorsque la valeur absolue de son coefficient d'asymétrie $\mathrm{G}(p)$ est inférieure à 0,2, ne peut souvent pas être utilisée directement parce que $\mathrm{G}(p)$ est inconnu. En fait, son estimateur sans biais $\mathrm{g}(p)$ a tendance à avoir un léger biais à la hausse en raison de la nature aléatoire de son dénominateur (dans l'équation 2.6) lorsque $\mathrm{G}(p)$ est positif. Par conséquent, il semble (d'après nos expériences de simulation limitées) généralement sûr de remplacer la règle brute par :

> l'intervalle de confiance à 95 % bilatéral standard pour une proportion estimée p (en fonction d'un échantillon complexe) est raisonnable lorsque la valeur absolue de son coefficient d'asymétrie $\mathrm{G}(p)$ est inférieure à 0,2,

avec la règle plus opérationnelle :

> l'intervalle de confiance à 95 % bilatéral standard pour une proportion estimée p est raisonnable lorsque la valeur absolue de son coefficient d'asymétrie estimé $\mathrm{g}(p)$ est inférieure à 0,2,

ou même

> l'intervalle de confiance à 95 % bilatéral standard pour une proportion estimée p est raisonnable lorsque la valeur absolue du coefficient d'asymétrie alternatif estimé $\mathrm{g}^*(p)$ est inférieure à 0,2 lorsque $\mathrm{g}(p)$ ne peut être calculé.

Ces versions opérationnelles de la règle de Cochran pour une proportion estimée à partir d'un échantillon complexe sont encore plus susceptibles d'être raisonnables lorsque l'intervalle de confiance à 95 % bilatéral standard pour p est exprimé comme $\mathrm{IC}_{60}(P) = p \pm 2\sqrt{\mathrm{var}(p)}$, et que $\mathrm{var}(p)$ a au moins 60 degrés nominaux de liberté.

$\mathrm{IC}_{60}(P) = p \pm 2\sqrt{\mathrm{var}(p)}$ est la version de l'intervalle de confiance à 95 % bilatéral standard pour P que de nombreux utilisateurs avertis calculent à l'interne lorsqu'on leur fournit uniquement une proportion estimée p et son erreur type estimée $\sqrt{\mathrm{var}(p)}$. Il en ressort la règle de suppression suivante pour une proportion estimée :

> Supprimer p lorsque la valeur absolue de son coefficient d'asymétrie estimé $\mathrm{g}(p)$ ou de son coefficient d'asymétrie estimé alternatif $\mathrm{g}^*(p)$ est supérieure à 0,2,

car c'est à ce moment-là que l'intervalle de confiance largement utilisé $\mathrm{IC}_{60}(P) = p \pm 2\sqrt{\mathrm{var}(p)}$ peut ne plus être raisonnable.

Bien que nous n'ayons pas étudié l'échantillonnage en grappes *en tant que tel*, les expériences à 90 strates peuvent être considérées comme représentant un plan d'échantillonnage à deux niveaux (ou plus) avec une corrélation parfaite dans chacune des 180 unités primaires d'échantillonnage (UPE), c'est-à-dire que chaque élément de chaque UPE a la même valeur d'enquête (0 ou 1) que tous les autres éléments de l'UPE. Dean et Pagano (2015) et leur documentation complémentaire (disponible auprès des auteurs) montrent que pour une version particulière d'un échantillon à deux degrés (30 UPE sélectionnées à l'aide de l'échantillonnage avec probabilité proportionnelle à la taille et sept éléments tirés avec une probabilité égale dans chaque UPE), l'intervalle de confiance à 95 % bilatéral standard couvre moins bien une petite P à mesure que la corrélation dans les UPE augmente (leur analyse s'est arrêtée lorsque la corrélation entre les grappes a atteint un sommet de 0,5). Même l'intervalle Clopper-Pearson « exact » couvre mal les cas où les corrélations entre les grappes sont élevées et où P est petit.

Bibliographie

Brown, L., Cai, T. et Dasgupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science* 16, 101-133.

Cochran, W.G. (1977). *Sampling Techniques* (3rd edition), New York: John Wiley & Sons, Inc.

Dean, N., et Pagano, M. (2015). Evaluating confidence interval methods for binomial proportions in clustered surveys. *Journal of Survey Statistics and Methodology*, 3, 484-503.

Evans, M., Hastings, N. et Peacock, B. (2000). *Statistical Distributions* (3rd edition), New York: John Wiley & Sons, Inc.

SAS (2020). *SAS/STAT 15.2 User's Guide.* SAS Institute Inc., Cary, NC.

Sugden, R.A., Smith, T.M.F. et Jones, R.P. (2000). Cochran's rule for simple random sampling. *Journal of the Royal Statistical Society* (B), 62, 4, 787-793.

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont fourni de l'aide ou ont fait la critique d'un article ou plus durant l'année 2024.

- C. Adrijo, *US Food and Drug Administration White Oak Campus*
- P. Andersson, Stockholm University
- N. Bates, *U.S. Census Bureau* (retraitée)
- B. Bell, *U.S. Census Bureau*
- E. Berg, *Iowa State University*
- C. Bocci, *Statistique Canada*
- H.-J. Boonstra, *Statistics Netherlands Heerlen*
- C. Boulet, *Statistique Canada*
- J. Breidt, *NORC at the University of Chicago*
- J.M. Brick, *Westat Inc.*
- P.J. Cantwell, *U.S. Census Bureau*
- G. Chauvet, *École nationale de la statistique et de l'analyse de l'information*
- L. Chen, *National Institute of Statistical Sciences*
- S. Chen, *University of Oklahoma Health Sciences Center*
- J. Chipperfield, *Australian Bureau of Statistics*
- R. Clark, *Australian National University*
- M. Cohen, *American Institutes for Research*
- T. Crossley, *European University Institute*
- M. Dagdoug, *Laboratoire de Mathématiques de Besançon*
- S. Das, *Maastricht University*
- M. DeBell, *Stanford University*
- M. del Mar Rueda, *University of Granada*
- J. Drechsler, *Institute for Employment Research*
- P. Duchesne, *Université de Montréal*
- J.L. Eltinge, *U.S. Census Bureau*
- N. English, *National Opinion Research Center*
- A. Erciulescu, *Westat Inc.*
- W.A. Fuller, *Iowa State University*
- J. Gambino, *Statistique Canada*
- G. Goh, *Kyungpook National University*
- D. Haziza, *University of Ottawa*
- D. Hedlin, *Stockholm University*
- M.A. Hidiroglou, *Statistique Canada*
- J. Jiang, *University of California*
- D. Judkins, *ABT Associates Inc Bethesda*
- Y. Kawakubo, *Chiba University*
- B. Kim, *University of Maryland*
- J.K. Kim, *Iowa State University*
- P.S. Kott, *RTI International*
- K. Larbi, *L'Institut national de la statistique et des études économiques*
- P. Lahiri, *University of Maryland*
- É. Lesage, *L'Institut national de la statistique et des études économiques*
- T. Lewis, *George Mason University*
- A. Manda, *University of Georgia*
- A. Matei, *Université de Neuchâtel*
- S. Matthews, *Government of Canada*
- K. McConville, *Reed College*
- M. McRoy, *NORC at the University of Chicago*
- E. Médous, *Nantes Université*
- T. Merly-Alpa, *Institut national d'études démographiques*
- I. Molina, *Universidad Complutense de Madrid*
- J. Moore, *University of Essex*
- R. Münnich, *University of Trier*
- J. Opsomer, *Westat Inc.*
- P. Parker, *University of California*
- J. Pascale, *U.S. Census Bureau*
- D. Pfeffermann, *University of Southampton*
- J.N.K. Rao, *Carleton University*
- P. Righi, *Italian National Institute of Statistics*
- L.-P. Rivest, *Université Laval Faculté des sciences et de génie*
- A. Ruiz-Gazen, *Toulouse School of Economics*
- F.J. Scheuren, *National Opinion Research Center*
- A. Sen, *University of Maryland at College Park*
- N. Shlomo, *The University of Manchester*
- P.L.d.N. Silva, *Escola Nacional de Ciências Estatísticas*
- E. Slud, *University of Maryland at College Park*
- P. Smith, *University of Southampton*
- D. Steel, *University of Wollongong*
- S. Sugasawa, *Keio University*
- H. Sun, *West Tennessee Healthcare*
- X. Tang, *The University of Arizona*
- M. Templ, *University of Applied Sciences and Arts Northwestern Switzerland*
- Y. Tillé, *University of Neuchatel Faculty of Sciences*
- R. Tiller, *U.S. Bureau of Labor Statistics*
- M. Torabi, *University of Manitoba*
- D. Toth, *U.S. Bureau of Labor Statistics*
- J. van den Brakel, *Statistics Netherlands*
- A. Veiga, *Brazilian Institute of Geography and Statistics*
- F. Verret, *Statistique Canada*
- G. Walejko, *U.S. Census Bureau*
- L. Wang, *University of Virginia*
- M. Williams, *RTI International*
- B. West, *University of Michigan*
- C. Wu, *University of Waterloo*
- D. Yang, *US Bureau of Labor Statistics*
- Y. You, *Statistique Canada*
- W. Yung, *Statistique Canada*
- L.-C. Zhang, *University of Southampton*

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2024 : Céline Ethier de la Division des Méthodes de la Statistique Économique; Patrick O'Leary de la Division des méthodes de la statistique sociale; Julie Bélanger et Catherine Pelletier de la Division de la diffusion officielle, de la publication et des services de création; l'équipe de la Division de la diffusion, en particulier Chantal Chalifoux, Isabelle Gravelle, Kathy Charbonneau, Ashley Perry, Travis Robinson et Karo-Lynn Audy ainsi que nos partenaires de la Division des communications.

ANNONCES

Demande de candidatures pour le prix Waksberg 2026

La revue *Techniques d'enquête* a mis sur pied une série annuelle de communications sollicitées en l'honneur de Joseph Waksberg, en reconnaissance des contributions exceptionnelles qu'il a faites à la statistique et méthodologie d'enquête. Chaque année, un éminent statisticien d'enquête est choisi pour rédiger un article où il examine l'évolution et l'état actuel d'un thème important du domaine de la statistique et méthodologie d'enquête. L'article reflète le mélange de théorie et de pratique caractéristique des travaux de Joseph Waksberg.

Le lauréat du prix Waksberg recevra une prime en argent et présentera la communication sollicitée Waksberg 2026 au Symposium de Statistique Canada prévu à l'automne 2026. L'article paraîtra dans un numéro futur de *Techniques d'enquête* (prévu pour décembre 2026).

L'auteur de l'article Waksberg 2026 sera choisi par un comité de quatre personnes désignées par *Techniques d'enquête* et l'*American Statistical Association*. **Les candidatures doivent être envoyées par courriel avant le 15 février 2025 au président du comité, Jae Kwang Kim (jkim@iastate.edu). Les candidatures doivent inclure un Curriculum Vitae et une lettre de candidature.** Les candidatures resteront actives pour une période de 5 ans.

Membres du comité de sélection de l'article Waksberg (2024-2025)

Jae Kwang Kim, *Iowa State University* (Président)
Paul Smith, *University of Southampton*
Alina Matei, *Université de Neuchatel*
Kristen Olson, *University of Nebraska-Lincoln*

Présidents précédents :

Graham Kalton (1999-2001)
Chris Skinner (2001-2002)
David A. Binder (2002-2003)
J. Michael Brick (2003-2004)
David R. Bellhouse (2004-2005)
Gordon Brackstone (2005-2006)
Sharon Lohr (2006-2007)
Robert Groves (2007-2008)
Leyla Mojadjer (2008-2009)
Daniel Kasprzyk (2009-2010)
Elizabeth A. Martin (2010-2011)
Mary E. Thompson (2011-2012)
Steve Heeringa (2012-2013)
Cynthia Clark (2013-2014)
Louis-Paul Rivest (2014-2015)
Tommy Wright (2015-2016)
Kirk Wolter (2016-2017)
Danny Pfeffermann (2017-2018)
Michael A. Hidiroglou (2018-2019)
Robert E. Fay (2019-2020)
Jean Opsomer (2020-2021)
Jack Gambino (2021-2022)
Maria Giovanna Ranalli (2022-2023)
Denise Silva (2023-2024)

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 40, No. 2, June 2024

Robust Statistical Estimation for Capture-Recapture Using Administrative Data
James O. Chipperfield, Randall Chu, Li-Chun Zhang and Bernard Baffour 215

Small-Sample Bias Correction of Inequality Estimators in Complex Surveys
Silvia De Nicolò, Maria Rosaria Ferrante and Silvia Pacei 238

Disaggregating Death Rates of Age-Groups Using Deep Learning Algorithms
Andrea Nigri, Susanna Levantesi and Salvatore Scognamiglio 262

A Computationally Efficient Approach to Fully Bayesian Benchmarking
Taylor Okonek and Jon Wakefield 283

Nonlinear Fay-Herriot Models for Small Area Estimation Using Random Weight Neural Networks
Paul A. Parker 317

Reliable Event Rates for Disease Mapping
Harrison Quick and Guangzi Song 333

Alternative Sources and Machine Learning for Official Statistics: A Review
Marco Puts 348

All inquires about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 40, No. 3, September 2024

An Application of a Small Area Procedure with Correlation Between Measurement Error and Sampling Error to the Conservation Effects Assessment Project
Emily Berg and Sepideh Mosaferi 355

Constructing Limited-Revisable and Stable CPPIs for Small Domains
Farley Ishaak, Pim Ouwehand and Hilde Remoy 380

A New Approach to Composite Estimation for Repeated Surveys with Rotating Panels
Takis Merkouris 409

Capitalization Accounting of Data Factor: Theoretical Mechanism, Methodological Path, and Statistical Measurement
Kaike Wang, Qiang He, Wuyi Zeng and Chunyun Wang 425

Comparing Long- Versus Short-Forms of Depression Scales in an Omnibus Longitudinal Survey
Qiong Wu and Haozhi Qian 457

State-Space Modeling Approach to Exploring the Index of Production in Construction for Türkiye
Özlem Yiğit 472

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 52, No. 1, March/mars 2024

Issue Information 1

Research Articles

PCA Rerandomization
Hengtao Zhang, Guosheng Yin, Donald B. Rubin 5

Subgroup analysis of linear models with measurement error
Yuan Le, Yang Bai, Guoyou Qin 26

Robust nonparametric hypothesis tests for differences in the covariance structure of functional data
Kelly Ramsay, Shoja'eddin Chenouri 43

A stable and adaptive polygenic signal detection method based on repeated sample splitting
Yanyan Zhao, Lei Sun 79

Penalized complexity priors for the skewness parameter of power links
José A. Ordoñez, Marcos O. Prates, Jorge L. Bazán, Victor H. Lachos 98

Asymptotic distribution of one-component partial least squares regression estimators in high dimensions
Jerónimo Basa, R. Dennis Cook, Liliana Forzani, Miguel Marcos 118

Segment regression model average with multiple threshold variables and multiple structural breaks
Pan Liu, Jialiang Li 131

Variable selection in additive models via hierarchical sparse penalty
Canhong Wen, Anan Chen, Xueqin Wang, Wenliang Pan, for the Alzheimer's Disease Neuroimaging Initiative 162

Regression model selection via log-likelihood ratio and constrained minimum criterion
Min Tsao 195

Method of model checking for case II interval-censored data under the additive hazards model
Yanqin Feng, Ming Tang, Jieli Ding 212

Volatility analysis for the GARCH-Itô model with option data
Huiling Yuan, Yong Zhou, Zhiyuan Zhang, Xiangyu Cui 237

Distributed sequential estimation procedures
Zhuojian Chen, Zhanfeng Wang, Yuan-chin Ivan Chang 271

Unweighted estimation based on optimal sample under measurement constraints
Jing Wang, HaiYing Wang, Shifeng Xiong 291

A class of space-filling designs with low-dimensional stratification and column orthogonality
Pengnan Li, Fasheng Sun 310

Acknowledgement of referees' services remerciements aux membres des jurys 327

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 52, No. 2, June/juin 2024

Issue Information 333

Research Article

Smoothed model-assisted small area estimation of proportions
Peter A. Gao, Jon Wakefield 337

Finite sample and asymptotic distributions of a statistic for sufficient follow-up in cure models
Ross Maller, Sidney Resnick, Soudabeh Shemehsavar 359

Analysis of Multivariate Survival Data under Semiparametric Copula Models
Wenqing He, Grace Y. Yi, Ao Yuan 380

Joint modelling of quantile regression for longitudinal data with information observation times and a terminal event
Weicai Pang, Yutao Liu, Xingqiu Zhao, Yong Zhou 414

New highly efficient high-breakdown estimator of multivariate scatter and location for elliptical distributions
Justin Fishbone, Lamine Mili 437

Identifiability constraints in generalized additive models
Alex Stringer 461

Nonparametric simulation extrapolation for measurement-error models
Dylan Spicker, Michael P. Wallace, Grace Y. Yi 477

Bayesian instrumental variable estimation in linear measurement error models
Qi Wang, Lichun Wang, Liqun Wang 500

Optimal multiwave validation of secondary use data with outcome and exposure misclassification
Sarah C. Lotspeich, Gustavo G. C. Amorim, Pamela A. Shaw, Ran Tao, Bryan E. Shepherd 532

A calibration method to stabilize estimation with missing data
Baojiang Chen, Ao Yuan, Jing Qin 555

A combined moment equation approach for spatial autoregressive models
Jiaxin Liu, Hongliang Liu, Yi Li, Huazhen Lin 577

On the correlation analysis of stocks with zero returns
Hamdi Raïssi 597

High-dimensional model averaging for quantile regression
Jinhan Xie, Xianwen Ding, Bei Jiang, Xiaodong Yan, Linglong Kong 618

Objective model selection with parallel genetic algorithms using an eradication strategy
Jean-François Plante, Maxime Larocque, Michel Adès 636

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Les auteurs désirant faire paraître un article sont invités à le soumettre en français ou en anglais via le **portail de *Techniques d'enquête* sur le site Web de ScholarOne Manuscripts** (https://mc04.manuscriptcentral.com/surveymeth). Avant de soumettre l'article, prière d'examiner un numéro récent de *Techniques d'enquête* et de noter les points ci-dessous. Les articles doivent être soumis en Word ou Latex, préférablement en Word avec MathType pour les expressions mathématiques. Une version pdf est également requise pour les formules et graphiques.

1. Présentation

1.1 Les textes doivent être écrits à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
1.3 Le nom (écrit au long) et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
1.4 Les remerciements doivent paraître à la fin du texte.
1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé et introduction

2.1 Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.
2.2 Le dernier paragraphe de l'introduction devrait contenir une brève description de chacune des sections.

3. Rédaction

3.1 Éviter les notes au bas des pages et les abréviations.
3.2 Limiter l'utilisation d'acronymes. Si un acronyme est utilisé, il doit être défini lors de sa première utilisation.
3.3 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme $\exp(\cdot)$ et $\log(\cdot)$, etc.
3.4 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées par un chiffre arabe à la droite si l'auteur y fait référence plus loin. Utiliser un système de numérotation à deux niveaux selon le numéro de la section. Par exemple, l'équation (4.2) est la deuxième équation importante de la section 4.
3.5 Des caractères gras devraient normalement être utilisés pour distinguer les vecteurs et les matrices des valeurs scalaires.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés avec des chiffres arabes et porter un titre aussi explicatif que possible en haut des tableaux ou des figures. Utiliser un système de numérotation à deux niveaux selon le numéro de la section. Par exemple, le tableau 3.1 est le premier tableau de la section 3.
4.2 Une description textuelle détaillée des figures pourrait être requise à des fins d'accessibilité si le message transmis par l'image n'est pas suffisamment expliqué dans le texte.

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple : Cochran (1977, page 164).
5.2 La première fois qu'une référence est citée dans le texte, le nom de chacun des auteurs doit être écrit. Pour les fois suivantes, le nom de chacun des auteurs peut être écrit à nouveau. Cependant, si la référence contient trois auteurs ou plus, les noms du deuxième auteur et des auteurs suivants peuvent être remplacés par « et coll. ».
5.3 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les articles d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

6. Communications brèves

6.1 Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots, incluant les tableaux, les figures et la bibliographie.