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

Techniques d’enquête

Techniques d’enquête
48-2

Date de diffusion : le 15 décembre 2022

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 2022

Tous droits réservés. 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 2022

•

Volume 48

•

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.*
- S. Cai, *Carleton University*
- 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. Dever, *RTI International*
- J.L. Eltinge, *U.S. Bureau of Labor Statistics*
- W.A. Fuller, *Iowa State University*
- D. Haziza, *University of Ottawa*
- M.A. Hidiroglou, *Statistique Canada*
- B. Hulliger, *University of Applied and Arts Sciences Northwestern, Switzerland*
- 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*
- 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 48, numéro 2, décembre 2022

Table des matières

Article sollicité Waksberg

Roderick J. Little
Bayes, étayé par des idées fondées sur le plan, est le meilleur paradigme global pour l'inférence en enquête par échantillonnage 279

Article spécial avec commentaires

Changbao Wu
Inférence statistique avec des échantillons d'enquête non probabiliste 307
Michael A. Bailey
Commentaires à propos de l'article « Inférence statistique avec des échantillons d'enquête non probabiliste » : Échantillons non probabilistes : évaluation et voie à suivre 339
Michael R. Elliott
Commentaires à propos de l'article « Inférence statistique avec des échantillons d'enquête non probabiliste » 347
Sharon L. Lohr
Commentaires à propos de l'article « Inférence statistique avec des échantillons d'enquête non probabiliste » 359
Xiao-Li Meng
Commentaires à propos de l'article « Inférence statistique avec des échantillons d'enquête non probabiliste » : La miniaturisation de la corrélation due à un défaut des données : une stratégie polyvalente de traitement des échantillons non probabilistes 369
Zhonglei Wang et Jae Kwang Kim
Commentaires à propos de l'article « Inférence statistique avec des échantillons d'enquête non probabiliste » 393
Changbao Wu
Réponse de l'auteur aux commentaires sur l'article « Inférence statistique avec des échantillons d'enquête non probabiliste » 399

Articles réguliers

Zhenhua Wang, Olanrewaju Akande, Jason Poulos et Fan Li
Les modèles d'apprentissage profond sont-ils plus efficaces pour l'imputation de données manquantes dans les enquêtes ? Une comparaison empirique fournit des éléments de preuve 407
Sumonkanti Das, Jan van den Brakel, Harm Jan Boonstra et Stephen Haslett
Modélisation de séries chronologiques multiniveaux de la couverture des soins prénataux au Bangladesh à des niveaux administratifs désagrégés 439
Takis Merkouris
Estimation linéaire optimale dans un échantillonnage à deux phases 481
Hee Cheol Chung et Gauri S. Datta
Modèles spatiaux bayésiens pour l'estimation des moyennes pour petites régions échantillonnées et non échantillonnées 507

Remerciements 537
Annonces 539
Autres revues 541

Techniques d'enquête, décembre 2022
Vol. 48, nº2, pp. 277-278
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 2024.

Ce numéro de *Techniques d'enquête* commence par le 22e article de la série du prix Waksberg. Le comité de rédaction remercie les membres du comité de sélection, composé de Jean Opsomer (président), Jack Gambino, Maria Giovanna Ranalli et Elizabeth Stuart, d'avoir choisi Roderick J. Little comme auteur de l'article du prix Waksberg de 2022.

Communication sollicitée pour le prix Waksberg 2022

Auteur : Roderick J. Little

Roderick Little a publié de nombreux travaux sur les méthodes pour l'analyse des données associées à des valeurs manquantes et sur les modes d'inférence d'enquête basée sur un modèle, ainsi que sur l'utilisation des statistiques dans divers domaines scientifiques, notamment la médecine, la démographie, l'économie, la psychiatrie, le vieillissement et l'environnement. Il est actuellement un professeur éminent de biostatistique Richard D. Remington à l'Université du Michigan, où il occupe également des postes au département des statistiques et à l'Institut de recherche sociale. Il a présidé le département de biostatistique dans cette université pendant 11 ans. De 2010 à 2012, Little a été le tout premier codirecteur de la recherche et de la méthodologie au Bureau du recensement des États-Unis, en plus d'y avoir joué le rôle de scientifique en chef. Ses contributions en matière de recherche lui ont valu des prix de distinction comme le prix du fondateur de l'*American Statistical Association* et la médaille Wilks. Il a également prononcé le message annuel du président en plus d'avoir été conférencier Fisher du COPSS lors des *Joint Statistical Meetings*. Little est un membre élu de l'Institut international de statistique, un *Fellow* de l'*American Statistical Association* et de l'*American Academy of Arts and Sciences*, ainsi qu'un membre de la *National Academy of Medicine* des États-Unis.

Gagnants du prix Waksberg et leurs articles sollicités depuis 2001

2023 Raymond **Chambers**, Manuscrit en préparation prévu pour le numéro de décembre 2023.

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 2022
Vol. 48, nº 2, pp. 279-306
Statistique Canada, nº 12-001-X au catalogue

Bayes, étayé par des idées fondées sur le plan, est le meilleur paradigme global pour l'inférence en enquête par échantillonnage

Roderick J. Little[1]

Résumé

Des arguments conceptuels et des exemples sont présentés qui suggèrent que l'approche d'inférence bayésienne pour les enquêtes permet de répondre aux défis nombreux et variés de l'analyse d'une enquête. Les modèles bayésiens qui intègrent des caractéristiques du plan de sondage complexe peuvent donner lieu à des inférences pertinentes pour l'ensemble de données observé, tout en ayant de bonnes propriétés d'échantillonnage répété. Les exemples portent essentiellement sur le rôle des variables auxiliaires et des poids d'échantillonnage, et les méthodes utilisées pour gérer la non-réponse. Le présent article propose 10 raisons principales de favoriser l'approche d'inférence bayésienne pour les enquêtes.

Mots-clés : Inférence bayésienne calée; inférence fondée sur le plan de sondage; modèles de mélange de schémas d'observation; modèles de superpopulation; pondération d'enquête; post-stratification; échantillonnage avec probabilité proportionnelle à la taille; propension à répondre; splines pénalisés.

1. Introduction

À mon avis, les inférences bayésiennes constituent le meilleur paradigme inférentiel pour l'inférence statistique à partir d'enquêtes, qu'il s'agisse d'échantillonnages probabilistes ou non probabilistes. Consultez, à titre d'exemple, Ericson (1969), Binder (1982), Rubin (1987), Ghosh et Meeden (1997), Little (2003ab, 2004, 2012, 2015), Sedransk (2008) et Fienberg (2011). Cependant, les propriétés fondées sur le plan des inférences bayésiennes sont importantes parce que « tous les modèles sont incorrects » et qu'une vaste acceptation des résultats nécessite des inférences ayant de bonnes caractéristiques d'exploitation dans un échantillonnage répété. En particulier, les modèles bayésiens doivent intégrer des caractéristiques complexes du plan de sondage pour donner lieu à des inférences approximativement calées, c'est-à-dire des intervalles de crédibilité qui sont près des niveaux nominaux lorsqu'ils sont traités comme des intervalles de confiance dans un échantillonnage répété (Rubin, 1984, 2019; Little, 2006). Dans les grands échantillons, les modèles de travail flexibles permettent d'éviter les hypothèses fortes au sujet des paramètres qui mèneraient à des estimations potentiellement biaisées.

Pour orienter les propos, considérons le problème de dérivation d'une estimation ponctuelle q d'une quantité de population finie Q, et d'une estimation par intervalles à 95 % $I_{0,95} = (l, u)$ qui saisit une incertitude de q; l'intervalle pourrait avoir une interprétation fréquentiste comme un intervalle de confiance à 95 %, ou une interprétation bayésienne comme un intervalle de crédibilité *a posteriori* à 95 % pour Q. Je crois que les scientifiques qui ne sont pas statisticiens ont l'habitude d'interpréter l'intervalle I de manière bayésienne, en tant qu'intervalle fixe saisissant l'incertitude à propos de Q. Je ne me

1. Roderick J. Little, Département de biostatistique, University of Michigan. Courriel : rlittle@umich.edu.

concentre toutefois pas indûment sur les différentes interprétations de $I_{0.95}$ selon les deux paradigmes. La valeur nominale de 95 % est établie par convention, et d'autres niveaux peuvent être choisis.

L'échantillonnage de la population finie présente une caractéristique intéressante qui consiste à travailler avec des quantités réelles (bien qu'inconnues). En cas d'inférence par enquête « analytique » où l'accent est mis sur les paramètres des modèles idéalisés de la population, comme les coefficients de régression dans un modèle de régressions multiples, établissons la quantité de population finie Q en tant qu'estimation du paramètre d'intérêt si le modèle était ajusté aux données pour toute la population, selon une certaine méthode d'ajustement convenue comme les moindres carrés ou le maximum de vraisemblance. Une des caractéristiques utiles de cette construction est que Q est une quantité réelle au lieu d'être la caractéristique d'un modèle hypothétique simplifié de la population (par exemple Little, 2004).

Dans le paradigme bayésien, l'inférence pour Q repose sur sa distribution prédictive *a posteriori* en fonction des données, pour des choix judicieux de modèle et de distribution *a priori* concernant des paramètres inconnus. Par conséquent, q pourrait constituer la moyenne prédictive *a posteriori* de Q, et $I_{0.95}$, la valeur du 2,5^{e} au 97,5^{e} centile de la distribution prédictive *a posteriori*, ou les limites de la fourchette de valeurs de Q pour la densité *a posteriori* la plus élevée, en supposant que la distribution prédictive *a posteriori* est unimodale. Une des caractéristiques utiles de l'approche bayésienne est l'intégration automatique des « corrections de la population finie » dans la distribution prédictive *a posteriori* des quantités de population finie – à mesure que l'échantillon converge vers une population finie, la variance *a posteriori* tend vers zéro.

L'accent est mis sur la création de modèles appropriés et de distributions *a priori*. Le calcul était autrefois un défi majeur et demeure une considération pratique, même si c'est moins le cas aujourd'hui, avec l'avènement des méthodes de Monte-Carlo par chaîne de Markov et les progrès rapides accomplis en matière de calculs bayésiens. Par conséquent, il s'avère plus difficile qu'il y a, disons, 30 ans de soutenir l'idée que Bayes est intéressant sur le plan conceptuel, mais trop difficile à mettre en œuvre.

Voici la structure de la suite du présent article. Dans la section 2, je présente une notation et je décris officiellement les modèles et les distributions *a priori* qui sont nécessaires dans l'approche bayésienne relative aux données d'enquête, avec et sans non-réponse. Je décris à la section 3 les caractéristiques habituellement souhaitables d'une inférence à propos de Q et j'indique les raisons qui me font croire que le paradigme bayésien pour des modèles choisis de manière appropriée peut se révéler plus efficace que l'approche fondée sur le plan de sondage pour obtenir ces caractéristiques. Je présente dans la section 4 une variété d'exemples visant à illustrer les points de la section 3. À titre de conclusion, dans la section 5, je propose 10 raisons d'adopter Bayes dans le cadre d'échantillonnage.

2. Notation et article fondamental

Dans la présente section, je présente la notation et un article fondamental qui sous-tend en grande partie la réflexion à l'origine du présent article. Supposons que $Y=(y_1,\ldots,y_N)$ et $S=(S_1,\ldots,S_N)$, où $N<\infty$ est le nombre d'unités dans la population, y_i est l'ensemble de variables d'enquête et S_i, l'indicateur de sélection pour la i^{e} unité, dont la valeur est 1 si l'unité i^{e} est sélectionnée, et 0 autrement. Supposons que Z correspond à l'information du plan de sondage, comme une strate ou des indicateurs de grappe, et que z_i correspond à la valeur de Z pour l'unité i. Examinons l'inférence à propos de la quantité de population finie $Q(Y,Z)$, par exemple le total de la population $Q(Y,Z)=\sum_{i=1}^{N} y_i$, où $Y=(y_1,\ldots,y_N)$. Une approche générale fondée sur un modèle traite S et Y comme des variables aléatoires, dont la distribution conjointe en fonction de Z est :

$$f_{S,Y|Z}(S,Y|Z,\theta,\psi) = f_{Y|Z}(Y|Z,\theta)\, f_{S|Y,Z}(S|Z,Y,\psi), \tag{2.1}$$

où $f_{Y|Z}$ représente la densité des variables d'enquête Y indexée par des paramètres θ inconnus, et où $f_{S|Y,Z}$ représente le modèle pour l'inclusion indexée par des paramètres ψ inconnus. Dans le cas d'un échantillonnage probabiliste sans non-réponse, la distribution de l'échantillonnage est connue et ne dépend pas de Y, c'est-à-dire que :

$$f_{S|Y,Z}(S|Z,Y,\psi) = f_{S|Z}(S|Z); \tag{2.2}$$

les méthodes fondées sur le plan de sondage fondent les inférences sur la distribution des statistiques dans un échantillonnage répété tiré de cette distribution.

Pour une enquête comportant de la non-réponse totale, une inclusion se produit lorsqu'une unité est sélectionnée, et puis répond, si sélectionnée. Ainsi, supposons que $R_i=1$ si l'unité sélectionnée i répond, et que $R_i=0$ dans le cas contraire. L'approche fondée sur un modèle modélise la distribution conjointe de S, R et Y compte tenu de Z sous la forme suivante :

$$f_{S,R,Y|Z}(S,R,Y|Z,\theta,\psi) = f_{Y|Z}(Y|Z,\theta)\, f_{S|Y,Z}(S|Z,Y,\psi)\, f_{R|S,Y,Z}(R|Z,Y,S,\phi), \tag{2.3}$$

en ajoutant à l'équation (2.1) un modèle correspondant à une non-réponse totale ayant une densité $f_{R|S,Y,Z}$. La modélisation des indicateurs pour les schémas de non-réponse partielle traite également les données manquantes partielles (par exemple Little, 2003b).

Le traitement de S, R et Y comme des variables aléatoires est une des principales caractéristiques de Rubin (1978) qui est, selon moi, un des articles phares de l'histoire de la statistique. L'article présente des conditions selon lesquelles les données manquantes et les mécanismes de sélection sont ignorables, c'est-à-dire qu'ils n'ont pas à être modélisés pour une inférence fondée sur la vraisemblance, en élargissant les définitions de l'ignorabilité aux données manquantes, dans Rubin (1976), tout en procurant un cadre pour

l'inférence en cas de sélection ou de données manquantes non ignorables. L'importance de l'article pour l'échantillonnage passe facilement inaperçue, puisque son axe principal est le rôle du mécanisme d'attribution de traitement en cas d'inférence relative à des effets de causalité. Le mécanisme d'attribution est ignorable dans une attribution de traitement aléatoire, comme dans des essais cliniques aléatoires. L'article expose ensuite le cadre général des inférences causales en comparant les traitements, caractéristique qui a valu à l'article sa popularité. Toutefois, il fournit également une justification bayésienne de l'échantillonnage aléatoire, en tant que moyen pour éviter le recours à un modèle pour procéder à la sélection.

Dans la modélisation de superpopulation fréquentiste (par exemple Valliant, Dorfman et Royall, 2000), les paramètres des modèles sont traités comme étant fixes; dans une modélisation bayésienne pour les enquêtes, ces paramètres sont attribués dans le cadre d'une distribution *a priori*, et les inférences pour $Q(Y)$ reposent sur sa distribution prédictive *a posteriori* en fonction des données. Dans les grands échantillons, la distribution *a priori* joue un rôle mineur, et les deux méthodes procurent des réponses semblables pour des modèles comparables; plus particulièrement, l'estimation du maximum de vraisemblance d'un paramètre constitue essentiellement le mode de distribution *a posteriori* dans le cas d'une distribution *a priori* uniforme, et a donc une interprétation bayésienne. Dans les petits échantillons, l'incertitude entourant les paramètres du modèle se propage lorsqu'ils sont intégrés à la distribution *a posteriori*. Cette approche de la propagation de l'erreur dans les paramètres permet aux inférences bayésiennes des modèles et des distributions *a priori* judicieusement choisis d'être mieux étalonnées que les inférences issues de la modélisation de superpopulation, dans le sens où elles ont de meilleures propriétés fréquentistes dans un échantillonnage répété (Rubin, 1978). Ainsi, à mon avis, la « modélisation de superpopulation est super, mais Bayes la surpasse ».

3. Inférence fondée sur le plan de sondage et inférence fondée sur un modèle

Les ouvrages portant sur l'échantillonnage font l'objet de nombreuses controverses (par exemple Smith, 1976, 1994; Kish, 1995; Brewer, 2013; Little, 2014) entre l'inférence « fondée sur le plan de sondage » qui repose sur la distribution de l'échantillonnage (2.2) et l'inférence « fondée sur un modèle » qui repose sur la distribution du modèle $f_{Y|Z}\left(Y \middle| Z, \theta\right)$ si la sélection est ignorable, ou sur la distribution complète du modèle (2.1) ou (2.3) si la sélection est non ignorable ou en présence de non-réponse. Je cherche des inférences fondées à la fois sur le plan et sur un modèle, par le fait qu'elles découlent de modèles bayésiens, mais qui ont de bonnes propriétés fondées sur le plan.

Rubin (1984) fait une distinction entre une *inférence statistique pour un ensemble de données particulier*, et les *propriétés* de cette inférence – convergence, couverture de confiance – dans un

échantillonnage répété. Pour être généralement crédible, l'inférence doit aussi afficher de bonnes propriétés d'échantillonnage répété. L'objectif de l'approche « fondée sur le plan de sondage » doit aussi en être un pour les modèles bayésiens pour les enquêtes – il faut choisir le modèle et la distribution *a priori* de façon à obtenir des inférences ayant de bonnes propriétés fondées sur le plan. Pour y parvenir, les caractéristiques du plan d'échantillonnage probabiliste complexe doivent faire partie du modèle – stratification et pondération intégrées à l'aide de covariables, échantillonnage à plusieurs degrés intégré à l'aide de modèles hiérarchiques. L'inclusion d'une distribution *a priori* dans la modélisation bayésienne, décriée par certains comme constituant une autre hypothèse à ajouter au modèle, me fournit un outil complémentaire à la modélisation de superpopulation. Elle procure une plus grande souplesse que la modélisation de superpopulation, ce qui restreint efficacement le nombre de distributions *a priori* uniformes possibles.

En plus d'avoir de bonnes propriétés fréquentistes, l'inférence qui repose sur q et I doit être appropriée – Rubin (2019) emploie le mot « pertinente » – à l'ensemble de données réalisé. Supposons que D désigne les données à la base de l'inférence et $\tilde{D}$, la réalisation particulière de D, les valeurs réellement obtenues de l'échantillon et du répondant. Que q et I découlent d'un modèle bayésien formel, d'une équation estimative ou d'une procédure algorithmique, ils doivent fournir une bonne inférence pour les données $\tilde{D}$, pas d'autres ensembles de données D qui seraient possiblement obtenus. Les méthodes bayésiennes ont tendance à présenter cette propriété en raison des conditions de distribution *a posteriori* appliquées à $\tilde{D}$; mais il faut également un intervalle de confiance approximativement valide s'il est considéré comme un intervalle crédible qui conditionne $\tilde{D}$, ne serait-ce que parce qu'il s'agit de la façon dont un non-statisticien aurait tendance à l'interpréter. Cette perspective manque parfois aux intervalles de confiance fondés sur le plan, comme le montrent les exemples 2 et 4 ci-dessous.

En somme, l'un des objectifs communs de l'inférence fondée sur le plan de sondage et de celle fondée sur un modèle consiste à obtenir une valeur de q qui utilise à bon escient les données, présente certaines propriétés comme la convergence par rapport au plan qui suppose qu'elle n'est pas trop éloignée de Q, et un intervalle I qui est aussi étroit que le permettent les renseignements contenus dans les données, tout en incluant Q selon une probabilité près de la valeur nominale de 95 %. Rubin (2019) associe ces propriétés à un statisticien « rempli de sagesse », dans son discours Waksberg.

Les méthodes fondées sur le plan sont souvent simplifiées pour éviter le recours à un modèle, parce que les propriétés comme la convergence par rapport au plan ne reposent pas sur un modèle, pour les données. Cependant, le rendement des méthodes fondées sur le plan dépend souvent d'un modèle implicite, et la modification de l'estimation reposant sur un modèle plus réaliste peut améliorer l'inférence, selon une perspective fondée sur le plan de sondage ou sur un modèle. Ce point est illustré dans les exemples 3 à 5 ci-dessous.

La question de l'ensemble de références approprié pour les propriétés d'un échantillonnage répété, comme les intervalles de confiance, est truffée de difficultés, surtout en ce qui concerne l'adéquation d'appliquer un conditionnement sur des statistiques ancillaires ou sur celles s'en rapprochant (par exemple Birnbaum, 1962; Berger et Wolpert, 1988; Ghosh, Reid et Fraser, 2010). L'évaluation des propriétés d'une inférence bayésienne pour un échantillonnage répété soulève également ces questions, mais ces dernières ne s'appliquent pas à l'inférence elle-même parce que la distribution *a posteriori* conditionne $\tilde{D}$.

L'approche fondée sur le plan de sondage, pour l'inférence pour les enquêtes par échantillonnage, comporte une portée trop limitée, ce qui empêche de traiter adéquatement bon nombre des problèmes liés à ce type d'enquêtes, dans la pratique. Elle comporte les limites suivantes :

1. L'inférence fondée sur le plan de sondage est asymptotique, et elle ne procure pas d'inférences valides dans de petits échantillons. Examinons l'exemple simple qui suit.

   Exemple 1. Inférence concernant la moyenne de la population par échantillonnage aléatoire simple. Considérons l'inférence concernant la moyenne de la population d'une variable Y tirée d'un échantillon aléatoire simple de taille n à partir d'une population de taille N. L'intervalle de confiance type à 95 %, fondé sur le plan, prend la forme suivante

   $$\bar{y} \pm 1{,}96\, s\sqrt{(1/n - 1/N)}, \tag{3.1}$$

   où $\bar{y}$ est la moyenne de l'échantillon et s, l'écart-type de l'échantillon. Cet intervalle est asymptotique et ne fournit pas d'inférences valides avec petit échantillon. Si, notamment, Y est continu, on obtient habituellement une meilleure inférence en remplaçant 1,96, le 97,5^{e} centile de la distribution normale, par le 97,5^{e} centile d'une distribution qui tient compte de l'incertitude à propos de l'estimation de la variance, comme la distribution t comportant $n-1$ degrés de liberté. La procédure présume toutefois une distribution normale pour Y et ainsi elle n'est pas fondée sur le plan. Si Y est binaire, avec les valeurs 0 et 1, alors $\bar{y}$ est la proportion de l'échantillon, $s = \sqrt{\bar{y}(1-\bar{y})}$ et (3.1) est l'intervalle de Wald asymptotique, qui a un très mauvais rendement avec les petits échantillons, surtout lorsque la proportion réelle se rapproche de 0 ou de 1. L'intervalle bayésien crédible pour l'*a priori* de Jeffreys ou uniforme présente de bien meilleures propriétés fréquentistes. Consultez les travaux de Dean et Pagano (2015) et de Franco, Little, Louis et Slud (2019) qui présentent des comparaisons des intervalles de Wald avec des solutions de rechange pour les plans de sondage complexes. L'inférence fondée sur le plan de sondage est souvent une mauvaise option pour les petits échantillons, plus particulièrement pour l'estimation sur petits domaines où on fait appel à un modèle pour Y afin d'« obtenir un renforcement par emprunt de données » aux différents domaines.

2. Une inférence fondée sur le plan de sondage ne gère pas d'unité d'enquête, de non-réponse partielle ou d'erreur de réponse, parce que ces problèmes nécessitent des modèles qui génèrent habituellement des résultats satisfaisants.
3. Une inférence fondée sur le plan de sondage n'est pas normative, d'une manière permettant de prescrire un choix approprié de méthode d'inférence pour les données à traiter. Le choix approprié d'un estimateur nécessite effectivement un modèle implicite, comme dans l'estimation « assistée par un modèle » (par exemple Särndal, Swensson and Wretman, 1992). L'estimateur de régression ou par quotient, par exemple, l'intégration d'information auxiliaire, ou l'estimateur de Horvitz-Thompson ou de Hájek pour l'insertion de poids d'enquête, reposent tous sur des modèles implicites, et si ce modèle est loin d'être réaliste, ces méthodes se révéleraient très sous-optimales – les éléphants de Basu (1971) constituant un exemple extrême et satirique. L'inférence bayésienne fondée sur des modèles plus souples tend à donner de meilleurs résultats, comme il sera établi dans l'exemple 5 ci-dessous.
4. Une inférence fondée sur le plan de sondage ne traite pas de la manière de générer des inférences pour des échantillonnages non probabilistes, qui sont de plus en plus utilisés en raison du coût et des difficultés liées à l'obtention de vrais échantillons aléatoires.

Le point 4 ne permet pas d'exclure le recours à des méthodes fondées sur le plan pour obtenir q et I, parce que nous pouvons toujours prétendre que nous disposons d'un échantillon probabiliste en adoptant un modèle pour les indicateurs de sélection S qui définit l'inclusion dans l'échantillon, et en estimant les paramètres inconnus dans ce modèle (par exemple Elliott et Valliant, 2017). Les statisticiens qui emploient des méthodes fondées sur le plan pour une inférence tirée d'échantillons aléatoires ont tendance à privilégier cette méthode de « quasi-randomisation ». Cette méthode comporte toutefois les mêmes limites que l'approche fondée sur le plan de sondage concernant les échantillons probabilistes, à savoir l'incapacité à gérer de petits échantillons, les données manquantes ou les erreurs de réponse; la boîte à outils bayésienne est beaucoup mieux garnie pour traiter ces problèmes.

4. Exemples

Une variété d'exemples sont présentés, certes plutôt axés sur mes propres travaux réalisés avec des collègues. J'évite les sujets comme l'estimation de petits domaines ou de séries chronologiques, parce que le besoin de modélisation est bien établi, dans ces cas.

Exemple 1 (suite). Échantillon aléatoire normal. Une des critiques souvent entendues à propos des méthodes fondées sur un modèle soutient que « si mon inférence repose sur un modèle et que ce modèle est incorrect, l'inférence le sera également. Puisque (pour paraphraser George Box) tous les modèles sont

incorrects, toutes les inférences fondées sur un modèle sont par conséquent incorrectes. Je préfère donc une inférence fondée sur le plan de sondage, qui ne demande pas d'hypothèses de modélisation ». Le raisonnement, quoique plausible, n'est pas si simple. La validité des méthodes fondées sur un modèle dépend du plan, de même que du degré et de la nature de l'erreur de spécification du modèle; le fait que les méthodes fondées sur le plan ne dépendent pas nécessairement des modèles ne constitue pas un signe incontestable de leur supériorité par rapport à ceux qui en dépendent.

Supposons qu'un échantillon aléatoire simple de taille n est prélevé à partir d'une distribution continue ayant une moyenne μ inconnue et un écart-type σ. Examinons trois estimations d'intervalle pour la moyenne de la population :

(i) L'intervalle A est l'intervalle de confiance type à 95 %, fondé sur le plan, équation (3.1).

(ii) L'intervalle B remplace le 97,5^{e} centile normal dans l'équation (3.1), à savoir 1,96, par le 97,5^{e} centile de la distribution t avec $n-1$ df.

(iii) L'intervalle C est l'intervalle de crédibilité *a posteriori* à 95 % fondé sur un modèle normal, avec une distribution $p(\mu,\sigma) \propto 1/\sigma$ selon l'*a priori* de Jeffreys.

D'un point de vue fréquentiste, lequel de ces intervalles est le plus juste ? L'intervalle A ne s'appuie sur aucune hypothèse de distribution pour Y, et B s'appuie sur une hypothèse normale – faut-il en déduire que A est supérieur ? Si n est grand, les intervalles seront presque les mêmes, et si n est petit, alors B est sans doute mieux que A, même si les données ne sont pas normales, puisqu'il tient compte de l'incertitude à propos de la variance.

Dans une enquête informelle, les deux tiers d'une classe récente de nos étudiants de troisième cycle bien formés ont préféré B à C, faisant valoir qu'il évite de choisir une distribution *a priori* et, par conséquent, fait moins d'hypothèses. Mais B et C sont également bons ou mauvais, puisqu'il s'agit de la même procédure! Que l'intervalle de confiance à 95 % soit précis selon la normalité et qu'il soit un intervalle de crédibilité à 95 % pour le choix d'*a priori* mentionné constitue simplement deux propriétés de la procédure. Juger une méthode par ses hypothèses apparentes est un excès de simplification.

Exemple 2. Échantillon aléatoire simple avec borne inférieure sur la variance (Little, 2006). Supposons dans l'exemple précédent que $n=7$, $\bar{y}=1, s=1$. L'intervalle t normal (en ignorant les corrections d'une population finie) serait

$$\bar{y} \pm t_{6\ 975}\, s/\sqrt{n} \;=\; 1 \pm 2{,}447/\sqrt{7} = 1 \pm 0{,}925 \tag{4.1}$$

si, en fait, nous savons que $\sigma =$ 1,5, nous aurions un meilleur intervalle en

$$\bar{y} \pm z_{0{,}975}\, \sigma/\sqrt{n} \;=\; 1 \pm 1{,}96 \times 1{,}5/\sqrt{7} = 1 \pm 1{,}11, \tag{4.2}$$

l'intervalle le plus vaste tenant compte du fait que σ est plus grand que la valeur particulière de s pour cet échantillon. L'intervalle t (4.1) a une couverture de confiance exacte, mais compte tenu de ce que nous savons à propos de σ, il s'agit de la mauvaise inférence pour cet ensemble de données en particulier : nous ne devrions pas le retenir au détriment de (4.2) parce qu'il est plus étroit!

Supposons maintenant que nous savons que $\sigma > 1{,}5$, en raison d'une certaine source de variation additionnelle dont on n'a pas tenu compte. L'approche bayésienne intègre ensuite ce renseignement à la distribution *a priori* pour σ, ce qui génère un intervalle crédible plus large que dans (4.2). Quelle est la réponse fréquentiste ? L'intervalle de confiance t a toujours une couverture exactement nominale dans un échantillonnage répété, mais il est nettement trop étroit en tant qu'inférence pour l'ensemble de données observé parce qu'il ne tient pas compte de ce qui est connu à propos de σ. L'intervalle (4.2) est anticonservateur, malgré le fait qu'il soit plus grand que (4.1) pour l'ensemble de données réalisé – une propriété qu'il est possible d'observer pour des intervalles de confiance, mais pas pour des intervalles de crédibilité selon un modèle particulier. Des méthodes fréquentistes asymptotiques ne sont d'aucune aide dans le cas présent, alors quelle est la solution de rechange à Bayes dans cet exemple ?

Un exemple connexe apparaît dans une analyse de variance à effets aléatoires au sein d'un groupe, lorsque l'estimation des moindres carrés de la variance entre les groupes est négative – une analyse bayésienne aborde ce problème à l'aide d'une distribution *a priori* pour la variance entre les groupes qui ne permet pas d'obtenir des valeurs négatives. Les modèles à effets aléatoires et à effets mixtes sont importants pour traiter la mise en grappe dans les enquêtes, milieu dans lequel les méthodes bayésiennes fonctionnent mieux que le maximum de vraisemblance.

Exemple 3. Poststratification sur une covariable catégorique. La prévision (dans une modélisation) constitue une approche générale plus fiable de l'inférence que la pondération (dans une inférence fondée sur le plan de sondage). J'illustre cet énoncé général à l'aide d'un simple exemple de poststratification sur une seule variable Z. Vous trouverez ci-dessus un exemple plus complexe (exemple 7).

Examinons un échantillonnage probabiliste à probabilité égale comportant une seule variable catégorique de poststratification Z, pour laquelle le chiffre de population connu N_h est disponible pour chaque strate *a posteriori* h, $h = 1, 2, \ldots, H$. Supposons que $\bar{y}_h$ est la moyenne de l'échantillon dans une strate *a posteriori* h, d'après une taille d'échantillon de n_h, et $n = \sum_{h=1}^{H} n_h$, $N = \sum_{h=1}^{H} N_h$. L'estimation normale de la moyenne de la population est la moyenne poststratifiée :

$$\bar{y}_{\mathrm{PS}} = \sum_{h=1}^{H} P_h \bar{y}_h, \tag{4.3}$$

où $P_h = N_h / N$ et N correspond à la taille de la population. Cela peut être vu comme une moyenne pondérée

$$\bar{y}_{\mathrm{PS}} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_i \, y_i,$$

où $w_i = N_h / (N n_h)$ est le poids de poststratification pour les unités échantillonnées dans une strate *a posteriori* h. Ces poids peuvent être très importants dans les strates « *a posteriori* » avec des petites tailles d'échantillon n_h, qui, contrairement à un échantillonnage stratifié qui repose sur Z, ne relèvent pas du contrôle de l'échantillonneur. Ces poids imposants peuvent entraîner une variabilité excessive dans $\bar{y}_{\mathrm{PS}}$. En fait, à proprement dit, $\bar{y}_{\mathrm{PS}}$ n'a pas de distribution dans un échantillonnage répété, car avec une probabilité positive, les tailles d'échantillon dans certaines strates *a posteriori* peuvent être nulles. Cela demeure vrai en cas de modification des strates *a posteriori* visant à assurer une valeur positive aux dénombrements d'échantillon des strates *a posteriori* de tout l'échantillon observé, par exemple par la combinaison de strates adjacentes.

L'approche type fondée sur le plan de sondage pour une variabilité excessive de $\bar{y}_{\mathrm{PS}}$ consiste à modifier les poids $\{w_i\}$, par exemple en éliminant les plus grands. Cependant, d'un point de vue prévisionnel, cette stratégie n'est pas une bonne idée. Le problème ne se situe pas au niveau des poids – les proportions de la population $\{P_h\}$ de chaque strate *a posteriori* sont connues, après tout – mais bien dans le manque de données de l'échantillon dans certaines strates *a posteriori* qui rend les estimations $\{\bar{y}_h\}$ peu fiables. Ce sont les estimations des strates *a posteriori* ayant peu de données qu'il faut modifier, non les poids $\{w_i\}$ rattachés aux unités échantillonnées. Selon les principes, le moyen de modifier $\{\bar{y}_h\}$ consiste à adopter un modèle qui relie Y et Z. L'approche fondée sur le plan de sondage, en évitant d'adopter ce modèle, mène au mauvais principe – modifier les poids au lieu des prévisions de valeurs non échantillonnées.

Point connexe, une approximation courante fondée sur le plan (Kish, 1992) mesure l'augmentation proportionnelle de la variance par rapport à la pondération en tant que $1 + \mathrm{cv}(w_i)$, où $\mathrm{cv}(w_i)$ est le coefficient de variation des poids; un élagage des poids réduit $\mathrm{cv}(w_i)$ ce qui entraîne par conséquent une augmentation proportionnelle. Cependant, cette règle empirique n'est valide que lorsque Y et Z ne sont pas liés, auquel cas une poststratification est inutile. Si Y et Z sont liés et que la taille de l'échantillon n'est pas trop petite, la variance de la moyenne poststratifiée est *plus petite*, non *plus grande*, que la variance de la moyenne non pondérée de l'échantillon (Holt et Smith, 1979; Little et Vartivarian, 2005). La règle empirique ne fonctionne pas parce que la relation entre Y et Z n'est pas modélisée.

Quelle est l'approche bayésienne face à une variabilité excessive de $\bar{y}_{\mathrm{PS}}$? Cette dernière représente la moyenne *a posteriori* du modèle normal stratifié

$$\left(y_{hi} \,|\, \mu_h, \sigma^2\right) \overset{\text{iid}}{\sim} G\left(\mu_h, \sigma^2\right), \tag{4.4}$$

$$p\left(\mu_h, \sigma\right) \propto 1/\sigma, \tag{4.5}$$

où $G\left(a, b^2\right)$ désigne la distribution normale (gaussienne) ayant une moyenne a, une variance b^2 et l'équation (4.5) est la distribution *a priori* de Jeffreys sur la moyenne et l'écart-type pour chaque strate *a posteriori*. Puisque l'échantillon compte quelques strates *a posteriori* ayant peu de données, la distribution *a priori* doit être modifiée afin de permettre un renforcement par emprunt d'information tirée d'autres strates. Une des méthodes consiste à adopter un modèle normal à effets aléatoires

$$p\left(\mu_h \,|\, \mu, \tau, \sigma\right) \overset{\text{iid}}{\sim} G\left(\mu, \tau^2\right),\ p\left(\mu, \sigma, \tau\right) \propto \sigma^{-1}, \tag{4.6}$$

qui traite les moyennes de strate comme des effets aléatoires. Il serait aussi possible de traiter les variances dans chaque strate *a posteriori* comme des effets aléatoires distincts et les soumettre à une distribution *a priori*, au lieu de les regrouper. La moyenne *a posteriori* de $\bar{Y}$ pour la distribution *a priori* (4.6) déplace le poids w_i des unités échantillonnées dans la strate *a posteriori* h vers la valeur 1, avec un degré de rétrécissement qui dépend de la taille relative des estimations de σ et de τ (Lazzeroni et Little, 1998).

La distribution *a priori* (4.6) donne lieu à l'hypothèse non négligeable que les moyennes des strates *a posteriori* sont échangeables. Il est possible de la modérer en limitant le modèle à effets aléatoires à un sous-ensemble de strates *a posteriori* ayant peu d'unités échantillonnées; ou en remplaçant la moyenne constante μ dans (4.6) par une régression des caractéristiques C_h connues des strates *a posteriori*, comme dans :

$$p\left(\mu_h \,|\, \beta_0, \beta_1, \tau, \sigma, c_h\right) \overset{\text{ind}}{\sim} G\left(\beta_0 + \beta_1 c_h, \tau^2\right),\ p\left(\beta_0, \beta_1, \sigma, \tau\right) \propto \sigma^{-1}, \tag{4.7}$$

qui limite l'hypothèse de l'échangeabilité aux erreurs de la régression de μ_h sur c_h. Pour des généralisations complètes de cet exemple de base, consultez les travaux de Gelman et Little (1997), d'Elliott et Little (2000), d'Elliott (2007), de Gelman (2007) et de Si, Trangucci, Gabry et Gelman (2020).

Exemple 4. Estimateur par la régression de la moyenne, compte tenu d'une variable auxiliaire de la population. Si la variable auxiliaire Z de l'exemple précédent est continue, une façon habituelle de l'intégrer à l'inférence consiste à utiliser une estimation par régression de la moyenne :

$$q = \bar{y}_{\text{REG}} = \bar{y} + \hat{\beta}_1\left(\bar{Z} - \bar{z}\right),$$

où $\hat{\beta}_1$ est l'estimation des moindres carrés de la pente de Y sur Z dans l'échantillon, alors que $\bar{z}$ et $\bar{Z}$ correspondent respectivement à la moyenne de l'échantillon et de la population de Z. Dans une étude de simulation de cinq populations réelles, Royall et Cumberland (1981, 1985) évaluent les inférences axées sur q, avec (a) l'erreur type fondée sur le plan reposant sur un échantillonnage aléatoire simple, à savoir :

$$I_{0{,}95D} = \bar{y}_{\text{reg}} \pm 1{,}96\hat{\text{se}}_D\left(\hat{\beta}\right),\ \hat{\text{se}}_D\left(\hat{\beta}\right) = \sqrt{\left(1-f\right)s_{Y.Z}/n},$$

où $s^2_{Y.Z}$ est la variance résiduelle de l'échantillon et $f = n/N$, la fraction d'échantillonnage; et (b) l'erreur type de la prévision, reposant sur le modèle de régression linéaire normal ayant une variance constante, à savoir :

$$
\begin{aligned}
I_{0,95M} &= \bar{y}_{\text{reg}} \pm t_{0,975,\,n-2} 1{,}96 \hat{\text{se}}_M\left(\hat{\beta}\right), \\
\hat{\text{se}}_M\left(\hat{\beta}\right) &= \hat{\text{se}}_D\left(\hat{\beta}\right)\sqrt{\left(1+\left(\bar{z}-\bar{Z}\right)^2\right)\Big/\left(1-f\right)\left(\sum_{i=1}^{n}\left(z_i-\bar{z}\right)^2/n\right)}.
\end{aligned}
$$

Les intervalles de confiance fondés sur le plan $I_{0,95D}$ affichent une très faible couverture de confiance conditionnelle lorsque le $\bar{z}$ observé dévie nettement de $\bar{Z}$. L'intervalle de confiance fondé sur un modèle $I_{0,95M}$ tient compte du manque d'équilibre concernant $\bar{z}$, mais est sensible à une erreur de spécification du modèle, surtout à un manque de linéarité dans la relation entre Y et Z ou à une variance résiduelle non constante. De solides estimations de l'erreur type, reposant sur l'estimateur sandwich ou l'estimateur jackknife, procurent des intervalles ayant de meilleures propriétés de couverture conditionnelle, même si on observe encore parfois une déviation par rapport aux taux de couverture nominaux. Une démarche de rechange serait l'inférence bayésienne, reposant sur un modèle plus souple reliant Y à Z, comme le modèle de spline pénalisée :

$$
\begin{aligned}
&\left(y_i \,|\, z_i, \beta, \sigma^2\right) \overset{\text{ind}}{\sim} G\left(\text{spline}\left(z_i, \beta\right), \sigma^2 z_i^{\alpha}\right), \\
&\text{spline}\left(z_i, \beta\right) = \beta_0 + \sum_{j=1}^{p} \beta_j z_i^j + \sum_{\ell=1}^{m} \beta_{\ell+p}\left(z_i - \kappa_\ell\right)_+^p, \\
&\left(\beta_{l+p} \,|\, \tau\right) \overset{\text{iid}}{\sim} N\left(0, \tau^2\right), l = 1, \ldots, m;\ \ p\left(\beta_0, \ldots, \beta_p, \alpha, \sigma, \tau\right) \propto 1/\sigma, 0 < \alpha < 2,
\end{aligned}
\tag{4.8}
$$

où les constantes $\kappa_1 < \ldots < \kappa_m$ sont des nœuds fixes sélectionnés, et $(u)_+^p = u^p$ si $u > 0$ et 0, dans les autres cas (consulter, à titre d'exemple, Ruppert, Wand et Carroll, 2003). Le paramètre α permet d'avoir une variété de formes courantes d'hétéroscédasticité. Les erreurs types bayésiennes tiennent alors compte du déséquilibre de la distribution de Z dans l'échantillon et la population, alors que la souplesse du modèle limite le biais causé par une erreur de spécification du modèle.

Supposons que la quantité ciblée n'est pas la moyenne de la population de Y, mais la pente de la droite des moindres carrés de Y sur Z dans la population. Une démarche judicieuse consiste à imputer des valeurs non échantillonnées de Y à l'aide du modèle (4.8), puis d'estimer la pente de la droite de Y sur Z en tant que pente de la droite des moindres carrés pour les données sur la population insérées au préalable. Il est possible de propager l'incertitude grâce à une imputation multiple (Rubin, 1987), une méthode qui repose sur des principes bayésiens. Dans ce contexte, Little (2004) fait une distinction entre le « modèle ciblé » qui détermine la quantité de la population cible présentant un intérêt, la régression linéaire de Y sur X dans le cas présent, et le « modèle de travail » (4.8) qui constitue le fondement d'une inférence et sert à prévoir les variables d'enquête pour les unités non échantillonnées et non répondantes de la population. La distinction entre ces deux modèles procure une solide forme d'inférence bayésienne pour les enquêtes.

Szpiro, Rice et Lumley (2010) appliquent un principe semblable dans le réglage d'une régression de la superpopulation, et Little (2019) avance que cela se révèle plus simple que de modifier l'interprétation du paramètre, la démarche adoptée par Buja, Berk, Brown, George, Pitkin, Zhan et Zhang (2019). Szpiro

et coll. (2010) montrent que la démarche mène à une interprétation bayésienne de l'estimateur sandwich de la variance dans une régression, qui équivaut sur le plan asymptotique aux estimations de la variance par réutilisation de l'échantillon, comme le bootstrap ou le jackknife, qui sont souvent appliquées dans le cadre d'une enquête par échantillonnage.

Exemple 5. Inférence pour des échantillons ayant des probabilités inégales de sélection. En ce qui concerne les plans ayant des probabilités inégales de sélection, les articles classiques critiquant la méthode de modélisation des enquêtes (Kish et Frankel, 1974; Hansen, Madow et Tepping, 1983) n'incluent pas les probabilités de sélection dans le modèle, ce qui donne lieu à des inférences sensibles à une erreur de spécification du modèle. Les probabilités de sélection jouent un rôle de premier plan dans la solidité d'une inférence fondée sur un modèle, mais en tant que covariables du modèle plutôt que poids d'échantillonnage.

Examinons, par exemple, une inférence relative à la moyenne de la population $\bar{Y}$. Si y_i est la valeur de la variable d'enquête Y et π_i est la probabilité de sélection de l'unité i, l'estimateur habituel de la moyenne de Y, fondé sur le plan, pondère les unités échantillonnées (disons les unités $i=1,\ldots,n$) par l'inverse de π_i qui nous donne l'estimation de Horvitz-Thompson (Horvitz et Thompson, 1952)

$$\bar{y}_{\mathrm{HT}} = N^{-1}\sum_{i=1}^{n} y_i/\pi_i, \tag{4.9}$$

si la taille de la population N est connue, ou l'estimation de Hájek (1971)

$$\bar{y}_{\mathrm{HK}} = \left(\sum_{i=1}^{n} y_i/\pi_i\right)\Big/\left(\sum_{i=1}^{n} 1/\pi_i\right), \tag{4.10}$$

si N est estimé. La pondération est une démarche universelle où les unités échantillonnées obtiennent le même poids, sans tenir compte de la relation entre π_i et y_i. Cela se révélera possiblement inefficace – si la relation est faible, la pondération aura simplement pour effet de réduire la précision de l'estimation en l'absence d'une réduction compensatoire du biais.

La méthode de modélisation intègre les probabilités de sélection en procédant à une régression de y_i sur π_i. La force de la relation entre y_i et π_i modère ensuite l'effet qu'aura la probabilité de sélection sur l'estimateur – si la relation est faible, le coefficient de régression de π_i est modeste et le poids d'échantillonnage a peu d'incidence. Cela procure des estimations plus efficaces.

Une régression linéaire de y_i sur π_i est sensible au biais en cas d'erreur de spécification de la linéarité, mais on peut réduire l'incidence de cette erreur en choisissant un modèle aboutissant à une estimation convergente par rapport au plan. De nombreux modèles satisfont à cette exigence. À titre d'exemple, voir Firth et Bennett (1998).

En cas d'échantillonnage stratifié à l'aide d'une variable de stratification Z, le modèle de régression naturel comporte des covariables qui consistent en variables indicatrices pour la strate. L'estimation résultante de la moyenne de la population d'une variable d'enquête continue Y est la moyenne stratifiée,

qui prend la même forme que dans l'équation (4.3), mais pour laquelle Z forme des strates au lieu de strates *a posteriori*. Dans cette situation, la pondération par l'inverse de la probabilité de sélection de chaque strate et la régression par variable indicatrice aboutiront toutes deux à l'estimateur (4.3). Les méthodes bayésiennes de réduction de la variance décrites dans l'exemple 3 tendent à être moins rentables dans le cas d'un échantillonnage stratifié qu'en poststratification, parce que l'échantillonneur a le contrôle sur la taille des échantillons de chaque strate.

Dans un échantillonnage de probabilité proportionnelle à la taille (PPT), la covariable Z correspond à la taille de l'unité et de $\pi_i = \min(cz_i, 1)$. Maintenant, puisque Z est une variable continue, la pondération et la régression pourraient donner des réponses différentes. On pourrait unifier les méthodes en prenant en considération les modèles qui procurent des estimations pondérées selon le plan de sondage lorsqu'ils sont utilisés pour prévoir les unités non échantillonnées. Plus particulièrement, en ignorant les corrections de la population finie, l'estimation de Horvitz-Thompson (4.9) est la moyenne *a posteriori* pour le « modèle Horvitz-Thompson ».

$$\left(y_i \,|\, z_i, \beta, \sigma^2\right) \overset{\text{ind}}{\sim} G\left(\beta z_i, \sigma^2 z_i^2\right),\ p(\beta, \sigma) \propto 1/\sigma, \tag{4.11}$$

et l'estimation Hájek (4.10), la moyenne *a posteriori* pour le « modèle Hájek » :

$$\left(y_i \,|\, z_i, \beta, \sigma^2\right) \overset{\text{ind}}{\sim} G\left(\beta, \sigma^2 z_i\right),\ p(\beta, \sigma) \propto 1/\sigma. \tag{4.12}$$

Ces modèles sous-jacents décrivent des situations dans lesquelles les estimations fondées sur le plan correspondantes sont optimales. Ils comportent toutefois de fortes hypothèses paramétriques. Une bonne méthode de modélisation bayésienne intègre ces modèles dans un modèle plus vaste, comme le modèle de spline pénalisée (4.8), comme l'ont proposé Zheng et Little (2003). Zheng et Little (2005) et Chen, Elliott, Haziza, Yang, Ghosh, Little, Sedransk et Thompson (2017) présentent des études de simulation indiquant que cette méthode de modélisation peut procurer des gains substantiels par rapport à une estimation de Horvitz-Thompson ou de Hájek, à la fois sur le plan de l'efficacité et du rapprochement de la couverture de confiance nominale (fréquentiste) dans des échantillons modérés – rappelons-nous que les résultats fondés sur le plan sont asymptotiques. Le modèle (4.8) s'agrandit facilement pour permettre l'insertion d'autres variables auxiliaires mesurées pour toutes les unités de population, et la souplesse des inférences de petits échantillons s'accroît grâce à l'inclusion de distributions *a priori* qui conviennent pour les paramètres du modèle.

Leon-Novelo et Savitsky (2019) examinent des modèles bayésiens pour la distribution conjointe de y_i et π_i, en appelant les modèles qui corrigent π_i des méthodes « d'extension ». Cependant, l'attribution d'une distribution à π_i me semble à la fois artificielle et inutile dans ce réglage; si on enregistre π_i pour toutes les unités de population, il peut être conditionné dans le modèle, comme dans des méthodes de régression classiques qui traitent les covariables comme des valeurs fixes. Même si les valeurs de π_i pour les unités non échantillonnées ne sont pas fournies aux analystes, il est possible de prévoir leurs valeurs à l'aide d'un théorème bayésien, comme dans Zangeneh et Little (2015).

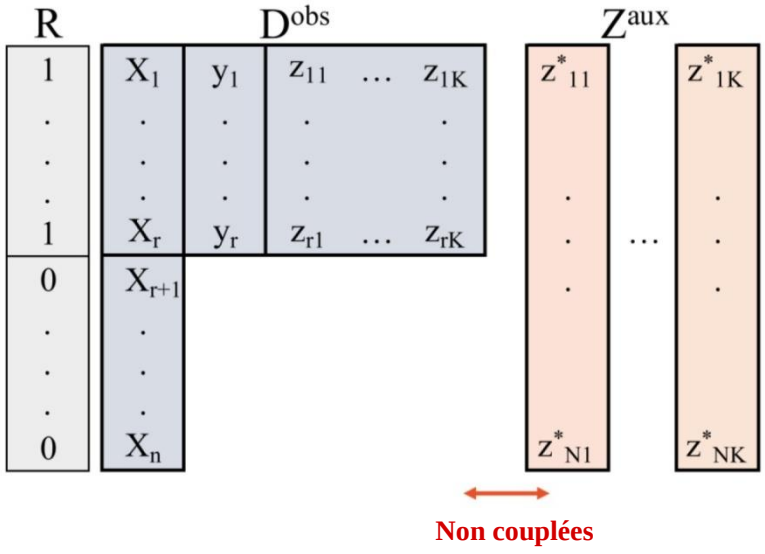
Exemple 6. Spline pénalisée de modèles de propension de réponse. Examinons la non-réponse totale pour une variable d'enquête Y, dans le cadre de l'observation d'un ensemble de variables, disons

$X_1, \ldots, X_p$ concernant les répondants et les non-répondants de l'échantillon. La propension de réponse pour l'unité i, $\theta_i = \Pr\left(R_i = 1 \middle| x_{i1}, \ldots, x_{ip}, \psi\right)$, où ψ représente des paramètres inconnus, joue un rôle important dans les méthodes de pondération et de prévision relatives à la non-réponse à l'enquête. Dans le cas de la pondération de la non-réponse, le poids d'échantillonnage est multiplié par le poids de non-réponse, de la façon suivante

$$\begin{aligned} w_i &= 1/\Pr(\text{l'unité } i \text{ est sélectionnée et répond}) \\ &= (1/\Pr(i \text{ sélectionnée})) \times (1/\Pr(i \text{ répond} \mid \text{sélectionnée})) \\ &\quad \text{poids d'échantillonnage} \times \text{poids de réponse.} \end{aligned}$$

Contrairement au poids d'échantillonnage, le poids de réponse est inconnu, et il faut indiquer clairement ce sur quoi la probabilité dépend. Little (2022) avance qu'elle doit conditionner les variables auxiliaires et d'enquête, mais pas d'autres variables qui pourraient avoir un effet sur elle. Comme pour les poids d'échantillonnage, une pondération par l'inverse de la propension de réponse estimée est une méthode universelle qui ne tient pas compte de la force de la relation entre R et Y. Une spline pénalisée de prévision de la propension (Zhang et Little, 2009) fait régresser Y sur la spline pénalisée de la propension de réponse estimée, alors que d'autres variables observées pour les répondants et les non-répondants apparaissent sur le plan paramétrique. La spline modélise une relation souple entre la propension de réponse et Y, ce qui donne de la robustesse à une erreur de spécification du modèle. La propriété équilibrante du score de propension procure également une double propriété de robustesse, en cela que la forme paramétrique de la régression des autres prédicteurs peuvent comporter une erreur de spécification sans entraîner de biais, à condition que la spline pénalisée saisisse la relation entre la propension et Y. Cette méthode fonctionne bien avec les méthodes de pondération, dans les simulations (Zhang et Little, 2009; Yang et Little, 2015), et la version exclusivement bayésienne avec distributions *a priori* sur les paramètres propage l'erreur dans l'estimation des propensions.

Figure 4.1 Structure des données dans l'exemple 7.



Exemple 7. Non-réponse totale avec poststratification. Un autre exemple dans lequel les approches bayésiennes et les approches fondées sur le plan de sondage qui ont trait à l'inférence diffèrent est celui de la non-réponse totale avec poststratification. Examinons le scénario de la figure 4.1, où Y est une variable d'enquête assujettie à une non-réponse, X est une variable observée pour toutes les unités de l'échantillon et où $Z^{\text{aux}} = (Z_1, \ldots, Z_K)$ est constitué d'un ensemble de variables auxiliaires catégoriques. La distribution conjointe de Z^{aux} est observée pour les répondants de l'enquête, et les distributions marginales de Z_k, $k = 1, \ldots, K$, sont aussi observées pour la population ou l'échantillon provenant de sources externes. Les unités des données auxiliaires ont une caractéristique distinctive, celle de ne pas être couplées aux unités de l'enquête. Ce scénario apparaît souvent dans les réglages où une poststratification sert à ajuster la non-réponse. Définissons R_i en tant qu'indicateur de réponse pour (y_i, z_i) dans une unité d'échantillonnage i, et supposons que

$$\Pr(R_i = 1 | x_i, z_i, y_i, \psi) = \Pr(R_i = 1 | x_i, z_i, \psi), \tag{4.13}$$

faisant en sorte que cette probabilité ne dépende pas de y_i. Il convient de souligner que si cette probabilité dépend de z_i, le mécanisme de réponse n'est pas manquant au hasard dans le sens de Rubin (1976), parce que les valeurs de z_i sont manquantes pour les non-répondants de l'enquête. Zangeneh et Little (2022) prennent en considération l'estimation bayésienne et l'estimation par maximum de vraisemblance pour les données présentant ce modèle. En considérant, pour simplifier, des modèles de type i.i.d. pour les unités i, la distribution conjointe de X, Z, Y et R est prise en compte de la façon suivante

$$\begin{aligned} f_{X,Z,Y,R}(x_i, z_i, y_i, r_i | \theta, \phi) &= f_{Y|X,Z,R}(y_i | x_i, z_i, \theta, R_i = r_i)\, f_{X,Z,R}(x_i, z_i, r_i | \phi) \\ &\overset{\text{selon l'éq. (4.13)}}{=} f_{Y|X,Z,R}(y_i | x_i, z_i, \theta)\, f_{X,Z,R}(x_i, z_i, r_i | \phi), \end{aligned}$$

où θ et ϕ sont des paramètres distincts (Little et Rubin, 2019, chapitre 6). Il est ensuite possible d'estimer les paramètres θ de cette factorisation à partir des données d'enquête des répondants $R_i = 1$, de même que les paramètres de la distribution conjointe de X et Z par la combinaison des données des répondants et des données auxiliaires pour ces variables.

Voici deux cas particuliers examinés par Zangeneh et Little (2022) :

(1) Aucune covariable X et une seule variable de stratification *a posteriori* Z. L'hypothèse relative aux données manquantes de l'équation (4.13) se réduit ensuite à $\Pr(R_i = 1 | Z_i, Y_i, \psi) = \Pr(R_i = 1 | Z_i, \psi)$. L'estimation des paramètres de la distribution conditionnelle de Y compte tenu de Z est réalisée à partir des répondants de l'enquête, et celle des paramètres de la distribution marginale de Z, à partir des données auxiliaires sur Z. Si notamment Z est multinomial avec $\Pr(z_i = j | \phi) = \phi_j$, l'estimation à l'aide du maximum de vraisemblance de ϕ_j est tout simplement la proportion des chiffres auxiliaires dans la strate *a posteriori* j. Si, en outre, Y compte tenu de $Z = j$ est présumé normal, avec une moyenne μ_j et une variance σ_j^2, l'estimation par maximum de vraisemblance de la moyenne de la population Y qui en découle

est simplement la moyenne stratifiée *a posteriori* obtenue dans l'équation (4.3). Cet exemple simple est intéressant sur le plan théorique parce que le mécanisme de réponse n'est pas manquant au hasard, mais ignorable en ce qui a trait à l'inférence de vraisemblance; le fait d'être manquant au hasard est une condition suffisante, mais pas toujours nécessaire en matière d'ignorabilité.

(2) X est une seule variable catégorique et Z, un stratificateur catégorique *a posteriori*. Maintenant, la distribution conjointe de X, Z et R n'est pas mentionnée dans un modèle saturé et nécessite d'autres hypothèses afin d'établir le modèle. Un modèle « RAKE » restreint, avec données manquantes non au hasard, part de l'hypothèse que les distributions marginales de Z_1 et Z_2 sont différentes pour les répondants et les non-répondants, mais les rapports de cotes de Z_1 et Z_2 sont les mêmes pour les uns et les autres. Il en résulte un modèle exactement identifié. Le ratissage du tableau de fréquences des répondants de Z_1 et Z_2 jusqu'aux marges auxiliaires de Z_1 et Z_2 procure des estimations selon le maximum de vraisemblance de ϕ, selon ce modèle RAKE (Little et Wu, 1991; Little, 1993). L'estimateur poststratifié de la moyenne de Y est alors

$$\bar{Y}_{\text{rake}} = \sum_{j=1}^{J}\sum_{k=1}^{K} P_{jk}\left(\hat{\phi}\right)\bar{Y}_{jk}\left(\hat{\theta}\right),$$

où $P_{jk}\left(\hat{\phi}\right)$ est la proportion de la population pour $X = j, Z = k$ tirée du ratissage des fréquences de répondants jusqu'aux marges, et $\bar{Y}_{jk}\left(\hat{\theta}\right)$ est la moyenne estimée de Y compte tenu de $X = j, Z = k$ tirée du modèle pour Y compte tenu de X et Z.

Voici quelques commentaires concernant cette démarche :

(1) Il convient de souligner ici aussi qu'il s'agit d'un modèle avec des données manquantes non au hasard, et que ce modèle est moins fiable qu'un modèle ayant des données manquantes au hasard qui présume que la réponse dépend de X, mais pas de Z. Ce faisant, la méthode a tendance à afficher moins de biais que d'autres méthodes alternatives qui reposent sur une hypothèse de données manquantes au hasard pour obtenir des estimations convergentes.

(2) Comme dans le cas de l'exemple 4, la méthode du maximum de vraisemblance qu'emploie ce modèle ne comporte pas de modification des poids appliqués aux données dans une cellule $X = j, Z = k$ à l'aide de fonctions de distance arbitraires – les estimations selon le maximum de vraisemblance sont pleinement efficaces dans le modèle théorique.

(3) Il est possible de pallier la rareté de données dans les cellules par l'ajout de distributions *a priori* appropriées pour les paramètres et par l'application de méthodes bayésiennes. À titre d'exemple, pour l'inférence ayant trait aux paramètres θ de la distribution de Y compte tenu de $X = j, Z = k$, il serait possible de présumer d'une distribution *a priori* uniforme sur les

principaux effets de X et Z, mais d'une distribution *a priori* normale sur les interactions, en les faisant rétrécir jusqu'à zéro. Cela permet d'obtenir un rétrécissement de la moyenne de la cellule $\bar{y}_{jk}$ jusqu'aux moyennes ajustées tirées du modèle additif. Dans la terminologie fréquentiste, on parlera d'un modèle mixte d'analyse de la variance à effets principaux fixes et interactions aléatoires.

(4) Le principe de ratissage des marges X et Z est bien connu, mais il convient de souligner que si X est un stratificateur et Z, un stratificateur *a posteriori*, la démarche habituelle consiste à effectuer une itération du ratissage, en couplant d'abord la marge X, puis la marge Z : dans le modèle théorique, un ratissage itératif constitue la bonne procédure. Le ratissage correspond dans le cas présent à un maximum de vraisemblance, mais les formes bayésiennes de ratissage peuvent aussi servir à propager l'incertitude du paramètre.

Exemple 8. Analyse indirecte par modèles de mélange de schémas d'observation. Une analyse indirecte par modèles de mélange de schémas d'observation est une méthode d'évaluation du biais de non-réponse pour la moyenne d'une variable d'enquête Y assujettie à une non-réponse, en présence d'un ensemble de covariables observées pour les non-répondants et les répondants. Dans le passé, la quantité de données manquantes, mesurées en tant que taux de réponse, constituait la mesure la plus souvent utilisée pour évaluer la qualité d'une enquête. Mais les taux de réponse font abstraction des renseignements que renferment des covariables auxiliaires observées pour les non-répondants. Des méthodes reposant sur la probabilité estimée de non-réponse, comme l'indicateur R (Schouten, Cobben et Bethlehem, 2009) et la mesure q2 (Särndal et Lundström, 2010), ne tiennent pas compte de la solidité de l'association de la variable d'intérêt de l'enquête et de la probabilité de réponse, qui devrait sans doute être prise en compte dans l'évaluation du biais. Ces mesures supposent la présence de données manquantes au hasard, mais si des variables auxiliaires sont disponibles pour l'ajustement de la non-réponse, ce sont les déviations par rapport aux données manquantes au hasard qui entraînent un biais.

Andridge et Little (2011) proposent une analyse indirecte par modèles de mélange de schémas d'observation, méthode qui repose sur un modèle de mélange de schémas d'observation pour les non-réponses dans lequel sont combinées de manière simple et intuitive les principales caractéristiques de l'ajustement des non-réponses. Supposons que Y_i désigne la valeur d'un résultat d'enquête continu et $Z_i = \left(Z_{i1}, Z_{i2}, \ldots, Z_{ip}\right)$, la valeur des covariables p pour l'unité i dans l'échantillon. Seulement r des n unités échantillonnées répondent, de sorte que les données observées correspondent à $\left(Y_i, Z_i\right)$ pour $i = 1, \ldots, r$ et à Z_i pour $i = r+1, \ldots, n$. Supposons que R désigne l'indicateur de réponse, de sorte que pour l'unité i $R_i = 1$ si Y_i est observé et $R_i = 0$ si Y_i est manquant. Afin de réduire la dimensionnalité, nous remplaçons Z par une seule variable substitut X ayant la plus forte corrélation avec Y dans l'échantillon de répondants. Il est possible d'estimer cette variable substitut en effectuant une régression

de Y sur Z à l'aide des données des répondants, en incluant d'importants prédicteurs de Y, de même que des interactions et des termes non linéaires, si cela convient. Nous adoptons plus précisément le modèle de régression $E\left(Y \mid Z, R=1\right)=\alpha_0+\alpha Z$, et supposons $X=\alpha Z$. La distribution conjointe de X, Y et R à l'aide du modèle de mélange de schémas d'observation suivant, qui prend une forme comparable à celle présentée dans Little (1994) :

$$\begin{aligned}
&\left(X, Y \mid R=r\right) \sim G_2\left((\mu_x^{(r)}, \mu_y^{(r)}), \Sigma^{(r)}\right) \\
&R \sim \text{Bernoulli}(\pi) \\
&\Sigma^{(r)}=\begin{pmatrix} \sigma_{xx}^{(r)} & \rho^{(r)}\sqrt{\sigma_{xx}^{(r)}\sigma_{yy}^{(r)}} \\ \rho^{(r)}\sqrt{\sigma_{xx}^{(r)}\sigma_{yy}^{(r)}} & \sigma_{yy}^{(r)} \end{pmatrix}
\end{aligned}$$

où N_2 désigne la distribution bivariée normale. Nous nous intéressons au biais apparaissant dans la moyenne marginale de Y, que nous pouvons inscrire de cette façon : $\mu_y=\pi\mu_y^{(1)}+\left(1-\pi\right)\mu_y^{(0)}$. Andridge et Little (2011) présument que

$$\Pr\left(R=1 \mid Y, X\right) = f\left(X^*+\lambda Y\right),$$

pour une certaine fonction non spécifiée f et une constante connue λ. Dans le cas présent, $X^*=X\sqrt{\sigma_{yy}^{(1)}/\sigma_{xx}^{(1)}}$ est le substitut X proportionné de manière a obtenir la même variance que pour Y, ce qui facilite l'interprétation de λ grâce au positionnement de X et de Y à la même échelle. Ce mécanisme fait appel à des données manquantes au hasard lorsque $\lambda=0$, et dévie progressivement de celles-ci à mesure que λ augmente. Dans cette hypothèse, les paramètres sont exactement identifiés et l'estimation selon le maximum de vraisemblance de la moyenne de Y, en faisant la moyenne des schémas d'observation, correspond à

$$\hat{\mu}_y = \bar{y}_1+\left(\frac{n-r}{n}\right)\left(\frac{\lambda+\hat{\rho}}{\lambda\hat{\rho}+1}\right)\left(\bar{x}_0-\bar{x}_1\right),$$

où $\bar{x}_1, \bar{y}_1$ sont les moyennes de répondants de X et Y et $\bar{x}_0$, la moyenne de X pour les non-répondants. L'indice de biais correspond alors à l'ajustement de la moyenne de l'échantillon $\bar{y}_1$ considérée dans cette estimation, à savoir

$$\left(\frac{n-r}{n}\right)\left(\frac{\lambda+\hat{\rho}}{\lambda\hat{\rho}+1}\right)\left(\bar{x}_0-\bar{x}_1\right), \tag{4.14}$$

où $\hat{\rho}$ est la corrélation de l'échantillon de répondants. Cet ajustement intègre de manière simple et intuitive trois facteurs clés qui influent sur le possible biais – le taux de non-réponse $\left(n-r\right)/n$, la corrélation $\hat{\rho}$ entre X et Y, et la déviation de $\bar{x}_0$ par rapport à $\bar{x}_1$. L'indice croît notamment en fonction du taux de non-réponse et de l'écart de moyenne de X pour les répondants et les non-répondants.

Les données ne fournissent aucun renseignement à propos du paramètre λ – ce qui est habituellement le cas dans l'emploi de méthodes qui modélisent des déviations du mécanisme de données manquantes au hasard. À la suite de Little (1994), Andridge et Little (2011) proposent une analyse de la sensibilité dans laquelle on génère des estimations pour une fourchette de valeurs de λ entre 0 et l'infini, particulièrement 0, 1 et l'infini; le choix de 0 correspond à des données manquantes au hasard, celui de 1 présuppose que le biais de Y est le même que le biais de la variable substitut X^*, alors que le choix de l'infini correspond à la déviation la plus marquée par rapport aux données manquantes au hasard; dans ce cas, les estimations présentent la variance la plus élevée. Puisque λ varie de 0 à l'infini, le facteur mitoyen $(\lambda+\hat{\rho})/(\lambda\hat{\rho}+1)$ varie de $\hat{\rho}$ (quand $\hat{\mu}_y$ correspond à l'estimateur par la régression classique de la moyenne) à $1/\hat{\rho}$ (quand $\hat{\mu}_y$ correspond à l'estimateur par la régression inverse proposée par Brown [1990]). La sensibilité de l'estimation pour le choix de λ est faible lorsque $\hat{\rho}$ se situe près de 1, ce qui indique la présence d'une variable substitut robuste, et forte lorsque $\hat{\rho}$ se situe près de 0, auquel cas, la variable substitut est faible. Il s'avère donc essentiel de disposer de variables auxiliaires constituant de bons prédicteurs des résultats de l'enquête.

Des versions bayésiennes du modèle de mélange de schémas d'observation sont déjà au point, ce qui permet la propagation de l'erreur dans les paramètres du modèle. Il est en outre possible d'appliquer le modèle de manière à créer des imputations multiples de valeurs de non-répondants, ce qui permet d'intégrer à l'indice les éléments d'un plan d'échantillonnage complexe.

Plus récemment, le modèle a été utilisé pour développer des indices d'écart par rapport à l'échantillonnage aléatoire, en l'appliquant à la sélection du modèle plutôt qu'à la non-réponse (Little, West, Boonstra et Hu, 2020; Boonstra, Little, West, Andridge et Alvaredo-Leiton, 2021). Le perfectionnement de ces travaux vise (a) à remplacer le paramètre λ par $\phi=\lambda/(1+\lambda)$, qui constitue une meilleure paramétrisation parce que ϕ va de 0 (données manquantes au hasard) à 1, et (b) (avec l'ajout d'autres hypothèses) à permettre aux données manquantes de dépendre également de variables auxiliaires orthogonales à X. L'expansion de la méthode a aussi permis de gérer des résultats binaires (Andridge, West, Little, Boonstra et Alvarado-Leiton, 2019) et des indices de biais potentiels dans les coefficients de régression (West, Little, Andridge, Boonstra, Ware, Pandit et Alvarado-Leiton, 2021).

5. Conclusion : 10 raisons d'adopter Bayes pour les inférences d'enquête

Mes exemples visent à donner une idée de l'ampleur des possibilités de la modélisation bayésienne pour les données d'enquête, sans pour autant se révéler un tant soit peu exhaustifs. Bayes est aussi utile dans des domaines que je n'ai pas abordés, notamment l'échantillonnage à plusieurs degrés, les séries chronologiques, modèles d'analyse de classes latentes et de facteurs, les erreurs de mesures, la

combinaison de données issues de différentes sources, la création de données synthétiques pour prévenir la divulgation, et ainsi de suite. Je termine par un résumé des raisons qui m'amènent à favoriser l'approche d'inférence bayésienne pour les enquêtes :

1. L'approche fondée sur le plan de sondage est asymptotique et trop limitée pour gérer les divers problèmes relatifs aux inférences tirées d'enquêtes, qu'elles soient de nature probabiliste ou non probabiliste.
2. L'approche bayésienne est à la fois unifiée et suffisamment souple pour gérer les divers problèmes rencontrés dans les enquêtes, en plus de comprendre une modélisation de la superpopulation sous forme d'inférence tirée d'un vaste échantillon.
3. Des modèles bayésiens choisis avec soin peuvent fournir des intervalles crédibles qui possèdent de bonnes propriétés fondées sur le plan de sondage dans un échantillonnage répété. Il est notamment possible de modéliser la pondération, la stratification et la stratification *a posteriori* à l'aide de covariables et d'intégrer la mise en grappes à partir de modèles bayésiens hiérarchiques. Des modèles souples qui incorporent des caractéristiques du plan rendent inutiles des démarches hybrides comme l'estimation assistée par un modèle (par exemple Särndal, Swensson et Wretman, 1992).
4. Les premières critiques de la méthode de modélisation portent sur des modèles n'intégrant aucune caractéristique du plan et, par conséquent, qui sont sensibles à l'erreur de spécification du modèle. Il est possible, voire souhaitable d'éviter le recours à ces modèles.
5. Les calculs bayésiens d'intégration de paramètres de nuisance aboutissent à des inférences ayant de bonnes propriétés fréquentistes à partir de petits comme de grands échantillons.
6. L'approche fondée sur le plan de sondage pour traiter certains problèmes et fondée sur un modèle dans d'autres situations donne lieu à des incohérences logiques (voir, par exemple, Little, 2012, section 4.3); l'approche bayésienne fournit des inférences unifiées et logiquement cohérentes.
7. Dans l'approche bayésienne, la spécification des distributions *a priori* constitue un point fort, non une faiblesse, parce qu'elle donne une plus grande souplesse à la modélisation. Pour certains problèmes, des *a priori* « objectives » faibles aboutissent à des résultats équivalant à ceux obtenus de solutions fréquentistes conventionnelles. Pour d'autres problèmes, des *a priori* « subjectives » plus robustes donnent des réponses utiles dans le cas de modèles non identifiés, comme la non-réponse des données manquantes non au hasard.
8. Les récentes avancées méthodologiques et l'augmentation de la puissance informatique ont grandement réduit les difficultés de calcul dans l'approche bayésienne.
9. La modélisation inscrit la recherche par enquête dans le processus principal de la modélisation statistique pour d'autres types de données. Le paradigme bayésien gère aisément les caractéristiques particulières de la recherche par enquête – complexité du plan de sondage et accent sur des quantités de population finie.

10. L'approche bayésienne ne nie pas l'utilité de l'échantillonnage probabiliste pour le plan, elle qui se révèle extrêmement précieuse pour l'obtention des inférences robustes limitant le recours à des hypothèses contestables concernant la représentativité de l'échantillon.

Remerciements

Je tiens à remercier le Comité du prix Waksberg de m'avoir donné la chance de présenter ces travaux, de même que Yajuan Si et un examinateur pour leurs commentaires constructifs.

Bibliographie

Andridge, R.H., et Little, R.J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27, 2, 153-180.

Andridge, R.R., West, B.T., Little, R.J.A., Boonstra, P.S. et Alvarado-Leiton, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, 68, 5, 1465-1483.

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part I. Dans *Foundations of Statistical Inference*, (Éds., V.P. Godambe et D.A. Sprott), Toronto: Holt, Rinehart and Winston, 203-242.

Berger, J.O., et Wolpert, R.L. (1988). The likelihood principle. *Institute of Mathematical Statistics Lecture Notes-Monograph Series*, 6, 1-199.

Binder, D.A. (1982). Non-parametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical Society, Series B,* 44,3, 388-393.

Birnbaum, A. (1962). On the foundations of statistical inference (avec discussion). *Journal of the American Statistical Association*, 57, 269-326.

Boonstra, P.S., Little, R.J., West, B.T., Andridge, R.R. et Alvaredo-Leiton, F. (2021). A simulation study of diagnostics for bias in non-probability samples. *Journal of Official Statistics*, 37, 3, 751-769.

Brewer, K. (2013). Trois controverses dans l'histoire de l'échantillonnage. *Techniques d'enquête*, 39, 2, 275-289. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2013002/article/11883-fra.pdf.

Brown, C.H. (1990). Protecting against nonrandomly missing data in longitudinal studies. *Biometrics*, 46, 143-155.

Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Zhan, L. et Zhang, K. (2019). Models as approximations 1: Consequences illustrated with linear regression. *Statistical Science*, 34, 4, 580-583.

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

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.

Elliott, M.R. (2007). Réduction bayésienne des poids pour les modèles de régression linéaire généralisée. *Techniques d'enquête*, 33, 1, 27-40. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2007001/article/9849-fra.pdf.

Elliott, M.R., et Little, R.J. (2000). Model-based approaches to weight trimming. *Journal of Official Statistics*, 16, 191-210.

Elliott, M.R., et Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32, 2, 249-264.

Ericson, W.A. (1969). Subjective Bayesian models in sampling finite populations (avec discussion). *Journal of the Royal Statistical Society, Series B*, 31, 195-233.

Fienberg, S.E. (2011). Bayesian models and methods in public policy and government settings. *Statistical Science*, 26, 2, 212-226.

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

Franco, C., Little, R.J., Louis, T.A. et Slud, E.V. (2019). Comparative study of confidence intervals for proportions in complex sample surveys. *Journal of Survey Statistics and Methodology*, 7, 3, 334-364.

Gelman, A. (2007). Struggles with survey weighting and regression modeling (avec discussion). *Statistical Science*, 22, 2, 153-188.

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

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

Ghosh, M., Reid, N. et Fraser, D.A.S. (2010). Ancillary statistics: A review. *Statistica Sinica*, 20, 1309-1332.

Hájek, J. (1971). Comment on a paper by D. Basu. Dans *Foundations of Statistical Inference*, (Éds., V.P. Godambe et D.A. Sprott), Toronto: Holt, Rinehart and Winston, 236.

Hansen, M.H., Madow, W.G. et Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys (avec discussion). *Journal of the American Statistical Association*, 78, 776-793.

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

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.

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

Kish, L. (1995). The hundred years' wars of survey sampling. *Statistics in Transition*, 2, 813-830. Reproduit comme chapitre 1 de *Leslie Kish: Selected Papers*, (2003, Éds., G. Kalton et S. Heeringa), New York: John Wiley & Sons, Inc.

Kish, L., et Frankel, M.R. (1974). Inferences from complex samples (avec discussion). *Journal of the Royal Statistical Society, Series B*, 36, 1-37.

Lazzeroni, L.C., et Little, R.J. (1998). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics*, 14, 61-78.

Leon-Novelo, L.G., et Savitsky, T.D. (2019). Fully Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 13, 1608-1645.

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

Little, R.J. (1994). A class of pattern-mixture models for normal missing data. *Biometrika*, 81, 3, 471-483.

Little, R.J. (2003a). The Bayesian approach to sample survey inference. Dans *Analysis of Survey Data*, (Éds., R.L. Chambers et C.J. Skinner), New York: John Wiley & Sons, Inc., 49-57.

Little, R.J. (2003b). Bayesian methods for unit and item nonresponse. Dans *Analysis of Survey Data*, (Éds., R.L. Chambers et C.J. Skinner), New York: John Wiley & Sons, Inc., 289-306.

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, 546-556.

Little, R.J. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *The American Statistician*, 60, 3, 213-223.

Little, R.J. (2012). Calibrated Bayes: An alternative inferential paradigm for official statistics (avec discussion). *Journal of Official Statistics*, 28, 3, 309-372.

Little, R.J. (2014). Survey sampling: Past controversies, current orthodoxies, and future paradigms. Dans *Past, Present and Future of Statistical Science*, COPSS 50th Anniversary Volume, (Éds., X. Lin, D.L. Banks, C. Genest, G. Molenberghs, D.W. Scott et J.-L. Wang), CRC Press.

Little, R.J. (2015). Calibrated Bayes, an inferential paradigm for official statistics in the era of big data. *Statistical Journal of the International Association of Survey Statisticians*, 31, 4, 555-563.

Little, R.J. (2019). Comment on "Models as approximations 1: Consequences illustrated with linear regression" by A. Buja et al. *Statistical Science*, 34, 4, 580-583.

Little, R.J. (2022). A note about the definition of propensity weights. À paraître dans le *Journal of Survey Statistics and Methodology*.

Little, R.J., et Rubin, D.B. (2019). *Statistical Analysis with Missing Data*, 3rd edition. New York: John Wiley & Sons, Inc.

Little, R.J., et Vartivarian, S. (2005). La pondération pour la non-réponse augmente-t-elle la variance des moyennes de sondage ? *Techniques d'enquête*, 31, 2, 175-183. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2005002/article/9046-fra.pdf.

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

Little, R.J., West, B.T., Boonstra, P.S. et Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of Survey Statistics and Methodology*, 8, 5, 932-964.

Royall, R.M., et Cumberland, W.G. (1981). The finite population linear regression estimator and estimator of its variance-an empirical study. *Journal of the American Statistical Association*, 76, 924-930.

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

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 53, 581-592.

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6, 1, 34-58.

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12, 1151-1172.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B. (2019). Le calage conditionnel et le sage statisticien. *Techniques d'enquête*, 45, 2, 199-210. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2019002/article/00010-fra.pdf.

Ruppert, D., Wand, M.P et Carroll, R.J. (2003). *Semiparametric Regression*, Cambridge University Press.

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/fr/pub/12-001-x/2010002/article/11376-fra.pdf.

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

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.

Sedransk, J. (2008). Assessing the value of Bayesian methods for inference about finite population quantities. *Journal of Official Statistics*, 24, 495-506.

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.

Smith, T.M.F. (1976). The foundations of survey sampling: A review (avec discussion). *Journal of the Royal Statistical Society, Series A*, 139, 183-204.

Smith, T.M.F. (1994). Sample surveys 1975-1990: An age of reconciliation? (avec discussion). *Revue Internationale de Statistique*, 62, 5-34.

Szpiro, A.A., Rice, K.M. et Lumley, T. (2010). Model-robust regression and a Bayesian "sandwich" estimator. *Annals of Applied Statistics*, 4, 4, 2099-2113.

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

West, B.T., Little, R.J., Andridge, R.R., Boonstra, P.S., Ware, E.B., Pandit, A. et Alvarado-Leiton, F. (2021). Measures of selection bias in regression coefficients estimated from non-probability samples. *The Annals of Applied Statistics*, 15(3), 1556-1581.

Yang, Y., et Little, R.J. (2015). A comparison of doubly robust estimators of the mean with missing data. *Journal of Statistical Computation and Simulation*, 85, 16, 3383-3403.

Zangeneh, S.Z., et Little, R.J. (2015). Bayesian inference for the finite population total in heteroscedastic probability proportional to size samples. *Journal of Survey Statistics and Methodology*, 3, 162-192.

Zangeneh, S.Z., et Little, R.J. (2022). Likelihood based estimation of the finite population mean with post-stratification information under nonignorable nonresponse. À paraître dans la *Revue Internationale de Statistique*.

Zhang, G., et Little, R.J. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65, 3, 911-918.

Zheng, H., et Little, R.J. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 2, 99-117.

Zheng, H., et Little, R.J. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from apenalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.

Techniques d'enquête, décembre 2022
Vol. 48, n° 2, pp. 307-338
Statistique Canada, n° 12-001-X au catalogue

Inférence statistique avec des échantillons d'enquête non probabiliste

Changbao Wu[1]

Résumé

Nous offrons un examen critique et quelques discussions approfondies sur des questions théoriques et pratiques à l'aide d'une analyse des échantillons non probabilistes. Nous tentons de présenter des cadres inférentiels rigoureux et des procédures statistiques valides dans le cadre d'hypothèses couramment utilisées et d'aborder les questions relatives à la justification et à la vérification d'hypothèses sur des applications pratiques. Certains progrès méthodologiques actuels sont présentés et nous mentionnons des problèmes qui nécessitent un examen plus approfondi. Alors que l'article porte sur des échantillons non probabilistes, le rôle essentiel des échantillons d'enquête probabilistes comportant des renseignements riches et pertinents sur des variables auxiliaires est mis en évidence.

Mots-clés : Information auxiliaire; estimateur de la variance bootstrap; méthode de calage; estimateur doublement robuste; équations d'estimation; pondération de la probabilité inverse; prédiction fondée sur un modèle; poststratification; pseudo-vraisemblance; score de propension; enquête par quotas; analyse de sensibilité; estimation de la variance.

1. Introduction

Le domaine de l'échantillonnage se distingue des autres domaines de la statistique en raison d'un certain nombre de caractéristiques uniques. La population cible est constituée d'un nombre fini d'unités et les paramètres de population peuvent être déterminés sans erreur, du moins conceptuellement, en menant un recensement. Les contraintes opérationnelles et l'avantage sur le plan administratif pour la collecte des données rendent nécessaire d'envisager la stratification, la catégorisation des résultats et la sélection à probabilités inégales. Depuis le document précurseur de Neyman (1934), les méthodes d'échantillonnage probabiliste sont devenues l'un des outils de collecte de données primaires pour les statistiques officielles et les chercheurs dans les domaines des sciences de la santé, des études sociales et économiques, des affaires et de la commercialisation, des stocks de ressources agricoles et naturelles, et autres domaines. Des échantillons d'enquête probabilistes ont également été utilisés pour effectuer des études analytiques comportant des modèles et des paramètres de modèles; voir, par exemple, Binder (1983), Godambe et Thompson (1986), Thompson (1997), Rao et Molina (2015), entre autres. Des échantillons d'enquête probabilistes et une inférence fondée sur le plan ont constitué une mesure couronnée de succès dans le cadre des sciences statistiques au cours des 80 dernières années.

Toutefois, au cours des dernières années, « on a observé un vent de changement et on considère de plus en plus d'autres sources de données » (Beaumont, 2020). Le succès des échantillons d'enquête probabilistes a mené à des plans d'études ambitieux, à de longs questionnaires compliqués et à un fardeau accru sur les répondants. Les taux de réponse ont diminué et le coût de la collecte des données a grimpé en flèche au fil des ans. Compte tenu des progrès de nouvelles technologies et de l'explosion de l'information

1. Changboa Wu, Département des statistiques et de l'actuariat, Université de Waterloo, Waterloo (Ontario) N2L 3G1. Courriel : cbwu@uwaterloo.ca.

sur Internet, il existe également un fort désir d'accéder à des statistiques en temps réel. Statistique Canada a lancé les initiatives dites de modernisation appelée « Aller au-delà de l'approche fondée sur les données d'enquête pour adopter de nouvelles méthodes et intégrer des données provenant de diverses sources existantes ».

Les échantillons non probabilistes représentent l'une de ces sources des données qui ont gagné en popularité au cours des dernières années. Les échantillons non probabilistes ne sont pas nouveaux dans le domaine de l'échantillonnage. Ils ont été utilisés depuis les débuts de la réalisation d'enquêtes. Par exemple, les enquêtes par quotas ont donné des échantillons non probabilistes et la méthode est largement utilisée et peut réussir dans certaines conditions; voir la section 5 pour consulter d'autres analyses. Les échantillons non probabilistes n'ont jamais pris un véritable élan dans le passé dans la pratique des enquêtes en raison de l'absence d'un cadre de travail théorique mature pour l'analyse des données. Néanmoins, il existe des données accessibles qui sont moins chères et plus rapides à obtenir et qui sont devenues courantes pour la recherche en ligne. Les entreprises commerciales d'enquête créent et tiennent à jour une longue liste de personnes, appelées « panels volontaires », qui ont accepté que l'on communique avec elles pour participer à des enquêtes comme volontaires ou grâce à des incitatifs. Les mécanismes précis d'inclusion des personnes dans le panel sont habituellement inconnus, se traduisant par des échantillons d'enquête non probabilistes fondés sur des panels.

Le principal problème des échantillons non probabilistes est qu'il s'agit d'échantillons biaisés et que ceux-ci ne sont pas représentatifs de la population cible. On peut soutenir que, outre les échantillons à unités indépendantes et identiquement distribuées, la plupart des échantillons sont biaisés, et même les échantillons d'enquête probabilistes le sont. Les probabilités d'inclusion connues du plan d'enquête sont la raison pour laquelle nous ne nous inquiétons pas de la nature biaisée des échantillons d'enquête probabilistes, car elles mènent à des méthodes d'estimation valides par l'entremise de procédures appropriées de pondération. Le principal enjeu véritable des échantillons non probabilistes est donc l'aspect inconnu de l'inclusion dans un échantillon ou des mécanismes de participation. Il ressortira clairement des analyses présentées à la section 4 que la nature biaisée des échantillons non probabilistes ne peut être corrigée au moyen de l'échantillon lui-même. Cela nécessite des renseignements auxiliaires sur la population cible.

Le présent article offre un examen critique et quelques discussions approfondies sur des questions théoriques et pratiques à l'aide d'une analyse d'échantillons non probabilistes. La section 2 brosse un tableau du contexte général, des hypothèses communément utilisées ainsi que des cadres inférentiels employés dans les procédures statistiques traitées dans l'article. La section 3 présente une approche de prédiction des échantillons non probabilistes fondée sur un modèle. La section 4 aborde l'estimation de scores de propension et la construction d'estimateurs fondés sur les scores de propension. La section 5 illustre les liens entre les estimateurs pondérés de probabilité inverse et les enquêtes par quotas avec des extensions à la poststratification. La section 6 met l'accent sur les techniques ainsi que sur les enjeux liés à

l'estimation de la variance. La section 7 aborde l'importante question touchant la manière de contrôler et de vérifier les hypothèses requises dans la pratique. Quelques conclusions sont présentées à la section 8.

2. Hypothèses et cadres inférentiels

Supposons que la population cible $U = \{1, 2, \ldots, N\}$ est constituée de N unités étiquetées. Associées à l'unité i, on retrouve les valeurs $\mathbf{x}_i$ et y_i pour les variables auxiliaires $\mathbf{x}$ et la variable étudiée y. Les discussions portent sur un seul y, mais les ensembles de données sont plus susceptibles de contenir de multiples variables étudiées. Soit $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$, la moyenne de la population, laquelle est le paramètre d'intérêt. Soit $\{(y_i, \mathbf{x}_i), i \in S_A\}$, l'ensemble de données pour l'échantillon non probabiliste S_A avec n_A unités participantes. Pour la plupart des scénarios pratiques, la moyenne simple de l'échantillon $\bar{y}_A = n_A^{-1}\sum_{i \in S_A} y_i$ est un estimateur biaisé de μ_y et, par conséquent, il n'est pas valide.

2.1 Hypothèses

Soit $R_i = I(i \in S_A)$, la variable indicatrice pour l'unité i comprise dans l'échantillon non probabiliste S_A. Il convient de souligner que la variable R_i est définie pour tous les i dans la population cible. Supposons que :

$$\pi_i^A = P\left(i \in S_A \,|\, \mathbf{x}_i, y_i\right) = P\left(R_i = 1 \,|\, \mathbf{x}_i, y_i\right), \quad i = 1, 2, \ldots, N.$$

Nous appelons la π_i^A les scores de propension, un terme emprunté aux ouvrages publiés sur les données manquantes (Rosenbaum et Rubin, 1983). Certains auteurs utilisent le terme « probabilités de participation ». Voir, par exemple, Beaumont (2020) et Rao (2021), entre autres. Les scores de propension π_i^A caractérisent l'inclusion dans un échantillon et les mécanismes de participation. Ils sont inconnus et nécessitent des hypothèses appropriées du modèle pour l'élaboration de méthodes d'estimation valides. Chen, Li, et Wu (2020) ont utilisé les trois hypothèses de base suivantes, lesquelles ont été adaptées à partir de la littérature sur les données manquantes.

A1 L'inclusion dans l'échantillon et l'indicateur de participation R_i, ainsi que la variable étudiée y_i sont indépendants compte tenu de l'ensemble de covariables $\mathbf{x}_i$, c'est-à-dire $\left(R_i \perp y_i\right) \,|\, \mathbf{x}_i$.

A2 Toutes les unités dans la population cible ont des scores de propension non nuls, c'est-à-dire $\pi_i^A > 0,\ i = 1, 2, \ldots, N$.

A3 Les variables indicatrices $R_1, R_2, \ldots, R_N$ sont indépendantes compte tenu de l'ensemble de variables auxiliaires, $\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right)$.

L'hypothèse A1 est semblable à l'hypothèse de répartition au hasard des données manquantes pour l'analyse des données manquantes. Selon A1, cela donne $\pi_i^A = P\left(R_i = 1 \,|\, \mathbf{x}_i, y_i\right) = P\left(R_i = 1 \,|\, \mathbf{x}_i\right) = \pi\left(\mathbf{x}_i\right)$. L'hypothèse A2 peut être problématique dans la pratique. Voir la section 7 pour obtenir d'autres analyses. L'hypothèse A3 se vérifie habituellement quand les participants sont approchés un à la fois, mais peut être

discutable lors de l'utilisation de sélections en grappes. Il est démontré à la section 4 que l'estimation de $\pi_i^A = \pi(\mathbf{x}_i)$ selon l'hypothèse A1 nécessite une information auxiliaire sur la population cible. Le scénario idéal est que l'information auxiliaire complète $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ soit disponible. Le scénario le plus pratique est que l'information auxiliaire peut être obtenue d'une enquête probabiliste existante.

A4 Il existe un échantillon d'enquête probabiliste S_B de taille n_B avec des renseignements sur les variables auxiliaires $\mathbf{x}$ (mais pas sur y) disponibles dans l'ensemble de données $\left\{(\mathbf{x}_i, d_i^B), i \in S_B\right\}$, où d_i^B sont les poids de sondage pour l'échantillon probabiliste S_B.

Le S_B est appelé « l'échantillon d'enquête probabiliste de référence ». La partie la plus essentielle de l'hypothèse A4 est que l'ensemble de variables auxiliaires $\mathbf{x}$ est observé dans l'échantillon non probabiliste S_A et dans l'objet probabiliste S_B. Un échantillon d'enquête probabiliste de référence est souvent disponible dans la pratique, mais l'ensemble de variables auxiliaires commun peut ne pas contenir toutes les composantes pour satisfaire l'hypothèse A1.

2.2 Cadre inférentiel

Il existe trois sources possibles de variation en vertu du contexte général de deux échantillons S_A et S_B : i) Le modèle q pour le score de propension sur l'inclusion dans l'échantillon et la participation dans l'échantillon d'enquête non probabiliste S_A; ii) le modèle ξ pour la régression des résultats $(y \mid \mathbf{x})$ ou l'imputation; et iii) le plan d'échantillonnage probabiliste p pour l'échantillon d'enquête probabiliste de référence S_B. Pour les trois approches de l'inférence qui seront abordées dans les sections 3 et 4, l'échantillon probabiliste de référence S_B est toujours mis à contribution. Chacune des trois approches nécessite un cadre de randomisation conjoint mettant en jeu p et un de (q, ξ).

a) Approche de prédiction fondée sur un modèle : le cadre ξp sous la randomisation conjointe du modèle de régression des résultats ξ et le plan d'échantillonnage avec probabilités p.

b) Pondération de la probabilité inverse à l'aide de scores de propension estimés : le cadre qp sous la randomisation conjointe du modèle de score de propension q et le plan d'échantillonnage avec probabilités p.

c) Inférence doublement robuste : le cadre qp ou le cadre ξp, sans spécification de l'un ou de l'autre.

Le cadre inférentiel est la base du développement théorique. La convergence des estimateurs ponctuels doit être établie en vertu de la randomisation conjointe appropriée. Les variances théoriques comportent habituellement deux composantes, une de chaque source de variation, et le calcul correct des deux composantes constitue la clé de la construction d'estimateurs de variance convergents selon le cadre inférentiel désigné.

3. Approche de prédiction fondée sur un modèle

Les méthodes de prédiction fondées sur un modèle pour les paramètres de population finie nécessitent deux ingrédients essentiels : la quantité de renseignements auxiliaires qui sont disponibles à l'étape de l'estimation et la fiabilité du modèle théorique d'inférence. En l'absence de toute information auxiliaire, le modèle moyen commun $E_\xi(y_i) = \mu_0,\ V_\xi(y_i) = \sigma^2,\ i = 1, \ldots, N$ peut être considéré comme raisonnable, mais l'estimateur de prédiction fondé sur un modèle $\hat{\mu}_y = \bar{y}_A = n_A^{-1}\sum_{i \in S_A} y_i,$ bien que non biaisé selon le modèle depuis $E_\xi\left(\bar{y}_A - \mu_y\right) = 0,$ n'est généralement pas un estimateur acceptable de μ_y. La variance σ^2 pour le modèle moyen commun est habituellement de grande taille et elle rend l'estimateur $\hat{\mu}_y = \bar{y}_A$ avec une variance de prédiction qui est trop importante pour être d'une utilité pratique.

3.1 Modèles de régression des résultats semi-paramétriques

Sans perte de généralité, nous supposons que $\mathbf{x}$ contient 1 comme sa première composante correspondant à l'ordonnée à l'origine d'un modèle de régression. Dans les conditions décrites dans la section 2, nous considérons le modèle semi-paramétrique suivant pour la population finie, désigné comme ξ :

$$E_\xi\left(y_i \mid \mathbf{x}_i\right) = m\left(\mathbf{x}_i, \boldsymbol{\beta}\right), \text{ et } V_\xi\left(y_i \mid \mathbf{x}_i\right) = v\left(\mathbf{x}_i\right)\sigma^2, \quad i = 1, 2, \ldots, N, \tag{3.1}$$

ou la fonction moyenne $m(\cdot,\cdot)$ et la fonction variance $v(\cdot)$ ont des formes connues et les y_i sont également présumés être conditionnellement indépendants compte tenu des $\mathbf{x}_i$. Soit $\boldsymbol{\beta}_0$ et $\sigma_0^2,$ les valeurs vraies des paramètres $\boldsymbol{\beta}$ et σ^2 du modèle selon le modèle théorique. La première conséquence importante de l'hypothèse A1 est que $E_\xi\left(y_i \mid \mathbf{x}_i, R_i = 1\right) = E_\xi\left(y_i \mid \mathbf{x}_i\right)$ et $V_\xi\left(y_i \mid \mathbf{x}_i, R_i = 1\right) = V_\xi\left(y_i \mid \mathbf{x}_i\right)$. Le modèle (3.1), qui est considéré pour la population finie, se vérifie également pour les unités dans l'échantillon d'enquête non probabiliste S_A. L'estimateur par le maximum de quasi-vraisemblance $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}_0$ est obtenu à l'aide de l'ensemble de données $\left\{(y_i, \mathbf{x}_i), i \in S_A\right\}$ pour l'échantillon d'enquête non probabiliste comme la solution des équations de quasi-score (McCullagh and Nelder, 1989) données par :

$$\mathrm{S}\left(\boldsymbol{\beta}\right) = \sum_{i \in S_A} \frac{\partial m\left(\mathbf{x}_i, \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} \left\{v\left(\mathbf{x}_i\right)\right\}^{-1} \left\{y_i - m\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\right\} = \mathbf{0}. \tag{3.2}$$

Le modèle semi-paramétrique (3.1) peut être élargi pour remplacer $v(\mathbf{x}_i)$ par une fonction de variance générale $v(\mu_i)$ où $\mu_i = m(\mathbf{x}_i, \boldsymbol{\beta})$. La théorie de l'estimation par la méthode du maximum de vraisemblance englobe les modèles de régression linéaires et non linéaires comportant des estimateurs par les moindres carrés pondérés, le modèle de régression logistique et d'autres modèles linéaires généralisés. Soit $m_i = m(\mathbf{x}_i, \boldsymbol{\beta}_0)$ et $\hat{m}_i = m(\mathbf{x}_i, \hat{\boldsymbol{\beta}}),\ i = 1, 2, \ldots, N.$

3.2 Deux formes générales d'estimateurs de prédiction

Il existe deux estimateurs de prédiction fondée sur un modèle pour μ_y en présence de renseignements auxiliaires complets $\left\{\mathbf{x}_1, \ldots, \mathbf{x}_N\right\}$; voir le chapitre 5 de Wu et Thompson (2020). Notons que $E_\xi(\mu_y) = N^{-1}\sum_{i=1}^N m_i.$ Les deux estimateurs de prédiction sont établis comme suit :

$$\hat{\mu}_{y1} = \frac{1}{N}\sum_{i=1}^{N}\hat{m}_i \quad \text{et} \quad \hat{\mu}_{y2} = \frac{1}{N}\left\{\sum_{i\in S_A} y_i - \sum_{i\in S_A}\hat{m}_i + \sum_{i=1}^{N}\hat{m}_i\right\}. \tag{3.3}$$

L'estimateur $\hat{\mu}_{y2}$ est établi en fonction de $\mu_y = N^{-1}\left\{\sum_{i\in S_A} y_i + \sum_{i\notin S_A} y_i\right\}$ et s'appuie sur $\sum_{i\notin S_A}\hat{m}_i = \sum_{i=1}^{N}\hat{m}_i - \sum_{i\in S_A}\hat{m}_i$ pour prédire le terme non observé $\sum_{i\notin S_A} y_i$. Dans un modèle de régression linéaire où $m(\mathbf{x},\boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}$, les deux estimateurs donnés dans (3.3) se réduisent à :

$$\hat{\mu}_{y1} = \boldsymbol{\mu}_{\mathbf{x}}'\hat{\boldsymbol{\beta}} \quad \text{et} \quad \hat{\mu}_{y2} = \frac{n_A}{N}\left(\bar{y}_A - \bar{\mathbf{x}}_A'\hat{\boldsymbol{\beta}}\right) + \boldsymbol{\mu}_{\mathbf{x}}'\hat{\boldsymbol{\beta}}, \tag{3.4}$$

où $\boldsymbol{\mu}_{\mathbf{x}} = N^{-1}\sum_{i=1}^{N}\mathbf{x}_i$ est le vecteur des moyennes de population des variables de $\mathbf{x}$ et $\bar{\mathbf{x}}_A = n_A^{-1}\sum_{i\in S_A}\mathbf{x}_i$ est le vecteur de la moyenne simple de l'échantillon de $\mathbf{x}$ de l'échantillon non probabiliste S_A. Si le modèle de régression linéaire contient une ordonnée à l'origine et $\hat{\boldsymbol{\beta}}$ est l'estimateur des moindres carrés ordinaires, nous avons $\hat{\mu}_{y2} = \hat{\mu}_{y1} = \boldsymbol{\mu}_{\mathbf{x}}'\hat{\boldsymbol{\beta}}$, puisque $\bar{y}_A - \bar{\mathbf{x}}_A'\hat{\boldsymbol{\beta}} = 0$ en raison de la somme nulle des résidus prédits. Les estimateurs de prédiction en (3.4) sous le modèle linéaire ne requièrent que la moyenne de population $\boldsymbol{\mu}_{\mathbf{x}}$ en plus de l'échantillon non probabiliste S_A. Dans les conditions décrites à la section 2 comportant des renseignements auxiliaires sur $\mathbf{x}$ fournies par un échantillon probabiliste de référence S_B, nous remplaçons simplement $\sum_{i=1}^{N}\hat{m}_i$ par $\sum_{i\in S_B} d_i^B\hat{m}_i$ pour les estimateurs en (3.3) et substituons $\boldsymbol{\mu}_{\mathbf{x}}$ par $\hat{\boldsymbol{\mu}}_{\mathbf{x}} = \hat{N}_B^{-1}\sum_{i\in S_B} d_i^B\mathbf{x}_i$ pour les estimateurs en (3.4), où $\hat{N}_B = \sum_{i\in S_B} d_i^B$. La taille de la population N apparaissant en (3.3) ou (3.4) doit également être remplacée par $\hat{N}_B$ même si elle est connue.

3.3 Imputation massive

Les estimateurs de prédiction fondée sur un modèle μ_y reposant sur un échantillon d'enquête non probabiliste sur $(y,\mathbf{x})$ et un échantillon d'enquête probabiliste de référence sur $\mathbf{x}$ ont été traditionnellement présentés comme *l'estimateur d'imputation massive*. La variable étudiée y n'est pas observée pour aucune des unités dans l'échantillon de l'enquête de référence S_B et donc peut être perçue comme manquante pour tous les $i\in S_B$. Soit y_i^*, une valeur imputée pour y_i, $i\in S_B$. L'estimateur d'imputation massive de μ_y est ensuite établi comme suit :

$$\hat{\mu}_{y\text{IM}} = \frac{1}{\hat{N}_B}\sum_{i\in S_B} d_i^B y_i^*, \tag{3.5}$$

où $\hat{N}_B$ est défini comme il l'est précédemment et l'indice « IM » signifie « imputation massive » (et non pas « imputation multiple »). Sous l'imputation par la régression déterministe où $y_i^* = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$, l'estimateur $\hat{\mu}_{y\text{IM}}$ se réduit à l'estimateur de prédiction fondée sur un modèle $\hat{\boldsymbol{\mu}}_{\mathbf{x}}'\hat{\boldsymbol{\beta}}$ comme il est mentionné à la section 3.2.

L'approche de l'imputation massive pour l'analyse d'échantillons d'enquête non probabilistes s'inscrit dans le même esprit que les méthodes de prédiction fondées sur un modèle, mais elle ouvre la porte à l'utilisation de modèles plus flexibles et de techniques d'imputation qui ont été mises au point dans la littérature existante sur les problèmes de données manquantes. Cette approche a été examinée la première fois par Rivers (2007) à l'aide de la méthode dite d'*appariement d'échantillons*. Pour chaque $i\in S_B$, le y_i

« manquant » est imputé comme $y_i^* = y_j$ pour quelques valeurs $j \in S_A$, où j est un donneur apparié de S_A sélectionné par la méthode du plus proche voisin, mesuré par la distance entre $\mathbf{x}_i$ et $\mathbf{x}_j$. Le modèle sous-jacent ξ pour la méthode d'imputation du plus proche voisin est non paramétrique, c'est-à-dire $E_\xi\left(y_i \mid \mathbf{x}_i\right) = m\left(\mathbf{x}_i\right)$ pour une fonction inconnue $m(\cdot)$. La valeur d'appariement y_j peut être perçue comme la valeur prédite du y_i manquant selon le modèle. Les propriétés théoriques des estimateurs fondées sur l'imputation du plus proche voisin ont été abordées par Chen et Shao (2000, 2001) pour les problèmes de données d'enquêtes manquantes.

Le modèle semi-paramétrique (3.1) peut être utilisé pour l'imputation massive par la régression déterministe. Dans l'hypothèse A1, un estimateur convergent $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ est d'abord obtenu de l'ensemble de données d'échantillons non probabilistes $\left\{(y_i, \mathbf{x}_i), i \in S_A\right\}$, et l'estimateur $\hat{\boldsymbol{\beta}}$ est ensuite utilisé pour calculer les valeurs imputées $y_i^* = m\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}}\right)$ pour $i \in S_B$. En d'autres mots, l'hypothèse A1 met en jeu ladite *transportabilité du modèle* par Kim, Park, Chen et Wu (2021) : le modèle qui est construit pour l'échantillon non probabiliste peut être utilisé pour une prédiction avec l'échantillon probabiliste de référence. L'estimateur d'imputation massive résultant $\hat{\mu}_{y\mathrm{IM}}$ est semblable à l'un des estimateurs de prédiction fondée sur un modèle présentés à la section 3.2. Les propriétés asymptotiques et l'estimation de la variance pour l'estimateur $\hat{\mu}_{y\mathrm{IM}}$ réalisée à l'aide du modèle semi-paramétrique (3.1) ont été décrites par Kim et coll. (2021).

Selon l'approche de l'imputation massive, le seul rôle joué par le y_i observé pour $i \in S_A$ est d'estimer les paramètres du modèle $\boldsymbol{\beta}$. L'estimateur $\hat{\mu}_{y\mathrm{IM}}$ est établi à l'aide du modèle prédit et des renseignements auxiliaires de l'échantillon probabiliste de référence S_B. Il semble que nous n'avons pas pleinement utilisé les renseignements sur le y_i observé étant donné que μ_y est le paramètre principal d'intérêt. Cela a mené à la question de recherche décrite au chapitre 17 de Wu et Thompson (2020) sur *l'appariement inversé d'échantillons*. L'estimateur proposé est construit comme suit : $\hat{\mu}_{y\mathrm{A}} = (\hat{N}^*)^{-1} \sum_{i \in S_A} d_i^* y_i$ en utilisant tous les y_i observés dans l'échantillon non probabiliste, où $\hat{N}^* = \sum_{i \in S_A} d_i^*$. Le d_i^* est un poids d'enquête apparié de S_B de sorte que $d_i^* = d_j^B$ avec $j \in S_B$ soit le plus proche voisin de $i \in S_A$, mesuré par $\left\| \mathbf{x}_i - \mathbf{x}_j \right\|$. Les propriétés théoriques de l'estimateur apparié inverse $\hat{\mu}_{y\mathrm{A}}$ reposant sur le plus proche voisin $j \in S_B$ pour coupler d_i^* avec d_j^B n'a pas fait l'objet d'une enquête formelle dans la littérature existante.

Wang, Graubard, Katki et Li (2020) ont proposé une approche de pondération par noyau pour inverser l'appariement des échantillons en utilisant $d_i^* \propto \sum_{j \in S_B} K_{ij} d_j$, où K_{ij} est une distance entre les noyaux $\hat{p}_i$ et $\hat{p}_j$; voir la méthode de pondération par la propension logistique ajustée à la fin de la section 4.1.1 sur le calcul de $\hat{p}_i$. Ils ont montré que l'estimateur $\hat{\mu}_{y\mathrm{A}}$ est convergent dans certaines conditions de régularité. Dans un récent document de travail sur arXiv par Liu et Valliant (2021), les auteurs ont traité des problèmes avec le biais et la variance de l'estimateur à appariement inverse dans différents cadres de randomisation mettant en jeu une source, deux sources ou les trois sources (p, q, ξ). Les auteurs ont également proposé une étape de calage sur les poids appariés, ce qui semble être une idée prometteuse. Des études supplémentaires sur ce sujet sont nécessaires.

L'approche par imputation massive pour l'analyse des échantillons d'enquête non probabilistes mène à une question de recherche intéressante faisant actuellement l'objet d'une étude par un étudiant au doctorat

de l'université de Waterloo : Est-il théoriquement possible et pratiquement utile de créer un ensemble de données obtenues par imputation massive $\left\{(y_i^*, \mathbf{x}_i, d_i^B), i \in S_B\right\}$ selon l'échantillon d'enquête probabiliste de référence qui peut être utilisé pour des inférences statistiques générales ? La réponse dépend clairement des types de problèmes inférentiels à effectuer sur les ensembles de données obtenues par imputation. Une exigence minimale est que la distribution conditionnelle de la variable étudiée y compte tenu des covariables $\mathbf{x}$ est préservée pour l'ensemble de données obtenues par imputation massive. La méthode d'imputation du plus proche voisin et la méthode d'imputation par la régression aléatoire peuvent être utiles à cette fin. L'imputation fractionnaire est une autre possibilité, plus particulièrement pour les variables étudiées binaires ou ordinales. L'imputation multiple est également potentiellement utile dans ce sens pour créer de multiples ensembles de données obtenues par imputation massive. Dans ce cas, l'indice « IM » pourrait devoir être changé pour « IM^2 », signifiant « imputation massive avec imputation multiple ».

4. Approche fondée sur les scores de propension

Les scores de propension $\pi_i^A = P\left(R_i = 1 \middle| \mathbf{x}_i, y_i\right)$ pour l'échantillon d'enquête non probabiliste S_A sont théoriquement définis pour toutes les unités dans la population cible. L'estimation des scores de propension pour les unités dans S_A, lequel joue le rôle le plus important pour les méthodes fondées sur les scores de propension, nécessite un modèle théorique sur les scores de propension et des renseignements auxiliaires au niveau de la population. Dans la présente section, nous abordons d'abord les procédures d'estimation pour les scores de propension en vertu des conditions et des hypothèses décrites à la section 2; nous fournirons ensuite un aperçu des méthodes d'estimation proposées dans la littérature récente sur la moyenne de la population μ_y comportant des estimations de scores de propension.

4.1 Estimation des scores de propension

Dans l'hypothèse A1, les scores de propension $\pi_i^A = P\left(R_i = 1 \middle| \mathbf{x}_i\right) = \pi\left(\mathbf{x}_i\right)$ sont une fonction des variables auxiliaires $\mathbf{x}_i$, mais la forme fonctionnelle peut être compliquée et est complètement inconnue. Trois formes paramétriques populaires $\pi_i^A = \pi\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$ pour traiter une réponse binaire peuvent être considérées : i) la fonction logit inverse $\pi_i^A = 1 - \left\{1 + \exp(\mathbf{x}_i'\boldsymbol{\alpha})\right\}^{-1}$; ii) la fonction probit inverse $\pi_i^A = \Phi(\mathbf{x}_i'\boldsymbol{\alpha})$, où $\Phi(\cdot)$ est la fonction de distribution cumulative de $N(0,1)$; iii) la fonction log-log complémentaire inverse $\pi_i^A = 1 - \exp\left\{-\exp(\mathbf{x}_i'\boldsymbol{\alpha})\right\}$. Des techniques non paramétriques, sans présumer une forme fonctionnelle explicite pour $\pi\left(\mathbf{x}\right)$, constituent des solutions de rechange intéressantes pour l'estimation des scores de propension.

4.1.1 La méthode par pseudo maximum de vraisemblance

Soit $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$, une forme paramétrique précisée avec des paramètres de modèle inconnus $\boldsymbol{\alpha}$. Dans la situation idéale où l'information auxiliaire complète $\left\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right\}$ est disponible avec l'hypothèse d'indépendance A3, la fonction du logarithme du rapport de vraisemblance complète sur $\boldsymbol{\alpha}$ peut être formulée comme suit (Chen et coll., 2020) :

$$\ell(\boldsymbol{\alpha}) = \log\left\{\prod_{i=1}^{N}\left(\pi_i^A\right)^{R_i}\left(1-\pi_i^A\right)^{1-R_i}\right\} = \sum_{i\in S_A}\log\left(\frac{\pi_i^A}{1-\pi_i^A}\right)+\sum_{i=1}^{N}\log\left(1-\pi_i^A\right). \tag{4.1}$$

L'estimateur du maximum de vraisemblance de $\boldsymbol{\alpha}$ est le maximiseur de $\ell(\boldsymbol{\alpha})$. Dans les conditions actuelles où l'information auxiliaire sur la population est fournie par un échantillon d'enquête probabiliste de référence S_B, nous remplaçons $\ell(\boldsymbol{\alpha})$ par une fonction de pseudo-logarithme du rapport de vraisemblance (Chen et coll., 2020) :

$$\ell^*(\boldsymbol{\alpha}) = \sum_{i\in S_A}\log\left(\frac{\pi_i^A}{1-\pi_i^A}\right) + \sum_{i\in S_B} d_i^B\log\left(1-\pi_i^A\right). \tag{4.2}$$

L'estimateur du maximum de pseudo-vraisemblance $\hat{\boldsymbol{\alpha}}$ est le maximiseur de $\ell^*(\boldsymbol{\alpha})$ et peut être obtenu comme la solution aux équations de pseudo-score données par $\mathbf{U}(\boldsymbol{\alpha}) = \partial\ell^*(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha} = \mathbf{0}$. Si la fonction logit inverse est présumée pour π_i^A, les fonctions de pseudo-score sont données par :

$$\mathbf{U}(\boldsymbol{\alpha}) = \sum_{i\in S_A}\mathbf{x}_i - \sum_{i\in S_B} d_i^B\pi(\mathbf{x}_i,\boldsymbol{\alpha})\,\mathbf{x}_i. \tag{4.3}$$

En général, les fonctions de pseudo-score $\mathbf{U}(\boldsymbol{\alpha})$ aux valeurs réelles des paramètres du modèle $\boldsymbol{\alpha}_0$ ne sont pas biaisées dans la randomisation qp conjointe dans le sens que $E_{qp}\{\mathbf{U}(\boldsymbol{\alpha}_0)\} = \mathbf{0}$, laquelle indique que l'estimateur $\hat{\boldsymbol{\alpha}}$ est qp-convergent pour $\boldsymbol{\alpha}_0$ (Tsiatis, 2006).

Valliant et Dever (2011) ont tenté plus tôt d'estimer les scores de propension en regroupant l'échantillon non probabiliste S_A avec l'échantillon probabiliste de référence S_B. Soit $S_{AB} = S_A \cup S_B$, l'échantillon groupé sans retirer toute unité dupliquée potentielle. Soit $R_i^* = 1$, si $i \in S_A$ et $R_i^* = 0$ si $i \in S_B$. Valliant et Dever (2011) ont proposé d'adapter un modèle de régression logistique pondéré d'enquête à l'ensemble des données groupées $\left\{(R_i^*, \mathbf{x}_i, d_i), i \in S_{AB}\right\}$, où les poids sont définis comme $d_i = 1$ si $i \in S_A$ et $d_i = d_i^B\left(1 - n_A/\hat{N}_B\right)$ si $i \in S_B$. La principale motivation derrière la création des poids d_i est que le poids total $\sum_{i\in S_{AB}} d_i = \sum_{i\in S_B} d_i^B = \hat{N}_B$ pour l'échantillon groupé correspond à la taille de la population estimée, et il reste à espérer que le modèle de régression logistique pondéré d'enquête entraînera des estimations valides pour les scores de propension. Chen et coll. (2020) ont démontré que l'approche de l'échantillon groupé de Valliant et Dever (2011) ne produit pas des estimateurs convergents pour les paramètres du modèle de scores de propension à moins que l'échantillon non probabiliste S_A ne soit un simple échantillon aléatoire de la population cible.

La méthode de Valliant et Dever (2011) révèle une difficulté fondamentale avec les approches fondées sur l'échantillon groupé S_{AB}. Si les unités dans l'échantillon non probabiliste S_A sont traitées comme des unités échangeables dans l'échantillon groupé S_{AB}, comme l'indiquaient les poids égaux $d_i = 1$ utilisés dans la méthode de Valliant et Dever (2011), les estimations résultantes des scores de propension ne seront pas valides à moins que S_A soit un simple échantillon aléatoire. Cette observation a des répercussions sur la validité des méthodes non paramétriques ou des méthodes à arborescence de régression qui seront abordées à la section 4.1.3.

Dans un article récent, Wang, Valliant et Li (2021) ont proposé une méthode de pondération par la propension logistique ajustée. La méthode comporte deux étapes de calcul des scores de propension estimés. Les estimations initiales, désignées comme S_{AB} pour $d_i = 1,$ sont obtenues en adaptant le modèle de régression logistique pondéré d'enquête à l'échantillon groupé $\hat{p}_i$ semblable à Valliant et Dever (2011), assorties des poids définis comme suit : $i \in S_A$ si $d_i = d_i^B$ et $i \in S_B$ si $i \in S_A$. Les scores de propension estimés définitifs sont calculés comme suit : $\hat{\pi}_i^A = \hat{p}_i / (1 - \hat{p}_i)$. Le principal argument théorique est que l'équation $\pi_i^A = p_i / (1 - p_i)$ où $\pi_i^A = P(i \in S_A \mid U)$, $p_i = P\left(i \in S_A^* \mid S_A^* \cup U\right)$ et S_A^*, est une copie de S_A, mais est considéré comme un ensemble différent. Toutefois, les arguments soulèvent des questions conceptuelles, car les probabilités $\pi_i^A = P(i \in S_A \mid U)$ sont définies dans le modèle de scores de propension théorique avec la population finie donnée U, et le modèle théorique ne mène pas à une interprétation significative des probabilités $p_i = P\left(i \in S_A^* \mid S_A^* \cup U\right)$. Ces dernières nécessitent un espace de probabilité différent et sont conditionnelles au S_A donné. En fait, on peut facilement faire valoir qu'en vertu du modèle de scores de propension théorique et conditionnel au S_A donné, on observe $p_i = 1$ si $i \in S_A$ et $p_i = 0$ autrement.

4.1.2 Estimation de méthodes fondées sur des équations

Les équations de pseudo-score $\mathbf{U}(\boldsymbol{\alpha}) = \mathbf{0}$ dérivées de la fonction de pseudo-vraisemblance $\ell^*(\boldsymbol{\alpha})$ peuvent être remplacées par un système d'équations d'estimation générales. Soit $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$, un vecteur personnalisé de fonctions avec la même dimension de $\boldsymbol{\alpha}$. Supposons que :

$$\mathbf{G}(\boldsymbol{\alpha}) = \sum_{i \in S_A} \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}) - \sum_{i \in S_B} d_i^B \pi(\mathbf{x}_i, \boldsymbol{\alpha}) \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}). \tag{4.4}$$

Il s'ensuit que $E_{qp}\{\mathbf{G}(\boldsymbol{\alpha}_0)\} = \mathbf{0}$ pour tout $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$ choisi. En principe, un estimateur $\hat{\boldsymbol{\alpha}}$ de $\boldsymbol{\alpha}$ peut être obtenu en résolvant $\mathbf{G}(\boldsymbol{\alpha}) = \mathbf{0}$ au moyen de la forme paramétrique choisie $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$ et des fonctions choisies $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$, et l'estimateur $\hat{\boldsymbol{\alpha}}$ est convergent.

L'estimateur $\hat{\boldsymbol{\alpha}}$ reposant sur des fonctions personnalisées arbitraires $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$ est habituellement moins efficace que celui fondé sur des fonctions de pseudo-score, en raison de l'optimalité de l'estimateur par le maximum de vraisemblance (Godambe, 1960). Quelques résultats empiriques limités montrent également que la solution à $\mathbf{G}(\boldsymbol{\alpha}) = \mathbf{0}$ peut se révéler instable pour certains choix de $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$. Néanmoins, les méthodes fondées sur des équations d'estimation procurent un outil utile pour l'estimation des scores de propension dans des scénarios plus restreints. Par exemple, si nous avons $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha}) = \mathbf{x} / \pi(\mathbf{x}, \boldsymbol{\alpha})$, les fonctions d'estimation données en (4.4) réduisent à :

$$\mathbf{G}(\boldsymbol{\alpha}) = \sum_{i \in S_A} \frac{\mathbf{x}_i}{\pi(\mathbf{x}_i, \boldsymbol{\alpha})} - \sum_{i \in S_B} d_i^B \mathbf{x}_i. \tag{4.5}$$

La forme de $\mathbf{G}(\boldsymbol{\alpha})$ en (4,5) ressemble à la version « déformée » des fonctions de pseudo-score données en (4.3) sous un modèle de régression logistique pour les scores de propension. Toutefois, la différence la

plus importante en pratique entre les deux versions est le fait que le $\mathbf{G}(\boldsymbol{\alpha})$ donné en (4.5) ne nécessite que des totaux de population estimés pour les variables auxiliaires $\mathbf{x}$. Il existe des scénarios où les totaux de population des variables auxiliaires $\mathbf{x}$ peuvent être consultés ou estimés à partir d'une source existante, mais les valeurs de $\mathbf{x}$ au niveau des unités pour l'ensemble de la population ou même juste un échantillon probabiliste ne sont pas disponibles. Grâce à l'utilisation des fonctions d'estimation $\mathbf{G}(\boldsymbol{\alpha})$ données (4.5), il a été possible d'obtenir des estimations valides des scores de propension pour les unités de l'échantillon non probabiliste. La section 6.3 donne un exemple dans lequel l'approche fondée sur des équations d'estimation mène à un estimateur de variance valide pour l'estimateur doublement robuste de la moyenne de la population.

4.1.3 Méthodes non paramétriques et méthodes fondées sur un arbre de régression

Les scores de propension $\pi_i^A = P\left(R_i = 1 \mid \mathbf{x}_i\right)$ sont la fonction de moyenne $E_q\left(R_i \mid \mathbf{x}_i\right) = \pi\left(\mathbf{x}_i\right)$ pour la réponse binaire R_i. Les méthodes non paramétriques pour l'estimation $\pi(\mathbf{x})$ peuvent constituer une solution de rechange intéressante. Le principal défi est d'élaborer des procédures d'estimation qui procurent des estimations valides des scores de propension. Comme il est mentionné à la section 4.1.1, les méthodes d'estimation fondées sur l'échantillon groupé $S_{AB} = S_A \cup S_B$ peuvent mener à des estimations invalides. Des stratégies semblables à celle utilisée par Chen et coll. (2020) peuvent être justifiées théoriquement dans le cadre *qp* conjoint, où les procédures d'estimation sont calculées d'abord à l'aide de données de l'ensemble de la population finie et où les quantités de populations inconnues sont ensuite remplacées par des estimations obtenues de l'échantillon probabiliste de référence.

Nous examinons l'estimateur par la régression par noyau de $\pi_i^A = \pi\left(\mathbf{x}_i\right)$. Supposons que l'ensemble de données $\left\{(R_i, \mathbf{x}_i), i = 1, 2, \ldots, N\right\}$ est disponible pour la population finie. Soit $K_h\left(t\right) = K\left(t/h\right)$, un noyau choisi ayant une largeur de bande h. L'estimateur par la régression par noyau Nadaraya-Watson (Nadaraya, 1964; Watson, 1964) de $\pi\left(\mathbf{x}\right)$ est donné par :

$$\tilde{\pi}\left(\mathbf{x}\right) = \frac{\sum_{j=1}^{N} K_h\left(\mathbf{x} - \mathbf{x}_j\right) R_j}{\sum_{j=1}^{N} K_h\left(\mathbf{x} - \mathbf{x}_j\right)}. \tag{4.6}$$

Un estimateur par noyau dans la forme de $\tilde{\pi}(\mathbf{x})$ donnée en (4.6) ne possède généralement pas de valeurs pratiques, car nous n'avons pas de renseignements auxiliaires complets pour la population finie. Il s'avère que, pour l'estimation des scores de propension, le numérateur en (4.6) n'a besoin que d'observations de l'échantillon non probabiliste provenant de la variable binaire R_j, et le dénominateur est un chiffre de population et peut être estimé en utilisant l'échantillon probabiliste de référence. L'estimateur par la régression par noyau non paramétrique des scores de propension est donné par (Yuan, Li et Wu, 2022) :

$$\hat{\pi}_i^A = \hat{\pi}\left(\mathbf{x}_i\right) = \frac{\sum_{j \in S_A} K_h\left(\mathbf{x}_i - \mathbf{x}_j\right)}{\sum_{j \in S_B} d_j^B K_h\left(\mathbf{x}_i - \mathbf{x}_j\right)}, \quad i \in S_A. \tag{4.7}$$

L'estimateur $\hat{\pi}_i^A$ donné en (4,7) est convergent en vertu du cadre qp conjoint et le modèle q pour les scores de propension est très flexible en raison de l'hypothèse non paramétrique sur $\pi(\mathbf{x})$. Les scores de propension estimés sont faciles à calculer quand la taille de $\mathbf{x}$ n'est pas trop élevée. Les problèmes avec un $\mathbf{x}$ de grandes dimensions et les choix du noyau $K_h(\cdot)$ et de la largeur de bande h subsistent, comme dans les applications générales des méthodes d'estimation à base de noyaux. Les résultats d'une simulation rapportés par Yuan et coll. (2022) montrent que la méthode de l'estimation par la méthode des noyaux donne de solides résultats pour les scores de propension à l'aide du noyau normal et de choix populaires pour la largeur de bande.

Chu et Beaumont (2019) ont examiné les méthodes fondées sur un arbre de régression. Leur méthode TrIPW proposée est une variante de l'algorithme CART (Breiman, Friedman, Olshen et Stone, 1984) et repose sur des données de l'échantillon combiné de l'échantillon non probabiliste et de l'échantillon probabiliste de référence. La méthode vise à construire un arbre de classification, les nœuds terminaux de l'arbre final étant traités comme des groupes homogènes en matière de scores de propension. L'estimateur de μ_y est construit en fonction de l'arbre final et de la poststratification. La section 5 contient d'autres renseignements sur les estimateurs stratifiés *a posteriori*.

Les techniques d'apprentissage statistique comme les arbres de classification et de régression et les forêts aléatoires d'arbres décisionnels ont été élaborées principalement à des fins de prédiction. Leur utilisation pour estimer les scores de propension d'échantillons non probabilistes nécessite des travaux supplémentaires. Il ne s'agit pas d'une approche souhaitable pour naïvement appliquer les méthodes sur l'échantillon groupé S_{AB} sans justification théorique sur la convergence des estimateurs finaux. Il faut encourager d'autres études dans ce sens.

4.2 Pondération de probabilité inverse

Soit $\hat{\pi}_i^A$, une estimation de $\pi_i^A = P\left(i \in S_A \mid \mathbf{x}_i\right)$ selon une méthode choisie pour l'estimation des scores de propension. Deux versions de l'estimateur pondéré de la probabilité inverse (PPI) de μ_y sont établies comme suit :

$$\hat{\mu}_{\text{PPI1}} = \frac{1}{N}\sum_{i\in S_A}\frac{y_i}{\hat{\pi}_i^A} \quad \text{et} \quad \hat{\mu}_{\text{PPI2}} = \frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i}{\hat{\pi}_i^A}, \tag{4.8}$$

où N est la taille de la population et $\hat{N}^A = \sum_{i\in S_A}(\hat{\pi}_i^A)^{-1}$ est la taille estimée de la population. L'estimateur $\hat{\mu}_{\text{PPI1}}$ est une version de l'estimateur de Horvitz-Thompson et $\hat{\mu}_{\text{PPI2}}$ correspond à l'estimateur Hájek comme en fait état la théorie de l'estimation fondée sur le plan. Il y a suffisamment de données probantes des justifications théoriques et des observations pratiques que l'estimateur Hájek $\hat{\mu}_{\text{PPI2}}$ donne de meilleurs résultats que l'estimateur Horvitz-Thrompson et qu'il doit être utilisé en pratique même si la taille de la population N est connue.

La validité des estimateurs PPI $\hat{\mu}_{\text{PPI1}}$ et $\hat{\mu}_{\text{PPI2}}$ dépend de la validité des scores de propension estimés. Selon les hypothèses A1 et A2 et le modèle paramétrique $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha}_0)$, la convergence de $\hat{\mu}_{\text{PPI1}}$ suit un

argument standard en deux étapes. Soit $\tilde{\mu}_{\text{PPI}} = N^{-1}\sum_{i\in S_A} y_i/\pi_i^A$, lequel n'est pas un estimateur calculable, mais un outil analytique utile à des fins asymptotiques. Il s'ensuit que $E_q\left(\tilde{\mu}_{\text{PPI}}\right) = \mu_y$ et l'ordre $V_q\left(\tilde{\mu}_{\text{PPI}}\right) = O\left(n_A^{-1}\right)$ se vérifient à la condition que $n_A\pi_i^A/N$ ait une borne de inférieure strictement positive. Par conséquent, nous avons $\tilde{\mu}_{\text{PPI}} \to \mu_y$ en probabilité quand $n_A \to \infty$. Dans le modèle correctement indiqué $\pi_i^A = \pi\left(\mathbf{x}_i, \boldsymbol{\alpha}_0\right)$ pour les scores de propension, l'ordre type n-racine $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 = O_p\left(n_A^{-1/2}\right)$ se vérifie pour les scénarios fréquemment rencontrés. Nous pouvons montrer qu'en traitant $\hat{\mu}_{\text{PPI1}}$ comme une fonction de $\hat{\boldsymbol{\alpha}}$ et en utilisant un développement en série de Taylor que $\hat{\mu}_{\text{PPI1}} = \tilde{\mu}_{\text{PPI}} + O_p\left(n_A^{-1/2}\right)$ dans certaines conditions peu contraignantes de moments finis. La convergence de $\hat{\mu}_{\text{PPI2}}$ peut être établie en utilisant des arguments standard pour un estimateur par le ratio (section 5.3, Wu et Thompson, 2020) où $N^{-1}\sum_{i\in S_A}\left(\pi_i^A\right)^{-1} = 1 + o_p(1)$.

4.3 Estimation doublement robuste

La dépendance de l'estimateur PPI sur la validité du modèle de scores de propension théorique est perçue comme une faiblesse de la méthode. Le problème n'est pas propre aux estimateurs PPI et se pose pour de nombreuses autres approches comportant un modèle statistique théorique. Les procédures robustes d'estimation qui procurent un certain degré de protection contre les erreurs de spécification du modèle ont été poursuivies par des chercheurs et lesdits estimateurs doublement robustes ont eu du succès depuis les travaux de Robins, Rotnitzky et Zhao (1994).

L'estimateur doublement robuste (DR) de μ_y est établi à l'aide du modèle de score de propension q et du modèle de régression des résultats ξ. L'estimateur DR avec les scores de propension donnés π_i^A, $i \in S_A$ et les réponses moyennes $m_i = E_\xi\left(y_i \mid \mathbf{x}_i\right)$, $i = 1, 2, \ldots, N$ présente la forme générale suivante :

$$\tilde{\mu}_{\text{DR}} = \frac{1}{N}\sum_{i\in S_A}\frac{y_i - m_i}{\pi_i^A} + \frac{1}{N}\sum_{i=1}^{N} m_i. \tag{4.9}$$

Le second terme du côté droit de l'équation (4.9) est la prédiction modéliste de μ_y. Le premier terme est un ajustement fondé sur le score de propension reposant sur les erreurs $\varepsilon_i = y_i - m_i$ du modèle de régression des résultats. L'importance du terme d'ajustement est négativement corrélée à la « qualité de l'ajustement » du modèle de régression des résultats. Il est possible de démontrer que $\tilde{\mu}_{\text{DR}}$ est un estimateur exactement sans biais de μ_y si l'un des deux modèles q et ξ est correctement précisé et ainsi doublement robuste. L'estimateur $\tilde{\mu}_{\text{DR}}$ possède une structure identique à l'estimateur par la différence généralisée de Wu et Sitter (2001). Il convient de souligner que la propriété de double robustesse de $\tilde{\mu}_{\text{DR}}$ ne nécessite pas de savoir lequel des deux modèles est correctement précisé. Il est également évident que l'estimateur $\tilde{\mu}_{\text{DR}}$ donné en (4.9) n'est pas calculable dans des applications pratiques.

Soient $\hat{\pi}_i^A$ et $\hat{m}_i$, les estimateurs de π_i^A et q respectivement, selon les modèles théoriques ξ et m_i. Dans le contexte de deux échantillons décrit à la section 2, les deux estimateurs DR de μ_y proposés par Chen et coll. (2020) sont donnés par :

$$\hat{\mu}_{\text{DR1}} = \frac{1}{N}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{N}\sum_{i\in S_B} d_i^B \hat{m}_i \tag{4.10}$$

et

$$\hat{\mu}_{\text{DR2}} = \frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{\hat{N}^B}\sum_{i\in S_B} d_i^B \hat{m}_i, \tag{4.11}$$

où d_i^B sont les poids déterminés par le plan d'échantillonnage pour l'échantillon probabiliste S_B, $\hat{N}^A = \sum_{i\in S_A}\left(\hat{\pi}_i^A\right)^{-1}$ et $\hat{N}^B = \sum_{i\in S_B} d_i^B$. L'estimateur $\hat{\mu}_{\text{DR2}}$ fondé sur la taille estimée de la population offre un meilleur rendement en matière de biais et d'erreur quadratique moyenne et devrait être utilisé dans la pratique.

Le plan d'enquête probabiliste p fait partie intégrante du cadre théorique pour évaluer les deux estimateurs $\hat{\mu}_{\text{DR1}}$ et $\hat{\mu}_{\text{DR2}}$. On suppose que S_A et S_B sont sélectionnés indépendamment, ce qui signifie que $E_p\left(\sum_{i\in S_B} d_i^B \hat{m}_i\right) = \sum_{i=1}^N \hat{m}_i$. La convergence des estimateurs $\hat{\mu}_{\text{DR1}}$ et $\hat{\mu}_{\text{DR2}}$ peut être établie en vertu de qp ou du cadre ξp. Il convient de souligner que même l'échantillon non probabiliste S_A est un simple échantillon aléatoire avec $\pi_i^A = n_A/N$, l'estimateur doublement robuste dans la forme de (4.9) ne se réduit pas à l'estimateur de prédiction modéliste $\hat{\mu}_{y2}$ donné en (3.3).

4.4 L'approche de la pseudo-vraisemblance empirique

Les méthodes de la pseudo-vraisemblance empirique (PEL pour *Pseudo Empirical Likelihood*) pour les échantillons d'enquête probabilistes sont en cours d'élaboration depuis deux décennies. Deux articles sur le sujet sont ceux de Chen et Sitter (1999) sur l'estimation ponctuelle intégrant de l'information auxiliaire, et celui de Wu et Rao (2006) sur les intervalles de confiance du rapport de PEL. Les approches de la PEL sont en outre utilisées pour des enquêtes à bases multiples (Rao et Wu, 2010a) et les inférences de Bayesian pour des données d'enquête (Rao et Wu, 2010b; Zhao, Ghosh, Rao et Wu, 2020b). L'utilisation des méthodes de la PEL pour des problèmes inférentiels généraux avec des enquêtes complexes a été étudiée dans deux articles récents (Zhao et Wu, 2019; Zhao, Rao et Wu, 2020a).

Chen, Li, Rao et Wu (2022) ont démontré que la PEL offre une solution de rechange intéressante à l'inférence avec des échantillons d'enquête non probabilistes. Soit $\hat{\pi}_i^A$, $i\in S_A$ étant les scores de propension estimés sous un modèle paramétrique ou non paramétrique théorique, q. La fonction de PEL pour l'échantillon d'enquête non probabiliste S_A est définie comme suit :

$$\ell_{\text{PEL}}(\mathbf{p}) = n_A\sum_{i\in S_A}\tilde{d}_i^A \log\left(p_i\right), \tag{4.12}$$

où $\mathbf{p} = (p_1,\ldots,p_{n_A})$ est une mesure de probabilité discrète sur les unités n_A sélectionnées dans S_A, $\tilde{d}_i^A = (\hat{\pi}_i^A)^{-1}/\hat{N}^A$ et $\hat{N}^A = \sum_{j\in S_A}(\hat{\pi}_j^A)^{-1}$ qui est définie précédemment dans la section 4. Sans utiliser de renseignements supplémentaires, maximiser $\ell_{\text{PEL}}\left(\mathbf{p}\right)$ sous la contrainte de normalisation suivante :

$$\sum_{i \in S_A} p_i = 1 \tag{4.13}$$

donne $\hat{p}_i = \tilde{d}_i^A$, $i \in S_A$. L'estimateur du maximum de PEL de μ_y est donné par $\hat{\mu}_{\text{PEL}} = \sum_{i \in S_A} \hat{p}_i y_i$, lequel est identique à l'estimateur PPI $\hat{\mu}_{\text{PPI2}}$ donné en (4.8).

L'approche de la PEL pour les échantillons d'enquête non probabilistes procure une flexibilité en combinant des renseignements du fait de contraintes supplémentaires et en construisant des intervalles de confiance et en menant des tests d'hypothèses à l'aide de statistiques sur le rapport de PEL. L'estimateur du maximum de PEL $\hat{\mu}_{\text{PEL}} = \sum_{i \in S_A} \hat{p}_i y_i$ est doublement robuste si $\left(\hat{p}_1, \ldots, \hat{p}_{n_A}\right)$ est le maximiseur de $\ell_{\text{PEL}}(\mathbf{p})$ sous la contrainte de normalisation et la contrainte de calage de modèle donné par :

$$\sum_{i \in S_A} p_i \hat{m}_i = \bar{m}^B, \tag{4.14}$$

où $\bar{m}^B = \left(\hat{N}^B\right)^{-1} \sum_{i \in S_B} d_i^B \hat{m}_i$ est calculé à l'aide des valeurs prédites $\hat{m}_i$, $i \in S_B$ d'un modèle de régression des résultats théorique, ξ. L'équation (4.14) est une version modifiée de la contrainte de calage de modèle originale de Wu et Sitter (2001) reposant sur l'échantillon probabiliste S_B. L'étude de Chen et coll. (2022) contient d'autres données détaillées sur les répartitions asymptotiques des statistiques sur le rapport de PEL et des études de simulation sur le rendement des intervalles de confiance du rapport de PEL sur une proportion de population finie.

5. Enquête par quotas et poststratification

Les enquêtes par quotas représentent l'une des plus vieilles méthodes d'échantillonnage non probabiliste qui sont encore utilisées en pratique de nos jours. Pour une taille globale prédéterminée d'échantillon n_A, les quotas des tailles d'échantillons sont établis pour des sous-populations qui sont définies par des variables démographiques et des indicateurs de l'état socioéconomique ou d'autres variables des caractéristiques appropriées pour la population cible. Les processus de collecte des données se poursuivent jusqu'à ce que les quotas pour chacune des sous-populations soient remplis. Les unités de la population sont habituellement abordées par tous les moyens pratiques existants, et il y a peu ou pas de contrôle sur la manière de sélectionner les unités pour l'échantillon définitif autre que des quotas prédéterminés.

La théorie des estimateurs PPI pour les échantillons d'enquêtes non probabiliste offre une occasion d'examiner des scénarios où les enquêtes par quotas peuvent réussir ou échouer. Pour faciliter la notation sans perte de généralité, supposons que S_A est l'échantillon d'enquête par quotas et que $\mathbf{x}$ est l'ensemble de variables catégoriques utilisé pour définir les sous-populations et fixer les quotas. L'échantillon global peut être divisé en $S_A = S_{A1} \cup \cdots \cup S_{AK}$, ce qui correspond au classement recoupé des unités échantillonnées au moyen des combinaisons de niveaux des variables $\mathbf{x}$. Par exemple, si $\mathbf{x} = (x_1, x_2)'$, x_1 ayant deux niveaux et x_2 ayant trois niveaux, nous avons au total $K = 2 \times 3 = 6$ sous-populations définies par $\mathbf{x}$. Supposons que n_k est la taille prédéterminée de S_{Ak} et que N_k est la taille de la sous-population

correspondante. Selon l'hypothèse A1, les scores de propension $\pi_i^A = \pi(\mathbf{x}_i)$ deviennent une constante pour des unités dans la même sous-population et sont donnés par $\pi_i^A = n_k / N_k$ pour la k^{e} sous-population. L'estimateur PPI $\hat{\mu}_{\text{PPI2}}$ donnée en (4.8) se réduit à :

$$\hat{\mu}_{\text{PPI2}} = \frac{1}{\hat{N}^A} \sum_{k=1}^{K} \sum_{i \in S_{Ak}} \frac{y_i}{\hat{\pi}_i^A} = \sum_{k=1}^{K} \hat{W}_k \bar{y}_k, \tag{5.1}$$

où $\bar{y}_k = n_k^{-1} \sum_{i \in S_{Ak}} y_i$. $\hat{W}_k = \hat{N}_k / \hat{N}^A$, $\hat{N}_k$ est la taille de la k^{e} sous-population obtenue ou estimée à partir de sources externes, et $\hat{N}^A = \sum_{k=1}^{K} \hat{N}_k$. Dans les conditions actuelles et compte tenu de la disponibilité d'un échantillon probabiliste de référence S_B, nous formons la même subdivision, recoupée selon les niveaux de $\mathbf{x}$ et obtenons $S_B = S_{B1} \cup \cdots \cup S_{BK}$. Nous pouvons alors utiliser $\hat{N}_k = \sum_{i \in S_{Bk}} d_i^B$.

L'estimateur donné en (5.1) est l'estimateur stratifié *a posteriori* standard de μ_y. Il nécessite l'information sur les « poids de strate » $\hat{W}_k$, $k = 1, \ldots, K$, qui n'est pas disponible à partir des données échantillon elles-mêmes. Les enquêtes par quotas, associées à l'utilisation de l'estimateur stratifié *a posteriori*, peuvent produire avec succès des estimations de population valides pour la variable de l'étude y si les conditions suivantes sont maintenues :

i) Les variables catégoriques $\mathbf{x}$ utilisées pour définir les sous-populations et établir des quotas offrent des caractérisations du comportement de participation des unités pour les enquêtes à participation volontaire.

ii) L'inclusion d'unités dans l'enquête est quelque peu aléatoire au sein de chaque sous-population et aucun groupe précis n'est exclu intentionnellement de l'enquête.

iii) L'information sur les poids de strate correspondant aux classements recoupés dans l'établissement des quotas peut être obtenue de manière fiable de sources externes.

iv) Les non-répondants irréductibles dans la population qui ne remplissent jamais d'enquêtes à participation volontaire présentent des traits semblables à ceux des répondants en ce qui a trait à la variable étudiée y.

Les estimateurs PPI $\hat{\mu}_{\text{PPI1}}$ et $\hat{\mu}_{\text{PPI2}}$ donnés en (4.8) peuvent être sensibles à des petites valeurs de scores de propension estimés. L'estimateur stratifié *a posteriori* dans la forme de (5.1) sert de solution de rechange fiable dans des scénarios généraux où la dimension de $\mathbf{x}$ n'est pas faible et où certaines composantes de $\mathbf{x}$ sont continues. Les strates K sont formées en fonction de groupes homogènes en matière de scores de propension. Supposons que $\hat{\pi}_i^A = \pi(\mathbf{x}_i, \hat{\boldsymbol{\alpha}})$, $i \in S_A$ est calculé en fonction d'un modèle paramétrique, q. Supposons également que $n_A = m_A K$ avec le K choisi où m_A est un nombre entier. Supposons que $\hat{\pi}_{(1)}^A \leq \cdots \leq \hat{\pi}_{(n_A)}^A$ représente les scores de propension estimés en ordre croissant. Supposons que S_{A1} est l'ensemble des premières unités m_A dans la séquence, S_{A2} représente les deuxièmes unités m_A dans la séquence, et ainsi de suite. L'estimateur stratifié *a posteriori* de μ_y est calculé comme $\hat{\mu}_{\text{PST}} = \sum_{k=1}^{K} \hat{W}_k \bar{y}_k$, qui a la même forme que l'estimateur donné en (5.1). Les estimations des poids de strate $\hat{W}_k$, $k = 1, 2, \ldots, K$ peuvent être obtenues en utilisant l'échantillon probabiliste de

référence S_B comme suit. Supposons que $b_k = \max\left\{\hat{\pi}_i^A : i \in S_{Ak}\right\}$, $k = 1, 2, \ldots, K-1$. Supposons que $b_0 = 0$ et que $b_K = 1$.

a) Calculer $\hat{\pi}_i = \pi(\mathbf{x}_i, \hat{\boldsymbol{\alpha}})$, $i \in S_B$.

b) Définir $S_{Bk} = \left\{i \,\middle|\, i \in S_B, b_{k-1} < \hat{\pi}_i \le b_k\right\}$, $k = 1, 2, \ldots, K$.

c) Calculer $\hat{N}_k = \sum_{i \in S_{Bk}} d_i^B$, $k = 1, 2, \ldots, K$.

Il est évident que $S_B = S_{B1} \cup \cdots \cup S_{BK}$ et $\sum_{k=1}^K \hat{N}_k = \hat{N}^B = \sum_{i \in S_B} d_i^B$. Les poids de strate estimés sont donnés par $\hat{W}_k = \hat{N}_k \big/ \hat{N}^B$.

Le choix de K doit refléter l'équilibre entre l'homogénéité des unités au sein de chaque strate *a posteriori* (en matière de scores de propension) et la stabilité de l'estimateur stratifié *a posteriori* (en matière de tailles d'échantillons de strate). Quand la taille de l'échantillon n_A est petite ou modérée, il faut utiliser un petit nombre comme $K = 5$. Pour des scénarios où n_A est grand, un plus grand K doit être utilisé de sorte que les unités au sein du même échantillon stratifié *a posteriori* S_{Ak} ont des scores de propension estimés similaires. Un guide pratique pour le choix de K est de s'assurer que $m_A \ge 30$ pour les échantillons stratifiés *a posteriori*. Pour ceux qui sont assez vieux, vous souvenez-vous du bon vieux temps quand « la taille de l'échantillon est grande » signifiait « $n \ge 30$ » ?

6. Estimation de la variance

L'estimation de la variance dans la configuration de deux échantillons S_A et S_B comporte au moins deux sources différentes de variation. Le plan d'échantillonnage avec probabilités pour l'échantillon de référence S_B reste l'une des sources employées, et ce, indépendamment des approches utilisées pour les échantillons d'enquête non probabilistes. L'estimation de la composante de variance provenant de l'utilisation de S_B nécessite des formules appropriées d'approximation de la variance ou des poids de rééchantillonnage comme faisant partie de l'ensemble de données de l'échantillon probabiliste de référence. Dans la présente section, notre analyse suppose que l'estimateur de variance fondé sur le plan pour l'estimateur ponctuel pondéré de l'enquête fondé sur S_B est disponible.

6.1 Estimation de la variance pour les estimateurs d'imputation massive

L'estimation de la variance pour l'estimateur de prédiction basé sur un modèle $\hat{\mu}_y$ comporte d'abord le calcul de la formule de la variance asymptotique pour $\mathrm{Var}\left(\hat{\mu}_y - \mu_y\right)$ selon le modèle de régression des résultats théorique ou le modèle d'imputation ξ et le plan d'échantillonnage avec probabilités p, et ensuite l'utilisation des diverses quantités inconnues de population.

L'estimateur d'imputation massive $\hat{\mu}_{y\mathrm{IM}} = \hat{N}_B^{-1} \sum_{i \in S_B} d_i^B y_i^*$ donné en (3.5) est un type spécial d'estimateur de prédiction basé sur un modèle, où le modèle ξ fait référence à celui utilisé pour l'imputation et n'est pas nécessairement le même que le modèle de régression des résultats. La méthode de

l'imputation joue un rôle clé dans le calcul de la formule de la variance asymptotique et l'estimateur de la variance doit être établi en conséquence. Il convient de souligner que $\hat{\mu}_{y\mathrm{IM}}$ est un estimateur de type Hájek en raison de l'utilisation de la taille de population estimée $\hat{N}_B$ et que les calculs de la formule de la variance asymptotique commencent en mettant la valeur réelle N en premier et en traitant ensuite $\hat{\mu}_{y\mathrm{IM}}$ comme un estimateur par le ratio. Kim et coll. (2021) ont examiné l'estimation de la variance pour $\hat{\mu}_y = N^{-1}\sum_{i\in S_B} d_i^B y_i^*$, où $y_i^* = m\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}}\right)$ est la valeur imputée pour y_i selon le modèle semi-paramétrique (3.1). La formule de la variance asymptotique est élaborée en deux étapes. Premièrement, une version linéarisée de $\hat{\mu}_y$ est obtenue en utilisant un développement en série de Taylor à $\boldsymbol{\beta}^*$, où $\boldsymbol{\beta}^*$ est la limite de probabilité de $\hat{\boldsymbol{\beta}}$ de sorte que $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + O_p\left(n_A^{-1/2}\right)$. Deuxièmement, deux composantes de variance sont calculées pour $\mathrm{Var}\left(\hat{\mu}_y - \mu_y\right)$ en fonction de la version linéarisée en utilisant le modèle semi-paramétrique (3.1) et le plan d'échantillonnage pour S_B. Le processus est fastidieux, ce qui est le cas pour la plupart des méthodes d'estimation de la variance basées sur un modèle. Un estimateur de la variance bootstrap se révèle être plus attrayant pour des applications pratiques. Voir Kim et coll. (2021) pour obtenir d'autres renseignements détaillés.

6.2 Estimation de la variance pour les estimateurs PPI

L'estimateur PPI couramment utilisé $\hat{\mu}_{\mathrm{PPI2}}$ donné en (4.8) est valide dans le modèle théorique q pour les scores de propension. Une formule de variance asymptotique pour $\hat{\mu}_{\mathrm{PPI2}}$ peut être obtenue selon le cadre qp conjoint quand les scores de propension sont estimés à l'aide de la méthode du pseudo-maximum de vraisemblance ou d'une méthode fondée sur une équation d'estimation comme en fait état la section 4.1. L'outil théorique est la formule de variance de type sandwich pour les estimateurs ponctuels définis comme la solution à un système combiné d'équations d'estimation pour μ_y et $\boldsymbol{\alpha}_0$.

Examinons la forme paramétrique $\pi_i^A = \pi\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$ pour les scores de propension, où les paramètres du modèle $\boldsymbol{\alpha}$ sont estimés à l'aide des équations d'estimation (4.4) avec des fonctions personnalisées $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$. La première étape majeure dans le calcul de la formule de la variance asymptotique pour $\hat{\mu}_{\mathrm{PPI2}}$ est de formuler le système des équations d'estimation conjointes pour μ_y et $\boldsymbol{\alpha}_0$. Soit $\boldsymbol{\eta} = \left(\mu, \boldsymbol{\alpha}'\right)'$, le vecteur des paramètres combinés. L'estimateur $\hat{\boldsymbol{\eta}} = \left(\hat{\mu}_{\mathrm{PPI2}}, \boldsymbol{\alpha}'\right)'$ est la solution au système d'équations d'estimation conjointes $\boldsymbol{\Phi}_n\left(\boldsymbol{\eta}\right) = \mathbf{0}$, où :

$$\boldsymbol{\Phi}_n\left(\boldsymbol{\eta}\right) = \begin{pmatrix} N^{-1}\sum_{i=1}^N R_i\left(y_i - \mu\right)/\pi_i^A \\ N^{-1}\sum_{i=1}^N R_i \mathbf{h}\left(\mathbf{x}_i, \boldsymbol{\alpha}\right) - N^{-1}\sum_{i\in S_B} d_i^B \pi_i^A \mathbf{h}\left(\mathbf{x}_i, \boldsymbol{\alpha}\right) \end{pmatrix}. \tag{6.1}$$

Le facteur N^{-1} est redondant, mais utile pour faciliter les ordres asymptotiques. Les fonctions d'estimation définie par (6.1) sont sans biais selon le cadre qp conjoint, c'est-à-dire $E_{qp}\left\{\boldsymbol{\Phi}\left(\boldsymbol{\eta}_0\right)\right\} = \mathbf{0}$, où $\boldsymbol{\eta}_0 = \left(\mu_y, \boldsymbol{\alpha}_0'\right)'$. Il y a deux conséquences majeures à l'absence de biais du système d'équations d'estimation. Premièrement, on peut soutenir la convergence de l'estimateur $\hat{\boldsymbol{\eta}}$ en utilisant la théorie des fonctions d'estimation générales semblables à celles présentées à la section 3.2 de Tsiatis (2006).

Deuxièmement, la matrice de covariance-variance asymptotique de $\hat{\boldsymbol{\eta}}$, désignée comme $\mathrm{AV}(\hat{\boldsymbol{\eta}})$, a la forme standard en sandwich et est donnée par :

$$\mathrm{AV}(\hat{\boldsymbol{\eta}}) = \left[E\{\boldsymbol{\phi}_n(\boldsymbol{\eta}_0)\}\right]^{-1} \mathrm{Var}\{\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0)\}\left[E\{\boldsymbol{\phi}_n(\boldsymbol{\eta}_0)\}'\right]^{-1},$$

où $\boldsymbol{\phi}_n(\boldsymbol{\eta}) = \partial\boldsymbol{\Phi}_n(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$, laquelle dépend des formes de $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$ et $\mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha})$. Le terme $\mathrm{Var}\{\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0)\}$ comporte deux composantes, une en raison du modèle du score de propension q et l'autre en raison du plan d'échantillonnage avec probabilités pour S_B. Plus particulièrement, nous avons $\mathrm{Var}\{\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0)\} = V_q(\mathbf{A}_1) + V_p(\mathbf{A}_2)$, où $V_q(\cdot)$ désigne la variance sous le modèle du score de propension q et $V_p(\cdot)$ représente la variance fondée sur le plan sous le plan d'échantillonnage avec probabilités p, et

$$\mathbf{A}_1 = \frac{1}{N}\sum_{i=1}^{N} R_i \begin{pmatrix} (y_i - \mu)/\pi_i^A \\ \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}) \end{pmatrix}, \quad \mathbf{A}_2 = \frac{1}{N}\sum_{i\in S_B} d_i^B \begin{pmatrix} 0 \\ \pi_i^A \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}) \end{pmatrix}.$$

L'expression analytique pour $V_q(\mathbf{A}_1)$ découle immédiatement de $V_q(R_i) = \pi_i^A(1-\pi_i^A)$ et l'indépendance parmi $R_1, \ldots, R_N$. La composante de variance fondée sur le plan $V_p(\mathbf{A}_2)$ nécessite des renseignements supplémentaires sur le plan d'enquête pour S_B ou une formule d'approximation de la variance appropriée avec le plan donné.

La formule de la variance asymptotique pour l'estimateur PPI $\hat{\mu}_{\mathrm{PPI2}}$ est le premier élément en diagonale de la matrice $\mathrm{AV}(\hat{\boldsymbol{\eta}})$. L'estimateur définitif de la variance pour $\hat{\mu}_{\mathrm{PPI2}}$ peut alors être obtenu en remplaçant diverses quantités de population par des estimateurs des moments fondés sur l'échantillon. Dans leur étude, Chen et coll. (2020) ont présenté l'estimateur de la variance avec des expressions explicites quand $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$ sont modélisés par la régression logistique et $\hat{\boldsymbol{\alpha}}$ est obtenu par la méthode du pseudo-maximum de vraisemblance.

6.3 Estimation de la variance pour les estimateurs doublement robustes

Il s'avère que l'estimation de la variance pour l'estimateur doublement robuste est un problème difficile. Alors que la double robustesse est une propriété souhaitable pour l'estimation ponctuelle, elle pose un dilemme pour l'estimation de la variance. L'estimateur $\hat{\mu}_{\mathrm{DR2}}$ donné en (4.11) est convergent si le modèle de score de propension q ou le modèle de régression des résultats ξ est correctement précisé. Il n'existe aucun besoin de savoir quel modèle est correctement précisé, ce qui constitue la partie la plus cruciale derrière la double robustesse. Toutefois, cette caractéristique ambiguë devient un problème pour une estimation de la variance. La formule de la variance asymptotique selon le modèle q est habituellement différente de celle selon le modèle ξ et, par conséquent, il est difficile de concevoir un estimateur de variance cohérent avec des scénarios inconnus sur les spécifications du modèle.

Plusieurs stratégies ont été proposées dans les ouvrages portant sur l'estimation de la variance pour les estimateurs doublement robustes. Une approche naïve consiste à utiliser l'estimateur de la variance dérivé dans le modèle de score de propension théorique q et de prendre le risque qu'un tel estimateur de la

variance puisse avoir des biais non négligeables dans le modèle de régression des résultats. Une bonne nouvelle est que, dans le modèle de score de propension, l'estimation des paramètres $\boldsymbol{\beta}$ pour le modèle de régression des résultats n'a aucune incidence asymptotiquement sur la variance des estimateurs doublement robustes. Nous pouvons observer cela en utilisant $\hat{\mu}_{\text{DR1}}$ de (4.10) comme un exemple. Soit $\hat{m}_i = m\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}}\right)$, où $\hat{\boldsymbol{\beta}}$ est obtenu en fonction du modèle de travail (3.1) qui n'est pas nécessairement correct. Soit $\boldsymbol{\beta}^*$, la limite de probabilité de $\hat{\boldsymbol{\beta}}$ de sorte que $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + O_p\left(n_A^{-1/2}\right)$ indépendamment du vrai modèle de régression des résultats (White, 1982). Supposons que $m_i^* = m\left(\mathbf{x}_i, \boldsymbol{\beta}^*\right)$ et que $\mathbf{a}(\mathbf{x}, \boldsymbol{\beta}) = \partial m(\mathbf{x}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$. Nous pouvons constater que :

$$\frac{1}{N}\sum_{i\in S_B} d_i^B \hat{m}_i \; - \; \frac{1}{N}\sum_{i\in S_A} \frac{\hat{m}_i}{\hat{\pi}_i^A} \; = \; \frac{1}{N}\sum_{i\in S_B} d_i^B m_i^* \; - \; \frac{1}{N}\sum_{i\in S_A} \frac{m_i^*}{\hat{\pi}_i^A} \; + \; \left\{\mathbf{B}\left(\boldsymbol{\beta}^*\right)\right\}' \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \; + \; o_p\left(n_A^{-1/2}\right),$$

où

$$\mathbf{B}\left(\boldsymbol{\beta}^*\right) \; = \; \frac{1}{N}\sum_{i\in S_B} d_i^B \mathbf{a}\left(\mathbf{x}_i, \boldsymbol{\beta}^*\right) \; - \; \frac{1}{N}\sum_{i\in S_A} \frac{\mathbf{a}\left(\mathbf{x}_i, \boldsymbol{\beta}^*\right)}{\hat{\pi}_i^A}. \tag{6.2}$$

Comme les deux termes sur le côté droit de (6.2) sont tous les deux des estimateurs de $N^{-1}\sum_{i=1}^N \mathbf{a}\left(\mathbf{x}_i, \boldsymbol{\beta}^*\right)$, nous concluons que $\mathbf{B}\left(\boldsymbol{\beta}^*\right) = o_p(1)$ et

$$\frac{1}{N}\sum_{i\in S_B} d_i^B \hat{m}_i \; - \; \frac{1}{N}\sum_{i\in S_A} \frac{\hat{m}_i}{\hat{\pi}_i^A} \; = \; \frac{1}{N}\sum_{i\in S_B} d_i^B m_i^* \; - \; \frac{1}{N}\sum_{i\in S_A} \frac{m_i^*}{\hat{\pi}_i^A} \; + \; o_p\left(n_A^{-1/2}\right).$$

Il s'ensuit que :

$$\hat{\mu}_{\text{DR1}} \; = \; \frac{1}{N}\sum_{i\in S_A} \frac{y_i - m_i^*}{\hat{\pi}_i^A} \; + \; \frac{1}{N}\sum_{i\in S_B} d_i^B m_i^* \; + \; o_p\left(n_A^{-1/2}\right).$$

Les mêmes arguments s'appliquent à $\hat{\mu}_{\text{DR2}}$. Nous pouvons traiter $\hat{\boldsymbol{\beta}}$, car il est fixe dans le calcul de la variance asymptotique pour $\hat{\mu}_{\text{DR1}}$ et $\hat{\mu}_{\text{DR2}}$ sous le modèle de score de propension théorique. Les techniques décrites à la section 6.2 peuvent être utilisées directement lorsque la première fonction d'estimation dans (6.1) est remplacée par celle servant à définir $\hat{\mu}_{\text{DR1}}$ ou $\hat{\mu}_{\text{DR2}}$. Voir le théorème 2 de Chen et coll. (2020) pour obtenir d'autres renseignements détaillés. Toutefois, l'estimateur de la variance calculé selon le modèle du score de propension théorique présente généralement un biais dans le modèle de régression des résultats.

Dans l'étude de Chen et coll. (2020), on a également décrit une technique faisant appel à une idée originale présentée par Kim et Haziza (2014) pour la construction du soi-disant estimateur de variance doublement robuste. Cette technique est délicate et présente un certain attrait théorique, mais pose divers problèmes pour des applications pratiques. Nous utilisons $\hat{\mu}_{\text{DR1}}$ comme un exemple pour illustrer les étapes de la construction de l'estimateur de variance doublement robuste. Supposons que :

$$\hat{\mu}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} R_i \frac{y_i - m(\mathbf{x}_i,\boldsymbol{\beta})}{\pi(\mathbf{x}_i,\boldsymbol{\alpha})} + \frac{1}{N}\sum_{i\in S_B} d_i^B m(\mathbf{x}_i,\boldsymbol{\beta}).$$

Il s'ensuit que $\hat{\mu}_{\text{DR1}} = \hat{\mu}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})$ si $\hat{\boldsymbol{\alpha}}$ et $\hat{\boldsymbol{\beta}}$ sont les méthodes d'estimation originales. La première étape consiste à modifier l'estimation de $\boldsymbol{\alpha}$ et de $\boldsymbol{\beta}$ de manière à obtenir $\hat{\boldsymbol{\alpha}}$ et $\hat{\boldsymbol{\beta}}$ comme solutions à :

$$\frac{\partial\hat{\mu}(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial\boldsymbol{\alpha}} = \mathbf{0} \quad \text{et} \quad \frac{\partial\hat{\mu}(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \mathbf{0}. \tag{6.3}$$

Dans le modèle de régression logistique q où $\text{logit}\{\pi(\mathbf{x}_i,\boldsymbol{\alpha})\} = \mathbf{x}_i'\boldsymbol{\alpha}$ et le modèle de régression linéaire ξ où $m(\mathbf{x}_i,\boldsymbol{\beta}) = \mathbf{x}_i'\boldsymbol{\beta}$, le système d'équations (6.3) devient :

$$\frac{1}{N}\sum_{i=1}^{N} R_i \left\{\frac{1}{\pi(\mathbf{x}_i,\boldsymbol{\alpha})} - 1\right\}(y_i - \mathbf{x}_i'\boldsymbol{\beta})\,\mathbf{x}_i = \mathbf{0}, \tag{6.4}$$

$$\frac{1}{N}\sum_{i=1}^{N} \frac{R_i\mathbf{x}_i}{\pi(\mathbf{x}_i,\boldsymbol{\alpha})} - \frac{1}{N}\sum_{i\in S_B} d_i^B\mathbf{x}_i = \mathbf{0}. \tag{6.5}$$

Les équations d'estimation dans (6.5) sont sans biais selon le cadre qp conjoint. Elles sont identiques à l'équation (4.5) abordée à la section 4.1.2. Les équations d'estimation en (6.4) sont également sans biais dans le modèle de régression des résultats, mais elles sont différentes des équations de quasi-score qui figurent à l'équation (3.2). Les estimateurs $\hat{\boldsymbol{\alpha}}$ et $\hat{\boldsymbol{\beta}}$ obtenus comme solutions à (6.4) et (6.5) sont moins stables que ceux des méthodes standard. En outre, le système d'équations (6.4) et (6.5) n'aura pas de solution si $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ n'ont pas la même dimension, car le nombre d'équations en (6.4) est déterminé par la dimension de $\boldsymbol{\alpha}$ et le nombre d'équations en (6.5) est le même que la dimension de $\boldsymbol{\beta}$. L'estimateur définitif $\hat{\mu}_{\text{DR}} = \hat{\mu}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})$ subit également des pertes d'efficacité quand $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ sont estimés en résolvant (6.4) et (6.5).

La raison expliquant l'utilisation du système d'équations (6.3) est purement technique. Cela peut être démontré par une expansion de premier ordre de Taylor que les estimateurs $\hat{\boldsymbol{\alpha}}$ et $\hat{\boldsymbol{\beta}}$ tirés de (6.3) n'ont pas d'effet asymptotiquement sur la variance de $\hat{\mu}_{\text{DR}} = \hat{\mu}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})$. Cette manœuvre technique permet que des expressions explicites simples pour la variance $V_{qp}(\hat{\mu}_{\text{DR}})$ selon le cadre qp et pour la variance de prédiction $V_{\xi p}(\hat{\mu}_{\text{DR}} - \mu_y)$ selon le cadre ξp puissent aisément être obtenues. La construction d'un estimateur de variance doublement robuste pour $\hat{\mu}_{\text{DR}}$ commence par l'estimateur par substitution pour $V_{qp}(\hat{\mu}_{\text{DR}})$ dans le modèle de scores de propension q. Un terme de correction de biais est ensuite ajouté pour obtenir un estimateur valide pour $V_{\xi p}(\hat{\mu}_{\text{DR}} - \mu_y)$ dans le modèle de régression des résultats ξ. L'aspect positif est que le terme de correction du biais possède une forme analytique $N^{-2}\sum_{i=1}^{N}(R_i/\pi_i^A - 1)\sigma_i^2$ où $\sigma_i^2 = E_\xi(y_i \mid \mathbf{x}_i)$, lequel est négligeable dans le modèle de scores de propension q. L'estimateur de variance corrigé pour le biais est valide dans le modèle de scores de propension ou le modèle de régression des résultats.

Un estimateur de variance doublement robuste pour le $\hat{\mu}_{\mathrm{DR2}}$ couramment utilisé ne figure pas dans la littérature. Une solution pratique est d'utiliser des méthodes bootstrap. Chen et coll. (2022) ont démontré que les procédures bootstrap standard avec remise appliquées séparément à S_A et à S_B procuraient des intervalles de confiance doublement robustes en utilisant une approche fondée sur la pseudo-vraisemblance empirique à des échantillons d'enquête non probabilistes quand l'échantillon de référence est sélectionné par des plans d'échantillonnage avec probabilités inégales à un seul degré. Des complications surviennent quand l'échantillon probabiliste S_B s'appuie sur des méthodes d'échantillonnage à plusieurs degrés, un défi connu pour l'estimation de la variance dans le cadre d'enquêtes complexes. La construction d'estimateurs de variance doublement robustes pour l'estimateur doublement robuste $\hat{\mu}_{\mathrm{DR2}}$ dans des conditions générales mérite que l'on y consacre des efforts dans de futurs travaux de recherche.

7. Hypothèses réexaminées

Nos commentaires portant sur les procédures d'estimation pour des échantillons d'enquête non probabilistes figurent dans la section sur les hypothèses A1 à A4 et l'accent est mis sur la validité et l'efficacité des estimateurs pour la moyenne de population finie selon trois cadres inférentiels. Les résultats théoriques sur la prédiction fondée sur un modèle, la pondération de probabilité inverse et une estimation doublement robuste ont été rigoureusement établis dans le cadre de ces hypothèses. Il semble que les chercheurs ont du succès lorsqu'ils traitent du domaine émergent des sources de données non probabilistes. Toutefois, comme l'a souligné le président de l'ASA 2021 Robert Santos dans sa tribune intitulée « Utiliser nos superpouvoirs pour contribuer à l'intérêt public » (en anglais, Armstat News, mai 2021), « Nos superpouvoirs ne sont aussi fiables que leurs hypothèses sous-jacentes, hypothèses qui sont trop souvent acceptées avec assurance, sans pourtant pouvoir être prouvées » [traduction]. La façon de vérifier les hypothèses A1 à A4 dans des applications pratiques des méthodes est une question à laquelle on ne peut jamais répondre complètement et ils restent des étapes à suivre pour renforcer la confiance en utilisant les résultats théoriques. Il est également important de comprendre les conséquences possibles quand certaines hypothèses deviennent très discutables.

7.1 Hypothèse A1

L'hypothèse A1 stipule que $\pi_i^A = P\left(R_i = 1 \middle| \mathbf{x}_i, y_i\right) = P\left(R_i = 1 \middle| \mathbf{x}_i\right)$. Il s'agit de l'hypothèse la plus importante pour la validité de l'estimateur du pseudo-maximum de vraisemblance de Chen et coll. (2020) et l'estimateur obtenu par lissage par la méthode du noyau non paramétrique présenté à la section 4.1.3 pour les scores de propension, bien que toutes les autres hypothèses soient également en cause. Elle est équivalente à l'hypothèse des données manquantes au hasard pour les ouvrages publiés sur les données manquantes. Il est bien compris que l'hypothèse des données manquantes au hasard ne peut être testée à l'aide des données elles-mêmes de l'échantillon. Le même énoncé vaut pour l'hypothèse A1 comportant des échantillons d'enquête non probabilistes.

En résumé, l'hypothèse A1 indique que les variables auxiliaires $\mathbf{x}$ comprises dans l'échantillon non probabiliste caractérisent complètement le comportement de participation ou le mécanisme d'inclusion des échantillons pour des unités dans la population. Une attention suffisante doit être donnée à l'étape de conception de l'étude avant la collecte des données, si une telle étape existe, afin d'examiner les facteurs et les caractéristiques potentiels des unités qui peuvent être liées à la participation et à l'inclusion des échantillons. Pour des populations humaines, les facteurs et les caractéristiques peuvent comprendre des variables démographiques, des indicateurs sociaux et économiques ainsi que des variables géographiques.

L'hypothèse A1 mène à la conclusion que la répartition conditionnelle de y étant donné $\mathbf{x}$ pour des unités de l'échantillon non probabiliste est la même que la répartition conditionnelle de y étant donné $\mathbf{x}$ pour les unités de la population cible. Cela suppose que les variables auxiliaires $\mathbf{x}$ doivent comprendre des prédicteurs pertinents pour la variable étudiée y. Pour ce qui est des ensembles de données établis S_A et S_B, une analyse de sensibilité par comparaison des répartitions marginales et des modèles conditionnels peut être utile pour renforcer la confiance à l'égard de l'hypothèse A1. Pour les variables qui sont présentées dans S_A et S_B, on peut comparer les fonctions de répartition empirique (ou moments) de S_A aux fonctions de répartition empirique pondérés de l'enquête (ou moments) de S_B. Des différences marquées entre les deux indiquent que S_A est un échantillon non probabiliste avec des scores de propension inégaux. Une analyse de sensibilité possible sur l'hypothèse A1 consiste à sélectionner une variable z qui présente certaines similarités avec y, et un ensemble de variables auxiliaires $\mathbf{u}$ avec z et $\mathbf{u}$ disponibles de S_A et de S_B. Nous avons ajusté un modèle conditionnel $z \mid \mathbf{u}$ en utilisant des données de S_A et un modèle conditionnel pondéré d'enquête $z \mid \mathbf{u}$ en utilisant des données de S_B. Si $\mathbf{u}$ comprend toutes les variables auxiliaires principales pour l'hypothèse A1, nous devrons voir que les deux versions des modèles ajustés sont semblables. D'importantes différences entre les deux modèles ajustés représentent un fort signe que le z est lui-même une variable auxiliaire importante de l'hypothèse A1 ou que l'hypothèse est discutable.

7.2 Hypothèse A2

Un coup d'œil à l'hypothèse A2 peut donner à penser qu'elle pourrait facilement être satisfaite en pratique, car une hypothèse similaire est largement utilisée dans les analyses sur les données manquantes et les inférences causales. Il s'avère que l'hypothèse peut se révéler très problématique et pour des scénarios où l'hypothèse ne peut se vérifier, la population cible est différente de celle supposée pour les méthodes d'estimation. Elle est semblable au sous-dénombrement incomplet de la base de sondage et à des problèmes de non-réponse, lesquels sont longuement étudiés dans l'échantillonnage probabiliste.

L'hypothèse A2 stipule que $\pi_i^A = P\left(R_i = 1 \mid \mathbf{x}_i, y_i\right) > 0$ pour tous les i. Elle équivaut à déclarer que chaque unité de la population cible présente une probabilité non nulle d'être incluse dans l'échantillon non probabiliste. Si l'échantillon était prélevé par une méthode d'échantillonnage probabiliste, il s'agirait d'un scénario où la base de sondage est complète et où il n'y a pas de non-répondants irréductibles. Pour la plupart des échantillons non probabilistes, le concept de « *base de sondage* » est souvent non pertinent ou

simplement, une liste pratique et le processus de sélection et d'inclusion d'unités pour l'échantillon peut se révéler non structuré. Dans sa présentation à l'atelier CANSSI-NISS de 2021, Mary Thompson mentionnait que « l'énoncé que l'indicateur d'inclusion d'un échantillon R est une variable aléatoire est lui-même une hypothèse » [traduction] pour les échantillons d'enquête non probabilistes.

Supposons que U est l'ensemble de N unités pour la population cible. Soit $U_0 = \left\{ i \,\middle|\, i \in U \text{ et } \pi_i^A > 0 \right\}$. Il est évident que $U_0 \subset U$ et $U_0 \neq U$ quand l'hypothèse A2 n'est pas respectée. On retrouve deux scénarios types en pratique. On peut appeler le premier un *sous-dénombrement stochastique*, où l'échantillon non probabiliste S_A est sélectionné d U_0 et U_0 lui-même peut être perçu comme un échantillon aléatoire de U. Par exemple, la liste de personnes à joindre d'une enquête probabiliste existante est utilisée pour approcher des unités de la population pour une participation dans l'échantillon non probabiliste. Dans ce cas, U_0 est constitué d'unités de l'échantillon probabiliste. Un autre exemple est une enquête auprès de volontaires où la population cible est constituée d'adultes dans une ville ou une région particulière, mais où les participants sont recrutés parmi les visiteurs de grands centres commerciaux de la région au cours d'une période donnée. La sous-population U_0 comprend des visiteurs aux emplacements choisis au cours de la période d'échantillonnage et il est raisonnable de supposer que U_0 est un échantillon aléatoire de la population cible. Soit $D_i = 1$, si $i \in U_0$ et $D_i = 0$ autrement, $i = 1, 2, \ldots, N$. Nous observons :

$$P\left(R_i = 1 \,\middle|\, \mathbf{x}_i, y_i, D_i = 1\right) > 0 \quad \text{et} \quad P\left(R_i = 1 \,\middle|\, \mathbf{x}_i, y_i, D_i = 0\right) = 0$$

pour $i = 1, 2, \ldots, N$. Si la sous-population U_0 est formée avec un mécanisme stochastique sous-jacent tel que $P\left(D_i = 1 \,\middle|\, \mathbf{x}_i, y_i\right) > 0$ pour tous les $i \in U$, nous avons :

$$\pi_i^A = P\left(R_i = 1 \,\middle|\, \mathbf{x}_i, y_i\right) = P\left(R_i = 1 \,\middle|\, \mathbf{x}_i, y_i, D_i = 1\right) P\left(D_i = 1 \,\middle|\, \mathbf{x}_i, y_i\right) > 0$$

pour $i = 1, 2, \ldots, N$. En d'autres mots, l'hypothèse A2 est valide selon le scénario du sous-dénombrement stochastique pour les échantillons non probabilistes.

Le second scénario est appelé *sous-dénombrement déterministe* où les unités présentant certaines caractéristiques ne seront jamais incluses dans l'échantillon non probabiliste. Supposons que la participation dans l'enquête non probabiliste nécessite un accès à Internet et une adresse de courriel valide, et que 20 % de la population n'a pas accès à Internet ni à une adresse de courriel, nous avons un exemple où 20 % de la population aura des scores de propension nuls. Il n'existe pas de solution simple aux procédures inférentielles élaborées sous l'hypothèse A2. La thèse de doctorat de Yilin Chen de l'université de Waterloo (Chen, 2020) contenait un chapitre traitant de certains aspects précis du scénario.

7.3 Hypothèse A3

Parmi toutes les hypothèses, celle-ci est la moins essentielle à la validité des procédures inférentielles proposées. Dans l'hypothèse A3, la fonction de vraisemblance complète pour les scores de propension est

donnée dans (4.1). Pour tout modèle paramétrique sur $\pi_i^A = \pi\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$, la fonction de logarithme du rapport de quasi-vraisemblance $\ell^*(\boldsymbol{\alpha})$ donnée en (4.2) mène aux fonctions de quasi-scores $\mathbf{U}(\boldsymbol{\alpha}) = \partial \ell^*(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}$, lesquelles restent sans biais même si l'hypothèse A3 n'est pas respectée. Il pourrait y avoir certaines pertes d'efficacité sans l'hypothèse A3 lors de l'estimation des paramètres du modèle $\boldsymbol{\alpha}$, mais les méthodes d'estimation sont encore valides dans les trois autres hypothèses.

7.4 Hypothèse A4

Il n'est pas difficile de trouver un échantillon probabiliste existant de la même population cible. Toutefois, il peut s'avérer très difficile d'avoir un échantillon d'enquête probabiliste qui contient les variables auxiliaires souhaitables. Des enquêtes probabilistes existantes sont conçues à des fins précises et des objectifs scientifiques, et les variables auxiliaires comprises dans l'enquête ne sont pas nécessairement adaptées à l'analyse d'un échantillon d'enquête non probabiliste particulier. L'objectif ultime pour satisfaire l'hypothèse A4 est d'établir et d'avoir accès à un échantillon d'enquête probabiliste existant avec une riche collecte de variables démographiques, d'indicateurs sociaux et économiques et de variables géographiques.

Un problème des personnes riches (quand une personne a trop d'argent) pour l'hypothèse A4 peut également survenir quand au moins deux échantillons d'enquête probabilistes sont disponibles. La façon de combiner toutes ces données pour réaliser une analyse efficace des échantillons d'enquête non probabilistes est un sujet de recherche qui mérite plus d'attention. Certains conseils pratiques sur le choix d'un échantillon probabiliste de référence de solutions de rechange possibles comprennent les éléments suivants.

i) Vérifier la disponibilité de variables auxiliaires importantes qui présentent un intérêt pour la caractérisation du comportement de participation ou pour donner un pouvoir prédictif à l'étude des variables dans l'échantillon non probabiliste;

ii) Accorder la priorité à l'un d'un plus grand ensemble de variables qui sont communes à l'échantillon non probabiliste;

iii) Attribuer une seconde préférence à l'échantillon probabiliste ayant une plus grande taille d'échantillon;

(iv) Finalement, utiliser l'échantillon probabiliste pour lequel le mode de collecte des données est le même que celui pour l'échantillon non probabiliste.

Chen et coll. (2020) ont démontré que deux échantillons d'enquête probabilistes de référence ayant le même ensemble de variables auxiliaires communes ont tendance à produire des estimateurs PPI très similaires, mais que celui ayant la plus grande taille d'échantillon mène à de meilleurs estimateurs d'imputation massive.

8. Conclusions

Au début du 21e siècle, les enquêtes en ligne ont commencé à être populaires, ce qui a suscité un intérêt important pour la recherche sur ce sujet (Tourangeau, Conrad et Couper, 2013). Les enjeux et les défis posés par les échantillons d'enquête en ligne et autres enquêtes non probabilistes ont mené à la publication de « Summary Report of the AAPOR Task Force on Non-probability Sampling » par Baker, Brick, Bates, Bataglia, Couper, Dever, Gile et Tourangeau (2013). Entre autres, le rapport indiquait que i) contrairement à l'échantillonnage probabiliste, il n'y avait pas de cadre unique qui englobait adéquatement tous les échantillonnages non probabilistes; ii) faire des inférences pour toute enquête probabiliste ou non probabiliste nécessitait de s'appuyer sur des hypothèses de modélisation; iii) si les échantillons non probabilistes devaient être mieux acceptés chez les spécialistes de la recherche sur les enquêtes, il devait y avoir un cadre plus cohérent et des ensembles de mesures d'accompagnement pour évaluer leur qualité.

Les spécialistes de l'échantillonnage ont répondu à l'appel en menant des études accrues sur l'inférence statistique avec des échantillons d'enquête non probabilistes. La configuration actuelle de deux échantillons S_A et S_B, avec l'échantillon non probabiliste S_A ayant des mesures sur la variable étudiée y et les variables auxiliaires $\mathbf{x}$ et l'échantillon probabiliste S_B fournissant des renseignements sur $\mathbf{x}$, a été examiné en premier par Rivers (2007) sur un échantillon correspondant à l'imputation du plus proche voisin, laquelle est l'idée originale menant à la méthode de l'imputation massive (Kim et coll., 2021). La régression logistique pondérée reposant sur l'échantillon groupé pour estimer les scores de propension proposés par Valliant et Dever (2011) fut le premier essai sérieux sur le sujet, lequel sert de motivation pour la méthode du pseudo-maximum de vraisemblance élaborée par Chen et coll. (2020). Brick (2015) a examiné l'inférence du modèle compositionnel dans les mêmes conditions. Elliot et Valliant (2017) ont tenu des débats éclairés sur l'inférence des échantillons non probabilistes. Yang, Kim et Song (2020) ont traité des problèmes associés à des données de grande dimension en combinant des échantillons d'enquête probabilistes et non probabilistes.

L'inférence statistique avec des échantillons d'enquête non probabilistes fait partie du thème plus général sur la combinaison de données de multiples sources. Le terme « intégration des données » est fréquemment utilisé dans ce contexte. Combiner les renseignements d'échantillons d'enquête probabilistes indépendants a été longuement étudié dans les ouvrages publiés portant sur les enquêtes; voir, par exemple, Wu (2004), Kim et Rao (2012) et les références connexes. L'inférence avec des échantillons de l'enquête à base multiple représente un autre sujet qui a été étudié en profondeur par les statisticiens d'enquête; voir Lohr et Rao (2006) et Rao et Wu (2010a) et les références connexes. Dans son récent article sollicité du prix Waksberg, Lohr (2021) a offert un aperçu des enquêtes à cadres multiples et quelques analyses fascinantes de l'utilisation d'une structure à cadres multiples pour servir de principe organisateur pour d'autres méthodes de combinaison de données. Compte tenu des nouvelles sources de données émergentes et des modifications de la façon dont les sources des données traditionnelles comme les dossiers administratifs sont considérés, l'intégration des données est devenue un domaine très vaste qui nécessite la poursuite de travaux de recherche. D'autres analyses sont fournies par Lohr et Raghunathan

(2017) sur la combinaison de données d'enquête avec d'autres sources de données, ainsi que par Thompson (2019) sur la combinaison de sources nouvelles et traditionnelles dans les enquêtes sur les populations. Kim et Tam (2021), ainsi que Yang, Kim et Hwang (2021), ont abordé l'intégration des données en combinant des mégadonnées et des données d'échantillons d'enquête pour l'inférence de population finie. L'article de Yang et Kim (2020) contenait un examen de l'intégration de données statistiques dans un échantillonnage.

Un des messages essentiels que le présent article transmet a trait aux concepts de *validité* et d'*efficacité* dans l'analyse des échantillons d'une enquête non probabiliste. La validité fait référence à la convergence des estimateurs ponctuels, et l'efficacité est mesurée par la variance asymptotique de l'estimateur ponctuel. La validité est la principale préoccupation, et la quête de l'efficacité est un objectif secondaire quand d'autres approches valides sont disponibles. Les analyses sur la validité et l'efficacité nécessitent un cadre inférentiel approprié et la mise au point rigoureuse de procédures statistiques, ce qui constitue un autre message principal du présent article. Les échantillons non probabilistes ne conviennent pas au cadre inférentiel fondé sur un plan ou fondé sur un modèle pour les échantillons d'enquête probabilistes. Toutefois, des concepts statistiques standard et des procédures inférentielles peuvent être conçus dans un cadre approprié pour une inférence valide et efficace avec des échantillons d'enquête non probabilistes.

Les échantillons non probabilistes peuvent avoir une très grande taille. Une grande taille d'échantillon constitue une épée à double tranchant : quand les procédures inférentielles sont valides, une grande taille d'échantillon mène à une inférence plus efficace; quand les estimateurs sont biaisés, une grande taille d'échantillon rend le biais encore plus prononcé. Un échantillon d'enquête non probabiliste présentant une fraction de sondage de 80 % par rapport à la population ne procure pas nécessairement de meilleurs résultats pour l'estimation qu'un petit échantillon probabiliste (Meng, 2018).

Une grande taille d'échantillon produit également des échantillons non probabilistes associés aux problèmes modernes des mégadonnées. Le rôle des méthodes statistiques traditionnelles dans l'ère des mégadonnées a été plaidé de façon convaincante par Richard Lockhart (2018) : « Les ressources massives de calcul n'éliminent pas le besoin d'une modélisation soignée, d'une évaluation honnête de l'incertitude, ou de la conception d'un bon devis expérimental. Les idées classiques de statistique jouent un rôle crucial pour que l'analyse de données demeure honnête, efficace et efficiente » [traduction].

Jean-François Beaumont (2020) a soulevé la question suivante : « Les enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles ? ». La réponse courte est que les méthodes d'échantillonnage probabiliste et les échantillons d'enquête probabilistes continueront d'être un important outil de collecte des données pour de nombreux domaines, y compris les statistiques officielles, et l'inférence fondée sur un plan jouera un rôle crucial pour tout cadre inférentiel en évolution. La tendance actuelle qui consiste à utiliser des échantillons non probabilistes et des données d'autres sources se poursuivra. Une inférence statistique valide et efficace avec des échantillons non probabilistes nécessite une information auxiliaire de la population cible. Quelques enquêtes probabilistes nationales de haute

qualité comportant des variables d'enquête soigneusement conçues peuvent jouer un rôle central dans l'analyse des échantillons d'enquête non probabilistes.

Remerciements

La présente étude a reçu des subventions du Conseil de recherches en sciences naturelles et en génie du Canada et de l'Institut canadien des sciences statistiques. Une première version du document a été présentée à la rencontre annuelle de 2021 de la Société statistique du Canada (SSC) en tant qu'allocution spéciale de l'invité du président par le Groupe des méthodes d'enquête de la SSC. L'auteur remercie le rédacteur en chef de *Techniques d'enquête*, Jean-François Beaumont, pour l'invitation et pour l'organisation des discussions sur le sujet émergent de l'inférence statistique avec des échantillons d'enquête non probabilistes. Il y a également lieu de remercier les deux réviseurs anonymes qui ont offert des commentaires constructifs sur la première ébauche, ce qui a permis d'améliorer l'article.

Bibliographie

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. et Tourangeau, R. (2013). Report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1, 90-143.

Beaumont, J.-F. (2020). Les enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles ? *Techniques d'enquête*, 46, 1, 1-30. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2020001/article/00001-fra.pdf.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Breiman, L., Friedman, J.H., Olshen, R.A. et Stone, C.J. (1984). *Classification and Regression Trees*, second edition. Wadsworth & Brooks/Cole Advanced Books & Software.

Brick, J.M. (2015). Compositional model inference. Dans Proceedings of the Survey Research Methods Section, Joint Statistical Meetings, American Statistical Association, Alexandrie, Virginie, 299-307.

Chen, J., et Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*,16, 113-131.

Chen, J., et Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association*, 96, 260-269.

Chen, J., et Sitter, R.R. (1999). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.

Chen, Y. (2020). *Statistical Analysis with Non-probability Survey Samples*, Thèse de doctorat, Department of Statistics and Actuarial Science, University of Waterloo.

Chen, Y., Li, P. et Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Chen, Y., Li, P., Rao, J.N.K. et Wu, C. (2022). Pseudo empirical likelihood inference for non-probability survey samples. *The Canadian Journal of Statistics*, accepté.

Chu, K.C.K., et Beaumont, J.-F. (2019). The use of classification trees to reduce selection bias for a non-probability sample with help from a probability sample. Proceedings of the Survey Methods Section of SSC.

Elliott, M., et Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32, 249-264.

Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 31, 1208-1212.

Godambe, V.P., et Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *Revue Internationale de Statistique*, 54, 127-138.

Kim, J.K., et Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, 24, 375-394.

Kim, J.K., et Rao, J.N.K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika*, 99, 85-100.

Kim, J.K., et Tam, S. (2021). Data integration by combining big data and survey sample data for finite population inference. *Revue Internationale de Statistique*, 89, 382-401.

Kim, J.K., Park, S., Chen, Y. et Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society, Series A*, 184, 941-963.

Liu, Z., et Valliant, R. (2021). Investigating an alternative for estimation from a nonprobability sample: Matching plus calibration. arXiv:2112.00855v1 [stat.ME]. Déc. 2021.

Lockhart, R. (2018). Special issue on big data and the statistical sciences: Guest editor's introduction. *The Canadian Journal of Statistics*, 46, 4-9.

Lohr, S.L. (2021). Les enquêtes à bases de sondage multiples pour un monde fait de sources de données multiples. *Techniques d'enquête*, 47, 2, 247-285. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021002/article/00008-fra.pdf.

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

Lohr, S.L., et Rao, J.N.K. (2006). Estimation in multiple frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.

McCullagh, P., et Nelder, J.A. (1989). *Generalized Linear Models*, second edition, New York: Chapman and Hall.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics*, 12, 685-726.

Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9, 141-142.

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, 558-606.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*, 83, 242-272.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation*, second Edition. Hoboken, NJ: Wiley.

Rao, J.N.K., et Wu, C. (2010a). Pseudo empirical likelihood inference for multiple frame surveys. *Journal of the American Statistical Association*, 105, 1494-1503.

Rao, J.N.K., et Wu, C. (2010b). Bayesian pseudo empirical likelihood intervals for complex surveys. *Journal of the Royal Statistical Society, Series B*, 72, 533-544.

Rivers, D. (2007). Sampling for web surveys. Dans *Proceedings of the Survey Research Methods Section*, Joint Statistical Meetings, American Statistical Association, Alexandrie, Virginie, 1-26.

Robins, J.M., Rotnitzky, A. et Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846-866

Rosenbaum, P.R., et Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Tourangeau, R., Conrad, F.G. et Couper, M.P. (2013). *The Science of Web Surveys*, first edition. Oxford: Oxford University Press.

Thompson, M.E. (1997). *Theory of Sample Surveys*. Londres : Chapman & Hall.

Thompson, M.E. (2019). Combining data from new and traditional sources in population surveys. *Revue Internationale de Statistique*, 87, S79-89.

Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

Valliant, R., et Dever, J.A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40, 105-137.

Wang, L., Graubard, B.I., Katki, H.A. et Li, Y. (2020). Improving external validity of epidemiologic cohort analysis: A kernel weighting approach. *Journal of the Royal Statistical Society*, *Series A*, 183, 1293-1311.

Wang, L., Valliant, R. et Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40, 5237-5250.

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā A*, 26, 359-372.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1-25.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *The Canadian Journal of Statistics*, 32, 15-26.

Wu, C., et Rao, J.N.K. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics*, 34, 359-375.

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, 185-193.

Wu, C., et Thompson, M.E. (2020). *Sampling Theory and Practice*. Springer, Cham.

Yang, S., et Kim, J.K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3, 625-650.

Yang, S., Kim, J.K. et Hwang, Y. (2021). Intégration de données d'enquêtes probabilistes et de mégadonnées aux fins d'inférence de population finie au moyen d'une imputation massive. *Techniques d'enquête*, 47, 1, 33-64. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2021001/article/00004-fra.pdf.

Yang, S., Kim, J.K. et Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society, Series B*, 82, 445-465.

Yuan, M., Li, P. et Wu, C. (2022). Nonparametric estimation of propensity scores for non-probability survey samples. Document de travail.

Zhao, P., et Wu, C. (2019). Some theoretical and practical aspects of empirical likelihood methods for complex surveys. *Revue Internationale de Statistique*, 87, S239-256.

Zhao, P., Rao, J.N.K. et Wu, C. (2020a). Empirical likelihood methods for public-use survey data. *Electronic Journal of Statistics*, 14, 2484-2509.

Zhao, P., Ghosh, M., Rao, J.N.K. et Wu, C. (2020b). Bayesian empirical likelihood inference with complex survey data. *Journal of the Royal Statistical Society, Series B*, 82, 155-174.

Techniques d'enquête, décembre 2022
Vol. 48, nº 2, pp. 339-345
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article « Inférence statistique avec des échantillons d'enquête non probabiliste » : Échantillons non probabilistes : évaluation et voie à suivre

Michael A. Bailey[1]

Résumé

Les enquêtes non probabilistes jouent un rôle croissant dans la recherche par enquête. L'étude de Wu rassemble de façon compétente les nombreux outils disponibles lorsqu'on suppose que la non-réponse est conditionnellement indépendante de la variable étudiée. Dans le présent exposé, j'étudie la façon d'intégrer les idées de Wu dans un cadre plus large qui englobe le cas dans lequel la non-réponse dépend de la variable étudiée, un cas qui est particulièrement dangereux dans les sondages non probabilistes.

Mots-clés : Échantillonnage d'enquête; sondages non probabilistes.

1. Introduction

Les enquêtes sont en train de subir des changements importants. Nous sommes loin de l'époque où une enquête par composition aléatoire de numéros de téléphone pouvait produire de manière fiable des échantillons représentatifs. À l'heure actuelle, peu de gens répondent aux appels téléphoniques ou même aux courriels. Les enquêteurs ont réagi en proposant une myriade de nouvelles façons ingénieuses de générer des réponses aux enquêtes dans un environnement peu accueillant.

L'innovation la plus répandue est sans aucun doute l'utilisation d'échantillons non probabilistes, souvent par Internet. Bien que la mise en œuvre varie, l'approche consiste habituellement à recueillir les coordonnées d'un grand nombre de personnes qui sont disposées à répondre, puis à sélectionner un sous-ensemble de ce bassin pour une enquête donnée. Ces enquêtes se sont révélées rentables et ont souvent, quoique peut-être pas toujours, produit des résultats utilisables.

Cela dit, ces enquêtes sont-elles crédibles ? La plupart des enquêtes n'ont pas de « vérité fondamentale » qui pourrait être utilisée pour évaluer les résultats; l'absence de renseignement de ce genre est après tout la raison qui a poussé quelqu'un à mener l'enquête. Les échantillons probabilistes permettent de surmonter ce problème en s'appuyant sur la théorie, car les propriétés de ce genre d'enquête sont bien comprises. Pour les échantillons non probabilistes, cependant, la pratique a largement dépassé la théorie, ce qui signifie que les raisons de se fier aux résultats sont plutôt spéculatives.

L'étude de Wu est donc une contribution utile à notre compréhension des enquêtes non probabilistes. Wu se concentre sur la classe d'estimateurs qui supposent une non-réponse ignorable; il les met en contexte les uns par rapport aux autres et il détermine des pistes pour les études à venir.

Wu soulève un point important en mentionnant qu'« il devait y avoir un cadre plus cohérent et des ensembles de mesures d'accompagnement pour évaluer leur qualité » (page 332). Je suis tout à fait

1. Michael A. Bailey, Georgetown University. Courriel : baileyma@georgetown.edu.

d'accord. Dans le présent exposé, j'approfondis ce point de trois façons. À la section 2, j'étudie la façon de le faire dans le cadre des travaux sur lesquels portent ses études. À la section 3, je cherche à élargir la portée d'un tel cadre, en soulignant que les conséquences des violations des principales hypothèses sont tellement plus graves dans un contexte non probabiliste que nous devrions construire notre cadre qui comprend les violations des principales hypothèses des données manquantes au hasard (DMH). À la section 4, j'étudie ensuite ce que nous pouvons faire, si jamais quoi que ce soit peut être fait. Enfin, à la section 5, je présente quelques observations finales.

2. Enquêtes non probabilistes lorsque les données sont manquantes au hasard

Wu fonde son analyse sur une présentation claire des quatre hypothèses sous-jacentes aux modèles qu'il examine. L'hypothèse la plus importante est que les données sont manquantes au hasard, ce qui signifie que pour un ensemble donné de covariables, la variable étudiée est indépendante de la décision de répondre. (Bien que la nomenclature soit normalisée dans la littérature, je ne peux pas m'empêcher d'exprimer mon malaise à l'égard du terme « données manquantes au hasard ». Bien sûr, les données sont manquantes au hasard, ce qui est vrai même pour son « contraire », tout aussi mal nommé, « données non manquantes au hasard ». Je rêve d'un jour où la nomenclature correspondra à la définition, peut-être en remplaçant DMH par le terme « indépendance conditionnelle », qui serait un meilleur nom. Cependant, je reconnais à quel point il est difficile de changer les termes acceptés que les gens utilisent.)

Compte tenu de ces hypothèses, Wu divise les approches selon celles basées sur des modèles, celles basées sur la pondération par l'inverse du score de propension (PISP) et celles en modèles doublement robustes. Dans les approches basées sur les modèles, nous voyons la gamme d'efforts à imputer à partir de l'échantillon observé, y compris l'imputation de masse qui, au sens large, comprend des approches souples d'appariement des échantillons qui nous permettent de représenter une plus grande population en fonction des points de données observés qui sont « proches », un mot ayant diverses définitions. La PISP s'appuie sur les mêmes hypothèses. Les estimateurs doublement robustes ont tendance à être plus récents et attrayants en raison de leur capacité à offrir aux spécialistes deux chances de parvenir à des hypothèses correctes. Wu documente habilement les maux de tête que ces modèles provoquent lorsqu'on les utilise pour réaliser une inférence.

Bien que Wu ait montré les différences dans ces approches, il est utile de comprendre qu'après tout, il n'a fait que « pêcher dans un seul coin de la mare ». Tous les modèles utilisent des renseignements semblables de façons semblables : ils supposent tous que les données sont manquantes au hasard et ils fournissent des outils pour modéliser ou imputer le comportement de personnes non observées sous forme d'extrapolations directes tirées des données observées. Si les titulaires d'un diplôme collégial diffèrent des autres diplômés et que nous avons trop de titulaires d'un diplôme collégial, toutes les approches fondées sur l'hypothèse des DMH consisteront à extrapoler à la population générale directement à partir des

données de l'échantillon sur deux groupes au prorata de la présence de ces groupes dans la population cible.

Mon intuition me dit que les modèles envisagés par Wu sont à peu près aussi utiles, et aussi plus ou moins vulnérables aux violations de l'hypothèse des DMH. Ou y a-t-il des contextes dans lesquels nous nous attendons à ce que les différences entre les méthodes soient considérables ? Il n'est pas facile de répondre à cette question, certes, mais je serais fasciné d'apprendre le point de vue de Wu sur les situations où la principale « action » se trouve dans les échantillons non probabilistes et quels modèles il considère les mieux adaptés pour tenir compte de tels problèmes.

Une des approches possibles serait d'étudier la flexibilité entre les modèles. À ce stade, mon intuition me dit que même si, en théorie, ces différences peuvent être substantielles, en pratique, ces différences sont relativement modestes. Cela est particulièrement vrai si un spécialiste expérimenté possédant des connaissances dans le domaine définit un modèle paramétrique avec doigté, y compris les bonnes interactions et ainsi de suite.

3. Enquêtes non probabilistes lorsque les données ne sont pas manquantes au hasard

Nous devrions prendre très au sérieux l'appel de Wu en faveur d'un cadre plus cohérent pour analyser les échantillons non probabilistes. Et nous devrions viser bien haut, parce qu'un paradigme pour les échantillons non probabilistes est, essentiellement, un paradigme pour l'ensemble du champ de recherche étant donné l'importance et la trajectoire des échantillons non probabilistes.

Alors que nous réfléchissons à l'élaboration d'un cadre de sondage, il est utile de se rappeler le célèbre aphorisme de George Box : « Puisque tous les modèles sont faux, le scientifique doit être vigilant face à ce qu'il y a de plus faux. Il ne convient pas de s'inquiéter des souris quand les tigres sont à l'affût » (Box, 1976). Le tigre dans les échantillons non probabilistes ne rôde pas entre l'échantillonnage par quotas et les modèles de PISP. En fait, le tigre se trouve sans nul doute dans l'hypothèse des DMH. La violation de cette hypothèse constitue la faiblesse emblématique des hypothèses de DMH, et tout cadre pour les enquêtes non probabilistes devrait donc commencer là.

Le problème réside dans le fait que, malgré les violations des hypothèses de DMH qui constituent un problème dans l'échantillonnage probabiliste (découlant de la non-réponse chez les personnes avec lesquelles on communique au hasard), les violations des hypothèses de DMH sont plus graves dans un monde non probabiliste. Meng (2018) précise cette idée et définit l'erreur glissée dans une enquête :

$$\bar{Y}_n - \bar{Y}_N = \underbrace{\rho_{R,Y}}_{\text{qualité des données}} \underbrace{\sqrt{\frac{N-n}{n}}}_{\text{quantité des données}} \underbrace{\sigma_Y}_{\text{difficulté des données}}. \tag{3.1}$$

Le premier segment de l'équation est $\rho_{R,Y}$, soit la corrélation dans la population entre R et Y. Cette quantité peut être utilisée pour refléter la qualité de données relatives à l'échantillonnage. Le deuxième segment dans l'équation de Meng, $\sqrt{N-n/n}$, se rapporte à la taille de la population (N majuscule) et à la taille de l'échantillon (n minuscule). Le troisième segment de l'équation de Meng est σ_Y, l'écart type de Y.

Quand $\rho_{R,Y} \neq 0$, la moyenne échantillonnée sera non nulle sauf si $n = N$ (c'est-à-dire que l'échantillon représente la population tout entière) ou si $\sigma_Y = 0$ (c'est-à-dire que la valeur de Y est pareille pour tous les membres de la population). Dans un cas comme dans l'autre, aucun ne constitue un contexte de sondage intéressant.

Il s'agit d'une identité, par conséquent, même lorsque la valeur prévue de $\rho_{R,Y} = 0$, il en résulte un certain nombre d'erreurs (comme dans le cas de l'échantillonnage aléatoire). Cependant, lorsque nous passons à l'échantillonnage non aléatoire, nous pouvons nous attendre à ce que la corrélation réalisée de R et Y s'élargisse. Plus la valeur de $\rho_{R,Y}$ est grande, plus l'erreur d'échantillonnage est grande, et son ampleur exacte interagira avec les autres segments.

L'implication la plus explosive de l'équation de Meng découle de l'interaction des deux premiers segments. Quand des données manquantes non au hasard (DMNH) (c'est-à-dire qu'il y aura une raison précise de s'attendre à ce que $\rho_{R,Y} \neq 0$ parce que R dépend de Y), l'erreur réelle dépend de la population totale. Ce résultat est choquant en ce qui concerne le caractère délicat des sondages dans le monde actuel, mais il est essentiel d'en tenir compte dans le contexte de l'échantillonnage non aléatoire.

Nous pouvons construire un monde simple constitué de deux pays pour expliquer en détail le fonctionnement. Supposons que les taux d'infection à la COVID-19 représentent notre variable étudiée et, pour notre exemple, que les taux d'infection sont les mêmes dans les deux pays. Le premier pays est immense (comme la Chine) et l'autre est petit (comme le Luxembourg). Si nous échantillonnions *au hasard* 1 000 personnes dans chaque pays, nous pourrions produire des estimations ayant la même précision pour chaque pays, malgré leurs énormes différences sur le plan de la population.

Que se passe-t-il si nous avons affaire à un échantillon *non aléatoire* de 1 000 personnes dans chaque pays ? Supposons, par souci de simplicité, que l'empressement des gens à se faire tester ne soit qu'une fonction de leurs symptômes et que les personnes présentant plus de symptômes soient plus susceptibles d'avoir la COVID-19. Cela crée un échantillonnage de DMNH parce que le choix de faire partie de l'échantillon sera associé à des valeurs attendues plus élevées pour notre variable étudiée.

En Chine, nous obtiendrons les 1 000 personnes les plus malades. Ces personnes seront vraiment malades, car elles feront partie du 0,00001^{e} centile supérieur ou quelque chose du genre. Au Luxembourg, nous aurons aussi les 1 000 personnes les plus malades, mais ces personnes n'ont pas à être aussi malades pour faire partie de cet ensemble par rapport à un ensemble d'un pays beaucoup plus grand. Cela signifie que les 1 000 personnes les plus malades au Luxembourg seront dans environ le 0,2^{e} centile supérieur; elles seront encore très malades par rapport à la population, mais elles ne seront pas aussi biaisées qu'en

Chine. En bref, les DMNH produiront une erreur proportionnelle à la taille de la population pour une taille d'échantillon donnée.

(Il convient de noter que les échantillons véritablement aléatoires sont extrêmement rares, étant donné la non-réponse parmi les gens avec lesquels on communique au hasard. La pratique réelle des échantillons probabilistes peut être décrite comme un contact aléatoire, défini comme des enquêtes dans le cadre desquelles on communique avec les gens au hasard, même si la réponse des personnes avec lesquelles on communique peut être non aléatoire. Les enquêtes par communication aléatoire peuvent contrevenir à l'hypothèse des DMH, mais ont néanmoins de fortes vertus. Bradley, Kuriwaki, Isakov, Sejdinovic, Meng et Flaxman [2021] et Bailey [2023] montrent la façon dont l'erreur d'enquête dans les enquêtes par communication aléatoire est proportionnelle au taux de réponse plutôt qu'à la taille de la population.)

Les violations de l'hypothèse des DMH dans l'échantillonnage non probabiliste entraînent des erreurs qui sont proportionnelles à la taille de la population. Pour utiliser la métaphore de Box, c'est là que se trouvent les tigres. C'est pourquoi, alors que nous poursuivons l'exhortation de Wu concernant une plus grande cohérence dans la façon dont nous évaluons les nouvelles formes de sondages, nous devrions nous entendre pour adopter un cadre qui comprend la possibilité de violations de l'hypothèse des DMH plutôt qu'un cadre qui prétend que ce problème n'existe pas.

4. Que faire au sujet des violations de l'hypothèse des données manquantes au hasard ?

Wu suit une grande partie de la littérature en se détournant des modèles des DMNH. Cela est en partie en raison de la perception selon laquelle la non-réponse des DMNH est essentiellement insoluble. Par exemple, Wu fait remarquer de façon quelque peu pessimiste qu'« il est bien compris que l'hypothèse des données manquantes au hasard ne peut être testée à l'aide des données elles-mêmes de l'échantillon » (page 328) et que « la nature biaisée des échantillons non probabilistes ne peut être corrigée au moyen de l'échantillon lui-même » (page 308).

Pour ce qui est des conseils aux spécialistes de la recherche sur les enquêtes qui s'inquiètent des violations de l'hypothèse des DMH, Wu n'offre qu'un test modeste, qui consiste essentiellement à trouver une autre variable semblable à la variable étudiée, mais qui est disponible pour l'ensemble de la population. Si seulement c'était si facile! Des générations d'enquêteurs ont passé les données au peigne fin pour trouver de telles variables et demeurent néanmoins préoccupées par les DMNH, surtout lorsque la réponse est non aléatoire.

Le cadre de Wu sous-estime ce que nous pouvons faire à propos des violations de l'hypothèse des DMH. Ces efforts exigeront des hypothèses, certes, mais au moins nous pourrons assouplir l'hypothèse intransigeante imposée par le concept des DMH. Le lien avec les points précédents est essentiel : comme nous aurons besoin d'hypothèses, il est important que nous disposions d'un cadre pour réfléchir aux hypothèses les plus importantes afin de pouvoir concentrer nos efforts de façon appropriée. L'équation de

Meng montre de quelle façon les violations de l'hypothèse des DMH jouent un rôle central dans la création d'erreurs dans un monde d'échantillonnage non probabiliste et, par conséquent, nous devrions faire tout ce qui est en notre pouvoir pour régler ce problème.

Le modèle de sélection de Heckman (1979) est un exemple bien connu d'un modèle qui peut s'attaquer à l'hypothèse des DMNH. Ce modèle permet de tenir compte des violations de l'hypothèse des DMH et d'estimer même leur ampleur. Il n'est pas sans problème, bien sûr. En pratique, une restriction d'exclusion est exigée (une hypothèse selon laquelle une ou plusieurs variables ont une incidence sur la réponse, mais pas sur la variable étudiée) et de nombreux spécialistes modernes sont naturellement prudents à l'égard de la forte hypothèse paramétrique du modèle de Heckman.

Les spécialistes ont réalisé des progrès considérables au-delà du modèle de Heckman pour traiter les violations de l'hypothèse des DMH (Bailey, 2023). Il est facile d'assouplir l'hypothèse paramétrique grâce aux fonctions de copule (Gomes, Radice, Brenes et Marra, 2019). Si nous voulons étudier les déterminants de Y, il existe une documentation abondante et grandissante qui applique des fonctions de contrôle très souples aux contextes des DMNH (Das, Newey et Vella, 2003; Liu et Yu, 2022). Et si nous pouvions déterminer les variables qui ont une incidence sur la propension à réagir, mais pas sur le résultat d'intérêt, plusieurs méthodes permettraient de modéliser et de neutraliser l'échantillonnage des DMNH (Peress, 2010; Sun, Liu, Miao, Wirth, Robins et Tchetgen-Tchetgen, 2018).

5. Conclusion

Dans son étude, Wu résume de manière compétente et utile l'état de la littérature en ce qui concerne l'analyse des données d'enquête non probabilistes selon l'hypothèse des DMH. Il souligne également qu'il est grandement nécessaire que le champ de recherche se rallie autour d'un cadre plus cohérent pour évaluer ces innovations et d'autres innovations en matière de sondage.

Dans la présente étude, je me base sur le travail de Wu pour proposer un cadre qui comprend non seulement les modèles des DMH qu'il a analysés, mais aussi les modèles des DMNH, car la violation de l'hypothèse des DMH est particulièrement pertinente et nuisible aux enquêtes non probabilistes.

Bibliographie

Bailey, M.A. (2023). *Polling at a Crossroads: Rethinking Modern Survey Research*, Cambridge University Press – sous contrat.

Box, G. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791-799.

Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L. et Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated us vaccine uptake. *Nature*, 600, 695-700.

Das, M., Newey, W.K. et Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, 70, 33-58.

Gomes, M., Radice, R., Brenes, J.C. et Marra, G. (2019). Copula selection models for nongaussian outcomes that are missing not at random. *Statistics in Medicine*, 38, 480-496.

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153-162.

Liu, R., et Yu, Z. (2022). Sample selection models with monotone control functions. *Journal of Econometrics*, 226, 321-342.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (1): Law of large populations, big data paradox, and the 2016 presidential election. *The Annals of Applied Statistics*, 12, 685-726.

Peress, M. (2010). Correcting for survey nonresponse using variable response propensity. *Journal of the American Statistical Association*, 105, 1418-1430.

Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J. et Tchetgen-Tchetgen, E.J. (2018). Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*, 28, 1965-1983.

Techniques d'enquête, décembre 2022
Vol. 48, nº 2, pp. 347-358
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article « Inférence statistique avec des échantillons d'enquête non probabiliste »

Michael R. Elliott[1]

Résumé

Cet exposé vise à approfondir l'examen de Wu sur l'inférence à partir d'échantillons non probabilistes, ainsi qu'à mettre en évidence les aspects qui constituent probablement d'autres pistes de recherche utiles. Elle se termine par un appel en faveur d'un registre organisé d'enquêtes probabilistes de grande qualité qui visera à fournir des renseignements utiles à l'ajustement d'enquêtes non probabilistes.

Mots-clés : Analyse de sensibilité; estimation doublement robuste; pseudo-pondération; score de propension.

1. Introduction

Je remercie M. Changbao Wu de son excellent examen des travaux précédents et des questions ouvertes pour l'inférence statistique à partir d'échantillons non probabilistes. Étant donné l'ampleur et l'évolution rapide des travaux dans ce domaine, il est clair que M. Wu n'a pas été en mesure d'être exhaustif. Ma propre analyse comporte elle aussi des limites, mais je vais aborder quelques approches supplémentaires qui se rapportent aux sujets examinés par mon confrère. Je me pencherai également sur la question de la modélisation par rapport à la pondération pour différentes cibles inférentielles, et me servirai de l'analyse et des conclusions de M. Wu pour souligner l'importance cruciale des échantillons probabilistes – notamment les études de grande qualité qui portent sur l'estimation de covariables pertinentes – dans le but d'améliorer l'inférence pour la profusion des échantillons non probabilistes utilisés pour remplacer les échantillons probabilistes traditionnels dans un grand nombre d'études et de contextes statistiques officiels. Pour éviter toute confusion de notation, toutes les notations suivront celles de Wu, sauf lorsque de nouvelles notations seront nécessaires.

La section 2 examine d'autres approches pour combiner les données tirées d'enquêtes probabilistes et celles d'enquêtes non probabilistes. La section 3 examine brièvement la question de la pondération par rapport à la modélisation au moment de rajuster des données d'enquête non probabilistes. La section 4 examine certaines découvertes récentes liées aux analyses de sensibilité des hypothèses types applicables à l'ajustement des données d'enquêtes non probabilistes à l'aide des données d'enquêtes probabilistes. Enfin, la section 5 se termine par un appel à concevoir systématiquement un ensemble d'enquêtes probabilistes dans le but explicite de rajuster les enquêtes non probabilistes.

2. Approches supplémentaires pour combiner les données d'enquêtes probabilistes et les données d'enquêtes non probabilistes

L'article de Wu suit le principe général fondé sur les trois points suivants : 1) appliquer une estimation de modèle suivie d'un calibrage des estimations des distributions de covariables d'échantillons

1. Michael R. Elliott, Département de biostatistique, University of Michigan, M4124 SPH II, 1415 Washington Heights, Ann Arbor, MI 48109. Courriel : mrelliot@umich.edu.

probabilistes; 2) élaborer des estimations de scores de propension selon les écarts entre les données d'échantillons probabilistes et celles d'échantillons non probabilistes; 3) trouver des méthodes doublement robustes qui combinent les points 1 et 2 pour faire en sorte qu'un seul des deux modèles sous-jacents soit exact.

2.1 Estimateurs de score de propension

Rivers (2007) semble avoir été le premier à suggérer d'estimer le score de propension en utilisant la régression logistique avec l'appartenance à l'échantillon non probabiliste comme résultat et en utilisant comme poids d'inclusion la réciproque des scores de propension résultants. Cette approche a été systématisée davantage dans Valliant et Dever (2011). Séparément, en utilisant des résultats simples du théorème de Bayes et de l'analyse discriminante décrite pour la première fois par Elliott et Davis (2005), Elliott, Resler, Flannagan et Rupp (2010) et Elliott (2013) ont élaboré un estimateur quelque peu différent prenant la forme

$$\hat{\pi}_i^A\left(\mathbf{x}_i, \boldsymbol{\alpha}\right) \;=\; \hat{P}\left(i \in S_A\right) \propto P\left(i \in S_B\right) \frac{\hat{P}\left(i \in S_A \,\middle|\, i \in S_A \text{ ou } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha}\right)}{\hat{P}\left(i \in S_B \,\middle|\, i \in S_A \text{ ou } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha}\right)}. \tag{2.1}$$

Le segment $\hat{P}\left(i \in S_A \,\middle|\, i \in S_A \text{ ou } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha}\right)$ peut être obtenu par une régression logistique, ou au moyen d'une série d'approches qui reposent sur l'apprentissage automatique, comme les machines à vecteurs de support (Soentpiet, 1999), l'estimation pondérée du maximum de vraisemblance (Van Der Laan et Rubin, 2006) ou les arbres de régression additifs bayésiens (Chipman, George et McCulloch, 2010), et où le segment $\hat{P}\left(i \in S_A \,\middle|\, i \in S_B \text{ ou } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha}\right)$ est obtenu en tant que $1 - \hat{P}\left(i \in S_A \,\middle|\, i \in S_A \text{ ou } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha}\right)$. En principe, le segment $P\left(i \in S_B\right) = 1/d_i^B$ est connu, puisque les probabilités d'échantillonnage sont connues pour tous les éléments de la population, y compris ceux de l'échantillon non probabiliste, mais dans la pratique, les analystes qui n'ont accès qu'aux données à grande diffusion pourraient aussi avoir besoin d'estimer ce paramètre. (En outre, d_i^B peut inclure des ajustements de calage et de non-réponse qui ne sont pas connus pour les éléments de l'échantillon non probabiliste.) Ce dernier point est essentiel, car l'utilisation de l'échantillon probabiliste pour établir des scores de propension en utilisant uniquement les écarts entre l'échantillon non probabiliste et l'échantillon probabiliste sera faussée à moins que l'échantillon probabiliste ne repose sur un plan à probabilités égales *epsem* pour *equal probability selection methods*, comme l'a souligné Wu.

Par contre, Chen, Li et Wu (2020) montrent que l'utilisation d'une approche de pseudo-vraisemblance pour estimer $\hat{\pi}_i^A\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$ directement à partir de la vraisemblance de population pour les indicateurs $I\left(i \in S_A\right)$ en fonction de $\mathbf{x}_i$ produit un estimateur qui ne requiert pas le segment $P\left(i \in S_B\right)$ pour les éléments de l'échantillon non probabiliste, à condition que $\pi_i^A\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$ suive un modèle linéaire généralisé avec un lien canonique, c'est-à-dire une régression logistique.

(Aucune de ces approches ne fournit la bonne ordonnée permettant d'obtenir un score de propension réel. Cependant, comme l'indique l'article de Wu, une estimation pondérée repose généralement sur des estimateurs de type Hájek [qui utilisent des poids pour estimer un total de population pour les

dénominateurs; Hájek, 1971], de sorte que des scores de propension estimés jusqu'à l'obtention d'une constante de normalisation sont suffisants.)

2.2 Estimateurs doublement robustes

Si l'inférence est centrée sur une variable particulière Y offerte seulement dans l'échantillon non probabiliste, on peut alors revenir aux estimateurs assistés par un modèle qui remontent à Cassel, Särndal et Wretman (1976) et qui posent comme postulat un modèle d'espérance $E\left(y_i \mid \mathbf{x}_i\right) = m_i$. La combinaison de ce modèle avec les estimations du score de propension de la probabilité de faire partie de l'échantillon non probabiliste (que nous traiterons comme un « échantillon probabiliste inconnu » et que nous examinerons plus en détail dans la section sur les hypothèses) permet d'obtenir les estimateurs de la formule

$$\frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{\hat{N}^B}\sum_{i\in S_B} d_i^B \hat{m}_i \tag{2.2}$$

correspondant au segment $\hat{\mu}_{\mathrm{DR}2}$ de la formule (4.11) indiquée dans l'article de Wu. On pourrait penser que tout biais découlant d'une erreur de spécification du modèle dans l'estimation de m_i dans $\frac{1}{\hat{N}^B}\sum_{i\in S_B} d_i^B \hat{m}_i$ sera égal et de signe opposé à $\frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A}$ si le modèle pour π_i^A est correctement spécifié. À l'inverse, si le modèle pour π_i^A est mal spécifié alors que m_i est correctement spécifié, $y_i - \hat{m}_i$ sera iid avec une moyenne nulle et, par conséquent, $\frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A}$ aura aussi une moyenne de zéro, ce qui donnera un estimateur sans biais. Chen, Valliant et Elliott (2019) ont utilisé la méthode LASSO pour les prévisions, de pair avec des estimateurs par la régression généralisée (McConville, Breidt, Lee et Moisen, 2017) lorsque $\mathbf{X}$ est de grande dimension. Comme Wu le fait remarquer, Wu et Sitter (2001) montrent l'équivalence entre l'estimateur par la régression généralisée appliqué aux valeurs prédites et les estimateurs doublement robustes de la formule dans la section (2.2), qui indique que l'approche adoptée par Chen et coll. (2019) est équivalente à la formule de la section 2.2 avec estimation de la méthode LASSO pour m_i et une hypothèse d'échantillonnage aléatoire simple pour l'échantillon non probabiliste.

Un inconvénient d'utiliser la formule (2.1) par rapport à la méthode de Chen et coll. (2020) comme estimateur de π_i^A, et donc de d_i^A, tient au fait qu'on doive connaître les poids de l'échantillon probabiliste d_i^B, ou qu'il faille au moins les estimer pour l'échantillon non probabiliste. L'un des avantages de la formule (2.1) est que les modèles non linéaires et les méthodes d'apprentissage automatique peuvent être utilisés dans l'estimation. Rafei, Flannagan et Elliott (2020) utilisent les arbres de régression additifs bayésiens pour estimer m_i et π_i^A, ce qui permet de réduire l'incidence d'une erreur de spécification du modèle. Les simulations ont montré une amélioration considérable de la réduction du biais et de la variance par rapport à la méthode de Chen et coll. (2020) en cas de spécification erronée des modèles linéaires. On peut effectuer l'estimation de la variance en adaptant les règles d'imputation multiple de Rubin : à partir de M tirages indépendants selon les arbres de régression additifs bayésiens, la moyenne des variances calculées traitant le tirage de d_i^A comme connu reposant sur des estimateurs classiques d'un

plan d'échantillonnage complexe et ajouté à $\frac{M+1}{M}$ fois la variance des estimations ponctuelles calculées pour les tirages de d_i^A donne un estimateur de variance approximativement sans biais.

Une autre approche de l'estimation doublement robuste consiste à utiliser le score de propension comme « score d'équilibrage » le plus grossier possible qui contient toute l'information sur l'association entre l'indicateur d'échantillonnage et le résultat visé. Cela a mené à la mise au point d'estimateurs de la moyenne qui utilisent des fonctions lisses de pondération pour produire des estimateurs convergents susceptibles d'être plus efficaces lorsque les poids sont très variables ou faiblement liés au résultat (Elliott et Little, 2000; Zheng et Little, 2005). Zhou, Elliott et Little (2019) ont étendu cette idée à l'inférence causale dans les études non randomisées, où la probabilité d'attribution à un traitement ou à une exposition (score de propension) est estimée comme une fonction des covariables $P_Z\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$ utilisant la régression logistique, puis les résultats possibles non observés Y^z au volet de traitement $z_i' \neq z_i$ pour le traitement observé z_i sont imputés à partir de

$$Y_i^Z \sim N\left(s\left(\hat{P}_Z^*\left(\mathbf{x}_i, \hat{\boldsymbol{\alpha}}\right) \middle| \boldsymbol{\theta}_Z\right)\right) + g_Z\left(\hat{P}^*\left(\mathbf{x}_i, \hat{\boldsymbol{\alpha}}\right), \mathbf{x}_i \middle| \boldsymbol{\beta}_Z\right), \sigma^2 \tag{2.3}$$

où P^* est la transformation logit de P, $s(\hat{P}_Z^* \,|\, \boldsymbol{\theta}_Z)$ désigne une spline pénalisée avec nœuds fixes de propension (Eilers et Marx, 1996), et $g_Z(\hat{P}^*, \mathbf{x}_i \,|\, \boldsymbol{\beta}_Z)$ représente une fonction générale des covariables, y compris les scores de propension. L'estimateur qui en résulte est doublement robuste en ce sens que si $P_Z\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$ ou $E(Y^z) = g_Z(\hat{P}^*, \mathbf{x}_i \,|\, \boldsymbol{\beta}_Z)$ est correctement spécifié, $Y^{(z)}$ sera approximativement sans biais (voir Zhang et Little, 2009). Il peut être mis en œuvre dans l'échantillon non probabiliste en remplaçant $\hat{P}_Z\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$ dans le modèle de moyenne pour (2.3) par la valeur $\hat{\pi}_i^A$ estimée à l'aide de (2.1) pour obtenir un tirage de $Y_i^{(b)}$. (Veuillez noter que cela nécessite l'obtention de $\hat{\pi}_i^A$ pour les éléments de l'échantillon probabiliste nécessitant une prédiction.) L'inférence peut se faire en obtenant des tirages $b = 1, \ldots, B$ à partir de la distribution postérieure de la quantité estimée de population étudiée, par exemple pour la moyenne de la population

$$Y^{(b)} = \frac{\sum_{i \in S_R} N_i^{(b)} Y_i^{(b)} + \sum_{i \in S_A} \left(y_i - Y_i^{(b)}\right)}{N}$$

où $N_i^{(b)}$ est maintenant une estimation de la population représentée par le poids d_i^R obtenu à partir du bootstrap bayésien en population finie (Little et Zheng, 2007). Des extensions plus complètes du bootstrap bayésien en population finie à des plans d'échantillonnage complexes qui comprennent des mises en grappes et des stratifications sont consultables dans Dong, Elliott et Raghunathan (2014).

Comme dans l'estimation de (2.1), la composante non paramétrique (spline) de (2.3) peut être remplacée par d'autres estimateurs d'apprentissage automatique; voir le chapitre 4 de Rafei (2021) pour la mise en œuvre à l'aide de processus gaussiens. En outre, les applications aux modèles non normaux sont directes, bien qu'elles ne soient pas nécessairement faciles à calculer.

2.3 Estimateurs poststratifiés

Wu décrit également l'utilisation d'estimateurs poststratifiés dans le contexte d'échantillonnage par quota, qui est non seulement une très vieille forme d'échantillonnage non probabiliste, mais en fait la norme avant que Neyman défende l'idée de l'échantillonnage aléatoire stratifié (Neyman, 1934). La section 5 de l'article de Wu propose une solution de rechange solide aux estimations de score de propension obtenues en classant les observations tirées de l'échantillon probabiliste par $\hat{\pi}_i$, en les stratifiant en K strates fondées sur ce classement, et en calculant la proportion prédite de la population appartenant à la k^{e} strate comme proportion des poids d'échantillonnage W_k de cette strate en utilisant l'échantillon probabiliste à l'aide de la formule

$$\hat{\mu}_{\text{PST}} = \sum_k \hat{W}_k \bar{y}_k \tag{2.4}$$

où $\bar{y}_k$ est la moyenne dans la k^{e} strate de l'échantillon non probabiliste. Wu souligne le compromis entre le choix d'un K suffisamment grand pour conserver l'homogénéité à l'intérieur des unités, mais assez petit pour obtenir des estimations stables de $\bar{y}_k$, en indiquant 30 comme ancienne « règle empirique » pour les « tailles d'échantillon [suffisamment] grandes ». J'ajouterais qu'une approche plus officielle dont il est question dans Little (1986) consiste à adopter une méthode de création de strates (dans un contexte d'ajustement pour la non-réponse) qui limite l'erreur quadratique moyenne en maximisant la variance entre les strates à la variance dans la strate. Il semblerait qu'une telle approche conviendrait aussi à l'estimateur poststratifié des échantillons non probabilistes.

Une approche plus directe pour obtenir des estimations à l'aide d'un estimateur poststratifié est la régression multiniveau et la poststratification (Wang, Rothschild, Goel et Gelman, 2015; Downes et Carlin, 2020). Dans le cas présent, seules les données de l'échantillon non probabiliste sont utilisées dans le modèle de résultat

$$E\left(Y_{k[i]}\right) = \beta_0 + \mathbf{x}_k^T \boldsymbol{\beta} + \sum_j a_{l[k]}^j \tag{2.5}$$

où $k = 1, \ldots, K$ indexe la poststrate créée à partir des variables $j = 1, \ldots, J$, $a_{l[k]}^j \sim N(0, \sigma_j^2)$ pour $l = 1, \ldots, L_j$ et $l[k]$ met en correspondance la cellule k de la postrate avec la catégorie l de la variable j appropriée. L'estimateur poststratifié est toujours donné par (2.4), $\hat{W}_k$ étant maintenant remplacé par les totaux W_k de population connus; l'inférence postérieure est obtenue par les tirages postérieurs de β_0, $\boldsymbol{\beta}$ et $a_{l[k]}^j$ pour obtenir un tirage de

$$\hat{\mu}_{\text{PST}}^{(b)} = \sum_k W_k \left[\frac{1}{n_k} \sum_{i \in k} \left(\beta_0^{(b)} + \mathbf{x}_k^T \boldsymbol{\beta}^{(b)} + \sum_j a_{l[k]}^{j(b)} \right) \right].$$

Bien que cet estimateur ne soit pas doublement robuste au sens strict, il a démontré son bon fonctionnement dans certaines applications où J est suffisamment grand pour saisir tous les écarts importants entre un échantillon probabiliste et un échantillon non probabiliste, et lorsque l'échantillon non

probabiliste est suffisamment grand pour permettre une estimation plutôt exacte de $a_{l[k]}^{j}$. En l'absence de distributions conjointes connues de $\mathbf{X}$ à grande dimensionnalité, cette approche présente la faiblesse de reposer sur des distributions estimées qui sont instables. Une solution de rechange possible consisterait à remplacer $\bar{y}_k$ par (2.5) dans l'estimateur poststratifié de Wu (2.4), en partant du principe que les poids d'échantillonnage d_i^R résument l'information au sujet de $\mathbf{X}$ dans l'échantillon probabiliste semblable à celle du score de propension pour l'échantillon non probabiliste.

3. Pondération ou modélisation pour l'utilisateur général

L'article de Wu et les addendas ci-dessus ont tendance à suivre les sentiers battus au chapitre du choix entre la pondération et la modélisation dans le contexte de l'inférence de population finie, qui remonte au moins à Hansen, Madow et Tepping (1983). En pensant à ce choix, je crois qu'il est important de faire la distinction entre les modèles utilisés pour en déduire les paramètres dits descriptifs – au sens de Kalton (1983) – et les modèles dignes d'intérêt, les paramètres dits d'analyse dans les modèles de régression, l'analyse des classes latentes, etc. Dans le premier cas, distinguer une cible descriptive d'intérêt Y des covariables de modélisation potentielles $\mathbf{X}$ présente l'avantage de créer des estimateurs doublement robustes qui sont axés sur un seul paramètre descriptif. Cela nécessite également des hypothèses comme A1 à la section 2.1 (le score de propension ne dépend pas de la conditionnalité de Y par rapport à $\mathbf{X}$). Lorsque les modèles eux-mêmes sont les cibles d'intérêt, il est possible que l'élaboration de poids à l'aide de scores de propension pour tenir compte du biais de sélection et, comme Wu le fait remarquer, l'utilisation d'équations d'estimation pondérées standards soit le choix le plus sensé, étant donné qu'on peut habituellement envisager un grand nombre de modèles. Ce choix nécessite une double robustesse, puisqu'il n'y a généralement aucune tentative de modéliser directement les paramètres d'analyse. Trouver des façons d'étendre la double robustesse à une vaste catégorie d'estimations de paramètres de modèle est un exercice qui peut s'avérer fructueux.

4. Hypothèses non vérifiables : découvertes récentes au chapitre de l'analyse de sensibilité

Wu pose quatre hypothèses clés nécessaires pour corriger le biais de sélection dans les enquêtes non probabilistes en utilisant des données d'enquêtes probabilistes : Il s'agit essentiellement d'une « sélection aléatoire » (les covariables des échantillons non probabilistes expliquent la probabilité de sélection dans les échantillons non probabilistes); la « positivité » (tous les éléments de la population ont une probabilité non nulle de sélection dans l'échantillon non probabiliste); l'« indépendance » (les éléments sont sélectionnés de façon indépendante dans l'échantillon non probabiliste); et les « covariables courantes » (il existe une enquête probabiliste avec des covariables dont le sous-ensemble correspond aux covariables nécessaires pour que l'hypothèse de données manquantes au hasard reste valable). Il conviendrait de

mentionner que les deux premières hypothèses nécessitent que l'enquête non probabiliste soit une enquête probabiliste « déguisée », c'est-à-dire qu'il y a réellement des probabilités non nulles de sélection, dans l'enquête non probabiliste, de tous les éléments de la population, mais en tant qu'analystes, nous ne connaissons pas la nature de ces éléments.

Dans la pratique, il se peut qu'aucune de ces hypothèses ne soit vérifiable avec précision. Certaines études récentes ont porté sur l'échec de la première hypothèse, soit celle de la « sélection aléatoire ». Certaines mesures existantes tirées de la littérature sur la non-réponse ont été adaptées dans le cas présent, comme la mesure de l'indicateur R (Schouten, Cobben et Bethlehem, 2009), qui dans ce contexte est la mesure de la variabilité des probabilités de sélection dans l'échantillon non probabiliste :

$$\hat{R} = 1 - 2\sqrt{\frac{1}{n_a - 1}\sum_{i=1}^{n_A}\left(\hat{\pi}_i^A - \sum_{j=1}^{n_A}\hat{\pi}_j^A / n_a\right)^2}$$

$\hat{R}$ varie entre 0 et 1, où la valeur 1 est atteinte lorsque la probabilité de sélection est constante, ce qui permet de penser à un simple échantillon aléatoire présentant un risque moindre de biais de sélection, et la valeur 0 indiquant que tous les éléments sont soit inclus dans la probabilité de 1 ou de 0, ce qui augmente le risque de biais de sélection.

Bien entendu, en l'absence du résultat Y dans l'échantillon probabiliste, il n'y a aucun moyen d'évaluer directement le biais de sélection. De ce fait, des travaux récents ont étendu le champ d'action de l'étude de Andridge et Little (2011) qui permet d'élaborer une analyse de sensibilité à l'aide d'un modèle de mélange de schémas d'observation, où la sélection dans l'échantillon non probabiliste a la possibilité de dépendre entièrement d'une réduction scalaire des covariables $\mathbf{X}$, entièrement du résultat Y, ou d'une combinaison convexe de ces éléments. Little, West, Boonstra et Hu (2020), Andridge, West, Little, Boonstra et Alvarado-Leiton (2019), et West, Little, Andridge, Boonstra, Ware, Pandit et Alvarado-Leiton (2021) considèrent la sensibilité en lien avec cette hypothèse dans l'estimation de la moyenne d'une variable normalement distribuée, la moyenne d'un résultat binaire et les paramètres de régression d'un modèle de régression linéaire, respectivement, dans les échantillons non probabilistes. En variant le paramètre de mélange convexe ϕ, il est possible d'évaluer la sensibilité à l'hypothèse de « sélection aléatoire ». Boonstra, Little, West, Andridge et Alvarado-Leiton (2021) constatent que ces « mesures normalisées du biais » soutiennent favorablement la comparaison avec d'autres solutions, comme $\hat{R}$ dans une étude de simulation. Il est important de mentionner que les méthodes qui complètent l'étude de Andridge et Little (2011) ne dépendent pas de l'hypothèse de covariables communes dans un échantillon probabiliste. Cela donne à penser que les méthodes qui utilisent les renseignements disponibles dans l'échantillon probabiliste pour évaluer l'hypothèse de la « sélection aléatoire » sont un domaine laissant place à l'évolution.

La seconde hypothèse, la positivité, est également peu susceptible d'être applicable avec précision dans de nombreux contextes pratiques. Mes propres travaux dans ce domaine ont porté sur des études de conduite en situation réelle tirées le plus souvent d'échantillons de commodité dans une région

géographique limitée. Par exemple, le *Second Strategic Highways Research Program* a recruté des conducteurs dans six régions géographiques précises des États-Unis (*Transportation Research Board* de la *National Academy of Sciences*, 2013). Cela correspond au deuxième scénario énoncé par Wu dans la section 7.2, où seule une sous-population a la possibilité d'être sélectionnée dans l'échantillon non probabiliste, ce qui, comme il le fait remarquer, n'a « pas de solution simple ». En suivant sa notation de D indiquant l'appartenance à une sous-population, il semblerait que si $D_i \perp \mathbf{X}_i, Y_i \mid \pi_i^A$ – c'est-à-dire, si la distribution de $\mathbf{X}, Y$ est la même pour $D = 0$ et $D = 1$ après pondération avec π_i^A à l'intérieur de la strate $D = 1$ –, alors l'absence de positivité n'aurait aucune incidence sur l'inférence. Il s'agit probablement d'un défi de taille dans les contextes les plus généraux, mais la positivité pourrait être plutôt bien estimée si l'analyse d'intérêt comprend un sous-ensemble de $\mathbf{X}, Y$ qui n'est que faiblement associé à D, même avant ajustement.

Enfin, en ce qui concerne la quatrième hypothèse, soit l'existence d'un échantillon probabiliste avec $\mathbf{X}$ disponible, je suis tout à fait d'accord avec l'observation de Wu voulant que les méthodes permettant de tirer parti d'enquêtes probabilistes multiples doivent être développées davantage. Cependant, il reste plus probable qu'un chercheur s'efforce de trouver un seul échantillon probabiliste avec suffisamment de covariables que de se débattre avec une surabondance d'options (ce que Wu appelle le « dilemme de la personne riche »). À cette fin, je terminerai par un appel à l'action lancé par les spécialistes des enquêtes.

5. L'échantillonnage probabiliste au 21e siècle : maintenant plus que jamais

J'ai appris les statistiques, plus particulièrement les statistiques issues d'enquêtes, vers la fin du 20e siècle, lorsque l'échantillonnage probabiliste était la pierre de touche incontestée du plan d'enquête. J'ai été initié pour la première fois au problème d'inférence à partir d'échantillons non probabilistes à la fin des années 2000, dans un contexte d'analyse des blessures faisant appel aux données du *Crash Injury Research Engineering Network* des États-Unis, où les analystes traitaient un échantillon très restreint de personnes accidentées dans des véhicules automobiles comme s'il s'agissait d'un échantillon aléatoire de victimes d'accidents de la route et qu'ils obtenaient par conséquent des résultats non réalistes (Elliott et coll., 2010). Vers la même époque, la popularité des enquêtes en ligne explosait et les statisticiens d'enquête ne savaient pas vraiment comment tirer des conclusions à partir de ces données. J'avoue avoir eu une attitude plutôt paternaliste à l'époque; j'ai presque évité de faire de la recherche dans ce domaine parce que je pensais que cela ne ferait qu'encourager un « mauvais comportement » à l'égard des plans d'échantillonnage. Je ne pensais pas pouvoir arrêter cette tendance à moi tout seul, mais je ne voulais pas participer à ce que j'ai perçu comme la dévalorisation de la science. J'en suis néanmoins venu à reconnaître que bon nombre de ces nouvelles sources de données présentent des avantages qui vont au-delà de ce que l'on peut obtenir au moyen d'un échantillon probabiliste traditionnel, surtout avec des budgets limités. Cela va bien au-delà des défis croissants que représente la mise en œuvre d'enquêtes

probabilistes, notamment auprès de la population générale, en raison des non-réponses, de l'absence de bases de sondage adéquates, etc.

Pour autant, je reste inquiet à l'idée qu'en ayant conçu des méthodes pour faire face aux limites des enquêtes non probabilistes, l'échantillonnage probabiliste est devenu chose du passé chez les scientifiques et les décideurs dont la formation statistique est limitée, et ce, malgré des efforts comme ceux déployés par Bradley, Kuriwaki, Isakov, Sejdinovic, Meng et Flaxman (2021), et Marek, Tervo-Clemmens, Calabro et coll. (2022). Cependant, comme Wu le souligne dans son analyse, l'absence d'échantillons probabilistes rend l'échantillon non probabiliste moins vulnérable à la possibilité d'un calage même partiel ou d'autres méthodes d'ajustement (bien que les analyses de sensibilité comme les approches de « mesures normalisées du biais » mentionnées ci-dessus ne nécessitent pas d'échantillons probabilistes de référence). Par conséquent, je crois qu'il est de plus en plus essentiel de mettre en place des enquêtes probabilistes structurées et idéalement financées par le gouvernement pour les collectes de données courantes. Ce type d'enquêtes existe déjà – l'*American Community Survey* du *United States Census Bureau* et la *National Health Interview Survey* du *National Center for Health Statistics* en font notamment partie –, mais à l'avenir, je crois qu'il serait utile pour les organismes statistiques de s'entendre sur la nécessité de mettre les enquêtes probabilistes de grande qualité à la disposition des enquêtes non probabilistes à des fins d'analyse, et non de les diffuser comme de simples publications autonomes. Cela signifie qu'il faut réfléchir attentivement aux covariables importantes dans diverses fonctions de santé publique et de sciences sociales dans lesquels les données d'enquête jouent un rôle. Il faudra faire des choix compte tenu des contraintes budgétaires limitées et, en même temps, prévoir un financement suffisant pour conserver la qualité nécessaire aux ajustements. Enfin, bien que certaines méthodes ne nécessitent pas de microdonnées et peuvent se servir de mesures agrégées, comme celles qu'on retrouve dans l'*American Communities Survey*, d'autres nécessiteront ce type de données, ce qui signifie que de nouveaux domaines de recherche seront à explorer au chapitre des renseignements personnels et de la confidentialité en ce qui concerne la combinaison de données tirées d'enquêtes probabilistes et non probabilistes.

Bibliographie

Andridge, R.R., et Little, R.J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27, 153-180.

Andridge, R.R., West, B.T., Little, R.J., Boonstra, P.S. et Alvarado-Leiton, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *Journal of the Royal Statistical Society*, C68, 1465-1483.

Boonstra, P.S., Little, R.J., West, B.T., Andridge, R.R. et Alvarado-Leiton, F. (2021). A simulation study of diagnostics for selection bias. *Journal of Official Statistics*, 37, 751-769.

Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.L. et Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600, 695-700.

Cassel, C.M., Särndal, C.-E. et Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

Chen, J.K.T., Valliant, R.L. et Elliott, M.R. (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society*, 68, 657-681.

Chen, Y., Li, P. et Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Chipman, H.A., George, E.I. et McCulloch, R.E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4, 266-298.

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.

Downes, M., et Carlin, J.B. (2020). Multilevel regression and poststratification as a modeling approach for estimating population quantities in large population health studies: A simulation study. *Biometrical Journal*, 62, 479-491.

Eilers, P.H., et Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89-121.

Elliott, M.R. (2013). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2(6).

Elliott, M.R., et Davis, W.W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: Combining data from the Behavioral Risk Factor Surveillance Survey and the National Health Interview Survey. *Journal of the Royal Statistical Society*, C54, 595-609.

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

Elliott, M.R., Resler, A., Flannagan, C.A. et Rupp, J.D. (2010). Appropriate analysis of CIREN data: Using NASS-CDS to reduce bias in estimation of injury risk factors in passenger vehicle crashes. *Accident Analysis and Prevention*, 42, 530-539.

Hájek, J. (1971). Comment on a paper by D. Basu. *Foundations of Statistical Inference*, 236.

Hansen, M.H., Madow, W.G. et Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

Kalton, G. (1983). Models in the practice of survey sampling. *Revue Internationale de Statistique*, 51, 175-188.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54 139-157.

Little, R.J.A., et Zheng, H. (2007). The Bayesian approach to the analysis of finite population surveys. *Bayesian Statistics*, 8, 1-20.

Little, R.J.A., West, B.T., Boonstra, P.S. et Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of Survey Statistics and Methodology*, 8, 932-964.

Marek, S., Tervo-Clemmens, B., Calabro, F.J. et coll. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, sous presse.

McConville, K.S., Breidt, F.J., Lee, T. et Moisen, G.G. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology*, 5, 131-158.

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, 558-625.

Rafei, A. (2021). Robust and efficient Bayesian inference for large-scale non-probability samples. Thèse de l'University of Michigan. Accessible sur https://www.overleaf.com/project/6228db145a47be05f8da3777.

Rafei, A, Flannagan, C.A.C. et Elliott, M.R. (2020). Big data for finite population inference: Applying quasi-random approaches to naturalistic driving data using Bayesian additive regression trees. *Journal of Survey Statistics and Methodology*, 8, 148-180.

Rivers, D. (2007). Sampling for web surveys. *Proceedings of the Joint Statistical Meetings*. Disponible sur https://static.texastribune.org/media/documents/Rivers_matching4.pdf.

Schouten, B., Cobben, F. et Bethlehem, J. (2009). Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête*, 35, 1, 101-113. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2009001/article/10887-fra.pdf.

Soentpiet, R. (1999). *Advances in Kernel Methods: Support Vector Learning*. Boston: MIT Press.

Transportation Research Board of the National Academy of Sciences (2013). *The 2nd Strategic Highway Research Program Naturalistic Driving Study Dataset.*

Valliant, R., et Dever, J.A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods and Research*, 40, 105-137.

Van Der Laan, M.J., et Rubin, D.R. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

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

West, B.T., Little, R.J.A., Andridge, R.R., Boonstra, P.S., Ware, E.B., Pandit, A. et Alvarado-Leiton, F. (2021). Assessing selection bias in regression coefficients estimated from nonprobability samples with applications to genetics and demographic surveys. *The Annals of Applied Statistics*, 15, 1556-1581.

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, 185-193.

Zhang, G., et Little, R.J.A. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65, 911-918.

Zheng, H., et Little, R.J.A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.

Zhou, T., Elliott, M.R., et Little, R.J.A. (2019). Penalized spline of propensity methods for treatment comparison. *Journal of the American Statistical Association*, 114, 1-19.

Techniques d'enquête, décembre 2022
Vol. 48, nº 2, pp. 359-367
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article « Inférence statistique avec des échantillons d'enquête non probabiliste »

Sharon L. Lohr[1]

Résumé

Des hypothèses solides sont nécessaires pour faire des inférences au sujet d'une population finie à partir d'un échantillon non probabiliste. Les statistiques d'un échantillon non probabiliste devraient être accompagnées de preuves que les hypothèses sont respectées et que les estimations ponctuelles et les intervalles de confiance sont propres à l'utilisation. Je décris certains diagnostics qui peuvent être utilisés pour évaluer les hypothèses du modèle, et je discute des questions à prendre en considération au moment de décider s'il convient d'utiliser les données d'un échantillon non probabiliste.

Mots-clés : Échantillon de commodité; diagnostic; imputation; échantillon probabiliste; qualité de l'enquête; poids d'enquête.

1. Introduction

Un grand merci à Changbao Wu pour son examen stimulant et son évaluation des méthodes permettant de faire des inférences à partir d'échantillons non probabilistes. J'apprécie particulièrement son examen réfléchi des hypothèses solides nécessaires pour établir le biais et la variance des estimations.

M. Wu passe en revue trois approches pour estimer la moyenne de population finie μ_y d'une variable y qui est mesurée dans un échantillon S_A non probabiliste de taille n_A. Étant donné que cet échantillon n'est pas représentatif de la population (et que, par conséquent, la moyenne de l'échantillon $\bar{y}_A$ est probablement biaisée pour estimer μ_y), chaque approche repose sur l'information provenant d'un échantillon probabiliste S_B de grande qualité et de taille n_B : S_B ne mesure pas y, mais contient un ensemble de variables auxiliaires $\mathbf{x}$ qui sont également observées dans S_A.

Dans l'approche prédictive basée sur un modèle, un modèle est créé sur S_A pour prédire y à partir de $\mathbf{x}$. L'estimateur d'imputation massive (IM), par exemple, utilise le modèle pour imputer une estimation y_i^* de y_i pour chaque membre de l'échantillon probabiliste S_B. Ensuite, le total de la population de y est estimé par $\sum_{i \in S_B} d_i^B y_i^*$ où d_i^B est le poids d'enquête de l'unité i dans S_B.

Dans l'approche de la pondération par l'inverse de la propension (PIP), un modèle est élaboré pour prédire la probabilité π_i^A que l'unité de population i apparaisse dans S_A en fonction de $\mathbf{x}$. L'unité i dans S_A reçoit ensuite la pondération $w_i^A = 1/\hat{\pi}_i^A$ et le total de la population est estimé par $\sum_{i \in S_A} w_i^A y_i$.

M. Wu se penche également sur un estimateur « doublement robuste » de μ_y qui, en combinant les estimateurs prédictifs et de PIP, est approximativement sans biais selon les hypothèses que l'un ou l'autre des modèles est correctement indiqué. Dans cet exposé, je me concentrerai sur les approches prédictives et de PIP, parce que ces méthodes se généralisent plus facilement pour les analyses multivariées et l'estimation des caractéristiques des populations autres que les moyennes.

1. Sharon Lohr est professeure émérite à l'Arizona State University. Courriel : sharon.lohr@asu.edu.

À la section 2, j'examinerai les hypothèses nécessaires à l'inférence à partir d'échantillons non probabilistes et des diagnostics pour les évaluer. Ensuite, à la section 3, je me pencherai sur certaines questions à poser au moment de décider quelle approche (le cas échéant) utiliser pour l'inférence.

2. Hypothèses et diagnostics de modèle

L'échantillonnage probabiliste a été largement utilisé après l'élaboration de la théorie dans les années 1930 et 1940 parce qu'il offrait une solution mathématiquement justifiée au problème de la généralisation d'un échantillon à une population. Selon des hypothèses minimales, un échantillon probabiliste avec réponse complète produit des estimations approximativement sans biais des quantités de population, accompagnées d'intervalles de confiance qui ont des probabilités de couverture approximativement exactes. C'est la *seule* méthode qui garantit la création d'intervalles de confiance exacts sans qu'il soit nécessaire de poser d'hypothèses sur les membres non échantillonnés d'une population donnée. Un échantillon probabiliste est représentatif en raison de la procédure utilisée pour l'établir.

Toutes les autres méthodes exigent d'énormes hypothèses. Les principales hypothèses applicables aux méthodes prédictives et de PIP, qui sont présentées à la section 2.1 de l'article de M. Wu, sont les suivantes : (A1) y et la variable aléatoire indiquant la participation à S_A sont indépendants étant donné $\mathbf{x}$; (A2) chaque unité de la population se caractérise par $\pi_i^A > 0,$ et (A3) les variables aléatoires indiquant la participation à S_A sont indépendantes étant donnée $\mathbf{x}$. Ces hypothèses laissent entendre que l'information auxiliaire $\mathbf{x}$ est suffisamment riche pour élaborer des poids de propension inverse qui éliminent le biais de sélection pour $y,$ et qu'un modèle élaboré sur S_A pour prédire y à partir de $\mathbf{x}$ s'appliquera également aux unités qui ne sont pas dans S_A.

Les propriétés statistiques des estimateurs sont élaborées en supposant que les hypothèses (A1) à (A3) sont vraies et que les modèles adoptés pour la pondération ou l'imputation sont correctement indiqués. Dans ces conditions, la moyenne estimée de la population est approximativement sans biais avec la variance obtenue par le théorème approprié. Mais, comme M. Wu le souligne dans son article, cette estimation de la variance est conditionnelle au respect des hypothèses. Autrement, l'erreur quadratique moyenne réelle sera sérieusement sous-estimée et la fiabilité de l'estimation donnera une impression trompeuse. Si n_A et n_B sont importants, mais que (A1) n'est pas respectée, le biais pourrait être de 10 points de pourcentage, mais l'erreur-type déclarée d'une estimation d'IM ou de PIP sera proche de zéro. Dans la pratique, de nombreux échantillons non probabilistes ne respecteront pas les hypothèses. Mercer, Lau et Kennedy (2018) ont constaté, en pondérant les échantillons d'enquêtes volontaires en ligne avec de l'information auxiliaire riche, que « même la stratégie de rajustement la plus efficace n'a permis d'éliminer qu'environ 30 % du biais initial ».

Il est impossible de mettre les hypothèses pleinement à l'essai parce qu'elles se caractérisent pas des données manquantes : des membres de la population sont absents de $S_A,$ et des valeurs y sont absentes de S_B. Toutefois, comme pour les rajustements effectués pour compenser la non-réponse dans les

échantillons probabilistes (Lohr, 2022, chapitre 8), il est possible d'effectuer des vérifications et des diagnostics de modèle à l'aide des renseignements disponibles, tout en étant conscient que cela risque de ne pas déceler toutes les lacunes du modèle.

Comparaison des statistiques de l'échantillon non probabiliste avec celles d'autres sources de données

M. Wu suggère de comparer les fonctions de distribution empirique des variables contenues dans $\mathbf{x}$ à partir de S_A avec les fonctions de distribution empirique pondérées par des poids d'enquête tirés de S_B. Les différences peuvent indiquer que les observations dans S_A ont des scores de propension inégaux, ou que les variables $\mathbf{x}$ dans S_A et dans S_B sont mesurées de façon distincte (voir la section 3). Il est également possible de comparer les distributions empiriques de S_A avec celles de S_C d'une autre enquête probabiliste.

Si l'on utilise la PIP, on peut aussi comparer les fonctions de distribution empirique pondérées par le score de propension de S_A avec celles de S_B et d'autres enquêtes. Cela ne devrait être fait que pour les variables qui ne sont pas utilisées dans la pondération, puisque les poids des scores de propension ont déjà été rajustés pour tenir compte des déséquilibres dans les variables de pondération. Dutwin et Buskirk (2017), par exemple, ont construit des poids de propension pour un échantillon non probabiliste en ratissant des totaux marginaux, puis en comparant les tableaux croisés de ces variables de ratissage.

M. Wu suggère également de traiter une variable z qui est mesurée à la fois dans S_A et S_B en tant que variable réponse, et de comparer les modèles conditionnels de $z \mid \mathbf{u}$ ajustés sur S_A et S_B, où $\mathbf{u}$ est un sous-ensemble de $\mathbf{x}$ (à l'exclusion de z). Les différences entre les deux modèles peuvent indiquer que z est nécessaire en tant que variable auxiliaire, et peuvent également soulever la question de savoir dans quelle mesure l'ensemble de variables auxiliaires mesurées répond à l'hypothèse (A1).

Dans un exemple de Kim, Park, Chen et Wu (2021), le pourcentage estimé de personnes qui s'étaient portées volontaires était de 24,8 % selon la *Current Population Survey* (estimation de référence), mais les estimations de l'IM et de la PIP tirées de S_A étaient toutes deux proches de 50 % avec une erreur-type déclarée de moins d'un point de pourcentage. L'erreur-type, calculée selon les hypothèses du modèle, n'a pas tenu compte du biais de sélection de S_A en lien avec la participation volontaire, un biais qui ne pouvait pas être éliminé en utilisant les données démographiques, l'accession à la propriété et l'assurance médicale comme covariables du modèle.

Comparer les résultats des approches de la PIP et de l'IM

Une solution de rechange à l'utilisation d'un estimateur doublement robuste aux fins d'analyse consiste à utiliser chaque modèle pour déterminer les lacunes possibles de l'autre. Les examens possibles consistent à comparer la distribution empirique de y tirée de S_A (à l'aide des poids de propension inverse) avec la distribution empirique de y^* tirée de S_B (en utilisant les valeurs imputées et les poids d'enquête). De même, comme l'ont suggéré Chipperfield, Chessman et Lim (2012), on peut comparer les moyennes de domaine estimées à partir de S_A et S_B pour un ensemble de domaines $d = 1, \ldots, D$. On

pourrait aussi comparer les imputations pour l'ajustement de y à l'ensemble de données non pondérées S_A avec des imputations élaborées pour S_A avec des poids de propension inverse.

Les études de simulation sont utiles pour vérifier le comportement d'un petit échantillon lorsque les hypothèses sont respectées, mais elles ont un intérêt limité pour étudier la sensibilité aux hypothèses du modèle. Ces études permettent d'analyser les écarts entre les modèles conçus par les chercheurs, mais les enquêtes réelles peuvent diverger du modèle de nombreuses façons imprévues.

Effectuer des diagnostics de modèle

Bien entendu, qu'il s'agisse de la PIP ou de l'approche prédictive basée sur un modèle, les analystes devraient utiliser des diagnostics de régression standard, comme l'examen des résidus et les observations influentes pour étudier l'ajustement du modèle et la sensibilité aux valeurs aberrantes, et garder une trace des vérifications ainsi effectuées.

Dans le cas de l'approche de PIP, il est également souhaitable d'examiner les caractéristiques des poids finaux. Le coefficient de variation des poids fournit une mesure approximative de l'ampleur des ajustements qui ont été nécessaires pour rendre l'échantillon S_A représentatif. Toutefois, un faible coefficient de variation ne signifie pas nécessairement que l'échantillon est représentatif; cela peut simplement refléter l'insuffisance d'information auxiliaire pour élaborer des poids. Par exemple, supposons qu'un échantillon par quota d'un groupe à participation volontaire sur Internet est extrait pour le faire correspondre à la population en ce qui concerne les variables auxiliaires. Les poids de propension inverse connaîtront peu de variation parce que les variables $\mathbf{x}$ ont été utilisées pour former les catégories de quotas, mais l'échantillon peut tout de même produire des estimations biaisées des variables y, comme l'utilisation d'Internet ou la participation volontaire.

Les méthodes graphiques proposées par Makela, Si et Gelman (2014) pour évaluer les ajustements de poids dans les enquêtes peuvent également être utilisées avec la PIP. Brick (2015) suggère d'examiner l'ampleur des ajustements de la PIP dans les classes de pondération. On peut également examiner la répartition des poids dans les domaines d'intérêt.

Les poids de propension inverse peuvent également fournir des renseignements sur l'hypothèse (A2). Un domaine qui a des pondérations élevées par rapport à d'autres domaines peut signaler un sous-dénombrement dans S_A. Dever (2018) a proposé d'examiner l'hypothèse (A2) en identifiant les répondants dans S_B qui n'ont pas de correspondance étroite dans S_A.

Bondarenko et Raghunathan (2016) ont examiné et proposé des outils de diagnostic graphiques et numériques pour évaluer et améliorer les modèles d'imputation. Cependant, aucun de ces diagnostics ne permet de valider l'hypothèse selon laquelle l'ajustement du modèle de régression sur S_A s'applique aux unités qui ne sont pas dans S_A. Tout comme $\bar{y}_A$ pourrait être un estimateur biaisé de μ_y, les coefficients de régression dérivés de S_A peuvent également être biaisés, et le modèle construit sur S_A pour prédire y à partir de $\mathbf{x}$ pourrait ne pas s'appliquer à d'autres parties de la population.

Prendre un petit échantillon probabiliste pour analyser les hypothèses

Les étapes précédentes peuvent permettre de déceler certaines lacunes du modèle, mais ne peuvent pas valider entièrement les hypothèses (A1) et (A2). Il est néanmoins possible de mettre à l'essai le modèle d'imputation en obtenant des données sur y à partir d'un sous-échantillon probabiliste de S_B. De même, on pourrait prendre un échantillon probabiliste de membres de la population qui ne sont pas présents dans S_A pour vérifier les inférences de l'approche de PIP, ou observer y sur un sous-échantillon d'unités dans S_B qui sont semblables à celles assorties d'une pondération élevée dans S_A, ou qui n'ont pas de correspondance étroite dans S_A.

3. Quand faut-il utiliser des échantillons non probabilistes ?

M. Wu décrit les méthodes pour combiner l'information des échantillons probabilistes avec celle des échantillons non probabilistes après qu'une décision a été prise à cette fin. Or, une première question est de savoir s'il faut réellement effectuer cette opération. On pourrait vouloir utiliser un échantillon non probabiliste parce qu'il n'existe pas d'échantillon probabiliste de grande qualité pour mesurer y, et qu'on pense que toute information est préférable à un manque d'information. Mais est-ce vrai ?

Supposons que, malgré l'ajustement minutieux et la vérification du modèle, les statistiques clés demeurent biaisées. La déclaration d'une statistique erronée pourrait-elle être pire que la déclaration d'aucune statistique ? Les mauvaises statistiques, une fois publiées, peuvent circuler longtemps, même après que des études plus rigoureuses montrent qu'elles sont biaisées. En 1975, la chroniqueuse Ann Landers a demandé à ses lecteurs de répondre à la question suivante : « Si c'était à refaire, auriez-vous des enfants ? » Environ 70 % des 10 000 personnes qui ont envoyé une réponse par la poste ont déclaré qu'elles n'auraient pas eu d'enfants si elles en avaient le choix. Cette statistique est toujours citée presque 50 ans plus tard, même si elle provient d'un échantillon de commodité et qu'elle a été contredite par de nombreuses autres études (Lohr, 2022). Il est également peu probable que la modélisation prédictive ou la PIP aurait corrigé le biais de sélection touchant la statistique de Landers, qui porte sur tous les groupes démographiques.

En gardant ces considérations à l'esprit, voici quelques questions qu'il conviendrait de poser au moment de décider d'utiliser des estimations d'un échantillon non probabiliste et, le cas échéant, quelle méthode statistique il faudrait adopter pour faire des inférences.

- Comment les statistiques seront-elles utilisées ? Les estimations d'un échantillon non probabiliste peuvent être utiles à l'élaboration d'une stratégie de marketing ou pour une étude sociologique exploratoire, mais elles pourraient ne pas être jugées suffisamment fiables pour estimer le chômage ou le nombre de personnes ayant besoin d'aide alimentaire. Les statistiques d'un échantillon non probabiliste devraient être accompagnées de preuves que les hypothèses sont propres à l'utilisation.

- Quelle est la qualité des données dans S_A ? Les dossiers administratifs, comme les dossiers fiscaux, ont un profil de qualité différent de celui d'une enquête sur des volontaires recrutés au moyen d'une annonce sur Internet.

  Si la population de S_A est bien définie (il s'agit par exemple de déclarants fiscaux), il peut être préférable de déclarer des statistiques pour cette population plutôt que d'essayer de généraliser à la population de S_B. En ce qui concerne les dossiers fiscaux, de nombreuses personnes dont le revenu est inférieur aux seuils de revenu préétablis se caractérisent par la valeur $\pi_i^A = 0,$ ce qui enfreint l'hypothèse (A2). On pourrait plutôt adopter une approche à bases multiples, où une source de données différente est utilisée afin d'estimer μ_y pour les parties de la population qui ne sont pas dans S_A (Lohr, 2021).

  Étant donné que tous les modèles reposent sur des données auxiliaires $\mathbf{x}$, il est important que S_A et S_B mesurent les variables $\mathbf{x}$ de la même façon. Si le revenu est utilisé comme variable auxiliaire, les mêmes questions doivent être utilisées pour définir le revenu dans les deux enquêtes, et le revenu doit être mesuré pour la même unité (personne ou ménage).

  Kennedy (2022) indique que certains répondants qui choisissent de participer à des sondages en ligne peuvent fournir des renseignements démographiques inexacts ou de fausses réponses aux questions et que si cela se produit, les prévisions du modèle seront erronées. Il est même possible que des intervenants externes souhaitant obtenir un résultat particulier manipulent les données contenues dans S_A. Par exemple, une organisation pourrait s'arranger pour qu'une enquête soit attribuée à un ensemble de volontaires dont les caractéristiques démographiques prétendues correspondent à celles de la population, mais qui fournissent la réponse « souhaitée » pour y. Certains partisans des échantillons non probabilistes soutiennent que les échantillons probabilistes à faible taux de réponse nécessitent également des ajustements de poids ou une imputation, mais il y a une différence importante, à savoir que l'enquête probabiliste peut se caractériser par des non-réponses, mais l'échantillon initial est sélectionné au hasard et ne peut pas être manipulé par des organisations externes.

  Si les données contenues dans S_A sont de faible qualité, vaut-il la peine de construire des modèles ? Comme l'a dit Louis (2016), « les procédures de l'ère spatiale ne sauveront pas les données de l'âge de pierre ».

- Dans quelle mesure les renseignements auxiliaires sont-ils détaillés ? Si S_A est importante et que l'information auxiliaire est suffisamment précise pour permettre de repérer des enregistrements précis, alors le couplage des enregistrements entre S_A et S_B serait une meilleure méthode pour combiner les données. On se servira de l'imputation ou de la PIP si l'information auxiliaire est suffisamment riche pour donner de bonnes prédictions de y_i ou de π_i^A, mais pas assez riche pour permettre un couplage précis. Cependant, s'il y a peu d'information auxiliaire, on peut s'attendre à une faible variation des scores de propension ou

des valeurs imputées, et les méthodes peuvent donner de mauvaises prévisions, étant donné le peu d'information permettant de diagnostiquer des problèmes éventuels.

- Quelles sont les analyses souhaitées ? M. Wu discute de l'estimation de la moyenne de la population, mais l'analyste pourrait aussi vouloir examiner les relations entre y et d'autres variables, ou estimer les moyennes ou les médianes pour les sous-groupes. Le choix de la méthode dépend en partie des variables disponibles dans S_A et S_B. Si S_A contient de nombreuses variables de réponse dont la relation présente un intérêt, l'approche de PIP pourrait être privilégiée.

  Si l'on souhaite explorer les relations entre y et les variables mesurées uniquement dans S_B, l'imputation pourrait être un meilleur choix. Dans ce cas, cependant, l'analyste doit prendre soin de reconnaître l'imputation lorsqu'il présente les résultats. Si, par exemple, la régression linéaire est utilisée pour l'imputation, la corrélation calculée sur S_B n'est pas entre les variables u et y, mais entre u et $\mathbf{x}^T\hat{\boldsymbol{\beta}}$.

- Quelles sont les répercussions sur l'équité des données ? Jagadish, Stoyanovich et Howe (2021) ont défini l'« équité en matière de représentation » comme une « augmentation de la visibilité des groupes sous-représentés qui ont été historiquement désavantagés ou supprimés dans l'enregistrement de données ».

  Les échantillons non probabilistes peuvent améliorer l'équité des données. Ils peuvent accroître la taille de l'échantillon et la visibilité des sous-groupes de population rares – un grand ensemble de données S_A pourrait contenir 10 000 membres d'un sous-groupe, tandis que même une enquête probabiliste à réponse complète de $n_B = 60\,000$ pourrait en contenir seulement 10. Ou encore, l'échantillon non probabiliste pourrait contenir des membres de la population qui sont sous-représentés dans l'enquête probabiliste parce qu'ils sont hors du champ de l'enquête, sous-dénombrés dans la base de sondage ou qu'ils présentent une tendance à la non-réponse. Dans ces situations, S_A fournit des renseignements sur les groupes qui ne sont pas aussi bien représentés dans l'enquête probabiliste.

  Par ailleurs, les groupes historiquement défavorisés pourraient être sous-représentés dans toutes les sources de données, y compris dans S_A. Par exemple, un grand échantillon non probabiliste de dossiers de santé électroniques pourra produire des estimations pour un plus grand nombre de sous-groupes de population qu'un petit échantillon probabiliste sur la santé. Or, les personnes sans assurance-maladie ou qui n'ont pas accès aux soins médicaux seront sous-représentées. Dans cette situation, le fait d'avoir recours à S_A pour la production d'estimations démographiques est susceptible de renforcer les inégalités. Si des estimations sont utilisées pour répartir les ressources, alors, au fur et à mesure de la mise en œuvre du programme, davantage de données seront recueillies dans les régions qui obtiennent ces ressources et valideront leurs besoins, alors qu'aucun suivi ne sera effectué dans les régions qui sont désignées à tort comme

ne recevant aucune ressource. La boucle de rétroaction propagera la représentation inéquitable dans les sources de données.

Les méthodes d'IM et de PIP ont des répercussions différentes sur l'équité des données. L'imputation attribue une valeur prédite de y à chaque observation dans S_B, et la valeur imputée y peut différer de la valeur y que le répondant aurait fournie si on lui posait la question, en particulier si le répondant fait partie d'un sous-groupe qui est sous-représenté ou mal représenté dans S_A. Le modèle fournira-t-il des prédictions exactes pour les sous-groupes historiquement sous-représentés ? Les répondants de S_B ont-ils donné leur consentement éclairé pour permettre l'imputation de y ?

La PIP part du principe que les scores de propension peuvent être estimés à partir de renseignements auxiliaires. Cette information est-elle suffisamment riche pour donner des poids exacts ? Certains sous-groupes sont-ils sous-représentés dans S_A ? Il peut être utile de comparer les résultats des deux méthodes et d'autres sources de données, si elles sont disponibles, pour les sous-groupes de population historiquement sous-représentés.

L'examen critique de M. Wu soulève de nombreuses questions importantes pour les personnes qui envisagent d'utiliser des échantillons non probabilistes pour faire des inférences au sujet de la population. J'apprécie particulièrement son évaluation des hypothèses solides nécessaires pour les méthodes fondées sur des modèles, et je salue l'accent mis sur la résolution de ces problèmes à l'étape des plans d'enquête.

Bibliographie

Bondarenko, I., et Raghunathan, T. (2016). Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Statistics in Medicine,* 35(17), 3007-3020.

Brick, J.M. (2015). Compositional model inference. Dans *Proceedings of the Survey Research Methods Section*, Alexandrie, Virginie: American Statistical Association, 299-307.

Chipperfield, J., Chessman, J. et Lim, R. (2012). Combining household surveys using mass imputation to estimate population totals. *Australian & New Zealand Journal of Statistics,* 54, 223-238.

Dever, J.A. (2018). Combining probability and nonprobability samples to form efficient hybrid estimates: An evaluation of the common support assumption. Dans *Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference*. https://nces.ed.gov/FCSM/pdf/A4_Dever_2018FCSM.pdf.

Dutwin, D., et Buskirk, T.D. (2017). Apples to oranges or gala versus golden delicious? Comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly,* 81(S1), 213-239.

Jagadish, H.V., Stoyanovich, J. et Howe, B. (2021). COVID–19 brings data equity challenges to the fore. *Digital Government: Research and Practice*, 2(2), 1-7.

Kennedy, C. (2022). Exploring the assumption that online opt-in respondents are answering in good faith. Communication présentée à la conférence 2022 Morris Hansen, 1er mars 2022.

Kim, J.-K., Park, S., Chen, Y. et Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society, Series A*, 184, 941-963.

Lohr, S.L. (2021). Les enquêtes à bases de sondage multiples pour un monde fait de sources de données multiples. *Techniques d'enquête,* 47,2, 247-285. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2021002/article/00008-fra.pdf.

Lohr, S.L. (2022). *Sampling: Design and Analysis, Third Edition*. Boca Raton, FL: CRC Press.

Louis, T.A. (2016). Discussion of combining information from survey and non-survey data sources: Challenges and opportunities. Séminaire public de la 130e réunion du CNSTAT; Washington, DC. https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_172505.pdf.

Makela, S., Si, Y. et Gelman, A. (2014). Statistical graphics for survey weights. *Revista Colombiana de Estadística*, 37(2), 285-295.

Mercer, A., Lau, A. et Kennedy, C. (2018). *For Weighting Online Opt-In Samples, What Matters Most?* Washington, DC: PewResearch.

Techniques d'enquête, décembre 2022
Vol. 48, n° 2, pp. 369-392
Statistique Canada, n° 12-001-X au catalogue

Commentaires à propos de l'article « Inférence statistique avec des échantillons d'enquête non probabiliste » : La miniaturisation de la corrélation due à un défaut des données : une stratégie polyvalente de traitement des échantillons non probabilistes

Xiao-Li Meng[1]

Résumé

Il n'est pas possible de tirer parti de la puissante *probabilité du plan* pour établir l'inférence fondée sur la randomisation à partir d'échantillons non probabilistes. Cela nous incite à exploiter une *probabilité divine* naturelle qui accompagne toute population finie. Dans cette perspective, un des paramètres principaux est la *corrélation due à un défaut des données (cdd)*, qui est la corrélation de la population finie sans modèle entre l'indicateur d'inclusion de l'échantillon de la personne et la caractéristique de la personne échantillonnée. Un mécanisme de génération de données équivaut à un échantillonnage probabiliste, en ce qui concerne l'effet de plan, si et seulement si la *cdd* correspondante est de l'ordre (stochastique) $N^{-1/2}$, où N est la taille de la population (Meng, 2018). Par conséquent, les méthodes d'estimation linéaire valides existantes pour les échantillons non probabilistes peuvent être converties en plusieurs stratégies de miniaturisation de la *cdd* jusqu'à l'ordre $N^{-1/2}$. Les méthodes quasi fondées sur le plan permettent d'accomplir cette tâche en réduisant la variabilité entre les N propensions d'inclusion au moyen d'une pondération. L'approche fondée sur un modèle de superpopulation permet d'atteindre le même objectif par la réduction de la variabilité des caractéristiques des N personnes en les remplaçant par leurs résidus issus d'un modèle de régression. Les estimateurs doublement robustes doivent la propriété dont ils portent le nom au fait qu'une corrélation est nulle chaque fois qu'une des variables corrélées est constante, quelle qu'elle soit. Comprendre les points communs de ces méthodes au moyen de la *cdd* nous aide à voir clairement la possibilité d'une « robustesse plus que double », c'est-à-dire une estimation valide qui ne dépend pas de la pleine validité du modèle de régression ni de la propension d'inclusion estimée, qui ne sont garanties ni l'une ni l'autre parce que les deux reposent sur la *probabilité du procédé*. Les renseignements générés par la *cdd* incitent également à un *sous-échantillonnage de contrebalancement*, une stratégie visant à créer une miniature de la population à partir d'un échantillon non probabiliste, et comportant un compromis de qualité et de quantité favorable parce que les erreurs quadratiques moyennes sont beaucoup plus sensibles à la *cdd* qu'à la taille de l'échantillon, en particulier pour les populations de grande taille.

Mots-clés : Biais de non-réponse; estimateurs d'enquête assistés par un modèle; indice de défaut des données; inférence fondée sur le plan de sondage; probabilité divine; probabilité du plan; probabilité du procédé.

1. Distinguer la probabilité du plan, la probabilité divine et la probabilité du procédé

1.1 Que peuvent indiquer les statistiques et les statisticiens au sujet des échantillons non probabilistes ?

Le traitement des échantillons non probabilistes est une affaire délicate, particulièrement pour les statisticiens. Ceux qui estiment que les statistiques reposent sur le raisonnement probabiliste et l'inférence peuvent se demander si les statistiques ont quelque chose d'utile à offrir au monde non probabiliste. Bien

1. Xiao-Li Meng, Département de statistique, Université Harvard, Cambridge, MA 02138. Courriel : meng@stat.harvard.edu.

que ce questionnement puisse refléter une certaine ignorance, voire de l'hostilité à l'égard des statistiques, du point de vue conceptuel, il mérite une introspection et une extrospection de la part des statisticiens. À quel genre de probabilité faisons-nous référence quand l'échantillon est non probabiliste ? Toute la théorie et toutes les méthodes d'échantillonnage probabiliste reposent sur le caractère aléatoire introduit par de puissants mécanismes d'échantillonnage, qui permettent d'obtenir ensuite le magnifique cadre inférentiel fondé sur le plan sans qu'il soit nécessaire de *concevoir* que quoi que ce soit d'autre soit aléatoire (Kish, 1965; Wu et Thompson, 2020; Lohr, 2021). Quand cette puissance et cette beauté nous sont enlevées, que reste-t-il aux statisticiens ?

La réponse philosophique de certains statisticiens serait de rejeter complètement la question en déclarant qu'il n'existe pas d'échantillon probabiliste dans le monde réel. (Andrew Gelman m'a rappelé cette opinion quand je lui ai demandé de commenter la présente étude; voir une analyse sur le sujet dans https://statmodeling.stat.columbia.edu/2014/08/06.) Quand les données arrivent à notre bureau ou dans notre disque dur, même le plan d'échantillonnage probabiliste le mieux conçu est compromis par les imperfections de l'exécution, qui vont des défauts (incontrôlables) dans les bases de sondage aux non-réponses à différentes étapes, en passant par les erreurs de mesure dans les réponses. En ce sens, la notion d'échantillon probabiliste est toujours théorique, de la même manière que l'hypothèse de l'efficience du marché en économie, qui offre un cadre mathématiquement élégant pour l'idéalisation et les approximations, mais qui ne doit jamais être prise au pied de la lettre (par exemple Lo, 2017).

L'article d'actualité du professeur Changbao Wu (Wu, 2022) fournit une réponse plus pratique, en présentant la façon dont les statisticiens ont traité les échantillons non probabilistes dans la longue littérature sur les enquêtes par échantillons et (bien entendu) les études d'observation, en particulier concernant l'inférence causale; voir Elliott et Valliant (2017) et Zhang (2019) pour obtenir deux aperçus complémentaires portant sur le même défi. Pour mieux comprendre l'utilité de la théorie des probabilités en présence d'échantillons non probabilistes, il est important de reconnaître (au moins) trois types de concepts probabilistes aux fins d'inférence statistique, qui sont présentés dans la section 1.2. Les échantillons non probabilistes permettent de n'en retirer qu'un seul des trois, ainsi ils contraignent en général les spécialistes à une plus grande dépendance aux deux autres.

Une fois ces questions conceptuelles clarifiées, les autres sections traitent d'une stratégie unifiée de traitement des échantillons non probabilistes. La section 2 porte sur une identité fondamentale pour l'erreur d'estimation, qui a mené à l'élaboration de la corrélation due à un défaut des données (Meng, 2018). La section 3 traite ensuite de la façon dont ce concept peut inspirer une stratégie unifiée. La section 4 présente la stratégie pour les configurations qp et ξp, respectivement, dans l'étude de Wu (2022). Dans la section 5, on applique ensuite la stratégie aux deux configurations simultanément pour donner un aperçu immédiat de la célèbre double robustesse, laquelle est examinée dans l'étude de Wu (2022). À partir de ce même concept, la section 6 repose sur *l'échantillonnage de contrebalancement* comme stratégie de remplacement de la pondération. La section 7 se termine par un appel général à traiter la théorie de l'échantillonnage probabiliste comme une aspiration plutôt que comme l'élément central de la recherche sur les enquêtes et l'échantillonnage.

1.2 Un trio de concepts de probabilité

Le premier des trois concepts nommés ci-dessous, la probabilité du plan, se passe d'explications. Il est au cœur de la théorie de l'échantillonnage et il a été réifié par la mise en œuvre pratique, aussi imparfaite qu'elle puisse être. Bien que la distinction entre les deux concepts suivants, la probabilité divine et la probabilité du procédé, puisse être plus nuancée, en particulier sur le plan pratique, leurs différences conceptuelles ne sont pas moins importantes que la distinction entre un paramètre et un estimateur. Comme il se doit, l'indicateur d'enregistrement ou d'inclusion des données, lequel représente une quantité importante dans la modélisation des échantillons non probabilistes, fournit une illustration concrète des trois concepts de probabilité; voir le paragraphe principal de la section 4.

Probabilité du plan. Les répliques randomisées (Craiu, Gong et Meng, 2022) sont un concept et un outil primordiaux pour les statistiques et les sciences en général. En concevant et en exécutant un mécanisme probabiliste pour générer des répliques randomisées, nous créons des données probabilistes qui peuvent directement servir à faire des énoncés inférentiels vérifiables. De plus, l'échantillonnage probabiliste dans les enquêtes, la randomisation dans les essais cliniques, les bootstraps aux fins d'évaluation de la variabilité, les tests de permutation pour les vérifications d'hypothèses et les simulations Monte Carlo pour le calcul sont tous des exemples de méthodes statistiques élaborées à partir de la probabilité du plan. Les échantillons non probabilistes, par définition, n'ont pas de probabilité de plan, du moins pas de probabilité de plan définie. Par conséquent, le terme « échantillons non probabilistes » devrait être considéré comme une formulation abrégée d'« échantillons sans concept de probabilité du plan défini ».

Il faut toutefois nous rappeler que les probabilités du plan peuvent reprendre une place importante, surtout pour les grands ensembles de données non probabilistes, comme les données administratives, en raison de l'adoption de la confidentialité différentielle (Dwork, 2008), par exemple par le Bureau du recensement des États-Unis (voir l'éditorial de Gong, Groshen et Vadhan, 2022, et le numéro spécial du *Harvard Data Science Review* qu'il présente). Les méthodes de confidentialité différentielle permettent d'introduire du bruit aléatoire bien conçu dans les données dans le but de protéger la confidentialité des données sans sacrifier indûment l'utilité des données. Comme la probabilité du plan utilisée pour l'échantillonnage probabiliste, le fait que le mécanisme d'introduction de bruit soit conçu par le curateur des données et qu'il soit rendu public permet de fournir la transparence essentielle pour que l'utilisateur des données produise une inférence statistique valide (Gong, 2022). La question de savoir comment analyser correctement les données non probabilistes en cas de protection différentielle de la vie privée est largement ouverte. Il est encore plus fascinant de connaître la façon de tenir compte des défauts existants dans les données non probabilistes lors de la conception de mécanismes de protection probabiliste pour la confidentialité des données, afin d'éviter d'ajouter du bruit inutile. Les lecteurs qui souhaitent avoir une vue d'ensemble des problèmes statistiques découlant de la confidentialité des données devraient consulter l'excellent article de synthèse de Slavkovic et Seeman (2022) sur le domaine de la confidentialité des données en général.

Probabilité divine. En l'absence de probabilité du plan pour procéder à l'inférence fondée sur la randomisation, nous nous appuyons habituellement sur le concept selon lequel les données à notre disposition sont la réalisation d'un mécanisme probabiliste génératif donné par la nature ou par Dieu, afin d'effectuer une inférence statistique (classique). (J'ai entendu le terme « modèle de Dieu » pendant ma formation de doctorat et je l'ai considéré comme une expression de la foi ou comme quelque chose qui échappe au contrôle de l'homme, plutôt que comme le reflet des croyances religieuses d'une personne. L'adjectif « divine » est adopté dans la présente étude avec une connotation similaire.) Nous le faisons que nous croyions ou non que le monde est intrinsèquement déterministe ou stochastique (par exemple voir David Peat, 2002; Li et Meng, 2021). Nous devons supposer cette probabilité divine principalement en raison de la nature restrictive du cadre probabiliste auquel nous sommes si habitués. Par exemple, pour invoquer l'hypothèse de répartition au hasard des données manquantes, nous devons évoquer un mécanisme probabiliste dans lequel le concept de « données manquantes au hasard » (Rubin, 1976) peut être mis en forme. Comme l'ont souligné Elliott et Valliant (2017), la méthode de quasi-randomisation, qui correspond au cadre *qp* de Wu (2022), « suppose que l'échantillon non probabiliste ait en fait un mécanisme d'échantillonnage probabiliste, bien que ce soit un mécanisme comportant des probabilités qui doivent être estimées selon des hypothèses d'identification ». Cela signifie que nous remplaçons la probabilité du plan par une probabilité divine à laquelle nous croyons et qui est alors généralement traitée comme la « vérité » ou du moins comme un paramètre.

Sur le plan conceptuel, nous devons donc reconnaître que l'hypothèse d'une sorte particulière de probabilité divine n'est pas involontaire, car sinon nous n'aurions pas besoin de dépendre de notre foi pour continuer à travailler. Ce n'est pas non plus toujours nécessaire. Toute population finie fournit un histogramme naturel pour toute caractéristique quantifiable, ou un tableau de contingence pour toute caractéristique classable de ses éléments et, par conséquent, elle induit une probabilité divine sans faire référence à un quelconque caractère aléatoire, conceptualisé ou réalisé, *si notre cible inférentielle est la population finie elle-même* (et non une superpopulation qui permet de la générer, par exemple). La méthode de la vraisemblance empirique tire parti de ce cadre de probabilité naturel, qui se révèle également fondamental pour quantifier la qualité des données au moyen de la corrélation due à un défaut des données (voir Meng, 2018). Zhang (2019), pour qui le critère unifié était fondé sur la même identité pour établir la corrélation due à un défaut des données, fait la même constatation; voir la section 2 ci-dessous.

Probabilité du procédé. La plupart des probabilités utilisées dans la modélisation statistique sont, de loin, des procédés servant à exprimer notre croyance, nos connaissances *a priori*, nos hypothèses, nos idéalisations, nos compromis, voire notre désespoir (comme quand on impose une distribution *a priori* pour assurer l'identifiabilité, puisque rien d'autre ne fonctionne). Bien que la littérature statistique ait toujours mis l'accent sur la modélisation de la réalité, nous sommes inévitablement contraints de réaliser une variété de simplifications, d'approximations et, parfois, de distorsions délibérées afin de composer avec des contraintes pratiques (par exemple l'utilisation de l'inférence variationnelle à des fins d'efficacité du calcul; voir Blei, Kucukelbir et McAuliffe, 2017). Par conséquent, pour bon nombre de ces probabilités

du procédé, rien ne contraint à ce qu'elles soient réalisables ni même mathématiquement cohérentes (par exemple l'emploi de distributions de probabilité conditionnelle incompatibles pour l'imputation à chaînes multiples; voir Van Buuren et Oudshoorn, 1999). Il n'est pas non plus facile de valider ces probabilités, ni même possible, comme Zhang (2019) l'a étudié et fait valoir dans un contexte d'échantillonnage non probabiliste, en particulier avec la méthode de modélisation de superpopulation, qui correspond au cadre ξp de Wu (2022). Néanmoins, les probabilités de procédé offrent le meilleur rendement pour procéder à des inférences statistiques. La méthode de quasi-randomisation et la modélisation de superpopulation reposent toutes deux sur des probabilités du procédé pour fonctionner, comme le montre Wu (2022) et comme il est expliqué en détail dans les sections 4 et 5 de la présente étude. L'absence de probabilité du plan ne peut que favoriser la progression des probabilités de procédé. Pour paraphraser une citation célèbre de Box, « tous les modèles sont faux, mais certains sont utiles ». Cela signifie que toutes les probabilités de procédé posent des problèmes, mais que certaines permettent de résoudre des problèmes.

1.3 La réduction des données inexactes menant à des paquets d'intelligence artificielle

Pour résumer, les concepts probabilistes sont plus nécessaires pour les échantillons non probabilistes que pour les échantillons probabilistes, précisément parce qu'ils n'ont pas de probabilité de plan. La difficulté de traitement des échantillons non probabilistes ne représente pas un nouveau défi pour les statisticiens. S'il y a quelque chose de nouveau, il s'agit de la quantité considérable de grands ensembles de données non probabilistes, comme les données administratives et les données provenant des médias sociaux, et du besoin accéléré de combiner plusieurs sources de données, qui sont pour la plupart intrinsèquement non probabilistes parce qu'elles ne sont pas recueillies à des fins d'inférence statistique (par exemple Lohr et Rao, 2006; Meng, 2014; Buelens, Burger et van den Brakel, 2018; Beaumont et Rao, 2021). Contrairement à la croyance courante, la grande taille des « mégadonnées » peut empirer notre inférence, en raison du « paradoxe des mégadonnées » (Meng, 2018; Msaouel, 2022), quand nous ne tenons pas compte de la qualité des données lors de l'évaluation des erreurs et des incertitudes dans nos analyses; voir la section 6.1.

Il est donc plus urgent que jamais de sensibiliser et d'informer à grande échelle à propos de l'importance cruciale de la qualité des données et de la façon dont nous pouvons utiliser les méthodes et les théories statistiques pour contribuer à réduire le défaut des données. Dans le cadre de la présente étude, la préoccupation centrale va au-delà de la mise en garde habituelle au sujet du dicton « à données inexactes, résultats erronés ». En effet, si des données sont reconnues comme étant mauvaises, elles seront probablement traitées comme telles (probablement, mais pas toujours, car, comme Andrew Gelman me l'a rappelé, « de nombreux chercheurs croient fermement à la *procédure* plutôt qu'à la *mesure*, et pour ces personnes, le plus important consiste à respecter les règles, et non pas à s'interroger sur la provenance de leurs données »). L'objectif est d'empêcher que des données inexactes mènent à un paquet d'intelligence artificielle (Meng, 2021), c'est-à-dire que des données de faible qualité sont traitées automatiquement au

moyen de procédures génériques afin de créer un paquet d'intelligence artificielle esthétiquement attrayant qui serait vendu à des consommateurs non informés, ou pire encore, à ceux qui cherchent des « données probantes » pour induire en erreur ou désinformer. De toute évidence, le traitement adéquat des échantillons non probabilistes ne résout pas tous les problèmes de qualité des données, mais il contribue grandement à régler le problème de plus en plus courant et nuisible du manque de contrôle de la qualité des données en science des données.

Je remercie donc le professeur Changbao Wu pour son analyse opportune et complète des « incontournables » de la grande usine de fabrication de saucisses qu'est le traitement d'échantillons non probabilistes. Elle apporte des vues beaucoup plus détaillées et nuancées que celles de l'analyse générale d'Elliott et Valliant (2017), qui illustrait de manière très fine de nombreuses formes d'échantillons non probabilistes ainsi que leurs méfaits. Elle présente également des jalons théoriques et méthodologiques qui nous permettront de mieux apprécier ceux exposés dans l'analyse intellectuelle de Zhang (2019), qui met les statisticiens et les scientifiques des données en général au défi de mieux comprendre la qualité, ou plutôt l'absence de qualité, des produits que nous fabriquons et promouvons. Ensemble, ces trois articles de synthèse donnent un aperçu informé à quiconque souhaiterait se joindre aux spécialistes cherchant à traiter les difficultés de plus en plus grandes causées par les données non probabilistes. Le mieux serait peut-être de commencer par l'étude d'Elliott et de Valliant (2017), qui brosse un portrait général de la situation, puis de lire l'étude de Wu (2022), qui s'attache principalement à exposer les méthodologies, et de terminer par l'étude de Zhang (2019), qui suscite des réflexions plus profondes sur certains défis particuliers. Pour en savoir plus sur d'autres méthodes courantes de traitement des échantillons non probabilistes, comme la modélisation multiniveau et la poststratification, nous invitons les lecteurs à lire Gelman (2007), Wang, Rothschild, Goel et Gelman (2015) et Liu, Gelman et Chen (2021).

En tant que chercheur et enseignant, je me suis penché sur ce domaine, mais j'ai souvent été frustré par le manque de temps ou d'énergie pour l'étudier en profondeur. Je suis donc particulièrement reconnaissant envers le rédacteur en chef, Jean-François Beaumont, de m'avoir invité à contribuer à ce que les messages du professeur Wu soient clairement entendus : les données ne peuvent pas être traitées comme si elles étaient représentatives à moins que les données observées ne soient véritablement des échantillons probabilistes (ce qui est extrêmement rare). De nombreux remèdes ont été proposés et essayés, mais il faut en élaborer et en évaluer bien davantage. Parmi ces remèdes, le concept de corrélation due à un défaut des données est une mesure générale prometteuse qui doit être étudiée et élaborée, comme nous le montrons ci-dessous.

2. Une identité déterministe de population finie pour l'erreur réelle

Pour démontrer la richesse du cadre de population finie, considérons l'estimation de la moyenne de la population, indiquée par $\bar{G}$, de $\{G_i = G(X_i): i \in \mathcal{N}\}$, où $\mathcal{N} = \{1, \ldots, N\}$ indexe une population finie, et les X_i sont des données recueillies sur une personne i. Pour chaque i, supposons que $R_i = 1$ si G_i (ou plutôt X_i) est enregistré dans notre échantillon, et que $R_i = 0$ sinon. La taille de l'échantillon est alors

$n_R = \sum_{i=1}^{N} R_i$. Nous insistons sur le fait qu'il s'agit d'un indicateur global, qui peut (et devrait) être décomposé en $R_i = r_i^{(1)}, \ldots, r_i^{(J)}$, quand la collecte des données consiste en J étapes (par exemple $r_i^{(1)}$ indique si la i^{e} personne a été échantillonnée et $r_i^{(2)}$, si la personne a répondu ou non une fois qu'elle a été échantillonnée).

Supposons que $\{W_i, i \in S\}$ est un ensemble de poids à déterminer où l'indice est paramétré à $S = \{i : R_i = 1\}$, de sorte que $\sum_{i \in S} W_i > 0$. Supposons que $\bar{G}_W$ est la moyenne pondérée de l'échantillon, qu'on peut exprimer de trois façons :

$$\bar{G}_W = \frac{\sum_{i \in S} W_i G_i}{\sum_{i \in S} W_i} = \frac{\sum_{i=1}^{N} R_i W_i G_i}{\sum_{i=1}^{N} R_i W_i} = \frac{\mathrm{E}_I(\tilde{R}_I G_I)}{\mathrm{E}_I(\tilde{R}_I)}, \tag{2.1}$$

où $\tilde{R}_I = R_I W_I$, et E_I est prise par rapport à la distribution uniforme de l'indice paramétré $\mathcal{N}$. La première expression dans l'équation (2.1) définit simplement une moyenne pondérée de l'échantillon. À l'aide de R_i, la deuxième expression permet de transformer les moyennes de l'échantillon en moyennes de population finie. Cette nouvelle expression banale est fondamentale parce qu'elle explique le rôle de R_i dans l'influence sur le comportement de $\bar{G}_W$ en tant qu'estimateur de $\bar{G}$. La troisième expression révèle une probabilité divine au moyen de I, la variable de l'indice de population finie (IPF), grâce au fait que le calcul de la moyenne revient à prendre en compte l'espérance d'un indice aléatoire uniformément distribué I. Tous les moments de population finie peuvent alors être exprimés au moyen de E_I.

En particulier, nous pouvons exprimer l'erreur réelle de $\bar{G}_W$ par l'identité suivante, dont la première expression remonte à Hartley et Ross (1954), qui l'ont utilisée pour exprimer les biais dans des estimateurs par le ratio. La deuxième expression a été donnée dans Meng (2018), mais elle comportait une expression légèrement différente (mais équivalente) :

$$\bar{G}_W - \bar{G} = \frac{\mathrm{Cov}_I(\tilde{R}_I, G_I)}{\mathrm{E}_I[\tilde{R}_I]} = \rho_{\tilde{R},G} \times \sqrt{\frac{N - n_W}{n_W}} \times \sigma_G. \tag{2.2}$$

Dans cette équation, $\rho_{\tilde{R},G} = \mathrm{Corr}_I(\tilde{R}_I, G_I)$ est la *corrélation de population finie* entre $\tilde{R}_I$ et G_I, σ_G^2 est la variance de la population finie de G_I et n_W est la taille d'échantillon efficace en raison de l'utilisation des poids (Kish, 1965),

$$n_W = \frac{n_R}{1 + \mathrm{CV}_W^2}, \tag{2.3}$$

où CV_W est le coefficient de variation (c'est-à-dire l'écart-type ou la moyenne) de $\{W_i, i \in S\}$.

L'expression de l'équation (2.2) est une identité algébrique parce qu'elle se vérifie pour toute instance de $\{(G_i, R_i W_i), i \in \mathcal{N}\}$. Ainsi, aucune hypothèse de modèle n'est imposée, pas même l'hypothèse que R (ou toute quantité) est aléatoire, ce qui rappelle le commentaire de Mary Thompson cité dans l'étude de Wu (2022), selon lequel « le fait que l'indicateur d'inclusion dans l'échantillon R est une variable aléatoire est en soi une hypothèse ». La seule exigence est que la valeur de G_i enregistrée soit identique à celle de G_i dans la population cible. (Il faut mentionner toutefois que cette exigence comporte deux

éléments : 1) il n'y a pas de surdénombrement, c'est-à-dire que chaque personne dans l'échantillon appartient à la population cible, par exemple aucun électeur non admissible n'est sondé quand la population cible est celle des électeurs admissibles; 2) il n'y a pas d'erreur de mesure. Il peut y avoir des extensions de cas comportant des erreurs de mesure, mais elles ne sont pas examinées dans la présente étude.) Quand nous utilisons des poids égaux, les trois facteurs du membre de droite de l'équation (2.2) représentent, respectivement (de gauche à droite), le défaut des données, l'insuffisance des données et la difficulté du problème, comme l'explique Meng (2018) et comme l'illustrent en détail Bradley, Kuriwaki, Isakov, Sejdinovic, Meng et Flaxman (2021) dans le contexte des enquêtes sur la vaccination contre la COVID-19.

En particulier, quand tous les poids sont égaux, $\rho_{\tilde{R},G}$ est appelée *corrélation due à un défaut des données* (*cdd*) dans Meng (2018) parce qu'elle permet de mesurer le manque de représentativité de l'échantillon en saisissant la dépendance de l'indicateur d'inclusion ou d'enregistrement aux caractéristiques : plus la dépendance est élevée, plus la moyenne de l'échantillon est biaisée quand il faut estimer les moyennes de population. Quand l'on utilise les stratégies de base de l'échantillonnage probabiliste ou de la pondération de probabilité inverse, la *cdd* est nulle en moyenne parce que $\mathrm{E}(W_i R_i) = 1,$ et elle est de l'ordre $O_p(N^{-1/2})$ parce qu'il s'agit essentiellement d'une moyenne de N termes indépendants (Meng, 2018). Notre objectif général est donc de ramener la *cdd* à $O_p(N^{-1/2})$ pour les échantillons non probabilistes, ce que nous appellerons « miniaturiser la *cdd* » parce que $N^{-1/2}$ est généralement un nombre minuscule dans la pratique.

Quand nous utilisons des poids, le premier terme $\rho_{\tilde{R},G}$ saisit le défaut des données qui existe toujours après l'ajustement de la pondération, puisqu'aucun poids n'est parfait en pratique. L'identité dans l'équation (2.2) montre l'incidence des poids sur la qualité et la quantité des données. L'incidence sur la taille d'échantillon efficace *nominale* n_W n'est jamais positive, car $n_W \le n_R$ comme on peut le voir dans l'équation (2.3). Par ailleurs, l'exactitude de l'équation (2.3) révèle qu'en fait, cette expression bien connue n'est pas une approximation (ce qui est souvent attribué à Kish, 1965), mais une formule exacte de réduction de la taille de l'échantillon en raison de la pondération *si la pondération n'a pas d'incidence sur la cdd*. Cependant, la pondération peut avoir une incidence positive importante sur la réduction de l'erreur globale quand on choisit judicieusement des poids pour diminuer considérablement la *cdd*, bien qu'apparemment cela se fasse au prix de $n_W < n_R$. Bien entendu, c'est exactement ce que vise le cadre de quasi-randomisation dont il est question ci-dessous. Plus important encore, l'équation (2.2) donne un aperçu unifié de la variété des méthodes examinées dans l'étude de Wu (2022), notamment une explication intuitive de la propriété doublement robuste, qui fait l'objet d'une attention accrue aux fins d'intégration des sources de données, concernant à la fois des échantillons probabilistes et non probabilistes (par exemple Yang, Kim et Song, 2020).

En effet, Zhang (2019, section 3.1) a utilisé la première expression dans l'équation (2.2) pour définir une hypothèse de non-informativité asymptotique non paramétrique unifiée, qui exige que le numérateur $\mathrm{Cov}_I(\tilde{R}_I, G_I)$ passe à zéro, tout en gardant le dénominateur $\mathrm{E}_I[\tilde{R}_I]$ positif, quand $N \to \infty$. Cette

unification a permis à Zhang (2019) d'évaluer la méthode de quasi-randomisation et la modélisation par la régression au moyen d'un critère commun. Comme le montre la section 3, le cadre de la *cdd* fait écho à cette unification. La section 4 met plutôt l'accent sur le message général de Zhang (2019). La section 5 traite d'un autre avantage simple de la formulation de la *cdd* qui fournit une explication immédiate de la célèbre double robustesse. La section 6 aborde quant à elle le domaine beaucoup plus difficile de l'élaboration d'un sous-échantillon plus représentatif à partir d'un grand échantillon non représentatif, soit un compromis précieux, puisque la qualité des données est beaucoup plus importante que la quantité (Meng, 2018), comme nous le voyons brièvement ci-dessous.

3. Une stratégie unificatrice fondée sur la corrélation due à un défaut des données

Dans la configuration de Wu (2022), pour chaque personne i, nous avons un ensemble de caractéristiques $A_i = \{y_i, \mathbf{x}_i\}$, où y est la caractéristique d'intérêt et $\mathbf{x}$ est une variable auxiliaire, ce qui est utile de deux façons. Premièrement, la réduction du biais d'échantillonnage attribuable à l'échantillonnage non probabiliste devient possible quand le mécanisme non probabiliste peut être (entièrement) expliqué par $\mathbf{x}$. Deuxièmement, en tirant parti des relations entre y_i et $\mathbf{x}_i$, nous pouvons améliorer l'efficacité de notre estimation. Comme point de départ, Wu (2022) suppose que nous avons deux sources de données disponibles, que nous désignons au moyen de deux indicateurs d'enregistrement, R et R^*. La source principale des données est un échantillon non probabiliste dans lequel nous observons à la fois y_i et $\mathbf{x}_i$ lorsque $i \in S \equiv \{i : R_i = 1\}$, mais l'indicateur d'enregistrement R_i est déterminé par un mécanisme non contrôlé par une probabilité de plan (connue). La deuxième source est (supposée être) un échantillon probabiliste dans lequel nous observons seulement $\mathbf{x}_i$ lorsque $i \in S^* \equiv \{i : R_i^* = 1\}$. Ce deuxième échantillon fournit des renseignements pour estimer des renseignements auxiliaires sur la population, qui sont utiles dans l'estimation des quantités de la population de y, par exemple sa moyenne. Par conséquent, cette configuration est étroitement liée à la configuration où $S \cup S^* = \mathcal{N}$; voir Tan (2013).

Pour toute fonction $m(\mathbf{x})$, considérons que $z_i = y - m(\mathbf{x}_i), i \in \mathcal{N}$. Il est clair que nous pouvons estimer la moyenne de la population $\bar{y}_N = \mathrm{E}_I(y_I)$ en estimant $\bar{z} = \mathrm{E}_I(z_I)$ et $\bar{m} = \mathrm{E}_I\left[m(\mathbf{x}_I)\right]$. À partir du deuxième échantillon, $\bar{m}$ peut être estimé sans biais puisqu'il ne concerne que $\mathbf{x}$. Nous pouvons alors nous concentrer sur l'estimation de $\bar{z}$, tout en reconnaissant qu'une méthode plus fondée sur des principes nous amènerait à établir un modèle de probabilité ou un modèle bayésien pour estimer conjointement toutes les quantités inconnues (Pfeffermann, 2017). L'application de l'identité dans l'équation (2.2) où $G = z$ nous indique alors que notre tâche centrale consiste à choisir le poids $\{W_i, i \in S\}$ ou la fonction m pour miniaturiser la *cdd*, $\rho_{\tilde{R},z}$. Dans la présente étude, il est plus facile de tout expliquer au moyen de la covariance

$$c_{\tilde{R},z} \equiv \mathrm{Cov}_I(\tilde{R}_I, z_I) = \mathrm{Cov}_I(W_I R_I, y_I - m(\mathbf{x}_I)) = \frac{1}{N}\sum_{i=1}^{N} W_i R_i (z_i - \bar{z}) \tag{3.1}$$

au lieu de la corrélation $\rho_{\tilde{R},z}$ parce que $\mathrm{Cov}_I(\tilde{R}_I, z_I)$ est une fonction bilinéaire dans R_I et z_I. Toutefois, sur le plan théorique et à des fins de modélisation, $\rho_{\tilde{R},z}$ est plus attrayante en raison de sa normalisation; voir les sections 6 et 7.

L'expression dans l'équation (3.1) nous indique immédiatement la façon de la rendre nulle dans les espérances sur le plan opérationnel, et dans quel sens conceptuel. Quelle que soit la probabilité que nous imposions à R_i (à préciser dans les dernières sections), supposons que $\pi_i = \Pr(R_i = 1 \mid \mathbf{A})$, que nous assumons dépendra de A_i seulement. Alors, la linéarité de l'opérateur de covariance implique que la covariance moyenne pour ce qui est du caractère aléatoire dans R_i est obtenue par

$$\mathrm{E}[c_{\tilde{R},z} \mid \mathbf{A}] = \mathrm{Cov}_I\left(W_I\pi_I, y_I - m(\mathbf{x}_I)\right), \tag{3.2}$$

où $\mathbf{A} = \{A_i, i \in \mathcal{N}\}$. De même, si l'on est prêt à postuler un modèle conjoint pour $\{(R_i, y_i), i \in \mathcal{N}\}$ conditionné sur $\mathbf{X}$ sous forme d'indépendance $\Pi_{i=1}^{N} P\left(R_i, y_i \mid \mathbf{x}_i\right)$, alors

$$\mathrm{E}[c_{\tilde{R},z} \mid \mathbf{X}] = \mathrm{Cov}_I\left(W_I\pi_I, \mathrm{E}(y_I \mid \mathbf{x}_I) - m(\mathbf{x}_I)\right). \tag{3.3}$$

De façon très intuitive, on peut assurer une covariance ou une corrélation nulle entre deux variables en faisant de l'une des deux une constante. Les deux choix mèneraient alors respectivement à la méthode de quasi-randomisation si l'on fait de $W_I\pi_I \propto 1$ et à la méthode de la superpopulation si l'on fait de $\mathrm{E}[y_I \mid \mathbf{x}_I] - m(\mathbf{x}_I)$ une constante (par exemple zéro). La double robustesse naît du fait que l'une ou l'autre suffise à rendre la covariance nulle (dans le modèle conjoint), puisque la variable n'a pas d'importance. Cependant, il est évident que ce ne sont pas les seules méthodes permettant d'obtenir une corrélation ou une covariance nulle, ou une double robustesse, comme le soulignent Kang et Schafer (2007) dans leur volonté de démystifier la double robustesse (Robins, Rotnitzky et Zhao, 1994; Robins, 2000; Scharfstein, Rotnitzky et Robins, 1999). La question est aussi abordée dans l'étude de Tan (2007, 2010), qui porte sur plusieurs estimateurs et leur comparaison, y compris ceux qui correspondent seulement à la méthode de quasi-randomisation ou à la méthode de la superpopulation. Certains estimateurs sont doublement robustes.

En effet, parce que la formule (2.2) est une identité pour l'erreur réelle, tout estimateur asymptotiquement sans biais (linéaire) de la moyenne de population doit impliquer que la *cdd* correspondante est asymptotiquement sans biais pour les valeurs nulles, et vice versa, pour ce qui est du caractère aléatoire dans R ou dans $\{R, y\}$. Cependant, il est possible que la *cdd* soit asymptotiquement sans biais pour les valeurs nulles, sans supposer que le modèle est correctement précisé, comme l'illustre un exemple dans la section 5. (Cette « robustesse plus que double » est différente de la « robustesse multiple » de Han et Wang (2013), qui doit encore supposer la validité d'au moins un des modèles postulés.) Ces deux observations donnent à penser que toute stratégie générale suffisante et nécessaire qui assure des estimateurs asymptotiquement convergents ou sans biais (linéaires) pour la moyenne de la population équivaudrait à miniaturiser la *cdd*.

À titre d'exemple d'aperçu unifié qui autrement ne serait pas aussi intuitif, l'expression de l'équation (3.2) donne à penser que nous devrions inclure notre estimation de π_I comme élément du prédicteur dans le modèle de régression $m(\mathbf{x}_I)$, puisque cela peut aider à réduire la corrélation entre $W_I\pi_I$ et $z_I = y_I - m(\mathbf{x}_I)$, en particulier quand nous utilisons des poids constants W_I. En général, il est difficile de motiver l'utilisation de $\hat{\pi}_I$ comme prédicteur pour y uniquement du point de vue de la régression, surtout quand nous supposons que y et R sont indépendants étant donné $\mathbf{x}$ (ce qui est habituellement une condition nécessaire pour continuer, comme nous l'expliquons dans la section suivante). Cependant, l'expression de l'équation (3.2) nous indique que pour estimer la moyenne de y, il n'est pas absolument nécessaire d'ajuster le bon modèle de régression $m(\mathbf{x})$. En fait, il suffit de s'assurer que le « résidu » z_I est autant non corrélé à $W_I\pi_I$ quand I varie. Cependant, il est extrêmement important de reconnaître qu'il ne suffit pas d'assurer une corrélation nulle ou faible seulement dans les données observées, car $\mathrm{Cov}_I(W_I\pi_I, z_I \mid R_I = 1)$ nous informe peu sur $\mathrm{Cov}_I(W_I\pi_I, z_I \mid R_I = 0)$. Dans la configuration de Wu (2022), notre capacité à extrapoler de $R_I = 1$ à $R_I = 0$ dépend de la disponibilité des données auxiliaires (indépendantes) indexées par $R_I^* = 1$, ce qui nous permet d'observer certains x_I pour lesquels $R_I = 0$.

La littérature montre les avantages présentés par la stratégie consistant à inclure des estimations de la propension comme prédicteur. Par exemple, Little et An (2004) ont inclus le logit de $\hat{\pi}$ dans leur modèle d'imputation et ils ont constaté que cette inclusion a amélioré la robustesse de la moyenne imputée par rapport à la spécification erronée du modèle d'imputation. Zhang et Little (2009) et Tan, Flannagan et Elliott (2019) ont mis au point cette méthode et ils l'ont améliorée davantage; ils ont utilisé l'expression « robuste au carré » pour souligner la robustesse accrue. Dans un article plus récent portant sur une stratégie similaire pour les échantillons non probabilistes, Liu et coll. (2021) ont montré qu'il était important d'inclure la propension estimée $\hat{\pi}_i$ « comme prédicteur » dans $m(x, \hat{\pi})$ (en utilisant la notation de la présente étude). De plus, dans la littérature sur l'estimation par la méthode du maximum de vraisemblance ciblée (EMVC) pour les modèles semi-paramétriques de traitement des données non probabilistes (van der Laan et Rubin, 2006; Luque-Fernandez, Schomaker, Rachet et Schnitzer, 2018) (voir aussi Scharfstein et coll., 1999; Tan, 2010), les variables $R_I / \hat{\pi}_I$ et $(1 - R_I) / (1 - \hat{\pi}_I)$ sont appelées *covariables intelligentes* et sont utilisées dans les modèles de régression pour y_I. Les mises en œuvre et les théories de l'EMVC et celles de l'EMVC collaborative liée (van der Laan et Gruber, 2009 et 2010), sont mathématiquement plus impliquées que celles en contexte de population finie, comme nous le verrons plus bas, mais les résultats tirés des équations (3.2) et (3.3) peuvent nous permettre d'avoir des raisonnements intuitifs utiles sur la compréhension de l'essence de ces méthodes.

4. Quasi-randomisation *ou* mises en œuvre de la méthode de la superpopulation

En bref, la méthode de quasi-randomisation vise à faire de $W_I\pi_I$ une variable constante (induite par l'indice de population finie I). Lorsque notre échantillon est véritablement sélectionné selon un plan

probabiliste intentionnel, alors $\pi_i = \Pr(R_i = 1 \mid \mathbf{x}_i)$ lorsque $i \in \mathcal{N}$ est une probabilité du plan sans y_i, mais elle peut dépendre de $\mathbf{x}_i$, par exemple quand $\mathbf{x}_i$ comprend une variable de stratification. Quand la probabilité du plan n'est pas disponible, nous devons d'abord invoquer une probabilité divine. Il pourrait s'agir d'une probabilité naturelle donnée par la population finie, comme la propension $\pi_i = \Pr_I(R_I = 1 \mid A_I = A_i)$ induite par l'IPF, où $A_i = \{y_i, \mathbf{x}_i\}$, ou d'une probabilité imaginaire de superpopulation comme celle où R_i est généré indépendamment de $\text{Ber}(\pi_i)$, où $\pi_i = \Pr(R_i = 1 \mid A_i) > 0$. Cette hypothèse de positivité est nécessaire si la population finie est précisée au préalable, ou si son imposition définit la population finie qui peut être étudiée. (Cette considération est pertinente en pratique, comme dans le cas des sondages électoraux, où la population finie n'est pas toujours précisée au préalable, pas même en théorie.) Étant donné que ces probabilités divines sont inconnues et qu'elles constituent notre paramètre, nous devons supposer certaines probabilités de procédé, comme par l'intermédiaire d'un modèle linéaire généralisé $\pi_i = g(y_i, \mathbf{x}_i)$ pour poursuivre, même si nous ne croyons pas vraiment en un choix particulier de g.

Aux fins de notre analyse actuelle, nous supposons que notre probabilité divine est donnée par le modèle de superpopulation de Bernoulli, soit $n_R = \sum_{i=1}^{N} R_i$, et $\tilde{p}(\mathbf{A}) = \Pr(n_R > 0 \mid \mathbf{A}) = 1 - \Pi_{i \in N}(1 - \pi_i)$, où $\mathbf{A} = \{A_i, i \in \mathcal{N}\}$. Étant donné que, dans ce cas-ci, R_i est contrôlé par une probabilité divine, la taille de l'échantillon n_R n'est plus une variable du plan de sondage qui conditionne notre schéma de réplication; ce n'est généralement plus une statistique ancillaire. Néanmoins, nous devrions imposer comme condition que $n_R > 0$, ce qui représente une exigence universelle pour l'élaboration d'estimations fondées sur des données pour $\bar{G}$. Heureusement, cette condition ne crée pas de complications mathématiques à la simplicité accordée par l'indépendance de $\pi_i, i \in \mathcal{N}$ comme fonctions de A_i. Cela est dû au fait que $\tilde{\pi}_i(\mathbf{A}) \equiv \Pr(R_i = 1 \mid \mathbf{A}, n_R > 0) = \pi_i \,/\, \tilde{p}(\mathbf{A})$, mais la constante de normalisation $\tilde{p}(\mathbf{A})$, qui dépend de la totalité de $\mathbf{A}$, n'est pas pertinente pour les besoins de la présente étude, comme l'attribution de poids proportionnels à $\tilde{\pi}_i^{-1}(\mathbf{A})$.

Par conséquent, selon cette probabilité divine, qui correspond au (véritable modèle pour le) scénario du modèle q dans l'article de Wu (2022), nous obtenons, pour chaque W_I choisi selon l'équation (3.1),

$$\begin{aligned} \mathrm{E}(c_{\tilde{R},z} \mid \mathbf{A}, n_R > 0) &= \mathrm{Cov}_I(W_I \mathrm{E}[R_I \mid \mathbf{A}, n_R > 0], y_I - m(\mathbf{x}_I)) \\ &= \tilde{p}^{-1}(\mathbf{A})\, \mathrm{Cov}_I(W_I \pi_I, y_I - m(\mathbf{x}_I)), \end{aligned} \tag{4.1}$$

où E est choisi par rapport à la probabilité divine (inconnue) en lien avec R_I (pour une valeur I fixe). Ensuite, que nous voulions ou non garantir une espérance nulle dans l'équation (3.2) ou l'équation (4.1), nous imposerons $W_I \pi_I \propto 1$, c'est-à-dire $W_I \propto \pi_I^{-1}$, la pondération de probabilité inverse bien connue. Par conséquent, si notre modèle postulé q nous permet de saisir π_i de façon fiable dans la réalité, alors $c_{\tilde{R},z} = O_p(N^{-1/2})$ parce qu'il a une moyenne nulle (pour ce qui est de la probabilité divine), et qu'il s'agit d'une moyenne pondérée de N variables de Bernoulli essentiellement indépendantes, comme on le voit dans l'équation (3.1).

Il s'agit d'une approche axée sur la randomisation parce qu'elle traite toutes les valeurs des caractéristiques de la population finie $\mathbf{A}$ comme étant fixes, et les répliques hypothétiques sont générées uniquement au moyen de réalisations répétées de l'indicateur d'enregistrement R_I. Bien sûr, les valeurs de $\{\pi_i, i \in \mathcal{N}\}$ sont généralement inconnues, et pire encore, elles ne peuvent pas être estimées à partir d'un échantillon non probabiliste sans que l'on formule d'autres hypothèses. Pour poursuivre, nous posons des hypothèses comme l'hypothèse de répartition au hasard des données manquantes, c'est-à-dire $\Pr(R_i = 1 \mid A_i) = \Pr(R_i = 1 \mid \mathbf{x}_i)$, et l'exigence d'un échantillon auxiliaire afin que nous ayons certaines valeurs de $\mathbf{x}_i$ au moyen de $R_i = 0$. Nous avons également des choix sur la façon d'estimer la propension d'inclusion $\pi_i = \Pr(R_i = 1 \mid \mathbf{x}_i)$, de façon paramétrique ou non paramétrique. Ces hypothèses, ces exigences et ces méthodes d'estimation sont toutes essentielles à la mise en œuvre pratique, comme Wu (2022) en a fait l'évaluation et l'examen avec soin; voir aussi, dans l'article de Tan (2010), une comparaison détaillée des différentes stratégies d'estimation. Néanmoins, l'idée générale des méthodes de quasi-randomisation est de choisir W_I pour libérer $\tilde{R}_I = W_I R_I$ de I en espérance selon les répliques hypothétiques postulées, pour regagner la liberté garantie par l'échantillonnage probabiliste.

De façon complémentaire, les méthodes de superpopulation visent à miniaturiser $c_{\tilde{R}, z}$ en rendant l'autre variable dans $c_{\tilde{R}, z}$, c'est-à-dire z_I, exempte de I en espérance, mais selon un schéma de réplication hypothétique différent. Dans ce cas-ci, l'idée est de choisir un $m(\mathbf{x}_i)$ qui soit une bonne approximation de y_i de sorte que le résidu $z_i = y_i - m(\mathbf{x}_i)$ sera nul en espérance conditionnellement à $\mathbf{x}$. En général, on le fait en considérant un modèle conjoint pour $\{R_i, y_i\}$ étant donné $\mathbf{x}_i$ et avec un modèle de régression particulier $\xi(y \mid \mathbf{x})$, au moyen de la notation de Wu (2022). Il est important d'admettre que, bien que nous ne précisions que le modèle de régression y_i étant donné $\mathbf{x}_i$, nous devons inclure R_i dans les répliques afin de saisir la dépendance possible de R_i à la totalité de $A_i = \{y_i, \mathbf{x}_i\}$, ce qui est la principale préoccupation pour les échantillons non probabilistes. En effet, c'est cette spécification conjointe qui permet l'adoption de l'hypothèse de répartition au hasard des données manquantes pour réduire $P(y_i \mid \mathbf{x}_i, R_i) = P(y_i \mid \mathbf{x}_i)$, ce qui nous permet de nous concentrer sur la spécification d'un modèle de régression unique $\xi(y_i \mid \mathbf{x}_i)$ pour les personnes observées et non observées. Par conséquent, quand nous écrivons E_ξ, nous entendons l'espérance pour ce qui est de

$$P(R_i, y_i \mid \mathbf{x}_i) \;=\; P(R_i \mid \mathbf{x}_i)\, P(y_i \mid R_i, \mathbf{x}_i) \;=\; \pi_i^{R_i}\, (1-\pi_i)^{1-R_i}\, \xi(y_i \mid \mathbf{x}_i), \tag{4.2}$$

où $\pi_i = \Pr(R_i = 1 \mid \mathbf{x}_i)$ n'est pas précisé, contrairement à ce qui se passe dans la méthode de quasi-randomisation.

Il s'ensuit alors que, conditionnellement à $\mathbf{X} = \{\mathbf{x}_i, i \in \mathcal{N}\}$ et $n_R > 0$, qui ne modifie pas $P(y \mid \mathbf{X})$ parce que y et R sont indépendants étant donné $\mathbf{X}$, nous obtenons

$$\mathrm{E}(c_{\tilde{R},z} \mid \mathbf{X}, n_R > 0) \;=\; [\tilde{p}(\mathbf{X})]^{-1} \mathrm{Cov}_I\, (W_I \pi_I, \mathrm{E}[y_I \mid \mathbf{x}_I] - m(\mathbf{x}_I)). \tag{4.3}$$

Il est clair que l'équation (4.3) devient nulle quand nous choisissons $m(\mathbf{x}_I) = \mathrm{E}_\xi[y_I \mid \mathbf{x}_I]$ et que le modèle ξ (de premier ordre) est correctement précisé, c'est-à-dire que $\mathrm{E}_\xi[y_I \mid \mathbf{x}_I] = \mathrm{E}[y_I \mid \mathbf{x}_I]$. Cela résume la

méthode de la superpopulation et donne $c_{\tilde{R},z} = O_p(N^{-1/2})$ pour des raisons semblables à celles données pour le cadre de quasi-randomisation.

5. Quasi-randomisation *et* mise en œuvre de la méthode de superpopulation

Une fois qu'un modèle conjoint pour $\{R_i, y_i\}$ est configuré, nous pouvons bien sûr l'utiliser pour estimer à la fois π_i et la fonction de régression $m(\mathbf{x})$. Chaque estimation est rendue possible par la disponibilité de l'échantillon probabiliste auxiliaire et l'hypothèse de répartition au hasard des données manquantes. Mais comme nous l'avons montré précédemment, il suffit de préciser et d'estimer correctement l'une des deux pour miniaturiser $c_{\tilde{R},z}$. Cependant, à partir de l'équation (4.3), pour que la covariance ou la corrélation soit nulle, ni la correction multiplicative de π_I au moyen de W_I ni l'ajustement additif de $\mathrm{E}(y_I \,|\, \mathbf{x}_I)$ au moyen de $m(\mathbf{x}_I)$ n'ont besoin d'être corrects. Il faut seulement qu'après la correction ou l'ajustement, les éléments qui restent ne soient pas corrélés entre eux. Le cadre d'EMVC collaborative mentionné précédemment a essentiellement été élaboré à partir de cette perspective (par exemple voir la section 3.1 de van der Laan et Gruber, 2009), bien que les traitements mathématiques complexes présents dans la littérature aient pu décourager les lecteurs de chercher une compréhension aussi intuitive.

Pour donner un exemple simple, prenons une population finie qui est un échantillon indépendant et identiquement distribué d'un modèle de superpopulation :

$$\mathrm{E}[y \,|\, x] = \sum_{k=0}^{3} \beta_k x^k, \quad x \sim N(0,1). \tag{5.1}$$

L'échantillon non probabiliste est généré par un mécanisme R comme $\Pr\left(R=1 \,|\, y, x\right) = \pi\left(|x|\right)$, c'est-à-dire qu'il est déterminé par l'ampleur de x seulement. Supposons que nous avons précisé incorrectement la forme de fonction pour π (par exemple le modèle divin peut ne pas être monotone dans $|x|$, mais le modèle de procédé, comme le lien logistique classique, l'est), ainsi que le modèle de régression en choisissant $m(x) = b_0 + b_1 x + b_2 x^2$. Puisque x^2 n'est corrélé ni avec x ni avec x^3 selon $x \sim N(0,1)$, nous savons que notre estimateur par les moindres carrés pour b_2 serait toujours valide pour β_2 même dans le modèle de régression incorrectement précisé. Cela suffit pour assurer l'absence de biais asymptotique (quand $N \to \infty$) de l'estimateur « doublement robuste » suivant pour $\mu = \bar{y}_N$, la moyenne de population finie,

$$\hat{\mu}_+ = \frac{\sum_{i=1}^{N} R_i w\left(|x_i|\right)\left(y_i - \hat{m}(x_i)\right)}{\sum_{i=1}^{N} R_i w\left(|x_i|\right)} + \frac{\sum_{i=1}^{N} R_i^* \hat{m}(x_i)}{\sum_{i=1}^{N} R_i^*}, \tag{5.2}$$

où R^* indique l'échantillon auxiliaire (de $\mathbf{x}$ seulement). Ou encore, de façon équivalente,

$$\hat{\mu}_+ - \bar{y}_N = \frac{\mathrm{Cov}_I\left(R_I w(|x_I|), y_I - \hat{m}(x_I)\right)}{\mathrm{E}_I\left(R_I w(|x_I|)\right)} + \frac{\mathrm{Cov}_I\left(R_I^*, \hat{m}(x_I)\right)}{\mathrm{E}_I\left(R_I^*\right)}, \tag{5.3}$$

ce qui rend plus clair le fait que tout biais dans $\hat{\mu}_+$ est contrôlé par la covariance (ou la corrélation) comprenant R, puisque la covariance comprenant R^* est déjà miniaturisée par l'hypothèse selon laquelle l'échantillon auxiliaire est probabiliste (qui, par souci de simplicité, est supposé être un échantillon aléatoire simple).

Dans le cas présent, $w(x)$ représente toute fonction de pondération telle que $\mathrm{E}_\phi\left[|x|^3 w(|x|)\right] < \infty$, où l'espérance est prise par rapport à $x \sim N(0,1)$, et $\hat{m}(x) = b_0 + b_1 x + \hat{\beta}_2 x^2$, où $\hat{\beta}_2$ est l'estimateur par les moindres carrés pour β_2 à partir de l'échantillon biaisé, et où b_0 et b_1 peuvent être choisis arbitrairement. Parce que la covariance ou la corrélation de la population finie entre $\pi(|x_I|)w(|x_I|)$ et x_I^k est $O_p(N^{-1/2})$ quand $k=1$ et quand $k=3$, les parties ajustées de façon erronée pour π ou m ne contribuent pas (asymptotiquement) à la *cdd*, puisqu'elles ne sont pas corrélées les unes avec les autres dans le modèle de la superpopulation, ce qui entraîne une robustesse supplémentaire dépassant la « double robustesse ». Bien entendu, cela ne signifie pas que nous pouvons ajuster de façon erronée un modèle arbitrairement et tout de même obtenir des estimateurs valides, mais cela implique qu'avoir un modèle correct au moins est une condition suffisante, mais pas nécessaire, pour que les estimateurs doublement robustes soient valides.

Il est aussi intéressant de souligner que, dans la mise en forme du modèle de régression, nous n'avons pas nécessairement besoin d'invoquer une probabilité de procédé, par exemple un modèle de régression de la superpopulation, parce que la variable d'IPF fournit une régression de la population finie par l'application de la méthode des moindres carrés pour exécuter une régression de y_i sur $\mathbf{x}_i, i \in \mathcal{N}$. Cette régression qui s'autoajuste ne nous précise pas vraiment si la droite de régression $y = \hat{m}(\mathbf{x})$ qui en résulte est un bon ajustement de $(y_i, \mathbf{x}_i)$ ou non. Cependant, l'exemple ci-dessus indique que pour estimer la moyenne de population de y, l'absence d'ajustement peut ne pas avoir beaucoup d'importance, tant que le « résidu » $z_I = y_I - \hat{m}(\mathbf{x}_I)$ a peu de corrélation avec $W_I \pi_I$, en tant que deux fonctions de la variable d'IPF I. En effet, comme nous l'avons vu à la section 3, nous pouvons envisager d'inclure $\hat{\pi}_I$ dans le modèle de régression $\hat{m}(\mathbf{x}_I, \hat{\pi}_I)$. D'autres recherches devront examiner le degré d'efficacité générale de cette stratégie.

6. Sous-échantillonnage de contrebalancement

6.1 L'effet dévastateur du défaut des données sur la taille d'échantillon efficace

L'étude de la qualité des données aboutit à une constatation importante qui en a surpris plus d'un, à savoir la petite taille de nos « mégadonnées » si l'on tient compte du défaut des données. Pour la prouver mathématiquement, nous pouvons établir un rapport d'égalité entre l'erreur quadratique moyenne (EQM)

de $\bar{G}_W$ dans l'équation (2.1) et l'EQM d'un estimateur d'échantillonnage aléatoire simple de taille n_{eff}. Cela permet d'obtenir (voir le calcul dans l'article de Meng, 2018) :

$$n_{\text{eff}} \approx \frac{f_W}{1-f_W}\frac{1}{\mathrm{E}[\rho^2_{\tilde{R},G}]} \approx \frac{f_W}{1-f_W}\frac{1}{\rho^2_{\tilde{R},G}}, \tag{6.1}$$

où $f_W = n_W/N$ et l'espérance E est choisie par rapport à la distribution conditionnelle de $\tilde{R}$ étant donné n_W. Il convient de mentionner que cette distribution (conditionnelle) peut comprendre les trois types de probabilité abordés à la section 1.2, car les variations dans $\tilde{R}$ peuvent provenir de sources multiples. À titre d'exemple, dans les sondages d'opinion typiques, il y a 1) une probabilité du plan dans l'indicateur d'échantillonnage, 2) une probabilité divine dans la formulation du mécanisme de la non-réponse et 3) une probabilité de procédé pour l'estimation du mécanisme et des poids.

L'expression (6.1) est la version pondérée ou l'extension de l'expression fournie dans l'article de Meng (2018) comportant des poids égaux, ce qui révèle l'effet dévastateur d'une *cdd* apparemment minuscule. Supposons que notre échantillon représente 1 % de la population et qu'il est miné par une *cdd* de 0,5 %. Si nous appliquons (au moyen des poids égaux) f_W = 0,01 et $\rho_{\tilde{R},G}$ = 0,005 à l'expression (6.1), nous obtenons $n_{\text{eff}} \approx 404$, *quelle que soit la taille de l'échantillon* n_R. Dans le cas de l'élection présidentielle de 2020 aux États-Unis, 1 % de la population d'électeurs représente environ 1,55 million de personnes. La perte de la taille d'échantillon en raison d'une *cdd* de 0,5 % est ainsi d'environ 1 - (404/1 550 000) > 99,97 %. Ces pertes apparemment impossibles ont été relevées dans les études sur les élections (Meng, 2018) et les études sur la vaccination contre la COVID-19 (Bradley et coll., 2021). Une des conséquences les plus dévastatrices de ces pertes est le « paradoxe des mégadonnées » : plus la taille (apparente) des données est grande, plus nous nous leurrons parce que notre fausse confiance (au sens technique et littéral) augmente en fonction de la taille erronée des données, alors que la probabilité de couverture réelle des intervalles de confiance incorrectement construits devient extrêmement faible (Meng, 2018; Msaouel, 2022).

Cette réalité a tout de même une conséquence positive : nous pouvons échanger la grande quantité de données contre des données de qualité et finir par obtenir des estimations statistiquement plus exactes. Bien sûr, pour réduire le biais, il nous faudra des renseignements le concernant. Si nous avons des renseignements fiables sur la valeur de la *cdd*, nous pouvons corriger directement le biais dans l'estimation de la moyenne de la population correspondant à la *cdd*, par exemple par une approche bayésienne semblable à celle adoptée par Isakov et Kuriwaki (2020) dans leur analyse par scénario. De plus, si nous avons suffisamment de renseignements pour construire des poids fiables, nous pouvons utiliser les poids pour corriger les biais de sélection comme nous le faisons habituellement. Néanmoins, même dans ces cas, il est parfois utile de créer une miniature représentative de la population à partir d'un échantillon biaisé à des fins générales. Cela peut, par exemple, éliminer l'anxiété de nombreux spécialistes et les erreurs qu'ils sont susceptibles de commettre parce qu'ils ne maîtrisent pas l'utilisation des poids. De fait, peu de gens savent vraiment traiter les poids, parce que « la pondération des enquêtes est une véritable pagaille! » (Gelman, 2007).

Cependant, en général, la création d'une miniature représentative à partir d'un échantillon biaisé est une tâche difficile, surtout parce que la *cdd* peut (et va) varier selon la variable d'intérêt. Néanmoins, tout comme la pondération est un outil populaire, malgré toutes ses imperfections, nous étudierons la miniaturisation représentative pour voir jusqu'où cette idée peut nous mener. L'exemple suivant nous servira donc seulement de piste de réflexion. Nous examinerons un scénario courant, mais difficile, dans lequel nous disposons de renseignements ou d'une compréhension raisonnables sur l'orientation du biais, c'est-à-dire le signe de la *cdd*, mais de renseignements plutôt vagues sur son ampleur. Un bon exemple est la non-représentativité des sondages électoraux parce que les électeurs ont tendance à ne pas vouloir divulguer leurs préférences quand ils prévoient voter pour un candidat impopulaire dans la société. Nous connaissons donc l'orientation du biais, mais peu de choses au sujet de son degré, à part quelques conjectures (par exemple une fourchette de 10 points de pourcentage).

6.2 La création d'un sous-échantillon moins biaisé

L'idée de base consiste à utiliser ces renseignements partiels au sujet du biais de sélection pour concevoir un plan de sous-échantillonnage *biaisé* afin de *contrebalancer* le biais de l'échantillon original, de sorte que les sous-échantillons qui en résultent aient une *forte probabilité* d'être moins biaisés que l'échantillon original de notre population cible. Autrement dit, nous créons un indicateur de sous-échantillonnage S_I de façon à obtenir une probabilité élevée de réduction de la corrélation entre $S_I R_I$ et G_I, comparativement à la valeur originale de $\rho_{R,G}$, à un point tel qu'elle compensera la perte de la taille d'échantillon et réduira donc l'EQM de notre estimateur (par exemple la moyenne de l'échantillon). Nous parlons de *probabilité élevée*, dans un sens non technique, car en l'absence de renseignements complets sur le mécanisme de réponse ou d'enregistrement, nous ne pouvons jamais garantir qu'un sous-échantillonnage de contrebalancement (SECB) soit toujours meilleur. Il reste qu'une exécution judicieuse peut réduire la probabilité de commettre de graves erreurs.

À titre d'exemple, prenons le cas où y est binaire. Soit $\Delta = r_1 - r_0$, où r_y est la propension de réponse ou de déclaration chez les personnes dont les réponses auront la valeur y: $r_y = \Pr_I (R_I = 1 \,|\, y_I = y)$. Si l'échantillon est représentatif, alors comme $\rho_{R,G}$, Δ est miniaturisé, ce qui signifie qu'il est de l'ordre de $N^{-1/2}$. Cela se constate le plus clairement au moyen de l'identité facilement vérifiable (voir la formule [4.1] dans l'article de Meng, 2018)

$$\Delta = \frac{\mathrm{Cov}_I (y_I, R_I)}{p(1-p)} = \rho_{R,y} \sqrt{\frac{f_R (1-f_R)}{p(1-p)}}, \tag{6.2}$$

où $p = \Pr_I (y_I = 1)$ et $f_R = \Pr_I (R_I = 1)$, qui est le taux d'échantillonnage original. Un ingrédient clé du SECB consiste à déterminer $s_y = P_I (S_I = 1 \,|\, y_I = y, R_I = 1)$ pour $y = 0, 1$, c'est-à-dire, les probabilités de sous-échantillonnage des personnes qui ont déclaré $y = 1$ et $y = 0$, respectivement.

Pour déterminer les choix avantageux, supposons que $f_S = \Pr_I (S_I = 1 \,|\, R_I = 1)$ est le taux de sous-échantillonnage et que $\Delta_S = s_1 r_1 - s_0 r_0$. Ensuite, en appliquant l'équation (2.2) (au moyen de poids égaux)

et l'équation (6.2) à la moyenne de l'échantillon et à la moyenne du sous-échantillon, nous constatons que l'erreur de magnitude (réelle) du sous-échantillon est plus faible si, et seulement si

$$\left(\frac{\Delta_S}{f_S f_R}\right)^2 < \left(\frac{\Delta}{f_R}\right)^2 \Leftrightarrow f_S^2 > \left(\frac{\Delta_S}{\Delta}\right)^2. \tag{6.3}$$

Si l'on écrit $r = r_1/r_0$ et $s = s_1/s_0$, le deuxième membre de l'expression (6.3) devient

$$\left[sp^* + (1-p^*)\right]^2 > \left(\frac{rs-1}{r-1}\right)^2, \tag{6.4}$$

où $p^* = \Pr_I(y_I = 1 | R_I = 1)$ est observé dans l'échantillon original, ce qui devrait nous rappeler que p^* peut être plutôt différent du p que nous recherchons, en raison du mécanisme R biaisé.

Un choix immédiat pour satisfaire l'expression (6.4) consiste à établir $s = r^{-1}$, ce qui bien sûr est généralement irréaliste, car si nous connaissons la valeur de r, le problème serait beaucoup plus simple. Pour étudier la marge de manœuvre dont nous disposons pour nous écarter de ce choix idéal, nous supposons $\delta = r - 1$ pour ensuite montrer que l'expression (6.4) équivaut à

$$(s-1)\left\{[1+(1+p^*)\delta](s-1)+2\delta\right\} < 0. \tag{6.5}$$

Cela indique précisément les choix admissibles de s sans correction excessive (dans l'ampleur du biais qui en résulte) :

(i) Quand $r > 1$, c'est-à-dire $\delta > 0$, nous pouvons prendre tout s de sorte que

$$\frac{[1-(1-p^*)\delta]_+}{1+(1+p^*)\delta} \le s < 1; \tag{6.6}$$

(ii) Quand $r < 1$, c'est-à-dire $\delta < 0$, nous pouvons prendre tout s de sorte que

$$1 < s \le \frac{1-(1-p^*)\delta}{[1+(1+p^*)\delta]_+}. \tag{6.7}$$

Cette paire de résultats confirme un certain nombre de nos intuitions, mais offrent également des quantifications moins évidentes. Étant donné que nous sous-échantillonnons pour compenser le biais de l'échantillon original, s et r doivent rester du côté opposé à 1, c'est-à-dire $(s-1)(r-1) = (s-1)\delta < 0$, comme on le voit dans les expressions (6.6) et (6.7). Pour éviter les corrections excessives, il faut certaines limites, mais il est également possible que le biais initial soit si mauvais qu'aucun schéma de sous-échantillonnage ne puisse empirer les choses, ce qui se reflète dans la fonction de positivité $[x]_+$ des deux expressions mentionnées ci-dessus. Toutefois, pour les limites comme pour les seuils d'activation des fonctions de positivité, les expressions ne sont pas si évidentes. Il n'est pas non plus évident que ces expressions dépendent indirectement de la variable inconnue p par l'intermédiaire de la variable observée

p^*. Ainsi, seule la connaissance préalable de r est requise pour la mise en œuvre ou l'évaluation du SECB.

Cette observation donne à penser que nous pouvons mettre en œuvre un SECB avantageux quand nous pouvons emprunter des renseignements provenant d'autres enquêtes (ou études) dans lesquelles les comportements de réponse ou d'enregistrement sont de nature semblable. Par exemple, nous pourrions apprendre que pour une enquête antérieure semblable, $r = 1{,}5$ (par exemple les enquêtes pour lesquelles $y = 1$ avaient 6 % de chance d'être enregistrées et celles pour lesquelles $y = 0$ avaient seulement 4 % de chance). Compte tenu de l'incertitude de la similarité des deux enquêtes, nous pourrions sans hésitation poser (1,2; 1,8) comme fourchette plausible pour r dans la présente étude. Si nous observons que $p^* = 0{,}6$, cela signifie que le maximum, selon la fourchette $r \in (1{,}2;\ 1{,}8)$, de la limite inférieure de la variable s admissible, comme l'illustre l'expression (6.6), est

$$\frac{[1-(1-0{,}6)\,(r-1)]_+}{1+1{,}6(r-1)} = \frac{[1{,}4-0{,}4r]_+}{1{,}6r-0{,}6} \le \frac{1{,}4-0{,}4\times 1{,}2}{1{,}6\times 1{,}2-0{,}6} = 0{,}7. \qquad (6.8)$$

Par conséquent, tant que nous choisissons $s \in [0{,}7;1)$, il est peu probable que notre correction soit excessive. Le prix que nous payons pour cette robustesse est que le sous-échantillon qui en résulte n'est pas aussi bon qu'il pourrait l'être, par exemple quand la variable sous-jacente r de la présente étude est effectivement 1,5 (en espérance). Quel que soit le choix de $s \in [0{,}7;1)$, il ne donnera pas la correction complète fournie par $s = 1/r = 0{,}67$, c'est-à-dire que la moyenne du sous-échantillon aura toujours un biais positif, mais elle aura une EQM plus petite que la moyenne de l'échantillon d'origine. Il faut bien entendu soigneusement étudier la faisabilité et l'efficacité de ce type de SECB avant de pouvoir le recommander pour un usage général, en particulier au-delà d'un y binaire. La littérature sur l'échantillonnage inverse (Hinkins, Oh et Scheuren, 1997; Rao, Scott et Benhin, 2003) est d'une grande pertinence pour de telles investigations, car elle vise également à produire des échantillons aléatoires simples par sous-échantillonnage, bien qu'avec une motivation différente (pour transformer des enquêtes complexes en enquêtes simples afin de faciliter l'analyse).

7. L'échantillonnage probabiliste comme aspiration et non comme prescription

Comme la définition de la *cdd* devrait l'indiquer explicitement, il n'est pas possible de l'estimer directement à partir de l'échantillon biaisé seulement. C'est la raison pour laquelle on pourrait naturellement se demander (et l'on devrait se demander) dans quelle mesure la *cdd* est utile. En fait, la réponse se révèle de plus en plus longue étant donné que la *cdd* est sans modèle et donc, elle constitue une mesure polyvalente de la qualité des données pour les échantillons probabilistes et les échantillons non probabilistes. Son utilité pour produire des renseignements théoriques est démontrée par son rôle dans la quantification du compromis entre la qualité et la quantité des données au moyen de la taille d'échantillon

efficace, comme on l'a vu dans la formule (6.1), par la compréhension des erreurs de simulation dans la méthode de quasi-Monte Carlo, comme l'a étudiée Hickernell (2016), et par l'anticipation du phénomène de « robustesse plus que double » présenté dans la section 5. Ses emplois méthodologiques sont illustrés dans les analyses de scénarios concernant l'élection présidentielle américaine de 2020 (Isakov et Kuriwaki, 2020) et dans les évaluations de la vaccination contre la COVID-19 (Bradley et coll., 2021). Ses effets pratiques se trouvent dans des études épidémiologiques (Dempsey, 2020), dans des ouvrages sur la physique des particules (Courtoy, Houston, Nadolsky, Xie, Yan et Yuan, 2022) et dans des sondages politiques (Bailey, 2023).

Il n'est pas surprenant que, dans le cadre de ces applications pratiques, les spécialistes aient trouvé utiles la notion de *cdd* et la décomposition des erreurs sous-jacentes de l'équation (2.2) en raison des échantillons non probabilistes avec lesquels ils doivent composer, soit en raison de distorsions des échantillons probabilistes, par exemple celles causées par un mécanisme de non-réponse biaisé, ou en raison de biais de sélection, comme dans le cas des tests sélectifs de dépistage de la COVID-19. L'analyse du professeur Wu, ainsi que les nombreuses références indiquées dans son article et dans la présente étude, devrait clairement montrer que les échantillons non probabilistes sont *presque certainement* partout. Je n'ai pas recours à cette expression fortement probabiliste seulement pour sa valeur humoristique. Lorsque nous considérons le nombre inexplicable de valeurs possibles pour la moyenne de la *cdd*, la probabilité (peu importe la façon dont nous la construisons pour saisir le Far West des processus de collecte de données) qu'elle soit précisément de zéro doit être nulle. Cette moyenne nulle est une condition nécessaire pour que l'échantillon soit un échantillon probabiliste, car un échantillon probabiliste laisse entendre que la *cdd* soit de l'ordre $N^{-1/2}$ (Meng, 2018), ce qui est impossible quand sa moyenne n'est pas nulle (asymptotiquement). Cette observation nous donne à penser que nous devrions nous éloigner de la tradition consistant à traiter l'échantillonnage probabiliste comme une pièce maîtresse, puis à essayer de modéliser le monde beaucoup plus vaste des échantillons non probabilistes comme des « déviations » de celui-ci. Nous devrions plutôt commencer par étudier les échantillons à l'aide de mécanismes de collecte généraux et d'outils ou de concepts comme la *cdd*, puis nous devrions traiter les échantillons probabilistes (de plan) comme un cas idéal très particulier, qui doit toujours rester une aspiration, mais jamais la seule prescription à respecter.

Remerciements

Je tiens à remercier le rédacteur en chef, Jean-François Beaumont, de m'avoir invité à discuter de l'analyse fort opportune et stimulante de Changbao Wu. Je remercie aussi James Bailie, Radu Craiu, Adel Daoud, Andrew Gelman, Stas Kolenikov, Rod Little, Cory McCartan, Kelly McConville, James Robins, Zhiqiang Tan et Li-Chun Zhang de leur appui moral et de leurs critiques constructives. Je remercie également la Fondation nationale des sciences pour son soutien financier partiel et Steve Finch pour sa lecture d'épreuves minutieuse.

Bibliographie

Bailey, M.A. (2023). *Polling at a Crossroads – Rethinking Modern Survey Research*. Cambridge University Press.

Beaumont, J.-F., et Rao, J.N.K. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *Survey Statistician*, 83, 11-22.

Blei, D.M., Kucukelbir, A. et McAuliffe, J.D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518), 859-877.

Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, Z.-L. et Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600(7890), 695-700.

Buelens, B., Burger, J. et van den Brakel, J.A. (2018). Comparing inference methods for nonprobability samples. *Revue Internationale de Statistique*, 86(2), 322-343.

Courtoy, A., Houston, J., Nadolsky, P., Xie, K., Yan, M. et Yuan, C.-P. (2022). Parton distributions need representative sampling. *arXivpreprint arXiv:2205.10444.*

Craiu, R.V., Gong, R. et Meng, X.-L. (2022). Six statistical senses. *arXivpreprint arXiv:2204.05313.*

David Peat, F. (2002). *From Certainty to Uncertainty: The Story of Science and Ideas in the Twentieth Century*. Joseph Henry Press.

Dempsey, W. (2020). The hypothesis of testing: Paradoxes arising out of reported coronavirus case-counts. *arXiv preprint arXiv:2005.10425*.

Dwork, C. (2008). Differential privacy: A survey of results. Dans *International Conference on Theory and Applications of Models of Computation*, Springer, 1-19.

Elliott, M.R., et Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

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

Gong, R. (2022). Transparent privacy is principled privacy. *Harvard Data Science Review*, (Numéro spécial 2), 24 juin 2022. https://hdsr.mitpress.mit.edu/pub/ld4smnnf.

Gong, R., Groshen, E.L. et Vadhan, S. (2022). Harnessing the known unknowns: Differential privacy and the 2020 Census. *Harvard Data Science Review*, (Numéro spécial 2), 24 juin 2022. https://hdsr.mitpress.mit.edu/pub/fgyf5cne.

Han, P., et Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika*, 100(2), 417-430.

Hartley, H.O., et Ross, A. (1954). Unbiased ratio estimators. *Nature*, 174(4423), 270-271.

Hickernell, F.J. (2016). The trio identity for Quasi-Monte Carlo error. Dans *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, Springer, 3-27.

Hinkins, S., Oh, H.L. et Scheuren, F. (1997). Algorithmes de plan de sondage inverses. *Techniques d'enquête*, 23, 1, 13-24. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1997001/article/3101-fra.pdf.

Isakov, M., et Kuriwaki, S. (2020). Towards principled unskewing: Viewing 2020 election polls through a corrective Lens from 2016. *Harvard Data Science Review*, 2(4), 3 novembre 2020. https://hdsr.mitpress.mit.edu/pub/cnxbwum6.

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

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

Li, X., et Meng, X.-L. (2021). A multi-resolution theory for approximating infinite-*p*-zero-*n*: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *Journal of the American Statistical Association*, 116(533), 353-367.

Little, R., et An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, 14(3), 949-968.

Liu, Y., Gelman, A. et Chen, Q. (2021). Inference from non-random samples using Bayesian machine learning. *arXiv preprint arXiv:2104.05192*.

Lo, A.W. (2017). Adaptive markets. Dans *Adaptive Markets*. Princeton University Press.

Lohr, S., et Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101(475), 1019-1030.

Lohr, S.L. (2021). *Sampling: Design and Analysis*. Chapman and Hall/CRC.

Luque-Fernandez, M.A., Schomaker, M., Rachet, B. et Schnitzer, M.E. (2018). Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine*, 37(16), 2530-2546.

Meng, X.-L. (2014). A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it). Dans *Past, Present, and Future of Statistical Science*, (Éds., Lin et coll.), CRC Press.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i) Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, 12(2), 685-726.

Meng, X.-L. (2021). Enhancing (publications on) data quality: Deeper data minding and fuller data confession. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(4), 1161-1175.

Msaouel, P. (2022). The big data paradox in clinical practice. *Cancer Investigation*, 1-27.

Pfeffermann, D. (2017). Bayes-based non-bayesian inference on finite populations from non-representative samples: A unified approach. *Calcutta Statistical Association Bulletin*, 69(1), 35-63.

Rao, J.N.K., Scott, A.J. et Benhin, E. (2003). Défaire les structures des données d'enquête complexes : Théorie élémentaire et applications de l'échantillonnage inverse. *Techniques d'enquête*, 29, 2, 119-143. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2003002/article/6787-fra.pdf.

Robins, J.M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. Dans *Proceedings of the American Statistical Association*, Indianapolis, IN, 1999, 6-10.

Robins, J.M., Rotnitzky, A. et Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427), 846-866.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.

Scharfstein, D.O., Rotnitzky, A. et Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (avec discussions). *Journal of the American Statistical Association*, 94(448), 1096-1146.

Slavkovic, A., et Seeman, J. (2022). Statistical data privacy: A song of privacy and utility. *arXiv preprint arXiv:2205.03336*.

Tan, Y.V., Flannagan, C.A.C. et Elliott, M.R. (2019). "Robust-Squared" imputation models using Bart. *Journal of Survey Statistics and Methodology*, 7(4), 465-497.

Tan, Z. (2007). Comment: Understanding OR, PS and DR. *Statistical Science*, 22(4), 560-568.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3), 661-682.

Tan, Z. (2013). Simple design-efficient calibration estimators for rejective and high-entropy sampling. *Biometrika*, 100(2), 399-415.

Van Buuren, S., et Oudshoorn, K. (1999). *Flexible Multivariate Imputation by MICE*. Leiden: TNO.

van der Laan, M.J., et Gruber, S. (2009). Collaborative double robust targeted penalized maximum likelihood estimation. *UC Berkeley Division of Biostatistics Working Paper Series*, 246.

van der Laan, M.J., et Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 6(1).

van der Laan, M.J., et Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

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

Wu, C. (2022). Inférence statistique avec des échantillons d'enquête non probabiliste (avec discussions). *Techniques d'enquête*, 48, 2, 307-338. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2022002/article/00002-fra.pdf.

Wu, C., et Thompson, M.E. (2020). *Sampling Theory and Practice*. Springer.

Yang, S., Kim, J.K. et Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2), 445-465.

Zhang, G., et Little, R. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65(3), 911-918.

Zhang, L.-C. (2019). On valid descriptive inference from non-probability sample. *Statistical Theory and Related Fields*, 3(2), 103-113.

Techniques d'enquête, décembre 2022
Vol. 48, n° 2, pp. 393-398
Statistique Canada, n° 12-001-X au catalogue

Commentaires à propos de l'article « Inférence statistique avec des échantillons d'enquête non probabiliste »

Zhonglei Wang et Jae Kwang Kim[1]

Résumé

L'inférence statistique avec des échantillons d'enquête non probabilistes est un problème complexe bien connu en statistique. Dans la présente analyse, nous proposons deux nouvelles méthodes non paramétriques d'estimation des scores de propension pour pondérer les échantillons non probabilistes, à savoir la projection d'information et le calage uniforme dans un espace de Hilbert à noyau reproduisant.

Mots-clés : Projection d'information; calage uniforme des fonctions; intégration des données.

1. Introduction

Nous tenons à féliciter M. Changbao Wu pour son travail exceptionnel dans le domaine de l'échantillonnage non probabiliste. Bien que l'échantillonnage probabiliste ait servi d'outil de référence pour l'inférence de population finie au cours des dernières décennies, les faibles taux de réponse et les coûts élevés qui y sont associés ont récemment terni sa réputation. En revanche, l'échantillonnage non probabiliste gagne en popularité en raison de sa faisabilité et de son faible coût (Couper, 2000; Kaplowitz, Hadlock et Levine, 2004). Plus important encore, comparativement à l'échantillonnage probabiliste, l'échantillonnage non probabiliste, tel qu'utilisé dans les enquêtes en ligne, permet de recueillir plus rapidement les renseignements les plus récents. Cependant, comme le mécanisme de sélection des échantillons est inconnu dans le cas de l'échantillonnage non probabiliste, le fait de ne pas corriger le biais de sélection lors de l'analyse d'un échantillon non probabiliste peut entraîner une inefficacité, voire une inférence erronée. Par conséquent, la correction du biais de sélection dans les échantillons non probabilistes est une question fondamentale pour les chercheurs dans le domaine de l'échantillonnage d'enquête. Nous aborderons dans cette analyse les travaux les plus rigoureux s'étant penchés sur ce sujet.

L'article de recherche de Wu, en particulier, comprend un examen approfondi des techniques fondées sur les scores de propension. Ces techniques comportent cependant des inconvénients. Tout d'abord, même pour un modèle de score de propension correctement spécifié, l'estimateur de pondération de probabilité inverse peut être inefficace lorsque les scores de propension estimés sont faibles. Une solution de rechange est la poststratification, comme il est indiqué à la section 5 de l'article de Wu, bien qu'il n'y ait pas de directives claires sur la façon de choisir K. De plus, dans la pratique, il est difficile de spécifier correctement un modèle de score de propension. Même si l'estimation doublement robuste peut aider à corriger un modèle de score de propension erroné, l'estimateur final pose problème lorsque le modèle de score de propension et le modèle de régression sont inexacts (Kang et Schafer, 2007).

1. Zhonglei Wang, Wang Yanan Institute for Studies in Economics and School of Economics, Xiamen University, Xiamen, Fujian, People's République de Chine; Jae Kwang Kim, Iowa State University, Ames, IA 50011, États-Unis. Courriel : jkim@iastate.edu.

Pour corriger les erreurs de spécification du modèle de score de propension, Wu mentionne plusieurs méthodes non paramétriques, y compris une méthode des noyaux et une méthode à arborescence. Dans le cadre de la présente analyse, nous aimerions approfondir ces notions et présenter deux autres méthodes pour enrichir l'étude. La première est fondée sur un modèle de ratio de densité faisant appel à la projection d'information (Csiszár et Shields, 2004), et la deuxième est fondée sur le calage uniforme de fonctions dans un espace de Hilbert à noyau reproduisant. Comme l'explique Wahba (1990), au chapitre de l'approximation, les espaces de Hilbert à noyau reproduisant sont des espaces de fonctions très souples. Plutôt que d'estimer les scores de propension, on tente d'estimer les poids de sondage $\left\{(\pi_i^A)^{-1} : i \in S_A\right\}$ de sorte à éviter le risque d'inefficacité associé aux faibles scores de propension estimés.

S_A et S_B indiquent les ensembles d'indices de l'échantillon non probabiliste et de l'échantillon probabiliste de référence, respectivement, et n_A et n_B, les tailles d'échantillon correspondantes. Soit $\left\{(y_i, \mathbf{x}_i) : i \in S_A\right\}$ et $\left\{(\mathbf{x}_i, d_i^B) : i \in S_B\right\}$, que l'on suppose être disponibles, où y_i et $\mathbf{x}_i$ sont la variable étudiée et le vecteur auxiliaire pour la i^{e} unité et d_i^B est le poids de sondage pour $i \in S_B$.

La présente analyse est organisée comme suit. Dans la section 2, nous présentons la méthode fondée sur la projection d'information. Dans la section 3, nous définissons la notion de base qu'est le calage uniforme. Enfin, dans la section 4, nous présentons quelques conclusions.

2. Méthode fondée sur la projection d'information

Supposons que nous souhaitons estimer le paramètre $\boldsymbol{\theta}_0$ défini par $E_N\left\{U(\boldsymbol{\theta}; \mathbf{X}, Y)\right\} = 0$, où $E_N(\cdot)$ est la valeur attendue par rapport à la distribution empirique de la population $\Pr\left\{(\mathbf{X}, Y) = (\mathbf{x}_i, y_i)\right\} = N^{-1}$ pour $i = 1, \ldots, N$ et 0 autrement, et $U(\boldsymbol{\theta}; \mathbf{x}, y)$ est une fonction d'estimation donnée. Par exemple, $U(\theta; \mathbf{x}, y) = y - \theta$ correspond à $\mu_y = N^{-1}\sum_{i=1}^N y_i$ dans l'article. Nous souhaitons obtenir un estimateur de $(\pi_i^A)^{-1}$, $\pi_i^A = \Pr\left(R_i = 1 \mid \mathbf{x}_i, y_i\right)$ et $R_i = 1$ si $i \in S_A$ et 0 autrement.

Pour estimer $\left\{(\pi_i^A)^{-1} : i \in S_A\right\}$, nous pouvons utiliser la relation dans la fonction du ratio de densité. Tout d'abord, nous considérons un modèle de superpopulation ξ et supposons que $f_0(\mathbf{x}, y)$ et $f_1(\mathbf{x}, y)$ sont les fonctions de densité de $(\mathbf{x}, y)$, si $R = 0$ et $R = 1$, respectivement. Ainsi, nous définissons la fonction du ratio de densité comme suit :

$$r(\mathbf{x}, y) = \frac{f_0(\mathbf{x}, y)}{f_1(\mathbf{x}, y)},$$

D'après la formule de Bayes, nous avons :

$$(\pi_i^A)^{-1} = 1 + \frac{\Pr(R_i = 0)}{\Pr(R_i = 1)}\, r(\mathbf{x}_i, y_i). \tag{2.1}$$

Ainsi, il existe un rapport d'un à un entre $(\pi_i^A)^{-1}$ et $r(\mathbf{x}_i, y_i)$.

Dans l'hypothèse A1, nous pouvons montrer que $r(\mathbf{x}, y) = r(\mathbf{x})$. Dans cette section, nous posons l'hypothèse plus générale de l'existence de $\mathbf{b}(\mathbf{x}) = (b_1(\mathbf{x}), \ldots, b_L(\mathbf{x}))^{\mathrm{T}}$, de sorte que

$$R \perp Y \mid \mathbf{b}(\mathbf{x}). \tag{2.2}$$

Rosenbaum et Rubin (1983) ont défini $\mathbf{b}(\mathbf{x})$ dans l'équation (2.2) comme un score d'équilibrage.

Pour estimer la fonction du ratio de densité $r(\mathbf{x})$, nous minimisons la divergence de Kullback-Leibler

$$Q(f_0) \;=\; \int \log(f_0 / f_1)\; f_0 \, \mathrm{d}\mu \tag{2.3}$$

relativement à f_0 soumis à une certaine contrainte, où f_0 et f_1 sont absolument continus par rapport à une mesure σ-finie μ. Pour ce qui est de la contrainte, nous pouvons utiliser la suivante :

$$\Pr(R_i = 1) \int \mathbf{b}(\mathbf{x}) f_1(\mathbf{x}) \mu(\mathrm{d}\mathbf{x}) \;+\; \Pr(R_i = 0) \int \mathbf{b}(\mathbf{x}) f_0(\mathbf{x}) \mu(\mathrm{d}\mathbf{x}) \;=\; E_\xi\{\mathbf{b}(\mathbf{X})\}, \tag{2.4}$$

où $E_\xi(\cdot)$ est la valeur attendue en ce qui concerne le modèle de superpopulation ξ. Cela signifie qu'à partir de $f_1(\mathbf{x})$, nous pouvons trouver $f_0(\mathbf{x})$ pour minimiser (2.3) à l'aide d'une contrainte de calage par rapport à $\mathbf{b}(\mathbf{x})$.

Selon le lemme 3.1 de Wang et Kim (2021), la fonction de densité conditionnelle optimisée satisfait à

$$f_0^*(\mathbf{x}) \;=\; f_1(\mathbf{x}) \frac{\exp\{\boldsymbol{\lambda}_1^{\mathrm{T}} \mathbf{b}(\mathbf{x})\}}{E_1\left[\exp\{\boldsymbol{\lambda}_1^{\mathrm{T}} \mathbf{b}(\mathbf{x})\}\right]}, \tag{2.5}$$

où $\boldsymbol{\lambda}_1$ est choisi pour satisfaire à la contrainte (2.4). Il convient de souligner que la solution (2.5) est équivalente à

$$\log\{r(\mathbf{x};\, \boldsymbol{\lambda})\} \;=\; \lambda_0 + \boldsymbol{\lambda}_1^{\mathrm{T}} \mathbf{b}(\mathbf{x}) \tag{2.6}$$

pour la fonction du ratio de densité $r(\mathbf{x})$, où $\boldsymbol{\lambda} = (\lambda_0, \boldsymbol{\lambda}_1^{\mathrm{T}})^{\mathrm{T}}$, et λ_0 est une constante de normalisation satisfaisant à $\int r(\mathbf{x};\, \boldsymbol{\lambda})\, f_1(\mathbf{x})\, \mu(\mathrm{d}\mathbf{x}) = 1$. Par conséquent, la projection d'information permet de trouver le meilleur modèle pour la fonction du score de propension.

Une fois que nous avons déterminé le modèle, comme en (2.6), nous devons en estimer les paramètres. En raison des contraintes de moment en (2.4), l'équation d'estimation de la version d'échantillon pour $\boldsymbol{\lambda}$ est l'équation de calage obtenue par

$$\frac{n_A}{N} \sum_{i=1}^{N} R_i \left[1, \mathbf{b}(\mathbf{x}_i)\right] \left[1 + \frac{1 - n_A}{n_A} \exp\{\lambda_0 + \boldsymbol{\lambda}_1^{\mathrm{T}} \mathbf{b}(\mathbf{x}_i)\}\right] \;=\; \left[1, \frac{1}{N} \sum_{i \in S_B} d_i^B \mathbf{b}(\mathbf{x}_i)\right]. \tag{2.7}$$

Ici, comme $E_\xi\{\mathbf{b}(\mathbf{X})\}$ est inconnu, nous utilisons $N^{-1} \sum_{i \in S_B} d_i^B \mathbf{b}(\mathbf{x}_i)$ pour l'estimer. Une fois l'estimation des paramètres $\hat{\boldsymbol{\lambda}}$ obtenue, nous pouvons construire

$$\hat{\omega}_i \;=\; 1 \;+\; \frac{1 - n_A}{n_A} \exp\{\hat{\lambda}_0 + \hat{\boldsymbol{\lambda}}_1^{\mathrm{T}} \mathbf{b}(\mathbf{x}_i)\}$$

comme poids de score de propension finaux. On peut estimer le paramètre d'intérêt en résolvant $N^{-1}\sum_{i\in S_A}\hat{\omega}_i U(\boldsymbol{\theta};\, \mathbf{x}_i, y_i) = 0$ pour $\boldsymbol{\theta}$.

Wang et Kim (2021) ont élaboré ce cadre dans un contexte d'échantillonnage non probabiliste où $\mathbf{x}_i$ est disponible pour l'ensemble de la population finie. Il est possible d'établir la convergence et la normalité asymptotique en posant l'hypothèse selon laquelle $E\{U(\boldsymbol{\theta};\, \mathbf{x}, Y) \mid \mathbf{x}\}$ se trouve dans l'espace linéaire généré par $\{b_1(\mathbf{x}), \ldots, b_L(\mathbf{x})\}$. Au lieu de supposer la disponibilité de $\{\mathbf{x}_i : i = 1, \ldots, N\}$, comme dans Wang et Kim (2021), on utilise uniquement un échantillon probabiliste de référence $\{(\mathbf{x}_i, d_i^B) :\, i \in S_B\}$. Si l'échantillon probabiliste S_B est un recensement, alors la méthode ci-dessus équivaut à celle considérée par Wang et Kim (2021), à l'exception près que nous considérons un paramètre de population finie $\boldsymbol{\theta}_0$. Dans la section 11.2 de leur ouvrage, Kim et Shao (2021) appliquent la méthode fondée sur la projection d'information, qu'ils appellent la méthode d'entropie maximale, au problème d'intégration des données. Dans l'étude par simulations présentée à l'exemple 11.1 de l'ouvrage, la méthode fondée sur la projection d'information proposée génère de meilleurs résultats que les méthodes de Chen, Li et Wu (2020) et d'Elliott et Valliant (2017).

3. Méthode fondée sur le calage uniforme

Le calage est couramment utilisé pour améliorer la représentativité d'un échantillon non probabiliste, mais les méthodes existantes, y compris la méthode fondée sur la projection d'information mentionnée à la section 2, sont fondées sur le calage d'un ensemble de fonctions prédéfinies. Cependant, en pratique, il est difficile de préciser correctement ces fonctions aux fins du calage. Dans la présente section, nous proposons un cadre général pour le calage uniforme des fonctions dans un espace de Hilbert à noyau reproduisant. Au lieu de considérer une forme paramétrique pour $E_\xi(Y \mid \mathbf{x})$ dans l'équation (3.1), nous supposons seulement que $E_\xi(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i)$, où $m(\mathbf{x})$ est une fonction lisse satisfaisant à certaines conditions.

Nous considérons tout de même (2.1) selon l'hypothèse A1. Au lieu de supposer un ensemble de fonctions prédéfinies $\mathbf{b}(\mathbf{x})$, nous proposons d'estimer $\{r_i :\, i \in S_A\}$ au moyen de l'optimisation suivante :

$$\hat{\boldsymbol{\gamma}} \;=\; \underset{\boldsymbol{\gamma}\geq 0}{\operatorname{argmin}} \left[\sup_{u\in H} \left\{ \frac{S(\boldsymbol{\gamma}, u)}{\|u\|_2^2} \;-\; \lambda_1 \frac{\|u\|_H^2}{\|u\|_2^2} \right\} \;+\; \lambda_2 Q_A(\boldsymbol{\gamma}) \right], \tag{3.1}$$

où $\boldsymbol{\gamma} = (r_1, \ldots, r_N)$, $r_i = 0$ pour $i \notin S_A$, $\boldsymbol{\gamma} \geq 0$ est équivalent à $r_i \geq 0$ pour $i = 1, \ldots, N$, H est un espace de Hilbert à noyau reproduisant,

$$S(\boldsymbol{\gamma}, u) \;=\; \left[N^{-1} \sum_{i\in S_A} \left\{ 1 \;+\; \left(\frac{N}{n_A} \;-\; 1 \right) r_i \right\} u(\mathbf{x}_i) \;-\; N^{-1} \sum_{i\in S_B} d_i^B u(\mathbf{x}_i) \right]^2, \tag{3.2}$$

$\|u\|_2^2 \;=\; (n_A + n_B)^{-1} \sum_{i\in S_A \cup S_B} u(\mathbf{x}_i)^2$, $\|u\|_H$ est la norme associée à l'espace de Hilbert à noyau reproduisant, $Q_A(\boldsymbol{\gamma})$ est une pénalité générale imposée à $\boldsymbol{\gamma}$ pour éviter un surajustement, et λ_1 et λ_2 sont

deux paramètres de réglage. Voir Wahba (1990) pour une introduction détaillée à l'espace de Hilbert à noyau reproduisant.

L'intuition relative à l'optimisation (3.1) y est brièvement abordée. Premièrement, si r_i se rapproche suffisamment du ratio de densité réel $r(\mathbf{x}_i)$, le biais du premier segment dans l'équation (3.1) est négligeable pour l'estimation de $N^{-1}\sum_{i=1}^{N} u(\mathbf{x}_i)$ pour $u \in H$. De plus, $N^{-1}\sum_{i \in S_B} d_i^B u(\mathbf{x}_i)$ est sans biais par rapport au plan. Ainsi, $S(\boldsymbol{\gamma}, u)$ équilibre deux estimateurs pour $N^{-1}\sum_{i=1}^{N} u(\mathbf{x}_i)$, et il est faible si r_i équivaut approximativement à $r(\mathbf{x}_i)$ pour $i \in S_A$. Cependant, $S(\boldsymbol{\gamma}, u)$ n'est pas invariant par rapport à l'échelle, et nous avons $S(\boldsymbol{\gamma}, cu) = c^2 S(\boldsymbol{\gamma}, u)$ pour $c \in \mathbb{R}$. Par conséquent, nous utilisons $\|u\|_2^2$ pour en assurer l'invariabilité par rapport à l'échelle. Le segment $\lambda_1 \|u\|_H^2$ est utilisé pour pénaliser le lissage de la fonction u pour $u \in H$. En ce qui concerne $Q_A(\boldsymbol{\gamma})$, nous avons différentes options. Par exemple, nous pouvons pénaliser les valeurs extrêmes pour les poids de sondage au moyen de l'équation $Q_A(\boldsymbol{\gamma}) = \sum_{i \in S_A} \left\{1 + (N n_A^{-1} - 1)\, r_i\right\}^2$. Par ailleurs, Wong et Chan (2018) ont étudié un problème semblable en supposant la disponibilité de $\{\mathbf{x}_i : i = 1, \ldots, N\}$. L'optimisation (3.1) peut être considérée comme un problème « min-max », et si $m \in H$, les ratios de densité estimés $\left\{\hat{r}_i : i \in S_A\right\}$ peuvent conduire à un estimateur raisonnablement bon

$$\hat{\mu}_{uc} = N^{-1} \sum_{i \in S_A} \left\{1 + \left(\frac{N}{n_A} - 1\right) \hat{r}_i\right\} y_i. \tag{3.3}$$

Le calage uniforme est une nouvelle méthode d'échantillonnage non probabiliste, qui comporte certaines difficultés techniques en ce qui a trait à (3.1). Par exemple, la façon d'intégrer les propriétés de plan de S_B lors de l'établissement des propriétés théoriques de (3.3) n'a pas été étudiée de façon approfondie. Nous venons d'ailleurs de terminer un document de travail sur ce sujet (Wang, Mao et Kim, 2022). La méthode fondée sur le noyau est vorace en calcul, surtout lorsque la taille des échantillons est grande. Il pourrait être intéressant de proposer un algorithme plus efficace sur le plan du calcul pour résoudre le problème de calage uniforme. Considérer d'autres espaces fonctionnels, comme celui couvert par les B-splines, pourrait être une solution. De plus, il serait intéressant d'étudier la façon d'intégrer plus d'un échantillon probabiliste de référence et de formuler un problème de calage uniforme lorsqu'en présence de différentes covariables dans différents échantillons probabilistes de référence.

4. Conclusions

La pondération du score de propension est un important outil pour corriger le biais de sélection dans le contexte de l'échantillonnage non probabiliste. M. Changbao Wu a grandement contribué à l'avancement de la recherche sur ce sujet important. En plus des deux nouvelles méthodes proposées, l'approche de la vraisemblance empirique de Qin, Leung et Shao (2002) pourrait être un outil de pondération du score de propension utile. En particulier, la méthode de pondération fondée sur la vraisemblance empirique est applicable même dans le cadre d'un échantillonnage informatif. Cette approche sera étudiée plus amplement dans d'autres travaux.

Bibliographie

Chen, Y., Li, P. et Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Couper, M.P. (2000). Web surveys: A review of issues and approaches. *The Public Opinion Quarterly*, 64(4), 464-494.

Csiszár, I., et Shields, P.C. (2004). *Information Theory and Statistics: A Tutorial.*

Elliott, M., et Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

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

Kaplowitz, M.D., Hadlock, T.D. et Levine, R. (2004). A comparison of Web and mail survey response rates. *Public Opinion Quarterly*, 68(1), 94-101.

Kim, J.K., et Shao, J. (2021). *Statistical Methods for Handling Incomplete Data*, second edition. CRC press.

Qin, J., Leung, D. et Shao, J. (2002). Estimation with survey data under non-ignorable nonresponse or informative sampling. *Journal of the American Statistical Association*, 97, 193-200.

Rosenbaum, P.R., et Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphie.

Wang, H., et Kim, J.K. (2021). Information projection approach to propensity score estimation for handling selection bias under missing at random. *arXiv:2104.13469*, 1-34.

Wang, Z., Mao, X. et Kim, J.K. (2022). Functional calibration under non-probability survey sampling. Soumis (https://arxiv.org/abs/2204.09193).

Wong, R.K., et Chan, K.C.G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika*, 105(1), 199-213.

Techniques d'enquête, décembre 2022
Vol. 48, n° 2, pp. 399-406
Statistique Canada, n° 12-001-X au catalogue

Réponse de l'auteur aux commentaires sur l'article « Inférence statistique avec des échantillons d'enquête non probabiliste »

Changbao Wu[1]

Résumé

La présente réponse contient des remarques supplémentaires sur certaines questions soulevées par les participants à la discussion.

Mots-clés : Corrélation due à un défaut des données; double robustesse; pondération de probabilité inverse; hypothèses de modèle; prédiction fondée sur un modèle; échantillon de validation.

Permettez-moi tout d'abord de remercier le rédacteur en chef de *Techniques d'enquête*, Jean-François Beaumont, d'avoir organisé les discussions et d'avoir rassemblé un flamboyant ensemble d'intervenants. Chaque participant à la discussion s'est penché sur la question des échantillons d'enquête non probabilistes et de façon plus générale, sur les sujets de l'intégration des données et de la combinaison de données provenant de sources multiples, avec des points de vue uniques. Ces discussions stimulantes contribuent, selon moi, de façon importante au traitement des échantillons non probabilistes et d'autres types d'échantillons présentant un biais de sélection. Dans les lignes qui suivent, je ferai quelques observations supplémentaires sur certaines questions soulevées par les participants à la discussion.

Michael A. Bailey

Michael A. Bailey s'est concentré sur les limites des méthodes d'estimation que j'ai présentées dans le cadre des hypothèses A1 à A4, et il a appelé à poursuivre le développement en cas de violation de ces hypothèses ainsi que de l'hypothèse dite « de données manquantes au hasard » A1 en particulier. Bailey s'est servi de l'exemple du sondage non probabiliste pour soutenir que « la non-réponse dépend (peut en effet dépendre) de la variable étudiée » et que le danger de violation de l'hypothèse A1 est réel.

Bien que les critiques sur les limites des méthodes examinées dans mon article soient justes et honnêtes, les énoncés « (Wu) pêche dans un coin très précis de l'étang » et il « s'éloigne des modèles de données manquantes non au hasard » semblent montrer une sous-appréciation marquée de l'importance de l'élaboration de méthodes selon les hypothèses types A1 à A4, employées par plusieurs auteurs sur des échantillons d'enquête non probabilistes. Premièrement, l'hypothèse A1 porte sur le mécanisme de participation (ou d'inclusion ou de sélection) pour des échantillons non probabilistes, ce qui n'est pas la même chose que la « non-réponse ». Ces hypothèses peuvent se justifier dans de nombreux scénarios, surtout pour les enquêtes reposant sur des panels Web ou téléphoniques, dans lesquels la participation initiale dépend fortement de certaines variables démographiques. Deuxièmement, le comportement de

1. Changbao Wu, Département des sciences statistiques et actuarielles, Université de Waterloo, Waterloo (Ontario), N2L 3G1. Courriel : cbwu@uwaterloo.ca.

participation aux enquêtes non probabilistes peut être influencé par des facteurs de confusion, tels que certaines variables de l'étude pendant la collecte des données, comme ce que nous constatons dans les enquêtes probabilistes sur la non-réponse, ce qui correspond à la façon dont la littérature actuelle sur les enquêtes non probabilistes a évolué dans le traitement de ces questions. Troisièmement, toute avancée méthodologique dans le traitement des « modèles dits de données manquantes non au hasard » pour les enquêtes non probabilistes nécessiterait les fondements et la compréhension approfondie établis selon les hypothèses A1 à A4.

Michael A. Bailey a par ailleurs affirmé que « bien que les violations de l'hypothèse des données manquantes au hasard constituent un problème dans l'échantillonnage probabiliste (découlant de la non-réponse chez les personnes avec lesquelles on a communiqué au hasard), les violations de l'hypothèse des données manquantes au hasard sont plus graves dans un monde non probabiliste ». Je suis tout à fait d'accord avec cette analyse. De fait, les violations de l'hypothèse de positivité A2 sont aussi graves que les violations de l'« hypothèse de données manquantes au hasard » A1, et les deux sont interreliées. Les violations de l'hypothèse A2 impliquent que $\pi_i^A = P(i \in S_A \,|\, \mathbf{x}_i, y_i) = 0$ pour certaines unités de la population cible, ce qui entraîne un problème de sous-dénombrement qui est aussi connu que la non-réponse. S'il y a violation d'A2, mais qu'A1 se vérifie, on croit souvent que les estimateurs de prédiction fondés sur un modèle peuvent atténuer les biais dus au sous-dénombrement. Dans l'hypothèse A1, la variable de l'indicateur d'inclusion de l'échantillon R et la variable étudiée y sont conditionnellement indépendantes étant donné $\mathbf{x}$, ce qui signifie que :

$$E(y_i \,|\, \mathbf{x}_i, R_i = 1) \;=\; E(y_i \,|\, \mathbf{x}_i). \tag{1}$$

Il s'ensuit qu'un modèle de prédiction valide $y \,|\, \mathbf{x}$ peut être construit au moyen des données observées $\{(y_i, \mathbf{x}_i), i \in S_A\}$ (c'est-à-dire des unités avec $R_i = 1$). Malheureusement, l'équation (1) exige implicitement $P(R_i = 1 \,|\, \mathbf{x}_i) > 0,$ et les estimateurs fondés sur des prédictions ne sont pas à l'abri des biais potentiels dus au sous-dénombrement. L'appel de Michael A. Bailey en faveur d'un « cadre qui englobe la possibilité de violations de l'hypothèse de données manquantes au hasard » est conforme à certains travaux recherche actuels sur le traitement du sous-dénombrement et des mécanismes de participation « non ignorables » pour les échantillons d'enquête non probabilistes. Voir notamment Chen, Li et Wu (2023), Cho, Kim et Qiu (2022) et Yuan, Li et Wu (2022). En bref, pour obtenir des inférences statistiques valides à partir de ces scénarios, il faut des données externes, comme un échantillon de validation, ou des hypothèses supplémentaires, comme l'existence de variables instrumentales.

Je suis exactement sur la même longueur d'onde que Michael A. Bailey à propos de l'étiquette « manquant au hasard », puisque le terme pourrait être confondu avec « manquant aléatoirement » (Wu et Thompson, 2020, page 195). Le terme « ignorable » est également un mauvais choix de terminologie pour les données manquantes et la littérature sur l'inférence causale, car l'analyste des données ne peut certainement pas les ignorer (Rivers, 2007). J'utilise le terme courant « scores de propension » pour les échantillons non probabilistes, tandis que plusieurs autres auteurs lui préfèrent « probabilités de participation », y compris Beaumont (2020) et Rao (2021).

Michael R. Elliott

Michael R. Elliott a traité de plusieurs questions en utilisant des documents supplémentaires et une liste de références plus longue. Il s'agit d'ajouts importants au sujet actuel, en particulier les examens « d'autres approches pour combiner les données tirées d'enquêtes probabilistes et celles d'enquêtes non probabilistes » et l'analyse de sensibilité sur des « hypothèses non vérifiables ».

Les discussions d'Elliott concernant les distinctions entre les paramètres descriptifs et les paramètres analytiques ainsi que la pondération par rapport à la modélisation ont soulevé la question délicate de l'efficacité des estimateurs de pondération par l'inverse de la propension (PIP) dans la pratique. On sait, pour les échantillons d'enquête probabiliste, que l'estimateur de Horvitz-Thompson pondéré par la probabilité inverse du total de population T_y est extrêmement inefficace (en termes de grande variance) quand les probabilités de sélection de l'échantillon π_i sont inégales, mais ont une très faible corrélation avec la variable étudiée y, bien que l'estimateur demeure sans biais dans de tels scénarios. L'exemple de l'éléphant de Basu (Basu, 1971) montrait un « cas convaincant » dans lequel l'estimateur de Horvitz-Thompson pondéré par la probabilité inverse et sans biais a échoué lamentablement, ce qui a entraîné le congédiement du statisticien du cirque. Les discussions sur la pondération par rapport à la modélisation, c'est-à-dire les estimateurs pondérés par la probabilité inverse comparativement aux estimateurs de la prédiction fondés sur un modèle pour les paramètres descriptifs de la population, sont très pertinentes pour les développements théoriques et les applications pratiques. En tant que statisticiens, notre travail de traitement des échantillons d'enquête non probabilistes pourrait être très incertain si nous n'élaborons pas des lignes directrices et des outils de diagnostic solides nous permettant de choisir des méthodes adéquates selon l'ensemble de données à notre disposition et les problèmes d'inférence.

Michael R. Elliott fait écho à mon appel à réaliser quelques enquêtes probabilistes à grande échelle comportant une information riche sur les variables auxiliaires en affirmant « qu'il est de plus en plus essentiel de mettre en place des enquêtes probabilistes structurées et idéalement financées par le gouvernement pour les collectes de données courantes ». Ses commentaires à propos des nouveaux domaines de recherche sur les questions de protection de la vie privée et de confidentialité en raison du besoin de microdonnées dans le contexte de l'analyse d'échantillons d'enquêtes non probabilistes constituent un appel visionnaire, qui mérite une plus grande attention de la part du milieu de la recherche.

Zhonglei Wang et Jae Kwang Kim

Zhonglei Wang et Jae Kwang Kim présentent deux nouvelles approches de l'estimation fondée sur le score de propension : l'une utilise ce qu'on appelle la projection de renseignements au moyen d'un modèle de ratio de densité et l'autre emploie le calage uniforme de fonctions dans un espace de Hilbert à noyau reproduisant (RKHS, de l'appellation anglaise *Reproducing Kernel Hilbert Space*). Ces méthodes sont de nouvelles aventures dans notre domaine, et Kim et ses collaborateurs ont l'expérience et la puissance analytique nécessaires pour faire avancer la recherche dans ce sens.

Le point de départ des deux approches est l'équation suivante, qui relie les scores de propension aux ratios de densité :

$$\frac{1}{P(R_i = 1 \mid \mathbf{x}_i, y_i)} = 1 + \frac{P(R_i = 0)}{P(R_i = 1)} \frac{f_0(\mathbf{x}_i, y_i)}{f_1(\mathbf{x}_i, y_i)}.$$

Les scores de propension $\pi_i^A = P(R_i = 1 \mid \mathbf{x}_i, y_i)$ nécessitent seulement le modèle sur $R_i = 1$ étant donné $\mathbf{x}_i$ et y_i. Toutefois, la justification de l'équation donnée ci-dessus exige un cadre de randomisation conjoint comprenant à la fois le modèle q pour les scores de propension et le modèle de superpopulation ξ sur $(\mathbf{x}, y)$. Du point de vue de la cohérence concernant l'estimateur final de la moyenne de population finie de y, le cadre conjoint impose très peu de restrictions si les ratios de densité sont modélisés de façon non paramétrique. Cette méthode a une incidence considérable sur la variance et l'estimation de la variance. La variance d'un estimateur dans un cadre de randomisation conjoint comporte plus d'une composante, et l'estimation de la variance entraîne d'autres complications en cas de procédures non paramétriques. Les comparaisons d'efficacité entre les méthodes proposées et certaines des méthodes existantes doivent être effectuées dans des configurations appropriées. J'ai hâte de suivre les progrès à venir à partir des méthodes qui ont été avancées.

Sharon L. Lohr

La discussion approfondie de Sharon L. Lohr sur les outils de diagnostic aux fins d'évaluation des hypothèses du modèle est très précieuse pour le sujet qui nous intéresse. Ses analyses des idées et des méthodes existantes et les adaptations au contexte actuel mettent en évidence les questions apparemment différentes, mais profondément liées, auxquelles font face les échantillons d'enquête probabilistes et non probabilistes. L'une de ces questions est le problème du sous-dénombrement (c'est-à-dire les violations de l'hypothèse A2) et la conjugaison des hypothèses A1 et A2. Sharon L. Lohr était à juste titre préoccupée par les estimateurs fondés sur les prédictions dans lesquels le modèle de prédiction de y étant donné $\mathbf{x}$ est construit à partir de l'échantillon non probabiliste S_A et l'estimateur d'imputation de masse est calculé au moyen du $\mathbf{x}$ observé dans l'échantillon probabiliste de référence, un scénario dans lequel chacune des deux hypothèses A1 et A2 n'est pas autonome. Le problème de sous-dénombrement est un exemple où « les procédures de l'ère spatiale ne sauveront pas les données de l'âge de pierre ». Sharon L. Lohr a préconisé de « prendre un petit échantillon probabiliste pour analyser les hypothèses », ce qui est nécessaire en théorie puisqu'il faut des échantillons de validation pour des méthodes rigoureusement défendables dans certains scénarios. Cependant, l'élaboration de stratégies de compromis avec les sources de données disponibles, bien qu'il s'agisse d'une approche plus attrayante, est plus difficile à mettre en pratique.

L'observation de Sharon L. Lohr selon laquelle « les échantillons non probabilistes peuvent améliorer l'équité des données » est importante, car l'inclusion d'unités de groupes susceptibles d'être invisibles dans les échantillons probabilistes peut être favorisée relativement facilement pour les échantillons non

probabilistes. Elle observe par ailleurs que « les groupes historiquement défavorisés pourraient être sous-représentés dans toutes les sources de données, y compris dans les échantillons non probabilistes ». Aborder la question de l'équité des données en présence d'échantillons d'enquête non probabiliste présente à la fois des possibilités et des défis.

Il est difficile de répondre à la question de Sharon L. Lohr « quand faut-il utiliser des échantillons non probabilistes ? ». Cette même question peut être posée à propos de toute méthode statistique. Il semblerait que nous ne remettions pas toujours en question la validité des méthodes et l'utilité des résultats dans de nombreux autres scénarios, car nous sommes convaincus que les hypothèses requises semblent raisonnables. Pour les échantillons non probabilistes, nous nous retrouvons dans une situation de vulnérabilité plus importante en ce qui a trait aux hypothèses, et les évaluations et les diagnostics de ces hypothèses sont plus difficiles que les cas présentant des expériences contrôlées ou des données plus structurées. De ce point de vue, l'analyse approfondie de Sharon L. Lohr sur l'évaluation des hypothèses doit être lue avec attention et reconnaissance. Dans la pratique, un examen scrupuleux de « l'étape d'élaboration du plan » est crucial pour renforcer la confiance à l'égard des hypothèses, si cette étape peut être conçue avant la collecte des données sur des variables qui pourraient être liées au comportement de participation et que l'on inclut ces variables dans l'échantillon en étudiant davantage les sources de données existantes contenant ces variables.

Xiao-Li Meng

L'exposé de Xiao-Li Meng, intitulé « La miniaturisation de la corrélation due à un défaut des données : une stratégie polyvalente de traitement des échantillons non probabilistes »), devrait être un document de travail autonome. Xiao-Li Meng a passé en revue un certain nombre de problèmes dans l'estimation d'une moyenne de population finie en cas d'échantillon non probabiliste, et a examiné des stratégies et des orientations permettant de construire un estimateur approximativement sans biais au moyen d'un concept central dit de *corrélation due à un défaut des données (cdd)*. Ses éléments de discussion fascinants invitent à la réflexion et ils susciteront certainement davantage de discussions et de projets de recherche sur les implications de la *cdd*. J'aimerais saisir l'occasion pour commenter brièvement la relation de la *cdd* avec trois concepts de base de l'échantillonnage probabiliste, soit la *stratégie d'échantillonnage*, le *sous-dénombrement* et *l'estimation assistée par modèle*. Il ne s'agit pas de nostalgie du bon vieux temps où l'échantillonnage probabiliste était la norme par excellence, mais plutôt d'une appréciation de l'évolution de la recherche sur l'échantillonnage d'enquête et de l'utilité potentielle de la *cdd* dans le traitement des échantillons d'enquête non probabilistes.

Le terme *stratégie d'échantillonnage* désigne la paire constituée par le plan de sondage et la méthode d'estimation (Thompson, 1997, section 2.4; Rao, 2005, section 3.1). Ces deux composantes qui vont de pair constituent la colonne vertébrale de la théorie classique de l'échantillonnage probabiliste. Aux fins d'estimation du total de population T_y de la variable étudiée y au moyen d'un échantillon probabiliste S avec les probabilités d'inclusion de premier ordre π_i, l'estimateur de Horvitz-Thompson $\hat{T}_{\text{унг}} = \sum_{i \in S} d_i y_i$

avec le poids $d_i = \pi_i^{-1}$ est l'estimateur sans biais unique au sein d'une classe d'estimateurs linéaires (Wu et Thompson, 2020). L'argument théorique du résultat est simple en raison des probabilités d'inclusion connues π_i selon le plan d'échantillonnage donné. Si l'on utilise la notation de Meng, la *cdd* comprend trois variables, à savoir la variable étudiée G, la variable de poids W, l'indicateur d'inclusion de l'échantillon R, et elle est définie comme le coefficient de corrélation de la population finie entre $\tilde{R} = RW$ et G. La *cdd* pose implicitement R et W comme une paire inséparable pour toute *stratégie d'inférence*, R correspondant au « plan » inconnu et W à la « méthode d'estimation ». En ayant recours au « plan » inconnu caractérisé par les « probabilités divines » inconnues π_I pour l'échantillon non probabiliste, Meng a montré par son équation (3.3), que $W_I \propto \pi_I^{-1}$ est essentiellement une condition requise pour une estimation sans biais de $\bar{G}$ si rien n'est supposé sur le modèle de régression des résultats. Le résultat permet de justifier l'utilisation de l'estimateur de pondération par l'inverse de la propension (PIP) pour les échantillons non probabilistes comme seul choix raisonnable s'il n'y a pas de modèle de superpopulation de la variable étudiée.

Le problème du *sous-dénombrement* a été largement traité dans la littérature portant sur l'échantillonnage probabiliste. Pour les échantillons non probabilistes, la question est étroitement liée à la violation de l'hypothèse de positivité A2, abordée dans la section 7.2 de mon article et mes commentaires sur les discussions de Bailey, Elliott et Lohr. Des précisions supplémentaires sont données dans Chen et coll. (2023). Soit $U = U_0 \cup U_1$, où U_1 est la sous-population non couverte avec $\pi_i^A = P(R_i = 1 \,|\, \mathbf{x}_i, y_i) = 0$. Soit $N = N_0 + N_1$, où N_0 et N_1 sont respectivement les tailles des deux sous-populations U_0 et U_1. Supposons que Cov_I et $\mathrm{Cov}_I^{(0)}$ désignent respectivement la covariance par rapport à la distribution uniforme discrète sur U et U_0. On peut montrer que :

$$\mathrm{Cov}_I(\tilde{R}_I, G_I) = \omega_0 \left\{ \mathrm{Cov}_I^{(0)}(\tilde{R}_I, G_I) - \omega_1 (\bar{G}_1 - \bar{G}_0)\, \hat{N}_0 / N_0 \right\}, \tag{2}$$

où $\omega_k = N_k / N$ pour $k = 0, 1$, $\hat{N}_0 = \sum_{i \in S} W_i$, S est l'ensemble d'unités pour l'échantillon non probabiliste, et $\bar{G}_0$ et $\bar{G}_1$ sont respectivement les moyennes de population de U_0 et U_1 pour la variable étudiée G. L'équation (2) a deux conséquences immédiates. Premièrement, si la méthode d'estimation est valide dans le sens où la valeur de $\mathrm{Cov}_I^{(0)}(\tilde{R}_I, G_I)$ est petite, alors le biais de l'estimateur $\bar{G}_W$ dû au sous-dénombrement dépend de ω_1 (c'est-à-dire la taille de la sous-population non couverte U_1) et $\bar{G}_1 - \bar{G}_0$ (c'est-à-dire la différence entre U_0 et U_1), un énoncé qui a déjà été établi selon l'échantillonnage probabiliste. Deuxièmement, l'équation révèle un scénario de *contrepoids* potentiel : Un estimateur biaisé $\bar{G}_W$ pour la « moyenne de la population échantillonnée » $\bar{G}_0$ peut être moins biaisé pour la moyenne de la population cible $\bar{G}$ si $\mathrm{Cov}_I^{(0)}(\tilde{R}_I, G_I)$ et $\bar{G}_1 - \bar{G}_0$ ont le même signe plus ou moins.

Les discussions de Meng sur la quasi-randomisation ou la superpopulation au moyen de la *cdd* apportent une compréhension nettement plus fine de l'estimation doublement robuste. Historiquement, *l'estimation assistée par un modèle* a commencé à apparaître dans l'échantillonnage d'enquête au début des années 1970; cette méthode est dans le même esprit que la double robustesse. L'estimateur par la

différence généralisée de la moyenne de population $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$, comme le présentent Cassel, Särndal et Wretman (1976), est donné par :

$$\hat{\mu}_{y\text{GD}} = \frac{1}{N}\left\{\sum_{i\in S}\frac{y_i - c_i}{\pi_i} + \sum_{i=1}^{N} c_i\right\}, \tag{3}$$

où S est un échantillon probabiliste, les π_i sont les probabilités d'inclusion de premier ordre et $\{c_1, c_2, \ldots, c_N\}$ est une séquence arbitraire de nombres connus. L'estimateur $\hat{\mu}_{y\text{GD}}$ est exactement sans biais pour μ_y selon le plan d'échantillonnage probabiliste p pour toute séquence c_i; il est également sans biais par rapport au modèle si nous choisissons $c_i = m_i = E_\xi(y_i \mid \mathbf{x}_i)$. Cassel et coll. (1976) ont présenté un résultat théorique principal selon lequel le choix $c_i = m_i$ est optimal, ce qui conduit à une espérance minimale fondée sur un modèle de la variance fondée sur le plan $E_\xi\{V_p(\hat{\mu}_{y\text{GD}})\}$ quand le modèle a une certaine structure de variance. La première partie des résultats sur l'absence de biais est sous (p ou ξ); la deuxième partie sur l'optimalité se trouve sous (p et ξ). Il convient de souligner que l'estimateur $\hat{\mu}_{y\text{GD}}$ avec le choix $c_i = \hat{m}_i$ a exactement la même structure que l'estimateur doublement robuste dont il est abondamment question dans la littérature sur les données manquantes et l'inférence causale depuis les années 1990, les « probabilités divines » π_i étant inconnues et estimées dans ces derniers cas.

Dans la pratique, l'emploi de la *cdd* nécessite des renseignements supplémentaires tirés de la population. La proposition de Meng de créer une miniature représentative à partir d'un échantillon biaisé fait écho à la demande d'un échantillon de validation de petite taille, puisqu'un tel échantillon « peut (aussi) éliminer l'anxiété de nombreux praticiens et les erreurs qu'ils sont susceptibles de commettre parce qu'ils ne maîtrisent pas l'utilisation des poids ».

« Il n'existe pas d'échantillon probabiliste dans le monde réel » est probablement un énoncé défendable pour les populations humaines. Néanmoins, les échantillons probabilistes existent dans d'autres domaines, comme les enquêtes-entreprises et les enquêtes auprès des établissements, les enquêtes relatives à l'agriculture et à l'inventaire de ressources naturelles; voir Wu et Thompson (2020) pour en savoir plus. En revanche, pour les êtres humains, toutes les règles rigoureuses ou les procédures précises sont *presque certainement* une aspiration, mais non une prescription.

Bibliographie

Basu, D. (1971). An essay on the logical foundations of survey sampling. Part One. Dans *Foundations of Statistical Inference*, (Éds., V.P. Godambe et D.A. Sprott), Toronto, 203-242.

Beaumont, J.-F. (2020). Les enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles ? *Techniques d'enquête*, 46, 1, 1-30. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2020001/article/00001-fra.pdf.

Cassel, C.M., Särndal, C.-E. et Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

Chen, Y., Li, P. et Wu, C. (2023). Dealing with undercoverage for non-probability survey samples. *Techniques d'enquête*, à l'étude.

Cho, S., Kim, J.K. et Qiu, Y. (2022). *Multiple Bias Calibration for Valid Statistical Inference with Selection Bias*. Document de travail.

Rao, J.N.K. (2005). Évaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage. *Techniques d'enquête*, 31, 2, 127-151. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2005002/article/9040-fra.pdf.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*, 83, 242-272.

Rivers, D. (2007). Sampling for web surveys. Dans *Proceedings of the Survey Research Methods Section*, Joint Statistical Meetings, American Statistical Association, Alexandrie, Virginie, 1-26.

Thompson, M.E. (1997). *Theory of Sample Surveys*. Chapman & Hall, Londres.

Wu, C., et Thompson, M.E. (2020). *Sampling Theory and Practice*. Springer, Cham.

Yuan, M., Li, P. et Wu, C. (2022). *Inference with Non-Ignorable Sample Inclusion for Non-Probability Survey Samples*. Document de travail.

Techniques d'enquête, décembre 2022
Vol. 48, nº 2, pp. 407-437
Statistique Canada, nº 12-001-X au catalogue

Les modèles d'apprentissage profond sont-ils plus efficaces pour l'imputation de données manquantes dans les enquêtes ? Une comparaison empirique fournit des éléments de preuve

Zhenhua Wang, Olanrewaju Akande, Jason Poulos et Fan Li[1]

Résumé

L'imputation multiple est une approche populaire pour traiter les données manquantes découlant de la non-réponse dans les enquêtes-échantillons. L'imputation multiple au moyen d'équations en séries (MICE) est l'un des algorithmes d'imputation multiple les plus utilisés pour les données multivariées, mais son fondement théorique est insuffisant et elle exige beaucoup de calculs. Récemment, des méthodes d'imputation des données manquantes fondées sur des modèles d'apprentissage profond ont été élaborées, ce qui a donné des résultats encourageants dans de petites études. Cependant, peu de recherches ont été menées sur l'évaluation de leur rendement dans des contextes réalistes par rapport à la MICE, en particulier dans le cadre de grandes enquêtes. Nous menons de vastes études de simulation fondées sur un sous-échantillon de l'*American Community Survey* afin de comparer les propriétés d'échantillonnage répété de quatre méthodes d'apprentissage automatique fondées sur l'imputation multiple : MICE avec arbres de classification; MICE avec forêts aléatoires; réseaux antagonistes génératifs pour l'imputation; et imputation multiple à l'aide d'autoencodeurs débruiteurs. Nous constatons que les méthodes d'imputation fondées sur des modèles d'apprentissage profond sont plus efficaces que la MICE en ce qui a trait au temps de calcul. Cependant, étant donné le choix par défaut des hyperparamètres dans les progiciels communs, la MICE avec arbres de classification dépasse constamment, souvent de loin, les méthodes d'imputation fondées sur l'apprentissage profond quant au biais, à l'erreur quadratique moyenne et à la couverture dans une gamme de paramètres réalistes.

Mots-clés : Apprentissage profond; données d'enquête; données manquantes; études par simulation; imputation multiple au moyen d'équations en séries; sélection des hyperparamètres.

1. Introduction

De nombreuses enquêtes-échantillons comportent des données manquantes découlant de la non-réponse totale, lorsqu'un sous-ensemble de participants ne répond pas à l'enquête, ou de la non-réponse partielle, lorsque les valeurs manquantes se limitent à des questions particulières. Dans les sondages d'opinion, la non-réponse peut indiquer le refus de révéler une préférence ou l'absence de préférence (De Leeuw, Hox et Huisman, 2003). Si elles ne sont pas traitées adéquatement, les données manquantes peuvent mener à des analyses statistiques biaisées, surtout lorsqu'il y a des différences systématiques entre les données observées et les données manquantes (Rubin, 1976; Little et Rubin, 2019). L'analyse de cas complets d'unités pour lesquelles les données sont entièrement observées est souvent impossible et peut entraîner un biais important dans la plupart des situations (Little et Rubin, 2019). Par conséquent, de nombreux analystes tiennent compte des données manquantes en imputant les valeurs manquantes, puis en procédant comme si les valeurs imputées étaient des valeurs réelles.

1. Zhenhua Wang est étudiant au doctorat au Department of Statistics, University of Missouri, Columbia, MO 65211. Courriel : zhenhua.wang@mail.missouri.edu; Olanrewaju Akande est chercheur chez Meta Platforms, Inc. Courriel : akandelanre13@gmail.com; Jason Poulos est associé de recherche postdoctorale au Department of Health Care Policy, Harvard Medical School, Boston, MA. Courriel : poulos@hcp.med.harvard.edu; Fan Li est professeur au Department of Statistical Science, Box 90251, Duke University, Durham, Caroline du Nord 27708, C.P. 90251. Courriel : fl35@duke.edu.

L'imputation multiple (Rubin, 1987) est une approche répandue pour traiter les valeurs manquantes. Dans le cadre de l'imputation multiple, un analyste crée $L>1$ ensembles de données complets en remplaçant les valeurs manquantes dans les données de l'échantillon par des tirages plausibles générés à partir de la distribution prédictive de modèles probabilistes s'appuyant sur les données observées. Pour chaque ensemble de données complet, l'analyste peut ensuite calculer les estimations de l'échantillon pour les paramètres de la population d'intérêt et combiner les estimations de l'échantillon pour tous les L ensembles de données à l'aide des méthodes d'inférence de l'imputation multiple mises au point dans Rubin (1987) et, plus récemment, dans Rubin (1996), Barnard et Meng (1999), Reiter et Raghunathan (2007) et Harel et Zhou (2007). Dans l'imputation multiple, la variance estimée d'un paramètre consiste en des variances intra-imputation et inter-imputation, et tient donc compte de la variabilité inhérente des valeurs imputées. Il convient de prendre note que dans les études fondées sur des données d'enquête, l'imputation unique, par exemple par couplage ou régression, demeure courante pour le traitement des données manquantes, où la variance est estimée au moyen de la méthode delta ou des méthodes de rééchantillonnage (Chen et Haziza, 2019; Haziza et Vallée, 2020).

1.1 Imputation par modèle

Il existe deux stratégies générales de modélisation pour l'imputation multiple. La première stratégie, connue sous le nom de *modélisation conjointe*, consiste à préciser une distribution conjointe pour toutes les variables des données, puis à générer des imputations à partir des distributions (prédictives) conditionnelles implicites des variables pour lesquelles il manque des valeurs (Schafer, 1997). La stratégie de modélisation conjointe s'aligne sur le fondement théorique de Rubin (1987), mais il peut être difficile de spécifier un modèle conjoint ayant des variables de grande dimension de différents types. En effet, les méthodes de modélisation conjointe les plus reconnues, comme « PROC MI » dans SAS (Yuan, 2011), et « AMELIA » (Honaker, King et Blackwell, 2011) et « norm » dans `R` (Schafer, 1997), formulent une hypothèse simplificatrice selon laquelle les données suivent des distributions gaussiennes multivariées, même pour les variables catégoriques, ce qui peut entraîner un biais (Horton, Lipsitz et Parzen, 2003). Des études récentes ont permis de mettre au point des modèles de modélisation conjointe souples fondés sur des modèles bayésiens non paramétriques avancés comme le modèle par mélange selon le processus de Dirichlet (Manrique-Vallier et Reiter, 2014; Murray et Reiter, 2016). Cependant, ces méthodes sont coûteuses sur le plan des calculs et s'adaptent souvent mal aux cas ayant des variables de grande dimension.

La deuxième stratégie est la *spécification entièrement conditionnelle* (van Buuren, Brand, Groothuis-Oudshoorn et Rubin, 2006), où l'on précise séparément une distribution conditionnelle univariée pour chaque variable ayant des valeurs manquantes, compte tenu de toutes les autres variables, et où l'on impute les valeurs manquantes variable par variable de façon itérative, d'une manière s'apparentant à

l'échantillonneur de Gibbs. La méthode de la spécification entièrement conditionnelle la plus répandue est l'imputation multiple au moyen d'équations en séries (MICE) (van Buuren et Groothuis-Oudshoorn, 2011), habituellement mise en œuvre par la spécification des modèles linéaires généralisés (GLM) pour les distributions conditionnelles univariées (Raghunathan, Lepkowski, Van Hoewyk et Solenberger, 2001; Royston et White, 2011; Su, Gelman, Hill et Yajima, 2011). Des études récentes indiquent que la spécification des modèles conditionnels par arbres de classification et de régression (CART) (Breiman, Friedman, Olshen et Stone, 1984; Burgette et Reiter, 2010) dépasse largement la MICE par les GLM (Akande, Li et Reiter, 2017). Une extension naturelle de la MICE avec CART consiste à utiliser des méthodes s'appuyant sur un ensemble d'arbres, comme les forêts aléatoires, plutôt d'utiliser un seul arbre (Breiman, 2001; Doove, van Buuren et Dusseldorp, 2014).

La MICE est intéressante dans les données d'enquête à grande échelle parce qu'elle permet l'imputation de différents types de variables de manière simple et souple. Cependant, la MICE présente un inconvénient théorique clé, à savoir l'incompatibilité possible des distributions conditionnelles spécifiées, c'est-à-dire que les distributions ne correspondent pas à une distribution conjointe (Arnold et Press, 1989; Gelman et Speed, 1993; Li, Yu et Rubin, 2012). Malgré cet inconvénient, la MICE fonctionne remarquablement bien dans des applications réelles, et de nombreuses simulations ont démontré qu'elle est plus efficace que de nombreuses méthodes de modélisation conjointe théoriquement solides; voir van Buuren (2018) pour obtenir des études de cas. Cependant, la MICE nécessite aussi de nombreux calculs (White, Royston et Wood, 2011) et ne peut généralement pas être mise en parallèle. De plus, les progiciels reconnus pour la mise en œuvre de la MICE par les GLM, par exemple `mice` dans `R` (van Buuren et Groothuis-Oudshoorn, 2011), boguent souvent dans des contextes comportant des variables non continues de grande dimension, comme des variables catégoriques comportant de nombreuses catégories (Akande et coll., 2017).

1.2 Imputation au moyen de modèles d'apprentissage profond

Les récents progrès de l'apprentissage profond élargissent considérablement la portée des modèles complexes pour les données de grande dimension. Ces progrès apportent l'espoir qu'une nouvelle génération de méthodes d'imputation des données manquantes fondées sur des modèles d'apprentissage profond puisse tenir compte des limites théoriques et computationnelles des méthodes statistiques existantes. Par exemple, les modèles génératifs profonds comme les réseaux antagonistes génératifs (GAN) (Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville et Bengio, 2014) sont naturellement adaptés à la production d'imputations multiples parce qu'ils sont conçus pour générer des données qui ressemblent le plus possible aux données observées. Une méthode dans ce volet est le réseau antagoniste génératif pour l'imputation (GAIN) de Yoon, Jordon et Schaar (2018). L'imputation multiple à l'aide d'autoencodeurs débruiteurs (MIDA) (Gondara et Wang, 2018; Lu, Perrone et Unpingco, 2020)

est une autre méthode générative basée sur des réseaux neuronaux profonds formés à traiter des données d'entrée corrompues afin de forcer les réseaux à apprendre une représentation de faible dimension utile des données d'entrée plutôt que leur fonction d'identité (Vincent, Larochelle, Bengio et Manzagol, 2008; Vincent, Larochelle, Lajoie, Bengio, Manzagol et Bottou, 2010). Plusieurs méthodes ont été proposées pour l'imputation des valeurs manquantes dans les données chronologiques au moyen d'autoencodeurs variationnels (Fortuin, Baranchuk, Rätsch et Mandt, 2020) ou de réseaux neuronaux récurrents (Lipton, Kale et Wetzel, 2016; Monti, Bronstein et Bresson, 2017; Cao, Wang, Li, Zhou, Li et Li, 2018; Che, Purushotham, Cho, Sontag et Liu, 2018; Yoon, Zame et van der Schaar, 2018).

Les méthodes d'imputation multiple fondées sur l'apprentissage profond présentent plusieurs avantages, du moins théoriquement, par rapport aux modèles statistiques conventionnels, notamment : i) elles permettent d'éviter de faire des hypothèses de distribution; ii) elles permettent de facilement traiter des types de données mixtes; iii) elles permettent de modéliser des relations non linéaires entre les variables; iv) elles sont censées bien fonctionner dans des paramètres de grande dimension; et v) elles peuvent tirer parti de la puissance du processeur graphique pour accélérer le calcul. Plusieurs études font état d'un rendement encourageant des méthodes d'imputation multiple fondées sur l'apprentissage profond par rapport à la MICE (par exemple Yoon, Jordon et Schaar, 2018). Toutefois, les conclusions de ces études sont fondées sur des données probantes limitées. Premièrement, les études sont habituellement fondées sur de petites simulations ou sur plusieurs ensembles de données de « référence » publics bien étudiés, comme ceux décrits à la section 5, qui ne ressemblent pas aux données d'enquête. Deuxièmement, les évaluations sont habituellement fondées sur quelques mesures de rendement globales, comme l'erreur quadratique moyenne de prédiction globale ou l'exactitude. Il est possible que ces mesures ne donnent pas un tableau complet des comparaisons, et elles peuvent même être parfois trompeuses, comme nous le verrons plus loin. Troisièmement, étant donné l'incertitude du processus de données manquantes, il est essentiel d'examiner les propriétés d'échantillonnage répété des méthodes d'imputation, mais celles-ci ont rarement été évaluées. Enfin, le réglage des hyperparamètres est crucial pour les modèles d'apprentissage automatique, et différents réglages peuvent donner des résultats radicalement différents, par contre peu de renseignements sont fournis sur le réglage des hyperparamètres et ses conséquences sur le rendement des méthodes d'imputation.

Motivés par ces limites, nous effectuons, dans la présente étude, des simulations exhaustives fondées sur des données d'enquête réelles afin d'évaluer les méthodes d'imputation multiple au moyen d'une gamme de mesures de rendement. Plus précisément, nous effectuons des simulations fondées sur un sous-échantillon de l'*American Community Survey* pour comparer les propriétés d'échantillonnage répété de quatre méthodes d'imputation multiple mentionnées précédemment : MICE avec CART (MICE-CART), MICE avec des forêts aléatoires (MICE-RF), GAIN et MIDA. Nous avons découvert que les méthodes d'imputation multiple fondées sur l'apprentissage profond sont plus efficaces que la MICE en ce qui a

trait au temps de calcul. Cependant, la MICE-CART dépasse constamment, souvent de loin, les méthodes d'apprentissage profond quant au biais, à l'erreur quadratique moyenne et à la couverture dans une gamme de paramètres réalistes. Cette observation contredit les constatations antérieures dans la littérature sur l'apprentissage automatique et soulève des questions sur les mesures appropriées pour évaluer les méthodes d'imputation. Elle fait également ressortir l'importance d'évaluer les propriétés d'échantillonnage répété des méthodes d'imputation. Bien que nous mettions l'accent sur l'imputation multiple dans la présente étude, nous remarquons que les méthodes d'imputation multiple mentionnées précédemment sont facilement applicables pour générer une imputation unique lorsque L est établi à 1. Des preuves empiriques exhaustives indiquent que la variance intra-imputation domine habituellement la variance inter-imputation dans l'imputation multiple. Par conséquent, nous nous attendons à ce que les schémas entre les différentes méthodes d'imputation observées soient également valables si ces méthodes sont utilisées pour une seule imputation.

Le reste de la présente étude est organisé comme suit. À la section 2, nous examinons les quatre méthodes d'imputation multiple utilisées dans notre évaluation. À la section 3, nous décrivons un cadre comportant plusieurs mesures pour évaluer les méthodes d'imputation. À la section 4, nous décrivons le plan de simulation et les résultats à l'aide de données d'enquête à grande échelle, et à la section 5, nous résumons les résultats d'évaluation des ensembles de données de référence utilisés dans la documentation sur l'apprentissage automatique. Enfin, à la section 6, nous concluons en présentant un guide pratique pour la mise en œuvre dans des applications réelles.

2. Méthode d'imputation des données manquantes

Nous présentons d'abord la notation. Prenons un échantillon ayant n unités, dont chacune est associée à p variables. Si Y_{ij} correspond à la valeur de la variable j pour chaque i, où $j = 1, \ldots, p$ et $i = 1, \ldots, n$. Ici, Y peut être continu, binaire, catégorique ou binaire-continu. Pour chaque i, supposons que $\mathbf{Y}_i = \left(Y_{i1}, \ldots, Y_{ip}\right)$. Pour chaque variable j, supposons que $\mathbf{Y}_j = \left(Y_{1j}, \ldots, Y_{nj}\right)$. Supposons que $\mathbf{Y} = \left(\mathbf{Y}_1, \ldots, \mathbf{Y}_n\right)$ correspond à la matrice $n \times p$ comprenant les données pour tous les enregistrements inclus dans l'échantillon. Nous écrivons $\mathbf{Y} = \left(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}\right)$, où $\mathbf{Y}_{\text{obs}}$ et $\mathbf{Y}_{\text{mis}}$ sont les parties observées et manquantes de $\mathbf{Y}$, respectivement. Nous écrivons $\mathbf{Y}_{\text{mis}} = \left(\mathbf{Y}_{\text{mis},1}, \ldots, \mathbf{Y}_{\text{mis},p}\right)$, où $\mathbf{Y}_{\text{mis},j}$ représente toutes les valeurs manquantes pour la variable j, et $j = 1, \ldots, p$. De même, nous écrivons $\mathbf{Y}_{\text{obs}} = \left(\mathbf{Y}_{\text{obs},1}, \ldots, \mathbf{Y}_{\text{obs},p}\right)$ pour les données observées correspondantes.

Lors de l'imputation multiple, l'analyste génère les valeurs des données manquantes $\mathbf{Y}_{\text{mis}}$ au moyen de modèles prédéfinis estimés avec $\mathbf{Y}_{\text{obs}}$, ce qui donne un ensemble de données complet. L'analyste répète ensuite le processus pour générer L ensembles de données complets, $\left\{\mathbf{Y}^{(l)} : l = 1, \ldots, L\right\}$, qui sont accessibles aux fins d'inférence ou de diffusion. Pour l'inférence, l'analyste peut calculer des estimations

de l'échantillon pour les paramètres de la population dans chaque ensemble de données complet $\mathbf{Y}^{(l)}$ et il peut les combiner au moyen des règles d'inférence de l'imputation multiple élaborées par Rubin (1987), qui seront examinées à la section 3.

2.1 L'imputation multiple au moyen d'équations en séries et les modèles d'arbres de classification

Dans la MICE, l'analyste commence par préciser un modèle conditionnel univarié distinct pour chaque variable pour laquelle il manque des valeurs. L'analyste précise ensuite un ordre d'itération dans la séquence des modèles conditionnels, au moment de l'imputation. La liste ordonnée des variables s'écrit $\left(\mathbf{Y}_{(1)},\ldots,\mathbf{Y}_{(p)}\right)$. Ensuite, l'analyste initialise chaque $\mathbf{Y}_{\text{mis},(j)}$. Les options les plus populaires consistent à échantillonner à partir de : i) la distribution marginale de la variable $\mathbf{Y}_{\text{obs},(j)}$ correspondante; ou ii) la distribution conditionnelle de la variable $\mathbf{Y}_{(j)}$, compte tenu de toutes les autres variables, construites uniquement à l'aide des cas disponibles.

Après l'initialisation, l'algorithme de la MICE suit un processus itératif qui parcourt la séquence des modèles univariés. Pour chaque variable j à chaque itération t, on fait correspondre le modèle conditionnel $(\mathbf{Y}_{(j)} \,|\, \mathbf{Y}_{\text{obs},(j)}, \{\mathbf{Y}_{(k)}^{(t)}\colon k<j\}, \{\mathbf{Y}_{(k)}^{(t-1)}\colon k>j\})$. Ensuite, on remplace $\mathbf{Y}_{\text{mis},(j)}^{(t)}$ par des tirages à partir du modèle implicite $(\mathbf{Y}_{\text{mis},(j)}^{(t)} \,|\, \mathbf{Y}_{\text{obs},(j)}, \{\mathbf{Y}_{(k)}^{(t)}\colon k<j\}, \{\mathbf{Y}_{(k)}^{(t-1)}\colon k>j\})$. Le processus itératif se poursuit pendant T itérations totales jusqu'à la convergence, et les valeurs de l'itération finale constituent un ensemble de données complet $\mathbf{Y}^{(l)} = \left(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(T)}\right)$. L'ensemble du processus est ensuite répété L fois pour créer les L ensembles de données complets. Nous fournissons un pseudocode détaillant chaque étape de l'algorithme de la MICE dans la documentation supplémentaire.

Dans la MICE-CART, l'analyste utilise un CART (Breiman et coll., 1984) pour les modèles conditionnels univariés dans l'algorithme de la MICE. Le CART suit une structure d'arbre décisionnel qui s'appuie sur des parties fractionnées binaires récursives pour répartir l'espace de prévision en régions distinctes sans chevauchement. La partie supérieure de l'arbre représente souvent sa racine et chaque partie fractionnée binaire successive divise l'espace de prévision en deux nouvelles branches au fur et à mesure que l'on descend l'arbre. Le critère de fractionnement à chaque feuille est habituellement choisi pour réduire au minimum une mesure d'entropie théorique de l'information. Les parties fractionnées qui ne diminuent pas l'ajustement insuffisant d'une quantité raisonnable par rapport à un seuil fixe sont élaguées. L'arbre est ensuite construit jusqu'à ce qu'un critère d'arrêt soit satisfait; par exemple un nombre minimal d'observations dans chaque feuille.

Une fois l'arbre entièrement construit, on génère $\mathbf{Y}_{\text{mis},(j)}^{(t)}$ en faisant descendre l'arbre jusqu'à la feuille appropriée au moyen des combinaisons dans $(\{\mathbf{Y}_k^{(t)}\colon k<j\}, \{\mathbf{Y}_k^{(t-1)}\colon k>j\})$, puis en échantillonnant à partir des $Y_{(j)}^{\text{obs}}$ valeurs de cette feuille. Autrement dit, pour toute combinaison dans $(\{\mathbf{Y}_k^{(t)}\colon k<j\}, \{\mathbf{Y}_k^{(t-1)}\colon k>j\})$, on utilise la proportion de valeurs de $\mathbf{Y}_j^{\text{obs}}$ dans la feuille correspondante pour estimer la

distribution conditionnelle $(\mathbf{Y}_{(j)} \,|\, \mathbf{Y}_{\text{obs},(j)}, \{\mathbf{Y}_{(k)}^{(t)}: k < j\}, \{\mathbf{Y}_{(k)}^{(t-1)}: k > j\})$. Le processus itératif se poursuit à nouveau pendant T itérations totales, et les valeurs de l'itération finale constituent un ensemble de données complet.

Dans le cadre de la MICE-RF, on utilise plutôt des forêts aléatoires pour les modèles conditionnels univariés dans la MICE (par exemple Stekhoven et Bühlmann, 2012; Shah, Bartlett, Carpenter, Nicholas et Hemingway, 2014). Les forêts aléatoires (Ho, 1995; Breiman, 2001) constituent une méthode s'appuyant sur un ensemble d'arbres qui construit des arbres décisionnels multiples à partir des données, plutôt qu'un seul arbre comme le CART. Plus précisément, les forêts aléatoires construisent plusieurs arbres décisionnels au moyen d'échantillons bootstrap de l'arbre original, et ne s'appuient que sur un échantillon des facteurs prédictifs pour les partitions récursives dans chaque arbre. Cette approche peut réduire considérablement la prévalence des arbres instables ainsi que la corrélation entre les arbres individuels, puisqu'elle empêche les mêmes variables de dominer le processus de partitionnement dans l'ensemble des arbres. En théorie, cette décorrélation devrait se traduire par des prévisions ayant moins de variance (Hastie, Tibshirani et Friedman, 2009).

Pour l'imputation, l'analyste entraîne d'abord un modèle de forêts aléatoires pour chaque $\mathbf{Y}_{(j)}$ à l'aide des cas disponibles et compte tenu de toutes les autres variables. Ensuite, l'analyste génère des prévisions pour $\mathbf{Y}_{\text{mis},j}$ dans ce modèle. Plus précisément, pour toute valeur $\mathbf{Y}_{(j)}$ catégorique, et compte tenu de toute combinaison particulière dans $(\{\mathbf{Y}_k^{(t)}: k < j\}, \{\mathbf{Y}_k^{(t-1)}: k > j\})$, l'analyste génère d'abord des prévisions pour chaque arbre en fonction des valeurs $\mathbf{Y}_j^{\text{obs}}$ de la feuille correspondante pour cet arbre, puis il utilise le niveau majoritaire le plus courant de toutes les prévisions de tous les arbres. Pour une valeur $\mathbf{Y}_{(j)}$ continue, l'analyste utilise plutôt la moyenne de toutes les prédictions de tous les arbres. Le processus itératif se poursuit de nouveau pour toutes les variables, pendant T itérations totales, et les valeurs de l'itération finale constituent un ensemble de données complet. Un hyperparamètre particulièrement important dans les forêts aléatoires est le nombre maximum d'arbres d.

Pour nos évaluations, nous utilisons le progiciel `mice` R pour mettre en œuvre la MICE-CART et la MICE-RF, et nous conservons le réglage d'hyperparamètres par défaut dans le progiciel pour imiter la pratique courante dans des applications réelles. Plus précisément, nous avons fixé le nombre minimum d'observations dans chaque feuille terminale à 5 et le seuil d'élagage à 0,0001 dans la MICE-CART. Dans la MICE-RF, le nombre maximum d'arbres d est fixé à 10.

2.2 Réseau antagoniste génératif pour l'imputation

Le GAIN (Yoon, Jordon et Schaar, 2018) est une méthode d'imputation fondée sur les GAN (Goodfellow et coll., 2014), qui consiste en une fonction génératrice G et une fonction discriminante D. Pour toute matrice de données $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$, nous remplaçons $\mathbf{Y}_{\text{mis}}$ par du bruit aléatoire, Z_{ij}, échantillonné à partir d'une distribution uniforme. La fonction génératrice G permet d'entrer ces données

initialisées et une matrice de masque $\mathbf{M}$, dans laquelle $M_{ij} \in \{0,1\}$ indique les valeurs observées de $\mathbf{Y}$, et elle génère les valeurs estimées des données observées et des données manquantes, $\hat{\mathbf{Y}}$. La fonction discriminante D utilise $\hat{\mathbf{Y}} = \left(\mathbf{Y}_{\text{obs}}, \hat{\mathbf{Y}}_{\text{mis}}\right)$ et une matrice d'indice $\mathbf{H}$ de la même dimension pour déterminer les valeurs observées ou imputées par G, ce qui donne une matrice de masque prédite $\hat{\mathbf{M}}$. La matrice d'indice, échantillonnée à partir de la distribution de Bernoulli dans laquelle p est égal à un hyperparamètre de « taux d'indice », révèle à D des renseignements partiels à propos de $\mathbf{M}$ afin d'aider à guider G dans l'apprentissage de la distribution sous-jacente de $\mathbf{Y}$.

Nous entraînons d'abord D à réduire au minimum la fonction de perte, $L_D\left(\mathbf{M}, \hat{\mathbf{M}}\right)$, pour chaque mini-lot de taille n_i :

$$L_D\left(\mathbf{M}, \hat{\mathbf{M}}\right) = \sum_{i=1}^{n_i}\sum_{j=1}^{J} M_{ij} \log\left(\hat{M}_{ij}\right) + \left(1 - M_{ij}\right) \log\left(1 - \hat{M}_{ij}\right). \tag{2.1}$$

Ensuite, G est entraîné pour réduire au minimum la fonction de perte (2.2), qui se compose d'une perte de la fonction génératrice, $L_G\left(\mathbf{M}, \hat{\mathbf{M}}\right)$, et d'une perte de la fonction de reconstruction, $L_M\left(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{M}\right)$. La perte de la fonction génératrice (2.3) est réduite au minimum lorsque D définit incorrectement les valeurs imputées en valeurs observées. La perte de la fonction de reconstruction (2.4) est réduite au minimum lorsque les valeurs prédites sont semblables aux valeurs observées, et elle est pondérée par l'hyperparamètre β :

$$L\left(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{M}, \hat{\mathbf{M}}\right) = L_G\left(\mathbf{M}, \hat{\mathbf{M}}\right) + \beta L_M\left(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{M}\right), \tag{2.2}$$

$$L_G\left(\mathbf{M}, \hat{\mathbf{M}}\right) = \sum_{i=1}^{n_i}\sum_{j=1}^{J} M_{ij} \log\left(1 - \hat{M}_{ij}\right), \tag{2.3}$$

$$L_M\left(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{M}\right) = \sum_{i=1}^{n_i}\sum_{j=1}^{J} \left(1 - M_{ij}\right) L_{\text{rec}}\left(Y_{ij}, \hat{Y}_{ij}\right), \tag{2.4}$$

où

$$L_{\text{rec}}\left(Y_{ij}, \hat{Y}_{ij}\right) = \begin{cases} \left(\hat{Y}_{ij} - Y_{ij}\right)^2 & \text{si } Y_{ij} \text{ est continu} \\ -Y_{ij} \log \hat{Y}_{ij} & \text{si } Y_{ij} \text{ est catégorique.} \end{cases} \tag{2.5}$$

Dans nos expériences, nous modélisons à la fois G et D en tant que réseaux neuronaux entièrement connectés, chaque fonction ayant trois couches cachées, et θ unités cachées par couche cachée. Les poids des couches cachées sont initialisés uniformément au hasard selon la méthode d'initialisation de Xavier (Glorot et Bengio, 2010). Nous utilisons une fonction d'activation Leaky ReLU (Maas, Hannun et Ng, 2013) pour chaque couche cachée, et une fonction d'activation Softmax pour la couche de sortie de G dans le cas de variables catégoriques, ou une fonction d'activation sigmoïde dans le cas de variables

numériques et pour la couche de sortie de D. Nous facilitons ce choix de couche de sortie pour les variables numériques en transformant toutes les variables continues pour qu'elles soient dans la fourchette de $(0, 1)$ au moyen de la normalisation MinMax : $Y_{ij}^{*} = \{Y_{ij} - \min(Y_{\cdot j})\} \big/ \{\max(Y_{\cdot j}) - \min(Y_{\cdot j})\}$, où $\min(Y_{\cdot j})$ et $\max(Y_{\cdot j})$ sont respectivement le minimum et le maximum de la variable j. Après l'imputation, nous ramenons chaque valeur à son échelle d'origine. Nous générons des imputations multiples à l'aide de plusieurs exécutions du modèle et d'une imputation initiale variable des valeurs manquantes.

Pour mettre en œuvre le GAIN dans nos évaluations, nous utilisons la même architecture que celle de Yoon, Jordon et Schaar (2018). Nous établissons $\beta = 100$, θ étant égal au nombre de caractéristiques des données d'entrée, et nous ajustons le taux d'indice dans une seule simulation. Conformément à la pratique courante trouvée dans la littérature sur les GAN (Berthelot, Schumm et Metz, 2017; Ham, Jun et Kim, 2020), nous suivons l'évolution des pertes des fonctions génératrices et discriminantes du GAIN, et nous ajustons manuellement le taux d'indice de sorte que les deux pertes soient qualitativement similaires. Plus précisément, nous sélectionnons d'abord sommairement le taux d'indice parmi {0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9}. Nous déterminons ensuite la valeur finale en réalisant une étape d'ajustement supplémentaire. Dans le scénario des données manquantes au hasard (DMH), par exemple, après avoir observé que la valeur optimale se situe dans la fourchette (0,1; 0,2), nous effectuons une recherche parmi les valeurs {0,11; 0,12; 0,13; 0,14; 0,15; 0,16; 0,17; 0,18; 0,19}. Enfin, nous attribuons une valeur de 0,3 et 0,13 au taux d'indice optimal des scénarios de données manquantes complètement au hasard (DMCH) et de DMH, respectivement. Nous entraînons les réseaux pour 200 périodes au moyen de la descente par gradient stochastique et des mini-lots de taille 512 pour apprendre les poids des paramètres. Nous utilisons l'optimiseur Adam pour adapter le taux d'apprentissage, lequel se situe initialement à 0,001 (Kingma et Ba, 2014).

2.3 L'imputation multiple au moyen d'autoencodeurs débruiteurs

L'imputation multiple au moyen d'autoencodeurs débruiteurs (MIDA) (Gondara et Wang, 2018; Lu et coll., 2020) étend l'utilisation d'une classe de réseaux neuronaux, les encodeurs débruiteurs, à l'imputation multiple. Un autoencodeur est un modèle de réseau neuronal entraîné à apprendre la fonction de lien identité des données d'entrée. Les autoencodeurs débruiteurs corrompent intentionnellement les données d'entrée afin d'empêcher les réseaux d'apprendre la fonction de lien identité, pour qu'ils apprennent plutôt une représentation de faible dimension utile des données d'entrée. L'architecture de la MIDA se compose d'un encodeur et d'un décodeur, chacun étant modélisé comme un réseau neuronal entièrement connecté ayant trois couches cachées et θ unités cachées par couche cachée. Nous effectuons d'abord une imputation initiale des valeurs manquantes en utilisant la moyenne pour les variables continues et l'étiquette la plus fréquente pour les variables catégoriques, ce qui donne un ensemble de

données complet $\mathbf{Y}_0$. L'encodeur introduit $\mathbf{Y}_0$ et corrompt les données d'entrée en éliminant aléatoirement la moitié des variables. Les données d'entrée corrompues sont appliquées à une représentation dimensionnelle plus grande par l'ajout de Θ unités cachées à chaque couche cachée successive de l'encodeur. Le décodeur reçoit les données de sortie de l'encodeur et ajuste symétriquement le codage à la dimension d'entrée d'origine. Toutes les couches cachées s'appuient sur une fonction d'activation, la tangente hyperbolique (tanh), tandis que la couche de sortie du décodeur repose sur une fonction d'activation Softmax (sigmoïde) dans le cas de variables catégoriques (numériques). Les imputations multiples sont générées par plusieurs exécutions dans lesquelles les poids des couches cachées sont initialisés sous forme de variable aléatoire gaussienne.

Selon Lu et coll. (2020), nous entraînons la MIDA en deux phases : une phase primaire et une phase de mise au point. Dans la phase primaire, nous transmettons les données initialement imputées à la MIDA et nous l'entraînons pendant N_{prime} périodes. Dans la phase de mise au point, la MIDA est entraînée en fonction des données de sortie de la phase primaire pendant N_{tune} périodes, et elle produit le résultat. La fonction de perte est utilisée dans les deux phases et ressemble étroitement à la perte de la fonction de reconstruction dans le GAIN :

$$L\left(Y_{ij_0}, \hat{Y}_{ij}, M_{ij}\right) = \begin{cases} \left(1-M_{ij}\right)\left(Y_{ij_0}-\hat{Y}_{ij}\right)^2 & \text{si } Y_{ij} \text{ est continu} \\ -\left(1-M_{ij}\right) Y_{ij_0} \log \hat{\mathbf{Y}}_{ij} & \text{si } Y_{ij} \text{ est catégorique.} \end{cases} \tag{2.6}$$

Pour mettre en œuvre la MIDA dans nos évaluations, nous utilisons la même architecture et nous ajustons les hyperparamètres dans une seule simulation, comme dans Lu et coll. (2020). Nous traçons l'évolution de la fonction de perte L et sélectionnons le nombre d'unités supplémentaires Θ parmi {1; 2; 3; 4; 5; 6; 7; 8; 9; 10} pour réduire la perte. Dans nos expériences, nous établissons θ comme étant équivalent au nombre de caractéristiques des données d'entrée et nous ajoutons $\Theta = 7$ unités cachées à chacune des trois couches cachées de l'encodeur. Nous entraînons le modèle pendant $N_{\text{prime}} = 100$ périodes dans la phase primaire et pendant $N_{\text{tune}} = 2$ périodes dans la phase de mise au point. Comme dans le GAIN, nous apprenons les paramètres du modèle à l'aide de la descente par gradient stochastique comprenant des mini-lots de taille 512, et nous utilisons l'optimiseur Adam pour adapter le taux d'apprentissage au taux initial de 0,001.

3. Évaluation par simulation des méthodes d'imputation

Les méthodes d'imputation des données manquantes sont habituellement évaluées au moyen de simulations fondées sur des données réelles (van Buuren, 2018). Plus précisément, on crée des valeurs manquantes à partir d'un ensemble de données complet selon un mécanisme de données manquantes (Little et Rubin, 2014), on impute les valeurs manquantes à l'aide d'une méthode particulière, puis on compare ces valeurs imputées aux « vraies » valeurs originales en fonction de certaines mesures.

Nous allons d'abord passer rapidement en revue les règles de Rubin sur les combinaisons de l'imputation multiple. Supposons que Q soit le paramètre cible dans la population, et que $q^{(l)}$ et $u^{(l)}$ soient, respectivement, l'estimateur ponctuel et l'estimateur de la variance de Q en fonction du l^{e} ensemble de données imputé. L'estimateur ponctuel de l'imputation multiple de Q est $\bar{q}_L = \sum_{l=1}^{L} q^{(l)}/L$, et l'estimateur correspondant de la variance est égale à $T_L = (1+1/L)\, b_L + \bar{u}_L$, où $b_L = \sum_{l=1}^{L}\left(q^{(l)} - \bar{q}_L\right)^2/(L-1)$, et $\bar{u}_L = \sum_{l=1}^{L} u^{(l)}/L$. L'intervalle de confiance de Q est construit à l'aide de $(\bar{q}_L - Q) \sim t_v(0, T_L)$, où t_v est une distribution t ayant $v = (L-1)\left(1+\bar{u}_L/\left[(1+1/L)\, b_L\right]\right)^2$ degrés de liberté.

Premièrement, notre procédure d'évaluation par simulation consiste à choisir un ensemble de données ayant toutes les valeurs observées, qui est considéré comme la « population ». Nous choisissons ensuite un ensemble de paramètres cibles Q et nous calculons leurs valeurs à partir de ces données sur la population, qui sont considérées comme la « réalité du terrain ». Les paramètres sont habituellement des statistiques sommaires des variables ou des paramètres dans un modèle d'analyse en aval, par exemple un coefficient dans un modèle de régression (Tang, Song, Belin et Unützer, 2005; Huque, Carlin, Simpson et Lee, 2018). Deuxièmement, nous tirons au hasard sans remise H échantillons de taille n à partir des données de la population et, dans chacun des échantillons $(h = 1, \ldots, H)$, nous créons des données manquantes en fonction d'un mécanisme de données manquantes précis et d'une proportion préétablie de valeurs manquantes. Troisièmement, pour chaque échantillon simulé pour lequel il manque des données, nous créons L ensembles de données imputés à l'aide de la méthode d'imputation à l'étude et construisons l'estimateur ponctuel et l'estimateur par intervalles de chaque paramètre au moyen des règles de Rubin. Enfin, nous calculons les indicateurs de rendement de chaque paramètre à partir des quantités obtenues à l'étape précédente.

Dans l'application empirique, nous sélectionnons un grand sous-échantillon complet à partir de l'*American Community Survey* (ACS), une enquête nationale qui porte sur les caractéristiques distinctives de nombreuses grandes données d'enquête et qui constitue notre population. Étant donné que les variables discrètes sont répandues dans l'ACS, ainsi que dans la plupart des données d'enquête, nous nous concentrons sur les probabilités marginales des variables binaires et catégoriques, par exemple une variable catégorique ayant K catégories possède $K-1$ paramètres. Afin d'évaluer dans quelle mesure les méthodes d'imputation préservent les propriétés de distribution multivariée, comme dans Akande et coll. (2017), nous tenons également compte des probabilités bivariées de toutes les combinaisons bidirectionnelles de catégories dans les variables binaires et catégoriques. Une autre mesure utile est la corrélation par paires d'échantillons finis entre des variables continues. Pour les variables continues, les paramètres communs sont la moyenne, la médiane ou la variance. Pour faciliter des comparaisons importantes des résultats entre les variables catégoriques et continues, nous proposons de discrétiser chaque variable continue en K catégories en fonction des quantiles de l'échantillon. Nous évaluons

ensuite ces variables continues regroupées en classes en tant que variables catégoriques en fonction des paramètres des probabilités marginales et bivariées mentionnés précédemment.

Pour chaque paramètre Q, nous considérons trois mesures. La première mesure met l'accent sur le biais. Pour tenir compte des paramètres proches de zéro qui prévalent dans les probabilités des variables catégoriques, nous considérons le biais standardisé absolu (BSA) de chaque paramètre Q comme suit :

$$\text{BSA} = \sum_{h=1}^{H} \left| \bar{q}_L^{(h)} - Q \right| \Big/ \left(H \cdot Q \right), \tag{3.1}$$

où $\bar{q}_L^{(h)}$ est l'estimateur ponctuel de l'imputation multiple de Q dans la simulation h.

La deuxième mesure est l'erreur quadratique moyenne relative (EQMR), qui correspond au ratio entre l'EQM de l'estimation de Q à partir des données imputées et celle de l'estimation à partir des données échantillonnées avant l'introduction des données manquantes :

$$\text{EQMR} = \frac{\sum_{h=1}^{H} \left(\bar{q}_L^{(h)} - Q \right)^2}{\sum_{h=1}^{H} \left(\tilde{Q}^{(h)} - Q \right)^2}, \tag{3.2}$$

où $\bar{q}_L^{(h)}$ a été défini plus tôt, et $\tilde{Q}^{(h)}$ est l'estimateur prototype de Q, c'est-à-dire l'estimateur ponctuel de toutes les données échantillonnées dans la simulation h.

La troisième mesure est le taux de couverture, qui correspond à la proportion des intervalles de confiance $\alpha\,\%$ (par exemple $95\,\%$), indiquée par $\text{IC}_h^{\alpha} \left(h = 1, \ldots, H \right)$, dans les simulations H qui contiennent le vrai Q:

$$\text{Couverture} = \sum_{h=1}^{H} \mathbf{1} \left\{ Q \in \text{IC}_h^{\alpha} \right\} \Big/ H. \tag{3.3}$$

Nous recommandons d'effectuer un grand nombre de simulations (par exemple $H \geq 100$) pour obtenir des estimations fiables de l'EQM et de la couverture. Cela ne poserait pas de problème pour les algorithmes d'apprentissage profond, qui peuvent généralement être exécutés en quelques secondes, même avec des échantillons de grande taille. Cependant, cela peut être prohibitif sur le plan des calculs d'utiliser des algorithmes de la MICE lorsque chacune des données simulées est grande (par exemple $n = 100\,000$ dans certaines de nos simulations). Dans les cas où l'on ne doit se fier qu'à quelques simulations, voire une seule, pour réaliser l'évaluation, nous proposons une mesure modifiée du biais. Plus précisément, pour chaque variable catégorique ou chaque variable regroupée en classes j, nous définissons le biais absolu pondéré comme la somme du biais absolu pondéré par la probabilité marginale réelle dans chaque catégorie :

$$\text{Biais absolu pondéré} = \sum_{k=1}^{K} Q_{jk} \left| \bar{q}_{jk}^{(h)} - Q_{jk} \right|, \tag{3.4}$$

où K est le nombre total de catégories, Q_{jk} est la probabilité marginale de la population de la catégorie k dans la variable j, et $\bar{q}_{jk}^{(h)}$ est son estimateur ponctuel correspondant dans la simulation h. Nous pouvons également calculer la moyenne du biais absolu pondéré en fonction d'un certain nombre d'échantillons simulés à plusieurs reprises.

La procédure et les mesures détaillées ci-dessus diffèrent de la pratique courante dans la littérature sur l'apprentissage automatique. Par exemple, dans le cadre de nombreuses études sur l'apprentissage automatique qui portent sur l'imputation des données manquantes, des simulations sont effectuées sur des ensembles de données de référence, mais ces données ont souvent une structure et des caractéristiques très différentes des données d'enquête et sont donc moins informatives selon l'objectif de la présente étude. L'un de ces ensembles de données est l'ensemble de données sur le cancer du sein qui se trouve dans le *Machine Learning Repository* de l'Université d'Irvine en Californie (Dua et Graff, 2017), qui ne compte que 569 unités d'échantillonnage et aucune variable catégorique. De plus, ces simulations sont habituellement fondées sur la création aléatoire et répétitive de valeurs manquantes d'un seul ensemble de données plutôt que sur le fait de tirer des échantillons d'une population à plusieurs reprises, ce qui ne tient pas compte du mécanisme d'échantillonnage. Ces évaluations reposent sur des mesures axées sur l'exactitude des prévisions individuelles plutôt que sur les caractéristiques de distribution. Plus précisément, les mesures les plus couramment utilisées sont la racine carrée de l'erreur quadratique moyenne (REQM) et l'exactitude (Gondara et Wang, 2018; Yoon, Jordon et Schaar, 2018; Lu et coll., 2020). Les deux mesures peuvent être définies d'une manière globale ou propre à une variable, mais la littérature sur l'apprentissage automatique est habituellement axée sur la version globale. La REQM globale est définie comme suit :

$$\text{REQM} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j} M_{ij}\left(\hat{Y}_{ij} - Y_{ij}\right)^2}{\sum_{i=1}^{n}\sum_{j} M_{ij}}}, \tag{3.5}$$

où Y_{ij} est la valeur de la variable continue j pour une personne i dans les données complètes avant l'introduction des données manquantes, et $\hat{Y}_{ij}$ est la valeur imputée correspondante. Pour les valeurs non manquantes (c'est-à-dire $M_{ij} = 1$), $Y_{ij} = \hat{Y}_{ij}$. L'exactitude (globale) est définie pour les variables catégoriques, c'est-à-dire la proportion des valeurs imputées étant égale à la « vraie » valeur originale correspondante :

$$\text{Exactitude} = \frac{\sum_{i=1}^{n}\sum_{j\in S_{\text{cat}}} M_{ij}\mathbf{1}\left(\hat{Y}_{ij} = Y_{ij}\right)}{\sum_{i=1}^{n}\sum_{j\in S_{\text{cat}}} M_{ij}}, \tag{3.6}$$

où S_{cat} est l'ensemble des variables catégoriques.

Un certain nombre de mises en garde s'imposent pour la REQM et les mesures d'exactitude. Premièrement, elles sont habituellement calculées selon un seul échantillon imputé comme mesure globale

d'une méthode d'imputation, mais cela ne tient pas compte de l'incertitude des imputations. Deuxièmement, la REQM et l'exactitude sont des sommaires de valeurs uniques et ne saisissent pas la caractéristique de distribution multivariée des données. Troisièmement, la REQM ne tient pas compte des différentes échelles de variables et peut facilement être dominée par quelques valeurs aberrantes. De plus, elle est souvent calculée sans différenciation entre les variables continues et catégoriques. Enfin, lorsqu'il y a plusieurs (L) données imputées, une méthode courante consiste à utiliser la moyenne de la valeur imputée L en tant que $\hat{Y}_{ij}$ dans l'équation (3.5), mais la signification statistique des mesures résultantes n'est pas claire. Cela est particulièrement problématique pour les variables catégoriques. Pour ces raisons, nous exprimons une mise en garde contre l'utilisation de la REQM globale et l'exactitude comme seules mesures pour comparer les méthodes d'imputation. Il faut faire preuve de prudence lorsqu'on les interprète.

4. Évaluation fondée sur l'*American Community Survey*

Dans la présente section, nous évaluons les quatre méthodes d'imputation décrites à la section 2 en suivant la procédure et les mesures décrites à la section 3. Par souci de simplicité, dans les analyses suivantes, nous utilisons l'arbre de classification et de régression (CART) et l'algorithme de forêt aléatoire (RF) pour désigner respectivement la MICE-CART et la MICE-RF.

4.1 Les données sur la « population »

Nous utilisons l'échantillon de microdonnées à grande diffusion d'une durée d'un an de l'ACS de 2018 pour construire notre population. Les données de l'ACS de 2018 contiennent des variables au niveau du ménage, par exemple, si une maison est possédée ou louée, et des variables au niveau individuel, par exemple, l'âge, le revenu et le sexe des personnes dans chaque ménage. Étant donné que les personnes qui vivent dans un ménage sont souvent dépendantes et que les méthodes d'imputation que nous évaluons supposent généralement l'indépendance de toutes les observations, nous avons établi notre unité d'observation au niveau du ménage, où l'indépendance est plus susceptible de tenir. Nous retirons d'abord les unités correspondant aux maisons vacantes. Ensuite, nous supprimons les unités pour lesquelles il manque des valeurs, afin de ne conserver que les cas complets. Dans chaque ménage, nous conservons également les données au niveau individuel correspondant uniquement au chef de ménage et nous les fusionnons avec les variables au niveau du ménage, ce qui donne un ensemble riche de variables comportant des relations conjointes potentiellement complexes.

Il est souvent difficile de générer des imputations plausibles pour des variables ordinales à plusieurs niveaux lorsqu'il y a une très faible masse aux niveaux les plus élevés, comme c'est le cas pour certaines variables dans les données de l'ACS. Selon Li, Baccini, Mealli, Zell, Frangakis et Rubin (2014), nous

traitons les variables ordinales de plus de 10 niveaux comme des variables continues. Nous suivons également l'approche d'Akande et coll. (2017) qui consiste à exclure les variables binaires où les probabilités marginales ne correspondent pas à $np > 10$ ou à $n(1-p) > 10$; ceci permet d'éliminer les paramètres là où le théorème de la limite centrale risque de ne pas de tenir. Pour chaque variable catégorique comportant plus de deux niveaux, mais moins de 10 niveaux où cela pourrait également poser un problème, nous fusionnons les niveaux avec un petit nombre d'observations dans les données sur la population. Par exemple, pour la variable de la langue du ménage, nous reprogrammons les niveaux de cinq à trois (anglais, espagnol et autres), parce que la probabilité de ne parler ni l'anglais ni l'espagnol dans l'ensemble de la population est inférieure à 8,8 %.

Les données finales sur la population contiennent 1 257 501 unités, avec 18 variables binaires, 20 variables catégoriques ayant de 3 à 9 niveaux et 8 variables continues. Nous décrivons les variables plus en détail dans la documentation supplémentaire. Nous calculons les valeurs de population des paramètres Q décrits à la section 3, y compris toutes les probabilités marginales et bivariées de variables continues discrètes et regroupées en classes. Nous faisons varier la taille des échantillons simulés de 10 000 à 100 000 et simulons les données manquantes soit selon le mécanisme de données manquantes complètement au hasard (DMCH), soit selon le mécanisme de données manquantes au hasard (DMH) dans chacun de ces scénarios.

4.2 Simulations où n = 10 000

Nous tirons d'abord au hasard $H = 100$ échantillons d'une taille $n = 10\,000$ et nous déterminons qu'il manque 30 % de chaque échantillon selon le mécanisme de DMCH ou selon le mécanisme de DMH. Le CART ou la RF prend environ 2,8 et 9,2 heures, respectivement, pour créer $L = 10$ ensembles de données imputés qui comportent des paramètres par défaut sur un ordinateur de bureau standard ayant une seule unité centrale de traitement. Les méthodes d'apprentissage en profondeur sont beaucoup plus rapides parce qu'elles tirent parti de la puissance de calcul du processeur graphique lorsqu'elles sont mises en œuvre dans le cadre logiciel de TensorFlow optimisé pour le processeur graphique (Abadi, Agarwal, Barham, Brevdo, Chen, Citro, Corrado, Davis, Dean, Devin, Ghemawat, Goodfellow, Harp, Irving, Isard, Jia, Jozefowicz, Kaiser, Kudlur, Levenberg, Mané, Monga, Moore, Murray, Olah, Schuster, Shlens, Steiner, Sutskever, Talwar, Tucker, Vanhoucke, Vasudevan, Viégas, Vinyals, Warden, Wattenberg, Wicke, Yu et Zheng, 2015). Le GAIN prend environ 1,5 minute et la MIDA prend environ 4 minutes pour créer $L = 10$ ensembles de données complets au moyen d'un processeur graphique GeForce GTX 1660 Ti. Il convient de mentionner qu'il est impossible d'ajuster manuellement l'hyperparamètre dans chacune des 100 simulations de chaque scénario pour les modèles d'apprentissage profond. Donc, pour chaque scénario, nous avons choisi au hasard une simulation et nous avons ajusté les hyperparamètres en suivant la procédure décrite à la section 2. Nous avons ensuite appliqué ces hyperparamètres sélectionnés à toutes les simulations.

4.2.1 Scénario des données manquantes complètement au hasard

Pour créer le scénario de DMCH, nous établissons au hasard 30 % des valeurs de chaque variable en tant que valeurs manquantes de façon indépendante. Le tableau 4.1 présente les distributions du BSA estimé et de l'EQMR de toutes les probabilités marginales et bivariées dans les données imputées selon les quatre méthodes d'imputation.

Dans l'ensemble, pour les paramètres des probabilités marginales et bivariées des variables continues catégoriques et regroupées en classes, la MICE avec CART dépasse de façon importante les trois autres méthodes, produisant constamment le plus petit BSA et la plus petite EQMR. La RF se classe au deuxième rang, dépassant constamment les méthodes d'apprentissage profond. L'avantage des algorithmes de la MICE est particulièrement prononcé dans les quantiles supérieurs (par exemple 75 % et 90 %), ce qui indique que les imputations GAIN et MIDA présentent de grandes variations par rapport à des échantillons et des variables répétés. En effet, la MIDA et le GAIN mènent à des queues très longues dans l'estimation des statistiques sommaires des variables. Par exemple, pour les probabilités bivariées de variables continues regroupées en classes, le centile de 90 % du BSA de la MIDA et du GAIN est environ 20 et 27 fois plus grand que celui du CART, respectivement. L'écart est encore plus grand pour l'EQMR. Il n'y a pas de tendance cohérente lorsqu'on compare la MIDA et le GAIN. Plus précisément, pour les variables continues, la MIDA dépasse généralement le GAIN, mais la différence est faible, sauf pour les centiles supérieurs, où la GAIN a tendance à produire une EQMR et un biais très importants. Pour les variables catégoriques, le GAIN dépasse la MIDA la moitié du temps, mais, encore une fois, entraîne la plus grande variation des imputations entre les variables. De plus, une observation intéressante et quelque peu étonnante est que la MICE avec CART dépasse constamment la RF, parfois de beaucoup, quel que soit le choix du paramètre ou de la mesure.

Tableau 4.1
Distributions du biais standardisé absolu (BSA) (×100) et de l'erreur quadratique moyenne relative (EQMR) de toutes les probabilités marginales et bivariées selon les imputations découlant des quatre méthodes d'imputation multiple, lorsque *n* = 10 000 et que 30 % des valeurs sont manquantes complètement au hasard

	Quantiles		Probabilités marginales				Probabilités bivariées			
			CART	RF	GAIN	MIDA	CART	RF	GAIN	MIDA
BSA (×100)	Cat.	10 %	0,05	0,47	0,76	0,98	0,15	1,14	1,21	1,54
		25 %	0,13	1,25	1,48	2,22	0,40	2,83	3,08	3,93
		50 %	0,27	2,80	3,22	4,69	1,05	6,74	7,14	8,47
		75 %	0,64	5,86	7,18	8,86	2,51	13,59	17,03	15,23
		90 %	1,14	10,01	19,55	14,41	5,34	22,33	26,92	21,90
	Cont.	10 %	0,06	0,24	7,25	2,73	0,19	1,30	6,05	4,80
		25 %	0,10	1,05	12,86	8,36	0,43	3,24	17,61	12,01
		50 %	0,21	3,59	27,30	18,51	1,02	6,61	34,29	24,07
		75 %	0,43	5,43	30,21	26,84	1,90	11,76	49,38	39,54
		90 %	0,81	8,49	46,41	31,36	3,42	20,79	90,90	64,65

« Cat. » désigne les variables catégoriques et « Cont. » désigne les variables continues regroupées en classes.
CART = Classification and regression trees; RF = Random forests; GAIN = Generative adversarial imputation network;
MIDA = Multiple imputation using denoising autoencoders.

Tableau 4.1(suite)
Distributions du biais standardisé absolu (BSA) (×100) et de l'erreur quadratique moyenne relative (EQMR) de toutes les probabilités marginales et bivariées selon les imputations découlant des quatre méthodes d'imputation multiple, lorsque *n* = 10 000 et que 30 % des valeurs sont manquantes complètement au hasard

	Quantiles		Probabilités marginales				Probabilités bivariées			
			CART	RF	GAIN	MIDA	CART	RF	GAIN	MIDA
EQMR	Cat.	10 %	1,05	1,67	2,50	3,38	0,96	1,11	2,75	2,98
		25 %	1,16	2,40	4,97	9,03	1,08	1,61	4,33	4,75
		50 %	1,37	5,99	10,37	14,89	1,25	3,35	7,40	8,16
		75 %	1,49	10,25	27,73	26,16	1,48	9,07	14,87	15,80
		90 %	1,62	16,22	97,33	40,16	1,89	23,91	36,37	27,92
	Cont.	10 %	1,19	1,50	44,06	4,35	0,82	0,86	7,40	2,05
		25 %	1,30	1,77	74,42	13,82	0,92	1,11	14,80	4,90
		50 %	1,44	3,31	139,24	72,57	1,07	1,90	32,26	13,76
		75 %	1,55	6,71	284,00	150,35	1,26	4,09	88,78	47,56
		90 %	1,64	19,69	603,38	451,44	1,54	10,80	282,29	127,15

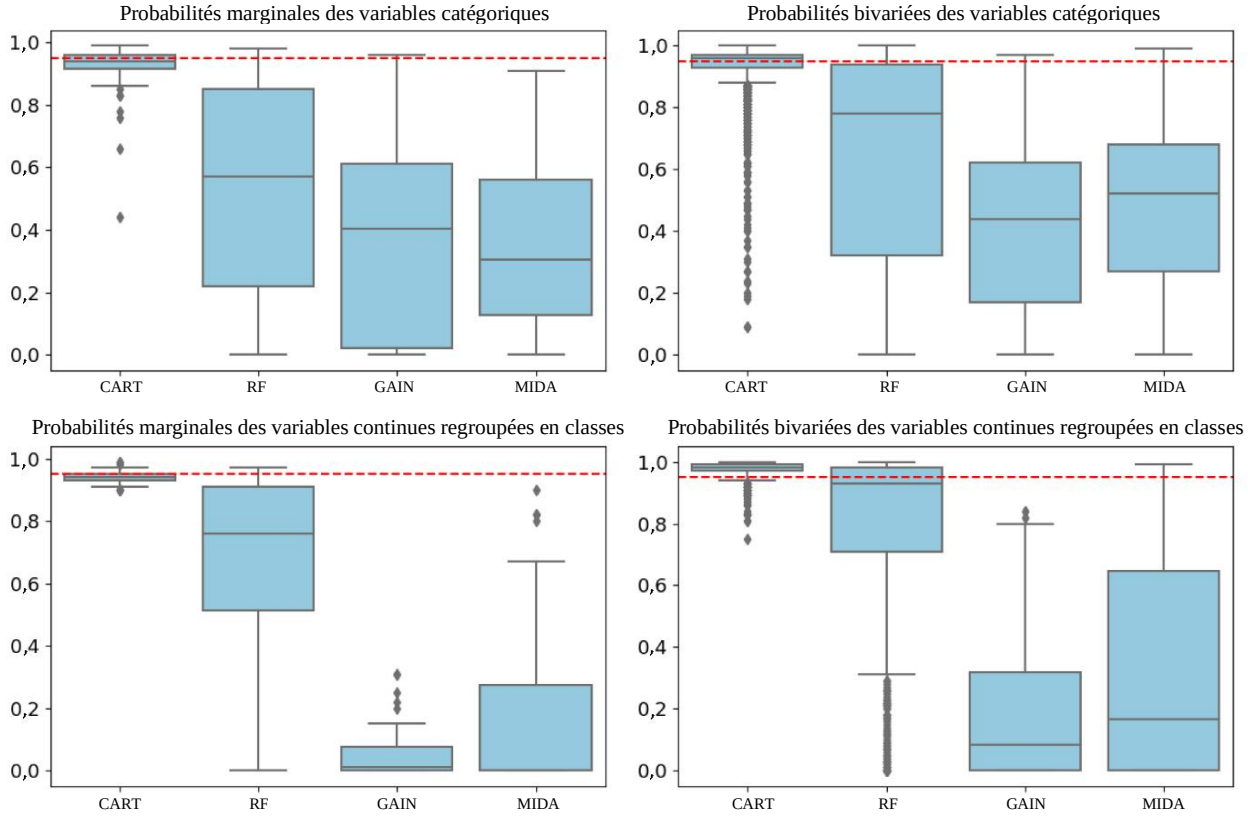
« Cat. » désigne les variables catégoriques et « Cont. » désigne les variables continues regroupées en classes.
CART = Classification and regression trees; RF = Random forests; GAIN = Generative adversarial imputation network;
MIDA = Multiple imputation using denoising autoencoders.

Toutes les méthodes donnent généralement des estimations moins biaisées (c'est-à-dire de plus petits BSA) des probabilités marginales que des probabilités bivariées. Ceci montre que la préservation des caractéristiques de distribution multivariée est plus difficile que celle des caractéristiques de distribution univariée. L'avantage du CART par rapport aux autres méthodes est comparativement plus important lorsqu'on estime des paramètres bivariés par rapport aux paramètres univariés. Fait intéressant, l'EQMR a tendance à être plus élevée pour les probabilités marginales que pour les probabilités bivariées. Cela est probablement dû au fait que le dénominateur de la définition de l'EQMR vu dans l'équation (3.2) correspond à l'EQM des données échantillonnées avant l'introduction des données manquantes, qui ont tendance à être plus faibles pour les probabilités marginales que pour les probabilités bivariées. Le CART donne des EQM qui sont très proches des EQM correspondantes à partir des données échantillonnées avant l'introduction des données manquantes; c'est-à-dire que l'EQMR est proche de 1. Au contraire, les deux méthodes d'apprentissage profond, et le GAIN en particulier, peuvent entraîner une EQMR extrêmement importante pour plusieurs paramètres.

Les figures 4.1 présentent les taux de couverture estimés des intervalles de confiance de 95 % pour les probabilités marginales et bivariées. Les tendances relatives à la couverture entre différentes méthodes sont semblables à celles relatives au biais et à l'EQM. Plus précisément, le CART tend à produire des taux de couverture qui se rapprochent du niveau nominal de 95 %, la médiane étant constamment autour de 95 % et correspondant à un écart interquartile serré. En revanche, les taux de couverture de la RF, du GAIN et de la MIDA sont beaucoup plus éloignés du niveau nominal de 95 %. Par exemple, les taux de couverture médians pour le GAIN et la MIDA sont tous inférieurs à 0,60, et même inférieurs à 0,30 pour les variables continues. Un examen plus approfondi de l'exactitude des prévisions de chaque variable révèle que le GAIN et la MIDA ont tendance à générer des imputations biaisées vers les niveaux les plus fréquents, et que le GAIN, en particulier, produit généralement des intervalles plus étroits que les autres

méthodes. Cela prouve une fois de plus que les méthodes d'apprentissage profond produisent des biais importants. Toutes les méthodes ont tendance à produire des taux de couverture médians plus élevés pour les probabilités bivariées que pour les probabilités marginales, bien que les queues à gauche soient généralement plus longues pour les probabilités bivariées que pour les probabilités marginales.

Figure 4.1 Taux de couverture de l'intervalle de confiance de 95 % pour toutes les probabilités marginales et bivariées obtenues à partir des quatre méthodes d'imputation dans les simulations comprenant *n* = 10 000 et 30 % de valeurs manquantes complètement au hasard.



La ligne pointillée rouge correspond à 0,95.
CART = Classification and regression trees; RF = Random forests; GAIN = Generative adversarial imputation network; MIDA = Multiple imputation using denoising autoencoders.

4.2.2 Scénario de données manquantes au hasard

Nous considérons également un scénario de DMH, qui est plus plausible que le scénario de DMCH en pratique. Nous avons établi six variables à observer en entier : l'âge, le sexe, l'état matrimonial, la race, le niveau de scolarité et la catégorie de travailleur. Il serait fastidieux de spécifier un mécanisme de DMH différent pour chacune des 40 variables restantes, alors nous les divisons aléatoirement en trois groupes, dont un composé de 10 variables et deux composés de 15 variables chacun. Nous précisons ensuite un modèle de non-réponse distinct que nous utiliserons pour générer les données manquantes pour les variables de chaque groupe. Plus précisément, nous postulons un modèle logistique par groupe,

conditionnel aux six variables entièrement observées, sur la base desquelles nous générons des indicateurs binaires de données manquantes pour chaque variable de ce groupe. Ce processus entraîne un taux de données manquantes d'environ 30 % pour chacune des 40 variables. Nous décrivons les modèles plus en détail dans la documentation supplémentaire.

Le tableau 4.2 présente les distributions du BSA et de l'EQMR de toutes les probabilités marginales et bivariées obtenues à partir des quatre méthodes. Toutes les méthodes donnent un BSA et une EQMR plus importants selon le scénario de DMH que selon le scénario de DMCH précédent. C'est normal, car le mécanisme de DMH correspond à une hypothèse plus forte que le mécanisme de DMCH, lequel exige le conditionnement d'un plus grand nombre de renseignements. Néanmoins, les tendances globales du rendement relatif entre les méthodes demeurent les mêmes que celles du mécanisme de DMCH. Plus précisément, le CART produit une fois de plus des estimations ayant les plus faibles BSA et EQMR, ce qui représente une marge encore plus grande que la méthode selon le mécanisme de DMCH parmi les quatre méthodes. Le CART est suivi de la RF, puis de la MIDA et du GAIN. Une observation notable est la détérioration du rendement des méthodes d'apprentissage profond, en particulier du GAIN, lors de l'imputation de variables continues, ce qui aboutit parfois à une multiplication par plusieurs centaines de l'EQMR par rapport au CART. Cela dénote les énormes incertitudes associées au GAIN lors de l'imputation de variables continues.

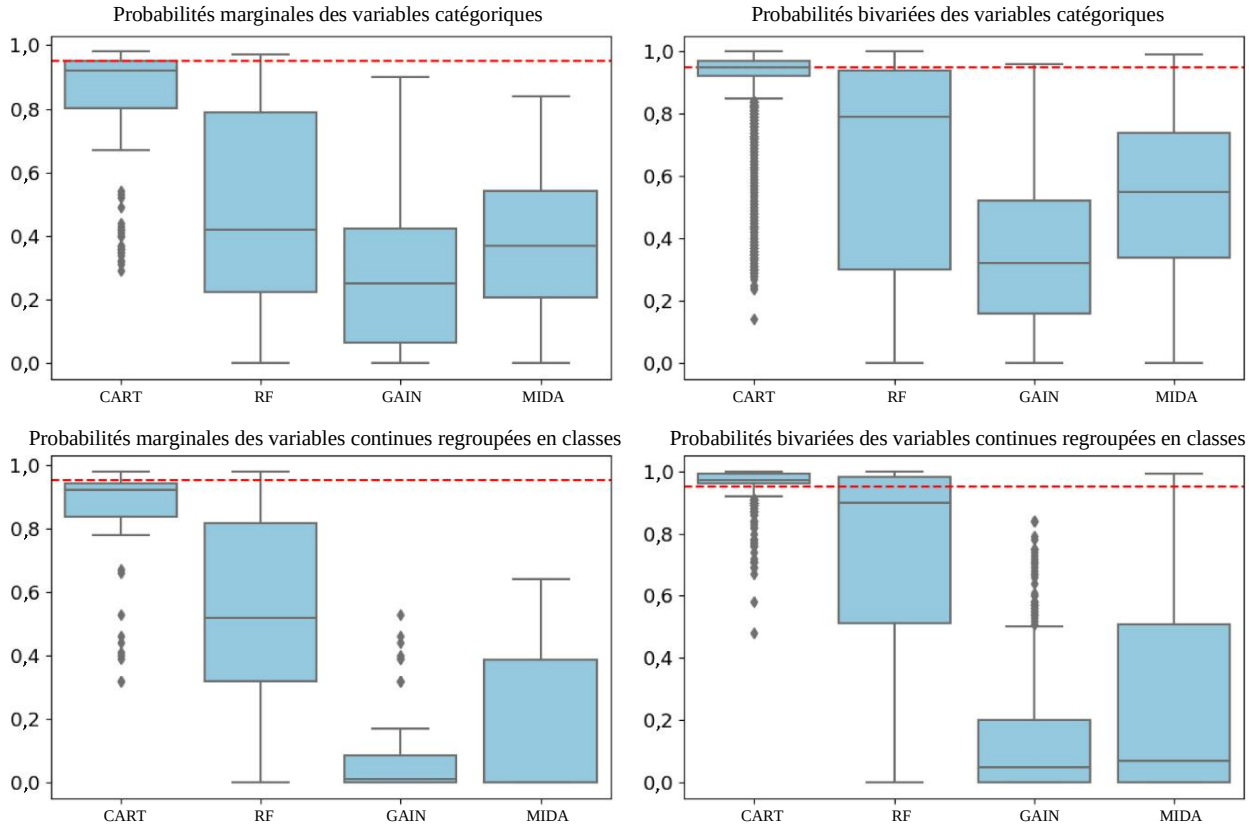
Tableau 4.2
Distributions du biais standardisé absolu (BSA) (×100) et de l'erreur quadratique moyenne relative (EQMR) pour toutes les méthodes, lorsque *n* = 10 000 et que 30 % des valeurs ne sont pas manquantes complètement au hasard, selon la totalité des probabilités marginales et bivariées possibles

	Quantiles		Probabilités marginales				Probabilités bivariées			
			CART	RF	GAIN	MIDA	CART	RF	GAIN	MIDA
BSA (×100)	Cat.	10 %	0,05	0,13	0,15	0,14	0,15	0,71	0,76	0,89
		25 %	0,11	0,44	0,62	0,61	0,40	2,23	2,55	3,20
		50 %	0,29	2,13	3,05	4,55	1,08	6,06	6,85	8,14
		75 %	1,04	4,98	6,63	10,22	2,49	13,43	16,78	16,19
		90 %	1,80	10,49	18,91	17,00	5,68	24,06	28,04	25,36
	Cont.	10 %	0,07	0,29	0,33	0,33	0,27	1,17	10,87	6,18
		25 %	0,17	1,07	9,64	3,13	0,69	3,49	23,67	16,26
		50 %	0,67	3,14	32,86	23,85	1,58	7,83	38,52	31,17
		75 %	1,20	6,95	39,57	36,09	3,40	15,20	53,59	47,34
		90 %	3,40	12,39	63,45	41,99	5,94	25,16	97,47	85,44
EQMR	Cat.	10 %	1,00	1,00	1,00	1,00	0,97	1,00	1,53	1,93
		25 %	1,08	1,82	2,56	4,75	1,04	1,39	3,78	4,03
		50 %	1,33	4,33	19,03	15,13	1,25	3,00	10,42	8,38
		75 %	1,72	13,08	55,07	33,36	1,59	9,56	27,45	16,95
		90 %	2,27	18,70	101,91	48,44	2,23	27,44	64,01	32,85
	Cont.	10 %	1,00	1,00	1,00	1,00	0,88	0,90	11,19	2,96
		25 %	1,38	1,83	90,98	8,49	1,00	1,16	20,15	6,87
		50 %	1,70	4,57	207,58	96,08	1,18	2,29	45,25	21,33
		75 %	2,12	11,47	692,67	239,69	1,50	6,95	125,39	70,90
		90 %	3,12	50,56	1342,23	806,43	2,12	18,07	459,78	205,14

« Cat. » désigne les variables catégoriques et « Cont. » désigne les variables continues regroupées en classes.
CART = Classification and regression trees; RF = Random forests; GAIN = Generative adversarial imputation network;
MIDA = Multiple imputation using denoising autoencoders.

Les figures 4.2 présentent les taux de couverture estimés des intervalles de confiance de 95 % pour les probabilités marginales et bivariées, selon chaque méthode. Comme dans le cas du biais et de l'EQM, toutes les méthodes entraînent généralement des taux de couverture inférieurs à ceux du mécanisme de DMH par rapport au mécanisme de DMCH. Les queues à gauche sont visiblement plus longues dans certains cas, mais les tendances globales, lorsqu'on compare les méthodes, demeurent les mêmes. Plus précisément, le CART a toujours tendance à donner des taux de couverture supérieurs à 90 %, tandis que les trois autres méthodes ont toujours un taux de couverture inférieur. En particulier, le GAIN et la MIDA se traduisent par des taux de couverture médians extrêmement bas, en dessous de 7 %, pour les variables continues. Ceci est étroitement lié à l'observation précédente de la grande incertitude des méthodes d'apprentissage profond dans l'imputation de variables continues.

Figure 4.2 Taux de couverture des intervalles de confiance de 95 % pour toutes les probabilités marginales et bivariées obtenues à partir des quatre méthodes dans les simulations comportant n = 10 000 et 30 % de valeurs manquantes au hasard.



La ligne pointillée rouge correspond à 0,95.
CART = Classification and regression trees; RF = Random forests; GAIN = Generative adversarial imputation network; MIDA = Multiple imputation using denoising autoencoders.

Enfin, pour illustrer le fait qu'il peut être trompeur d'évaluer seulement la REQM et les mesures d'exactitude globales, nous présentons les erreurs-types moyennes et empiriques de la REQM globale et de l'exactitude globale des 100 simulations au tableau 4.3, où les DMCH se trouvent dans la partie supérieure et les DMH, dans la partie inférieure. Dans le cas des deux mécanismes de données manquantes, pour les variables continues, la MIDA permet d'obtenir la plus petite REQM globale, suivie du CART ainsi que de la RF et du GAIN qui viennent en dernier. Pour les variables catégoriques, le CART et le GAIN permettent d'obtenir l'exactitude globale la plus élevée, la MIDA étant tout juste derrière et la RF apparaissant en dernier. Ces tendances, qui ne sont pas étonnantes, diffèrent de celles déclarées précédemment en fonction de probabilités marginales et bivariées et de mesures différentes. Comme nous avons vu à la section 3, la REQM et l'exactitude globales ne saisissent pas les caractéristiques de distribution des données multivariées ou les propriétés d'échantillonnage répété des méthodes d'imputation.

Tableau 4.3
Moyennes selon 100 simulations de la racine carrée de l'erreur quadratique moyenne (REQM) globale des variables continues et l'exactitude globale des variables catégoriques

Mécanisme	Mesure	CART	RF	GAIN	MIDA
DMCH	REQM	0,128 (0,002)	0,159 (0,003)	0,161 (0,008)	0,112 (0,002)
	Exactitude	0,785 (0,001)	0,658 (0,003)	0,782 (0,002)	0,752 (0,004)
DMH	REQM	0,130 (0,003)	0,154 (0,004)	0,145 (0,009)	0,110 (0,002)
	Exactitude	0,819 (0,001)	0,704 (0,003)	0,820 (0,002)	0,780 (0,007)

Les erreurs-types empiriques sont indiquées entre parenthèses.
Les données manquantes complètement au hasard (DMCH) se trouvent dans la partie supérieure et les données manquantes au hasard (DMH) se trouvent dans la partie inférieure, et 30 % de toutes les données sont manquantes.
CART = Classification and regression trees; RF = Random forests; GAIN = Generative adversarial imputation network;
MIDA = Multiple imputation using denoising autoencoders.

4.3 Simulations où $n = 100\,000$ et 30 % des valeurs sont manquantes complètement au hasard

Les modèles d'apprentissage profond exigent habituellement un échantillon de grande taille pour l'entraînement. Par conséquent, pour donner à la MIDA et au GAIN un cadre plus favorable et pour étudier la sensibilité de nos résultats aux variations de la taille de l'échantillon, nous générons un scénario de simulation de $H = 10$ échantillons et de $n = 100\,000$ selon le mécanisme de DMCH. Autrement dit, nous établissons au hasard 30 % des valeurs de chaque variable en tant que valeurs manquantes indépendamment. Dans la présente étude, nous ne générons que 10 simulations en raison des énormes coûts liés au calcul selon la MICE pour des échantillons de cette taille. Dans ce scénario, nous omettons la RF, car les résultats précédents de la section 4.2 ont montré que la RF est toujours inférieure au CART en ce qui a trait au rendement et au calcul. Nous utilisons le CART, le GAIN et la MIDA pour créer $L = 10$ ensembles de données complets.

Comme il faut habituellement un nombre beaucoup plus grand de simulations pour calculer de façon fiable l'EQM et la couverture, nous nous concentrons sur la mesure du biais absolu pondéré (3.4). Le tableau 4.4 présente les distributions du biais absolu pondéré estimé (dont la moyenne est établie sur 10 simulations) des probabilités marginales des variables catégoriques et des variables continues regroupées en classes. Dans l'ensemble, les tendances, lorsqu'on compare les quatre méthodes, demeurent conformes à celles observées à la section 4.2. Plus précisément, le CART entraîne de nouveau la plus petite différence absolue pondérée dans les variables catégoriques et continues, et l'avantage est particulièrement prononcé pour les variables continues. Par exemple, pour les variables catégoriques, la MIDA et le GAIN donnent une médiane du biais absolu pondéré au moins 9 et 11 fois plus élevée que le CART, respectivement. L'avantage du CART augmente à environ 30 et 60 fois par rapport aux variables continues avec la MIDA et le GAIN, respectivement. De plus, le CART donne de bons résultats pour l'ensemble des variables, comme en témoigne la faible variation du biais absolu pondéré, par exemple 0,07 pour le centile de 10 % et 0,33 pour le centile de 90 % des variables catégoriques. En revanche, les deux modèles d'apprentissage profond entraînent une variation beaucoup plus grande entre les variables, par exemple 0,57 pour le centile de 10 % et 2,92 pour le centile de 90 % des variables catégoriques selon la MIDA, et encore plus pour le GAIN. En résumé, en dehors du temps de calcul, la MICE avec CART dépasse de façon considérable la MIDA et le GAIN en ce qui concerne le biais et la variance, indépendamment de la taille de l'échantillon.

Tableau 4.4
Distributions du biais absolu pondéré ($\times 100$) et de la moyenne sur 10 échantillons simulés, où pour chacun n = 10 000 et 30 % des valeurs sont manquantes complètement au hasard

Quantiles	Variables catégoriques			Variables continues regroupées par classes		
	CART	GAIN	MIDA	CART	GAIN	MIDA
10 %	0,07	0,43	0,57	0,10	5,52	1,98
25 %	0,11	1,11	1,02	0,11	6,65	2,78
50 %	0,15	1,74	1,40	0,12	7,36	4,04
75 %	0,24	3,77	2,07	0,13	9,40	6,50
90 %	0,33	4,63	2,92	0,15	11,31	7,72

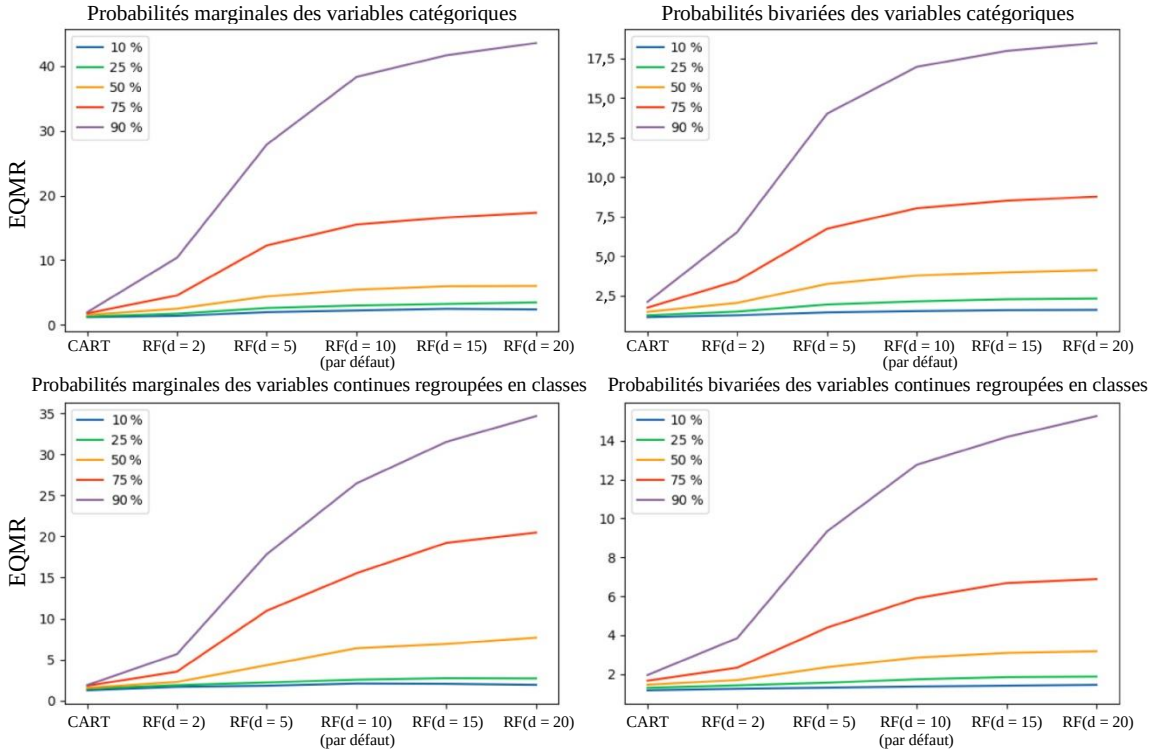
CART = Classification and regression trees; GAIN = Generative adversarial imputation network;
MIDA = Multiple imputation using denoising autoencoders.

4.4 Rôle des hyperparamètres dans l'imputation multiple au moyen d'équations en séries fondée sur l'arborescence

La tendance selon laquelle le CART dépasse la RF est étonnante, car on sait que les méthodes d'ensemble sont habituellement supérieures aux méthodes à arbre unique. Cependant, la même tendance a également été observée dans une autre étude récente (Wongkamthong et Akande, 2021). Nous étudions le rôle de l'hyperparamètre clé dans la RF, à savoir le nombre maximum d'arbres d, dans les simulations. Nous avons sélectionné au hasard des données simulées de taille $n = 10\,000$ et 30 % des entrées manquantes complètement au hasard. Nous utilisons le progiciel `mice` pour ajuster la RF en fonction des

différents nombres d'arbres : $d = 2, 5, 10, 15, 20,$ où $d = 10$ est le paramètre par défaut. L'EQMR des variables catégoriques imputées, ajustée à l'aide de chaque valeur d, ainsi que celle reposant sur le CART, est présentée sous forme de trajectoires à la figure 4.3. Ces trajectoires révèlent une tendance constante : les quantiles supérieurs de l'EQMR, en particulier ceux de plus de 50 %, se détériorent rapidement à mesure que le nombre maximal d'arbres de la RF augmente, tandis que les quantiles inférieurs, par exemple ceux de 10 % et de 25 %, demeurent stables. Nous avons constaté une tendance semblable en ce qui a trait à la mesure du biais standardisé et aux variables continues; par conséquent, les résultats sont omis dans la présente étude. Cela laisse entendre qu'un plus grand nombre d'arbres dans la RF (du moins tel que cela est mis en œuvre dans le progiciel `mice`) entraîne une queue beaucoup plus longue dans la distribution du biais et des EQM. Cela est probablement dû à un surajustement. Nous ne pouvons pas exclure la possibilité qu'un ajustement plus personnalisé des hyperparamètres de la RF puisse dépasser le CART dans certaines applications. Toutefois, un tel ajustement précis propre à un cas de l'algorithme MICE n'est généralement pas accessible à la grande majorité des utilisateurs de l'imputation multiple qui se fient à l'ajustement par défaut des progiciels populaires comme `mice`.

Figure 4.3 Quantiles de l'erreur quadratique moyenne relative (EQMR) pour l'ensemble des probabilités marginales et bivariées des variables catégoriques et continues regroupées en classes, selon les modèles par arbres de classification et de régression (CART) et l'algorithme de forêt aléatoire (RF) comprenant différents nombres d'arbres, pour un échantillon de simulation ayant *n* = 10 000 et 30 % des valeurs manquantes complètement au hasard.



5. Évaluation fondée sur des ensembles de données « de référence »

Pour vérifier les évaluations dans les études sur le GAIN et la MIDA (Gondara et Wang, 2018; Yoon, Jordon et Schaar, 2018; Lu et coll., 2020), nous avons également comparé les deux modèles d'apprentissage profond avec le CART en fonction des cinq ensembles de données de référence et de la procédure de simulation (différente de notre cadre proposé) utilisés dans ces études. Les précisions relatives à ces ensembles de données et à ces simulations sont présentées dans la documentation supplémentaire. Les tailles d'échantillon de ces données ne sont généralement pas assez grandes pour être considérées comme des données de population à partir desquelles nous pouvons échantillonner à répétition sans remise, de sorte que nous ne sommes pas en mesure de les évaluer de façon considérable à l'aide du biais standardisé absolu, de l'EQMR ou de la couverture. Nous évaluons donc les méthodes principalement en fonction de la mesure du biais absolu pondéré. En résumé, le CART dépasse encore de façon constante et importante la MIDA et le GAIN en ce qui concerne le biais absolu pondéré des variables catégoriques et continues, et ce, pour les cinq ensembles de données de référence. L'écart de rendement est particulièrement prononcé pour les variables continues. Nous avons également calculé l'EQM et l'exactitude globales comme les auteurs de ces études l'ont fait. À l'exception d'un ensemble de données, nous n'avons pas pu reproduire les résultats présentés dans ces études, même en utilisant le code des auteurs. Une raison possible est que la documentation du processus d'ajustement et de sélection des hyperparamètres du modèle serait obscure, ce qui est vrai dans le cas présent. Vous trouverez de plus amples renseignements dans la documentation supplémentaire en ligne.

6. Conclusion

Ces dernières années, de nombreuses méthodes d'imputation de données manquantes fondées sur l'apprentissage automatique ont vu le jour, ce qui a suscité l'espoir qu'il y ait des améliorations par rapport aux méthodes d'imputation plus conventionnelles comme la MICE. Cependant, les efforts pour évaluer ces méthodes dans des situations réelles demeurent rares. Dans la présente étude, nous adoptons un cadre d'évaluation des simulations reposant sur des données réelles. Nous menons de vastes études de simulation qui sont fondées sur l'*American Community Survey* afin de comparer les propriétés d'échantillonnage répété de deux méthodes de la MICE et de deux méthodes d'imputation par apprentissage profond s'appuyant sur le GAN (GAIN) et les autoencodeurs débruiteurs (MIDA).

Nous avons découvert que les modèles d'apprentissage profond ont un grand avantage sur le plan des calculs par rapport aux méthodes de la MICE, en partie parce qu'ils peuvent exploiter la puissance du processeur graphique pour effectuer des calculs de manière très efficace. Cependant, nos simulations ainsi que notre évaluation de plusieurs données de « référence » indiquent que la MICE avec la spécification du CART des modèles conditionnels dépasse constamment, et habituellement de façon substantielle, les modèles d'apprentissage profond en ce qui a trait au biais, à l'erreur quadratique moyenne et à la

couverture dans un large éventail de contextes réalistes. En particulier, le GAIN et la MIDA ont tendance à générer des imputations instables comportant d'énormes variations des échantillons répétés comparativement à la MICE. Une explication possible est que les réseaux neuronaux profonds excellent dans la détection de sous-structures complexes de mégadonnées, mais peuvent ne pas convenir aux données ayant une structure simple, comme les données simulées utilisées dans la présente étude. Une autre possibilité est que la taille des échantillons dans nos simulations ne soit pas suffisante pour former des réseaux neuronaux profonds, lesquels exigent habituellement beaucoup plus de données que les modèles statistiques conventionnels.

Ces résultats contredisent les constatations précédentes fondées sur la seule mesure de rendement de l'erreur quadratique moyenne globale qui proviennent de la littérature sur l'apprentissage automatique (par exemple Gondara et Wang, 2018; Yoon, Jordon et Schaar, 2018; Lu et coll., 2020). Cet écart met en évidence les pièges de la pratique courante de l'évaluation des méthodes d'imputation dans la littérature sur l'apprentissage automatique. Il démontre également l'importance d'évaluer les propriétés de l'échantillonnage répété selon plusieurs paramètres des méthodes d'imputation multiple. Une découverte intéressante est que les ensembles d'arbres (par exemple la RF) ne s'améliorent pas selon un seul arbre (par exemple le CART) dans le contexte de la MICE, ce qui correspond aux résultats d'une autre étude récente (Wongkamthong et Akande, 2021). En combinaison avec le fait que la méthode RF est plus intensive sur le plan des calculs que la méthode CART, nous recommandons d'utiliser en pratique la MICE avec CART plutôt que la RF.

Notre étude comporte quelques limites. Premièrement, il existe de nombreuses méthodes d'apprentissage profond qui peuvent être adaptées à l'imputation des données manquantes et elles peuvent toutes avoir des caractéristiques de fonctionnement différentes. Nous avons choisi le GAIN et la MIDA parce que les réseaux antagonistes génératifs et les autoencodeurs débruiteurs sont des méthodes d'apprentissage profond extrêmement reconnues, et les méthodes d'imputation fondées sur ces méthodes ont été reconnues comme étant supérieures à la MICE. Néanmoins, il serait souhaitable d'examiner d'autres méthodes d'imputation fondées sur l'apprentissage profond dans les travaux de recherche futurs. Deuxièmement, le rendement des méthodes d'apprentissage automatique dépend fortement de la sélection des hyperparamètres. On peut donc soutenir que le rendement inférieur du GAIN et de la MIDA peut être attribuable au moins en partie à la sélection d'hyperparamètres sous-optimaux. Cependant, les spécialistes s'appuieraient probablement sur les valeurs d'hyperparamètres par défaut pour toute méthode d'imputation par apprentissage automatique, ce qui est en fait ce que nous avons adopté dans nos simulations et ce qui représente donc la pratique réelle. Troisièmement, nous n'avons pas tenu compte de la distribution conjointe entre des variables catégoriques et continues, mais nos évaluations au sein de variables catégoriques et continues ont produit des conclusions cohérentes. Enfin, comme pour toute étude de simulation, il faut faire preuve de prudence lors de la généralisation des conclusions. En choisissant

soigneusement les données et les mesures, nous avons tenté d'imiter de près les contextes représentatifs des données d'enquête réelles afin que nos conclusions soient informatives pour les spécialistes qui doivent faire face à des situations semblables. D'autres études d'évaluation fondées sur différentes données sont souhaitées afin de mieux comprendre les caractéristiques opérationnelles et le rendement comparatif des différentes méthodes d'imputation des données manquantes. Les données, les codes et la documentation supplémentaire relatifs à la présente étude sont accessibles à l'adresse suivante : https://github.com/zhenhua-wang/MissingData_DL (en anglais seulement).

Remerciements

Les travaux de recherche de Poulos et Li sont soutenus par la *National Science Foundation* dans le cadre de la subvention DMS-1638521 accordée au *Statistical and Applied Mathematical Sciences Institute*.

Bibliographie

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. et Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems [Logiciel disponible sur tensorflow.org], https://www.tensorflow.org/.

Akande, O., Li, F., et Reiter, J. (2017). An empirical comparison of multiple imputation methods for categorical data. *The American Statistician*, 71(2), 162-170.

Arnold, B.C., et Press, S.J. (1989). Compatible conditional distributions. *Journal of the American Statistical Association*, 84, 152-156.

Barnard, J., et Meng, X.-L. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research*, 8(1), 17-36.

Berthelot, D., Schumm, T. et Metz, L. (2017). *BEGAN: Boundary Equilibrium Generative Adversarial Networks*. CoRR, abs/1703.10717. http://arxiv.org/abs/1703.10717.

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

Breiman, L., Friedman, J.H., Olshen, R.A. et Stone, C.J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworh, Inc.

Burgette, L., et Reiter, J.P. (2010). Multiple imputation via sequential regression trees. *American Journal of Epidemiology*, 172, 1070-1076.

Cao, W., Wang, D., Li, J., Zhou, H., Li, L. et Li, Y. (2018). BRITS: Bidirectional recurrent imputation for time series. *Advances in Neural Information Processing Systems*, 6775-6785.

Che, Z., Purushotham, S., Cho, K., Sontag, D. et Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1), 1-12.

Chen, S., et Haziza, D. (2019). Recent developments in dealing with item nonresponse in surveys: A critical review. *Revue Internationale de Statistique*, 87, S192-S218.

De Leeuw, E.D., Hox, J. et Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics*, Stockholm, 19(2), 153-176.

Doove, L., Van Buuren, S. et Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72, 92-104.

Dua, D., et Graff, C. (2017). *UCI Machine Learning Repository*. http://archive.ics.uci.edu/ml.

Fortuin, V., Baranchuk, D., Rätsch, G. et Mandt, S. (2020). GP-VAE: Deep probabilistic time series imputation. *International Conference on Artificial Intelligence and Statistics*, 1651-1661.

Gelman, A., et Speed, T.P. (1993). Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 55, 185-188.

Glorot, X., et Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Artificial Intelligence and Statistics*, 9, 249-256.

Gondara, L., et Wang, K. (2018). MIDA: Multiple imputation using denoising autoencoders. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 260-272.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. et Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2672-2680.

Ham, H., Jun, T.J. et Kim, D. (2020). Unbalanced Gans: Pre-Training the Generator of Generative Adversarial Network Using Variational Autoencoder. arXiv preprint arXiv:2002.02112.

Harel, O., et Zhou, X.-H. (2007). Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, 26(16), 3057-3077.

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

Haziza, D., et Vallée, A.-A. (2020). Variance estimation procedures in the presence of singly imputed survey data: A critical review. *Japanese Journal of Statistics and Data Science*, 3(2), 583-623.

Ho, T.K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1, 278-282.

Honaker, J., King, G. et Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7), 1-47.

Horton, N.J., Lipsitz, S.R. et Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician*, 57(4), 229-232.

Huque, M.H., Carlin, J.B., Simpson, J.A. et Lee, K.J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Medical Research Methodology*, 18(1), 1-16.

Kingma, D., et Ba, J. (2014). *Adam: A Method for Stochastic Optimization*. arXiv:1412.6980.

Li, F., Yu, Y. et Rubin, D. (2012). *Imputing Missing Data by Fully Conditional Models: Some Cautionary Examples and Guidelines*. Rapport technique, document de travail du Duke University Department of Statistical Science, 11-24.

Li, F., Baccini, M., Mealli, F., Zell, E.R., Frangakis, C.E. et Rubin, D.B. (2014). Multiple imputation by ordered monotone blocks with application to the anthrax vaccine research program. *Journal of Computational and Graphical Statistics*, 23(3), 877-892.

Lipton, Z.C., Kale, D.C. et Wetzel, R. (2016). Modeling missing data in clinical time series with RNNs. *Machine Learning for Healthcare*, 56.

Little, R.J., et Rubin, D.B. (2014). *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley & Sons, Inc.

Little, R.J., et Rubin, D.B. (2019). *Statistical Analysis with Missing Data*, 3rd edition. New York: John Wiley & Sons, Inc.

Lu, H.-M., Perrone, G. et Unpingco, J. (2020). *Multiple Imputation with Denoising Autoencoder Using Metamorphic Truth and Imputation Feedback*. arXiv preprint arXiv:2002.08338.

Maas, A.L., Hannun, A.Y. et Ng, A.Y. (2013). Rectifier nonlinearities improve neural network acoustic models. *Proc. ICML*, (1), 3.

Manrique-Vallier, D., et Reiter, J. (2014). Bayesian estimation of discrete multivariate truncated latent structure models. *Journal of Computational and Graphical Statistics*, 23, 1061-1079.

Monti, F., Bronstein, M. et Bresson, X. (2017). Geometric matrix completion with recurrent multi-graph neural networks. *Advances in Neural Information Processing Systems*, 3697-3707.

Murray, J.S., et Reiter, J.P. (2016). Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516), 1466-1479.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J. et Solenberger, P. (2001). Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression. *Techniques d'enquête*, 27, 1, 91-103. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2001001/article/5857-fra.pdf.

Reiter, J.P., et Raghunathan, T.E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480), 1462-1471.

Royston, P., et White, I.R. (2011). Multiple imputation by chained equations (mice): Implementation in Stata. *Journal of Statistical Software*, 45(4), 1-20.

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

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473-489.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Londres : Chapman & Hall.

Shah, A., Bartlett, J., Carpenter, J., Nicholas, O. et Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: A caliber study. *American Journal of Epidemiology*, 179, 764-74.

Stekhoven, D.J., et Bühlmann, P. (2012). Missforest – non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112-118.

Su, Y.-S., Gelman, A.E., Hill, J. et Yajima, M. (2011). Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. *Journal of Statistical Software*, 45.

Tang, L., Song, J., Belin, T.R. et Unützer, J. (2005). A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine*, 24(14), 2111-2128.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC Press LLC.

van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M. et Rubin, D.B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049-1064.

van Buuren, S., et Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3), 1-67.

Vincent, P., Larochelle, H., Bengio, Y. et Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning*, 1096-1103.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A. et Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(12).

White, I.R., Royston, P. et Wood, A.M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4), 377-399.

Wongkamthong, C., et Akande, O. (2021). A comparative study of imputation methods for multivariate ordinal data. *Journal of Survey Statistics and Methodology*, dans la presse.

Yoon, J., Jordon, J. et Schaar, M. (2018). Gain: Missing data imputation using generative adversarial nets. *International Conference on Machine Learning*, 5689-5698.

Yoon, J., Zame, W.R. et van der Schaar, M. (2018). Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering*, 66(5), 1477-1490.

Yuan, Y. (2011). Multiple imputation using SAS software. *Journal of Statistical Software*, 45(6), 1-25.

Techniques d'enquête, décembre 2022
Vol. 48, nº 2, pp. 439-479
Statistique Canada, nº 12-001-X au catalogue

Modélisation de séries chronologiques multiniveaux de la couverture des soins prénataux au Bangladesh à des niveaux administratifs désagrégés

Sumonkanti Das, Jan van den Brakel, Harm Jan Boonstra et Stephen Haslett[1]

Résumé

Des modèles de séries chronologiques multiniveaux sont appliqués pour estimer les tendances de séries chronologiques de la couverture des soins prénataux à plusieurs niveaux administratifs du Bangladesh, d'après les cycles répétés de la *Bangladesh Demographic and Health Survey* (BDHS, Enquête démographique et sur la santé du Bangladesh) pendant la période allant de 1994 à 2014. Les modèles de séries chronologiques multiniveaux sont exprimés dans un cadre bayésien hiérarchique et ajustés au moyen de simulations Monte Carlo par chaînes de Markov. Les modèles tiennent compte des intervalles variables de trois ou quatre ans entre les cycles de la BDHS et fournissent aussi des prédictions pour les années intermédiaires. Il est proposé d'appliquer les modèles transversaux de Fay-Herriot aux années d'enquête séparément au niveau des districts, soit l'échelle régionale la plus détaillée. Les séries chronologiques de ces prédictions pour petits domaines au niveau des districts et leurs matrices de variance-covariance sont utilisées comme séries de données d'entrée pour les modèles de séries chronologiques multiniveaux. Dans ces modèles, on examine les corrélations spatiales entre les districts, la pente et l'ordonnée à l'origine aléatoires au niveau des districts, ainsi que les différents modèles de tendance au niveau des districts et aux niveaux régionaux plus élevés pour l'emprunt d'information dans le temps et l'espace. Les estimations des tendances au niveau des districts sont obtenues directement à partir des résultats des modèles, tandis que les estimations des tendances à des échelons régionaux et nationaux plus élevés sont obtenues par agrégation des prédictions au niveau des districts, ce qui donne un ensemble cohérent d'estimations des tendances sur le plan numérique.

Mots-clés : Approche bayésienne hiérarchique; enquêtes démographiques et sur la santé; estimation sur petits domaines; modèle transversal de Fay-Herriot; simulation par la méthode de Monte Carlo par chaînes de Markov (MCMC).

1. Introduction

Dans plus de 90 pays, les enquêtes démographiques et sur la santé ont été largement utilisées pour estimer des indicateurs nationaux et infranationaux concernant la fécondité, la planification familiale, la mortalité infantile, la santé infantile, la santé maternelle et la nutrition des enfants et des femmes (DHS, 2021). Le plan de sondage de l'Enquête démographique et sur la santé du Bangladesh (BDHS) ne prend pas en compte les unités administratives inférieures à l'échelle infranationale (7 divisions), notamment 64 districts et plus de 450 sous-districts (deuxième et troisième niveaux de hiérarchie administrative respectivement). Par conséquent, la taille des échantillons est trop petite pour permettre d'estimer les indicateurs sous le niveau de la division au moyen d'estimateurs classiques fondés sur le plan de sondage. Au total, sept enquêtes ont été réalisées pour la période allant de 1994 à 2014. Elles ont fourni des séries chronologiques d'estimations directes à l'échelle nationale et au niveau de la division concernant les indicateurs susmentionnés, qui permettent de suivre les progrès de la diminution de la mortalité maternelle et néonatale au Bangladesh. Toutefois, pour optimiser l'affectation des ressources et l'élaboration des politiques, des renseignements statistiques plus détaillés et fiables à l'échelle régionale des districts sont

1. Sumonkanti Das, École de démographie, Australian National University, Département d'économie quantitative, Maastricht University; Jan van den Brakel, Département d'économie quantitative, Maastricht University, Département de méthodologie, Bureau central de la statistique des Pays-Bas (CBS). Courriel : ja.vandenbrakel@cbs.nl; Harm Jan Boonstra, Département de méthodologie, Bureau central de la statistique des Pays-Bas (CBS); Stephen Haslett, École des sciences fondamentales et Centre de recherche en santé publique, Massey University, École de recherche en finances, Études d'actuariat et statistique, Australian National University.

nécessaires. C'est à cette fin que des modèles d'estimation sur petits domaines sont élaborés dans le présent article. L'estimation sur petits domaines désigne une classe de procédures d'estimation fondées sur un modèle qui améliorent l'exactitude des estimations directes pour un domaine en augmentant la taille réelle de l'échantillon dans chaque domaine distinct au moyen de l'information de l'échantillon observée dans d'autres domaines ou dans la période de référence précédente. Cette méthode est souvent décrite comme un emprunt d'information dans l'espace ou dans le temps, respectivement (Rao et Molina, 2015).

La BDHS est réalisée de façon répétée avec des intervalles de trois ou quatre ans entre deux enquêtes consécutives. La présente étude porte sur sept éditions couvrant la période de 1994 à 2014. Dans le présent article, des modèles multivariés de séries chronologiques multiniveaux sont élaborés pour produire des estimations fiables des tendances de la couverture des soins prénataux au niveau des districts, des divisions et du pays. Ces modèles sont élaborés dans un cadre bayésien hiérarchique et ajustés au moyen de simulations Monte Carlo par chaînes de Markov. L'avantage d'une approche multivariée des séries chronologiques est qu'elle tire parti de toute l'information disponible en modélisant les corrélations transversales et temporelles entre districts et périodes de référence. Étant donné que les modèles sont définis à une fréquence annuelle, ils prennent en compte les intervalles variables de trois ou quatre ans entre cycles d'enquête. De plus, les modèles de séries chronologiques multiniveaux fournissent des prédictions pour les années sans données d'enquête.

Deux variables de réponses connexes sont examinées dans le présent article : si aucune consultation prénatale n'a été reçue ou si au moins quatre consultations prénatales ont été reçues, variables désignées respectivement par l'abréviation ANC0 (*antenatal care*) et ANC4. Les estimations directes et les estimations de la variance sont calculées à partir des données transversales des sept enquêtes BDHS à l'échelle régionale la plus détaillée des districts et sont utilisées comme données d'entrée dans le modèle de séries chronologiques multiniveaux. Cette méthode présente toutefois un inconvénient : d'autres renseignements auxiliaires, tirés de deux recensements, ne peuvent pas être inclus dans les modèles de séries chronologiques multiniveaux. Le recensement est mené tous les 10 ans. Cela signifie que les mêmes valeurs d'information auxiliaire, disponibles dans un recensement donné, sont utilisées dans deux ou même trois éditions ultérieures de la BDHS menées après ce recensement, mais avant le prochain. Cela produit des chocs dans les prédictions des séries chronologiques multiniveaux pendant les périodes pour lesquelles les renseignements d'un nouveau recensement sont publiés. On contourne ce problème en élaborant la méthode suivante en deux étapes. Dans un premier temps, on applique des modèles transversaux de Fay-Herriot (F-H) (Fay et Herriot, 1979) à chaque cycle d'enquête en utilisant les estimations directes au niveau du district et leurs erreurs-types lissées. L'information auxiliaire de recensement sert à améliorer ces modèles transversaux de F-H. Dans une deuxième étape, ces estimations transversales de F-H sont utilisées comme données d'entrée dans le modèle de séries chronologiques multiniveau. Notons que les prédictions transversales de F-H pour une année d'enquête donnée sont corrélées. Les modèles de séries chronologiques multiniveaux tiennent compte de cette corrélation en utilisant les matrices de variance-covariance complètes des prédictions de F-H transversales comme données d'entrée dans le modèle de séries chronologiques multiniveau. Cette approche en deux étapes présente l'avantage d'éliminer les grandes erreurs d'échantillonnage des estimations directes et de

stabiliser la série de données d'entrée pour les modèles de séries chronologiques multiniveaux. Cela repose toutefois sur l'hypothèse que les séries d'entrées des modèles de séries chronologiques ne présentent pas de biais en raison d'une erreur de spécification des modèles de F-H transversaux. Pour éviter cela, il faut un processus de sélection et d'évaluation minutieux des modèles transversaux de F-H à la première étape.

Les modèles de séries chronologiques multiniveaux empruntent de l'information dans le temps et l'espace de plusieurs manières. Les relations transversales sont modélisées au moyen d'effets fixes ainsi que de pentes et d'ordonnées à l'origine aléatoires au niveau du district, indépendantes ou corrélées. Les corrélations spatiales entre districts sont aussi prises en compte. Les tendances lisses et les tendances locales au niveau des districts, des divisions et du pays servent à modéliser les corrélations temporelles et transversales. Au lieu que soit définie une matrice de corrélation complète entre les termes de perturbation des tendances au niveau du district, les tendances sont définies au niveau de la division, et sont partagées par tous les districts sous-jacents. Les écarts par rapport à cette tendance générale sont modélisés au moyen des tendances au niveau du district. C'est une façon parcimonieuse de modéliser les relations transversales entre districts (Boonstra et van den Brakel, 2019). Les estimations des tendances au niveau des districts sont obtenues directement à partir des résultats des modèles, tandis que les tendances au niveau des divisions et du pays sont obtenues par agrégation des prédictions au niveau des districts. La production d'estimations pour des niveaux d'agrégation plus élevés par agrégation des prédictions à partir de l'échelle régionale la plus détaillée permet de disposer de tableaux de publication numériquement cohérents par définition. Les estimations relatives aux districts pour les années non couvertes par l'enquête et aux districts non couverts l'année de l'enquête sont aussi prédites à partir des modèles de séries chronologiques estimés.

Les modèles de séries chronologiques multiniveaux élaborés dans le présent article sont des extensions du modèle de F-H. Rao et Yu (1994) ont étendu le modèle de F-H pour emprunter de l'information dans le temps en supposant que les effets aléatoires propres à une zone suivent un modèle autorégressif de premier ordre Ar(1) dans le temps, de manière indépendante entre les domaines. Datta, Lahiri, Maiti et Lu (1999), et You et Rao (2000) ont généralisé une extension de la série chronologique du modèle de F-H à partir de Rao et Yu (1994) dans le cadre hiérarchique de Bayes en considérant que les termes d'erreur propres à un domaine suivent un modèle de marche aléatoire de premier ordre dans le temps plutôt qu'un processus Ar(1). Datta, Lahiri et Maiti (2002) ont également utilisé un modèle de marche aléatoire pour la composante temps afin d'estimer le revenu médian des familles de quatre personnes par État en s'appuyant sur des séries chronologiques et des données transversales en fonction de la méthode d'estimation empirique de Bayes. Marhuenda, Molina et Morales (2013) ont étendu le modèle de Rao et Yu (1994) à une version spatio-temporelle du modèle de F-H en intégrant une hypothèse supplémentaire selon laquelle les effets aléatoires au niveau du domaine suivent un processus autorégressif simultané de premier ordre (Pratesi et Salvati, 2008) pour inclure la corrélation spatiale entre les données des domaines voisins. Ces extensions sont très spécifiques à la seule composante des effets aléatoires au niveau du domaine. Dans le modèle spatio-temporel de F-H examiné dans la présente étude, il est possible de spécifier les effets aléatoires à divers niveaux de désagrégation à côté des domaines des niveaux détaillés cibles afin d'intégrer des corrélations spatiales, temporelles et spatio-temporelles entre données. À cet

égard, le modèle bayésien hiérarchique considéré est plus souple que l'autre extension du modèle de F-H. D'autres études pertinentes de modèles de séries chronologiques multiniveaux et de modèles d'espaces d'état qui étendent le modèle de F-H pour emprunter de l'information dans le temps et l'espace se trouvent notamment dans You, Rao et Gambino (2003); You (2008); Pfeffermann et Burck (1990); Pfeffermann et Tiller (2006); Bollineni-Balabay, van den Brakel, Palm et Boonstra (2017); Boonstra et van den Brakel (2022, 2019) et Boonstra, van den Brakel et Das (2021).

Le reste du présent l'article s'organise comme suit. La section 2 décrit le besoin de données statistiques fiables à un échelon régional peu élevé pour évaluer les objectifs de développement durable liés à la mortalité maternelle et néonatale au Bangladesh. La section 3 décrit brièvement les sources de données et le calcul des estimations directes et des estimations de la variance à partir des données de l'enquête BDHS, ainsi que les transformations des estimations directes et l'approche de la fonction généralisée de variance (FGV) aux fins de lissage des estimations de la variance, deux méthodes qui améliorent l'ajustement du modèle. La section 4 décrit le cadre bayésien hiérarchique de la modélisation multiniveau des séries chronologiques. Les modèles retenus pour ANC0 et ANC4 sont présentés à la section 5, accompagnés d'une brève discussion sur le processus de construction des modèles. La section 6 présente une analyse des estimations des tendances fondées sur les modèles élaborés et certains résultats d'évaluation de modèle sont illustrés à la section 7. Le document se termine par une discussion à la section 8.

2. La nécessité de statistiques régionales fiables sur la mortalité maternelle et néonatale au Bangladesh

Le Bangladesh a fait des progrès remarquables dans la réduction du taux de mortalité maternelle et du taux de mortalité néonatale conformément aux objectifs 4 et 5 du rapport intitulé *Objectifs du Millénaire pour le développement*. Cependant, les indicateurs du taux de mortalité maternelle (170 pour 100 000 naissances vivantes [OMS, UNICEF et autres, 2014]) et du taux de mortalité néonatale (28 pour 1 000 naissances vivantes [NIPORT, 2015a]) restent relativement hauts comparativement à l'objectif de développement durable (ODD) de réduction du taux de mortalité maternelle à 70 pour 100 000 naissances vivantes et du taux de mortalité néonatale à 12 décès pour 1 000 naissances vivantes au Bangladesh (BBS, 2020). Ces taux de mortalité élevés s'expliquent principalement par le recours insuffisant aux services de santé maternelle, comme les soins prénataux, les naissances assistées par du personnel soignant qualifié et les soins postnataux (NIPORT, 2016). En effet, il est important de recevoir suffisamment de soins prénataux pendant la grossesse, car cela augmente également le recours à du personnel soignant qualifié pour l'accouchement et à des soins postnataux (Mrisho, Obrist, Schellenberg, Haws, Mushi, Mshinda, Tanner et Schellenberg, 2009).

L'enquête auprès des ménages la plus récente indique que la majorité des femmes enceintes (75 %) au Bangladesh reçoivent des soins prénataux donnés par des fournisseurs ayant reçu une formation médicale. Cependant, la proportion de femmes qui reçoivent au moins 4 consultations prénatales comme recommandés par l'OMS est nettement moins élevée, soit 37 % (BBS et UNICEF, 2019). Ces données donnent à penser que le Bangladesh accuse un retard dans l'atteinte de l'objectif national de 50 % de

femmes qui reçoivent au moins 4 consultations prénatales d'ici 2016. Pour combler cet écart et atteindre l'objectif 3 de l'ODD, qui consiste à porter à plus de 98 % d'ici 2030 (NIPORT, 2015b) le nombre de femmes qui reçoivent au moins 4 consultations prénatales, le pays a besoin d'une stratégie globale et de jalons précis. Les tendances nationales de la couverture des soins prénataux indiquent que la proportion de femmes n'ayant pas reçu de soins prénataux (ANC0) s'est améliorée pour atteindre seulement 17,2 % en 2019, comparativement à 85 % en 1994, tandis que la proportion de femmes ayant reçu au moins quatre visites de soins prénataux (ANC4) a augmenté à 37 % en 2019, alors qu'elle était de 6 % en 1994. L'amélioration des indicateurs pendant cette période varie selon les divisions. L'amélioration la plus marquée est observée dans la division de *Khulna*, où ANC0 et ANC4 sont passés d'environ 70 % et 5 % à environ 12 % et 40 %, respectivement. L'évolution la plus faible a été observée dans la division de *Sylhet*.

Les installations proposant des services de soins prénataux présentent des variations considérables au Bangladesh. On trouve des cliniques communautaires et des centres de protection de la famille au niveau des unions (ainsi que des cliniques d'organisations non gouvernementales), des complexes de santé au niveau des upazila, qui sont des sous-districts, et des hôpitaux de district et tertiaires au niveau des districts. De plus, l'accès aux médecins privés varie selon le degré d'urbanisation ainsi que la distance entre le district ou le sous-district et les villes métropolitaines les plus proches, en particulier la capitale *Dhaka*. L'inégalité d'accès aux soins prénataux est aussi explicitement visible au niveau de la division. On peut s'attendre à ce que les inégalités soient encore plus grandes aux niveaux administratifs désagrégés, comme les districts et les sous-districts. Aucune étude n'a toutefois confirmé cette hypothèse, principalement en raison de l'absence de données d'enquête suffisamment détaillées à ces niveaux. Des données récentes tirées d'études à des niveaux désagrégés sur la pauvreté, la nutrition des enfants et la morbidité indiquent des niveaux élevés d'inégalité au niveau des districts et des sous-districts (Haslett et Jones, 2004; Haslett, Jones et Isidro, 2014; Das, Kumar et Kawsar, 2020; Hossain, Das et Chandra, 2020).

3. Sources des données et estimations des entrées

3.1 Sources des données

Depuis 1993-1994, la BDHS est réalisée sous l'autorité de l'Institut national de recherche et de formation en démographie (NIPORT) du ministère de la Santé et de la Protection de la famille (MOHFW). L'enquête évalue les programmes sociaux et sanitaires en place afin d'élaborer de nouvelles stratégies d'amélioration de l'état de santé des femmes et des enfants du pays. En 2018, huit enquêtes BDHS avaient été menées : en 1993-1994, 1996-1997, 2000, 2004, 2007, 2011, 2014 et 2017-2018. La présente étude s'appuie sur les données d'enquête relatives à la période de 1994 à 2014 puisque l'emplacement au niveau du district des grappes sondées n'est pas divulgué dans la BDHS la plus récente de 2017-2018. De 1994 à 2014, trois recensements de la population et du logement ont été réalisés, en 1991, 2001 et 2011. Les données complètes du recensement ne sont pas disponibles : seulement 10 % des données du Recensement de 1991, 10 % des données du Recensement de 2001 et 5 % des données du Recensement de 2011 sont accessibles au public par l'intermédiaire d'IPUMS-International (https://international.ipums.org). Un certain nombre de variables contextuelles au niveau des districts ont été générées et utilisées dans

l'élaboration de modèles transversaux de F-H afin de produire des estimations d'entrées pour les modèles de séries chronologiques multiniveaux.

3.2 Estimations directes

Les variables analysées dans le présent article sont ANC0 et ANC4. Le Bangladesh est divisé en sept régions infranationales appelées divisions. Ces divisions sont ensuite divisées en 64 districts, soit l'échelle régionale la plus détaillée examinée dans la présente étude. Dans un premier temps, les estimations et les estimations de la variance des deux variables cibles au niveau des districts sont obtenues à partir des données au niveau de l'unité de chaque année d'enquête au moyen d'un estimateur d'enquête direct classique fondé sur le plan (désigné par DIR ci-après), dans lequel les poids d'enquête servent à prendre en compte le plan de sondage et la non-réponse.

Dans la présente étude, les femmes en âge de procréer qui ont été mariées et qui ont accouché au cours des trois années précédant une année d'enquête sont considérées comme la population cible. Étant donné qu'on ne dispose pas de ces renseignements sur la grossesse dans la population du recensement, la taille de la population par domaine est estimée par le nombre de femmes en âge de procréer ayant déjà été mariées qui se trouve dans les trois recensements. Cela signifie que même si les tailles d'échantillon d'un domaine reposent sur un recensement, il y a une certaine incertitude à leur sujet, qui est ignorée dans les modèles d'estimation sur petits domaines (EPD). Pour une analyse plus détaillée sur la taille des populations propres aux divisions et aux districts, voir Das, van den Brakel, Boonstra et Haslett (2021).

La BDHS utilise un échantillon stratifié des ménages à deux degrés. Les strates sont formées à partir des divisions et sous-divisions selon leur caractérisation urbaine-rurale. Les unités primaires d'échantillonnage (UPE) sont les secteurs de dénombrement du recensement de la population et du logement créés pour comporter en moyenne environ 120 ménages (légère variation par rapport au recensement). À la première étape, les UPE sont sélectionnées avec des probabilités proportionnelles à la taille de l'UPE, c'est-à-dire au nombre de ménages. À la deuxième étape, le listage complet des ménages est effectué dans toutes les UPE sélectionnées, puis une trentaine de ménages sont sélectionnés dans chaque UPE au moyen d'un échantillonnage systématique. Le taux de réponse chez les femmes admissibles a dépassé 95 % toutes les années de la BDHS. Bien que la taille de l'échantillon des femmes qui ont été mariées soit supérieure à 10 000 dans toutes les enquêtes, la taille de l'échantillon de la présente étude est plus petite, car seules les femmes déjà mariées qui ont eu un enfant au cours des trois années précédant l'année d'enquête sont prises en compte. Au niveau des districts, la taille moyenne des échantillons varie entre 60 et 114, et certains districts comptent moins de 10 femmes observées, voire aucune.

Les poids de sondage sont calculés en fonction des probabilités de sélection. On a ensuite ajusté ces poids pour tenir compte de la non-réponse des ménages et des personnes. L'estimation directe de la proportion de la population dans un domaine i pour l'année d'enquête t est calculée comme étant la moyenne de l'échantillon.

$$\hat{Y}_{it} = \frac{\sum_{j \in s_{it}} w_{ijt} y_{ijt}}{\sum_{j \in s_{it}} w_{ijt}}, \tag{3.1}$$

où y est la variable de réponse d'intérêt, s_{it} est l'ensemble des femmes déjà mariées dans le domaine i pour lequel y est observé l'année t, et w_{ijt} est le poids de sondage pour la personne j vivant dans le domaine i l'année t. Notons que les poids w_{ijt} sont mis à l'échelle de façon à ce que la somme sur les poids dans l'échantillon soit égale à la taille nette de l'échantillon. Les estimations de la variance correspondantes sont obtenues par approximation comme suit :

$$\text{var}(\hat{Y}_{it}) = \frac{1}{n_{it}(n_{it}-1)} \sum_{j \in s_{it}} w_{ijt} \left(y_{ijt} - \hat{Y}_{it} \right)^2, \tag{3.2}$$

où n_{it} est le nombre de femmes qui ont été mariées observé dans le domaine i au cours de l'année de l'enquête t. Au départ, on estimait la variance en calculant la variance parmi les totaux estimés d'UPE comme si la sélection avait été réalisée au moyen d'un échantillonnage stratifié avec remise, appelé approximation de la variance de l'unité d'échantillonnage finale. Cela a donné des estimations de variance nulle pour quelques domaines. L'approximation de la variance (3.2) évite ces estimations de variance nulle, et donne des estimations de la variance comparables à l'approximation initiale où l'on supposait que les UPE étaient sélectionnées avec remise. Dans le premier modèle de séries chronologiques multiniveaux, représenté par MTS-I, ces estimations directes servent de séries de données d'entrée.

3.3 Estimations transversales de Fay-Herriot

Un des problèmes que pose le modèle MTS-I est qu'il utilise des données du recensement comme variables auxiliaires dans le modèle de séries chronologiques multiniveau. Étant donné que l'intervalle entre deux recensements est de 10 ans, alors que la BDHS est réalisée tous les trois ou quatre ans, les covariables du recensement restent identiques jusqu'à ce que de nouvelles données du recensement soient disponibles. L'inclusion de ces données de recensement comme covariables dans les modèles MTS-I faussera les estimations des tendances et des changements d'une période à l'autre. L'une des façons de tirer parti des données du recensement consiste à modéliser les estimations directes au niveau des districts dans des modèles transversaux de F-H distincts en utilisant des variables contextuelles pertinentes extraites des données de recensement. On s'attend aussi à ce que l'utilisation de variables auxiliaires de recensement disponibles à temps dans des modèles transversaux de F-H répétitifs puisse avoir une incidence sur les coefficients de régression et l'exactitude des prédictions de modèle de la variable dépendante, mais non sur les prédictions de la variable dépendante elle-même. Comparativement aux estimations directes utilisées dans MTS-I, ces modèles transversaux de F-H fournissent aussi de meilleures estimations parce qu'ils empruntent certaines informations dans les districts.

Les estimations transversales de F-H et leurs erreurs-types sont utilisées comme données d'entrée pour un deuxième modèle, désigné MTS-II. Les estimations transversales de F-H sont corrélées en raison de leurs composantes communes à effet fixe, qui sont ignorées dans MTS-II. Par conséquent, un troisième modèle de séries chronologiques multiniveaux, désigné MTS-III, est élaboré au moyen des estimations transversales de F-H et de leur matrice de covariance complète comme données d'entrée.

Les composantes à effet fixe et aléatoire pour les modèles transversaux de F-H spécifiques à l'enquête sont présentées dans les tableaux A.2 et A.3 de l'annexe. Pour tous les modèles, on suppose que les effets aléatoires suivent une loi normale. On a examiné des modèles non normaux pour les effets aléatoires (Laplace et *horseshoe*) et l'erreur d'échantillonnage (distribution t) comme solutions de rechange à la loi normale. Cela n'a toutefois pas amélioré l'ajustement du modèle.

3.4 Fonctions généralisées de variance

Dans les modèles de F-H et de séries chronologiques multiniveaux, les estimations de la variance des estimations directes sont en grande partie traitées comme des quantités fixes. Comme ces estimations de la variance peuvent présenter un bruit important, elles sont lissées au moyen d'une fonction généralisée de variance (FGV) avant d'être utilisées dans les modèles de F-H et de séries chronologiques multiniveaux. Il est entendu qu'un district sans données d'échantillon est considéré comme manquant et qu'il n'est donc pas pris en compte dans la méthode d'élaboration du modèle. Le modèle de F-H transversal peut produire des estimations et des erreurs-types pour ces domaines hors échantillon. On n'utilise toutefois pas ces estimations synthétiques dans l'élaboration des modèles MTS-II et MTS-III pour obtenir une meilleure comparaison avec le modèle MTS-I.

Les FGV sont des modèles de régression qui relient les estimations de la variance à des prédicteurs comme la taille de l'échantillon, les variables du plan de sondage et les estimations ponctuelles (Wolter (2007), chapitre 7). Pour les variables ANC0 et ANC4, on utilise la FGV suivante :

$$\log \mathrm{se}(\hat{Y}_{it}) = \alpha + \beta \log \tilde{Y}_{it} + \gamma \log(m_{it} + 1) + \delta\, \mathrm{Division} + \epsilon_{it}, \tag{3.3}$$

où $\mathrm{se}(\hat{Y}_{it})$ est l'erreur-type de $\hat{Y}_{it}$ dans (3.1), m_{it} le nombre d'unités d'échantillonnage contribuant au district i l'année t et Division est une variable catégorique ayant 7 niveaux. Étant donné que nous ne pouvons pas nous fier aux estimations directes pour les valeurs très petites de m_{it}, les $\tilde{Y}_{it}$ du côté droit de (3.3) sont des estimations lissées simples.

$$\begin{aligned} \tilde{Y}_{it} &= \lambda_{it} \hat{Y}_{it} + (1 - \lambda_{it})\, \bar{Y}_{d[i]t}, \\ \lambda_{it} &= \frac{m_{it}}{m_{it} + 1}, \end{aligned} \tag{3.4}$$

où $\bar{Y}_{d[i]t}$ désigne la moyenne pour la division d ($d = 1$ à 7) à laquelle appartient le district i, pour l'année t. Comme l'a mentionné un réviseur, un estimateur par la régression composite peut être utilisé comme solution de rechange pour (3.4).

Les erreurs de régression ϵ_{it}, sont censées être indépendantes et normalement distribuées avec un paramètre de variance commun σ^2. Les FGV sont ajustées uniquement aux districts pour lesquels les erreurs-types des estimations directes ne sont pas nulles. Les erreurs-types prédites (lissées) qui reposent sur les modèles ajustés sont

$$\mathrm{se}_{\mathrm{pred}}(\hat{Y}_{it}) = \exp\left(\hat{\alpha} + \hat{\beta} \log \tilde{Y}_{it} + \hat{\gamma} \log(m_{it} + 1) + \hat{\delta}\, \mathrm{Division} + \hat{\sigma}^2/2\right), \tag{3.5}$$

où $\hat{\sigma}$ est 0,03 pour ANC0 et 0,003 pour ANC4, respectivement. Les valeurs de R au carré pour les deux modèles sont assez élevées : 0,79 pour ANC0 et 0,99 pour ANC4. Notons que la rétro transformation exponentielle dans (3.5) comprend une correction de biais qui, dans ce cas, n'a qu'un petit effet. Cette méthode sert à obtenir des erreurs-types lissées pour les modèles de F-H transversaux et le modèle MTS-I.

3.5 Transformations des séries de données d'entrée

La transformation par la racine carrée, la transformation logarithmique et la transformation log-ratio sont considérées comme une transformation stabilisatrice de la variance, voir Sakia (1992). La transformation par la racine carrée est appliquée aux données de ANC4 (les modèles de séries chronologiques multiniveaux et les modèles de F-H transversaux), car cette transformation réduit la corrélation entre les estimations ponctuelles et leurs erreurs-types des séries de données d'entrée, diminue l'hétérogénéité, améliore la convergence de la simulation par la méthode de Monte Carlo par chaînes de Markov, et réduit l'asymétrie des données de proportion si elles prennent des valeurs proches de la borne inférieure de zéro. Pour ce qui est de ANC0, la transformation par la racine carrée est utilisée uniquement pour les modèles transversaux de F-H spécifiques aux années 2011 et 2014. Aucune transformation n'est appliquée aux autres années. Dans les trois modèles de séries chronologiques multiniveaux, aucune transformation n'est appliquée pour ANC0, car la transformation par la racine carrée des séries de données d'entrée augmente la dépendance entre les estimations directes et les erreurs-types.

Soit $\hat{\mathcal{Y}}_{it} = \sqrt{(\hat{Y}_{it} + \varepsilon)}$ désigne les estimations directes transformées par la racine carrée, où ε est un petit nombre (0,005), ce qui est nécessaire parce que pour certains districts, les estimations directes sont égales à zéro. Au moyen d'une approximation de Taylor du premier degré, on peut montrer que $\mathrm{se}(\hat{\mathcal{Y}}_{it}) \approx \mathrm{se}(\hat{Y}_{it}) \Big/ \left(2\sqrt{\hat{Y}_{it} + \varepsilon}\right)$.

Si la FGV (3.3) est appliquée aux erreurs-types des estimations directes non transformées, alors les erreurs-types pour les domaines comportant un très petit nombre d'unités d'échantillonnage peuvent devenir excessivement grandes en raison de l'approximation de la linéarisation. On évite ce problème en appliquant la FGV aux erreurs-types des estimations transformées, c'est-à-dire $\mathrm{se}(\hat{\mathcal{Y}}_{it})$.

4. Modélisation multiniveau de séries chronologiques

Dans la présente étude, les estimations directes et leurs erreurs-types sont disponibles pour les années d'enquête 1994, 1997, 2000, 2004, 2007, 2011 et 2014. Pour tenir compte des intervalles variables de trois ou quatre ans entre les années d'enquête, les modèles de séries chronologiques multiniveaux sont définis à une fréquence annuelle (c'est-à-dire que les valeurs désignent une période de référence d'un an) à l'échelle régionale la plus détaillée des 64 districts. Sur une durée de 21 ans, on a 1 344 combinaisons domaine-année. À partir des 7 années d'enquête disponibles, le modèle est adapté aux 448 observations domaine-année. Les années entre deux enquêtes consécutives sont définies comme étant manquantes dans le modèle. Ainsi, l'évolution de la tendance d'une période à l'autre est précisée correctement et le modèle fournit des prédictions pour les combinaisons domaine-année manquantes.

Pour plus de commodité, nous désignerons par $\hat{Y}_{it}$ la série d'entrées des modèles de séries chronologiques pour ANC0 ou ANC4 pour l'année t et le domaine i. Il peut s'agir des estimations directes non transformées, des estimations directes transformées par la racine carrée ou des prédictions du modèle obtenues au moyen de modèles transversaux de F-H. Dans le cas présent, l'indice de domaine i s'exécute de 1 à $M_d = 64$ et l'indice temporel t de 1 à $T = 21$. Nous combinons ensuite ces estimations dans un vecteur $\hat{Y} = (\hat{Y}_{11}, \ldots \hat{Y}_{M_d 1}, \ldots \hat{Y}_{1T}, \ldots \hat{Y}_{M_d T})'$, un vecteur de dimension $M = M_d T$.

4.1 Structure du modèle

Les modèles multiniveaux examinés prennent une forme additive linéaire générale

$$\hat{Y} = X\beta + \sum_{\alpha} Z^{(\alpha)} v^{(\alpha)} + e, \tag{4.1}$$

où X est une matrice de plan $M \times p$ pour un vecteur d'effets fixes β de dimension p et les $Z^{(\alpha)}$ sont des matrices de plan $M \times q^{(\alpha)}$ pour les vecteurs d'effets aléatoires $v^{(\alpha)}$ de dimension $q^{(\alpha)}$. Dans le cas présent, la somme sur α s'exécute sur plusieurs termes possibles d'effet aléatoire à différents niveaux, comme le niveau local et des tendances lisses au niveau du district et de la division, le bruit blanc au niveau le plus détaillé des domaines M, etc. Des explications plus détaillées sont données ci-dessous. Dans la formule (4.1), $e = (e_{11}, \ldots, e_{M_d 1}, \ldots e_{M_d T})'$ désigne, selon la série de données d'entrée, les erreurs d'échantillonnage des estimations directes ou les erreurs de prédiction du modèle transversal de F-H. Les erreurs d'échantillonnage sont considérées comme normalement distribuées, soit $e \sim N(0, \Sigma)$ où $\Sigma = \oplus_{t=1}^{T} \Sigma_t$. Si les séries de données d'entrée sont les estimations directes non transformées, alors Σ_t est la matrice de covariance pour les estimations directes non transformées observées l'année t. Si les séries de données d'entrée sont transformées, alors Σ_t est la matrice de covariance pour les estimations directes transformées, comme cela est décrit dans la sous-section 3.5. Si les séries de données d'entrée sont les prédictions qui reposent sur des modèles transversaux de F-H, alors Σ_t contient les erreurs quadratiques moyennes estimées des prédictions de F-H. Dans MTS-II, Σ_t est diagonale et ignore les corrélations entre les prédictions de domaine. Dans MTS-III, Σ_t est une matrice de covariance complète qui prend aussi en compte les corrélations entre les prédictions de domaine.

Selon la répartition des erreurs d'échantillonnage e dans (4.1), la fonction de vraisemblance conditionnelle à des paramètres d'effets fixes et aléatoires peut être définie comme étant

$$p\left(\hat{Y} \mid \eta, \Sigma\right) = N\left(\hat{Y} \mid \eta, \Sigma\right), \tag{4.2}$$

où $\eta = X\beta + \sum_{\alpha} Z^{(\alpha)} v^{(\alpha)}$ est le prédicteur linéaire. Pour les erreurs e, on peut considérer qu'une loi t de Student au lieu d'une loi normale donne moins de poids à un plus grand nombre d'observations aberrantes, d'après West (1984).

La partie à effets fixes de η peut contenir des composantes comme une ordonnée à l'origine, une tendance linéaire, les effets principaux pour le niveau de la division et du district et peut-être les interactions de second ordre pour les tendances linéaires et la division ou le district. On attribue au vecteur

β d'effets fixes une loi *a priori* normale $p(\beta) = N(0, 100 I_p)$, avec I_x la matrice d'identité de dimension $x \times x$. Cela n'est que très peu informatif, car une erreur-type de 10 est très grande par rapport aux échelles des estimations directes (transformées) et des covariables utilisées.

Le deuxième terme du côté droit de (4.1) consiste en une somme des contributions au prédicteur linéaire par des effets aléatoires ou des termes de coefficient variables. On suppose que les vecteurs d'effets aléatoires $v^{(\alpha)}$ pour différents α sont indépendants, mais que les composantes à l'intérieur d'un vecteur $v^{(\alpha)}$ peuvent être corrélées pour tenir compte de la corrélation temporelle ou transversale. Afin de décrire le modèle général de chaque vecteur d'effets aléatoires $v^{(\alpha)}$, nous supprimons l'exposant α dans ce qui suit aux fins de simplification de la notation.

On suppose que chaque vecteur d'effets aléatoires v est distribué ainsi

$$v \sim N(0, A \otimes V), \tag{4.3}$$

où V et A sont les matrices de covariance de dimension $d \times d$ et $l \times l$, respectivement, et $A \otimes V$ désigne le produit de Kronecker de A par V. La longueur totale de v est $q = dl$, et ces coefficients peuvent être considérés comme correspondant aux effets d autorisés à varier sur l niveaux d'une variable de facteur. Si, par exemple, V correspond à la division, alors V définit $d = 7$ différents effets aléatoires qui correspondent aux 7 catégories de division. Si ensuite A correspond au temps, alors $l = 21$ années. Dans ce cas, chacun des 7 effets peut varier sur ses 21 niveaux (des années en l'occurrence). Chaque effet aléatoire généré pour une combinaison division $\times$ année est partagé par tous les districts appartenant à cette division pour cette année particulière.

La matrice de covariance A décrit la structure de covariance entre les niveaux de la variable de facteur et est censée être connue. Au lieu des matrices de covariance, on utilise en fait les matrices de précision $Q_A = A^{-1}$, pour un calcul plus efficace (Rue et Held, 2005). La matrice de covariance V pour les d effets variables peut être paramétrée de l'une des trois façons suivantes : i) une matrice de covariance paramétrée complète; ii) une matrice diagonale avec des éléments diagonaux inégaux; iii) une matrice diagonale avec des éléments diagonaux égaux. Une loi *a priori* de Wishart inversée et mise à l'échelle est utilisée comme ce qui est proposé dans O'Malley et Zaslavsky (2008) et recommandé par Gelman et Hill (2007) quand on suppose une matrice de covariance complète, tandis que les lois *a priori* demi-Cauchy sont utilisées pour les écarts-types quand la matrice de covariance est supposée diagonale avec des éléments égaux ou inégaux. En cas de variances diagonales, les lois *a priori* demi-Cauchy sont de meilleures lois *a priori* par défaut que les lois *a priori* gamma inverses plus courantes (Gelman, 2006).

Les structures d'effet aléatoire suivantes sont examinées dans la procédure de sélection du modèle :

1. Ordonnées à l'origine aléatoires pour les domaines M_d. Dans ce cas, $A = I_{M_d}$ et V est un paramètre de variance scalaire. Cela suppose que $v_{it} = v_i, \forall t$ et $v_i \sim N(0, \sigma_I^2)$.
2. Marches aléatoires de premier ou de second ordre à différents niveaux d'agrégation. Une marche aléatoire de premier ordre ou une tendance locale au niveau du district est définie comme étant $v_{it} = L_{it}$, avec $L_{it} = L_{i,t-1} + \eta_{it}$ et $\eta_{it} \sim N(0, \sigma^2_{R1,i})$. Une marche aléatoire de second ordre ou un modèle de tendance lissée au niveau du district est définie comme étant $v_{it} = L_{it}$, avec $L_{it} = L_{i,t-1} + R_{i,t-1}$ et $\eta_{it} \sim N(0, \sigma^2_{R2,i})$. Les deux types de tendances peuvent être définis

similairement au niveau de la division ou à l'échelle nationale. Voir dans Rue et Held (2005) la spécification de la matrice de précision Q_A pour les marches aléatoires de premier et de second ordre. On peut considérer une matrice de covariance complète pour les innovations de tendance afin de permettre des corrélations transversales en plus des corrélations temporelles, ou une matrice diagonale avec des paramètres de variance différents ou égaux pour permettre des corrélations temporelles seulement. Dans le cas de variances égales, $\sigma^2_{R1,i} = \sigma^2_{R1}$ et $\sigma^2_{R2,i} = \sigma^2_{R2}, \forall i.$ Les composantes des marches aléatoires de premier et de second ordre au niveau du district sont désignées ci-dessous respectivement par RW1_District et RW2_District. Au niveau de la division, elles sont désignées par RW1_Division et RW2_Division.

3. Les marches aléatoires de premier ordre utilisées dans nos modèles ne peuvent pas saisir de niveau global, car les effets aléatoires correspondants sont soumis à une contrainte de somme nulle dans le temps. De même, les marches aléatoires de second ordre ne peuvent pas saisir à la fois le niveau et la tendance linéaire. Cela signifie que le niveau et la tendance linéaire doivent être pris en compte par d'autres termes du modèle, comme les effets fixes ou aléatoires. Les ordonnées à l'origine au niveau du district ont été traitées au point 1. Pour inclure aussi les tendances linéaires par district, cette composante peut être étendue aux ordonnées à l'origine et aux pentes aléatoires linéaires dans le temps. Dans ce cas, V peut être soit une matrice de covariance générale 2×2

$$V = \begin{pmatrix} \sigma_I^2 & \rho_{IS}\,\sigma_I\,\sigma_S \\ \rho_{IS}\,\sigma_I\,\sigma_S & \sigma_S^2 \end{pmatrix},$$

qui prend en compte les corrélations entre les ordonnées à l'origine et les pentes, ou une matrice diagonale avec des éléments diagonaux σ_I^2 et σ_S^2 les variances de l'ordonnée à l'origine et des pentes aléatoires, respectivement. Cette composante du modèle est désignée par RIS_District ci-dessous.

4. Effets aléatoires spatiaux : ordonnées à l'origine aléatoires qui varient selon l'emplacement spatial des districts selon un modèle autorégressif conditionnel intrinsèque (ICAR) (Besag et Kooperberg, 1995), défini comme étant $v_i \,|\, v_{-i} \sim N\left((\Sigma_{i' \in nb(i)} v_{i'}) \,/\, a_i\,,\, \sigma^2_{Sp} \,/\, a_i\right)$ pour chaque effet spatial conditionnel aux autres. Dans le cas présent, $nb(i)$ est l'ensemble des domaines voisins du domaine i et a_i le nombre de domaines voisins du domaine i. Voir dans Rue et Held (2005) la spécification de la matrice de précision Q_A. Cette composante spatiale est désignée ultérieurement par Spatial_District.

5. Bruit blanc : pour permettre une variation aléatoire inexpliquée, on peut inclure le bruit blanc au niveau de domaine le plus détaillé par année. Dans ce cas $A = I_M$ et V est un paramètre de variance scalaire. Cela suppose que $v_{it} \sim N(0, \sigma^2_W)$.

Nous avons aussi étudié des généralisations de (4.3) en répartitions anormales d'effets aléatoires en mettant en œuvre une loi de Student, une répartition *a priori horseshoe* (Carvalho, Polson et Scott, 2010)

et Laplace (Tibshirani, 1996; Park et Casella, 2008). Les queues plus lourdes de ces répartitions de rechange permettent de grands effets occasionnels. Cependant, ces répartitions n'ont pas amélioré les résultats pour les variables cibles considérées en ce qui concerne les critères d'information du modèle ainsi que les prédictions des tendances sous-jacentes. C'est pourquoi la loi normale est utilisée pour toutes les composantes à effet aléatoire. La présentation exacte des modèles de séries chronologiques multiniveaux définitifs pour ANC0 et ANC4 est précisée respectivement dans les sous-sections 5.1 et 5.2.

4.2 Estimation des modèles

Les modèles sont ajustés au moyen d'un échantillonnage par la méthode de Monte Carlo par chaînes de Markov (MCMC), en particulier l'échantillonneur de Gibbs (Geman et Geman, 1984; Gelfand et Smith, 1990). Voir dans Boonstra et van den Brakel (2022) la spécification des lois conditionnelles complètes. Les modèles précisés dans la sous-section 4.1 sont exécutés dans R (R Core Team, 2015) au moyen du progiciel `mcmcsae` (Boonstra, 2021). L'échantillonneur de Gibbs est exécuté en parallèle pour trois chaînes indépendantes avec des valeurs de départ générées aléatoirement. Dans l'étape de construction du modèle, 1 000 itérations sont utilisées, en plus d'une période de « rodage » de 100 itérations. Cela suffisait pour les estimations raisonnablement stables de Monte Carlo des paramètres du modèle et les prédictions de tendances. Pour le modèle sélectionné, nous utilisons un cycle plus long de 1 000 exécutions de rodage plus 5 000 itérations sur lesquelles les tirages d'une itération sur cinq sont stockés. Cela donne $3 \times 1\,000 = 3\,000$ tirages pour calculer les estimations et les erreurs-types. On évalue la convergence de la simulation par MCMC au moyen de tracés et de graphiques d'autocorrélation ainsi que du facteur de réduction d'échelle potentiel de Gelman-Rubin (Gelman et Rubin, 1992), qui décèle le mélange de chaînes. Pour la simulation plus longue du modèle sélectionné, tous les paramètres du modèle et les prédictions du modèle ont des facteurs de réduction d'échelle potentiels inférieurs à 1,01 et un nombre suffisamment efficace de tirages indépendants.

De nombreux modèles de la forme (4.1) ont été ajustés aux données. Afin de comparer les modèles au moyen des mêmes données d'entrée, nous utilisons le critère d'information largement applicable aussi appelé critère d'information Watanabe-Akaike (WAIC) (Watanabe, 2010, 2013) et le critère d'information de déviance (DIC) (Spiegelhalter, Best, Carlin et van der Linde, 2002). Nous comparons aussi les modèles graphiquement selon leurs ajustements et prédictions de tendances à trois niveaux d'agrégation.

5. Modèles sélectionnés et prédiction des modèles

5.1 Modèle de séries chronologiques multiniveaux pour ANC0

Aucune transformation n'est examinée pour les séries de données d'entrée des estimations directes ou des estimations de F-H. Les composantes à effet fixe suivantes sont incluses dans les modèles sélectionnés pour MTS-I, MTS-II et MTS-III :

$$1 + \text{Division} + yr.c + \text{Division} * yr.c, \tag{5.1}$$

où $yr.c$ désigne la variable d'année quantitative standardisée, qui définit une tendance linéaire à effet fixe. De même, Division * $yr.c$ définit une tendance linéaire à effet fixe pour chaque division distincte. Le tableau 5.1 présente la partie des effets aléatoires des trois modèles. Si de multiples effets variables sont modélisés, il y a donc un choix entre une matrice V scalaire, diagonale ou de covariance complète dans (4.3). Pour les variations au fil du temps, les marches aléatoires de second ordre RW2_Division et RW2_District ont finalement été sélectionnées. Les composantes de bruit blanc sont considérées, mais ne sont pas incluses dans le modèle final, car elles n'amélioraient pas l'ajustement du modèle.

Tableau 5.1
Résumé des composantes à effet aléatoire pour le modèle de séries chronologiques multiniveau sélectionné pour ANC0. Les deuxième et troisième colonnes font référence aux effets variables avec la matrice de covariance V dans (4.3), tandis que la quatrième colonne fait référence à la variable de facteur associée avec A dans (4.3). La dernière colonne contient le nombre total d'effets aléatoires pour chaque composante

Composante du modèle	Formule V	Structure de la variance	Facteur A	N^bre d'effets
RIS_District	$1+yr.c$	complète	District	128
RW2_Division	Division	scalaire	RW2(année)	147
RW2_District	District	scalaire	RW2(année)	1 344
District spatial	1	scalaire	Spatial(District)	64

Le prédicteur linéaire du modèle sélectionné peut être écrit, selon les éléments pour le district i et l'année t, comme suit :

$$\eta_{it} = \beta' x_{it} + \nu_i + z_t \nu_i^{(yr)} + u_{it} + u_{j[i]t}^{(div)} + s_i, \tag{5.2}$$

où β est le vecteur des effets fixes correspondant aux covariables x_{it} conformément à la spécification dans (5.1), ν_i désigne des ordonnées à l'origine aléatoires variant selon le district, z_t désigne la $yr.c$ variable pour l'année t, et $\nu_i^{(yr)}$ les pentes aléatoires correspondantes variant selon le district. Ces ordonnées à l'origine et pentes aléatoires sont distribuées conjointement comme suit :

$$\begin{pmatrix} \nu_i \\ \nu_i^{(yr)} \end{pmatrix} \overset{\text{iid}}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_I^2 & \rho\sigma_I\sigma_S \\ \rho\sigma_I\sigma_S & \sigma_S^2 \end{pmatrix} \right). \tag{5.3}$$

Les effets de marche aléatoire de second ordre au niveau du district et de la division sont distribués ainsi :

$$\begin{aligned} u_{it} - 2u_{i(t-1)} + u_{i(t-2)} &\overset{\text{iid}}{\sim} N(0, \sigma_{R2}^2) \\ u_{j[i]t}^{(div)} - 2u_{j[i](t-1)}^{(div)} + u_{j[i](t-2)}^{(div)} &\overset{\text{iid}}{\sim} N\left(0, (\sigma_{R2}^{(div)})^2\right), \end{aligned} \tag{5.4}$$

où $j[i]$ doit être lu comme la division j à laquelle le district i appartient. Enfin, les effets spatiaux s_i sont distribués comme suit :

$$s_i \mid s_{i' \neq i} \overset{\text{ind}}{\sim} N\left(\frac{1}{a_i} \sum_{i' \in nb(i)} s_{i'},\ \frac{1}{a_i} \sigma^2_{Sp} \right), \tag{5.5}$$

où a_i est la taille de l'ensemble $nb(i)$ des districts voisins du district i. Les distributions *a priori* pour la matrice de covariance de (5.3) et les autres paramètres de variance sont choisis de la façon décrite à la section 4.1. Pour que les composantes du modèle soient identifiables, les contraintes suivantes sont imposées :

$$\begin{aligned}
&\sum_{t=1}^{T} u_{it} = 0 \quad \text{et} \quad \sum_{t=1}^{T} t u_{it} = 0 \quad \text{pour tous les districts } i, \\
&\sum_{t=1}^{T} u^{(div)}_{j[i]t} = 0 \quad \text{et} \quad \sum_{t=1}^{T} t u^{(div)}_{j[i]t} = 0 \quad \text{pour toutes les divisions } j, \\
&\sum_{i=1}^{M_d} s_i = 0.
\end{aligned} \tag{5.6}$$

Notons que les tendances de RW2 sont précisées au niveau des divisions et des districts, chaque fois avec une structure de variance scalaire. Toute tendance au niveau de la division est partagée par tous les districts sous-jacents. Les écarts de chaque district par rapport à cette tendance au niveau de la division sont modélisés avec les tendances de la RW2 au niveau du district. Il s'agit d'une solution de rechange parcimonieuse permettant d'emprunter de l'information dans le temps et l'espace, comparativement à ce qu'apporterait la modélisation des tendances de RW2 au niveau du district seulement avec une matrice de covariance complète (Boonstra et van den Brakel, 2019).

5.2 Modèle de séries chronologiques pour ANC4

La transformation par la racine carrée est appliquée aux séries de données d'entrée des estimations directes et de F-H de ANC4 pour les modèles MTS-I, MTS-II et MTS-III. Pour MTS-I, on applique la FGV (3.3) aux erreurs-types transformées afin d'obtenir la matrice de la variance Σ, comme cela est expliqué à la fin de la sous-section 3.5. Pour la composante à effet fixe, une variable de facteur appelée « Région » a été créée en fonction du degré d'urbanisation d'après Rahman, Mohiuddin, Kafy, Sheel et Di (2019). La variable comporte quatre niveaux : 1) pour trois grandes villes, *Dhaka*, *Chittagong* et *Gazipur*; 2) pour neuf autres grandes villes régionales (*Barisal*, *Bogra*, *Comilla*, *Khulna*, *Mymensing*, *Narayanganj*, *Rajshahi*, *Rangpur*, *Sylhet*); 3) pour trois districts vallonnés (*Bandarban*, *Khagrachhari* et *Rangamati*); 4) pour les districts restants. Cette variable a surtout aidé dans l'ajustement des estimations pour les trois districts vallonnés sur lesquels les sept enquêtes examinées comportent très peu (voire pas) d'information. Le modèle final comprend les composantes à effets fixes suivantes :

$$1 + \text{Division} + yr.c + \text{Région}. \tag{5.7}$$

On constate que l'interaction entre « Division » et « yr.c » (comme dans le modèle ANC0) est négligeable dans le modèle ANC4. Les composantes à effet aléatoire du modèle ANC4 présentées au tableau 5.2 sont

très semblables à celles utilisées pour le modèle ANC0 (voir le tableau 5.1). On a considéré une tendance au niveau local plutôt qu'une tendance lisse au niveau de la division (RW1_Division dans le tableau 5.2) puisque la composante de tendance lisse (RW2_Division, comme dans le tableau 5.1) produisait un certain biais dans les tendances à l'échelle du pays et de la division. De plus, le modèle avec la composante RW1_Division donne de meilleurs résultats pour les critères d'information que le modèle avec la composante RW2_Division. Les composantes de bruit blanc sont considérées, mais ne sont pas incluses dans le modèle final, car elles n'amélioraient pas l'ajustement du modèle.

Tableau 5.2
Résumé des composantes à effet aléatoire pour le modèle de séries chronologiques multiniveau sélectionné pour ANC4. Les deuxième et troisième colonnes font référence aux effets variables avec la matrice de covariance V dans (4.3), tandis que la quatrième colonne fait référence à la variable de facteur associée avec A dans (4.3). La dernière colonne contient le nombre total d'effets aléatoires pour chaque terme

Composante du modèle	Formule V	Structure de la variance	Facteur A	N^bre d'effets
RIS_District	$1 + yr.c$	complète	District	128
RW1_Division	Division	scalaire	RW1(année)	147
RW2_District	District	scalaire	RW2(année)	1 344
District spatial	1	scalaire	Spatial(District)	64

Par ailleurs, le modèle peut être exprimé comme dans (5.2), où désormais β et x_{it} correspondent à la spécification des effets fixes (5.7). La seule autre différence est que les tendances au niveau de la division sont maintenant modélisées comme une marche aléatoire de premier ordre :

$$u_{j[i]t} - u_{j[i](t-1)} \overset{\text{iid}}{\sim} N\left(0, (\sigma_{R1}^{(div)})^2\right), \tag{5.8}$$

où pour des raisons d'identifiabilité, la contrainte $\sum_{t=1}^{T} u_{jt}^{(div)} = 0$ est imposée pour toute division j. Comme dans le cas de ANC0, les tendances de RW1 sont précisées au niveau des divisions et les tendances de RW2 au niveau des districts, les deux avec une structure de variance scalaire comme moyen parcimonieux d'emprunter de l'information dans le temps et l'espace.

5.3 Estimation des tendances

Les estimations des tendances sont calculées à partir des résultats de la simulation MCMC. Dans une première étape, pour chaque réplique de MCMC, un vecteur de dimension M contenant des prédictions au niveau le plus détaillé de toutes les combinaisons année-district est calculé comme suit :

$$\eta^{(r)} = X\beta^{(r)} + \sum_{\alpha} Z^{(\alpha)} v^{(\alpha,r)}, \tag{5.9}$$

où l'exposant (r) indexe les tirages MCMC retenus. Notons que $\eta^{(r)}$ comprend aussi les prédictions pour les années sans observations d'enquête. Étant donné qu'une transformation par la racine carrée a été

appliquée à aux séries de ANC4, la rétrotransformation suivante pour les vecteurs $\eta^{(r)}$ a été examinée, d'après Boonstra et coll. (2021) :

$$\theta^{(r)} = (\eta^{(r)})^2 + \left(\text{se}(\hat{\mathcal{Y}}_{it})\right)^2. \tag{5.10}$$

Le deuxième terme du côté droit est une correction de biais (relativement petite) utilisant les erreurs-types transformées et lissées. La correction du biais découle du fait que l'espérance par rapport au plan de sondage des estimations directes peut être écrite comme suit :

$$E(\hat{Y}) = E\left((\hat{\mathcal{Y}})^2\right) = E\left((\eta + \mathcal{E})^2\right) = \eta^2 + \text{var}(\mathcal{E}),$$

où $\mathcal{E}$ est le vecteur des erreurs d'échantillonnage après transformation, censé être normalement distribué avec des erreurs-types $\text{se}(\hat{\mathcal{Y}}_{it})$. Le problème posé par les données dont nous disposons est que la correction du biais ne peut être appliquée qu'aux années d'enquête, étant donné que les erreurs-types ne sont disponibles que pour ces années. L'application de la correction du biais uniquement pour les années d'enquête fausse les estimations des tendances, comme le montrent Das, van den Brakel, Boonstra et Haslett (2021). Dans le cas du modèle MTS-I, l'effet de cette correction de biais est plus évident pour les domaines pour lesquels il n'y a pas d'estimations directes, en particulier pour les districts vallonnés de *Chittagong*. L'effet de la correction du biais est moindre dans le cas des modèles MTS-II et MTS-III, puisque les erreurs-types estimées par les estimations de F-H sont déjà suffisamment lissées et convergentes. Toutefois, à l'échelle nationale et au niveau de la division, cette correction de biais entraîne une certaine surestimation dans certaines années d'enquête pour toutes les tendances qui reposent sur des modèles MTS. Par conséquent, la correction du biais pour la transformation par la racine carrée n'est pas appliquée dans les estimations de tendances, mais seulement utilisée dans le calcul des estimations transversales de F-H.

On obtient les estimations des tendances avec leurs erreurs-types au niveau le plus détaillé des districts pour toutes les années en prenant la moyenne et l'écart-type sur les répliques MCMC $\eta^{(r)}$, respectivement. Pour obtenir les tendances au niveau de la division et du pays, on agrège chaque réplique MCMC à partir de l'échelle régionale la plus détaillée des districts, en utilisant le nombre de femmes qui ont été mariées comme variable de pondération. Ensuite, on obtient les estimations de tendance et leurs erreurs-types en prenant la moyenne et l'écart-type sur ces répliques MCMC agrégées.

6. Résultats

Les tendances de ANC0 et de ANC4 présentées dans les figures comprennent cinq types d'estimations dont les intervalles de confiance sont d'environ 95 % : i) les estimations directes pondérées pour l'année d'enquête (ligne de barre d'erreur noire); ii) les estimations transversales de F-H pour l'année d'enquête (ligne de barre d'erreur verte); iii) les estimations qui reposent sur le modèle MTS-I (ligne rouge); iv) les estimations qui reposent sur le modèle MTS-II (ligne verte); v) les estimations qui reposent sur le modèle MTS-III (ligne bleue).

6.1 ANC0

Les tendances nationales de ANC0 sont présentées à la figure 6.1. La figure montre que les estimations directes et de F-H transversales sont très semblables pour les années d'enquête, selon un intervalle de confiance (IC) d'environ 95 %. On peut s'y attendre pour les chiffres à l'échelle nationale, puisque le gain de précision obtenu avec un modèle de prédiction sur petits domaines pour ce qui est d'un estimateur direct diminue à mesure que la taille de l'échantillon augmente. Pendant la première période allant de 1994 à 2000, la tendance nationale qui repose sur le modèle MTS-I suit les estimations directes et de F-H transversales, tandis que les tendances qui reposent sur les modèles MTS-II et MTS-III sont légèrement plus élevées. Pour la période allant de 2004 à 2010, la tendance qui repose sur le modèle MTS-I est légèrement supérieure aux tendances qui reposent sur les modèles MTS-II et MTS-III. Les différences sont cependant très minimes.

Les tendances au niveau de la division, illustrées à la figure 6.2, indiquent que les tendances dans MTS-I sont très semblables à celles qui reposent sur les modèles MTS-II et MTS-III, à quelques petites exceptions près dans les divisions de *Dhaka*, de *Khulna* et de *Rajshahi*. Les différences entre les divisions de *Dhaka* et de *Khulna* peuvent être à l'origine de la plupart des différences observées dans les tendances à l'échelle nationale.

Les tendances qui reposent sur les modèles MTS-II et MTS-III sont presque identiques au niveau du pays et de la division. Cela est appuyé par les composantes de la variance estimées de la composante aléatoire à tendance lisse au niveau de la division dans les deux modèles élaborés ($\hat{\sigma}_{R2}^{(div)}$: environ 0,020) présentés dans le tableau 6.1. En revanche, on constate des différences plus importantes dans les tendances des modèles MTS-II et MTS-III au niveau des districts, comme le montrent les figures 6.3 et 6.4. Voir Das, van den Brakel, Boonstra et Haslett (2021) qui présentent des graphiques pour tous les districts. Les tendances qui reposent sur le modèle MTS-III sont plus lisses que celles qui reposent sur le modèle MTS-II, ce qui est le résultat des plus petites valeurs de la composante de la variance estimée $\hat{\sigma}_{R2}$ dans MTS-III (voir le tableau 6.1).

Tableau 6.1
Moyennes *a posteriori* des paramètres d'écart-type des composantes aléatoires des modèles MTS-I, MTS-II, MTS-III pour ANC0. L'absence d'exposant fait référence au niveau du district, les exposants (*div*) font référence au niveau de la division

Modèle	$\hat{\sigma}_I$ (ET)	$\hat{\sigma}_S$ (ET)	$\hat{\rho}_{IS}$ (ET)	$\hat{\sigma}_{Sp}$ (ET)	$\hat{\sigma}_{R2}^{(div)}$ (ET)	$\hat{\sigma}_{R2}$ (ET)
MTS-I	0,083 (0,013)	0,054 (0,007)	0,168 (0,171)	0,068 (0,032)	0,019 (0,003)	0,024 (0,002)
MTS-II	0,069 (0,012)	0,033 (0,004)	0,254 (0,180)	0,071 (0,028)	0,020 (0,003)	0,013 (0,002)
MTS-III	0,062 (0,013)	0,027 (0,013)	0,227 (0,201)	0,067 (0,030)	0,020 (0,003)	0,009 (0,001)

Les tendances au niveau du district ont tendance à suivre la tendance du niveau de leur division respective illustrée à la figure 6.2. C'est particulièrement le cas pour les domaines ayant un nombre relativement faible d'observations, comme les districts de *Bandarban*, de *Khagrachnari* et de *Rangamati* à la figure 6.3, qui appartiennent à la division de *Chittagong* dans la figure 6.2. Pour réduire cette tendance, on a élaboré un modèle de séries chronologiques multiniveau en supprimant la composante de tendance lisse RW2_Division au niveau de la division dans le tableau 5.1. Cela a toutefois entraîné des tendances

irréalistes très lisses au niveau du pays et de la division. De même, pour examiner la nécessité d'une composante spatiale, on a élaboré des modèles de séries chronologiques multiniveaux en tenant compte et en ne tenant pas compte de la composante spatiale (Spatial_District dans le tableau 5.1). On constate que la composante spatiale rend les estimations plus plausibles pour les districts dont l'échantillon est petit ou nul. Observons par exemple les tendances des districts de *Bandarban* et de *Rangamati* de la division de *Chittagong*.

Le modèle MTS-I montre des tendances à la hausse pour certains districts pendant la période de 1994 à 2000. Ces évolutions ne sont pas probables d'un point de vue spécialisé et sont bien corrigées par les modèles MTS-II et MTS-III qui utilisent des estimations de F-H comme séries de données d'entrée. Voir par exemple les districts de *Noakhali*, de *Bandarban*, de *Rangamati*, de *Narayanganj*, de *Rajbari* et de *Narail* dans la figure 6.3. Certains districts présentent des tendances volatiles selon les estimations directes et le modèle MTS-I pendant toute la période, principalement en raison de la variation de la taille de l'échantillon. Voir par exemple les districts de *Bandarban*, de *Bhola*, de *Khagrachhari*, de *Kishoreganj* et de *Rangamati* à la figure 6.3, et de *Chapai Nababganj*, de *Feni*, de *Jhalokati*, de *Joypurhat* et de *Pabna* à la figure 6.4. D'un point de vue spécialisé, on attend une tendance lisse à la baisse pour la couverture de ANC0. Plus particulièrement, on ne s'attend pas aux points de retournement visibles dans plusieurs districts en 2007 et 2011. Les tendances qui reposent sur les modèles MTS-II et MTS-III ne tiennent pas compte de la plupart de ces volatilités et montrent des tendances lisses raisonnables pour ces districts; elles sont donc plus réalistes que celles de MTS-I. Néanmoins, les ajustements des trois modèles sont compatibles avec les données observées. MTS-II semble être un bon compromis entre les modèles I et III.

Figure 6.1 Tendances nationales d'ANC0 au Bangladesh : i) Estimations directes (ligne de barre d'erreur noire); ii) F-H transversale (ligne de barre d'erreur verte); iii) MTS-I (ligne rouge); iv) MTS-II (ligne verte); et v) MTS-III (ligne bleue).

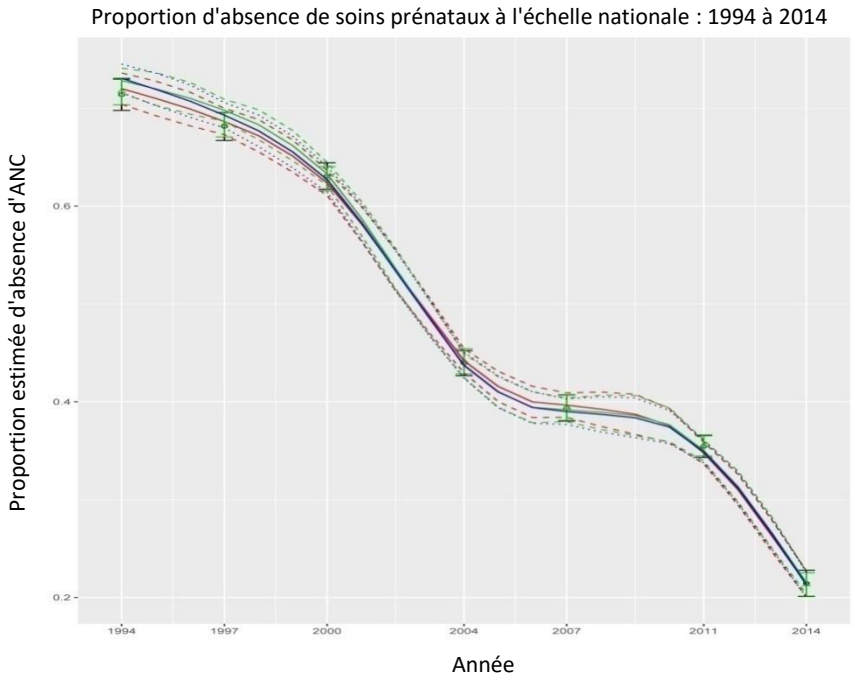


Figure 6.2 Tendances au niveau des divisions d'ANC0 au Bangladesh : i) Estimations directes (ligne de barre d'erreur noire); ii) F-H transversale (ligne de barre d'erreur verte); iii) MTS-I (ligne rouge); iv) MTS-II (ligne verte); v) MTS-III (ligne bleue).

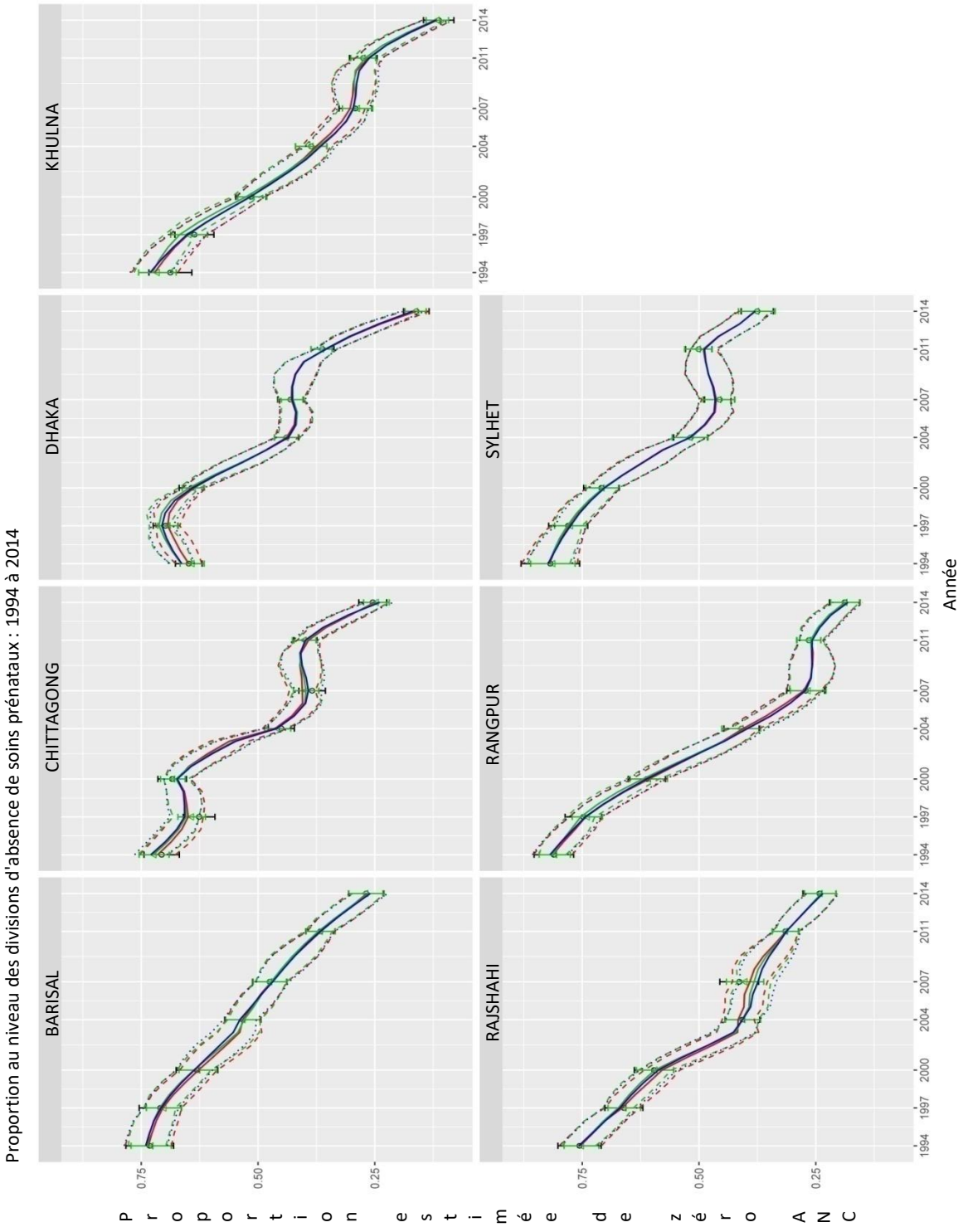


Figure 6.3 Tendances au niveau des districts d'ANC0 au Bangladesh : i) DIR (ligne de barre d'erreur noire); ii) F-H transversale (ligne de barre d'erreur verte); iii) MTS-I (ligne rouge); iv) MTS-II (ligne verte); v) MTS-III (ligne bleue).

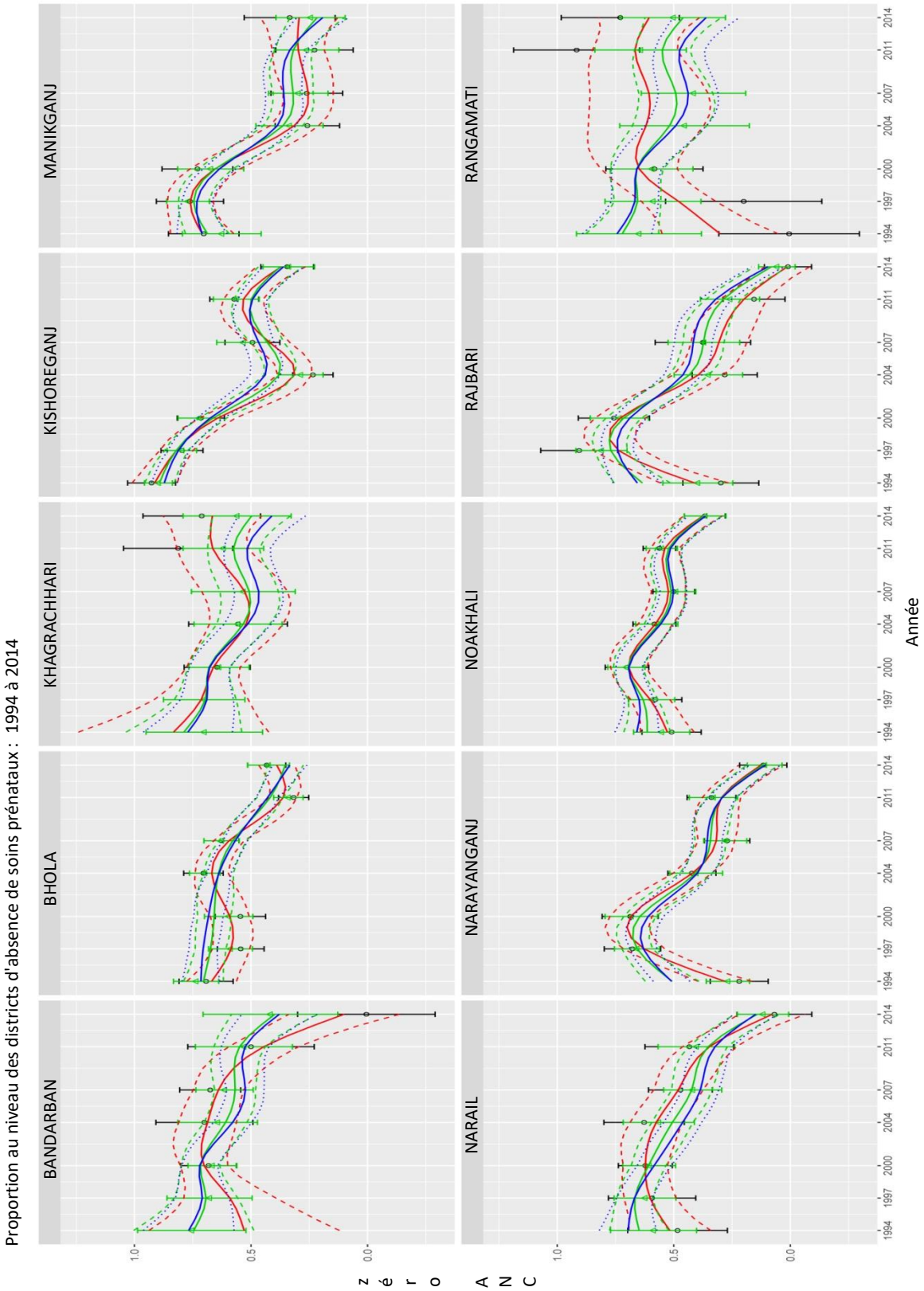
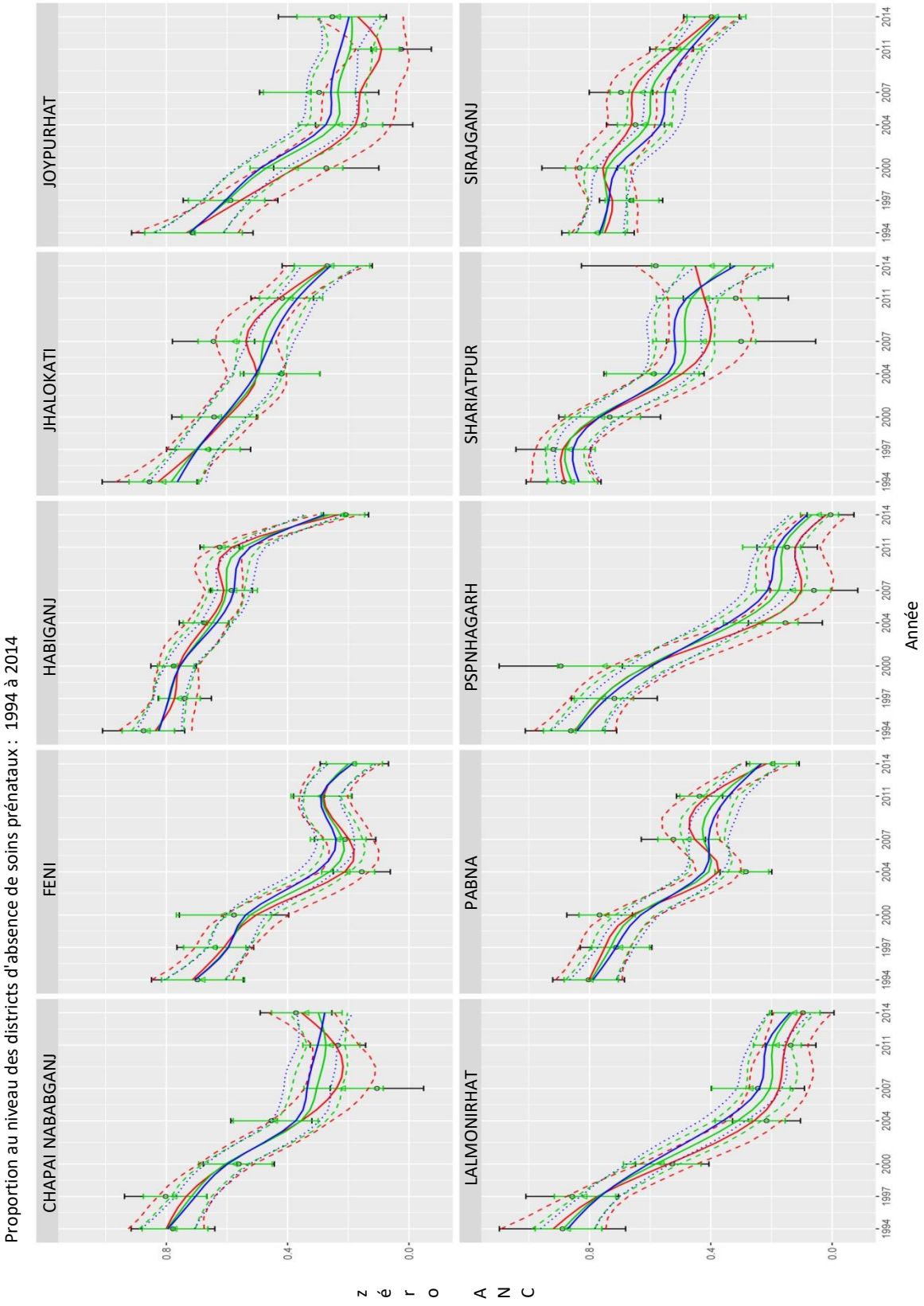


Figure 6.4 Tendances au niveau des districts d'ANC0 au Bangladesh : i) DIR (ligne de barre d'erreur noire); ii) F-H transversale (ligne de barre d'erreur verte); iii) MTS-I (ligne rouge); iv) MTS-II (ligne verte); v) MTS-III (ligne bleue).



Dans la plupart des cas, les modèles MTS-II et MTS-III se comportent de la même manière. Toutefois, le modèle MTS-III, qui prend en compte la corrélation dans les estimations de F-H transversales, surestime ANC0 pour certains districts (comme ceux de *Chapai Nababganj*, *Lalmonirhat* et *Shariatpur* dans la figure 6.4) et, de plus, sous-estime légèrement la tendance dans certains districts (comme ceux de *Khagrachari*, *Rangamati* et *Shirajganj* dans la figure 6.3) comparativement aux estimations de F-H transversales. Encore une fois, MTS-II cherche un compromis entre les tendances lisses de MTS-III et les tendances plus volatiles de MTS-I dans la plupart des districts et semble être le modèle privilégié aux fins d'estimation des tendances d'ANC0.

6.2 ANC4

La tendance nationale d'ANC4 illustrée à la figure 6.5 montre une augmentation linéaire de 6 % en 1994, à environ 31 % en 2014. Comme pour ANC0, les estimations directes et de F-H transversales d'ANC4 sont très semblables pour les années d'enquête, avec un IC d'environ 95 %. Les tendances estimées à partir de MTS-I (ligne rouge), MTS-II (ligne verte) et MTS-III (ligne bleue) présentent des tendances très similaires. Comparativement aux estimations directes et de F-H transversales, la tendance de MTS-I est légèrement inférieure en 2007 et 2014. Dans MTS-II et MTS-III, les tendances de l'année d'enquête 2011 sont un peu plus grandes que les estimations directes et de F-H transversales. Les tendances au niveau de la division sont présentées dans la figure 6.6. Les trois modèles de MTS donnent des estimations de tendance très similaires. Certaines différences sont constatées dans les divisions de *Chittagong*, de *Dhaka* et de *Rangpur*. MTS-I donne en effet une tendance légèrement plus élevée que les estimations directes et de F-H pour la division de *Rangpur* dans la période allant de 1994 à 2000. Pour MTS-II et MTS-III, la tendance est un peu plus élevée dans la division de *Rajshahi* dans la période allant de 2011 à 2014 comparativement aux estimations directes et de F-H. Les trois modèles de MTS montrent des bandes d'IC à 95 % légèrement en forme d'arc entre deux années d'enquête consécutives, ce qui indique une incertitude légèrement plus élevée pendant les années sans enquête comparativement aux années d'enquête.

Figure 6.5 Tendances nationales d'ANC4 au Bangladesh : i) Estimations directes (ligne de barre d'erreur noire); ii) F-H transversale (ligne de barre d'erreur verte); iii) MTS-I (ligne rouge); iv) MTS-II (ligne verte); v) MTS-III (ligne bleue).

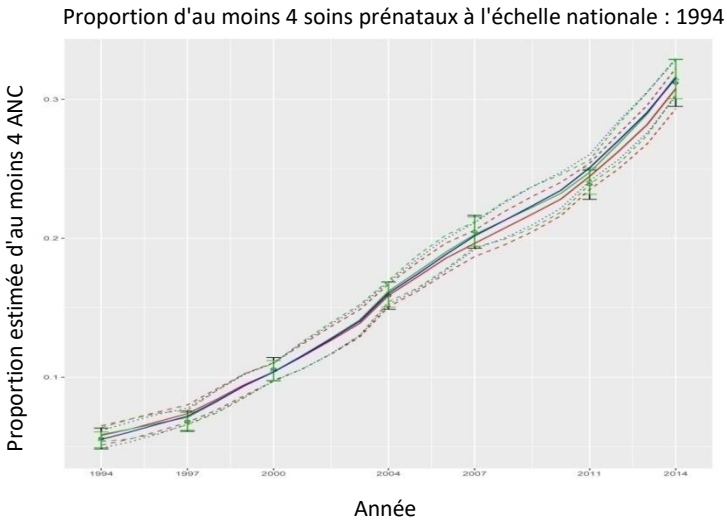
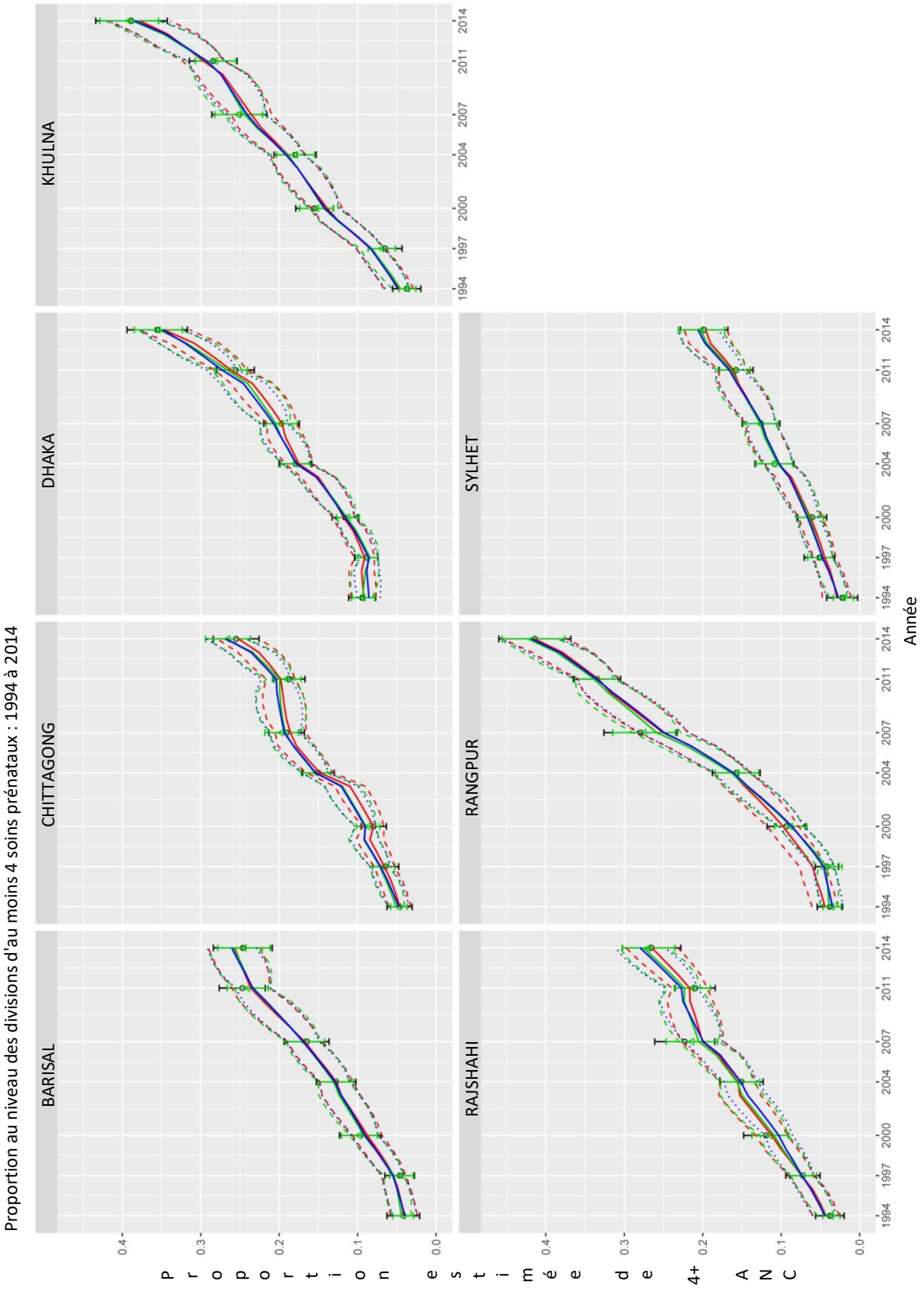


Figure 6.6 Tendances au niveau des divisions d'ANC4 au Bangladesh : i) Estimations directes (ligne de barre d'erreur noire); ii) F-H transversale (ligne de barre d'erreur verte); iii) MTS-I (ligne rouge); iv) MTS-II (ligne verte); v) MTS-III (ligne bleue).



Bien que les tendances qui reposent sur MTS-II et MTS-III soient presque identiques au niveau du pays et des divisions, les composantes de la variance estimée dans les deux modèles diffèrent considérablement comme le montre le tableau 6.2 ci-dessous. Ces différences entraînent des différences importantes dans les estimations des tendances au niveau des districts dans MTS-II et MTS-III. Les figures 6.7 et 6.8 présentent des graphiques pour certains districts. Das, van den Brakel, Boonstra et Haslett (2021) présentent des graphiques pour tous les districts. Comme pour ANC0, les tendances d'ANC4 dans MTS-III sont plus lisses que celles dans MTS-II. Les composantes de variance les plus petites de MTS-III entraînent aussi des bandes de confiance plus étroites que MTS-II.

Tableau 6.2
Moyennes *a posteriori* des paramètres d'écart-type des composantes aléatoires des modèles MTS-I, MTS-II, MTS-III pour ANC4. L'absence d'exposant fait référence au niveau du district, les exposants (*div*) font référence au niveau de la division

Modèle	$\hat{\sigma}_I$ (ET)	$\hat{\sigma}_S$ (ET)	$\hat{\rho}_{IS}$ (ET)	$\hat{\sigma}_{Sp}$ (ET)	$\hat{\sigma}_{R1}^{(div)}$ (ET)	$\hat{\sigma}_{R2}$ (ET)
MTS-I	0,060 (0,010)	0,033 (0,005)	0,428 (0,178)	0,047 (0,026)	0,012 (0,004)	0,009 (0,001)
MTS-II	0,046 (0,007)	0,022 (0,003)	0,501 (0,162)	0,035 (0,018)	0,016 (0,003)	0,004 (0,001)
MTS-III	0,038 (0,006)	0,018 (0,006)	0,522 (0,165)	0,027 (0,016)	0,014 (0,003)	0,002 (0,001)

Les estimations de tendance dans MTS-I sont volatiles et montrent des tendances à la baisse inattendues pour certains districts, voir par exemple les districts de *Bhola* et de *Pirojpur* de la division de Barisal, *Gazipur*, *Kishoreganj* et *Manikganj* de la division de *Dhaka*, *Bogra*, *Chapai Nababganj* et *Rajshahi* de la division de *Rajshahi*, et le district de *Habiganj* de la division de *Sylhet* dans la figure 6.7. D'un point de vue spécialisé, on n'attend pas de mouvements brusques ni de points de retournement dans la couverture d'ANC4. Il semblerait donc que MTS-I suive trop rigoureusement les estimations directes. Les tendances de MTS-II ignorent généralement ces volatilités et présentent des tendances raisonnablement stables pour ces districts. Dans MTS-III, les tendances sont encore plus lisses pour certains de ces districts, notamment les districts de *Bogra* et de *Habiganj* dans la figure 6.7, et les districts de *Mymensingh* et de *Sylhet* dans la figure 6.8.

La principale difficulté se pose pour les trois districts vallonnés de la division de *Chittagong*, à savoir *Khagrachhari*, *Rangamati* et *Lakshmipur* (les deux premiers districts sont représentés graphiquement dans la figure 6.8). MTS-I donne des estimations de tendance très médiocres pour ANC4 sur toute la période, principalement en raison d'estimations directes erratiques, qui sont soit nulles soit très incohérentes dans la plupart des enquêtes. Les estimations transversales de F-H sont plus robustes et, par conséquent, MTS-II et MTS-III présentent des tendances à la hausse raisonnables pour ANC4. On s'attend à ce que les femmes habitant dans des régions urbanisées et ayant de meilleures conditions socioéconomiques reçoivent plus de visites d'ANC que celles qui habitent dans des régions rurales aux conditions socioéconomiques défavorables. MTS-I donne, sur toute la période, des estimations inférieures dans certains districts et des estimations de tendance plus élevées que ce qui est attendu dans d'autres districts. Par exemple, la tendance obtenue avec MTS-I pour *Narsingdi*, district fortement urbanisé de la division de *Dhaka*, et présentée dans la figure 6.8 est plus basse que prévu. Ou encore, la tendance obtenue avec MTS-I pour *Munshiganj*, un district de *Dhaka* moins urbanisé, et présentée dans la figure 6.8 est plus élevée que ce qui est attendu. En outre, les estimations de tendance dans MTS-I sont plus élevées que ce

qui est attendu pour toute la période dans le district de *Meherpur* de la division de *Khulna* et les districts de *Lalmonirhat* et de *Panchagarh* de la division de *Rangpur*. Les estimations de tendances dans MTS-II et MTS-III semblent plus plausibles parce que les estimations transversales de F-H semblent plus réalistes que les estimations directes. Dans l'ensemble, comme pour ANC0, MTS-II est un bon compromis entre MTS-I et MTS-III.

Figure 6.7 Tendances au niveau des districts d'ANC4 au Bangladesh : i) Estimations directes (ligne de barre d'erreur noire); ii) F-H transversale (ligne de barre d'erreur verte); iii) MTS-I (ligne rouge); iv) MTS-II (ligne verte); v) MTS-III (ligne bleue).

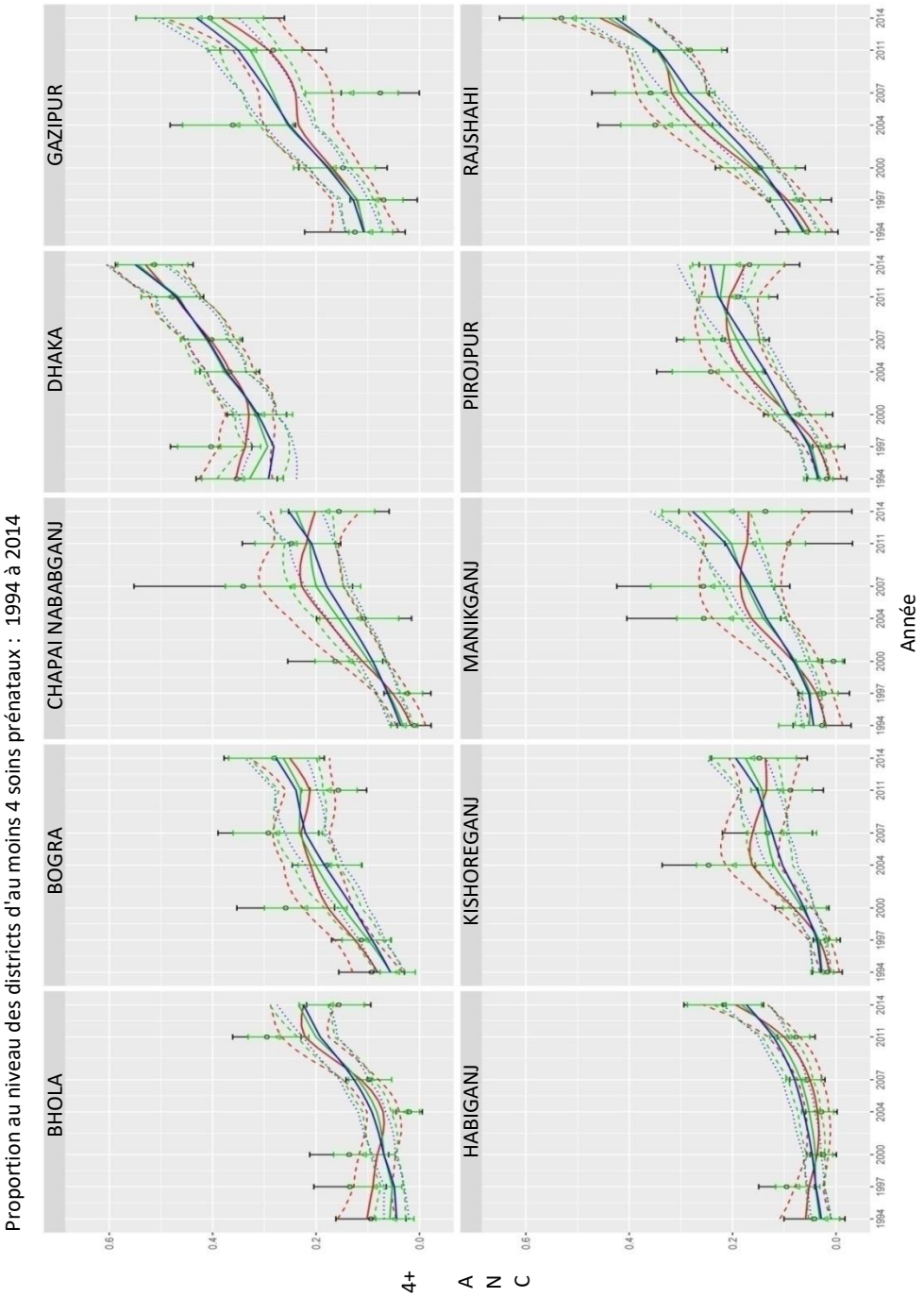
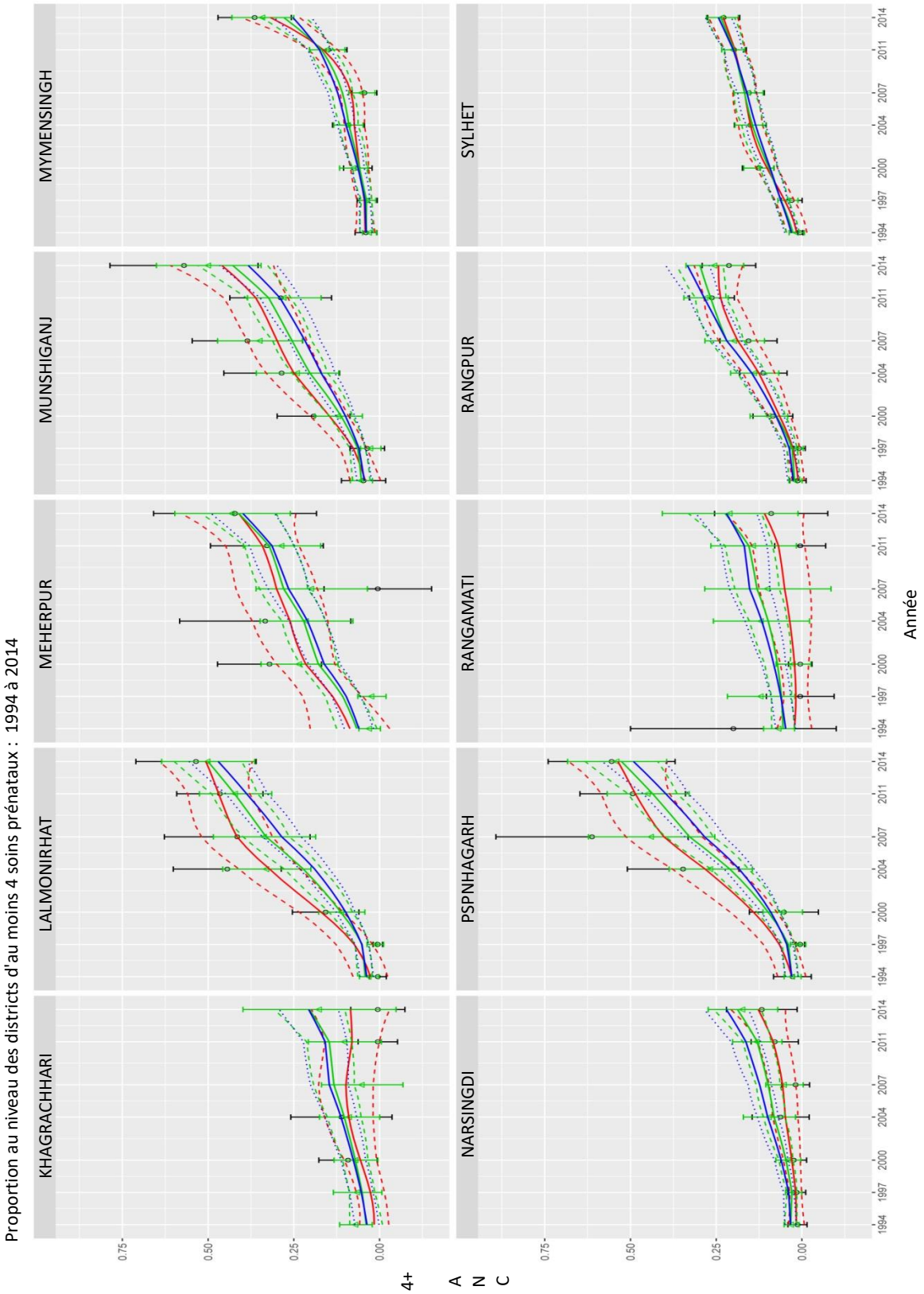


Figure 6.8 Tendances au niveau des districts d'ANC4 au Bangladesh : i) Estimations directes (ligne de barre d'erreur noire); ii) F-H transversale (ligne de barre d'erreur verte); iii) MTS-I (ligne rouge); iv) MTS-II (ligne verte); v) MTS-III (ligne bleue).



7. Évaluation des modèles

Dans la présente étude, les modèles ont été sélectionnés à partir du critère d'information Watanabe-Akaike (WAIC), du critère d'information de défiance (DIC) et de comparaisons graphiques de leurs prédictions de tendances à trois niveaux hiérarchiques. En plus de ces diagnostics de modèles, trois mesures de divergence sont définies pour évaluer et comparer les modèles de séries chronologiques multiniveaux. Les deux premières mesures sont le biais relatif (BR) et le biais relatif absolu (BRA), qui expriment les différences entre les estimations du modèle et les estimations directes, en pourcentage de ces dernières. Pour un modèle donné, BR_{it} et BRA_{it} pour le domaine i et l'année (d'enquête) t sont définis comme suit :

$$\mathrm{BR}_{it} = \frac{(\hat{\theta}_{it} - \hat{Y}_{it})}{\hat{Y}_{it}} \times 100\,\%, \tag{7.1}$$

$$\mathrm{BRA}_{it} = \frac{\left|\hat{\theta}_{it} - \hat{Y}_{it}\right|}{\hat{Y}_{it}} \times 100\,\% \tag{7.2}$$

$\hat{\theta}_{it}$ étant la prédiction du modèle et $\hat{Y}_{it}$ l'estimation directe. La troisième mesure de divergence est la réduction relative des erreurs-types (RRET), qui mesure le pourcentage de réduction de l'erreur-type des estimations fondées sur un modèle par rapport aux estimations directes, à savoir

$$\mathrm{RRET}_{it} = 100\,\% \times \left(\mathrm{se}(\hat{Y}_{it}) - \mathrm{se}(\hat{\theta}_{it})\right) \Big/ \mathrm{se}(\hat{Y}_{it}). \tag{7.3}$$

La mesure de la RRET ne doit pas être interprétée de façon trop stricte, étant donné que les erreurs-types fondées sur le plan et sur le modèle sont des quantités conceptuellement différentes. Il reste que toutes deux sont couramment utilisées comme mesures de l'incertitude, et une fois que des modèles raisonnables qui tiennent suffisamment compte des variations à tous les niveaux d'intérêt ont été sélectionnés, à partir d'autres critères, l'utilisation de la RRET comme une des mesures de comparaison est intéressante.

Ces trois mesures de divergence sont calculées au niveau du pays, de la division et du district (ce dernier étant le plus détaillé). Les distributions de ces mesures sont présentées selon la valeur minimale, le 1er quartile (Q_1), la médiane, la moyenne, le 3e quartile (Q_3) et la valeur maximale.

De plus, le taux de couverture observé (TC exprimé en %) pour l'intervalle de confiance de 95 % des modèles de MTS et de F-H transversaux considérés est calculé au niveau des divisions et des districts en déterminant si l'IC de 95 % estimé de $\hat{\theta}_{it}$ contient les estimations directes $(\hat{Y}_{it})$. La couverture au niveau du district est le pourcentage de combinaisons de district par année (environ 7×64 domaines) où l'estimation directe est incluse dans l'IC de $\hat{\theta}_{it}$. La couverture au niveau de la division est le pourcentage de combinaisons de division par année (environ 7×7 domaines) où l'estimation directe est incluse dans l'IC de $\hat{\theta}_{it}$. On définit les taux de couverture de la même façon pour chaque année d'enquête en établissant la moyenne sur tous les districts disponibles pour une année d'enquête donnée. Enfin, on calcule la couverture séparément pour chaque division en établissant la moyenne sur les sept années d'enquête.

Les distributions de BR_{it} (7.1), BRA_{it} (7.2) et $RRET_{it}$ (7.3) pour trois niveaux administratifs sont fournies dans les tableaux 7.1, 7.2 et 7.3 pour ce qui est d'ANC0 et ANC4 pour les modèles de F-H transversaux, MTS-I, MTS-II et MTS-III. Le tableau 7.1 montre que les modèles de F-H et de MTS-I fournissent un BR moyen inférieur pour ANC0 et ANC4 aux trois niveaux, tandis que MTS-II donne un BR moyen légèrement inférieur par rapport au modèle de MTS-III au niveau du district. Les distributions du BRA du tableau 7.2 montrent que les performances de MTS-II se situent entre celles de MTS-I et de MTS-III pour tous les niveaux administratifs, sauf à l'échelle nationale pour ANC4. Les valeurs du BRA les plus faibles s'observent dans le modèle de F-H transversal. On constate aussi que le BRA augmente à mesure que la taille de l'échantillon du domaine diminue. Le tableau 7.3 montre que MTS-II a les valeurs de RRET les plus élevées au niveau du pays et de la division, tandis qu'au niveau du district, ce modèle donne une RRET légèrement inférieure au modèle de MTS-III pour ANC0 et ANC4. La réduction de la variance augmente quand la taille des échantillons de domaine diminue. Les erreurs-types pour les tendances au niveau du pays et de la division sont plus petites dans MTS-II que dans MTS-III parce que dans MTS-II, les covariances entre les prédictions transversales de F-H au niveau du district dans les séries de données d'entrée sont ignorées. Ces covariances sont majoritairement positives et, par conséquent, les erreurs-types des tendances aux niveaux agrégés sont plus élevées et plus réalistes dans MTS-III. Les valeurs plus élevées de BR, de BRA et de RRET pour les modèles MTS-II et MTS-III sont une conséquence des tendances plus lisses obtenues au moyen des deux modèles. Les petites variances en présence de tendances lisses signifient un plus grand nombre de biais comparativement aux estimations directes. Comme nous l'avons vu à la section 6, ces tendances sont plus plausibles comparativement au modèle transversal de F-H et au modèle de MTS-I, puisque d'un point de vue spécialisé, on attend une diminution lisse pour ANC0 et une augmentation lisse pour ANC4.

Tableau 7.1
Statistiques sommaires du biais relatif (BR, en %) à différents niveaux d'agrégation pour les estimations sur petits domaines d'ANC0 et d'ANC4

Paramètre	Niveau d'agrégation	Modèle	Min.	Q_1	Médiane	Moyenne	Q_3	Max.
ANC0	Pays	F-H	-0,48	-0,20	0,23	0,08	0,37	0,47
		MTS-I	-1,84	-0,68	0,54	-0,05	0,73	0,88
		MTS-II	-1,16	-0,55	0,29	0,41	1,19	2,43
		MTS-III	-1,53	-0,80	-0,57	-0,02	0,60	2,35
	Division	F-H	-0,68	-0,48	-0,36	0,05	0,50	1,31
		MTS-I	-0,99	-0,50	-0,31	0,05	0,64	1,41
		MTS-II	-0,77	0,04	0,15	0,59	1,08	2,50
		MTS-III	-1,44	-0,37	0,13	0,15	0,89	1,35
	District	F-H	-8,77	-1,72	0,14	0,31	1,67	12,41
		MTS-I	-10,35	-1,24	-0,49	-0,66	0,30	1,87
		MTS-II	-7,87	-1,15	0,77	1,25	2,89	18,34
		MTS-III	-10,05	-2,63	0,89	1,34	3,91	21,43

Tableau 7.1(suite)
Statistiques sommaires du biais relatif (BR, en %) à différents niveaux d'agrégation pour les estimations sur petits domaines d'ANC0 et d'ANC4

Paramètre	Niveau d'agrégation	Modèle	Min.	Q_1	Médiane	Moyenne	Q_3	Max.
ANC4	Pays	F-H	-1,65	-0,62	0,07	-0,07	0,65	1,04
		MTS-I	-4,09	-1,60	0,05	1,00	3,19	7,88
		MTS-II	-1,85	0,27	1,98	1,91	3,80	5,07
		MTS-III	-2,00	-1,35	1,06	1,11	3,10	5,23
	Division	F-H	-1,33	-0,60	-0,13	0,24	0,43	3,47
		MTS-I	-1,17	-0,25	-0,04	-0,07	0,32	0,59
		MTS-II	-0,50	0,68	1,18	1,55	1,70	5,39
		MTS-III	-2,08	0,31	0,73	1,24	1,92	5,58
	District	F-H	-17,83	-4,85	0,40	2,08	6,78	64,77
		MTS-I	-16,32	-3,80	-0,56	-0,42	2,98	15,57
		MTS-II	-22,00	-5,30	0,57	4,57	12,47	84,31
		MTS-III	-29,92	-8,23	0,57	6,12	14,07	124,63

Cette conclusion est confirmée par les valeurs de TC indiquées dans le tableau 7.4. Les TC pour les modèles de F-H transversaux sont trop élevés, ce qui indique que les prédictions de F-H tendent trop vers les estimations directes. Les niveaux de TC sont raisonnablement bons pour MTS-I, considérablement plus bas pour MTS-II et les plus bas pour MTS-III. Les taux de couverture plus faibles de MTS-II et de MTS-III au niveau du district se traduisent par les valeurs correspondantes plus élevées du BRA et de la RRET. Ces résultats montrent que les prédictions du modèle de MTS-I sont plus volatiles et tendent vers les estimations directes, que les prédictions du modèle MTS-III sont fortement lissées et que celles du modèle MTS-II semblent constituer un compromis raisonnable entre les prédictions du modèle de MTS-I et du modèle de MTS-III, en particulier au niveau des districts.

Tableau 7.2
Statistiques sommaires du biais relatif absolu (BRA, en %) à différents niveaux d'agrégation pour les estimations sur petits domaines d'ANC0 et d'ANC4

Paramètre	Niveau d'agrégation	Modèle	Min.	Q_1	Médiane	Moyenne	Q_3	Max.
ANC0	Pays	F-H	0,04	0,27	0,42	0,34	0,46	0,48
		MTS-I	0,26	0,63	0,75	0,87	0,99	1,84
		MTS-II	0,29	0,44	0,58	1,05	1,59	2,43
		MTS-III	0,49	0,61	0,96	1,18	1,61	2,35
	Division	F-H	0,39	0,50	0,65	0,90	1,20	1,84
		MTS-I	0,48	0,66	0,78	1,39	2,13	2,90
		MTS-II	0,79	0,96	1,56	1,78	2,14	3,88
		MTS-III	1,00	1,14	1,41	1,88	2,43	3,61
	District	F-H	1,08	2,73	4,17	5,12	5,84	15,94
		MTS-I	1,48	3,93	6,58	7,53	9,02	26,67
		MTS-II	3,15	6,46	10,31	11,32	14,50	33,01
		MTS-III	4,15	8,65	12,54	13,49	16,98	38,16

Tableau 7.2(suite)
Statistiques sommaires du biais relatif absolu (BRA, en %) à différents niveaux d'agrégation pour les estimations sur petits domaines d'ANC0 et d'ANC4

Paramètre	Niveau d'agrégation	Modèle	Min.	Q_1	Médiane	Moyenne	Q_3	Max.
ANC4	Pays	F-H	0,07	0,25	0,92	0,76	1,08	1,65
		MTS-I	0,05	1,60	2,47	3,09	4,00	7,88
		MTS-II	0,97	1,68	1,98	2,71	3,80	5,07
		MTS-III	1,06	1,19	1,46	2,46	3,53	5,23
	Division	F-H	0,98	1,40	1,71	1,87	2,06	3,47
		MTS-I	1,96	3,06	4,31	4,07	4,64	6,82
		MTS-II	2,18	3,66	4,68	4,33	5,07	6,00
		MTS-III	3,66	4,60	5,36	5,27	5,63	7,46
	District	F-H	1,93	7,64	12,91	14,29	17,60	64,77
		MTS-I	3,86	14,27	18,72	20,61	28,10	53,45
		MTS-II	7,07	19,47	26,22	28,51	35,88	84,31
		MTS-III	8,62	21,36	29,32	33,13	41,00	124,63

Tableau 7.3
Statistiques sommaires de la réduction relative des erreurs-types (RRET en %) à différents niveaux d'agrégation pour les estimations sur petits domaines d'ANC0 et d'ANC4

Paramètre	Niveau d'agrégation	Modèle	Min.	Q_1	Médiane	Moyenne	Q_3	Max.
ANC0	Pays	F-H	-0,65	4,03	8,10	8,00	12,72	15,01
		MTS-I	-0,03	1,35	4,01	3,67	5,82	7,33
		MTS-II	4,07	7,90	13,71	12,89	17,47	21,68
		MTS-III	-3,52	1,02	3,30	3,69	7,15	9,67
	Division	F-H	2,99	5,82	7,56	7,03	8,64	9,75
		MTS-I	2,66	4,00	5,32	5,16	6,65	6,84
		MTS-II	8,47	12,74	13,70	13,12	14,34	15,53
		MTS-III	3,30	4,71	5,21	5,47	6,12	8,16
	District	F-H	-1,60	7,17	10,20	9,98	12,04	21,61
		MTS-I	7,91	15,16	17,81	18,06	21,15	27,47
		MTS-II	12,60	27,84	34,08	33,80	38,46	48,53
		MTS-III	19,48	32,61	38,40	37,79	41,55	52,71
ANC4	Pays	F-H	8,58	11,22	11,66	13,71	14,49	24,32
		MTS-I	6,64	12,16	14,60	14,75	18,50	20,66
		MTS-II	17,79	22,87	23,56	25,12	27,99	32,75
		MTS-III	10,33	16,58	19,45	18,15	21,04	22,04
	Division	F-H	11,08	11,80	14,07	14,23	16,39	18,08
		MTS-I	11,82	14,31	14,46	15,78	18,18	19,17
		MTS-II	20,32	24,96	27,39	26,34	28,15	30,45
		MTS-III	15,49	20,37	21,75	21,72	24,51	25,05
	District	F-H	0,34	11,62	16,77	17,63	22,60	38,62
		MTS-I	17,79	27,84	30,48	30,93	33,65	43,40
		MTS-II	29,58	43,37	46,86	48,10	54,96	66,75
		MTS-III	35,63	48,88	51,75	52,94	59,31	70,35

Tableau 7.4
Taux de couverture observé (TC en %) des prédictions du modèle pour l'intervalle de confiance de 95 % au niveau du district et de la division, ainsi qu'au niveau du district par année d'enquête pour les estimations sur petits domaines d'ANC0 et d'ANC4

Paramètre	Modèle	TC selon l'année au niveau du district							TC global par niveau	
		1994	1997	2000	2004	2007	2011	2014	District	Division
ANC0	F-H	100,00	98,33	100,00	100,00	100,00	98,36	100,00	99,53	100,00
	MTS-I	100,00	90,00	93,44	88,52	93,22	98,36	100,00	94,81	100,00
	MTS-II	88,33	63,33	70,49	67,21	71,19	75,41	91,53	75,10	95,92
	MTS-III	83,33	53,33	50,82	52,46	61,02	55,74	79,66	62,22	95,92
ANC4	F-H	98,15	98,28	100,00	100,00	100,00	100,00	90,20	98,36	100,00
	MTS-I	87,04	84,48	68,33	76,27	81,97	96,72	100,00	84,58	95,92
	MTS-II	44,44	51,72	50,00	52,54	62,30	65,57	76,47	57,55	97,96
	MTS-III	44,44	41,38	40,00	38,98	50,82	55,74	72,55	48,70	97,96

8. Discussion

Dans la présente étude, nous avons élaboré des modèles de séries chronologiques multiniveaux (MTS) pour obtenir le pourcentage de femmes ne recevant pas de consultation de soins prénataux (ANC0) et le pourcentage de femmes recevant au moins quatre consultations de soins prénataux (ANC4) au Bangladesh, en nous appuyant uniquement sur sept éditions de la *Bangladesh Demographic and Health Survey* (BDHS, Enquête démographique et sur la santé du Bangladesh) couvrant la période de 1994 à 2014. Les modèles de séries chronologiques sont définis à une fréquence annuelle pour laquelle les années sans édition d'enquête sont traitées comme manquantes. Ainsi, le modèle tient compte des intervalles entre éditions de la BDHS et produit des prédictions pour les années sans enquête-échantillon. Les tendances sont produites à trois échelons régionaux, à savoir l'échelle nationale, qui se décompose en 7 divisions, elles-mêmes décomposées en 64 districts.

Dans le premier modèle (MTS-I), des estimations directes selon le domaine et l'année et leurs erreurs-types servent de données d'entrée du modèle de séries chronologiques multiniveau. Les tendances obtenues au moyen de ce modèle sont plutôt volatiles puisque les estimations de tendances tendent à suivre les estimations directes. Un autre inconvénient du modèle MTS-I est qu'il entrave l'utilisation de l'information auxiliaire disponible dans les deux recensements, car les valeurs de l'information auxiliaire d'un recensement donné ne changent pas dans deux ou trois éditions consécutives de l'enquête. Pour utiliser ces données de recensement, on propose d'élaborer des modèles transversaux de Fay-Herriot (F-H) séparément pour chaque année d'enquête. Dans un deuxième modèle, dit MTS-II, ces estimations de F-H et leurs erreurs-types servent de séries de données d'entrée. Dans un troisième modèle, dit MTS-III, les estimations de F-H avec leurs matrices de covariance complètes, sont utilisées comme séries de données d'entrée. Ce modèle de MTS tient correctement compte des corrélations transversales entre les estimations de F-H. Le modèle global pour MTS-II et MTS-III est alors un processus non itératif en deux étapes, dans lequel la première étape consiste à produire les estimations de F-H.

Les modèles sont élaborés à l'échelle régionale la plus détaillée, celui des districts. Les tendances au niveau de la division et à l'échelle nationale sont estimées par agrégation des prédictions de tendances au niveau du district. Ainsi, les chiffres à différents niveaux d'agrégation sont numériquement cohérents par définition.

Comparativement à d'autres modèles d'estimation sur petits domaines de séries chronologiques proposés dans la littérature, nos modèles contiennent plus de structure, puisque les modèles de tendances dynamiques sont précisés à différents niveaux d'agrégation. Cela est nécessaire pour obtenir des prédictions agrégées exactes au niveau des divisions et du pays et constitue un moyen plus parcimonieux de modéliser les corrélations transversales. On a envisagé une régularisation supplémentaire du modèle par la spécification des distributions *a priori* globales et locales. Cela n'a cependant pas permis d'améliorer les ajustements du modèle.

Dans l'estimation sur petits domaines, les estimations pour un domaine sont souvent étalonnées sur les estimations directes à l'échelle nationale pour assurer la cohérence numérique et pour tenter de réduire le biais dans les prédictions par domaine fondées sur le modèle. Dans cette application, les estimations de tendances à l'échelle nationale selon les modèles MTS sont déjà très proches des estimations directes. C'est pourquoi nous n'avons pas examiné d'étape supplémentaire d'étalonnage.

Les trois modèles de séries chronologiques fournissent des estimations plus exactes. Étant donné que MTS-II ignore les corrélations majoritairement positives entre les séries de données d'entrée transversales de F-H, les erreurs-types des tendances aux niveaux agrégés sont de fait trop petites. Puisque MTS-III tient compte de ces corrélations, les erreurs-types pour les tendances au niveau du pays et des divisions sont plus grandes, mais aussi plus réalistes. Le modèle MTS-II semble toutefois fournir les tendances les plus plausibles pour les deux variables réponses, particulièrement au niveau des districts, en trouvant un compromis entre la volatilité des tendances dans le modèle MTS-I et le caractère plat des tendances dans le modèle MTS-III. Ce choix est appuyé par le fait que ces variables sont susceptibles d'être relativement lisses dans le temps. Dans une perspective conceptuelle, l'ajustement de ces modèles aux séries d'ANC0 et d'ANC4 est donc certainement approprié. Cela justifie aussi l'interpolation des tendances pour les années sans enquête-échantillon. De plus, notre méthode peut être utile à de nombreux pays en développement qui réalisent des enquêtes démographiques et sur la santé répétées, car elles sont généralement réalisées à des intervalles variables et dépendent principalement de données de recensement non mises à jour dans les deux ou trois éditions suivantes de l'enquête.

On propose comme solution pratique d'utiliser les prédictions des modèles transversaux de F-H comme séries de données d'entrée dans les modèles de séries chronologiques multiniveaux afin d'exploiter au mieux les données de recensement disponibles. Cette méthode présente un autre avantage : elle stabilise les séries de données d'entrée dans les modèles MTS en éliminant les grandes erreurs d'échantillonnage des estimations directes. Cela exige toutefois un processus de sélection et d'évaluation minutieux des modèles transversaux de F-H, car une erreur de spécification des modèles transversaux de F-H peut donner lieu à des séries de données d'entrée biaisées ayant des erreurs-types estimées qui sous-estiment l'incertitude réelle.

Une des limites de la présente étude est liée à la correction du biais pour la transformation par la racine carrée qui est appliquée à ANC4. La correction du biais ne peut être appliquée qu'aux estimations de tendances pour les années d'enquête. Cela entraîne d'étranges augmentations des estimations ponctuelles si l'erreur d'échantillonnage n'est pas suffisamment lissée, en particulier pour les domaines à petite taille d'échantillon. Cela entrave l'estimation des changements d'une période à l'autre entre les années d'enquête et les autres années. C'est pourquoi la correction de biais n'est appliquée qu'aux modèles de F-H transversaux et non aux modèles MTS.

Étant donné que la prévalence des visites de soins prénataux ANC0 et ANC4 est corrélée négativement, un modèle multivarié peut constituer une solution de rechange intéressante aux modèles univariés utilisés dans la présente étude. Les deux séries pourraient être combinées à la série de la catégorie restante dans un seul modèle multivarié. Toutefois, cela nécessite un modèle multinomial qui a l'avantage d'améliorer la précision des estimations et de garantir que les prédictions prennent des valeurs dans leur fourchette admissible, et que les prédictions dans les catégories s'additionnent pour correspondre à 100 %. La mise en œuvre du modèle multinomial n'est cependant pas facile. Dans la présente étude en particulier, la matrice de variance-covariance peut être difficile à estimer pour les districts ayant un petit nombre d'observations. En outre, Datta et coll. (2002) montrent que les modèles univariés peuvent fournir des résultats aussi bons que les modèles multivariés proposés dans Ghosh, Nangia et Kim (1996), tout en étant plus simples à mettre en œuvre. Nous avons préféré ne pas examiner l'extension de nos modèles univariés à un modèle multinomial, qui pourrait faire l'objet d'une autre recherche.

Pour ANC0, l'échelle nationale montre une tendance à la baisse. Cette baisse s'est arrêtée temporairement pour la période de 2004 à 2011. La tendance d'ANC4 indique une augmentation constante pendant la période examinée dans la présente étude. Les tendances au niveau des divisions pour ANC0 montrent une baisse constante dans toutes les divisions, sauf *Dhaka*, *Chittagong* et *Sylhet*. Les tendances de ces trois divisions sont restées stables pendant la période allant de 2004 à 2011, ce qui est la principale cause de la tendance stable à l'échelle nationale d'ANC0. Par ailleurs, au niveau de la division, ANC4 montre des tendances à la hausse presque linéaires pour la plupart des divisions, sauf *Dhaka* et *Chittagong*. La plus forte amélioration est observée dans les divisions de *Khulna* et de *Rangpur*, où les tendances d'ANC4 atteignent plus de 40 % en 2014. Les tendances au niveau des districts permettent de repérer les districts très vulnérables en ce qui concerne les deux variables réponses étudiées. Bien que la tendance nationale d'ANC0 diminue pour s'établir à environ 21 % en 2014, quelques districts enregistrent un pourcentage inférieur à 10 % (*Dhaka*, *Jhenaidaha* et *Meherpur*), tandis qu'un nombre considérable de districts ont encore une valeur d'ANC0 supérieure à 35 % (*Bhola*, *Cox's Bazar*, *Kishoregonj*, *Noakhali*, *Sunamganj*, *Sirajgonj* et trois districts des *Chittagong Hill Tracts*). Pour ANC4, certains districts ont des estimations supérieures à 50 % (*Dhaka*, *Nilphamari* et *Panchagarh*) et la plupart des districts ayant une ANC0 élevée ont des estimations d'ANC4 inférieures à 20 %. Ces tendances au niveau des districts pourraient aider les décideurs à se concentrer sur les lieux les plus vulnérables où les indicateurs ANC0 et ANC4 restent faibles. De toute évidence, des tendances déterminées à des niveaux détaillés pourraient aider les décideurs à prendre des mesures pour réduire les inégalités de niveaux désagrégés dans la course à la réalisation des objectifs de développement durable.

Remerciements

Nous tenons à remercier *Measure Evaluation* et le *National Institute of Population Research and Training* (NIPORT) qui ont rendu les données de la BDHS accessibles au public. Nos remerciements vont également à IPUMS qui a le mérite d'avoir donné accès aux données-échantillons des recensements du Bangladesh de 1991, de 2001 et de 2011. Les opinions exprimées dans le présent article sont celles des auteurs et ne reflètent pas nécessairement la politique du Bureau central de la statistique des Pays-Bas. Les auteurs remercient les deux réviseurs anonymes et le rédacteur adjoint pour leurs précieux commentaires sur une version précédente du présent article.

Annexe

Tableau A.1
Variables contextuelles au niveau du district générées à partir des données du Recensement de 1991, du Recensement de 2001 et du Recensement de 2011 pour ANC0

Variable	Définition
Division	Barishal, Chittagong, Dhaka, Khulna, Rajshahi, Rangpur, Sylhet.
Région	(1) Districts densément peuplés de Dhaka, Chittagong et Gazipur, (2) 9 districts régionaux avec de grandes villes, (3) 3 districts vallonnés (Bandarban, Khagrachhari et Rangamati), (4) 49 autres districts (régions moins urbanisées).
Chittagong	Division de Chittagong ?
Dhaka	Division de Dhaka ?
Khulna	Division de Khulna ?
Rangpur	Division de Rangpur ?
Rajshahi	Division de Rajshahi ?
P_U5	Proportion d'enfants de moins de 5 ans.
P_W	Proportion de femmes âgées de 15 à 49 ans.
P_MW	Proportion de femmes mariées âgées de 15 à 49 ans.
P_MW_Prim_Edu	Proportion de femmes mariées âgées de 15 à 49 ans qui ont suivi un enseignement primaire.
P_MW_Sec_Edu	Proportion de femmes mariées âgées de 15 à 49 ans qui ont suivi au moins un enseignement secondaire.
P_HH_No_Edu_W	Proportion de ménages comptant des femmes analphabètes âgées de 15 à 49 ans.
P_HH_Prim_Edu_W	Proportion de ménages comptant des femmes ayant suivi un enseignement primaire, âgées de 15 à 49 ans.
P_HH_High_Edu_W	Proportion de ménages comptant des femmes ayant suivi un enseignement supérieur, âgées de 15 à 49 ans.
P_HH_Sec_Edu_Head	Proportion de ménages ayant un chef de ménage ayant suivi au moins un enseignement secondaire.
P_Ru_HH_4^{+}	Proportion de ménages ruraux de 4 personnes et plus.
P_Ru_HH_Elec	Proportion de ménages ruraux ayant l'électricité.

Tableau A.1(suite)
Variables contextuelles au niveau du district générées à partir des données du Recensement de 1991, du Recensement de 2001 et du Recensement de 2011 pour ANC0

Variable	Définition
P_Ru_HH_Sing_Moth	Proportion de ménages ruraux comptant une mère célibataire.
P_HH_U5_Sec_Edu_W	Proportion de ménages comptant des enfants âgés de moins de 5 ans et des femmes âgées de moins de 49 ans ayant suivi au moins un enseignement secondaire.
P_HH_2^+_U5	Proportion de ménages ayant au moins deux enfants âgés de moins de cinq ans.
P_Ru_HH_U5	Proportion de ménages ruraux ayant des enfants de moins de cinq ans.
P_Ru_HH_2^+_U5	Proportion de ménages ruraux ayant au moins deux enfants âgés de moins de cinq ans.
P_Ur_HH_2^+_U5	Proportion de ménages urbains ayant au moins deux enfants âgés de moins de cinq ans.

Tableau A.2
Effets fixes et aléatoires des modèles de F-H spécifiques à une année d'enquête pour ANC0

Année d'enquête	Transformation	Effets fixes	Effets aléatoires	Données de recensement
1994	Non	1 + Division + P_HH_High_Edu_W + P_Ur_HH_2^+_U5	OOA : Niveau du district Ordonnée à l'origine aléatoire	1991
1997	Non	1 + Division + P_MW_Sec_Edu + P_HH_Sec_Edu_Head	OOA	1991
2000	Non	1 + Division + P_Ru_HH_U5 + P_W	OOA	1991
2004	Non	1 + Division + P_HH_U5_Sec_Edu_W + P_MW	OOA	2001
2007	Non	1 + Division + P_U5 + P_Ru_HH_Size_4^+	OOA	2001
2011	TPRC	1 + Division + sqrt (P_Ru_HH_U5) + sqrt (P_MW)	OOA	2011
2014	TPRC	1 + Division + sqrt (P_Ur_HH_2^+_U5) + sqrt (P_Ru_HH_Elec)	OOA	2011

Tableau A.3
Effets fixes et aléatoires des modèles de F-H spécifiques à une année d'enquête pour ANC4

Année d'enquête	Transformation	Effets fixes	Effets aléatoires	Données de recensement
1994	TPRC	1 + Division + sqrt (P_Ru_U5) + sqrt (P_HH_2^+_U5) + sqrt (P_Ur_HH_2^+_U5)	OOA : Niveau du district Ordonnée à l'origine aléatoire	1991
1997	TPRC	1 + Division + sqrt (P_HH_U5_Sec_Edu_W) + log (P_HH_Sec_Edu_Head)	OOA	1991
2000	TPRC	1 + Khulna + Region + sqrt (P_MW$_P$rim$_E$du) + sqrt (P_HH_W$_I$lli$_E$du)	OOA	1991
2004	TPRC	1 + Division + sqrt (P_HH_U5_Prim_Edu_W) + sqrt (P_W)	OOA	2001
2007	TPRC	1 + Rangpur + Region + sqrt (P_U5) + sqrt (P_Ru_HH_Size_4^+)	OOA	2001
2011	TPRC	1 + Rangpur + Chittagong + sqrt ((P_HH_U5_Sec_Edu_W) + sqrt (P_W)	OOA	2011
2014	TPRC	1 + Rangpur + Chittagong + Rajshahi + Region + sqrt (P_W) + sqrt (P_Ru_HH_Sing_Mot)	OOA	2011

Bibliographie

BBS (2020). *SDG Tracker: Bangladesh Development Mirror*. https://www.sdg.gov.bd/page/thirty_nine_plus_one_indicator/5#1, [En ligne; consulté le 18 décembre 2020].

BBS et UNICEF (2019). *Progotir Pathey Bangladesh: Bangladesh Multiple Indicator Cluster Survey 2019*. Rapport technique, Bangladesh Bureau of Statistics (BBS).

Besag, J., et Kooperberg, C. (1995). On conditional and intrinsic autoregression. *Biometrika*, 82(4), 733-746.

Bollineni-Balabay, O., van den Brakel, J., Palm, F. et Boonstra, H.J. (2017). Multilevel hierarchical Bayesian versus state space approach in time series small area estimation: The Dutch Travel Survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(4), 1281-1308.

Boonstra, H.J. (2021). *mcmcsae: Markov Chain Monte Carlo Small Area Estimation*. R package version 0.6.0.

Boonstra, H.J., et van den Brakel, J. (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/fr/pub/12-001-x/2019003/article/00005-fra.pdf.

Boonstra, H.J., et van den Brakel, J. (2022). Multilevel time series models for small area estimation at different frequencies and domain levels. *Annals of Applied Statistics*, 16, 4, 2314-2338, DOI: 10.1214/21-AOAS1592.

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.

Carvalho, C.M., Polson, N.G. et Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2), 465-480.

Das, S., Kumar, B. et Kawsar, L.A. (2020). Disaggregated level child morbidity in Bangladesh: An application of small area estimation method. *PLoS ONE*, 15(5), e0220164.

Das, S., van den Brakel, J., Boonstra, H.J. et Haslett, S. (2021). Multilevel time series modelling of antenatal care coverage in Bangladesh at disaggregated administrative levels. Document de discussion, Statistics Netherlands.

Datta, G.S., Lahiri, P. et Maiti, T. (2002). Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference*, 102(1), 83-97.

Datta, G.S., Lahiri, P., Maiti, T. et Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the States of the U.S. *Journal of the American Statistical Association*, 94(448), 1074-1082.

DHS (2021). *The DHS Program: Demographic and Health Surveys*. https://dhsprogram.com/, consulté le 01/11/2021.

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(366), 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, 398-409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3), 515-533.

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.

Geman, S., et Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, 6, 721-741.

Ghosh, M., Nangia, N. et Kim, D.H. (1996). Estimation of median income of four-person families: A Bayesian time series approach. *Journal of the American Statistical Association*, 91(436), 1423-1431.

Haslett, S., et Jones, G. (2004). *Local Estimation of Poverty and Malnutrition in Bangladesh*. Rapport technique, Bangladesh Bureau of Statistics and United Nations World Food Programme.

Haslett, S., Jones, G. et Isidro, M. (2014). *Local Estimation of Poverty and Malnutrition in Bangladesh*. Rapport technique, United Nations World Food Programme and International Fund for Agricultural Development.

Hossain, M.J., Das, S. et Chandra, H. (2020). Disaggregate level estimates and spatial mapping of food insecurity in Bangladesh by linking survey and census data. *PLoS ONE*, 15(4), e0230906.

Marhuenda, Y., Molina, I. et Morales, D. (2013). Small area estimation with spatio-temporal Fay–Herriot models. *Computational Statistics & Data Analysis*, 58, 308-325.

Mrisho, M., Obrist, B., Schellenberg, J.A., Haws, R.A., Mushi, A.K., Mshinda, H., Tanner, M. et Schellenberg, D. (2009). The use of antenatal and postnatal care: Perspectives and experiences of women and health care providers in rural Southern Tanzania. *BMC Pregnancy and Childbirth*, 9(1), 10.

NIPORT (2015a). *Bangladesh Demographic and Health Survey 2014: Key Indicators*. Rapport technique, National Institute of Population Research and Training.

NIPORT (2015b). *Bangladesh Demographic and Health Survey 2014: Policy Briefs*. Rapport technique, National Institute of Population Research and Training.

NIPORT (2016). *Bangladesh Demographic and Health Survey 2014*. Rapport technique, National Institute of Population Research and Training.

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.

OMS, UNICEF et autres (2014). *Trends in Maternal Mortality: 1990 to 2013: Estimates by WHO, UNICEF, UNFPA, the World Bank and the United Nations Population Division*. Rapport technique de l'Organisation mondiale de la santé.

Park, T., et Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482), 681-686.

Pfeffermann, D., et Burck, L. (1990). Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales. *Techniques d'enquête*, 16, 2, 229-249. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1990002/article/14534-fra.pdf.

Pfeffermann, D., et Tiller, R. (2006). Small area estimation with state-space models subject to benchmark constraints. *Journal of the American Statistical Association*, 101, 1387-1397.

Pratesi, M., et Salvati, N. (2008). Small area estimation: The EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications*, 17(1), 113-141.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienne, Autriche.

Rahman, M.S., Mohiuddin, H., Kafy, A.-A., Sheel, P.K. et Di, L. (2019). Classification of cities in Bangladesh based on remote sensing derived spatial characteristics. *Journal of Urban Management*, 8(2), 206-224.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation*. Wiley-Interscience.

Rao, J.N.K, et Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *The Canadian Journal of Statistics*, 22, 511-528.

Rue, H., et Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC, 56.

Sakia, R.M. (1992). The Box-Cox transformation technique: A review. *The Statistician*, 41, 169-178.

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 B*, 64(4), 583-639.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

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.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14, 867-897.

West, M. (1984). Outlier models and prior distributions in bayesian linear regression. *Journal of the Royal Statistical Society, Series B (Methodological)*, 431-439.

Wolter, K. (2007). *Introduction to Variance Estimation*. Springer.

You, Y. (2008). Une approche intégrée de modélisation de l'estimation du taux de chômage pour les régions infraprovinciales au Canada. *Techniques d'enquête*, 34, 1, 21-31. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2008001/article/10614-fra.pdf.

You, Y., et Rao, J.N.K. (2000). Estimation bayésienne hiérarchique des moyennes pour petites régions à l'aide de modèles à plusieurs niveaux. *Techniques d'enquête*, 26, 2, 197-206. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2000002/article/5537-fra.pdf.

You, Y., Rao, J.N.K. et Gambino, J. (2003). Estimation du taux de chômage fondée sur un modèle pour l'Enquête sur la population active du Canada : Une approche bayésienne hiérarchique. *Techniques d'enquête*, 29, 1, 27-36. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2003001/article/6602-fra.pdf.

Techniques d'enquête, décembre 2022
Vol. 48, n° 2, pp. 481-505
Statistique Canada, n° 12-001-X au catalogue

Estimation linéaire optimale dans un échantillonnage à deux phases

Takis Merkouris[1]

Résumé

L'échantillonnage à deux phases est un plan de sondage rentable couramment utilisé dans les enquêtes. Le présent article propose une méthode optimale d'estimation linéaire des totaux dans un échantillonnage à deux phases, qui exploite au mieux l'information auxiliaire de l'enquête. Tout d'abord, on calcule formellement un meilleur estimateur linéaire sans biais (MELSB) de tout total sous une forme analytique, et on démontre qu'il s'agit d'un estimateur par calage. Ensuite, la reformulation appropriée du MELSB et l'estimation de ses coefficients inconnus permettent de construire un estimateur par la régression « optimal », qui peut également être obtenu au moyen d'une procédure de calage adéquate. Ce calage présente une caractéristique distinctive : l'alignement des estimations des deux phases dans une procédure en une étape comprenant les échantillons combinés de la première et de la deuxième phase. L'estimation optimale est faisable pour certains plans à deux phases souvent employés dans les enquêtes à grande échelle. Pour les plans généraux à deux phases, une autre procédure de calage donne un estimateur par la régression généralisée comme estimateur optimal approximatif. L'approche générale proposée d'estimation optimale permet d'utiliser le plus efficacement possible l'information auxiliaire disponible dans toute enquête à deux phases. Les avantages de cette méthode par rapport aux méthodes existantes d'estimation dans un échantillonnage à deux phases sont démontrés théoriquement et au moyen d'une étude par simulations.

Mots-clés : Information auxiliaire; meilleure estimation linéaire sans biais; calage; estimation par la régression généralisée; échantillonnage double.

1. Introduction

Le plan d'échantillonnage à deux phases, aussi appelé échantillonnage double, a toujours été utilisé dans les enquêtes par sondage en tant que méthode d'enquête rentable. Pendant la première phase, un échantillon relativement important de la population cible est sélectionné afin de fournir des renseignements auxiliaires peu coûteux à obtenir. Cet échantillon constitue une base très informative à partir de laquelle un sous-échantillon est sélectionné à la deuxième phase pour recueillir des renseignements sur les éléments d'intérêt. De plus, l'échantillonnage à deux phases est de plus en plus utilisé comme mécanisme de traitement de la non-réponse. Särndal, Swensson et Wretman (1992) présentent un compte rendu exhaustif de ce type d'emploi de l'échantillonnage à deux phases. Groves et Heeringa (2006) ainsi que Brick et Tourangeau (2017) traitent du rôle important de l'échantillonnage à deux phases dans les plans de sondage réactifs quand des mesures coûteuses sont prises pour réduire le biais de non-réponse. D'autres applications de l'échantillonnage à deux phases – apparues récemment dans la pratique des enquêtes – comportent diverses formes d'intégration d'enquêtes distinctes. Dans une de ces formes, un échantillon de première phase sert de base de sondage à l'échantillon de deuxième phase pour une multitude d'enquêtes semblables (Turmelle et Beaucage, 2013). Dans un autre cas, une enquête primaire de grande envergure est utilisée comme base de sondage pour une autre enquête de plus petite envergure comportant un plus grand nombre d'éléments (Australian Bureau of Statistics, 2004).

1. Takis Merkouris, Department of Statistics, Athens University of Economics and Business, 2, rue Trias, Athènes 11362, Grèce. Courriel : merkouris@aueb.gr.

Dans un échantillonnage à deux phases, l'information auxiliaire peut se trouver à différents niveaux. Certains renseignements sont au niveau de l'ensemble de la population, alors que d'autres sont au niveau de l'échantillon de la première ou de la deuxième phase. De nombreuses études portent sur l'utilisation de ces renseignements en vue de l'amélioration de l'estimation des totaux ou des moyennes de la population; voir Särndal et coll. (1992), Hidiroglou et Särndal (1998), Hidiroglou (2001), Estevao et Särndal (2002, 2009), Wu et Luan (2003), Chen et Kim (2014), et les références qui s'y trouvent. En général, la littérature présente deux méthodes d'intégration de l'information auxiliaire au processus d'estimation : l'approche par régression généralisée et l'approche par calage; les deux phases de l'échantillonnage supposent deux modèles de régression ou deux calages successifs. Dans certaines conditions, les deux approches donnent des estimateurs identiques, mais en général, ce n'est pas le cas. L'estimation de la variance de ces estimateurs à deux phases a fait l'objet de nombreuses études; voir par exemple Sitter (1997), Fuller (1998), Kim et Sitter (2003), Kim, Navarro et Fuller (2006), Hidiroglou, Rao et Haziza (2008), Kim et Yu (2011), Beaumont, Beliveau et Haziza (2015).

Que ce soit une formulation par la régression ou par calage des procédures d'estimation existantes, les estimateurs obtenus pour une variable cible sont en fait des combinaisons linéaires d'estimateurs de Horvitz-Thompson de divers totaux (ou moyennes), y compris l'estimateur de la variable cible calculé à partir de l'échantillon de la deuxième phase et des estimateurs de variables auxiliaires calculés à partir à la fois des échantillons de première et de deuxième phase. Dans le présent article, en adoptant une approche formelle de l'estimation optimale, nous considérons la combinaison linéaire la plus efficace des estimateurs disponibles à partir des deux phases, selon le principe de la meilleure estimation linéaire sans biais. Nous montrons que le meilleur estimateur linéaire sans biais (MELSB) calculé sous une forme analytique possède une propriété d'orthogonalité utile et qu'il peut être construit autrement comme un estimateur par calage, qui est linéaire dans les valeurs de la variable associée et intègre l'information auxiliaire dans les poids de sondage calés. L'estimation des coefficients inconnus de ce MELSB, au moyen de toute l'information auxiliaire disponible des deux phases d'échantillonnage, donne un estimateur « optimal », analogue à l'estimateur par la régression optimale en une phase de Montanari (1987) et Rao (1994). Cet estimateur est une approximation à grand échantillon du MELSB, par des coefficients estimés minimisant sa variance approximative estimée (grand échantillon) et préservant la propriété d'orthogonalité du MELSB. Au moyen d'une reformulation adéquate du MELSB, l'estimateur optimal peut également être obtenu par une procédure de calage appropriée. La caractéristique distinctive de ce calage est la procédure pratique en une étape consistant à aligner les estimations des deux phases au moyen des échantillons combinés de la première et de la deuxième phase. L'estimation optimale est faisable pour certains plans à deux phases souvent employés dans les enquêtes à grande échelle. Pour les plans généraux, une autre procédure de calage à une étape donne un nouvel estimateur par la régression généralisée comme approximation pratique de l'estimateur optimal.

La méthode générale d'estimation proposée guide la construction d'estimateurs par calage dans tout cas particulier d'enquête à deux phases, en s'appuyant le plus efficacement possible sur l'information

auxiliaire disponible. Elle permet aussi de donner un aperçu des méthodes d'estimation moins efficaces quand elles sont placées dans le cadre d'une estimation optimale. Les avantages de la méthode proposée par rapport aux méthodes existantes sont démontrés théoriquement et au moyen d'une étude par simulations.

Le présent article est organisé comme suit. La structure du plan d'échantillonnage à deux phases et la notation sont présentées dans la section 2. Le calcul du MELSB pour l'information auxiliaire de type standard dans l'échantillonnage à deux phases et son autre construction comme estimateur par calage, sont décrits à la section 3. L'estimateur optimal à deux phases et son équivalent par calage sont présentés à la section 4. L'approximation de l'estimateur optimal par un estimateur par la régression généralisée est traitée dans la section 5. Des comparaisons avec les méthodes existantes sont données à la section 6. Une étude par simulations est présentée dans la section 7. L'article se conclut par une discussion à la section 8.

2. Plan d'échantillonnage à deux phases : structure et notation

Soit $U=\{1,\ldots,k,\ldots,N\}$ une population finie de N unités. Un échantillon de la première phase s_1 de taille n_1 est tiré de la population U, au moyen d'un plan de sondage qui définit la probabilité d'inclusion $\pi_{1k}=P(k\in s_1)$ pour l'unité $k\in U$ et la probabilité d'inclusion conjointe $\pi_{1kl}=P(k,l\in s_1)$ pour les unités $k,l\in U$. Ensuite, un échantillon de la deuxième phase s_2 de taille n_2 est tiré de s_1 au moyen d'un plan de sondage qui définit la probabilité d'inclusion conditionnelle $\pi_{2k}=P(k\in s_2\,|\,s_1)$ pour $k\in s_1$, et la probabilité d'inclusion conditionnelle conjointe $\pi_{2kl}=P(k,l\in s_2\,|\,s_1)$ pour les unités $k,l\in s_1$. En supposant que $\pi_{1k}>0$ pour tous les $k\in U$ et $\pi_{2k}>0$ pour tous les $k\in s_1$, le poids de sondage de la première phase pour $k\in s_1$ est $w_{1k}=1/\pi_{1k}$, le poids de sondage conditionnel de la deuxième phase pour $k\in s_2$ est $w_{2k}=1/\pi_{2k}$, et le poids de sondage global pour $k\in s_2$ est $w_k=w_{1k}w_{2k}$.

Le type standard de variables auxiliaires dans un échantillonnage à deux phases (voir, par exemple, Särndal et coll. (1992)) comporte un vecteur de variables auxiliaires $\mathbf{x}$, partitionné comme étant $\mathbf{x}=(\mathbf{x}_1',\mathbf{x}_2')'$ par p et q de ses composantes, avec un total de population $\mathbf{t}_\mathbf{x}=\sum_U\mathbf{x}_k$ et un total connu $\mathbf{t}_{\mathbf{x}_1}=\sum_U\mathbf{x}_{1k}$ de $\mathbf{x}_1$. La valeur $\mathbf{x}_k$ est observée pour chaque unité $k\in s_1$, alors que pour un vecteur à d dimensions de variables cibles $\mathbf{y}$, avec un total $\mathbf{t}_\mathbf{y}=\sum_U\mathbf{y}_k$, la valeur $\mathbf{y}_k$ est observée seulement pour les unités $k\in s_2$. Dans certaines enquêtes, les composantes du vecteur $\mathbf{x}_2$ sont aussi des variables cibles. Un estimateur sans biais du total $\mathbf{t}_\mathbf{y}$, l'estimateur commun de Horvitz-Thompson (HT), donné par $\tilde{\mathbf{t}}_\mathbf{y}=\sum_{s_2}w_k\mathbf{y}_k$, est obtenu au moyen de l'échantillon de la deuxième phase s_2, tandis que deux estimateurs de HT du total $\mathbf{t}_\mathbf{x}$, donnés par $\hat{\mathbf{t}}_\mathbf{x}=\sum_{s_1}w_{1k}\mathbf{x}_k$ et $\tilde{\mathbf{t}}_\mathbf{x}=\sum_{s_2}w_k\mathbf{x}_k$, sont obtenus respectivement au moyen des échantillons s_1 et s_2. Dans le calcul des résultats comprenant ces estimateurs, nous utilisons la notation de vecteur $\tilde{\mathbf{t}}_\mathbf{y}=\mathbf{Y}_2'\mathbf{w}$, $\hat{\mathbf{t}}_\mathbf{x}=\mathbf{X}_1'\mathbf{w}_1$, $\tilde{\mathbf{t}}_\mathbf{x}=\mathbf{X}_2'\mathbf{w}$, $\hat{\mathbf{t}}_{\mathbf{x}_1}=\mathbf{X}_{11}'\mathbf{w}_1$, où $\mathbf{w}_1$ et $\mathbf{w}$ désignent les vecteurs des poids de sondage pour les échantillons s_1 et s_2, respectivement, $\mathbf{X}_1$ et $\mathbf{X}_{11}$ désignent les matrices de l'échantillon s_1 de $\mathbf{x}$ et $\mathbf{x}_1$ de dimensions $n_1\times(p+q)$ et $n_1\times p$, respectivement, et $\mathbf{Y}_2$, $\mathbf{X}_2$ désignent les matrices de l'échantillon s_2 de $\mathbf{y}$ et $\mathbf{x}$ de dimensions $n_2\times d$ et $n_2\times(p+q)$.

La principale cible de l'estimation est le total $\mathbf{t}_\mathbf{y}$. Cependant, pour mieux comprendre la construction des estimateurs proposés, et puisque les composantes du vecteur $\mathbf{x}_2$ peuvent également être des variables cibles, nous adopterons une approche unifiée pour estimer à la fois $\mathbf{t}_\mathbf{y}$ et $\mathbf{t}_\mathbf{x}$.

3. Meilleure estimation linéaire sans biais dans un échantillonnage à deux phases

3.1 Forme analytique du meilleur estimateur linéaire sans biais

Pour estimer plus efficacement les totaux $\mathbf{t}_\mathbf{y}$ et $\mathbf{t}_\mathbf{x}$, en incorporant toute l'information disponible des deux phases par la corrélation de $\mathbf{y}$ et $\mathbf{x}$, nous examinons les meilleurs estimateurs linéaires sans biais (MELSB), désignés par $\hat{\mathbf{t}}_\mathbf{y}^B$ et $\hat{\mathbf{t}}_\mathbf{x}^B$, qui sont des combinaisons linéaires sans biais à variance minimale des quatre estimateurs $\tilde{\mathbf{t}}_\mathbf{y}$, $\hat{\mathbf{t}}_\mathbf{x}$, $\tilde{\mathbf{t}}_\mathbf{x}$, $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ et qui sont donnés sous forme matricielle par :

$$\left(\hat{\mathbf{t}}_\mathbf{y}^{B'}, \hat{\mathbf{t}}_\mathbf{x}^{B'}\right)' = \boldsymbol{\mathcal{P}}\left(\tilde{\mathbf{t}}_\mathbf{y}', \hat{\mathbf{t}}_\mathbf{x}', \tilde{\mathbf{t}}_\mathbf{x}', \mathbf{t}_{\mathbf{x}_1}' - \hat{\mathbf{t}}_{\mathbf{x}_1}'\right)', \tag{3.1}$$

où $\boldsymbol{\mathcal{P}} = (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{V}^{-1}$, la matrice $\mathbf{W}$ comporte des entrées 1 et 0 et satisfait à $E\left[(\tilde{\mathbf{t}}_\mathbf{y}', \hat{\mathbf{t}}_\mathbf{x}', \tilde{\mathbf{t}}_\mathbf{x}', \mathbf{t}_{\mathbf{x}_1}' - \hat{\mathbf{t}}_{\mathbf{x}_1}')'\right] = \mathbf{W}(\mathbf{t}_\mathbf{y}', \mathbf{t}_\mathbf{x}')'$, et où $\mathbf{V}$ est la matrice de covariance de $(\tilde{\mathbf{t}}_\mathbf{y}', \hat{\mathbf{t}}_\mathbf{x}', \tilde{\mathbf{t}}_\mathbf{x}', \mathbf{t}_{\mathbf{x}_1}' - \hat{\mathbf{t}}_{\mathbf{x}_1}')'$. Il s'ensuit que $\mathrm{Var}\left[(\hat{\mathbf{t}}_\mathbf{y}^{B'}, \hat{\mathbf{t}}_\mathbf{x}^{B'})'\right] = (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}$. Cette formulation typique de la meilleure estimation linéaire sans biais a été étudiée dans deux autres domaines de l'échantillonnage d'enquête, comme dans Wolter (1979), Jones (1980), Fuller (1990) et Chipperfield et Steel (2009). Dans le contexte actuel, une formulation plus pratique, qui conduit également à la représentation du MELSB comme estimateur par calage, se présente comme suit.

En formulant les deux combinaisons linéaires dans (3.1) sous une forme développée et en utilisant la condition d'absence de biais $E(\hat{\mathbf{t}}_\mathbf{y}^B) = \mathbf{t}_\mathbf{y}$ et $E(\hat{\mathbf{t}}_\mathbf{x}^B) = \mathbf{t}_\mathbf{x}$, on montre facilement que la matrice $\boldsymbol{\mathcal{P}}$ des coefficients dans ces combinaisons linéaires satisfait à :

$$\boldsymbol{\mathcal{P}} = \begin{pmatrix} \mathbf{B}_{1\mathbf{y}} & \mathbf{B}_{2\mathbf{y}} & \mathbf{B}_{3\mathbf{y}} & \mathbf{B}_{4\mathbf{y}} \\ \mathbf{B}_{1\mathbf{x}} & \mathbf{B}_{2\mathbf{x}} & \mathbf{B}_{3\mathbf{x}} & \mathbf{B}_{4\mathbf{x}} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{B}_{2\mathbf{y}} & -\mathbf{B}_{2\mathbf{y}} & \mathbf{B}_{4\mathbf{y}} \\ \mathbf{0} & \mathbf{B}_{2\mathbf{x}} & \mathbf{I} - \mathbf{B}_{2\mathbf{x}} & \mathbf{B}_{4\mathbf{x}} \end{pmatrix},$$

et qu'ensuite les deux composants du MELSB dans (3.1) s'expriment sous la forme d'une régression :

$$\begin{aligned} \hat{\mathbf{t}}_\mathbf{y}^B &= \tilde{\mathbf{t}}_\mathbf{y} + \mathbf{B}_{2\mathbf{y}}(\hat{\mathbf{t}}_\mathbf{x} - \tilde{\mathbf{t}}_\mathbf{x}) + \mathbf{B}_{4\mathbf{y}}(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}) \\ \hat{\mathbf{t}}_\mathbf{x}^B &= \tilde{\mathbf{t}}_\mathbf{x} + \mathbf{B}_{2\mathbf{x}}(\hat{\mathbf{t}}_\mathbf{x} - \tilde{\mathbf{t}}_\mathbf{x}) + \mathbf{B}_{4\mathbf{x}}(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}). \end{aligned} \tag{3.2}$$

Nous pouvons maintenant formuler (3.1) comme suit :

$$\begin{pmatrix} \hat{\mathbf{t}}_\mathbf{y}^B \\ \hat{\mathbf{t}}_\mathbf{x}^B \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{t}}_\mathbf{y} \\ \tilde{\mathbf{t}}_\mathbf{x} \end{pmatrix} + \boldsymbol{\mathcal{B}} \begin{pmatrix} \hat{\mathbf{t}}_\mathbf{x} - \tilde{\mathbf{t}}_\mathbf{x} \\ \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1} \end{pmatrix}, \tag{3.3}$$

où la matrice $\boldsymbol{\mathcal{B}}$ se compose des deuxième et quatrième colonnes de $\boldsymbol{\mathcal{P}}$, et a la valeur de minimisation de variance qui se calcule facilement :

$$\mathcal{B} = -\text{Cov}\left[\begin{pmatrix}\tilde{\mathbf{t}}_{\mathbf{y}}\\ \tilde{\mathbf{t}}_{\mathbf{x}}\end{pmatrix}, \begin{pmatrix}\hat{\mathbf{t}}_{\mathbf{x}}-\tilde{\mathbf{t}}_{\mathbf{x}}\\ \mathbf{t}_{\mathbf{x}_1}-\hat{\mathbf{t}}_{\mathbf{x}_1}\end{pmatrix}\right]\left[\text{Var}\begin{pmatrix}\hat{\mathbf{t}}_{\mathbf{x}}-\tilde{\mathbf{t}}_{\mathbf{x}}\\ \mathbf{t}_{\mathbf{x}_1}-\hat{\mathbf{t}}_{\mathbf{x}_1}\end{pmatrix}\right]^{-1}. \tag{3.4}$$

Nous posons ensuite :

$$\mathbf{w}^* = \begin{pmatrix}\mathbf{w}_1\\ \mathbf{w}\end{pmatrix}, \quad \boldsymbol{\mathcal{X}} = \begin{pmatrix}-\mathbf{X}_1 & \mathbf{X}_{11}\\ \mathbf{X}_2 & \mathbf{0}\end{pmatrix}, \quad \boldsymbol{\Psi} = \begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{Y}_2 & \mathbf{X}_2\end{pmatrix}, \tag{3.5}$$

de sorte que :

$$\boldsymbol{\mathcal{X}}'\mathbf{w}^* = \begin{pmatrix}\tilde{\mathbf{t}}_{\mathbf{x}}-\hat{\mathbf{t}}_{\mathbf{x}}\\ \hat{\mathbf{t}}_{\mathbf{x}_1}\end{pmatrix}, \qquad \boldsymbol{\Psi}'\mathbf{w}^* = \begin{pmatrix}\tilde{\mathbf{t}}_{\mathbf{y}}\\ \tilde{\mathbf{t}}_{\mathbf{x}}\end{pmatrix}, \tag{3.6}$$

et $\mathcal{B}$ peut ensuite être exprimé comme étant $\mathcal{B} = \text{Cov}(\boldsymbol{\Psi}'\mathbf{w}^*, \boldsymbol{\mathcal{X}}'\mathbf{w}^*)\left[\text{Var}(\boldsymbol{\mathcal{X}}'\mathbf{w}^*)\right]^{-1}$. Pour le calcul des variances et des covariances, nous définissons $\mathbf{w}^*$ au niveau de la population comme étant $\mathbf{w}_U^* = (\mathbf{w}_{1U}', \mathbf{w}_U')'$, où le k^{e} élément de $\mathbf{w}_{1U}$ est $w_{1U_k} = (1/\pi_{1k})\, I_{1k}$, la variable indicatrice I_1 désignant l'inclusion d'une unité de population dans s_1, et le k^{e} élément de $\mathbf{w}_U$ est $w_{U_k} = \left[1/(\pi_{1k}\pi_{2k})\right] I_{1k} I_{2k}$, la variable indicatrice I_2 désignant l'inclusion d'une unité de population dans s_2 conditionnellement à la sélection de l'échantillon s_1. Nous pouvons maintenant écrire $\boldsymbol{\mathcal{X}}'\mathbf{w}^* = \boldsymbol{\mathcal{X}}_U'\mathbf{w}_U^*$ et $\boldsymbol{\Psi}'\mathbf{w}^* = \boldsymbol{\Psi}_U'\mathbf{w}_U^*$, où $\boldsymbol{\mathcal{X}}_U$ et $\boldsymbol{\Psi}_U$ sont les contreparties de population de $\boldsymbol{\mathcal{X}}$ et $\boldsymbol{\Psi}$, respectivement; toutes les sous-matrices dans $\boldsymbol{\mathcal{X}}$ et $\boldsymbol{\Psi}$ sont élargies au niveau de la population et ont N lignes. Ensuite, en indiquant $\hat{\mathbf{t}}_{\boldsymbol{\Psi}} = \boldsymbol{\Psi}'\mathbf{w}^*$ et $\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}} = \boldsymbol{\mathcal{X}}'\mathbf{w}^*$, nous obtenons :

$$\mathcal{B} = \text{Cov}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})\left[\text{Var}(\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})\right]^{-1} = \boldsymbol{\Psi}_U'\,\text{Var}(\mathbf{w}_U^*)\,\boldsymbol{\mathcal{X}}_U\left[\boldsymbol{\mathcal{X}}_U'\,\text{Var}(\mathbf{w}_U^*)\,\boldsymbol{\mathcal{X}}_U\right]^{-1}. \tag{3.7}$$

Nous obtenons par la suite une expression plus analytique utile de $\mathcal{B}$ en utilisant le lemme suivant. La démonstration se trouve en annexe.

Lemme 1

$$\text{Var}(\mathbf{w}_U^*) = \begin{pmatrix}\text{Var}(\mathbf{w}_{1U}) & \text{Var}(\mathbf{w}_{1U})\\ \text{Var}(\mathbf{w}_{1U}) & \text{Var}(\mathbf{w}_U)\end{pmatrix}, \tag{3.8}$$

où $\text{Var}(\mathbf{w}_{1U}) = \left\{(\pi_{1kl} - \pi_{1k}\pi_{1l})/\pi_{1k}\pi_{1l}\right\}$, $\text{Var}(\mathbf{w}_U) = \left\{(\pi_{1kl}\pi_{2kl} - \pi_{1k}\pi_{2k}\pi_{1l}\pi_{2l})/\pi_{1k}\pi_{2k}\pi_{1l}\pi_{2l}\right\}$.

Au moyen de (3.7) et (3.8), il est facile de démontrer que (3.4) s'exprime comme suit :

$$\mathcal{B} = \begin{bmatrix}-\text{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}})\left[\text{Var}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}})\right]^{-1} & \text{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1})\left[\text{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1})\right]^{-1}\\ \mathbf{I} & \text{Cov}(\tilde{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1})\left[\text{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1})\right]^{-1}\end{bmatrix}. \tag{3.9}$$

Dans cette représentation de $\mathcal{B}$, la propriété $\text{Cov}(\tilde{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}}) = \text{Var}(\hat{\mathbf{t}}_{\mathbf{x}})$ est implicite, découlant de (3.8), ce qui implique que $\text{Var}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \text{Var}(\tilde{\mathbf{t}}_{\mathbf{x}}) - \text{Var}(\hat{\mathbf{t}}_{\mathbf{x}})$, et la propriété $\text{Cov}(\tilde{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) = \text{Cov}(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1})$, ce qui

implique $\mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$ (cette covariance étant le bloc hors diagonale de $\boldsymbol{\mathcal{X}}_U' \mathrm{Var}(\mathbf{w}_U^*) \boldsymbol{\mathcal{X}}_U$). Ensuite, (3.2) peut s'exprimer explicitement comme suit :

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{y}}^B &= \tilde{\mathbf{t}}_{\mathbf{y}} - \mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \right]^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1}) \right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}) \\ \hat{\mathbf{t}}_{\mathbf{x}}^B &= \hat{\mathbf{t}}_{\mathbf{x}} + \mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1}) \right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}). \end{aligned} \tag{3.10}$$

Compte tenu de la propriété $\mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$, il s'ensuit immédiatement que :

$$\begin{aligned} \mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{y}}^B) &= \mathrm{Var}(\tilde{\mathbf{t}}_{\mathbf{y}}) - \mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \right]^{-1} \mathrm{Cov}'(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad - \mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1}) \right]^{-1} \mathrm{Cov}'(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) \\ \mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}}^B) &= \mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}}) - \mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1}) \right]^{-1} \mathrm{Cov}'(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1}). \end{aligned} \tag{3.11}$$

Remarque 3.1. Chaque composante ou combinaison linéaire de composantes de $\hat{\mathbf{t}}_{\mathbf{y}}^B$ est le MELSB pour le total correspondant. De plus, comme le montre (3.11), l'efficacité de $\hat{\mathbf{t}}_{\mathbf{y}}^B$, par rapport à $\tilde{\mathbf{t}}_{\mathbf{y}}$, dépend de la force de la corrélation de $\mathbf{y}$ avec $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ ainsi que de la différence de taille de l'échantillon (et peut-être du plan de sondage) pour les échantillons s_1 et s_2.

Remarque 3.2. En raison de la propriété d'orthogonalité $\mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$, le coefficient de l'un ou l'autre des termes $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ et $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ dans (3.10) ne changerait pas si l'autre était égal à $\mathbf{0}$ dans (3.2). Par exemple, le MELSB pour $\mathbf{t}_{\mathbf{y}}$ basé sur $(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}}, \tilde{\mathbf{t}}_{\mathbf{x}})$ serait $\hat{\mathbf{t}}_{\mathbf{y}}^B$ comme dans (3.10), mais sans le dernier terme. Ceci se résout facilement comme un cas particulier de la configuration complète $(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}}, \tilde{\mathbf{t}}_{\mathbf{x}}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$. Cette propriété d'orthogonalité explique la réduction additive de la variance observée dans la première équation de (3.11).

Remarque 3.3. Le MELSB $\hat{\mathbf{t}}_{\mathbf{x}}^B$ dans (3.10) peut également être produit au moyen de la configuration réduite $(\hat{\mathbf{t}}_{\mathbf{x}}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ dans (3.1). Le même meilleur estimateur linéaire, pour une seule variable cible, a été calculé différemment dans un contexte d'échantillonnage général à une phase par Fuller et Isaki (1981) et Montanari (1987). En particulier, pour la variable auxiliaire $\mathbf{x}_1$, nous avons $\hat{\mathbf{t}}_{\mathbf{x}_1}^B = \mathbf{t}_{\mathbf{x}_1}$. Ensuite, il est facile de vérifier que le MELSB dans (3.1) peut par contre être calculé en deux étapes de meilleure estimation linéaire sans biais au moyen de la configuration $(\tilde{\mathbf{t}}_{\mathbf{y}}^B, \hat{\mathbf{t}}_{\mathbf{x}}^B, \tilde{\mathbf{t}}_{\mathbf{x}}^B)$, où $\tilde{\mathbf{t}}_{\mathbf{y}}^B = \tilde{\mathbf{t}}_{\mathbf{y}} + \mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1}) \right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ et $\tilde{\mathbf{t}}_{\mathbf{x}}^B = \tilde{\mathbf{t}}_{\mathbf{x}} + \mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1}) \right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ sont les MELSB produits respectivement par les configurations à une phase $(\tilde{\mathbf{t}}_{\mathbf{y}}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ et $(\tilde{\mathbf{t}}_{\mathbf{x}}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$. De fastidieux calculs algébriques permettraient de démontrer que $(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_2}, \tilde{\mathbf{t}}_{\mathbf{x}_2}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}, \mathbf{t}_{\mathbf{x}_1} - \tilde{\mathbf{t}}_{\mathbf{x}_1})$ est une autre configuration du MELSB, qui est plus explicite et équivaut à celle de (3.1). Cela prouve que la configuration compacte de (3.1) fournit l'estimation linéaire la plus efficace de $\mathbf{t}_{\mathbf{y}}$ et $\mathbf{t}_{\mathbf{x}}$ au moyen de toutes les estimations pertinentes disponibles.

3.2 Le meilleur estimateur linéaire sans biais à deux phases comme estimateur par calage

En utilisant la notation conduisant à (3.7), et la configuration $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{B} = (\hat{\mathbf{t}}_{\mathbf{y}}^{B'}, \hat{\mathbf{t}}_{\mathbf{x}}^{B'})'$ et $\boldsymbol{\Delta} = \mathrm{Var}(\mathbf{w}_U^*)$, nous pouvons exprimer le MELSB dans (3.3) comme étant $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{B} = \hat{\mathbf{t}}_{\boldsymbol{\Psi}} + \boldsymbol{\mathcal{B}}(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$, où $\boldsymbol{\mathcal{B}} = \boldsymbol{\Psi}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U (\boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1}$ et $\mathbf{t}_{\boldsymbol{\mathcal{X}}} = (\mathbf{0}', \mathbf{t}_{\mathbf{x}_1}')'$, ou sous une forme plus évocatrice :

$$\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{B} \;=\; \boldsymbol{\Psi}_U' \Big[\mathbf{w}_U^* + \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U (\boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1} \big(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}_U' \mathbf{w}_U^* \big) \Big]. \tag{3.12}$$

D'après (3.12), il semble que $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{B}$ a la forme d'un estimateur par calage, qui comporte un vecteur de population de poids calés $\mathbf{c}_U^* = \mathbf{w}_U^* + \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U (\boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1} \big(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}_U' \mathbf{w}_U^* \big)$ et un vecteur de totaux de calage $\mathbf{t}_{\boldsymbol{\mathcal{X}}}$. Cette constatation est mise en forme dans le théorème suivant et démontrée en annexe.

Théorème 1. *Le vecteur* $\mathbf{c}_U^* = \mathbf{w}_U^* + \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U (\boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1} \big(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}_U' \mathbf{w}_U^* \big)$ *minimise la distance des moindres carrés généralisés* $(\mathbf{c}_U^* - \mathbf{w}_U^*)' \boldsymbol{\Delta}^{-1} (\mathbf{c}_U^* - \mathbf{w}_U^*)$ *soumise aux contraintes* $\boldsymbol{\mathcal{X}}_U' \mathbf{c}_U^* = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$, *c'est-à-dire* $\mathbf{X}_U' \mathbf{c}_{1U} = \mathbf{X}_U' \mathbf{c}_U$ *et* $\mathbf{X}_{1U}' \mathbf{c}_{1U} = \mathbf{t}_{\mathbf{x}_1}$, *où* $(\mathbf{c}_{1U}, \mathbf{c}_U)$ *correspond à* $(\mathbf{w}_{1U}, \mathbf{w}_U)$.

Le théorème 1 montre que la meilleure estimation linéaire sans biais au moyen de la configuration $(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}}, \tilde{\mathbf{t}}_{\mathbf{x}}, \mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ est essentiellement une procédure de calage par laquelle les deux estimations $\hat{\mathbf{t}}_{\mathbf{x}}$ et $\tilde{\mathbf{t}}_{\mathbf{x}}$ de $\mathbf{t}_{\mathbf{x}}$ sont calées l'une sur l'autre, c'est-à-dire qu'elles sont alignées, et que l'estimation est calée sur le total $\hat{\mathbf{t}}_{\mathbf{x}_1}$. Nous pouvons maintenant exprimer formellement le MELSB $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{B}$ comme un estimateur par calage $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{B} = \boldsymbol{\Psi}_U' \mathbf{c}_U^*$, ses deux composantes étant données dans les formes linéaires simples $\hat{\mathbf{t}}_{\mathbf{y}}^{B} = \mathbf{Y}_U' \mathbf{c}_U$ et $\hat{\mathbf{t}}_{\mathbf{x}}^{B} = \mathbf{X}_U' \mathbf{c}_U$.

L'autre solution de construction en deux étapes du MELSB mentionnée dans la remarque 3.3 ci-dessus peut également être réalisée au moyen d'une procédure de calage en deux étapes comprenant $\mathbf{w}_U^*$ dans les deux étapes. En effet, en partitionnant $\boldsymbol{\mathcal{X}}_U$ par ses deux sous-matrices de colonnes en tant que $\boldsymbol{\mathcal{X}}_U = (\boldsymbol{\mathcal{X}}_{12U}, \boldsymbol{\mathcal{X}}_{1U})$ et en relevant que $\boldsymbol{\mathcal{X}}_{12U}' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_{1U} = \mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$, nous pouvons facilement décomposer le vecteur $\mathbf{c}_U^*$ comme suit :

$$\begin{aligned} \mathbf{c}_U^* \;&= \mathbf{w}_U^* + \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_{12U} (\boldsymbol{\mathcal{X}}_{12U}' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_{12U})^{-1} \big(\mathbf{0} - \boldsymbol{\mathcal{X}}_{12U}' \mathbf{w}_U^* \big) \\ &\quad + \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_{1U} (\boldsymbol{\mathcal{X}}_{1U}' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_{1U})^{-1} \big(\mathbf{t}_{\mathbf{x}_1} - \boldsymbol{\mathcal{X}}_{1U}' \mathbf{w}_U^* \big). \end{aligned} \tag{3.13}$$

Dans le membre de droite de (3.13), la somme des premier et deuxième termes résulte d'un calage avec une contrainte $\boldsymbol{\mathcal{X}}_{12U}' \mathbf{c}_U^* = \mathbf{X}_U' \mathbf{c}_{1U} - \mathbf{X}_U' \mathbf{c}_U = \mathbf{0}$ seulement, tandis que la somme des premier et troisième termes résulte d'un calage avec une contrainte $\boldsymbol{\mathcal{X}}_{1U}' \mathbf{c}_U^* = \mathbf{t}_{\mathbf{x}_1}$ seulement.

Maintenant, si nous posons $\boldsymbol{\Delta}_1 = \mathrm{Var}(\mathbf{w}_{1U})$ et $\boldsymbol{\Delta}_2 = \mathrm{Var}(\mathbf{w}_U)$, ces variances étant précisées dans (3.8), il découle facilement de (3.13) que les estimateurs par calage optimaux $\hat{\mathbf{t}}_{\mathbf{y}}^{B}$ et $\hat{\mathbf{t}}_{\mathbf{x}}^{B}$ dans (3.10) peuvent s'exprimer sous une forme explicite, qui sera reprise plus tard :

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{y}}^{B} \;&= \tilde{\mathbf{t}}_{\mathbf{y}} + \big[\mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U - \mathbf{Y}_U' \boldsymbol{\Delta}_1 \mathbf{X}_U \big] \big[\mathbf{X}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U - \mathbf{X}_U' \boldsymbol{\Delta}_1 \mathbf{X}_U \big]^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \mathbf{Y}_U' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} (\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U})^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}) \\ \hat{\mathbf{t}}_{\mathbf{x}}^{B} \;&= \hat{\mathbf{t}}_{\mathbf{x}} + \mathbf{X}_U' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} (\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U})^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}). \end{aligned} \tag{3.14}$$

4. Estimation linéaire optimale dans un échantillonnage à deux phases

4.1 L'estimateur optimal à deux phases

La matrice $\mathcal{B}$ de (3.7) comprend les variances et les covariances qui doivent être estimées. Compte tenu de $\mathrm{Var}(\hat{\mathbf{t}}_{\mathcal{X}}) = \mathcal{X}_U' \boldsymbol{\Delta} \mathcal{X}_U$ et $\mathrm{Cov}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\mathcal{X}}) = \boldsymbol{\Psi}_U' \boldsymbol{\Delta} \mathcal{X}_U$, et en reprenant (3.8), les estimations sans biais évidentes sont $\widehat{\mathrm{Var}}(\hat{\mathbf{t}}_{\mathcal{X}}) = \mathcal{X}' \hat{\boldsymbol{\Delta}} \mathcal{X}$ et $\widehat{\mathrm{Cov}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\mathcal{X}}) = \boldsymbol{\Psi}' \hat{\boldsymbol{\Delta}} \mathcal{X}$, où la matrice $(n_1 + n_2) \times (n_1 + n_2)$ $\hat{\boldsymbol{\Delta}} = \widehat{\mathrm{Var}}(\mathbf{w}_U^*)$ comporte des blocs diagonaux $\hat{\boldsymbol{\Delta}}_1 = \{(\pi_{1kl} - \pi_{1k}\pi_{1l})/(\pi_{1k}\pi_{1l}\pi_{1kl})\}$, $\hat{\boldsymbol{\Delta}}_2 = \{(\pi_{1kl}\pi_{2kl} - \pi_{1k}\pi_{2k}\pi_{1l}\pi_{2l})/(\pi_{1k}\pi_{2k}\pi_{1l}\pi_{2l}\pi_{1kl}\pi_{2kl})\}$, et des blocs non diagonaux $\hat{\boldsymbol{\Delta}}_{12}$, $\hat{\boldsymbol{\Delta}}_{21} = \hat{\boldsymbol{\Delta}}_{12}'$ avec $\hat{\boldsymbol{\Delta}}_{12} = \{(\pi_{1kl} - \pi_{1k}\pi_{1l})/(\pi_{1k}\pi_{1l}\pi_{1kl}\pi_{2l})\}$, et où $\mathcal{X}$, $\boldsymbol{\Psi}$ sont les matrices d'échantillon de (3.5).

Nous obtenons alors, comme éléments des matrices $\widehat{\mathrm{Var}}(\hat{\mathbf{t}}_{\mathcal{X}})$ et $\widehat{\mathrm{Cov}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\mathcal{X}})$, les estimations sans biais des toutes les variances et covariances de (3.9), c'est-à-dire $\widehat{\mathrm{Var}}(\hat{\mathbf{t}}_{\mathbf{x}}) = \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1$, $\widehat{\mathrm{Var}}(\tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{X}_2' \hat{\boldsymbol{\Delta}}_2 \mathbf{X}_2$, $\widehat{\mathrm{Var}}(\hat{\mathbf{t}}_{\mathbf{x}_1}) = \mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}$, $\widehat{\mathrm{Cov}}(\tilde{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) = \mathbf{X}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_{11}$, $\widehat{\mathrm{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}}) = \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_1$, $\widehat{\mathrm{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1}) = \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_{11}$, $\widehat{\mathrm{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_2 \mathbf{X}_2$. Cependant, la matrice $\widehat{\mathrm{Var}}(\hat{\mathbf{t}}_{\mathcal{X}})$ comprend aussi les éléments $\widehat{\mathrm{Cov}}(\hat{\mathbf{t}}_{\mathbf{x}}, \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_{12} \mathbf{X}_2$ et $\widehat{\mathrm{Cov}}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1 - \mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_{12} \mathbf{X}_2$, qui manifestement ne retiennent pas les propriétés $\mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}}, \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}})$ et $\mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$, respectivement. Les estimations sans biais pour les variances et les covariances de (3.9) pourraient être utilisées directement, mais l'estimation de la forme simple $\mathcal{B}$ dans (3.9) ne pourrait pas être exprimée comme étant $\boldsymbol{\Psi}' \hat{\boldsymbol{\Delta}} \mathcal{X} (\mathcal{X}' \hat{\boldsymbol{\Delta}} \mathcal{X})^{-1}$; donc l'estimateur qui en résulte ne conservait pas la forme de calage du MELSB dans (3.12). Cette complication est contournée par la reformulation suivante. Redéfinissons $\mathbf{w}^*$, $\mathcal{X}$ et $\boldsymbol{\Psi}$ comme étant :

$$\mathbf{w}^* = \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w} \\ \mathbf{w}_1 \end{pmatrix}, \qquad \mathcal{X} = \begin{pmatrix} -\mathbf{X}_1 & \mathbf{0} \\ \mathbf{X}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{11} \end{pmatrix}, \qquad \boldsymbol{\Psi} = \begin{pmatrix} -\mathbf{Y}_1 & -\mathbf{X}_1 \\ \mathbf{Y}_2 & \mathbf{X}_2 \\ \mathbf{Y}_1 & \mathbf{X}_1 \end{pmatrix}, \tag{4.1}$$

où les matrices d'échantillon $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_{11}$ et $\mathbf{Y}_2$ sont les mêmes que précédemment et $\mathbf{Y}_1$ est la matrice de $\mathbf{y}$ pour l'échantillon s_1 avec les valeurs fictives $\mathbf{y}_k$ pour $k \notin s_2$. Il apparaît clairement que $\mathcal{X}'\mathbf{w}^*$ et $\boldsymbol{\Psi}'\mathbf{w}^*$ sont exactement comme dans (3.6). Ensuite, en ayant comme précédemment $\hat{\mathbf{t}}_{\mathcal{X}} = \mathcal{X}'\mathbf{w}^*$ et $\hat{\mathbf{t}}_{\boldsymbol{\Psi}} = \boldsymbol{\Psi}'\mathbf{w}^*$, nous obtenons encore $\mathcal{B} = \mathrm{Cov}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\mathcal{X}}) \left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathcal{X}})\right]^{-1}$, où $\mathrm{Var}(\hat{\mathbf{t}}_{\mathcal{X}}) = \mathcal{X}_U' \mathrm{Var}(\mathbf{w}_U^*) \mathcal{X}_U$ et $\mathrm{Cov}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\mathcal{X}}) = \boldsymbol{\Psi}_U' \mathrm{Var}(\mathbf{w}_U^*) \mathcal{X}$, comme dans (3.7), mais $\mathbf{w}_U^*$, $\mathcal{X}_U$ et $\boldsymbol{\Psi}_U$ étant les contreparties de population des valeurs redéfinies $\mathbf{w}^*$, $\mathcal{X}$, $\boldsymbol{\Psi}$. Une extension du lemme 1 à la variable redéfinie $\mathbf{w}^*$ donne :

$$\mathrm{Var}(\mathbf{w}_U^*) = \begin{pmatrix} \mathrm{Var}(\mathbf{w}_{1U}) & \mathrm{Var}(\mathbf{w}_{1U}) & \mathrm{Var}(\mathbf{w}_{1U}) \\ \mathrm{Var}(\mathbf{w}_{1U}) & \mathrm{Var}(\mathbf{w}_U) & \mathrm{Var}(\mathbf{w}_{1U}) \\ \mathrm{Var}(\mathbf{w}_{1U}) & \mathrm{Var}(\mathbf{w}_{1U}) & \mathrm{Var}(\mathbf{w}_{1U}) \end{pmatrix},$$

où $\mathrm{Var}(\mathbf{w}_{1U})$ et $\mathrm{Var}(\mathbf{w}_U)$ sont les mêmes que dans le lemme 1. Il est maintenant facile de vérifier qu'encore une fois $\mathcal{B}$ peut être exprimé analytiquement comme dans (3.9) et que les deux composantes

du MELSB sont identiques à celles données par (3.10). Plus important encore, il découle de cette forme spéciale de $\text{Var}(\mathbf{w}_U^*)$ que nous avons de nouveau $\text{Var}(\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) = \boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U$ et $\text{Cov}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) = \boldsymbol{\Psi}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U$, où maintenant $\boldsymbol{\Delta} = \text{diag}(-\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{\Delta}_1)$ et $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$ comme cela a été défini précédemment. Ainsi, nous obtenons encore le MELSB sous la forme de calage de (3.12), et la décomposition orthogonale retenue du vecteur des poids calés dans (3.13) mène facilement à l'expression (3.14). La propriété d'orthogonalité $\boldsymbol{\mathcal{X}}_{12U}' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_{1U} = \mathbf{0}$ est induite par la structure diagonale par blocs de la variable $\boldsymbol{\mathcal{X}}_U$ redéfinie, plutôt que par la structure spéciale de la matrice initiale $\boldsymbol{\Delta}$ utilisée dans (3.12).

Pour ce qui est du MELSB reconstruit, nous avons maintenant les estimations sans biais $\widehat{\text{Var}}(\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) = \boldsymbol{\mathcal{X}}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}}$ et $\widehat{\text{Cov}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) = \boldsymbol{\Psi}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}}$, où $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\Psi}$ sont les matrices d'échantillon dans (4.1), et $\hat{\boldsymbol{\Delta}} = \text{diag}(-\hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2, \hat{\boldsymbol{\Delta}}_1)$ avec $\hat{\boldsymbol{\Delta}}_1$, $\hat{\boldsymbol{\Delta}}_2$ tels qu'ils sont définis au début de la section. À partir de là, nous recalculons facilement les estimations sans biais des variances et covariances de (3.9), mais deux des éléments de la matrice d'échantillon $\boldsymbol{\Psi}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}}$ qui comprennent $\mathbf{Y}_1$, à savoir $\mathbf{Y}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1$ et $\mathbf{Y}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}$, nécessitent une attention particulière. Les valeurs fictives (non observées) $\mathbf{y}_k$ pour $k \notin s_2$, nécessaires à l'élargissement de $\mathbf{Y}_1$ à la matrice de population $\mathbf{Y}_U$ dans le MELSB reconstruit, sont fixées égales à zéro, et les valeurs $\mathbf{y}_k$ pour $k \in s_2$ sont alors nécessairement pondérées par $1/\pi_{2k}$. Ensuite $\mathbf{Y}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1$ et $\mathbf{Y}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}$ se réduisent à $\mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_1$ et $\mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_{11}$, qui sont respectivement les estimations sans biais $\widehat{\text{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}})$ et $\widehat{\text{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \hat{\mathbf{t}}_{\mathbf{x}_1})$. La variable $\boldsymbol{\mathcal{B}}$ estimée dans (3.9) est maintenant donnée par :

$$\hat{\boldsymbol{\mathcal{B}}} = \begin{bmatrix} \left[\mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_2 \mathbf{X}_2 - \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_1\right] \left[\mathbf{X}_2' \hat{\boldsymbol{\Delta}}_2 \mathbf{X}_2 - \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1\right]^{-1} & \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_{11} \left[\mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}\right]^{-1} \\ \mathbf{I} & \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11} \left[\mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}\right]^{-1} \end{bmatrix}.$$

Le MELSB $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^B = \hat{\mathbf{t}}_{\boldsymbol{\Psi}} + \boldsymbol{\mathcal{B}}(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$ comportant une variable $\boldsymbol{\mathcal{B}}$ estimée sera appelé « estimateur linéaire sans biais optimal », « estimateur optimal » en abrégé, et sera représenté par $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O = \hat{\mathbf{t}}_{\boldsymbol{\Psi}} + \hat{\boldsymbol{\mathcal{B}}}(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$ et ses deux composantes seront données par :

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{y}}^O &= \tilde{\mathbf{t}}_{\mathbf{y}} + \left[\mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_2 \mathbf{X}_2 - \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_1\right] \left[\mathbf{X}_2' \hat{\boldsymbol{\Delta}}_2 \mathbf{X}_2 - \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_1\right]^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \mathbf{Y}_2' \hat{\boldsymbol{\Delta}}_{21} \mathbf{X}_{11} \left[\mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}\right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}) \qquad (4.2) \\ \hat{\mathbf{t}}_{\mathbf{x}}^O &= \hat{\mathbf{t}}_{\mathbf{x}} + \mathbf{X}_1' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11} \left[\mathbf{X}_{11}' \hat{\boldsymbol{\Delta}}_1 \mathbf{X}_{11}\right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}). \end{aligned}$$

Il s'agit de la version d'échantillon des MELSB dans (3.14), avec des coefficients estimés. En particulier, $\hat{\mathbf{t}}_{\mathbf{x}}^O$ est l'estimateur optimal à une phase habituel de $\mathbf{t}_{\mathbf{x}}$ reposant sur $\mathbf{x}_1$ comme variable auxiliaire et les données de l'échantillon complet de première phase s_1; voir Montanari (1987) et Rao (1994).

Remarque 4.1. Quand n_2 est très proche de n_1, l'estimateur optimal $\hat{\mathbf{t}}_{\mathbf{y}}^O$ peut être assez instable en raison de la quasi-singularité de la matrice inversée dans le coefficient de $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$, et ainsi devenir très inefficace; voir cependant la remarque 6.1 ultérieure concernant les plans à deux phases dans lesquels cette instabilité ne pose pas de problème. En général, il ne s'agit pas d'une configuration réaliste dans un échantillonnage à deux phases, où n_2 est généralement beaucoup plus petit que n_1.

Après la construction de $\mathbf{Y}_2'\hat{\boldsymbol{\Delta}}_{21}\mathbf{X}_1$ et $\mathbf{Y}_2'\hat{\boldsymbol{\Delta}}_{21}\mathbf{X}_{11}$ en tant que deux des estimations dans $\hat{\boldsymbol{\mathcal{B}}}$, il apparaît que ces deux formes bilinéaires peuvent s'exprimer autrement sous la forme $\breve{\mathbf{Y}}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_1$ et $\breve{\mathbf{Y}}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_{11}$, respectivement, où $\breve{\mathbf{Y}}_1$ est une version pondérée de $\mathbf{Y}_1$ dans laquelle $\breve{\mathbf{y}}_k = \mathbf{y}_k / \pi_{2k}$ si $k \in s_2$ et $\breve{\mathbf{y}}_k = 0$ si $k \notin s_2$. Alors $\hat{\mathbf{t}}_{\boldsymbol{\Psi}} = \boldsymbol{\Psi}'\mathbf{w}^* = \breve{\boldsymbol{\Psi}}'\mathbf{w}^*$, où $\breve{\boldsymbol{\Psi}}$ est $\boldsymbol{\Psi}$ dans (4.1), avec $\breve{\mathbf{Y}}_1$ à la place de $\mathbf{Y}_1$, et $\hat{\boldsymbol{\mathcal{B}}}$ peut être formulé de façon compacte comme étant $\hat{\boldsymbol{\mathcal{B}}} = \breve{\boldsymbol{\Psi}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}})^{-1}$, où $\hat{\boldsymbol{\Delta}} = \text{diag}(-\hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2, \hat{\boldsymbol{\Delta}}_1)$. À partir de maintenant, $\hat{\boldsymbol{\Delta}}$ désignera la matrice $\text{diag}(-\hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2, \hat{\boldsymbol{\Delta}}_1)$.

Comme c'est le cas dans Montanari (1987) et Rao (1994) pour l'estimateur optimal à une phase, pour les grands échantillons s_1 et s_2, l'estimateur optimal $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O = \hat{\mathbf{t}}_{\boldsymbol{\Psi}} + \hat{\boldsymbol{\mathcal{B}}}(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$ est une approximation du MELSB $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^B$, et il est donc approximativement sans biais. De plus, la variance de $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O$ est une approximation de celle de $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^B$, qui se résout facilement pour donner $\text{Var}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^B) = \text{Var}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}) - \text{Cov}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})\left[\text{Var}(\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})\right]^{-1}\text{Cov}'(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$, c'est-à-dire la forme compacte de (3.11). Ensuite, au moyen des estimations $\widehat{\text{Var}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}})$ et $\widehat{\text{Cov}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$, calculées auparavant, nous obtenons la variance approximative estimée de $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O$ comme étant $\widehat{\text{VA}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O) = \widehat{\text{Var}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}) - \widehat{\text{Cov}}(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})\left[\widehat{\text{Var}}(\hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})\right]^{-1}\widehat{\text{Cov}}'(\hat{\mathbf{t}}_{\boldsymbol{\Psi}}, \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$. À partir de là, nous calculons les expressions pratiques pour le calcul $\widehat{\text{VA}}(\hat{\mathbf{t}}_{\text{y}}^O) = \mathbf{Y}_2'\hat{\boldsymbol{\Delta}}_2\mathbf{Y}_2 - \breve{\boldsymbol{\Psi}}_1'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}})^{-1}\boldsymbol{\mathcal{X}}'\hat{\boldsymbol{\Delta}}\breve{\boldsymbol{\Psi}}_1$, où $\breve{\boldsymbol{\Psi}}_1$ est la première sous-matrice de colonnes de $\breve{\boldsymbol{\Psi}}$, et $\widehat{\text{VA}}(\hat{\mathbf{t}}_{\text{x}}^O) = \mathbf{X}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_1 - \mathbf{X}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_{11}\left[\mathbf{X}_{11}'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_{11}\right]^{-1}\mathbf{X}_{11}'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_1$.

4.2 L'estimateur optimal à deux phases comme estimateur par calage

L'estimateur optimal $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O = \hat{\mathbf{t}}_{\boldsymbol{\Psi}} + \hat{\boldsymbol{\mathcal{B}}}(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}})$, avec $\hat{\boldsymbol{\mathcal{B}}} = \breve{\boldsymbol{\Psi}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}})^{-1}$, prend la forme

$$\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O = \breve{\boldsymbol{\Psi}}'\left[\mathbf{w}^* + \hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}})^{-1}(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}'\mathbf{w}^*)\right],$$

d'un estimateur par calage, avec un vecteur de totaux de calage $\mathbf{t}_{\boldsymbol{\mathcal{X}}}$ et un vecteur d'échantillon de poids calés $\mathbf{c}^* = \mathbf{w}^* + \hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}})^{-1}\left(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}'\mathbf{w}^*\right)$ satisfaisant à $\boldsymbol{\mathcal{X}}'\mathbf{c}^* = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$. Cela est établi formellement par le théorème suivant, dont la démonstration est omise, car elle est semblable à celle du théorème 1.

Théorème 2. *Le vecteur* $\mathbf{c}^* = \mathbf{w}^* + \hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\hat{\boldsymbol{\Delta}}\boldsymbol{\mathcal{X}})^{-1}(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}'\mathbf{w}^*)$ *minimise la distance des moindres carrés généralisés* $(\mathbf{c}^* - \mathbf{w}^*)'\hat{\boldsymbol{\Delta}}^{-1}(\mathbf{c}^* - \mathbf{w}^*)$ *soumise aux contraintes* $\boldsymbol{\mathcal{X}}'\mathbf{c}^* = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$, *c'est-à-dire* $\mathbf{X}_1'\mathbf{c}_1 = \mathbf{X}_2'\mathbf{c}$ *et* $\mathbf{X}_{11}'\mathbf{c}_1 = \mathbf{t}_{\mathbf{x}_1}$, *où* $(\mathbf{c}_1, \mathbf{c})$ *correspond à* $(\mathbf{w}_1, \mathbf{w})$.

Le vecteur d'échantillon $\mathbf{c}^*$ admet la même décomposition orthogonale que sa contrepartie de population $\mathbf{c}_U^*$ dans (3.13). Nous pouvons maintenant exprimer formellement l'estimateur optimal $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O$ en tant qu'estimateur par calage $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O = \breve{\boldsymbol{\Psi}}'\mathbf{c}^*$ qui, compte tenu de $\boldsymbol{\mathcal{X}}'\mathbf{c}^* = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$, est généré par le calage simultané des deux estimations $\hat{\mathbf{t}}_{\mathbf{x}}$ et $\tilde{\mathbf{t}}_{\mathbf{x}}$ de $\mathbf{t}_{\mathbf{x}}$ l'une sur l'autre, et de l'estimation $\hat{\mathbf{t}}_{\mathbf{x}_1}$ sur le total $\mathbf{t}_{\mathbf{x}_1}$.

Sous une forme développée, le vecteur $\mathbf{c}^*$ est :

$$\mathbf{c}^* = \begin{pmatrix}\mathbf{c}_1\\ \mathbf{c}_2\\ \mathbf{c}_3\end{pmatrix} = \begin{pmatrix}\mathbf{w}_1 + \hat{\boldsymbol{\Delta}}_1\mathbf{X}_1[\mathbf{X}_2'\hat{\boldsymbol{\Delta}}_2\mathbf{X}_2 - \mathbf{X}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_1]^{-1}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}})\\ \mathbf{w} + \hat{\boldsymbol{\Delta}}_2\mathbf{X}_2[\mathbf{X}_2'\hat{\boldsymbol{\Delta}}_2\mathbf{X}_2 - \mathbf{X}_1'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_1]^{-1}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}})\\ \mathbf{w}_1 + \hat{\boldsymbol{\Delta}}_1\mathbf{X}_{11}(\mathbf{X}_{11}'\hat{\boldsymbol{\Delta}}_1\mathbf{X}_{11})^{-1}(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})\end{pmatrix}. \tag{4.3}$$

Ensuite, au moyen de la partition $\boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{X}}_{12}, \boldsymbol{\mathcal{X}}_1)$, où $\boldsymbol{\mathcal{X}}_{12}$ et $\boldsymbol{\mathcal{X}}_1$ sont les deux sous-matrices de colonnes orthogonales de $\boldsymbol{\mathcal{X}}$ figurant dans (4.1), les deux contraintes s'expriment par $\boldsymbol{\mathcal{X}}_{12}' \mathbf{c}^* = \mathbf{X}_2' \mathbf{c}_2 - \mathbf{X}_1' \mathbf{c}_1 = \mathbf{0}$ et $\boldsymbol{\mathcal{X}}_1' \mathbf{c}^* = \mathbf{X}_{11}' \mathbf{c}_3 = \mathbf{t}_{\mathbf{x}_1}$. Il découle aussi de (4.3) que $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O = \breve{\boldsymbol{\Psi}}' \mathbf{c}^*$ implique (4.2). Concernant les deux composantes de $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^O$, nous observons que $\hat{\mathbf{t}}_{\mathbf{x}}^O = -\mathbf{X}_1' \mathbf{c}_1 + \mathbf{X}_2' \mathbf{c}_2 + \mathbf{X}_1' \mathbf{c}_3 = \mathbf{X}_1' \mathbf{c}_3$ et que

$$\hat{\mathbf{t}}_{\mathbf{y}}^O = \breve{\mathbf{Y}}_1' (\mathbf{c}_3 - \mathbf{c}_1) + \mathbf{Y}_2' \mathbf{c}_2 = \sum\nolimits_{s_2} \left[(c_{3k} - c_{1k}) / \pi_{2k} + c_{2k} \right] \mathbf{y}_k .$$

L'expression explicite de $\hat{\mathbf{t}}_{\mathbf{y}}^O$, en termes d'unités d'échantillonnage, est :

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{y}}^O \quad &= \tilde{\mathbf{t}}_{\mathbf{y}} + \left[\sum\nolimits_{s_2} \sum\nolimits_{s_2} \hat{\Delta}_{2kl} \, \mathbf{y}_k \, \mathbf{x}_l' - \sum\nolimits_{s_2} \sum\nolimits_{s_1} \hat{\Delta}_{1kl} \, \breve{\mathbf{y}}_k \, \mathbf{x}_l' \right] \times \\ &\quad \left[\sum\nolimits_{s_2} \sum\nolimits_{s_2} \hat{\Delta}_{2kl} \, \mathbf{x}_k \, \mathbf{x}_l' - \sum\nolimits_{s_1} \sum\nolimits_{s_1} \hat{\Delta}_{1kl} \, \mathbf{x}_k \, \mathbf{x}_l' \right]^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \left(\sum\nolimits_{s_2} \sum\nolimits_{s_1} \hat{\Delta}_{1kl} \, \breve{\mathbf{y}}_k \, \mathbf{x}_{1l}' \right) \left(\sum\nolimits_{s_1} \sum\nolimits_{s_1} \hat{\Delta}_{1kl} \, \mathbf{x}_{1k} \, \mathbf{x}_{1l}' \right)^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}), \end{aligned} \tag{4.4}$$

où $\hat{\Delta}_{1kl}$ et $\hat{\Delta}_{2kl}$ sont les kl^{e} éléments de $\hat{\boldsymbol{\Delta}}_1$ et $\hat{\boldsymbol{\Delta}}_2$, respectivement. La formule (4.4) est simplifiée dans certains plans à deux phases employés dans d'importantes enquêtes à grande échelle. Des exemples en sont donnés dans Hidiroglou et Särndal (1998) et Turmelle et Beaucage (2013). Plus précisément, c'est le cas quand un échantillonnage indépendant (Poisson, ou Poisson stratifié) est utilisé dans l'une des deux phases, c'est-à-dire quand $\pi_{1kl} = \pi_{1k} \pi_{1l}$ ou $\pi_{2kl} = \pi_{2k} \pi_{2l}$. La simplification est considérable quand les deux phases ont un échantillonnage indépendant. Alors, $\hat{\boldsymbol{\Delta}}_1$ et $\hat{\boldsymbol{\Delta}}_2$ sont toutes deux diagonales, avec des éléments diagonaux $\hat{\Delta}_{1kk} = (1/\pi_{1k})((1/\pi_{1k}) - 1)$ et $\hat{\Delta}_{2kk} = (1/\pi_{1k}\pi_{2k})((1/\pi_{1k}\pi_{2k}) - 1)$, respectivement, et (4.4) comporte uniquement des sommations simples. D'autres plans de sondage à deux phases dans lesquels (4.4) comporte seulement des sommations simples, bien que $\hat{\boldsymbol{\Delta}}_1$ et $\hat{\boldsymbol{\Delta}}_2$ ne soient pas des diagonales, comportent un échantillonnage aléatoire simple ou un échantillonnage aléatoire simple stratifié dans l'une ou l'autre des phases. Voir un exemple d'enquête avec ce type de plan d'échantillonnage à deux phases dans Hidiroglou (2001). Toutefois, en général, l'estimateur optimal peut se révéler peu commode, car il nécessite l'utilisation des probabilités d'inclusion conjointes de la première phase et de la deuxième phase π_{1kl} et π_{2kl}, qui ne sont pas connues pour certains plans de sondage complexes. Même quand ces probabilités conjointes sont connues, mais que les matrices $\hat{\boldsymbol{\Delta}}_1$ et $\hat{\boldsymbol{\Delta}}_2$ ne sont pas diagonales, le coefficient estimé $\hat{\boldsymbol{\mathcal{B}}}$ et, par conséquent, l'estimateur optimal peuvent être instables en cas de très petits échantillons, surtout si la dimension du vecteur auxiliaire $\mathbf{x}$ est grande. Il est possible de surmonter ces difficultés, avec une certaine perte d'optimalité, en utilisant des approximations simples des variances et des covariances dans $\hat{\boldsymbol{\mathcal{B}}}$; pour ce qui est des estimations de la variance approximatives fondées uniquement sur des probabilités d'inclusion du premier ordre, voir par exemple Haziza, Mecatti et Rao (2008) et les références qui s'y trouvent. Une approximation de $\hat{\boldsymbol{\mathcal{B}}}$ très pratique pour le calcul permettant d'obtenir un estimateur à deux phases qui appartient à la classe des estimateurs par la régression généralisée est décrite dans la section suivante.

5. Un estimateur par la régression généralisée en deux phases

Une variante de $\hat{\mathcal{B}} = \breve{\boldsymbol{\Psi}}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{X}}' \hat{\boldsymbol{\Delta}} \boldsymbol{\mathcal{X}})^{-1}$, efficace pour ce qui est du calcul, mais qui n'est généralement pas optimale, est le coefficient de régression généralisée (GREG) $\hat{\mathcal{B}}^{\mathrm{GR}} = \boldsymbol{\Psi}' \boldsymbol{\Lambda} \boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{X}}' \boldsymbol{\Lambda} \boldsymbol{\mathcal{X}})^{-1}$, où $\boldsymbol{\Psi}$ est comme dans (4.1) avec $\mathbf{y}_k = 0$ si $k \notin s_2$, et $\boldsymbol{\Lambda}$ est la matrice de « pondération » $\mathrm{diag}(\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \boldsymbol{\Lambda}_1)$, étant donné que $\boldsymbol{\Lambda}_1 = \mathrm{diag}\{w_{1k} / q_{1k}\}$ et $\boldsymbol{\Lambda}_2 = \mathrm{diag}\{w_k / q_{2k}\}$, et que q_{1k}, q_{2k} sont des constantes positives. Cela donne l'estimateur par la régression généralisée (GREG) :

$$\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{\mathrm{GR}} = \hat{\mathbf{t}}_{\boldsymbol{\Psi}} + \hat{\mathcal{B}}^{\mathrm{GR}} (\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \hat{\mathbf{t}}_{\boldsymbol{\mathcal{X}}}) = \hat{\mathcal{B}}^{\mathrm{GR}} \mathbf{t}_{\boldsymbol{\mathcal{X}}} + (\boldsymbol{\Psi} - \boldsymbol{\mathcal{X}} \hat{\mathcal{B}}^{\mathrm{GR}\prime})' \mathbf{w}^{*}. \tag{5.1}$$

Notons que $\hat{\mathcal{B}}^{\mathrm{GR}}$ est optimal dans le sens des moindres carrés, c'est-à-dire qu'il minimise la distance quadratique $(\boldsymbol{\Psi} - \boldsymbol{\mathcal{X}} \hat{\mathcal{B}}^{\mathrm{GR}\prime})' \boldsymbol{\Lambda} (\boldsymbol{\Psi} - \boldsymbol{\mathcal{X}} \hat{\mathcal{B}}^{\mathrm{GR}\prime})$, impliquant les résidus $\boldsymbol{\Psi} - \boldsymbol{\mathcal{X}} \hat{\mathcal{B}}^{\mathrm{GR}\prime}$ dans $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{\mathrm{GR}}$, alors que le coefficient $\hat{\mathcal{B}}$ minimise $(\boldsymbol{\Psi} - \boldsymbol{\mathcal{X}} \hat{\mathcal{B}})' \hat{\boldsymbol{\Delta}} (\boldsymbol{\Psi} - \boldsymbol{\mathcal{X}} \hat{\mathcal{B}})'$, la variance approximative estimée de l'estimateur optimal $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{O}$. En ce sens, $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{\mathrm{GR}}$ est une approximation de $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{O}$. Les deux composants de $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{\mathrm{GR}}$, similaires dans leur structure aux composantes de $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{O}$ dans (4.2), sont :

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}} &= \tilde{\mathbf{t}}_{\mathbf{y}} + \left[\mathbf{Y}_2' \boldsymbol{\Lambda}_2 \mathbf{X}_2 + \mathbf{Y}_1' \boldsymbol{\Lambda}_1 \mathbf{X}_1\right] \left[\mathbf{X}_2' \boldsymbol{\Lambda}_2 \mathbf{X}_2 + \mathbf{X}_1' \boldsymbol{\Lambda}_1 \mathbf{X}_1\right]^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \mathbf{Y}_1' \boldsymbol{\Lambda}_1 \mathbf{X}_{11} \left[\mathbf{X}_{11}' \boldsymbol{\Lambda}_1 \mathbf{X}_{11}\right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}) \\ \hat{\mathbf{t}}_{\mathbf{x}}^{\mathrm{GR}} &= \hat{\mathbf{t}}_{\mathbf{x}} + \mathbf{X}_1' \boldsymbol{\Lambda}_1 \mathbf{X}_{11} \left[\mathbf{X}_{11}' \boldsymbol{\Lambda}_1 \mathbf{X}_{11}\right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}). \end{aligned} \tag{5.2}$$

L'estimateur GREG $\hat{\mathbf{t}}_{\mathbf{x}}^{\mathrm{GR}}$ est l'estimateur GREG standard à une phase basé sur s_1 et la variable auxiliaire $\mathbf{x}_1$. L'estimateur GREG $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}$, dont les deux termes de régression orthogonale sont indiqués dans (5.2), est exprimé explicitement en termes d'unités d'échantillon comme suit :

$$\begin{aligned} \hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}} &= \tilde{\mathbf{t}}_{\mathbf{y}} + \left[\sum_{s_2} (\Lambda_{1k} + \Lambda_{2k})\, \mathbf{y}_k\, \mathbf{x}_k'\right] \left[\sum_{s_2} \Lambda_{2k}\, \mathbf{x}_k\, \mathbf{x}_k' + \sum_{s_1} \Lambda_{1k}\, \mathbf{x}_k\, \mathbf{x}_k'\right]^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \left(\sum_{s_2} \Lambda_{1k}\, \mathbf{y}_k\, \mathbf{x}_{1k}'\right) \left(\sum_{s_1} \Lambda_{1k}\, \mathbf{x}_{1k}\, \mathbf{x}_{1k}'\right)^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}), \end{aligned}$$

où $\Lambda_{1k} = w_{1k} / q_{1k}$ et $\Lambda_{2k} = w_k / q_{2k}$ sont le k^{e} élément de $\boldsymbol{\Lambda}_1$ et $\boldsymbol{\Lambda}_2$, respectivement. Il faut spécifier les constantes q_{ik} comme étant $q_{ik} = n_i$ pour tenir compte de la différentielle dans la taille d'échantillon de s_i; voir dans Merkouris (2004) une justification dans le contexte d'un calage d'échantillons combinés. Un ajustement équivalent des poids dans Λ_{1k} et Λ_{2k} peut être effectué par la multiplication de w_{1k} dans Λ_{1k} par $\phi = n_2 / n_1$. Les valeurs de q_{ik} qui convertissent l'estimateur GREG $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}$ en l'estimateur optimal $\hat{\mathbf{t}}_{\mathbf{y}}^{O}$ peuvent être spécifiées pour les plans d'échantillonnage à deux phases pour lesquels une estimation optimale est possible, comme dans le contexte similaire de l'échantillonnage matriciel (Merkouris, 2015). Pour l'exemple simple comprenant un échantillonnage de Poisson dans les deux phases, cette spécification est $q_{1k} = \pi_{1k} / (1 - \pi_{1k})$ et $q_{2k} = \pi_{1k}\, \pi_{2k} / (1 - \pi_{1k}\, \pi_{2k})$, rendant $\boldsymbol{\Lambda}_1$ et $\boldsymbol{\Lambda}_2$ identiques à $\hat{\boldsymbol{\Delta}}_1$ et $\hat{\boldsymbol{\Delta}}_2$.

Le vecteur des poids calés associés à l'estimateur GREG $\hat{\mathbf{t}}_{\boldsymbol{\Psi}}^{\mathrm{GR}}$ est $\mathbf{c}^{\mathrm{GR}} = \mathbf{w}^{*} + \boldsymbol{\Lambda} \boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{X}}' \boldsymbol{\Lambda} \boldsymbol{\mathcal{X}})^{-1} (\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}' \mathbf{w}^{*})$. Il a la même forme que $\mathbf{c}^{*}$ dans (4.3), mais avec $\boldsymbol{\Lambda}_1$ et $\boldsymbol{\Lambda}_2$ à la place de $-\boldsymbol{\Lambda}_1$ et $\boldsymbol{\Lambda}_2$, et il minimise la distance des moindres carrés généralisés $(\mathbf{c}^{\mathrm{GR}} - \mathbf{w}^{*})' \boldsymbol{\Lambda}^{-1} (\mathbf{c}^{\mathrm{GR}} - \mathbf{w}^{*})$ soumise

aux contraintes $\boldsymbol{\mathcal{X}}'\mathbf{c}^{\mathrm{GR}} = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$. La partition $\boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{X}}_{12}, \boldsymbol{\mathcal{X}}_1)$, définie après (4.3), permet la décomposition orthogonale du vecteur $\mathbf{c}^*$:

$$\begin{aligned}\mathbf{c}^{\mathrm{GR}} &= \mathbf{w}^* + \boldsymbol{\Lambda}\boldsymbol{\mathcal{X}}_{12}(\boldsymbol{\mathcal{X}}_{12}'\boldsymbol{\Lambda}\boldsymbol{\mathcal{X}}_{12})^{-1}(\mathbf{0} - \boldsymbol{\mathcal{X}}_{12}'\mathbf{w}^*) \\ &\quad + \boldsymbol{\Lambda}\boldsymbol{\mathcal{X}}_1(\boldsymbol{\mathcal{X}}_1'\boldsymbol{\Lambda}\boldsymbol{\mathcal{X}}_1)^{-1}(\mathbf{t}_{\mathbf{x}_1} - \boldsymbol{\mathcal{X}}_1'\mathbf{w}^*).\end{aligned} \tag{5.3}$$

Dans le membre de droite de (5.3), la somme des premier et deuxième termes résulterait d'un calage avec une contrainte $\boldsymbol{\mathcal{X}}_{12}'\mathbf{c}^{\mathrm{GR}} = \mathbf{0}$ seulement, tandis que la somme du premier et du troisième terme résulterait d'un calage avec une contrainte $\boldsymbol{\mathcal{X}}_1'\mathbf{c}^{\mathrm{GR}} = \mathbf{t}_{\mathbf{x}_1}$ seulement. En pratique, cela signifie que le vecteur $\mathbf{c}^*$ pourrait être formé en concaténant les vecteurs de poids générés par deux calages distincts, c'est-à-dire le calage de $(\mathbf{w}_1', \mathbf{w}')'$ au moyen de $(-\mathbf{X}_1', \mathbf{X}_2')'$ suivi du calage de $\mathbf{w}_1$ au moyen de $\mathbf{X}_{11}$. Cependant, la procédure de calage en une étape générant $\mathbf{c}^{\mathrm{GR}}$ est plus pratique.

À partir de sa linéarisation en séries de Taylor, l'estimateur GREG $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}$ dans (5.1) est approximativement (pour les grands échantillons) sans biais. De plus, si l'on désigne par $\mathbf{e}$ la matrice des résidus d'échantillon $\boldsymbol{\Psi} - \boldsymbol{\mathcal{X}}\hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}'}$, la variance approximative estimée de $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}} = \hat{\boldsymbol{\mathcal{B}}}^{\mathrm{GR}}\mathbf{t}_{\boldsymbol{\mathcal{X}}} + \mathbf{e}'\mathbf{w}^*$ est la variance estimée de $\mathbf{e}'\mathbf{w}^*$, c'est-à-dire $\widehat{\mathrm{VA}}(\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}) = \widehat{\mathrm{Var}}(\mathbf{e}'\mathbf{w}^*) = \mathbf{e}'\hat{\boldsymbol{\Delta}}\mathbf{e}$, tandis que la variance estimée de l'estimateur de HT $\tilde{\mathbf{t}}_{\mathbf{y}}$ est $\widehat{\mathrm{Var}}(\tilde{\mathbf{t}}_{\mathbf{y}}) = \widehat{\mathrm{Var}}(\boldsymbol{\Psi}_1'\mathbf{w}^*) = \mathbf{Y}_2'\hat{\boldsymbol{\Delta}}_2\mathbf{Y}_2$, où $\boldsymbol{\Psi}_1$ est la première sous-matrice de colonnes de $\boldsymbol{\Psi}$.

En utilisant maintenant la forme de calage $\boldsymbol{\Psi}_1'\mathbf{c}^{\mathrm{GR}}$ de $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}$ et la décomposition orthogonale (5.3) de $\mathbf{c}^{\mathrm{GR}}$, nous obtenons facilement la décomposition $\mathbf{e} = \boldsymbol{\Psi}_1 - \boldsymbol{\mathcal{X}}_{12}\hat{\boldsymbol{\beta}}_{\mathbf{x}}' - \boldsymbol{\mathcal{X}}_1\hat{\boldsymbol{\beta}}_{\mathbf{x}_1}'$, où $\hat{\boldsymbol{\beta}}_{\mathbf{x}} = \boldsymbol{\Psi}_1'\boldsymbol{\Lambda}\boldsymbol{\mathcal{X}}_{12}(\boldsymbol{\mathcal{X}}_{12}'\boldsymbol{\Lambda}\boldsymbol{\mathcal{X}}_{12})^{-1}$ et $\hat{\boldsymbol{\beta}}_{\mathbf{x}_1} = \boldsymbol{\Psi}_1'\boldsymbol{\Lambda}\boldsymbol{\mathcal{X}}_1(\boldsymbol{\mathcal{X}}_1'\boldsymbol{\Lambda}\boldsymbol{\mathcal{X}}_1)^{-1}$ sont les coefficients de $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ et $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$, respectivement. Mentionnons que $\boldsymbol{\Psi}_1 - \boldsymbol{\mathcal{X}}_{12}\hat{\boldsymbol{\beta}}_{\mathbf{x}}'$ est la matrice des résidus dans l'estimateur GREG $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}|\mathbf{x}} = \tilde{\mathbf{t}}_{\mathbf{y}} + \hat{\boldsymbol{\beta}}_{\mathbf{x}}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}})$ qui résultent du calage comprenant seulement $\boldsymbol{\mathcal{X}}_{12}$, et $\boldsymbol{\Psi}_1 - \boldsymbol{\mathcal{X}}_1\hat{\boldsymbol{\beta}}_{\mathbf{x}_1}'$ est la matrice des résidus dans l'estimateur GREG $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}|\mathbf{x}_1} = \tilde{\mathbf{t}}_{\mathbf{y}} + \hat{\boldsymbol{\beta}}_{\mathbf{x}_1}(\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1})$ qui résulte du calage comprenant seulement $\boldsymbol{\mathcal{X}}_1$. Ensuite, au moyen de l'orthogonalité de $\boldsymbol{\mathcal{X}}_1$ et $\boldsymbol{\mathcal{X}}_{12}$, nous démontrons sans difficulté que :

$$\widehat{\mathrm{VA}}(\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}}) - \widehat{\mathrm{Var}}(\tilde{\mathbf{t}}_{\mathbf{y}}) = \widehat{\mathrm{VA}}(\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}|\mathbf{x}}) - \widehat{\mathrm{Var}}(\tilde{\mathbf{t}}_{\mathbf{y}}) + \widehat{\mathrm{VA}}(\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{GR}|\mathbf{x}_1}) - \widehat{\mathrm{Var}}(\tilde{\mathbf{t}}_{\mathbf{y}}),$$

ce qui implique que la réduction de la variance due à l'utilisation des deux variables auxiliaires $\mathbf{x}_1$ et $\mathbf{x}$ dans la procédure de régression (ou de calage) est additive. Ainsi, si l'on reprend la remarque 3.2, l'estimateur par la régression généralisée retient cette propriété d'additivité du MELSB de $\mathbf{t}_{\mathbf{y}}$.

6. Comparaisons avec les méthodes existantes

Hidiroglou (2001) décrit une méthode antérieure de l'estimation linéaire optimale dans les plans de sondage à deux phases, qui comprend le type standard d'information auxiliaire examiné dans les sections 2 à 5. La formulation commence par poser le postulat d'une forme de régression pour l'estimateur de $\mathbf{t}_{\mathbf{y}}$, pour un y univarié , qui est identique à la forme de l'estimateur $\hat{\mathbf{t}}_{\mathbf{y}}^{B}$ à la première ligne de (3.2);

ensuite, les deux coefficients inconnus sont déterminés de façon à minimiser la variance de cet estimateur. Dans l'estimation des deux coefficients, les identités $\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathrm{Var}(\tilde{\mathbf{t}}_{\mathbf{x}}) - \mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}})$ et $\mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}_1}, \hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) = \mathbf{0}$ n'ont pas été prises en compte dans les premier et deuxième coefficients, respectivement, et les variances et covariances dans les deux coefficients comportant des estimateurs de la première phase ont été estimées au moyen de l'échantillon de la deuxième phase seulement, faisant ainsi abstraction des renseignements pertinents provenant de la plus grande partie de l'échantillon de la première phase. L'estimateur qui en résulte ne s'est pas révélé être un estimateur par calage. De fait, cette version de l'estimateur optimal ne peut pas être construite comme un estimateur par calage. Comme variante pouvant s'appliquer à cela, Hidiroglou (2001) a considéré un estimateur GREG dont les deux coefficients (de $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ et $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$) peuvent être justifiés soit en supposant des modèles de régression différents pour chaque phase, soit par deux calages successifs. Ce même estimateur GREG avait été proposé précédemment par Hidiroglou et Särndal (1998), mais sans référence à l'estimation optimale. Dans l'approche par calage d'Hidiroglou et de Särndal (1998), l'estimateur $\hat{\mathbf{t}}_{\mathbf{x}_1}$ est d'abord calé sur son total $\mathbf{t}_{\mathbf{x}_1}$, au moyen de s_1, et l'estimateur GREG $\hat{\mathbf{t}}^{\mathrm{GR}}_{\mathbf{x}|\mathbf{x}_1}$ de $\mathbf{t}_{\mathbf{x}}$ est ensuite généré au moyen des poids calés représentés par $\tilde{w}_{1k}$. Alors, le poids total de $k \in s_2$ est formé comme étant $\tilde{w}_k = \tilde{w}_{1k}\, w_{2k}$. Dans un deuxième calage comprenant s_2 et $\tilde{w}_k$, l'estimateur $\tilde{\mathbf{t}}_{\mathbf{x}}$ est calé sur $\hat{\mathbf{t}}^{\mathrm{GR}}_{\mathbf{x}|\mathbf{x}_1}$. Les poids calés de s_2 qui en résultent servent ensuite à générer l'estimateur GREG de $\mathbf{t}_{\mathbf{y}}$, désigné ici par $\hat{\mathbf{t}}^{\mathrm{HS}}_{\mathbf{y}}$.

Estevao et Särndal (2002, 2009) ont proposé une version plus simple de l'estimateur $\hat{\mathbf{t}}^{\mathrm{HS}}_{\mathbf{y}}$, dans laquelle les poids de sondage globaux $w_k = w_{1k}\, w_{2k}$ pour $k \in s_2$ sont utilisés dans le deuxième calage. Au moyen de la notation actuelle, cet estimateur, représenté ici par $\hat{\mathbf{t}}^{\mathrm{ES}}_{\mathbf{y}}$, peut être exprimé sous forme de régression comme suit :

$$\begin{aligned}\hat{\mathbf{t}}^{\mathrm{ES}}_{\mathbf{y}} &= \tilde{\mathbf{t}}_{\mathbf{y}} + \mathbf{Y}_2' \boldsymbol{\Lambda}_2 \mathbf{X}_2 \left(\mathbf{X}_2' \boldsymbol{\Lambda}_2 \mathbf{X}_2\right)^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \mathbf{Y}_2' \boldsymbol{\Lambda}_2 \mathbf{X}_2 \left(\mathbf{X}_2' \boldsymbol{\Lambda}_2 \mathbf{X}_2\right)^{-1} \mathbf{X}_1' \boldsymbol{\Lambda}_1 \mathbf{X}_{11} \left[\mathbf{X}_{11}' \boldsymbol{\Lambda}_1 \mathbf{X}_{11}\right]^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}).\end{aligned} \tag{6.1}$$

Ici, les matrices de pondération standards $\boldsymbol{\Lambda}_1 = \mathrm{diag}\{w_{1k}\}$ et $\boldsymbol{\Lambda}_2 = \mathrm{diag}\{w_k\}$ sont utilisées. Estevao et Särndal (2009) ont démontré que cet estimateur est asymptotiquement équivalent à l'estimateur $\hat{\mathbf{t}}^{\mathrm{HS}}_{\mathbf{y}}$. Pour l'estimateur $\hat{\mathbf{t}}^{\mathrm{ES}}_{\mathbf{y}}$ figurant dans (6.1), Estevao et Särndal (2002) fournissent deux représentations de régression linéaire correspondant aux deux étapes de calage. En remplaçant $\mathbf{y}$ par $\mathbf{x}$ dans (6.1), nous obtenons $\hat{\mathbf{t}}^{\mathrm{ES}}_{\mathbf{x}}$, qui est identique à $\hat{\mathbf{t}}^{\mathrm{GR}}_{\mathbf{x}}$ dans (5.2).

À titre de comparaison, l'estimateur par la régression proposé dans la section 5 est motivé par la structure de calage à une seule étape de l'estimateur optimal à deux phases, dont il est une approximation utilisée à des fins de commodité. Comparativement aux estimateurs par la régression concurrents évalués dans la présente section, il tire son efficacité statistique et computationnelle d'une procédure de calage à une seule étape comprenant les échantillons combinés de la première phase et de la deuxième phase, et dans laquelle les totaux estimés de la première phase et de la deuxième phase sont calés les uns sur les autres. Par conséquent, les coefficients de régression des termes $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ et $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ intègrent l'information de l'échantillon complet s_1, comme dans l'estimateur optimal, et pour cette raison, ils sont des

estimations plus stables de leurs contreparties de population. Une comparaison empirique de l'estimateur par la régression proposé avec les estimateurs par la régression concurrents est présentée dans l'étude par simulations à la section 7.

Le remplacement de $\boldsymbol{\Lambda}_1$ et $\boldsymbol{\Lambda}_2$ par $\hat{\boldsymbol{\Delta}}_1$ et $\hat{\boldsymbol{\Delta}}_2$ dans (6.1) convertit le coefficient de l'estimateur GREG $\hat{\mathbf{t}}_{\mathbf{x}|\mathbf{x}_1}^{\mathrm{GR}}$ généré par le calage de la première étape en coefficient $\widehat{\mathrm{Cov}}(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1})[\widehat{\mathrm{Var}}(\hat{\mathbf{t}}_{\mathbf{x}_1})]^{-1}$ de l'estimateur par la régression optimal à une seule phase $\hat{\mathbf{t}}_{\mathbf{x}|\mathbf{x}_1}^{O}$, et le coefficient de l'estimateur GREG $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{ES}}$ généré par le calage de la deuxième étape en coefficient $\widehat{\mathrm{Cov}}(\tilde{\mathbf{t}}_{\mathbf{y}}, \tilde{\mathbf{t}}_{\mathbf{x}})[\widehat{\mathrm{Var}}(\tilde{\mathbf{t}}_{\mathbf{x}})]^{-1}$. Ce dernier coefficient peut être considéré comme pseudo-optimal puisque $\hat{\mathbf{t}}_{\mathbf{x}|\mathbf{x}_1}^{O}$ est traité comme une constante dans le calage de la deuxième étape, ce qui génère un estimateur pseudo-optimal $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSO}}$. Ensuite, si à la place des matrices d'échantillon $\hat{\boldsymbol{\Delta}}_1$ et $\hat{\boldsymbol{\Delta}}_2$ dans (6.1), nous utilisons les matrices de population $\boldsymbol{\Delta}_1$ et $\boldsymbol{\Delta}_2$, nous construisons le pseudo-MELSB :

$$\begin{aligned}\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSB}} &= \tilde{\mathbf{t}}_{\mathbf{y}} + \mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U \left(\mathbf{X}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U\right)^{-1} (\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}) \\ &\quad + \mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U \left(\mathbf{X}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U\right)^{-1} \mathbf{X}_U' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} \left(\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}\right)^{-1} (\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}),\end{aligned} \tag{6.2}$$

où les coefficients de $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ et $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ sont respectivement $\mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \tilde{\mathbf{t}}_{\mathbf{x}})\left[\mathrm{Var}(\tilde{\mathbf{t}}_{\mathbf{x}})\right]^{-1}$ et $\mathrm{Cov}(\tilde{\mathbf{t}}_{\mathbf{y}}, \tilde{\mathbf{t}}_{\mathbf{x}})\left[\mathrm{Var}(\tilde{\mathbf{t}}_{\mathbf{x}})\right]^{-1}\mathrm{Cov}(\hat{\mathbf{t}}_{\mathbf{x}}, \hat{\mathbf{t}}_{\mathbf{x}_1})\left[\mathrm{Var}(\hat{\mathbf{t}}_{\mathbf{x}_1})\right]^{-1}$. Ainsi, l'estimateur GREG (6.1) peut être considéré comme une approximation de $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSO}}$, qui est l'estimateur $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSB}}$ avec des coefficients estimés (par analogie avec la relation entre l'estimateur optimal $\hat{\mathbf{t}}_{\mathbf{y}}^{O}$ et le MELSB $\hat{\mathbf{t}}_{\mathbf{y}}^{B}$, dans (4.2) et (3.14)). L'estimateur pseudo-MELSB $\hat{\mathbf{t}}_{\mathbf{x}}^{\mathrm{PSB}}$, obtenu à partir de (6.2) en remplaçant $\mathbf{y}$ par $\mathbf{x}$, est identique au MELSB $\hat{\mathbf{t}}_{\mathbf{x}}^{B}$, dans (3.14). Par ailleurs, les estimateurs $\hat{\mathbf{t}}_{\mathbf{y}}^{B}$ et $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSB}}$ sont identiques seulement à condition de respecter la proposition suivante (voir la démonstration en annexe).

Proposition 1. Les estimateurs $\hat{\mathbf{t}}_{\mathbf{y}}^{B}$ et $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSB}}$ sont identiques si et seulement si $\boldsymbol{\Delta}_1 = \delta\boldsymbol{\Delta}_2$ pour une constante δ.

Remarque 6.1. La condition de la proposition 1 se vérifie si le même plan avec probabilités d'inclusion égales est utilisé dans les deux phases; la constante δ est alors une fonction des probabilités d'inclusion de l'échantillon. Les plans à deux phases qui satisfont à cette condition sont l'échantillonnage aléatoire simple et l'échantillonnage de Bernoulli dans les deux phases, ainsi que leurs versions de stratification avec stratification identique et répartition proportionnelle de l'échantillon dans les deux phases. Cette condition a une importance pratique : pour ces plans, les contreparties d'échantillon de $\hat{\mathbf{t}}_{\mathbf{y}}^{B}$ et $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSB}}$, c'est-à-dire $\hat{\mathbf{t}}_{\mathbf{y}}^{O}$ et $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSO}}$, seront presque identiques en cas de grands échantillons. De plus, $\boldsymbol{\Delta}_1 = \delta\boldsymbol{\Delta}_2$ laisse supposer que le signe moins dans le coefficient de $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ dans (3.14) et (4.2) pourrait devenir un signe plus, $1-\delta$ étant exclus, ce qui annulerait le problème de singularité mentionné dans la remarque 4.1.

Remarque 6.2. On peut vérifier simplement que $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSO}}$ est un estimateur par calage, construit par une procédure de calage en deux étapes (comme pour le $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{ES}}$ estimateur). De plus, comme $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{ES}}$, l'estimateur $\hat{\mathbf{t}}_{\mathbf{y}}^{\mathrm{PSO}}$ est formé au moyen des poids calés de l'échantillon de la deuxième phase seulement.

Enfin, il faut mentionner que Chen et Kim (2014) ont proposé un « estimateur par la régression optimal par rapport au plan », sans la propriété de calage, pour une application et une configuration de variable auxiliaire particulières. De plus, Wu et Luan (2003) ont proposé un estimateur par calage qui est optimal dans un cadre assisté par modèle et qui a $\mathbf{x}_2$ comme seul vecteur auxiliaire.

7. Étude par simulations

Nous avons réalisé une étude par simulations pour évaluer les performances des estimateurs à deux phases du total t_y proposés, pour les variables scalaires y, x_1 et x_2, et les comparer avec les estimateurs par régression concurrents examinés ci-dessus. Les distributions de ces variables ont été spécifiées comme suit. La distribution de x_1 est une loi log-normale comportant les paramètres de moyenne et de variance $(\mu_{x_1} = 4, \sigma^2_{x_1} = 4)$. La distribution de x_2 est spécifiée par le modèle linéaire $x_2 = 5 + x_1 + \epsilon$, où $\epsilon \sim N(0, \sigma^2_\epsilon)$, et la distribution de y est spécifiée par le modèle linéaire $y = 10 + 2x_1 + 3x_2 + \eta$, où $\eta \sim N(0, \sigma^2_\eta)$. La valeur de σ^2_ϵ détermine la relation linéaire entre x_2 et x_1, telle qu'elle est définie par le carré du coefficient de corrélation de la population $r^2_{x_1, x_2}$, et la valeur de σ^2_η détermine la relation linéaire de y avec x_1 et x_2, telle qu'elle est définie par le coefficient de détermination $r^2 = [r^2_{y,x_1} + r^2_{y,x_2} - 2r_{y,x_1} r_{y,x_2} r_{x_1,x_2}] / (1 - r^2_{x_1,x_2})$.

Trois valeurs, 0, 0,25 et 0,75 ont été précisées pour $r^2_{x_1,x_2}$, et deux valeurs, 0,25 et 0,75, pour r^2, ce qui donne six combinaisons de valeurs $(r^2_{x_1,x_2}, r^2)$. Pour la valeur $r^2_{x_1,x_2} = 0$, en particulier, la distribution log-normale bivariée pour (x_1, x_2) avec les paramètres $(\mu_{x_1} = 4, \sigma^2_{x_1} = 4)$, $(\mu_{x_2} = 9, \sigma^2_{x_2} = 9)$ et une corrélation nulle a été utilisée. Les valeurs requises de σ^2_ϵ et σ^2_η sont facilement déterminées, tandis que les valeurs pour r^2_{y,x_1} et r^2_{y,x_2} sont implicitement précisées. Pour chacune de ces six combinaisons, nous avons simulé une population de taille $N =$ 50 000 en générant des valeurs à partir des distributions des composantes du vecteur (y, x_1, x_2). Quatre combinaisons de tailles d'échantillon de la première phase et de la deuxième phase (n_1, n_2) comportant une valeur n_1 fixe et une valeur n_2 variable ont été spécifiées, à savoir (3 000; 2 000), (3 000; 1 500), (3 000; 1 000), (3 000; 500), ce qui a créé un total de 24 configurations de simulation.

Trois plans de sondage à deux phases différents ont été examinés. Un échantillonnage aléatoire simple (EAS) sans remise a d'abord été utilisé dans les deux phases. Pour ce plan de sondage, désigné par (EAS, EAS), le MELSB $\hat{t}^B_y$ dans (3.10) et sa variance exacte dans (3.11) peuvent être calculés. En s'appuyant sur le fait que dans un EAS, la corrélation des estimateurs de HT pour deux totaux est identique au coefficient de corrélation des variables associées, des calculs algébriques fastidieux mais simples donnent la différence relative des variances (DRV) des estimateurs $\hat{t}^B_y$ et $\tilde{t}_y$ comme étant :

$$\frac{\mathrm{Var}(\tilde{t}_y) - \mathrm{Var}(\hat{t}^B_y)}{\mathrm{Var}(\tilde{t}_y)} = \frac{N(n_1 - n_2)}{n_1(N - n_2)} r^2 + \frac{n_2(N - n_1)}{n_1(N - n_2)} r^2_{y,x_1}.$$

La DRV en pourcentage est la mesure de l'efficacité du MELSB $\hat{t}^B_y$ par rapport à l'estimateur de HT $\tilde{t}_y$. Cette efficacité maximale exacte servira à mesurer le degré de correspondance de l'approximation du

MELSB par l'estimateur optimal, pour les différentes tailles d'échantillon, ainsi que l'efficacité des autres estimateurs concurrents par rapport à l'estimateur de HT. Nous constatons que quand n_2 tend vers zéro, la DRV tend vers r^2 et que quand n_2 tend vers n_1, la DRV tend vers r^2_{y,x_1} (l'efficacité du MELSB basé sur s_1 et x_1). Le deuxième plan à deux phases, représenté par (EASS, EAS), était un échantillonnage aléatoire simple stratifié (EASS) et un EAS à la première et à la deuxième phase, respectivement. Les populations simulées étaient stratifiées selon la taille de la variable y, avec trois strates de tailles $N_1 = 30\,000$, $N_2 = 15\,000$, $N_3 = 5\,000$ et une répartition proportionnelle de l'échantillon s_1 aux trois strates, ce qui donne des probabilités d'inclusion égales dans chacune des deux phases. Pour ce plan également, il est possible de calculer le MELSB $\hat{t}_y^B$ et sa variance exacte. Le troisième plan à deux phases, représenté par (EAS, ESPPT), comportait un EAS dans la première phase et un échantillonnage systématique avec probabilité proportionnelle à la taille (ESPPT) dans la deuxième phase, où la mesure de taille utilisée était la transformation simple $z_2 = 15 + 0{,}5x_2$ de la variable x_2; si x_2 était utilisé comme taille, nous obtiendrions $\hat{t}_{x_2} = \tilde{t}_{x_2}$. Dans ce cas, il est impossible de calculer le MELSB $\hat{t}_y^B$ (et l'estimateur optimal $\hat{t}_y^O$), en raison des probabilités inconnues π_{2kl}. Toutefois, les estimateurs GREG peuvent être calculés.

Pour chacun de ces trois plans à deux phases, et pour les 24 configurations de simulation, l'échantillonnage a été répété 30 000 fois; à chacune des répétitions, nous avons calculé les estimateurs $\tilde{t}_y$, $\hat{t}_y^O$, $\hat{t}_y^{\mathrm{GR}}$, $\hat{t}_y^{\mathrm{ES}}$ et $\hat{t}_y^{\mathrm{HS}}$ pour obtenir leur biais empirique et leur variance. Dans tous ces cas, la simulation a montré que le biais de tous les estimateurs était négligeable, y compris pour les tailles de sous-échantillons n_2 les plus petites. C'est pourquoi leur comparaison est fondée sur leurs variances par rapport à la variance de référence de l'estimateur de HT $\tilde{t}_y$. Plus précisément, l'efficacité de chacun des estimateurs concurrents $\hat{t}_y^O$, $\hat{t}_y^{\mathrm{GR}}$, $\hat{t}_y^{\mathrm{ES}}$ et $\hat{t}_y^{\mathrm{HS}}$ est évaluée par la différence relative en pourcentage entre sa variance empirique et la variance empirique de l'estimateur $\tilde{t}_y$; par exemple, pour $\hat{t}_y^{\mathrm{GR}}$ la différence relative est $\left[\mathrm{Var}(\tilde{t}_y) - \mathrm{Var}(\hat{t}_y^{\mathrm{GR}})\right] \big/ \mathrm{Var}(\tilde{t}_y)$. La différence relative montre la réduction de la variance de l'estimateur particulier par rapport à la variance de l'estimateur de base $\tilde{t}_y$.

Dans le plan (EAS, EAS), l'efficacité exacte du MELSB $\hat{t}_y^B$ par rapport à l'estimateur de HT $\tilde{t}_y$ augmente à mesure que n_2 diminue et que nous passons à des valeurs plus élevées de (r^2_{y,x_2}, r^2), ce qui corrobore l'information fournie dans la remarque 3.1; voir la colonne 2 du tableau 7.1. Il est aussi confirmé que l'efficacité de $\hat{t}_y^B$ tend vers r^2 quand n_2 diminue, plus rapidement pour une valeur de $r^2_{x_1,x_2}$ plus élevée. Une approximation proche de cette efficacité maximale est donnée par l'efficacité empirique de $\hat{t}_y^O$, même pour les tailles de sous-échantillons les plus petites n_2; voir la colonne 3 du tableau 7.1. Pour le plan (EASS, EAS), l'efficacité exacte de $\hat{t}_y^B$, indiquée dans la colonne 6 du tableau 7.1, présente une tendance semblable à celle observée dans le plan (EAS, EAS). Une approximation proche de cette efficacité est donnée par l'efficacité empirique de $\hat{t}_y^O$; voir la colonne 7 du tableau 7.1. Dans les deux plans (EAS, EAS) et (EASS, EAS), l'approximation de $\hat{t}_y^B$ par $\hat{t}_y^O$ est un peu plus faible dans certaines configurations comprenant la plus grande valeur de n_2, pour la raison donnée dans la remarque 4.1.

Bien que l'estimateur $\hat{t}_y^O$ puisse être calculé dans les plans (EAS, EAS) et (EASS, EAS), les performances des estimateurs par calage (GREG) $\hat{t}_y^{\mathrm{GR}}$ et $\hat{t}_y^{\mathrm{ES}}$, plus pratiques et d'application générale, présentent un grand intérêt. Pour (EAS, EAS), les efficacités empiriques de ces estimateurs sont indiquées

dans les colonnes 4 et 5 du tableau 7.1. Le signe négatif indique une perte d'efficacité par rapport à l'estimateur de HT. L'efficacité de $\hat{t}_y^{\mathrm{GR}}$ est une approximation proche de l'efficacité de $\hat{t}_y^{O}$, sauf pour les quatre configurations indiquées par $r^2_{x_1,x_2} = 0$, $r^2 = 0{,}25$; 0,75 et $n_2 = 2\,000$; 1 500; plus particulièrement, quand $n_2 = 2\,000$ l'estimateur $\hat{t}_y^{\mathrm{GR}}$ est légèrement moins efficace que l'estimateur $\tilde{t}_y$. En revanche, l'estimateur $\hat{t}_y^{\mathrm{ES}}$ est moins efficace que l'estimateur $\tilde{t}_y$ dans six configurations, quand $r^2_{x_1,x_2} = 0$, $r^2 =$ 0,25; 0,75 et $n_2 = 2\,000$; 1 500; 1 000 et beaucoup moins efficace quand $n_2 = 2\,000$; 1 500. Le point important dans les colonnes 4 et 5 est que l'estimateur $\hat{t}_y^{\mathrm{GR}}$ est nettement plus efficace que l'estimateur $\hat{t}_y^{\mathrm{ES}}$ dans toutes les configurations, et encore plus pour les valeurs les plus élevées de n_2 et les valeurs les plus élevées de (r^2_{y,x_2}, r^2); cela indique que $\hat{t}_y^{\mathrm{GR}}$ est plus efficace lorsqu'il s'agit d'utiliser l'information tirée du complément de s_2 et d'exploiter les corrélations élevées de y avec x_1 et x_2. Étant donné que l'efficacité de l'estimateur $\hat{t}_y^{\mathrm{HS}}$ était pratiquement identique à celle de $\hat{t}_y^{\mathrm{ES}}$ dans les trois plans, elle n'est pas indiquée dans le tableau 7.1. Pour (EASS, EAS), les efficacités empiriques des estimateurs par calage $\hat{t}_y^{\mathrm{GR}}$ et $\hat{t}_y^{\mathrm{ES}}$ sont indiquées dans les colonnes 8 et 9 du tableau 7.1. Il convient de mentionner que les corrélations à l'intérieur des strates sont beaucoup plus faibles que les corrélations pour l'ensemble de la population (voir le tableau 7.1). De plus, l'estimateur de HT $\tilde{t}_y$ est très efficace en raison de la stratification, surtout pour les plus grandes valeurs de n_2. L'estimateur $\hat{t}_y^{\mathrm{GR}}$ est moins efficace que l'estimateur $\tilde{t}_y$ dans 3 des 24 configurations (qui comprennent $n_2 = 2\,000$), alors que pour les autres, son efficacité augmente considérablement à mesure que n_2 diminue, se rapprochant de l'efficacité de $\hat{t}_y^{O}$. L'estimateur $\hat{t}_y^{\mathrm{ES}}$ est moins efficace que l'estimateur $\tilde{t}_y$ dans 12 configurations. L'estimateur $\hat{t}_y^{\mathrm{GR}}$ est beaucoup plus efficace que l'estimateur $\hat{t}_y^{\mathrm{ES}}$ dans toutes les configurations, et encore plus pour les valeurs les plus élevées de n_2 et à mesure que nous passons de $r^2 = 0{,}25$ à $r^2 = 0{,}75$, et considérablement plus que dans le plan d'échantillonnage (EAS, EAS).

Tableau 7.1
Pourcentage d'efficacité de $\hat{t}_y^{B}$, $\hat{t}_y^{O}$, $\hat{t}_y^{\mathrm{GR}}$, $\hat{t}_y^{\mathrm{ES}}$ par rapport à $\tilde{t}_y$

	(EAS, EAS)				(EASS, EAS)				(EAS, ESPPT)	
n_2	$\hat{t}_y^{B}$	$\hat{t}_y^{O}$	$\hat{t}_y^{\mathrm{GR}}$	$\hat{t}_y^{\mathrm{ES}}$	$\hat{t}_y^{B}$	$\hat{t}_y^{O}$	$\hat{t}_y^{\mathrm{GR}}$	$\hat{t}_y^{\mathrm{ES}}$	$\hat{t}_y^{\mathrm{GR}}$	$\hat{t}_y^{\mathrm{ES}}$
	$\sigma^2_\eta = 292{,}41$; $r^2_{x_1,x_2} = 0{,}00$; $r^2_{y,x_1} = 0{,}04$; $r^2_{y,x_2} = 0{,}21$; $r^2 = 0{,}25$									
2 000	11,54	10,09	-1,66	-26,88	16,74	13,69	-23,49	-79,94	-5,51	-30,38
1 500	15,00	13,84	10,91	-13,61	20,01	17,80	7,22	-38,46	4,57	-20,69
1 000	18,41	17,68	17,79	-0,71	22,22	21,20	19,34	-11,76	9,69	-10,22
500	21,74	20,62	20,77	11,29	23,81	22,34	22,75	7,84	11,24	0,67
	$\sigma^2_\eta = 32{,}15$; $r^2_{x_1,x_2} = 0{,}00$; $r^2_{y,x_1} = 0{,}13$; $r^2_{y,x_2} = 0{,}62$; $r^2 = 0{,}75$									
2 000	34,35	31,21	-4,23	-80,22	52,31	48,02	-66,06	-232,15	-21,68	-107,41
1 500	44,84	42,34	33,09	-41,49	61,53	59,02	24,66	-108,30	15,95	-74,74
1 000	55,10	53,80	53,83	-2,25	67,58	65,48	59,17	-30,84	36,49	-39,03
500	65,16	63,87	63,90	34,31	71,85	70,53	70,41	27,83	45,55	2,00
	$\sigma^2_\epsilon = 12{,}11$; $\sigma^2_\eta = 632{,}52$; $r^2_{x_1,x_2} = 0{,}25$; $r^2_{y,x_1} = 0{,}12$; $r^2_{y,x_2} = 0{,}24$; $r^2 = 0{,}25$									
2 000	16,70	16,89	16,69	9,88	17,14	15,56	2,08	-23,18	10,24	3,03
1 500	18,85	19,02	19,52	13,57	20,26	19,45	16,46	-3,21	13,83	7,08
1 000	20,97	20,79	20,52	16,50	22,38	21,07	21,04	6,84	13,03	8,04
500	23,04	22,28	21,67	19,64	23,91	22,84	23,41	16,40	11,57	9,38

EAS = échantillonnage aléatoire simple; EASS = échantillonnage aléatoire simple stratifié;
ESPPT = échantillonnage systématique avec probabilité proportionnelle à la taille.

Tableau 7.1(suite)
Pourcentage d'efficacité de $\hat{t}_y^B$, $\hat{t}_y^O$, $\hat{t}_y^{\mathrm{GR}}$, $\hat{t}_y^{\mathrm{ES}}$ par rapport à $\tilde{t}_y$

	(EAS, EAS)				(EASS, EAS)				(EAS, ESPPT)	
n_2	$\hat{t}_y^B$	$\hat{t}_y^O$	$\hat{t}_y^{\mathrm{GR}}$	$\hat{t}_y^{\mathrm{ES}}$	$\hat{t}_y^B$	$\hat{t}_y^O$	$\hat{t}_y^{\mathrm{GR}}$	$\hat{t}_y^{\mathrm{ES}}$	$\hat{t}_y^{\mathrm{GR}}$	$\hat{t}_y^{\mathrm{ES}}$
σ_ϵ^2 =12,11; σ_η^2 = 70,68; $r^2_{x_1,x_2}$ = 0,25; r^2_{y,x_1} = 0,36; r^2_{y,x_2} = 0,71; r^2 = 0,75										
2 000	49,70	48,33	46,89	25,33	53,11	50,78	20,11	-38,15	35,49	11,64
1 500	56,23	55,48	56,46	38,08	61,86	60,63	53,81	8,36	46,33	22,98
1 000	62,63	62,20	61,35	48,69	67,71	66,38	66,10	34,90	47,91	30,19
500	68,90	68,09	65,58	59,65	71,90	70,99	71,00	55,27	47,68	38,93
σ_ϵ^2 =1,33; σ_η^2 = 340,40; $r^2_{x_1,x_2}$ = 0,75; r^2_{y,x_1} = 0,22; r^2_{y,x_2} = 0,24; r^2 = 0,25										
2 000	23,36	23,67	23,01	10,81	18,09	15,47	-6,07	-46,54	16,62	3,38
1 500	23,78	23,63	24,19	13,85	20,83	19,60	14,52	-17,17	19,77	7,95
1 000	24,20	23,83	23,15	16,97	22,68	21,67	21,58	0,86	17,04	10,02
500	24,61	23,54	22,24	19,92	24,01	22,39	22,98	13,24	14,52	11,77
σ_ϵ^2 =1,33; σ_η^2 = 37,82; $r^2_{x_1,x_2}$ = 0,75; r^2_{y,x_1} = 0,67; r^2_{y,x_2} = 0,72; r^2 = 0,75										
2 000	69,84	67,98	65,10	26,96	60,26	56,75	32,34	-27,50	56,65	13,24
1 500	71,17	69,57	70,70	38,91	66,25	64,49	59,73	14,39	65,49	26,11
1 000	72,47	71,26	69,17	49,62	70,17	68,80	68,69	40,44	61,00	35,28
500	73,74	72,19	67,58	60,66	72,94	71,12	71,10	56,90	54,67	44,54

EAS = échantillonnage aléatoire simple; EASS = échantillonnage aléatoire simple stratifié;
ESPPT = échantillonnage systématique avec probabilité proportionnelle à la taille.

Pour (EAS, ESPPT), les efficacités empiriques des estimateurs par calage $\hat{t}_y^{\mathrm{GR}}$ et $\hat{t}_y^{\mathrm{ES}}$ sont présentées dans les colonnes 10 et 11 du tableau 7.1. La tendance de ces efficacités est très semblable à celle du plan (EAS, EAS). C'est particulièrement le cas pour l'efficacité de $\hat{t}_y^{\mathrm{GR}}$ par rapport à $\hat{t}_y^{\mathrm{ES}}$, qui n'est pas présentée dans le tableau 7.1, mais qui peut se calculer facilement au moyen des efficacités de $\hat{t}_y^{\mathrm{GR}}$ et $\hat{t}_y^{\mathrm{ES}}$ par rapport à $\tilde{t}_y$ indiquées. L'estimateur de HT $\tilde{t}_y$ lui-même est plus efficace dans ce plan de sondage à deux phases, ce qui explique que l'efficacité des deux estimateurs par calage $\hat{t}_y^{\mathrm{GR}}$ et $\hat{t}_y^{\mathrm{ES}}$ par rapport à $\tilde{t}_y$ est quelque peu inférieure à celle observée dans les plans (EAS, EAS) et (EASS, EAS).

Toute l'étude par simulations a été répétée dans la population simulée pour le vecteur (y, x_1, x_2) généré à partir d'une loi log-normale trivariée comportant les structures de corrélation spécifiées. Pour les trois plans (EAS, EAS), (EAS, ESPPT) et (EASS, EAS), les résultats (non présentés ici) étaient très semblables à ceux basés sur le modèle linéaire pour y utilisé précédemment.

Il est intéressant d'examiner la configuration des variables auxiliaires dans laquelle la variable scalaire x_1 est augmentée pour s'établir à $(1, x_1)$, avec des totaux connus (N, t_{x_1}). Ensuite, dans le plan de sondage (EAS, EAS), dans lequel la construction du MELSB $\hat{t}_y^B$ et de l'estimateur optimal $\hat{t}_y^O$ est réalisable, l'utilisation de la configuration complète $(1, x_1, x_2)$ dans le calage donne le même $\hat{t}_y^B$ et pratiquement le même $\hat{t}_y^B$ que l'utilisation de (x_1, x_2). Cela convertirait également l'estimateur par la régression $\hat{t}_y^{\mathrm{GR}}$ en $\hat{t}_y^O$ (quand nous utilisons le même ajustement $1/\pi_{2k}$ de $\mathbf{Y}_1$ comme dans $\hat{t}_y^O$), et l'estimateur par la régression $\hat{t}_y^{\mathrm{ES}}$ en estimateur pseudo-optimal $\hat{t}_y^{\mathrm{PSO}}$ (défini à la section 6). Ces propriétés sont calculées à partir d'une théorie connue, voir par exemple Merkouris (2004, 2015), plus directement pour $\hat{t}_y^{\mathrm{ES}}$ et le second terme de la régression de $\hat{t}_y^{\mathrm{GR}}$ et l'estimateur optimal $\hat{t}_y^O$, indépendamment de toute relation fonctionnelle précise de y avec $(1, x_1, x_2)$. Les trois estimateurs fondés sur l'échantillon

présenteraient alors un comportement empirique pratiquement identique. Cela est conforme à la proposition 1, qui donne la condition (satisfaite par des plans spécifiques, notamment (EAS, EAS)) selon laquelle l'estimateur par la régression pseudo-optimal $\hat{t}_y^{\mathrm{PSO}}$ est asymptotiquement équivalent à l'estimateur optimal proposé $\hat{t}_y^O$. Des calculs expérimentaux ont confirmé cette équivalence. Dans le plan de sondage (EASS, EAS), l'utilisation de $(1, x_1, x_2)$ donne les mêmes $\hat{t}_y^B$ et $\hat{t}_y^O$ que l'utilisation de (x_1, x_2), et convertit les estimateurs $\hat{t}_y^{\mathrm{GR}}$ et $\hat{t}_y^{\mathrm{ES}}$ respectivement en estimateurs $\hat{t}_y^O$ et $\hat{t}_y^{\mathrm{PSO}}$. Cependant, selon la proposition 1, l'équivalence des deux derniers estimateurs, et par conséquent de $\hat{t}_y^{\mathrm{GR}}$ et $\hat{t}_y^{\mathrm{ES}}$, ne se vérifie pas dans ce plan de sondage.

8. Discussion

La méthode décrite d'estimation optimale et par la régression pour un échantillonnage à deux phases comprend un calage en une étape des poids des échantillons combinés de la première et de la deuxième phase. il est ainsi possible d'obtenir, au moyen d'un seul ensemble de poids calés qui comprend toute l'information disponible des deux phases, une estimation grandement améliorée de la valeur totale d'une variable cible, comme le démontre l'étude par simulations. Ces poids pourraient servir à calculer d'autres statistiques pondérées, y compris les moyennes, les ratios, les quantiles et les coefficients de régression. Le cadre de la méthode est suffisamment général pour englober des plans complexes à plusieurs degrés et une stratification différente aux deux phases, ainsi que différents types de variables auxiliaires connues au niveau de la population ou de l'échantillon; Estevao et Särndal (2002) définissent ainsi 10 cas différents d'information auxiliaire. De plus, la méthode peut être étendue aux plans de sondage à phases multiples au moyen d'une configuration de calage appropriée.

L'estimation d'un total pour tout domaine (sous-population) d'intérêt $U_d \subset U$ peut facilement être réalisée au moyen des poids calés et en additionnant les valeurs d'échantillon pondérées de la variable d'intérêt sur U_d. Pour que l'estimateur de domaine qui en résulte soit l'estimateur linéaire optimal, il faut combiner linéairement les estimations pour un domaine de $\mathbf{t}_{\mathbf{y}}$, $\mathbf{t}_{\mathbf{x}_1}$ et $\mathbf{t}_{\mathbf{x}_2}$ en effectuant un calage optimal au niveau du domaine avec les totaux de calage de domaine et avec la modification appropriée de la matrice $\boldsymbol{\mathcal{X}}$. Plusieurs options de calage, concernant l'utilisation de l'information auxiliaire disponible au niveau de la population, du domaine et de l'échantillon à deux phases, peuvent être prises en compte pour déterminer l'estimation la plus efficace des totaux de domaine dans une application donnée. Les travaux de Merkouris (2010) à ce propos seraient utiles dans ce contexte.

Les variances approximatives estimées de l'estimateur optimal à deux phases et de l'estimateur par la régression à deux phases, fondées sur la linéarisation en séries de Taylor, ont été données aux sections 4.1 et 5, respectivement. Pour l'estimateur par la régression à deux phases, les méthodes de rééchantillonnage d'estimation de la variance, comme la méthode du jackknife ou la méthode bootstrap, pourraient être appliquées autrement ou elles seraient la seule possibilité quand les probabilités d'inclusion de la première ou de la deuxième phase ne sont pas connues. Une littérature abondante traite de ces méthodes de

rééchantillonnage appliquées aux estimateurs par la régression existants dans un échantillonnage à deux phases. À ces fins, la fonction de calage en une seule étape de la méthode d'estimation par la régression proposée peut se révéler utile. Cette question, qui dépasse la portée du présent article, mériterait d'être approfondie.

Remerciements

L'auteur remercie le rédacteur en chef, le rédacteur associé et les deux arbitres pour leurs commentaires et suggestions qui ont mené à une amélioration importante de l'article.

Annexe

Démonstration du lemme 1

La matrice symétrique $\text{Var}(\mathbf{w}_U^*)$ a la forme de (3.8), mais elle a $\text{Cov}(\mathbf{w}_{1U}, \mathbf{w}_U)$ comme bloc hors diagonale. Le kl^{e} élément de la matrice $\text{Var}(\mathbf{w}_{1U})$ est :

$$\text{Cov}(w_{1U_k}, w_{1U_l}) = [E(I_{1k}\, I_{1l}) - E(I_{1k})\, E(I_{1l})] / \pi_{1k}\, \pi_{1l} = (\pi_{1kl} - \pi_{1k}\pi_{1l}) / \pi_{1k}\, \pi_{1l}\,.$$

Le kl^{e} élément de la matrice $\text{Var}(\mathbf{w}_U)$ est :

$$\begin{aligned}
\text{Cov}(w_{U_k}, w_{U_l}) &= [E(I_{1k}\, I_{2k}\, I_{1l}\, I_{2l}) - E(I_{1k}\, I_{2k})\, E(I_{1l}\, I_{2l})] / \pi_{1k}\, \pi_{2k}\, \pi_{1l}\, \pi_{2l} \\
&= [E_1(I_{1k}\, I_{1l}\, E_2(I_{2k}\, I_{2l})) - E_1(I_{1k}\, E_2(I_{2k}))\, E_1(I_{1l}\, E_2(I_{2l}))] / \pi_{1k}\, \pi_{2k}\, \pi_{1l}\, \pi_{2l} \\
&= [\pi_{1kl}\, \pi_{2kl} - \pi_{1k}\, \pi_{2k}\, \pi_{1l}\, \pi_{2l}] / \pi_{1k}\, \pi_{2k}\, \pi_{1l}\, \pi_{2l}\,,
\end{aligned}$$

où E_1 et E_2 désignent l'espérance pour la première et la deuxième phase de l'échantillonnage, respectivement. À partir d'arguments similaires, il s'ensuit que le kl^{e} élément de la matrice $\text{Cov}(\mathbf{w}_{1U}, \mathbf{w}_U)$ est :

$$\text{Cov}(w_{1U_k}, w_{U_l}) = [E(I_{1k}\, I_{1l}\, I_{2l}) - E(I_{1k})\, E(I_{1l}\, I_{2l})] / \pi_{1k}\, \pi_{1l}\, \pi_{2l} = (\pi_{1kl} - \pi_{1k}\, \pi_{1l}) / \pi_{1k}\, \pi_{1l}\,.$$

Cela montre que $\text{Cov}(w_{1U_k}, w_{U_l}) = \text{Cov}(w_{1U_k}, w_{1U_l})$ et donc que $\text{Cov}(\mathbf{w}_{1U}, \mathbf{w}_U) = \text{Var}(\mathbf{w}_{1U})$, ce qui termine la démonstration.

Démonstration du théorème 1

La matrice $\boldsymbol{\Delta} = \text{Var}(\mathbf{w}_U^*)$ est non singulière si et seulement si $\text{Var}(\mathbf{w}_U) - \text{Var}(\mathbf{w}_{1U})$ est non singulier. Cela découle d'un résultat général sur les inverses des matrices partitionnées (voir Harville, 2008, page 98). Toutefois, $\text{Var}(\mathbf{w}_U) - \text{Var}(\mathbf{w}_{1U}) = \text{Var}(\mathbf{w}_{1U} - \mathbf{w}_U)$, parce que $\text{Cov}(\mathbf{w}_{1U}, \mathbf{w}_U) = \text{Var}(\mathbf{w}_{1U})$, et par conséquent $\text{Var}(\mathbf{w}_U) - \text{Var}(\mathbf{w}_{1U})$ est non singulier, étant une matrice variance-covariance. Alors, pour trouver le vecteur $\mathbf{c}_U^*$ qui minimise $(\mathbf{c}_U^* - \mathbf{w}_U^*)' \boldsymbol{\Delta}^{-1} (\mathbf{c}_U^* - \mathbf{w}_U^*)$ soumis aux contraintes $\boldsymbol{\mathcal{X}}_U' \mathbf{c}_U^* = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$, nous considérons la fonction $\mathbf{F} = (\mathbf{c}_U^* - \mathbf{w}_U^*)' \boldsymbol{\Delta}^{-1} (\mathbf{c}_U^* - \mathbf{w}_U^*) - \boldsymbol{\lambda}' \boldsymbol{\mathcal{X}}_U' \mathbf{c}_U^*$ où $\boldsymbol{\lambda}$ est un vecteur des multiplicateurs de Lagrange. Nous obtenons alors le système d'équations :

$$\frac{\partial \mathbf{F}}{\partial \mathbf{c}_U^*} = 2\boldsymbol{\Delta}^{-1}(\mathbf{c}_U^* - \mathbf{w}_U^*) - \boldsymbol{\mathcal{X}}_U \boldsymbol{\lambda} = \mathbf{0}$$
$$\boldsymbol{\mathcal{X}}_U' \mathbf{c}_U^* - \mathbf{t}_{\boldsymbol{\mathcal{X}}} = \mathbf{0}.$$

En multipliant la première équation par $\boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta}$, en utilisant $\boldsymbol{\mathcal{X}}_U' \mathbf{c}_U^* = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$ et en résolvant $\boldsymbol{\lambda}$, nous obtenons $\boldsymbol{\lambda} = 2(\boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1}\left(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}_U' \mathbf{w}_U^*\right)$. Si nous l'insérons dans la première équation et que nous résolvons $\mathbf{c}_U^*$, nous obtenons $\mathbf{c}_U^* = \mathbf{w}_U^* + \boldsymbol{\Delta}\boldsymbol{\mathcal{X}}_U(\boldsymbol{\mathcal{X}}_U' \boldsymbol{\Delta} \boldsymbol{\mathcal{X}}_U)^{-1}\left(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}_U' \mathbf{w}_U^*\right)$.

Démonstration de la proposition 1

De toute évidence, les coefficients de $\hat{\mathbf{t}}_{\mathbf{x}} - \tilde{\mathbf{t}}_{\mathbf{x}}$ dans (3.14) et (6.2) sont identiques si $\boldsymbol{\Delta}_1 = \delta \boldsymbol{\Delta}_2$. Ensuite, en utilisant la partition $\mathbf{X}_U = (\mathbf{X}_{1U}, \mathbf{X}_{2U})$, nous exprimons le coefficient de $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ dans (6.2) comme suit. Nous obtenons d'abord :

$$\begin{aligned}\mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U \left(\mathbf{X}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U\right)^{-1} \mathbf{X}_U' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} &= \mathbf{Y}_U' \boldsymbol{\Delta}_2 (\mathbf{X}_{1U}, \mathbf{X}_{2U}) \begin{pmatrix} \mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U} & \mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{2U} \\ \mathbf{X}_{2U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U} & \mathbf{X}_{2U}' \boldsymbol{\Delta}_2 \mathbf{X}_{2U} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} \\ \mathbf{X}_{2U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} \end{pmatrix} \\ &= \mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_{1U} \left[A_{11}(\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) + A_{12}(\mathbf{X}_{2U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U})\right] \\ &\quad + \mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_{2U} \left[A_{21}(\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) + A_{22}(\mathbf{X}_{2U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U})\right],\end{aligned}$$

où A_{11}, A_{12}, A_{21}, A_{22} sont calculés algébriquement à partir des inverses des matrices partitionnées. En particulier,

$$A_{11} = (\mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})^{-1} - A_{12}(\mathbf{X}_{2U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})(\mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})^{-1}$$

et $A_{21} = -A_{22}(\mathbf{X}_{2U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U}) \times (\mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})^{-1}$. Alors,

$$\begin{aligned} A_{11}(\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) + A_{12}(\mathbf{X}_{2U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) &= (\mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})^{-1} \mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} + A_{12}\mathbf{B} \\ A_{21}(\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) + A_{22}(\mathbf{X}_{2U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}) &= A_{22}\mathbf{B}, \end{aligned}$$

où

$$\mathbf{B} = \mathbf{X}_{2U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} - \mathbf{X}_{2U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U} (\mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})^{-1} (\mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}).$$

Il est alors facile de vérifier que si $\boldsymbol{\Delta}_1 = \delta \boldsymbol{\Delta}_2$, nous avons $(\mathbf{X}_{1U}' \boldsymbol{\Delta}_2 \mathbf{X}_{1U})^{-1} \mathbf{X}_{1U}' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} = \delta \mathbf{I}$ et $\mathbf{B} = \mathbf{0}$. Il s'ensuit que $\mathbf{Y}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U \left(\mathbf{X}_U' \boldsymbol{\Delta}_2 \mathbf{X}_U\right)^{-1} \mathbf{X}_U' \boldsymbol{\Delta}_1 \mathbf{X}_{1U} = \mathbf{Y}_U' \boldsymbol{\Delta}_1 \mathbf{X}_{1U}$, et que les coefficients de $\mathbf{t}_{\mathbf{x}_1} - \hat{\mathbf{t}}_{\mathbf{x}_1}$ dans (3.14) et (6.2) sont aussi identiques.

Bibliographie

Australian Bureau of Statistics (2004). Estimation for the household income and expenditure survey. Document de recherche 1352.0.55.063.

Beaumont, J.-F., Beliveau, A. et Haziza, D. (2015). Clarifying some aspects of variance estimation in two-phase sampling. *Journal of Survey Statistics and Methodology*, 3, 524-542.

Brick, J.M., et Tourangeau, R. (2017). Responsive survey designs for reducing nonresponse bias. *Journal of Official Statistics*, 33, 735-752.

Chen, S., et Kim, J.K. (2014). Two-phase sampling experiment for propensity score estimation in self-selected samples. *The Annals of Applied Statistics*, 3, 1492-1515.

Chipperfield, J.O., et Steel, D.G. (2009). Design and estimation for split questionnaire surveys. *Journal of Official Statistics*, 25, 227-244.

Estevao, V.M., et Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics*, 18, 233-255.

Estevao, V.M., et Särndal, C.-E. (2009). Un nouveau visage pour l'échantillonnage à deux phases avec estimateurs par calage. *Techniques d'enquête*, 35, 1, 3-16. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2009001/article/10880-fra.pdf.

Fuller, W.A. (1990). Analyse d'enquêtes à passages répétés. *Techniques d'enquête*, 16, 2, 177-190. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1990002/article/14537-fra.pdf.

Fuller, W.A. (1998). Replication variance estimation for two-phase sampling. *Statistica Sinica*, 8, 1153-1164.

Fuller, W.A., et Isaki, C.T. (1981). Survey design under superpopulation models. Dans *Current Topics in Survey Sampling*, (Éds., D. Krewski, J.N.K. Rao et R. Platek), New York: Academic Press, 199-226.

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, 439-457.

Harville, D.A. (2008). *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

Haziza, D., Mecatti, F. et Rao, J.N.K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *METRON-International Journal of Statistics*, vol LXVI, 91-108.

Hidiroglou, M.A. (2001). L'échantillonnage double. *Techniques d'enquête*, 27, 2, 157-169. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2001002/article/6091-fra.pdf.

Hidiroglou, M.A., et Särndal, C.-E. (1998). Emploi des données auxiliaires dans l'échantillonnage à deux phases. *Techniques d'enquête*, 24, 1, 11-20. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1998001/article/3905-fra.pdf.

Hidiroglou, M.A., Rao, J.N.K. et Haziza, D. (2008). Variance estimation in two-phase sampling. *Australian and New Zealand Journal of Statistics*, 51, 127-141.

Jones, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, Ser. B*, 42, 221-226.

Kim, J.K., et Sitter, R.R. (2003). Efficient replication variance estimation for two-phase sampling. *Statistica Sinica*, 13, 641-653.

Kim, J.K., et Yu, C.L. (2011). Estimation de la variance par répliques sous échantillonnage à deux phases. *Techniques d'enquête*, 37, 1, 73-81. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2011001/article/11448-fra.pdf.

Kim, J.K., Navarro, A. et Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 311-320.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99, 1131-1139.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for more efficient small domain estimation. *Journal of the Royal Statistical Society, Ser. B*, 72, 27-48.

Merkouris, T. (2015). Une méthode d'estimation efficace pour l'échantillonnage matriciel. *Techniques d'enquête*, 41, 1, 249-276. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2015001/article/14174-fra.pdf.

Montanari, G.E. (1987). Post-sampling efficient QR-prediction in large-scale surveys. *Revue Internationale de Statistique*, 55, 191-202.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

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

Turmelle, C., et Beaucage, Y. (2013). Le Programme intégré de la statistique des entreprises : utilisation d'un plan de sondage à deux phases pour produire des estimations fiables. *Recueil : Symposium 2013, Produire des estimations fiables à partir de bases imparfaites.*

Wolter, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

Wu, C., et Luan, Y. (2003). Optimal calibration estimators under two-phase sampling. *Journal of Official Statistics*, 2, 119-131.

Techniques d'enquête, décembre 2022
Vol. 48, n° 2, pp. 507-535
Statistique Canada, n° 12-001-X au catalogue

Modèles spatiaux bayésiens pour l'estimation des moyennes pour petites régions échantillonnées et non échantillonnées

Hee Cheol Chung et Gauri S. Datta[1]

Résumé

Dans de nombreuses applications, les moyennes de population des petites régions géographiquement adjacentes présentent une variation spatiale. Si les variables auxiliaires disponibles ne tiennent pas suffisamment compte de la configuration spatiale, la variation résiduelle sera incluse dans les effets aléatoires. Par conséquent, l'hypothèse de distribution indépendante et identique sur les effets aléatoires du modèle Fay-Herriot échouera. De plus, des ressources limitées empêchent souvent l'inclusion de nombreuses sous-populations dans l'échantillon; il en résulte de petites régions non échantillonnées. Le problème peut être exacerbé au moment de prédire les moyennes de petites régions non échantillonnées à l'aide du modèle de Fay-Herriot ci-dessus, car les prévisions seront faites uniquement en fonction des variables auxiliaires. Pour remédier à ce problème, nous considérons les modèles spatiaux bayésiens à effets aléatoires qui peuvent prendre en compte de multiples régions non échantillonnées. Dans des conditions légères, nous déterminons si les distributions *a posteriori* de divers modèles spatiaux sont adaptées à une catégorie utile de densités *a priori* incompatibles avec les paramètres du modèle. L'efficacité de ces modèles spatiaux est évaluée à partir de données simulées et réelles. Plus précisément, nous examinons les prévisions du revenu médian des familles de quatre personnes à l'échelle de l'État fondées sur la « Current Population Survey » (enquête sur l'état de la population) de 1990 et le « Census for the United States of America » (recensement mené aux États-Unis d'Amérique) de 1980.

Mots-clés : Autorégression conditionnelle; Current Population Survey; modèle Fay-Herriot; autorégression intrinsèque; effets mixtes; autorégression simultanée; petites régions non échantillonnées.

1. Introduction

Les enquêtes par sondage fournissent des données utiles pour estimer les diverses caractéristiques d'une population d'intérêt. Les enquêtes sont généralement conçues de sorte que les estimateurs fondés sur le plan de sondage présentent une exactitude adéquate. Toutefois, lorsqu'il s'agit d'estimer une caractéristique d'une sous-population, une estimation directe fondée sur le plan de sondage et uniquement sur les données de cette sous-population est habituellement inexacte, car l'échantillon accessible est petit, voire inexistant. On appelle « petites régions » les sous-populations dont la taille d'échantillon n'est pas suffisante pour produire des estimations directes fiables. De plus, des ressources limitées empêchent souvent la sélection de nombreuses sous-populations dans l'échantillon, ce qui crée de petites régions non échantillonnées. Par exemple, l'American Community Survey (ACS) vise à produire des statistiques fiables pour les comtés des États-Unis. Cependant, l'ACS échantillonne habituellement environ le tiers des comtés, ce qui donne de nombreuses petites régions non échantillonnées.

Afin d'accroître l'exactitude des estimations directes pour petites régions, une approche fondée sur un modèle a été largement utilisée pour faciliter l'emprunt d'information à partir des estimations directes d'autres domaines et d'autres données auxiliaires. Dans de nombreuses applications, des renseignements supplémentaires provenant d'autres enquêtes et de données administratives fournissent des covariables

1. Hee Cheol Chung, Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, Caroline du Nord. Courriel : hchung13@uncc.edu; Gauri S. Datta, Department of Statistics, University of Georgia, Athens, GA et Center for Statistical Research and Methodology, U.S. Census Bureau, Suitland, MD.

utiles. Une estimation d'une région, fondée sur un modèle, est produite en réduisant de façon appropriée son estimation directe (si elle est disponible) à une estimation par régression synthétique fondée sur des variables auxiliaires. L'amélioration de la prévision dépend grandement de la mesure dans laquelle les moyennes de sous-population pour la caractéristique sont liées aux variables auxiliaires. Si une petite région n'a pas d'estimation directe, le modèle traditionnel indépendant à effets aléatoires de Fay et Herriot (1979) permet d'estimer la moyenne par une estimation par régression synthétique seulement.

Fay et Herriot (1979) ont proposé un modèle utile permettant d'élaborer des estimations de moyennes pour petites régions, fondé sur des estimations directes d'enquête (si disponibles) et des estimations par régression synthétique calculées à partir de variables auxiliaires. Ce modèle, qui est essentiellement un modèle linéaire mixte, est communément appelé le modèle Fay-Herriot (FH) dans l'estimation pour petites régions. Pour $i = 1, \ldots, m,$ supposons que Y_i est l'estimation directe de la caractéristique sur petites régions θ_i obtenue à partir d'une enquête. Supposons aussi que $\mathbf{x}_i$ et $\boldsymbol{\beta}$ sont les vecteurs à p composantes des covariables et les coefficients de régression correspondants, respectivement. En indiquant l'erreur d'échantillonnage de Y_i comme étant e_i, le modèle FH à effets aléatoires indépendants peut être formulé comme suit :

$$Y_i = \theta_i + e_i, \quad \theta_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + v_i, \quad i = 1, \ldots, m, \tag{1.1}$$

où les e_i et les effets aléatoires v_i $i = 1, \ldots, m,$ sont tous répartis indépendamment avec $e_i \sim N(0, D_i)$, et $v_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_v^2)$. Les variances d'échantillonnage D_i, $i = 1, \ldots, m$ sont considérées comme connues, tandis que le paramètre de régression β et la variance d'erreur du modèle σ_v^2, appelés paramètres de modèle, sont des quantités inconnues. Pour les régions non échantillonnées comportant des variables auxiliaires, seule la deuxième partie de (1.1) est vérifiée pour θ_i.

De nombreuses études ont été menées sur le modèle FH à effets aléatoires indépendants et ses nombreuses variantes. Alors que Fay et Herriot (1979) ont utilisé l'approche empirique de Bayes (EB), par la suite, Prasad et Rao (1990), Datta et Lahiri (2000) et Datta, Rao et Smith (2005) ont utilisé l'approche fréquentiste et ont dérivé l'erreur quadratique moyenne (EQM) de deuxième ordre du meilleur prédicteur linéaire sans biais empirique (MPLSBE) de θ_i et divers estimateurs approximatifs sans biais de deuxième ordre de l'EQM (voir Datta et Lahiri, 2000). Cependant, Ghosh (1992) a proposé une approche hiérarchique bayésienne (HB) pour le modèle Fay-Herriot (voir aussi Datta et coll. [2005]). Dans le cadre bayésien, le modèle FH dans (1.1) peut être exprimé comme le modèle HB suivant :

$$Y_i \,|\, \theta_1, \ldots, \theta_m, \boldsymbol{\beta}, \sigma_v^2 \overset{\text{ind}}{\sim} N(\theta_i, D_i), \quad i = 1, \ldots, m, \tag{1.2}$$

$$\theta_i \,|\, \boldsymbol{\beta}, \sigma_v^2 \overset{\text{ind}}{\sim} N(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \sigma_v^2), \quad i = 1, \ldots, m, \tag{1.3}$$

$$\pi(\boldsymbol{\beta}, \sigma_v^2) \propto g(\boldsymbol{\beta}, \sigma_v^2), \tag{1.4}$$

où $g(\cdot)$ est une fonction correctement choisie de $\boldsymbol{\beta}$ et σ_v^2, qui exprime une fonction de densité de probabilité (fdp) *a priori* pour ces paramètres. Fay et Herriot (1979) ont élaboré à l'origine un prédicteur

EB pour θ_i, qui ne nécessite pas de fdp *a priori*, comme dans (1.4). Bien qu'une approche EB standard sous-estime habituellement la mesure de l'incertitude de l'estimateur EB de θ_i, l'approche HB facilite la quantification de l'incertitude en raison de l'estimation des paramètres de modèle inconnus, $\boldsymbol{\beta}$ et σ_v^2. L'incertitude est entièrement représentée par la distribution *a posteriori* des paramètres du modèle.

Dans les estimations fondées sur un modèle, les effets aléatoires sont d'une grande importance pour bien rendre compte de la variabilité restante des θ_i qui n'est pas expliquée par le modèle de régression. Dans des applications réelles, les petites régions comportent généralement des caractéristiques comme la taille de la population, l'origine ethnique, le groupe d'âge et le niveau de scolarité, lesquelles peuvent avoir une incidence sur la variabilité des effets sur les petites régions. De plus, lorsque les taux de prévalence de la maladie sont intéressants, il est raisonnable de supposer que les effets aléatoires de petites régions adjacentes sont corrélés d'une certaine façon. Dans de tels cas, le modèle FH présenté en (1.1), que nous appelons le modèle FH à effets aléatoires indépendants, est une simplification excessive et une erreur de spécification de la distribution des effets aléatoires en supposant une distribution commune et indépendante. Opsomer, Claeskens, Ranalli, Kauermann et Breidt (2008) et Rao, Sinha et Dumitrescu (2014) ont proposé des modèles non paramétriques d'estimation pour petites régions, qui permettent de cerner l'effet de proximité spatiale à l'aide de la fonction p-spline (spline pénalisée). Toutefois, ces approches exigent des coûts informatiques supplémentaires pour l'inférence du modèle et la quantification de l'incertitude.

Dans le cadre de la présente étude, nous proposons des modèles FH spatiaux qui tiennent effectivement compte de l'hétéroscédasticité et de la dépendance spatiale des effets sur de petites régions. Nous adoptons une approche entièrement bayésienne en spécifiant une classe de distributions *a priori* non informatives sur les paramètres du modèle et incompatibles avec ceux-ci, et nous modélisons la dépendance spatiale des effets aléatoires sur petites régions par quatre structures d'autocorrélation largement utilisées. Il s'agit notamment de modèles autorégressifs simultanés et de trois types de modèles autorégressifs conditionnels. Il existe une abondante littérature sur les modèles spatiaux dans le cadre bayésien. Sun, Tsutakawa et Speckman (1999) ont étudié un modèle HB avec les modèles autorégressifs conditionnels et intrinsèques sur les effets aléatoires. Speckman et Sun (2003) ont examiné les mêmes modèles dans le contexte du lissage bayésien de la spline. Pour l'estimation sur petites régions, You et Zhou (2011) ont modélisé les effets sur petites régions à l'aide d'un modèle autorégressif conditionnel. Dans le prolongement du modèle de séries chronologiques FH (Datta, Lahiri, Maiti et Lu, 1999), Torabi (2012) a proposé un modèle spatio-temporel ayant des effets aléatoires autorégressifs intrinsèques. Porter, Holan, Wikle et Cressie (2014) ont proposé une extension du modèle FH ayant des covariables fonctionnelles et des effets aléatoires autorégressifs intrinsèques. Porter, Wikle et Holan (2015) ont intégré les effets aléatoires conditionnels autorégressifs au modèle FH multivarié.

Les modèles bayésiens actuels d'estimation spatiale sur petites régions considèrent une distribution *a priori* adéquate sur σ_v^2, même si la spécification d'une telle distribution *a priori* nécessitera des connaissances spécialisées. De plus, tous les modèles actuels supposent une structure autorégressive conditionnelle sur les effets aléatoires. Les principales contributions du présent article sont les suivantes. Premièrement, à notre connaissance, les modèles proposés à la section 2 (section 2.1) comprennent la plupart des structures spatiales couramment utilisées. Deuxièmement, à la section 2.2, nous élargissons

davantage les modèles spatiaux pour estimer les moyennes de plusieurs petites régions non échantillonnées sans estimations directes. La moyenne θ_i de la zone non échantillonnée est estimée par l'emprunt d'information des variables auxiliaires de la région et, pour les modèles spatiaux, des résidus de régression des régions avoisinantes. Troisièmement, pour tous les modèles proposés, nous fournissons, à la section 2.3, des conditions suffisantes pour la propriété *a posteriori* pour une classe de distributions *a priori* non informatives sur les paramètres du modèle et incompatibles avec ceux-ci. Il est intéressant de mentionner que les conditions suffisantes ne dépendent pas du modèle spatial présumé, à condition que le modèle produise une matrice de covariance définie positive pour les effets aléatoires. Nous fournissons des étapes d'échantillonnage de rejet pour la simulation à partir de la distribution *a posteriori* des modèles proposés à la section 3. L'efficacité des modèles spatiaux proposés est démontrée aux sections 4 et 5. Nous appliquons les modèles spatiaux à des ensembles de données simulées et à des données d'enquête réelles issues de la Current Population Survey (CPS). Nous comparons divers modèles spatiaux à la section 5 pour estimer le revenu médian des familles de quatre personnes pour les 49 États contigus des États-Unis, d'après les données de la CPS, les covariables appropriées du recensement précédent et des données administratives. Nos analyses de données et nos études de simulation révèlent que les modèles spatiaux proposés améliorent considérablement l'exactitude des prévisions et réduisent la mesure de l'incertitude, l'écart-type *a posteriori*. Nous formulons des observations finales à la section 6. Toutes les précisions techniques sont fournies à l'annexe.

2. Certaines options spatiales en remplacement du modèle Fay-Herriot à effets aléatoires indépendants

2.1 Intégration des effets spatiaux aléatoires

Soit $\mathbf{Y} = (Y_1, \ldots, Y_m)^{\mathrm{T}}$ le vecteur à m composantes avec les estimations directes de m petites régions, et soit $\mathbf{D} = \mathrm{diag}\{D_i\}_{i=1}^m$ la matrice diagonale $m \times m$ avec les variances d'échantillonnage des estimations directes. Nous désignons par $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^{\mathrm{T}}$ le vecteur à m composantes des moyennes de petites régions. De plus, supposons que $\mathbf{x}_i \in \mathbb{R}^p$ est le vecteur à p composantes des variables auxiliaires (y compris le terme d'ordonnée à l'origine) pour la i^{e} petite région, et que $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^{\mathrm{T}}$. Un cas particulier du modèle HB donné en (1.2) à (1.4) peut s'exprimer comme suit :

$$\mathbf{Y} \,|\, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2 \;\sim\; N_m(\boldsymbol{\theta}, \mathbf{D}), \tag{2.1}$$

$$\boldsymbol{\theta} \,|\, \boldsymbol{\beta}, \sigma_v^2 \;\sim\; N_m(\mathbf{X}\boldsymbol{\beta}, \sigma_v^2 \mathbf{I}_m), \tag{2.2}$$

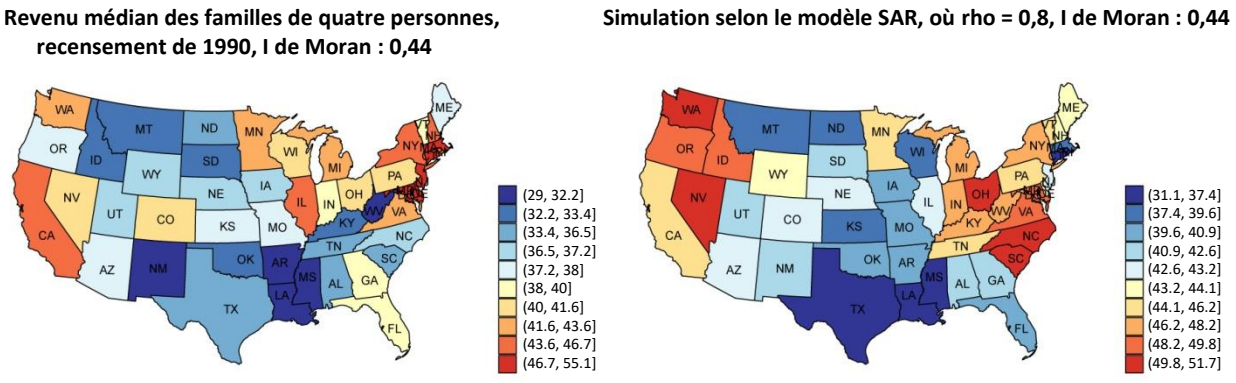
$$\pi(\boldsymbol{\beta}, \sigma_v^2) \;\propto\; 1, \tag{2.3}$$

où $\boldsymbol{\beta}$ est le vecteur du coefficient de régression à p composantes, σ_v^2 est la variance de l'erreur du modèle et $\mathbf{I}_m$ est la matrice d'identité d'ordre m. La distribution *a priori* uniforme (2.3) sur les paramètres du modèle est un précédent non informatif largement utilisé, et la fdp *a posteriori* qui en

résulte est adéquate, à condition que $m > p + 2$. Voir Berger (1985) et Datta et Smith (2003) pour obtenir une analyse détaillée.

Le modèle (2.2) suppose que θ_i, $i = 1, \ldots, m$, sont répartis indépendamment sur les petites régions avec une variance des effets aléatoires communs de σ_v^2. Toutefois, dans de nombreux problèmes d'estimation sur petites régions, la caractéristique de la région d'intérêt est étroitement liée à des facteurs géographiques comme la taille de la population, l'ethnicité, le groupe d'âge et le niveau de scolarité. Lorsque les covariables disponibles n'expliquent pas entièrement cette association spatiale, les hypothèses d'indépendance et de variance égale des effets aléatoires échouent, et l'inférence fondée sur le modèle hiérarchique indiqué en (2.1) à (2.3) peut générer des estimations non fiables, ce qui entraîne des décisions erronées. La figure 2.1 illustre les associations spatiales du revenu médian des familles de quatre personnes issues du recensement de 1990 (à l'échelle de 1 000 $) pour $m = 49$ États des États-Unis, y compris le District de Columbia. Les données simulées ayant la même valeur I de Moran sont également affichées à des fins de comparaison, où la valeur I de Moran est une mesure d'autocorrélation spatiale. Les données simulées sont générées selon le modèle autorégressif simultané (SAR pour *simultaneous autoregressive* en anglais) (défini ci-dessous), $\rho = 0{,}8$ correspondant à l'emplacement et à l'échelle des données du recensement. Les deux volets montrent l'existence d'une dépendance spatiale dans le revenu médian des familles de quatre personnes du recensement de 1990. En pratique, les covariables capables de rendre compte pleinement de la variation spatiale existante ne sont pas toujours disponibles, et le problème peut être exacerbé s'il existe des variables cachées, car elles introduisent une variabilité supplémentaire qui ne peut être expliquée par les effets aléatoires indépendants et identiquement distribués (i.i.d.).

Figure 2.1 Illustrations graphiques du revenu médian des familles de quatre personnes issues du recensement de 1990 pour $m = 49$ États des États-Unis et de données simulées.



- Les données simulées sont générées selon le modèle autorégressif simultané (SAR pour *simultaneous autoregressive* en anglais), où $\rho = 0{,}8$ et l'estimation du I de Moran est la même.
- AL = Alabama, AZ = Arizona, AR = Arkansas, CA = Californie, CO = Colorado, CT = Connecticut, DE = Delaware, DC = District de Columbia, FL = Floride, GA = Géorgie, ID = Idaho, IL = Illinois, IN = Indiana, IA = Iowa, KS = Kansas, KY = Kentucky, LA = Louisiane, ME = Maine, MD = Maryland, MA = Massachusetts, MI = Michigan, MN = Minnesota, MS = Mississippi, MO = Missouri, MT = Montana, NE = Nebraska, NV = Nevada, NH = New Hampshire, NJ = New Jersey, NM = Nouveau-Mexique, NY = New York, NC = Caroline du Nord, ND = Dakota du Nord, OH = Ohio, OK = Oklahoma, OR = Oregon, PA = Pennsylvanie, RI = Rhode Island, SC = Caroline du Sud, SD = Dakota du Sud, TN = Tennessee, TX = Texas, UT = Utah, VT = Vermont, VA = Virginie, WA = Washington, WV = Virginie-Occidentale, WI = Wisconsin, WY = Wyoming.

Pour régler ce problème, nous proposons d'utiliser des effets aléatoires spatialement corrélés. Supposons que $\mathbf{W} = \{w_{ij}\}_{ij}$, $1 \le i, j \le m$, soit la matrice de contiguïté qui joue un rôle important dans la prise en compte de la dépendance spatiale. En particulier, $w_{ij} = 1$ si les petites régions i^{e} et j^{e} sont connectées par des limites géographiques ou par d'autres mécanismes (par exemple le trafic aérien), et sinon $w_{ij} = 0$. Également, $w_{ii} = 0$ pour $i = 1, \ldots, m$. Les entrées extradiagonales, w_{ij}, ne doivent pas nécessairement être binaires; elles peuvent prendre d'autres valeurs positives, comme la « longueur » de la frontière géographique ou les volumes de trafic aérien entre les deux régions. La matrice de contiguïté $\mathbf{W}$ étant symétrique, ses valeurs propres sont réelles. Nous désignons la i^{e} plus grande valeur propre de $\mathbf{W}$ par $\lambda_i(\mathbf{W})$, de sorte que $\lambda_m(\mathbf{W}) \le \ldots \le \lambda_1(\mathbf{W})$. Puisque $\mathbf{W}$ n'est pas nulle et que $\sum_{i=1}^{m} w_{ii} = 0$, nous obtenons donc le résultat $\lambda_m(\mathbf{W}) < 0 < \lambda_1(\mathbf{W})$. Supposons que $w_{i.} = \sum_{j=1}^{m} w_{ij}$ est la somme de la i^{e} ligne de $\mathbf{W}$ et $\mathbf{L} = \text{diag}\{w_{i.}\}_{i=1}^{m}$. En supposant que les éléments diagonaux de $\mathbf{L}$ sont positifs, c'est-à-dire que toutes les petites régions ont au moins une région voisine, nous définissons $\tilde{\mathbf{W}} = \mathbf{L}^{-1}\mathbf{W}$. Puisque $\tilde{\mathbf{W}}$ est une matrice stochastique à lignes, toutes ses valeurs propres sont comprises entre -1 et 1, dont au moins une est 1. Par conséquent, $\lambda_1(\tilde{\mathbf{W}}) = 1$. De plus, $\tilde{\mathbf{W}}$ et $\text{diag}\{w_{i.}^{-1/2}\}_{i=1}^{m}\mathbf{W}\,\text{diag}\{w_{i.}^{-1/2}\}_{i=1}^{m}$ ont le même ensemble de valeurs propres, et cette dernière matrice est symétrique. Donc, toutes les valeurs propres de $\tilde{\mathbf{W}}$ sont réelles et la valeur $\lambda_m(\tilde{\mathbf{W}})$ sera négative. Nous considérons quatre dépendances spatiales de rechange associées à des effets aléatoires, qui sont représentées par les matrices de précision définies positives (à l'exclusion du paramètre d'échelle σ_v^2) :

$$\text{SAR}: \quad \mathbf{\Omega}_2(\rho) = (\mathbf{I}_m - \rho\tilde{\mathbf{W}})^{\top}(\mathbf{I}_m - \rho\tilde{\mathbf{W}}), \quad \rho \in (-1, 1), \tag{2.4}$$

$$\text{SCAR}: \quad \mathbf{\Omega}_3(\rho) = \mathbf{I}_m - \rho\mathbf{W}, \quad \rho \in (\lambda_m(\mathbf{W})^{-1}, \lambda_1(\mathbf{W})^{-1}), \tag{2.5}$$

$$\text{CAR}: \quad \mathbf{\Omega}_4(\rho) = \mathbf{L} - \rho\mathbf{W}, \quad \rho \in (\lambda_m(\tilde{\mathbf{W}})^{-1}, \lambda_1(\tilde{\mathbf{W}})^{-1}), \tag{2.6}$$

$$\text{LCAR}: \quad \mathbf{\Omega}_5(\rho) = \rho\mathbf{R} + (1-\rho)\,\mathbf{I}_m, \quad \rho \in (0, 1), \tag{2.7}$$

où ρ est le paramètre de dépendance spatiale qui représente la force de la dépendance spatiale (Hodges, 2019, chapitre 5.2) et $\mathbf{R}$ est défini comme $\mathbf{R} = \mathbf{\Omega}_4(1) = \mathbf{L} - \mathbf{W}$. Étant donné que les valeurs propres de $\mathbf{I}_m - \tilde{\mathbf{W}}$ sont comprises entre 0 (la plus petite valeur propre) et $1 - \lambda_m(\tilde{\mathbf{W}})$ (la plus grande valeur propre, > 1), la matrice $\mathbf{R}$ est définie non négative. Chaque matrice de précision est garantie positive et définie tant que ρ se situe dans l'intervalle précisé dans la définition respective.

La matrice de contiguïté $\tilde{\mathbf{W}}$ du modèle autorégressif simultané (SAR) (Whittle, 1954) est normalisée par ligne, de sorte que ρ peut varier de -1 à 1 tout en préservant la définition positive (Banerjee, Carlin et Gelfand, 2003, chapitre 4.4). Le modèle (2.5) est une version simple du modèle autorégressif conditionnel (CAR pour *conditional autoregressive* en anglais) (Rao et Molina, 2015, chapitre 9.6.2), où les entrées diagonales de la matrice de précision sont toutes égales à 1. Même si 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 entraîne l'hétéroscédasticité des effets aléatoires. Nous appelons ce modèle le modèle autorégressif conditionnel simple (SCAR pour *simple conditional autoregressive* en anglais). Le modèle (2.6) est un modèle autorégressif conditionnel largement utilisé (CAR; Banerjee et coll., 2003; Besag et

Kooperberg, 1995; You et Zhou, 2011), où les entrées diagonales de la matrice de précision sont le nombre de quartiers de la région correspondante. La limite supérieure de ρ est $\lambda_1(\tilde{\mathbf{W}})^{-1}=1$, et dans le cas de $\rho=1$, le modèle avec $\mathbf{\Omega}_4(1)$ est appelé le modèle autorégressif intrinsèque (IAR pour *intrinsic autoregressive* en anglais) (Banerjee et coll., 2003, chapitre 4.3). Le modèle (2.7) est un modèle autorégressif conditionnel, appelé autorégressif conditionnel de Leroux (LCAR pour *Leroux's conditional autoregressive* en anglais), dont la matrice de précision est donnée par la combinaison convexe de $\mathbf{R}=\mathbf{\Omega}_4(1)$ et $\mathbf{I}_m$. Ce modèle a été examiné par Leroux, Lei et Breslow (2000); MacNab (2003); You et Zhou (2011), où le i^{e} élément diagonal de $\mathbf{R}$ est le nombre de quartiers de la i^{e} petite région et le $(i, j)^{\text{e}}$ élément extradiagonal est de -1 si les i^{e} et j^{e} petites régions sont connectées, sinon il est de 0.

Les modèles autorégressifs conditionnels, le modèle SCAR, le modèle CAR et le modèle LCAR, supposent que θ_i dépend uniquement des moyennes de petites régions voisines. En d'autres termes, θ_i est corrélé avec les θ_j, $j\neq i$, uniquement par les moyennes des régions environnantes. Au contraire, le modèle SAR suppose que θ_i dépend de tous les autres θ_j en même temps, $j\neq i$, mais a des corrélations plus fortes (plus faibles) pour les régions avoisinantes (éloignées). Le modèle FH à effets aléatoires indépendants peut être considéré comme un cas spécial du modèle SAR, SCAR ou LCAR où $\rho=0$. Pour des raisons de commodité, nous incluons le modèle FH à effets aléatoires indépendants dans notre modèle en adoptant sa matrice de précision $\mathbf{\Omega}_1(\rho)=\mathbf{I}_m$, bien qu'elle soit exempte de ρ.

Nous considérons les modèles spatiaux HB suivants en intégrant les cinq dépendances spatiales définies en (2.4) à (2.7) :

$$\mathbf{Y}\,|\,\boldsymbol{\theta},\boldsymbol{\beta},\sigma_v^2,\rho \;\sim\; N_m(\boldsymbol{\theta},\mathbf{D}), \tag{2.8}$$

$$\boldsymbol{\theta}\,|\,\boldsymbol{\beta},\sigma_v^2,\rho \;\sim\; N_m(\mathbf{X}\boldsymbol{\beta},\sigma_v^2\{\mathbf{\Omega}_k(\rho)\}^{-1}),\quad k=1,\ldots,5, \tag{2.9}$$

$$\pi(\boldsymbol{\beta},\sigma_v^2,\rho)\;\propto\; g(\sigma_v^2)\,h(\rho),\quad \boldsymbol{\beta}\in\mathbb{R}^p,\sigma_v^2>0,\; l_k<\rho<u_k, \tag{2.10}$$

où σ_v^2 est le paramètre de l'échelle du modèle, $g(\sigma_v^2)$ et $h(\rho)$ sont des fonctions appropriées de σ_v^2 et ρ, l_k et u_k sont les limites inférieure et supérieure de ρ sous le k^{e} modèle. Nous évitons le terme « variance d'erreur du modèle » pour σ_v^2, car les entrées diagonales de $\mathbf{\Omega}_k(\rho)$ varient d'une petite région à l'autre et n'égalent pas nécessairement toutes 1.

2.2 Estimation des moyennes de population pour les petites régions non échantillonnées

Dans la présente section, nous examinons le cas où, dans l'enquête, plusieurs petites régions non échantillonnées n'ont pas d'estimations directes. Dans de nombreuses applications, les ressources limitées empêchent souvent l'inclusion de nombreuses sous-populations dans l'échantillon; il en résulte de petites régions non échantillonnées. Les petites régions non échantillonnées sont parfois appelées « régions non harmonisées » (Trevisani et Gelfand, 2013) lorsqu'elles découlent d'un manque de concordance des domaines entre l'estimation directe et les variables auxiliaires. Pour n'importe laquelle de ces régions non échantillonnées, la prédiction de sa moyenne à partir de n'importe quel modèle non spatial est fondée

uniquement sur son estimateur synthétique. Nous proposons d'exploiter les dépendances spatiales dans la prédiction des moyennes de petites régions non échantillonnées. Les prédictions des modèles proposés sont obtenues en modifiant l'estimateur synthétique, à l'aide du vecteur des résidus de régression, en mettant davantage l'accent sur les résidus de régression des régions avoisinantes.

Sans perte de généralité, supposons qu'il y ait m_1 petites régions non échantillonnées et que $Y_{m_1+1}, \ldots, Y_m$ sont les estimations directes de $m_2 = m - m_1$ petites régions échantillonnées. En nous fondant sur les estimations directes de m_2 régions échantillonnées, nous considérons les modèles HB suivants :

$$\mathbf{Y}_{(2)} \,|\, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho \;\sim\; N_{m_2}(\boldsymbol{\theta}_{(2)}, \mathbf{D}_{(2)}), \tag{2.11}$$

$$\boldsymbol{\theta} \,|\, \boldsymbol{\beta}, \sigma_v^2, \rho \;\sim\; N_m\left(\mathbf{X}\boldsymbol{\beta}, \sigma_v^2 \{\boldsymbol{\Omega}_k(\rho)\}^{-1}\right), \quad k = 1, \ldots, 5, \tag{2.12}$$

$$\pi(\boldsymbol{\beta}, \sigma_v^2, \rho) \;\propto\; g(\sigma_v^2)\, h(\rho), \quad \boldsymbol{\beta} \in \mathbb{R}^p, \sigma_v^2 > 0,\; l_k < \rho < u_k, \tag{2.13}$$

où $\mathbf{Y}_{(2)} = (Y_{m_1+1}, \ldots, Y_m)^\top$, $\mathbf{D}_{(2)} = \operatorname{diag}\{D_i\}_{i=m_1+1}^m$ et $\boldsymbol{\theta}_{(2)} = (\theta_{m_1+1}, \ldots, \theta_m)^\top$, qui est le sous-vecteur de $\boldsymbol{\theta}$ correspondant aux régions échantillonnées.

2.3 Propriété des distributions *a posteriori*

Dans la présente section, nous établissons la pertinence des distributions *a posteriori* des modèles spatiaux de petites régions données en (2.8) à (2.10) et (2.11) à (2.13). Supposons que $I(\cdot)$ est la fonction indicatrice prenant la valeur 1 lorsque son argument est vrai, sinon 0. Nous fournissons d'abord les conditions générales de la pertinence *a posteriori* des modèles proposés.

Théorème 1. Pour tous les modèles spatiaux HB donnés en (2.8) à (2.10) et (2.11) à (2.13), les fonctions de densité de probabilité *a posteriori* sont appropriées si les conditions suivantes se maintiennent pour une certaine constante positive $c > 0$:

(a) $\int_0^\infty g(\sigma_v^2)\, I(\sigma_v^2 \le c)\, d\sigma_v^2 < \infty.$

(b) $\int_0^\infty (\sigma_v^2)^{-(m^* - p)/2}\, g(\sigma_v^2)\, I(\sigma_v^2 > c)\, d\sigma_v^2 < \infty.$

(c) $\int_{l_k}^{u_k} h(\rho)\, d\rho < \infty,$

où $m^* = m$ pour (2.8) à (2.10), et $m^* = m - m_1$ pour (2.11) à (2.13).

Si $g(\cdot)$ est une fdp appropriée, alors (a) est vrai automatiquement et (b) est satisfait si $m^* \ge p$. La condition $m^* \ge p$ est évidente, car au moins p observations sont nécessaires pour estimer p composantes de $\boldsymbol{\beta}$ lorsque aucune information de fond à son sujet n'est disponible. De plus, toute fonction bornée de ρ satisfait à (c) dans le théorème 1, étant donné que leurs supports sont toutes bornées. En particulier, dans la famille populaire des distributions *a priori* non informatives, les fdp *a posteriori* sont adéquates dans les conditions suivantes.

$$\pi(\boldsymbol{\beta}, \sigma_v^2, \rho) \;\propto\; (\sigma_v^2)^{-\alpha}\, I(l_k < \rho < u_k), \quad \boldsymbol{\beta} \in \mathbb{R}^p, \sigma_v^2 > 0, \tag{2.14}$$

Corollaire 1. Pour n'importe lequel des modèles spatiaux HB donnés en (2.8) et (2.9) et (2.11) et (2.12) avec la distribution *a priori* mentionnée en (2.14), la fdp *a posteriori* est adéquate, tant que $\alpha < 1$ et $m^* > p + 2 - 2\alpha.$

Pour ce qui est de la distribution *a priori* uniforme comportant $\alpha = 0$ (qui sera utilisée dans le présent article), la pertinence *a posteriori* des distributions pour les modèles (2.8) et (2.9) est garantie tant que le nombre de petites régions est supérieur à $p + 2.$ Pour les modèles incorporant des régions non échantillonnées indiquées en (2.11) et (2.12), la deuxième condition du corollaire 1.1 devient $m - m_1 > p + 2,$ et par conséquent, les fdp *a posteriori* sont adéquates tant que le nombre de régions non échantillonnées est inférieur à $m - p - 2,$ ou qu'au moins $p + 3$ régions ont un échantillon.

3. Simulation de distributions *a posteriori*

Dans la présente section, nous illustrons les étapes d'échantillonnage de rejet pour obtenir des échantillons indépendants *a posteriori* à partir des distributions *a posteriori* des modèles proposés. Nous supposons que les composantes du vecteur de la moyenne de petites régions $\boldsymbol{\theta}$ sont disposées de façon que $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)}^\top, \boldsymbol{\theta}_{(2)}^\top)^\top,$ où $\boldsymbol{\theta}_{(1)} \in \mathbb{R}^{m_1}$ et $\boldsymbol{\theta}_{(2)} \in \mathbb{R}^{m_2}$ sont les vecteurs de la moyenne de petites régions correspondant aux régions non échantillonnées et échantillonnées, respectivement. Pour des raisons de commodité, nous désignons la matrice de précision du k^{e} modèle spatial par $\boldsymbol{\Omega} = (\sigma_v^2)^{-1} \boldsymbol{\Omega}_k(\rho)$ et l'intervalle admissible de ρ par (l, u) en supprimant l'indice du modèle k.

Nous calculons d'abord la densité *a posteriori* marginale de (σ_v^2, ρ) et nous fournissons les procédures d'échantillonnage subséquentes. Supposons que $\mathbf{0}_{m_2 \times m_1}$ est la matrice nulle $m_2 \times m_1$ et que $\mathbf{M} = [\mathbf{0}_{m_2 \times m_1}, \mathbf{I}_{m_2}],$ de sorte que $\boldsymbol{\theta}_{(2)} = \mathbf{M}\boldsymbol{\theta}.$ Supposons également que $\mathbf{X}_{(2)} = \mathbf{M}\mathbf{X}.$ Si l'on n'intègre pas $\boldsymbol{\theta}$ dans le modèle (2.11) et (2.12), nous avons $\mathbf{Y}_{(2)} \,|\, \boldsymbol{\beta}, \sigma_v^2, \rho \sim N_{m_2}(\mathbf{X}_{(2)}\boldsymbol{\beta}, \boldsymbol{\Delta}),$ où $\boldsymbol{\Delta} = \mathbf{D}_{(2)} + \mathbf{M}\boldsymbol{\Omega}^{-1}\mathbf{M}^\top.$ La marginalisation subséquente de $\boldsymbol{\beta}$ donne la densité *a posteriori* marginale $p(\sigma_v^2, \rho \,|\, \mathbf{y}_{(2)})$ comme suit :

$$p(\sigma_v^2, \rho \,|\, \mathbf{y}_{(2)}) \propto \frac{\exp\left[-\frac{1}{2}\mathbf{y}_{(2)}^\top \boldsymbol{\Delta}^{-1}\left\{\boldsymbol{\Delta} - \mathbf{X}_{(2)}(\mathbf{X}_{(2)}^\top \boldsymbol{\Delta}^{-1}\mathbf{X}_{(2)})^{-1}\mathbf{X}_{(2)}^\top\right\}\boldsymbol{\Delta}^{-1}\mathbf{y}_{(2)}\right]}{|\boldsymbol{\Delta}|^{1/2}\left|\mathbf{X}_{(2)}^\top \boldsymbol{\Delta}^{-1}\mathbf{X}_{(2)}\right|^{1/2}} I(l < \rho < u). \tag{3.1}$$

De plus, nous avons des distributions *a posteriori* conditionnelles de $\boldsymbol{\beta}$ et $\boldsymbol{\theta}$ qui sont :

$$\boldsymbol{\beta} \,|\, \sigma_v^2, \rho, \mathbf{y} \sim N_p(\boldsymbol{\gamma}, \boldsymbol{\Gamma}), \tag{3.2}$$

$$\boldsymbol{\theta} \,|\, \boldsymbol{\beta}, \sigma_v^2, \rho, \mathbf{y} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Psi}), \tag{3.3}$$

où $\boldsymbol{\Gamma} = (\mathbf{X}_{(2)}^\top \boldsymbol{\Delta}^{-1}\mathbf{X}_{(2)})^{-1},$ $\boldsymbol{\gamma} = \boldsymbol{\Gamma}\mathbf{X}_{(2)}^\top \boldsymbol{\Delta}^{-1}\mathbf{y}_{(2)},$ $\boldsymbol{\mu} = \mathbf{y}_* - \boldsymbol{\Psi}\boldsymbol{\Omega}(\mathbf{y}_* - \mathbf{X}\boldsymbol{\beta}),$ $\mathbf{y}_* = (\mathbf{0}_{m_1}^\top, \mathbf{y}_{(2)}^\top)^\top,$ et $\boldsymbol{\Psi}^{-1} = \mathbf{M}^\top \mathbf{D}_{(2)}^{-1}\mathbf{M} + \boldsymbol{\Omega}.$ Par conséquent, nous pouvons obtenir un échantillon indépendant *a posteriori* par échantillonnage de rejet à partir de (3.1) et d'échantillonnages subséquents à partir de (3.2) et de (3.3). Pour les données sans région non échantillonnée, nous obtenons des procédures d'échantillonnage souhaitées en établissant que $\mathbf{M} = \mathbf{I}_m$ et $\mathbf{y}_* = \mathbf{y}.$

4. Étude par simulations

Dans la présente section, nous comparons les performances de prédiction du modèle FH à effets aléatoires indépendants et des quatre modèles spatiaux en l'absence de covariables informatives ayant de multiples régions non échantillonnées. À l'exclusion d'Hawaï et de l'Alaska, nous prenons en compte les $m = 49$ États contigus des États-Unis, y compris le District de Columbia. Pour évaluer la qualité de la prédiction en l'absence d'estimations directes, nous ne simulons pas d'estimations directes de $m_1 = \lfloor 0{,}15m \rfloor = 7$ États choisis au hasard. Ces régions sont le Delaware, le Massachusetts, le Michigan, le Nebraska, le Rhode Island, le Dakota du Sud et le Texas. Il en résulte $m_2 = 42$ régions qui ont des estimations directes.

Pour rendre les contextes de simulation réalistes, nous reproduisons les données de 1989 sur le revenu médian familial de quatre personnes (revenu médian) décrites à la section 5. Nous générons des ensembles de données répétés de façon à ce que les valeurs I de Moran pour chaque moyenne de petite région répétée, $\theta_1, \ldots, \theta_m$, soient approximativement centrées autour de 0,44, soit la valeur I de Moran pour le revenu médian du recensement de 1990. Les estimations directes sont générées à l'aide des variances d'échantillonnage $D_1, \ldots, D_{m_2}$ des estimations de la Current Population Survey (CPS) de 1990. Ces variances d'échantillonnage de petites régions échantillonnées vont de 1,95 à 25,03 et la moyenne est de 9,08, lorsque les montants en dollars sont mis à l'échelle de 1 000 $. Pour chaque paramètre, nous considérons $S = 100$ ensembles de données répétés.

Génération de données : Supposons que $\bar{D} = m_2^{-1} \sum_{i=1}^{m_2} D_i$ et que $\mathbf{D}_{(2)} = \text{diag}\{D_i\}_{i=1}^{m_2}$. Nous établissons que $\rho =$ 0,85 et $\sigma_v^2 = \bar{D}/2$ et nous considérons deux covariables indépendantes $\mathbf{x}_1$ et $\mathbf{x}_2$ avec une dépendance spatiale SAR, c'est-à-dire $\mathbf{x}_1, \mathbf{x}_2 \sim N_m\left(\mathbf{0}_m, \{\boldsymbol{\Omega}_3(\rho)\}^{-1}\right)$. Ensuite, nous supposons que $(\beta_1, \beta_2)^{\top} = (2,1)^{\top}$ et $\boldsymbol{\mu} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2$, nous générons des moyennes de petites régions et des estimations directes à partir du modèle FH à effets aléatoires indépendants suivant :

$$\boldsymbol{\theta} \sim N_m(\boldsymbol{\mu}, \sigma_v^2 \mathbf{I}_m), \quad \mathbf{Y}_{(2)} \,|\, \boldsymbol{\theta}_{(2)} \sim N_{m_2}(\boldsymbol{\theta}_{(2)}, \mathbf{D}_{(2)}),$$

où les composantes de $\mathbf{Y}_{(2)}$ et $\boldsymbol{\theta}_{(2)}$ correspondent aux m_2 petites régions échantillonnées, tel que nous l'avons défini dans l'équation (2.13) ci-dessous. La covariable $\mathbf{x}_1(\mathbf{x}_2)$ introduit un schéma spatial plus fort (plus faible) par rapport aux θ_i, et par conséquent, nous appelons $\mathbf{x}_1$ ($\mathbf{x}_2$) la covariable forte (faible). Les valeurs I de Moran pour 100 petites régions répétées vont de 0,115 à 0,713, et la moyenne est de 0,449.

Nous considérons deux cadres de covariables différents pour examiner la façon dont les modèles spatiaux peuvent saisir la variabilité supplémentaire introduite par la dépendance spatiale à partir d'une covariable manquante, c'est-à-dire $\mathbf{X} = [\mathbf{1}_m, \mathbf{x}_2]$ et $\mathbf{X} = [\mathbf{1}_m, \mathbf{x}_1]$, où $\mathbf{1}_m$ représente le vecteur à m composantes des valeurs 1. L'exclusion des covariables du modèle ajusté laissera la variation spatiale de cette covariable au niveau résiduel. Nous ne considérons pas que le modèle complet comprend les deux covariables, puisque ce modèle permettra de cerner pleinement $\boldsymbol{\mu}$ et ne laissera aucune variabilité spatiale inexpliquée; par conséquent, le modèle FH à effets aléatoires indépendants sera suffisant pour bien saisir la variabilité des effets aléatoires i.i.d.

Simulations a posteriori : Pour tous les modèles proposés, des échantillons indépendants *a posteriori* peuvent être obtenus au moyen du plan d'échantillonnage de rejet décrit à la section 3. La procédure d'échantillonnage commence par l'échantillonnage de rejet de la distribution marginale *a posteriori* de (σ_v^2, ρ) et se poursuit par des échantillonnages successifs du reste des paramètres issus des distributions conditionnelles *a posteriori*. Cependant, lorsque la densité marginale *a posteriori* de (σ_v^2, ρ) est concentrée au niveau ou autour des limites de ρ, une loi instrumentale doit être soigneusement définie pour avoir un taux d'acceptation suffisamment élevé, ce qui peut nécessiter une spécification adaptative d'une loi instrumentale pour chaque ensemble de données répété. Pour éviter de telles difficultés, nous utilisons l'algorithme Monte Carlo hamiltonien avec le progiciel `rstan` en `R` (Stan Development Team, 2018). Nous adaptons le modèle HB (2.11) à (2.13) à chaque combinaison de covariables pour $k = 1, \ldots, 5$. Pour chaque modèle, nous utilisons quatre chaînes Monte Carlo hamiltoniennes parallèles (échantillonneur sans demi-tour) pour 2 500 itérations après 5 000 itérations de rodage. Nous conservons chaque dixième itération et concaténons les quatre chaînes pour obtenir un échantillon *a posteriori* d'une taille de 10 000. Les codes `R` qui mettent en œuvre l'étape d'échantillonnage de rejet décrite à la section 3 et les modèles `stan` sont disponibles à l'adresse https://github.com/heech31/spatial_sae.

Mesures de la performance : Au moyen de l'échantillon *a posteriori* pour chaque modèle, nous prédisons le vrai vecteur de moyenne sur petites régions, $\boldsymbol{\theta}^{(s)} = (\theta_1^{(s)}, \ldots, \theta_m^{(s)})^\top$, du s^{e} ensemble de données répété en utilisant la moyenne *a posteriori*, que nous désignons par $\hat{\boldsymbol{\theta}}^{(s)} = (\hat{\theta}_1^{(s)}, \ldots, \hat{\theta}_m^{(s)})^\top$. Supposons que $\mathcal{A}$ est un sous-ensemble de $\{1, \ldots, m\}$, qui est déterminé uniquement par les indices de petites régions échantillonnées ou non échantillonnées. Pour un sous-ensemble donné $\mathcal{A}$, nous calculons l'erreur quadratique moyenne de prédiction, $\text{EQMP}^{(s)} = \sum_{i \in \mathcal{A}} (\hat{\theta}_i^{(s)} - \theta_i^{(s)})^2 \big/ m_{\mathcal{A}}$, où $m_{\mathcal{A}} = |\mathcal{A}|$ est le nombre de régions dans $\mathcal{A}$. Nous calculons ensuite l'$\text{EQMP}^{(s)}$ sur S répétitions pour calculer la moyenne empirique de l'erreur quadratique moyenne de prédiction (meEQMP), où :

$$\text{meEQMP} = \frac{1}{S} \sum_{s=1}^{S} \text{EQMP}^{(s)} = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{m_{\mathcal{A}}} \sum_{i \in \mathcal{A}} (\hat{\theta}_i^{(s)} - \theta_i^{(s)})^2. \tag{4.1}$$

Nous évaluons également l'incertitude des prévisions à l'aide de l'écart-type moyen *a posteriori* (ETMP) défini comme $S^{-1} \sum_{s=1}^{S} m_{\mathcal{A}}^{-1} \sum_{i \in \mathcal{A}} \text{é.-t.}(\theta_i^{(s)})$, où $\text{é.-t.}(\theta_i^{(s)})$ est l'écart-type *a posteriori* de θ_i. En établissant le modèle FH à effets aléatoires indépendants comme modèle de référence, nous considérons les ratios suivants :

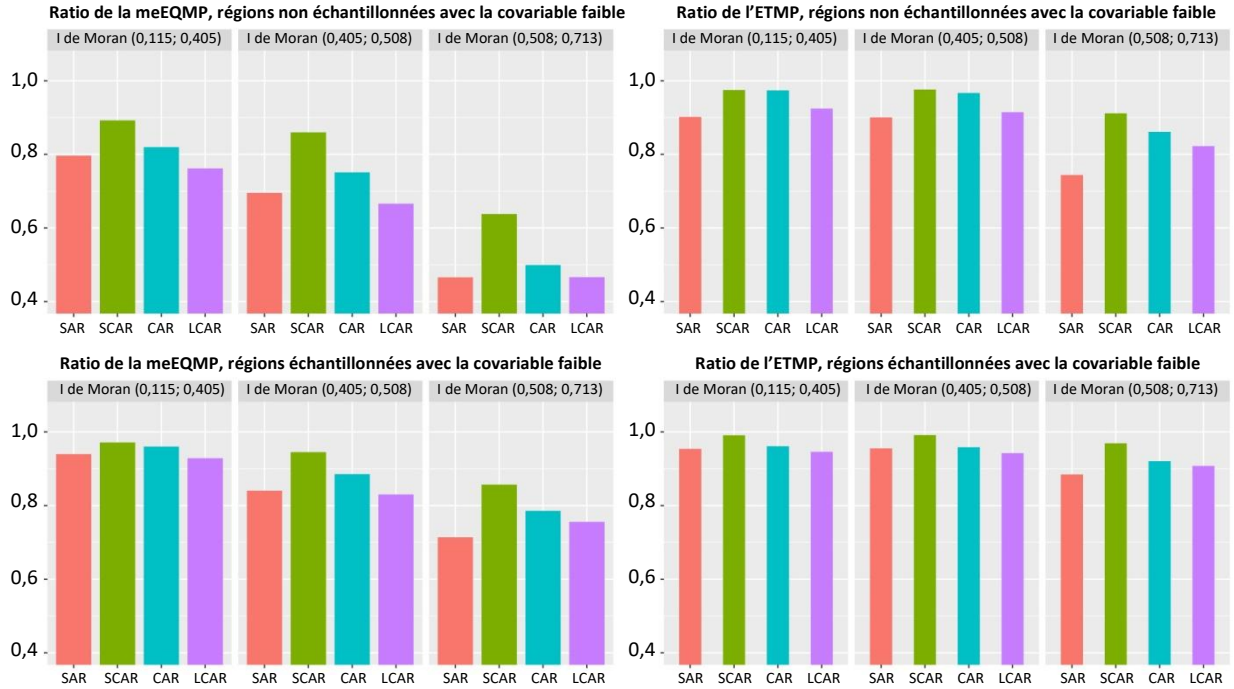
$$\text{meEQMP} - \text{Ratio}_k = \frac{\text{meEQMP}_k}{\text{meEQMP}_1}, \quad \text{ETMP} - \text{Ratio}_k = \frac{\text{ETMP}_k}{\text{ETMP}_1}, \tag{4.2}$$

où l'indice k indique que la quantité est calculée à partir de l'échantillon *a posteriori* selon le k^{e} modèle. Ces ratios mesurent les améliorations de la meEQMP et de l'ETMP obtenues en adaptant un modèle spatial au modèle FH à effets aléatoires indépendants. Un rapport inférieur à 1 indique la supériorité du modèle spatial, sinon le modèle FH à effets aléatoires indépendants est meilleur. Plus le rapport est petit, plus le modèle spatial est supérieur.

Comparaison de modèles : Divers graphiques à la figure 4.1 résument les ratios lorsque la covariable forte $\mathbf{x}_1$ est exclue des modèles ajustés. Nous classons en trois groupes les ratios meEQMP et ETMP selon

chaque répétition, en fonction des valeurs I de Moran des θ_i, c'est-à-dire les tiers inférieur, intermédiaire et supérieur, respectivement. La première ligne résume les résultats des prévisions pour les sept régions non échantillonnées. Comme prévu, les modèles spatiaux montrent une amélioration remarquable de la meEQMP et de l'ETMP; les améliorations sont plus importantes lorsque les valeurs I de Moran sont plus élevées. En ce qui concerne la meEQMP, les modèles SAR et LCAR produisent au moins 20 %, 30 % et 50 % de prédictions plus précises lorsque les valeurs I de Moran se trouvent dans le premier, le deuxième et le troisième groupe, respectivement. Pour ce qui est de l'incertitude de prédiction, les prédictions des modèles SAR et LCAR ont un ETMP 10 % inférieur pour le premier et le deuxième groupe. Dans le troisième groupe, le modèle SAR montre une réduction de l'ETMP de plus de 25 %.

Figure 4.1 Ratio de la moyenne empirique de l'erreur quadratique moyenne de prédiction (meEQMP) et ratio de l'écart-type moyen *a posteriori* (ETMP) pour les prédictions avec la covariable *faible*.



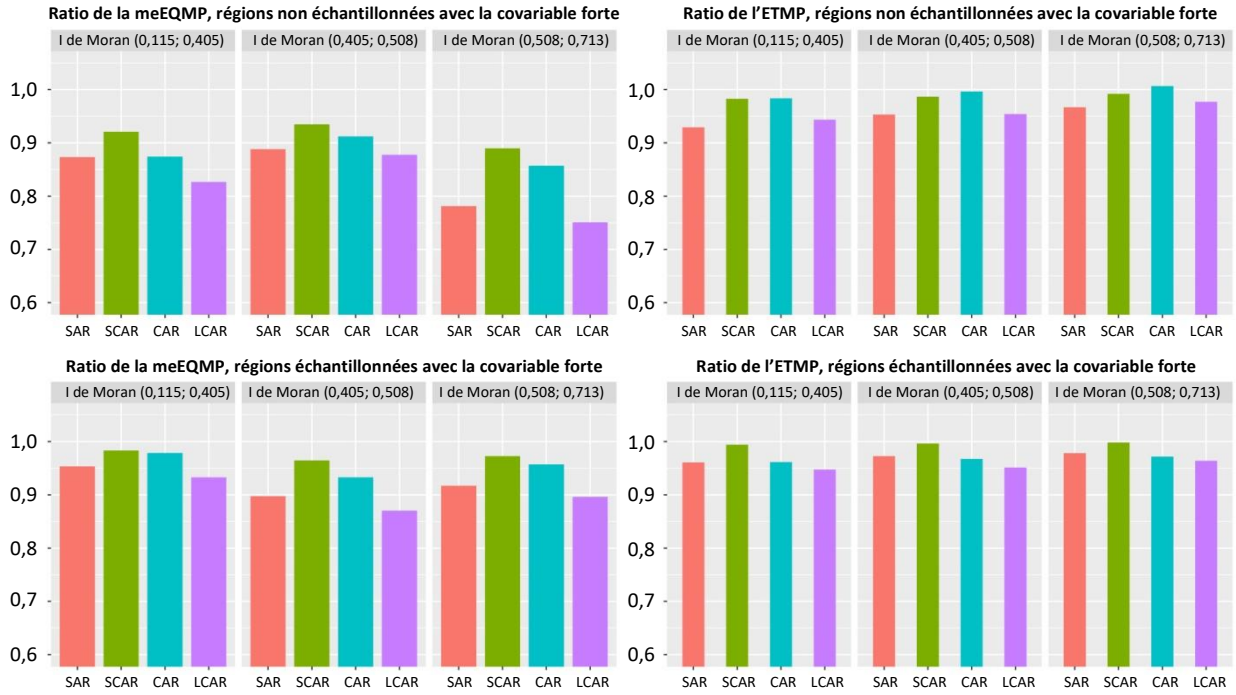
- Des barres verticales sont dessinées pour les ratios : une barre plus courte que 1 représente une meilleure (plus petite) meEQMP ou ETMP pour le modèle correspondant par rapport au modèle Fay-Herriot (FH) à effets aléatoires indépendants.
- SAR = autorégressif simultané, SCAR = autorégressif conditionnel simple, CAR = autorégressif conditionnel, LCAR = autorégressif conditionnel de Leroux, meEQMP = Moyenne empirique de l'erreur quadratique moyenne de prédiction, ETMP = Écart-type moyen *a posteriori*.

Pour les régions échantillonnées ayant des estimations directes, les améliorations de la meEQMP sont inférieures à 10 % pour le premier groupe, mais supérieures à 15 % et à 25 % pour les deuxième et troisième groupes (regroupés selon les valeurs I de Moran), respectivement. De plus, les prévisions des modèles spatiaux ont un niveau d'incertitude plus faible, et pour le troisième groupe, les prévisions des modèles SAR ont un ETMP inférieur de plus de 10 %.

De même, divers graphiques de la figure 4.2 résument les ratios lorsque la covariable faible $\mathbf{x}_2$ est exclue des modèles ajustés. En général, les modèles spatiaux continuent de générer de meilleures prédictions par rapport au modèle FH à effets aléatoires indépendants. Le modèle LCAR donne la

meilleure performance globale pour les sept régions non échantillonnées, ce qui se traduit par une réduction de plus de 10 % de la meEQMP pour le premier et le deuxième groupe, respectivement, et d'environ 25 % pour le troisième groupe. Pour ce qui est de l'incertitude, les modèles spatiaux montrent un ETMP inférieur de plus de 5 %, mais de moins de 10 %. Pour les petites régions échantillonnées, les modèles SAR et LCAR montrent une réduction de la meEQMP d'environ 5 % à 13 %, mais les réductions de l'ETMP sont inférieures à 5 %. Contrairement aux résultats précédents comportant une covariable faible, les améliorations apportées à la meEQMP et à l'ETMP sont comparables dans trois groupes qui sont classés selon les valeurs I de Moran. Ceci est dû au fait que les valeurs I de Moran pour les moyennes de petites régions sont principalement déterminées par la covariable forte, et une fois que celle-ci est présente dans le modèle pour expliquer la variabilité spatiale, la variabilité spatiale des résidus ne varie pas de façon marquée entre les trois groupes. Nous regroupons les ratios en fonction des valeurs I de Moran pour les résidus obtenus par régression de la covariable forte sur les moyennes de petites régions, et nous résumons les ratios à la figure 4.3. Dans le cadre de cette catégorisation, le modèle LCAR présente la meilleure performance, montrant des réductions de 5 %, 10 % et 15 % de la meEQMP pour le premier, le deuxième et le troisième groupe, respectivement. On peut aussi voir que plus la valeur I de Moran est grande, plus les modèles spatiaux permettent des améliorations.

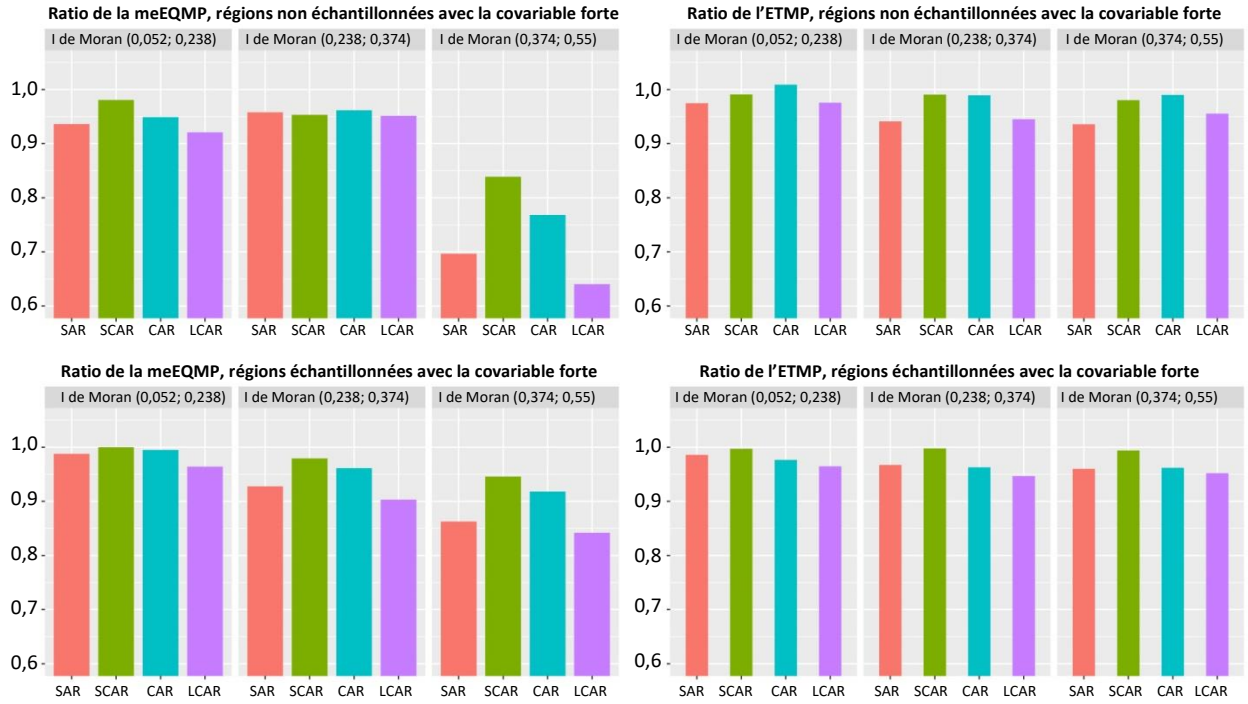
Figure 4.2 Ratio de la moyenne empirique de l'erreur quadratique moyenne de prédiction (meEQMP) et ratio de l'écart-type moyen *a posteriori* (ETMP) pour les prédictions de moyennes de petites régions non échantillonnées et échantillonnées avec la covariable *forte*.



– Des barres verticales sont dessinées pour les ratios : une barre plus courte que 1 représente une meilleure (plus petite) meEQMP ou ETMP pour le modèle correspondant par rapport au modèle Fay-Herriot (FH) à effets aléatoires indépendants.

– SAR = autorégressif simultané, SCAR = autorégressif conditionnel simple, CAR = autorégressif conditionnel, LCAR = autorégressif conditionnel de Leroux, meEQMP = Moyenne empirique de l'erreur quadratique moyenne de prédiction, ETMP = Écart-type moyen *a posteriori*.

Figure 4.3 Ratio de la moyenne empirique de l'erreur quadratique moyenne de prédiction (meEQMP) et ratio de l'écart-type moyen *a posteriori* (ETMP) pour les prédictions de moyennes de petites régions non échantillonnées et échantillonnées avec la covariable *forte*.



- Les résultats sont regroupés selon la valeur I de Moran pour les résidus, où les résidus sont obtenus par une régression de la covariable forte sur les moyennes de petites régions. Des barres verticales sont dessinées pour les ratios : une barre plus courte que 1 représente une meilleure (plus petite) meEQMP ou ETMP pour le modèle correspondant par rapport au modèle Fay-Herriot (FH) à effets aléatoires indépendants.
- SAR = autorégressif simultané, SCAR = autorégressif conditionnel simple, CAR = autorégressif conditionnel, LCAR = autorégressif conditionnel de Leroux, meEQMP = Moyenne empirique de l'erreur quadratique moyenne de prédiction, ETMP = Écart-type moyen *a posteriori*.

5. Application aux données de la Current Population Survey

Dans la présente section, nous évaluons l'exactitude des prédictions des modèles spatiaux pour certains revenus médians de la population au niveau de l'État. Le Department of Health and Human Service (le département de la Santé et des Services sociaux des États-Unis) avait besoin chaque année de données exactes sur les revenus médians pour que les États mettent en œuvre un programme d'aide sociale. Bien que des données exactes sur le revenu médian national soient disponibles à partir de la Current Population Survey (CPS), celles-ci ne fournissent pas de renseignements exacts sur le revenu médian au niveau de l'État. Pour fournir des statistiques exactes au Department of Health and Human Service, le U.S. Census Bureau a envisagé des méthodes d'estimation sur petites régions fondées sur des modèles, au moyen des données auxiliaires provenant d'autres programmes fédéraux. Nous appliquons les modèles spatiaux proposés pour estimer le revenu médian familial de quatre personnes en 1989 pour les 49 États américains contigus, y compris le District de Columbia. Nous utilisons les estimations directes de la CPS de 1990 et comparons nos prédictions avec les statistiques plus fiables issues du *questionnaire complet* du recensement de 1990, c'est-à-dire que nous considérons les statistiques du questionnaire complet du recensement de 1990 comme constituant des valeurs réelles. Les performances de prédiction sont mesurées à l'aide de toutes les petites régions et d'un sous-ensemble de régions après avoir exclu de multiples estimations directes.

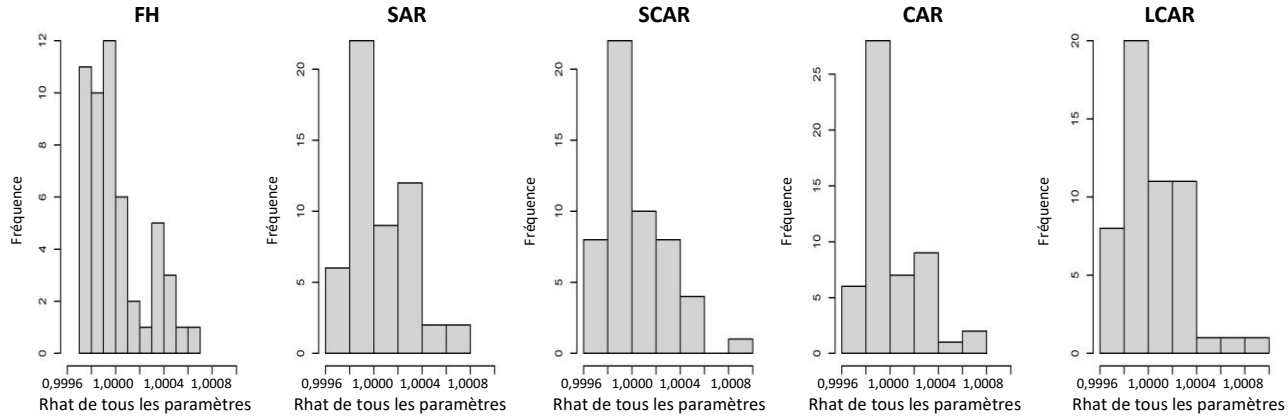
5.1 Estimation du revenu médian familial de quatre personnes

Supposons que θ_i est le véritable revenu médian familial de quatre personnes du i^{e} État pour l'année 1989, où $i = 1, \ldots, 49$. Les États de l'Alaska et d'Hawaï sont exclus parce qu'ils ne sont pas géographiquement rattachés au continent. Supposons que Y_i est l'estimation directe de θ_i selon la CPS de 1990. Les covariables d'intérêt sont le revenu médian du recensement de 1980 x_{i1} et le revenu médian rajusté du recensement de 1980 x_{i2}. Le revenu médian rajusté du recensement x_{i2} est défini comme $\left(\text{PCI}_{i,1989} \,/\, \text{PCI}_{i,1979}\right) x_{i1}$, $i = 1, \ldots, m$, où $\text{PCI}_{i,1979}$ et $\text{PCI}_{i,1989}$ sont les revenus par habitant de 1979 et 1989 du i^{e} État fournis par le Bureau of Economic Analysis du département du Commerce des États-Unis. On sait que le revenu médian rajusté du recensement est une bonne covariable qui tient très efficacement compte de la variabilité du revenu médian des petites régions.

Dans la distribution *a priori* non informative (2.14) où $\alpha = 0$, nous intégrons les cinq modèles décrits en (2.8) à (2.10) avec $\mathbf{X} = [\mathbf{1}_m, \mathbf{x}_1, \mathbf{x}_2]$ et $\mathbf{X} = [\mathbf{1}_m, \mathbf{x}_1]$, où pour le deuxième paramètre de covariable, nous excluons du modèle ajusté le revenu médian rajusté du recensement. Pour chaque modèle considéré, nous exécutons 4 chaînes Monte Carlo hamiltoniennes parallèles pour 2 500 itérations après 5 000 itérations de rodage à l'aide du progiciel `rstan` (Stan Development Team, 2018). Nous conservons chaque 10^{e} itération et concaténons les quatre chaînes pour obtenir un échantillon *a posteriori* de 10 000. Pour l'ensemble des modèles et des paramètres, les facteurs potentiels de réduction d'échelle ($\hat{R}$; Gelman et Rubin, 1992) sont tous des facteurs qui indiquent qu'il n'y a pas de manque de convergence. Les facteurs potentiels de réduction d'échelle sont présentés à la figure 5.1 et au tableau 5.1.

À l'aide des moyennes *a posteriori*, $\hat{\theta}_i$, nous calculons les erreurs quadratiques de prédiction à partir des moyennes respectives θ_i et obtenons l'erreur quadratique moyenne de prédiction (EQMP), définie à la section 4, en faisant la moyenne des $m = 49$ écarts quadratiques. Les écarts-types moyens *a posteriori* (ETMP) associés aux $\hat{\theta}_i$ sont utilisés pour quantifier l'incertitude des prévisions, et le critère d'information largement applicable (WAIC pour *widely applicable information criterion*; Watanabe et Opper, 2010) est utilisé pour évaluer et comparer les modèles, où une valeur WAIC plus petite indique un meilleur ajustement du modèle.

Figure 5.1 Le facteur potentiel de réduction d'échelle $\hat{R}$ de tous les paramètres, lorsqu'il n'y a aucune région non échantillonnée.



- Toutes les valeurs correspondent essentiellement à 1, ce qui indique qu'il n'y a aucune preuve d'absence de convergence.
- FH = Fay-Herriot, SAR = autorégressif simultané, SCAR = autorégressif conditionnel simple, CAR = autorégressif conditionnel, LCAR = autorégressif conditionnel de Leroux.

Tableau 5.1
Le facteur potentiel de réduction d'échelle $\hat{R}$ de l'hyperparamètre σ_v^2 et ρ et la limite de confiance supérieure correspondante de 95 % pour l'ensemble de données sans région non échantillonnée

Hyperparamètre	Covariable comprise	Facteur potentiel de réduction d'échelle (limite de confiance supérieure de 95 %)				
		FH	SAR	SCAR	CAR	LCAR
σ_v^2	x_1, x_2	0,995 (1,017)	0,987 (1,010)	0,995 (1,018)	0,985 (1,007)	1,008 (1,031)
	x_1	0,993 (1,015)	0,991 (1,012)	0,999 (1,022)	0,987 (1,009)	0,987 (1,01)
ρ	x_1, x_2	–	1,005 (1,028)	0,997 (1,023)	0,998 (1,036)	0,994 (1,017)
	x_1	–	1,005 (1,025)	0,997 (1,016)	0,998 (1,017)	1,019 (1,039)

Note : Toutes les limites supérieures se situent à l'intérieur de l'intervalle acceptable, en deçà de 1,1 (Gelman, Carlin, Stern, Dunson, Vehtari et Rubin, 2013, chapitre 11.5), ce qui n'indique aucune preuve d'absence de convergence.
– FH = Fay-Herriot, SAR = autorégressif simultané, SCAR = autorégressif conditionnel simple, CAR = autorégressif conditionnel, LCAR = autorégressif conditionnel de Leroux.

Le tableau 5.2 résume les diverses mesures d'évaluation que nous avons prises en compte et les améliorations en pourcentage (AP) respectives de l'EQMP et de l'ETMP. Lorsque les deux covariables sont disponibles, le modèle LCAR a une EQMP d'environ 14 % inférieure et un ETMP de 4 % inférieure par rapport au modèle FH à effets aléatoires indépendants. En ce qui concerne l'EQMP, le deuxième modèle le plus performant est le SAR qui présente une EQMP inférieure d'environ 9,5 %. Lorsque seule x_1 (covariable de la semaine) est incluse dans le modèle ajusté, le modèle SAR a une erreur quadratique moyenne *a posteriori* d'environ 40 % inférieure et un écart-type moyen *a posteriori* de 14 % inférieur par rapport au modèle FH à effets aléatoires indépendants. Les modèles CAR et LCAR montrent des performances concurrentielles avec une EQMP d'environ 36 % inférieure et un ETMP de 13 % inférieur par rapport au modèle FH à effets aléatoires indépendants. En supprimant la covariable forte x_2 du modèle complet, l'EQMP des modèles SAR et LCAR augmente respectivement d'environ 66 % et 84 %, respectivement, tandis que l'EQMP du modèle FH à effets aléatoires indépendants augmente de plus de 150 %.

Tableau 5.2
Erreur quadratique moyenne de prédiction (EQMP), écart-type moyen *a posteriori* (ETMP), et amélioration en pourcentage (AP) des modèles spatiaux par rapport au modèle Fay-Herriot (FH) à effets aléatoires indépendants

Covariable comprise	x_1, x_2					x_1				
	EQMP	EQMP-AP	ETMP	ETMP-AP	WAIC	EQMP	EQMP-AP	ETMP	ETMP-AP	WAIC
FH	2,88	–	1,93	–	*259,06 (7,13)*	7,27	–	2,31	–	267,75 (8,44)
SAR	2,61	*9,55 %*	1,94	0,34 %	261,46 (7,47)	4,34	*40,22 %*	1,98	*14,25 %*	265,76 (8,16)
SCAR	3,03	-5,14 %	1,95	-0,91 %	259,37 (7,01)	5,62	22,62 %	2,22	3,52 %	263,41 (7,29)
CAR	2,64	8,47 %	1,91	1,24 %	261,61 (7,86)	4,62	*36,35 %*	2,01	*12,97 %*	*263,32 (7,96)*
LCAR	2,47	*14,50 %*	1,85	*4,19 %*	261,79 (8,01)	4,54	*37,51 %*	1,97	*14,36 %*	263,35 (8,08)

– WAIC = critère d'information largement applicable, SAR = autorégressif simultané, SCAR = autorégressif conditionnel simple, CAR = autorégressif conditionnel, LCAR = autorégressif conditionnel de Leroux.

Pour ce qui est de la qualité de l'ajustement, le modèle FH à effets aléatoires indépendants montre le meilleur ajustement (la plus petite valeur WAIC) lorsque les deux covariables sont comprises. À l'inverse, lorsque seule x_1 (covariable de la semaine) est comprise dans le modèle ajusté, le modèle FH à effets aléatoires indépendants montre le meilleur ajustement ayant la plus grande valeur WAIC. Toutefois, compte tenu des erreurs-types indiquées entre parenthèses, il n'y a pas de différence significative dans l'ajustement du modèle.

Le tableau 5.3 résume les distributions *a posteriori* de ρ en ce qui concerne la moyenne *a posteriori*, le mode et l'écart-type. Lorsque toutes les covariables sont comprises dans le modèle ajusté, les distributions *a posteriori* de ρ n'indiquent pas une forte dépendance spatiale, les moyennes *a posteriori* étant centrées autour de zéro avec de grands écarts-types. En revanche, lorsque seule x_1 (covariable de la semaine) est comprise dans le modèle ajusté, la valeur ρ devient très significative, illustrant les distributions *a posteriori* concentrées près de la limite supérieure de sa corroboration.

En résumé, lorsque la faible covariable n'explique pas adéquatement la variation spatiale existante, les modèles spatiaux produisent de bien meilleures prédictions dans lesquelles la variation spatiale est prise en compte. Lorsqu'il ne reste aucune variation spatiale importante dans le résidu, les prédictions sont légèrement meilleures que celles du modèle FH à effets aléatoires indépendants, sans qu'il soit nécessaire de sacrifier l'ajustement du modèle.

Tableau 5.3
Moyenne et mode *a posteriori* (écart-type) de ρ

Covariable comprise	SAR	SCAR	CAR	LCAR
x_1, x_2	0,10/0,40 (0,48)	-0,06/0,04 (0,14)	0,21/0,83 (0,55)	0,57/0,80 (0,27)
x_1	0,76/0,80 (0,14)	0,14/0,17 (0,04)	0,93/0,99 (0,09)	0,85/0,97 (0,13)

Note : La première et la deuxième ligne du tableau résument respectivement les distributions *a posteriori* de ρ lorsque les deux covariables sont comprises et lorsque seule x_1 est comprisedans le modèle ajusté.

– SAR = autorégressif simultané, SCAR = autorégressif conditionnel simple, CAR = autorégressif conditionnel, LCAR = autorégressif conditionnel de Leroux.

5.2 Estimation de certaines moyennes d'États non échantillonnés en excluant leurs valeurs de la Current Population Survey

Dans la présente section, nous évaluons l'exactitude des prévisions des modèles spatiaux pour les petites régions non échantillonnées en utilisant le revenu médian du recensement de 1980 x_1. Plus précisément, nous excluons aléatoirement les estimations de la CPS (estimations directes) de plusieurs États à chaque cas et faisons des prédictions pour les θ_i des États exclus. Comme il y a 49 petites régions (États), nous avons créé 12 ensembles de données qui n'ont pas d'estimations directes pour $m_1 = 4$ ou 5 régions, où m_1 est le nombre de petites régions non échantillonnées, comme à la section 2.2. Les États exclus pour chaque ensemble de données sont énumérés au tableau 5.4.

Pour chaque ensemble de données, nous intégrons le modèle FH à effets aléatoires indépendants et les quatre modèles spatiaux indiqués en (2.11) à (2.13) à la distribution *a priori* non informative (2.13), $\alpha = 0$ exécutant des chaînes Monte Carlo hamiltoniennes dans le même contexte qu'à la section 5.1. Les valeurs $\hat{R}$ ne montrent aucune preuve d'absence de convergence, alors que des valeurs détaillées sont fournies à la figure 5.2 et au tableau 5.5. Pour chaque région non échantillonnée, l'erreur quadratique de prédiction $\text{EQP}_i = (\hat{\theta}_i - \theta_i)^2$ et l'écart-type *a posteriori* é.-t.(θ_i) sont obtenus pour chaque modèle. Si nous nous appuyons sur ce qui précède, nous constatons que la performance de prédiction est comparée au ratio suivant : pour $i = 1, \ldots, m$ et $k = 2, \ldots, 5$,

$$\text{EQP} - \text{Ratio}_{ki} = \frac{\text{EQP}_{ki}}{\text{EQP}_{1i}}, \quad \text{ETP} - \text{Ratio}_{ki} = \frac{\text{é.-t.}_k(\theta_i)}{\text{é.-t.}_1(\theta_i)} \tag{5.1}$$

où é.-t.$_k$ (θ_i) est l'écart-type *a posteriori* (ETP) de θ_i selon le k^{e} modèle. Une valeur de $\text{SEQP} - \text{Ratio}_{ki}$ ($\text{ETP} - \text{Ratio}_{ki}$) inférieure à 1 indique que le k^{e} modèle spatial présente une erreur quadratique de prédiction plus petite (écart-type *a posteriori*) que le modèle FH à effets aléatoires indépendants. Dans les figures 5.3 et 5.4, nous affichons les ratios en utilisant un schéma de couleurs rouge et bleu pour indiquer les ratios supérieurs et inférieurs à 1, respectivement, où une couleur plus foncée représente une valeur plus élevée.

Dans l'ensemble, les modèles SAR, SCAR, CAR et LCAR présentent des erreur quadratique de prédiction (EQP) plus petites dans 35, 41, 36 et 36 États, respectivement. Le modèle SCAR présente le plus grand nombre d'États dans lesquels les prédictions donnent de meilleurs résultats que celles du modèle FH à effets aléatoires indépendants, mais les améliorations globales sont les moins importantes. Dans plus de 35 États, les modèles SAR, CAR et LCAR produisent des prédictions plus précises que le modèle FH à effets aléatoires indépendants, et dans trois États (Nouveau-Mexique, Oregon et Wisconsin), les EQP sont plus de 100 fois plus petites. Pour la Californie, le Minnesota et la Caroline du Sud, tous les modèles spatiaux livrent des prédictions pires que le modèle FH à effets aléatoires indépendants. La Californie et le Minnesota ont des revenus médians beaucoup plus élevés que les États voisins, tandis que la Caroline du Sud a un revenu médian beaucoup plus bas. Parmi les 49 États, ces trois États ont la deuxième, la septième et la dix-neuvième valeur I de Moran locale la plus petite. Cela montre que si la moyenne de petites régions non échantillonnées est très différente de la moyenne de régions environnantes, les modèles spatiaux peuvent produire des prédictions inférieures. Le modèle qui démontre la meilleure correspondance en ce qui concerne le WAIC est le modèle SCAR, CAR ou LCAR; les chiffres exacts sont fournis au tableau 5.4.

Tableau 5.4
États dont les estimations de la Current Population Survey sont exclues pour chaque ensemble de données et critère d'information largement applicable (WAIC pour widely applicable information criterion) correspondant

États exclus	FH	SAR	SCAR	CAR	LCAR
AZ MS OK SD	246,15 (7,79)	244,37 (7,50)	*241,71 (6,67)*	242,46 (7,50)	242,21 (7,57)
AR CO DE TN	245,79 (7,83)	241,23 (7,69)	241,05 (6,70)	*239,70 (7,61)*	239,85 (7,69)
MD MI NV WV	246,06 (7,83)	242,23 (7,58)	241,34 (6,78)	240,13 (7,32)	*240,02 (7,43)*
MT NC NE NY	248,91 (8,10)	245,97 (9,51)	243,85 (6,87)	242,88 (8,34)	*242,40 (8,36)*
DC GA ID ND	245,07 (7,91)	241,02 (6,61)	240,09 (6,44)	*239,16 (6,75)*	239,48 (6,75)
AL MO VT WY	245,57 (8,31)	245,13 (7,67)	*242,74 (7,38)*	242,89 (7,65)	243,36 (7,66)
FL LA UT WA	247,38 (7,39)	245,64 (7,29)	*243,08 (6,44)*	243,25 (7,20)	243,48 (7,33)
MA MN SC TX	248,32 (9,73)	242,43 (8,95)	244,75 (8,83)	240,99 (8,69)	*240,34 (8,56)*
KY RI VA WI	243,86 (8,09)	241,74 (7,06)	240,05 (6,76)	239,04 (6,91)	*238,87 (6,90)*
IL IN NH PA	244,69 (7,10)	245,12 (7,41)	*243,31 (6,59)*	244,45 (7,60)	244,39 (7,68)
CA ME NJ OH	248,62 (8,54)	244,06 (8,26)	245,28 (7,68)	*242,00 (7,78)*	242,01 (7,90)
CT IA KS NM OR	239,86 (7,95)	239,81 (7,76)	*237,28 (7,29)*	237,78 (7,75)	237,75 (7,60)

Note : Les chiffres entre parenthèses sont les erreurs-types des estimations du WAIC.

– FH = Fay-Herriot, SAR = autorégressif simultané, SCAR = autorégressif conditionnel simple, CAR = autorégressif conditionnel, LCAR = autorégressif conditionnel de Leroux.

– AL = Alabama, AZ = Arizona, AR = Arkansas, CA = Californie, CO = Colorado, CT = Connecticut, DE = Delaware, DC = District de Columbia, FL = Floride, GA = Géorgie, ID = Idaho, IL = Illinois, IN = Indiana, IA = Iowa, KS = Kansas, KY = Kentucky, LA = Louisiane, ME = Maine, MD = Maryland, MA = Massachusetts, MI = Michigan, MN = Minnesota, MS = Mississippi, MO = Missouri, MT = Montana, NE = Nebraska, NV = Nevada, NH = New Hampshire, NJ = New Jersey, NM = Nouveau-Mexique, NY = New York, NC = Caroline du Nord, ND = Dakota du Nord, OH = Ohio, OK = Oklahoma, OR = Oregon, PA = Pennsylvanie, RI = Rhode Island, SC = Caroline du Sud, SD = Dakota du Sud, TN = Tennessee, TX = Texas, UT = Utah, VT = Vermont, VA = Virginie, WA = Washington, WV = Virginie-Occidentale, WI = Wisconsin, WY = Wyoming.

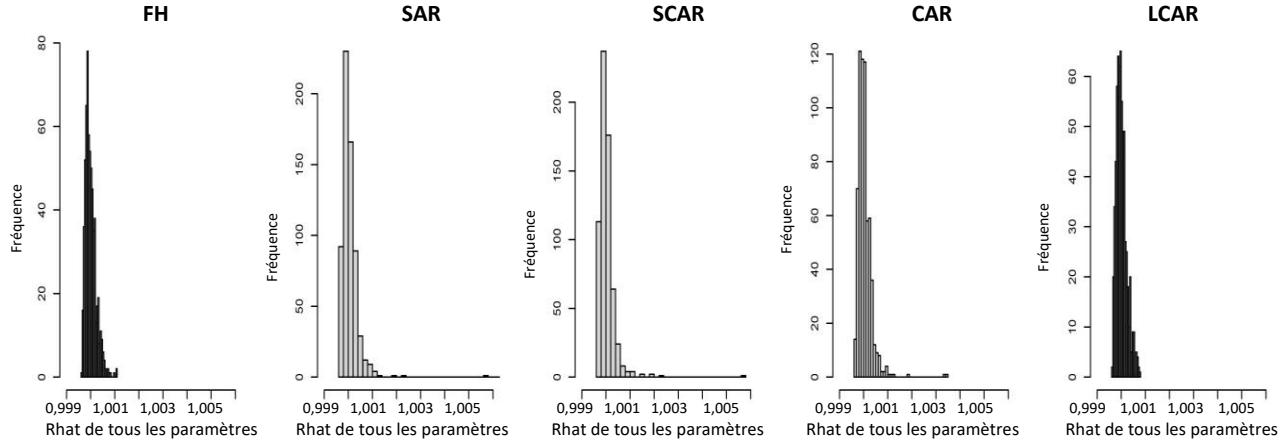
Tableau 5.5
Le facteur potentiel de réduction d'échelle $\hat{R}$ de l'hyperparamètre σ_v^2 et ρ et la limite de confiance supérieure correspondante de 95 % pour les 12 ensembles de données avec régions non échantillonnées

	États exclus	FH	SAR	SCAR	CAR	LCAR
Facteur potentiel de réduction d'échelle (limite de confiance supérieure de 95 %) de σ_v^2	AZ MS OK SD	0,991 (1,013)	1,012 (1,036)	0,995 (1,017)	1,005 (1,028)	1,005 (1,030)
	AR CO DE TN	0,997 (1,019)	1,010 (1,034)	0,999 (1,023)	0,985 (1,008)	0,984 (1,007)
	MD MI NV WV	0,992 (1,015)	1,007 (1,030)	1,023 (1,047)	1,001 (1,025)	1,011 (1,035)
	MT NC NE NY	0,990 (1,012)	0,984 (1,009)	0,997 (1,020)	0,999 (1,024)	0,983 (1,006)
	DC GA ID ND	1,002 (1,024)	1,007 (1,032)	1,001 (1,024)	1,011 (1,036)	0,997 (1,020)
	AL MO VT WY	0,997 (1,019)	1,006 (1,030)	0,998 (1,021)	1,002 (1,026)	1,019 (1,045)
	FL LA UT WA	1,014 (1,037)	1,005 (1,027)	1,002 (1,025)	0,999 (1,023)	1,019 (1,043)
	MA MN SC TX	1,008 (1,031)	1,001 (1,025)	1,003 (1,025)	0,994 (1,019)	1,007 (1,032)
	KY RI VA WI	1,011 (1,034)	1,005 (1,029)	1,009 (1,033)	0,989 (1,011)	1,011 (1,035)
	IL IN NH PA	0,992 (1,013)	1,018 (1,041)	0,992 (1,013)	0,989 (1,014)	0,989 (1,012)
	CA ME NJ OH	1,002 (1,025)	1,009 (1,033)	1,001 (1,024)	1,011 (1,035)	1,003 (1,026)
	CT IA KS NM OR	0,997 (1,018)	1,008 (1,032)	1,007 (1,030)	1,009 (1,034)	1,012 (1,041)
Facteur potentiel de réduction d'échelle (limite de confiance supérieure de 95 %) de ρ	AZ MS OK SD	–	1,005 (1,026)	1,001 (1,029)	1,004 (1,040)	0,995 (1,018)
	AR CO DE TN	–	1,019 (1,042)	0,993 (1,018)	1,000 (1,034)	1,005 (1,028)
	MD MI NV WV	–	1,005 (1,027)	0,992 (1,017)	0,988 (1,021)	1,002 (1,026)
	MT NC NE NY	–	1,001 (1,024)	0,999 (1,030)	0,986 (1,023)	0,998 (1,023)
	DC GA ID ND	–	0,999 (1,020)	0,995 (1,024)	0,991 (1,026)	1,000 (1,024)
	AL MO VT WY	–	1,003 (1,026)	1,011 (1,037)	0,989 (1,025)	1,008 (1,031)
	FL LA UT WA	–	1,004 (1,025)	0,996 (1,026)	1,000 (1,033)	1,005 (1,029)
	MA MN SC TX	–	1,002 (1,041)	1,014 (1,045)	0,984 (1,018)	0,996 (1,02)
	KY RI VA WI	–	1,010 (1,032)	0,990 (1,016)	0,993 (1,031)	0,986 (1,009)
	IL IN NH PA	–	1,030 (1,055)	0,993 (1,016)	0,994 (1,029)	1,003 (1,025)
	CA ME NJ OH	–	1,003 (1,026)	0,997 (1,026)	0,996 (1,032)	0,989 (1,012)
	CT IA KS NM OR	–	1,017 (1,040)	0,997 (1,022)	1,009 (1,044)	0,981 (1,003)

Note : Toutes les limites supérieures se situent à l'intérieur de la plage acceptable, en deçà de 1,1 (Gelman et coll., 2013, chapitre 11.5), ce qui n'indique aucune preuve d'absence de convergence.

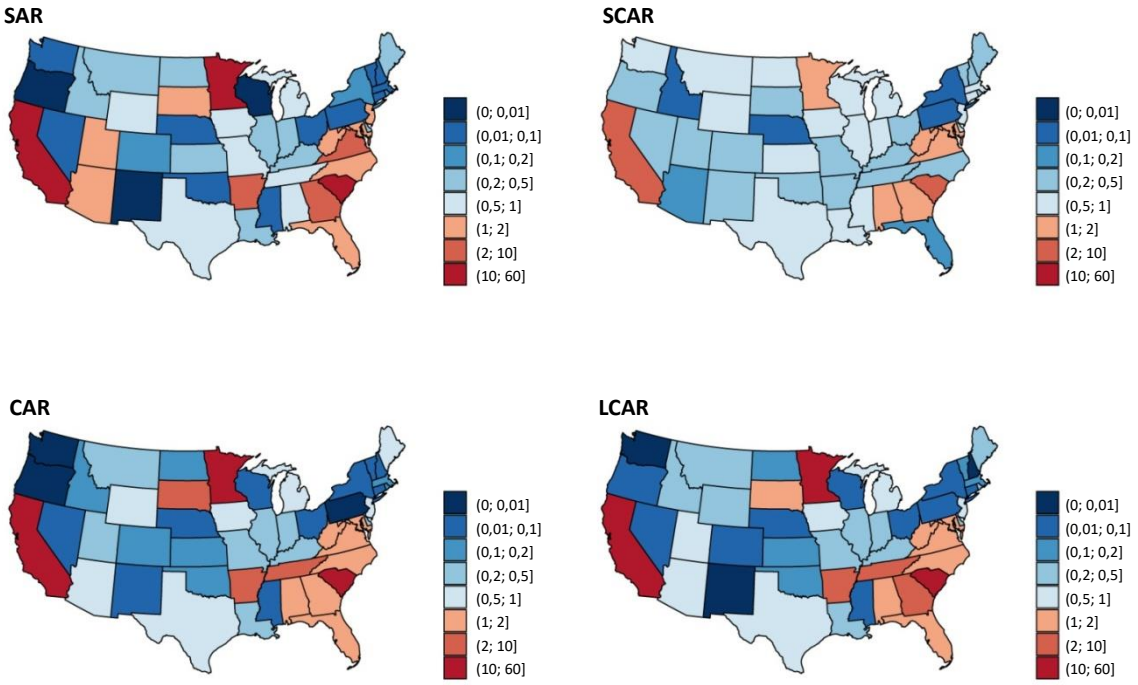
- FH = Fay-Herriot, SAR = autorégressif simultané, SCAR = autorégressif conditionnel simple, CAR = autorégressif conditionnel, LCAR = autorégressif conditionnel de Leroux.
- AL = Alabama, AZ = Arizona, AR = Arkansas, CA = Californie, CO = Colorado, CT = Connecticut, DE = Delaware, DC = District de Columbia, FL = Floride, GA = Géorgie, ID = Idaho, IL = Illinois, IN = Indiana, IA = Iowa, KS = Kansas, KY = Kentucky, LA = Louisiane, ME = Maine, MD = Maryland, MA = Massachusetts, MI = Michigan, MN = Minnesota, MS = Mississippi, MO = Missouri, MT = Montana, NE = Nebraska, NV = Nevada, NH = New Hampshire, NJ = New Jersey, NM = Nouveau-Mexique, NY = New York, NC = Caroline du Nord, ND = Dakota du Nord, OH = Ohio, OK = Oklahoma, OR = Oregon, PA = Pennsylvanie, RI = Rhode Island, SC = Caroline du Sud, SD = Dakota du Sud, TN = Tennessee, TX = Texas, UT = Utah, VT = Vermont, VA = Virginie, WA = Washington, WV = Virginie-Occidentale, WI = Wisconsin, WY = Wyoming.

Figure 5.2 Le facteur potentiel de réduction d'échelle $\hat{R}$ de tous les paramètres, lorsque les valeurs $\hat{R}$ des 12 ensembles de données sont toutes combinées.



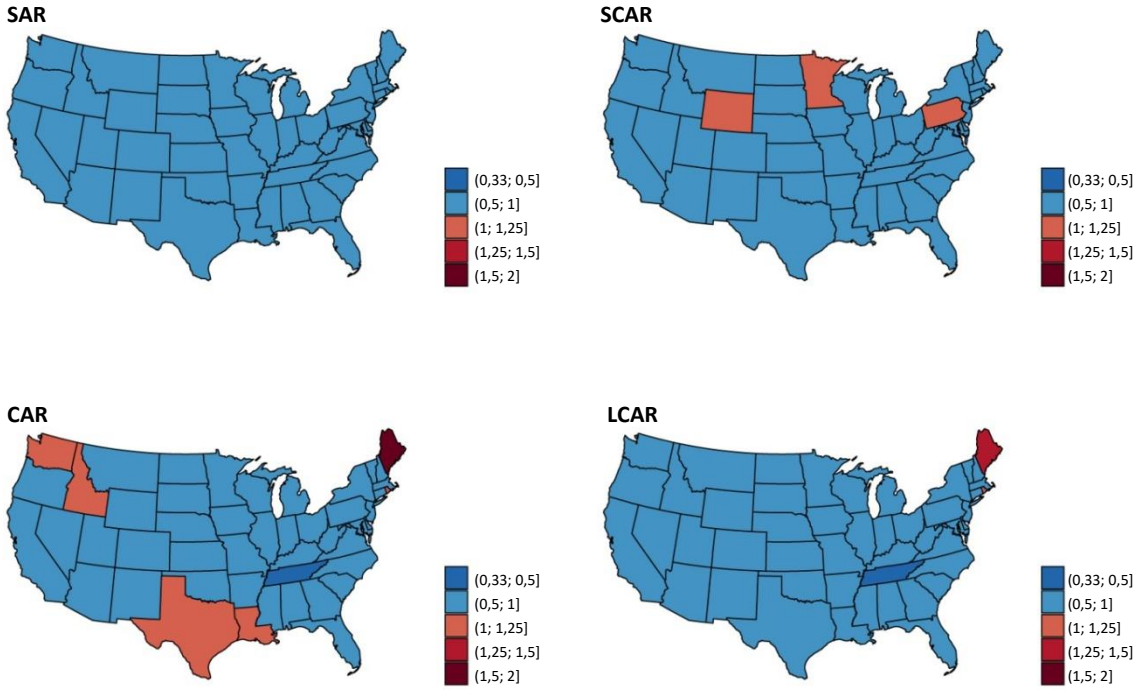
- Toutes les valeurs correspondent essentiellement à 1, ce qui indique qu'il n'y a aucune preuve d'absence de convergence.
- FH = Fay-Herriot, SAR = autorégressif simultané, SCAR = autorégressif conditionnel simple, CAR = autorégressif conditionnel, LCAR = autorégressif conditionnel de Leroux.

Figure 5.3 Ratios de l'erreur quadratique de prédiction.



- Les valeurs inférieures (supérieure) à 1 sont représentées en bleu (rouge) et indiquent un modèle spatial ayant un écart quadratique plus petit (plus grand).
- SAR = autorégressif simultané, SCAR = autorégressif conditionnel simple, CAR = autorégressif conditionnel, LCAR = autorégressif conditionnel de Leroux.

Figure 5.4 Ratios de l'écart-type *a posteriori*.



- Les valeurs inférieures (supérieure) à 1 sont représentées en bleu (rouge) et indiquent un modèle spatial ayant un écart-type *a posteriori* plus petit (plus grand).
- SAR = autorégressif simultané, SCAR = autorégressif conditionnel simple, CAR = autorégressif conditionnel, LCAR = autorégressif conditionnel de Leroux.

6. Conclusions

Dans le présent article, nous avons suivi une approche bayésienne pour étudier quatre modèles spatiaux à effets aléatoires comme solutions de rechange au modèle indépendant de Fay-Herriot pour estimer les moyennes de petites régions. En particulier, nous avons examiné quatre modèles spatiaux avec différentes structures d'autocorrélation. Nous avons élargi les modèles spatiaux afin de permettre la présence de multiples petites régions sans estimations directes pour prédire les moyennes de petites régions pour toutes les régions. Pour une classe de distributions *a priori* non informatives, nous avons établi la pertinence des densités *a posteriori* des modèles proposés pour les deux configurations.

Une étude de simulation à la section 4 montre que l'exactitude des prévisions peut être grandement améliorée si l'on tient compte des modèles spatiaux lorsque des covariables efficaces ne sont pas disponibles. Datta, Hall et Mandal (2011) ont souligné que l'exactitude des prévisions des modèles d'estimation sur petites régions dépend en grande partie de la disponibilité de bonnes covariables. Autrement dit, lorsque des covariables appropriées ne sont pas disponibles, le modèle indépendant de Fay-Herriot pourrait ne pas offrir un avantage important par rapport aux estimations directes. Les résultats de la simulation ont indiqué que, dans de tels cas, les modèles spatiaux permettent d'augmenter considérablement la précision des prédictions en exploitant l'information provenant de régions adjacentes.

Nous avons appliqué les modèles spatiaux à effets aléatoires proposés pour estimer le revenu médian familial de quatre personnes. Même en présence d'une bonne covariable, les modèles spatiaux ont montré des améliorations notables quant à l'écart quadratique moyen et aux écarts-types moyens *a posteriori*. Lorsqu'une bonne covariable n'est pas disponible, les modèles spatiaux fournissent des prédictions du revenu médian beaucoup plus précises ayant une variabilité beaucoup plus faible, ce qui concorde avec les résultats de la simulation. De plus, les modèles SAR et LCAR fournissent des estimations plus précises sur petites régions lorsque les estimations directes de certains États sont exclues dans l'ajustement du modèle.

En résumé, les modèles spatiaux examinés dans le présent article donnent de meilleurs résultats que le modèle indépendant de Fay-Herriot. On peut s'attendre à une amélioration importante lorsque des covariables efficaces ne sont pas disponibles. Comme les covariables utiles ne sont pas toujours disponibles, l'utilité des modèles proposés pour l'estimation sur petites régions peut être substantielle. Notre étude de simulation et notre analyse de données réelles ne montrent aucun gagnant clair parmi les modèles proposés. Néanmoins, les modèles SAR et LCAR montrent une meilleure performance que d'autres modèles spatiaux. De plus, le modèle LCAR fonctionne très bien avec des données simulées issues du modèle SAR et des données réelles dont la dépendance spatiale est inconnue. Par conséquent, dans le contexte d'applications réelles où la véritable dépendance est inconnue, nous recommandons le modèle LCAR.

Les travaux de la présente étude supposent que toutes les régions comptent au moins un quartier. Dans les applications réelles, cependant, il y a de nombreuses situations où les données contiennent de petites régions sans quartiers (régions autonomes). Bien que les modèles proposés puissent prendre en compte des

régions autonomes en ajustant les entrées diagonales des matrices de précision comme dans Brown, Datta et Lazar (2017), nous constatons que cette approche donne lieu à une distribution *a priori* contre-intuitive, quand les régions autonomes ont des variances *a priori* des effets aléatoires plus petites que les régions comptant des quartiers. De plus, nous constatons que cette distribution *a priori* peut considérablement altérer les prédictions des régions autonomes. C'est un problème important dans la pratique, car de nombreux pays ont des îles; il s'agira peut-être de notre prochain filon de recherche.

Avis de non-responsabilité et remerciements

Le présent rapport vise à informer les parties intéressées des travaux de recherche en cours et à favoriser la discussion. Les opinions exprimées sur les questions statistiques, méthodologiques, techniques ou opérationnelles sont celles des auteurs et non celles du U.S. Census Bureau ou de l'université de Géorgie. Les auteurs tiennent à remercier William R. Bell pour ses commentaires judicieux au sujet d'une version antérieure de ces travaux qui ont mené à l'amélioration du manuscrit.

Annexe

A. Preuve de la fonction de densité de probabilité *a posteriori*

Preuve du théorème 1. Pour des raisons de commodité, nous désignons $\mathbf{\Omega}_k(\rho)$ par $\mathbf{\Omega}_k$ et, pour une matrice quadratique donnée $\mathbf{A}$, le déterminant de $\mathbf{A}$ est indiqué par $|\mathbf{A}|$. Nous utilisons K pour désigner une constante positive générique, indépendante des variables que nous intégrons.

Supposons que $m_1 \geq 0$ est le nombre de petites régions sans estimation directe et supposons que $m_2 = m - m_1$. De plus, supposons que $\mathbf{Y}_{(2)}$ est le vecteur $m_2 \times 1$ comptant des estimations directes des petites régions échantillonnées. Sans perte de généralité, nous supposons que $\theta_1, \ldots, \theta_m$ sont disposés de sorte que $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)}^\top, \boldsymbol{\theta}_{(2)}^\top)^\top$. Supposons que $\mathbf{D}_{(2)} = \{D_i\}_{i=m_1+1}^{m}$ est la matrice diagonale avec variances d'échantillonnage correspondant aux composantes de $\mathbf{Y}_{(2)}$ et $\delta = \max_{m_1 < i \leq m} D_i < \infty$.

La fonction de densité de probablité (fdp) mixte de $\mathbf{Y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2$ et ρ est donnée par :

$$f(\mathbf{y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho) = N_{m_2}(\mathbf{y}_{(2)} \,|\, \boldsymbol{\theta}_{(2)}, \mathbf{D}_{(2)})\, N_m(\boldsymbol{\theta} \,|\, \mathbf{X}\boldsymbol{\beta}, \sigma_v^2 \mathbf{\Omega}^{-1})\, g(\sigma_v^2)\, h(\rho), \tag{A.1}$$

où $N_{m_2}(\mathbf{y}_{(2)} \,|\, \boldsymbol{\theta}_{(2)}, \mathbf{D}_{(2)})$ est la fdp normale comportant une moyenne $\boldsymbol{\theta}_{(2)}$ et une matrice de covariance $\mathbf{D}_{(2)}$. La fdp *a posteriori* $\pi(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho \,|\, \mathbf{y}_{(2)})$ sera adéquate si et seulement si la fonction $f(\mathbf{y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho)$ est intégrable relativement à $\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2$ et ρ. Puisque :

$$N_{m_2}(\mathbf{y}_{(2)} \,|\, \boldsymbol{\theta}_{(2)}, \mathbf{D}_{(2)}) \leq K \exp\left\{-\frac{1}{2\delta}(\mathbf{y}_{(2)} - \boldsymbol{\theta}_{(2)})^\top (\mathbf{y}_{(2)} - \boldsymbol{\theta}_{(2)})\right\},$$

nous avons, selon (A.1) :

$$f(\mathbf{y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho) \;\le\; K \exp\left\{-\frac{1}{2\delta}(\mathbf{y}_{(2)} - \boldsymbol{\theta}_{(2)})^\top (\mathbf{y}_{(2)} - \boldsymbol{\theta}_{(2)})\right\} N_m(\boldsymbol{\theta} \mid \mathbf{X}\boldsymbol{\beta}, \sigma_v^2 \boldsymbol{\Omega}^{-1})\, g(\sigma_v^2)\, h(\rho)$$

$$= K \int \exp\left\{-\frac{1}{2\delta}(\mathbf{y} - \boldsymbol{\theta})^\top (\mathbf{y} - \boldsymbol{\theta})\right\} d\mathbf{y}_{(1)}\, N_m(\boldsymbol{\theta} \mid \mathbf{X}\boldsymbol{\beta}, \sigma_v^2 \boldsymbol{\Omega}^{-1})\, g(\sigma_v^2)\, h(\rho), \quad \text{(A.2)}$$

où $\mathbf{y} = (\mathbf{y}_{(1)}^\top, \mathbf{y}_{(2)}^\top)^\top$. En intégrant les deux côtés de (A.2) relativement à $\boldsymbol{\theta}$, nous obtenons :

$$\int f(\mathbf{y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho)\, d\boldsymbol{\theta} \;\le\; Kg(\sigma_v^2)\, h(\rho) \int N_m(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta}, \delta \mathbf{I}_m + \sigma_v^2 \boldsymbol{\Omega}^{-1})\, d\mathbf{y}_{(1)}. \quad \text{(A.3)}$$

Subdivisons $\mathbf{X}$ en tant que $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]^\top$, où $\mathbf{X}_1^\top$ est $m_1 \times p$ et $\mathbf{X}_2^\top$ est $m_2 \times p$. Nous supposons que la ligne $(\mathbf{X}_2) = p$. Supposons que $\mathbf{d} = (\mathbf{0}_{m_1}^\top, \mathbf{y}_{(2)}^\top)^\top$, $\boldsymbol{\phi} = (\mathbf{y}_{(1)}^\top, \boldsymbol{\beta}^\top)^\top$ et

$$\mathbf{G} \;=\; \begin{bmatrix} -\mathbf{I}_{m_1} & \mathbf{X}_1^\top \\ \mathbf{0}_{m_2, m_1} & \mathbf{X}_2^\top \end{bmatrix}.$$

Alors nous pouvons utiliser la notation :

$$\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{d} - \mathbf{G}\boldsymbol{\phi},$$

où $\mathbf{G}$ est $m \times (m_1 + p)$, $\boldsymbol{\phi}$ est $(m_1 + p) \times 1$. Donc, l'équation (A.3) peut être formulée comme suit :

$$\int f(\mathbf{y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho)\, d\boldsymbol{\theta} \;\le\; Kg(\sigma_v^2)\, h(\rho) \int N_m(d \mid \mathbf{G}\boldsymbol{\phi}, \delta \mathbf{I}_m + \sigma_v^2 \boldsymbol{\Omega}^{-1})\, d\mathbf{y}_{(1)}. \quad \text{(A.4)}$$

En intégrant les deux côtés de (A.4) relativement à $\boldsymbol{\beta}$, nous obtenons :

$$\int f(\mathbf{y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho)\, d\boldsymbol{\theta}\, d\boldsymbol{\beta} \;\le\; Kg(\sigma_v^2)\, h(\rho) \int N_m(\mathbf{d} \mid \mathbf{G}\boldsymbol{\phi}, \delta \mathbf{I}_m + \sigma_v^2 \boldsymbol{\Omega}^{-1})\, d\boldsymbol{\phi}. \quad \text{(A.5)}$$

Puisque la ligne $(\mathbf{X}_2) = p$, il s'ensuit immédiatement que la ligne $(\mathbf{G}) = m_1 + p$. Par conséquent, $\mathbf{G}$ est dite « de plein rang–colonne ». Nous indiquons $m_1 + p$ par q. Pour $k = 2, \ldots, 5$, nous établissons maintenant les limites supérieures pour :

$$\left| \delta \mathbf{I}_m + \sigma_v^2 \boldsymbol{\Omega}_k^{-1} \right|^{-1/2} \int \exp\left\{-\frac{1}{2}(\mathbf{d} - \mathbf{G}\boldsymbol{\phi})^\top (\delta \mathbf{I}_m + \sigma_v^2 \boldsymbol{\Omega}_k^{-1})^{-1} (\mathbf{d} - \mathbf{G}\boldsymbol{\phi})\right\} d\boldsymbol{\phi}$$

qui pourront être intégrables relativement à σ_v^2 et ρ comme suit.

A.1 Renseignements sur le modèle autorégressif conditionnel simple (SCAR)

Prenons d'abord le modèle CAR où $k = 2$. Supposons que $\mathbf{P}_\mathbf{W}$ est une matrice orthogonale, de sorte que $\mathbf{P}_\mathbf{W}^\top \mathbf{W} \mathbf{P}_\mathbf{W} = \operatorname{diag}\{\lambda_i\}_{i=1}^m = \boldsymbol{\Lambda}$. Alors $\boldsymbol{\Omega}_2(\rho)^{-1} = \mathbf{P}_\mathbf{W} \{\mathbf{I} - \rho\boldsymbol{\Lambda}\}^{-1} \mathbf{P}_\mathbf{W}^\top$, et donc,

$$(\mathbf{d} - \mathbf{G}\boldsymbol{\phi})^\top (\delta \mathbf{I}_m + \sigma_v^2 \boldsymbol{\Omega}_2^{-1})^{-1} (\mathbf{d} - \mathbf{G}\boldsymbol{\phi}) \;= (\mathbf{P}_\mathbf{W}^\top \mathbf{d} - \mathbf{P}_\mathbf{W}^\top \mathbf{G}\boldsymbol{\phi})^\top \left(\delta \mathbf{I}_m + \sigma_v^2 \{\mathbf{I} - \rho\boldsymbol{\Lambda}\}^{-1}\right)^{-1} (\mathbf{P}_\mathbf{W}^\top \mathbf{d} - \mathbf{P}_\mathbf{W}^\top \mathbf{G}\boldsymbol{\phi})$$

$$= (\mathbf{d}_* - \mathbf{G}_* \boldsymbol{\phi})^\top \left(\delta \mathbf{I}_m + \sigma_v^2 \{\mathbf{I} - \rho\boldsymbol{\Lambda}\}^{-1}\right)^{-1} (\mathbf{d}_* - \mathbf{G}_* \boldsymbol{\phi}),$$

où $\mathbf{d}_* = \mathbf{P}_\mathbf{W}^\top \mathbf{d}$ et $\mathbf{G}_* = \mathbf{P}_\mathbf{W}^\top \mathbf{G}$. Supposons que les lignes de $\mathbf{G}_*$ correspondant à des indices q distincts $\{i_1, \ldots, i_q\} \subseteq \{1, \ldots, m\}$ sont linéairement indépendantes. Nous désignons ces lignes par $\mathbf{g}_{i_k^*}^\top$, $k = 1, \ldots, q$. Supposons que $\mathbf{A}$ est $q \times q$ la matrice non singulaire $[\mathbf{g}_{i_1^*}, \ldots, \mathbf{g}_{i_q^*}]^\top$ et que $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_q)^\top = \mathbf{A}\boldsymbol{\phi}$. Il convient de souligner que

$$(\mathbf{d} - \mathbf{G}\boldsymbol{\phi})^\top (\delta \mathbf{I}_m + \sigma_v^2 \boldsymbol{\Omega}_2^{-1})^{-1} (\mathbf{d} - \mathbf{G}\boldsymbol{\phi}) \geq \sum_{k=1}^{q} \frac{(d_{i_k^*} - \eta_{i_k})^2}{\delta + \sigma_v^2 (1 - \rho \lambda_{i_k})^{-1}}.$$

Ainsi, nous obtenons :

$$\begin{aligned}
\int \exp\left\{-\frac{1}{2}(\mathbf{d}-\mathbf{G}\boldsymbol{\phi})^\top (\delta \mathbf{I}_m + \sigma_v^2 \boldsymbol{\Omega}_2^{-1})^{-1} (\mathbf{d}-\mathbf{G}\boldsymbol{\phi})\right\} d\boldsymbol{\phi} &\leq \int \exp\left\{-\frac{1}{2}\sum_{k=1}^{q} \frac{(d_{i_k^*} - \eta_{i_k})^2}{\delta + \sigma_v^2 (1-\rho\lambda_{i_k})^{-1}}\right\} d\boldsymbol{\phi} \\
&= \int \exp\left\{-\frac{1}{2}\sum_{k=1}^{q} \frac{(d_{i_k^*} - \eta_{i_k})^2}{\delta + \sigma_v^2 (1-\rho\lambda_{i_k})^{-1}}\right\} d\boldsymbol{\eta} \left| \mathbf{A}^\top \mathbf{A} \right|^{-1/2} \\
&= K \prod_{k=1}^{q} \left\{\delta + \sigma_v^2 (1-\rho\lambda_{i_k})^{-1}\right\}^{1/2}, \qquad \text{(A.6)}
\end{aligned}$$

où $K > 0$ est une constante adéquatement finie. De plus, nous savons que

$$\left| \delta \mathbf{I}_m + \sigma_v^2 \boldsymbol{\Omega}_2^{-1} \right|^{-1/2} = \prod_{i=1}^{m} \left\{\delta + \sigma_v^2 (1-\rho\lambda_i)^{-1}\right\}^{-1/2}. \qquad \text{(A.7)}$$

Suivant (A.6) et (A.7) nous obtenons :

$$\begin{aligned}
\left| \delta \mathbf{I}_m + \sigma_v^2 \boldsymbol{\Omega}_2^{-1} \right|^{-1/2} \int \exp\left\{-\frac{1}{2}(\mathbf{d}-\mathbf{G}\boldsymbol{\phi})^\top (\delta \mathbf{I}_m + \sigma_v^2 \boldsymbol{\Omega}_2^{-1})^{-1} (\mathbf{d}-\mathbf{G}\boldsymbol{\phi})\right\} d\boldsymbol{\phi} \\
\leq K \prod_{i \notin \{i_1, \ldots, i_q\}} \left\{\delta + \sigma_v^2 (1-\rho\lambda_i)^{-1}\right\}^{-1/2} \\
\leq K \left\{ I(\sigma_v^2 < N) + (\sigma_v^2)^{-(m-q)/2} \prod_{i \notin \{i_1, \ldots, i_q\}} (1-\rho\lambda_i)^{1/2} I(\sigma_v^2 > N) \right\} \qquad \text{(A.8)}
\end{aligned}$$

pour tout nombre positif N. Il ne faut pas oublier que $\lambda_m^{-1} < \rho < \lambda_1^{-1}$. Nous savons que $1 - \rho\lambda_i$ est une valeur propre de $\boldsymbol{\Omega}_2$. Donc, pour $\lambda_m^{-1} < \rho < \lambda_1^{-1}$, pour $i = 1, \ldots, m$, $1 - \rho\lambda_i > 0$. De plus, $\sum_{i=1}^{m} (1 - \rho\lambda_i) = m$. Cela suppose que $0 < 1 - \rho\lambda_i < m$. En suite, suivant (A.8), nous obtenons :

$$\begin{aligned}
\left| \delta \mathbf{I}_m + \sigma_v^2 \boldsymbol{\Omega}_2^{-1} \right|^{-1/2} \int \exp\left\{-\frac{1}{2}(\mathbf{d}-\mathbf{G}\boldsymbol{\phi})^\top (\delta \mathbf{I}_m + \sigma_v^2 \boldsymbol{\Omega}_2^{-1})^{-1} (\mathbf{d}-\mathbf{G}\boldsymbol{\phi})\right\} d\boldsymbol{\phi} \\
\leq K \left\{ I(\sigma_v^2 < N) + (\sigma_v^2)^{-(m-q)/2} I(\sigma_v^2 > N) \right\}. \qquad \text{(A.9)}
\end{aligned}$$

Suivant (A.5) et (A.9), il s'ensuit que, dans les conditions établies par le théorème, l'intégrale $\int f(\mathbf{y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho)\, d\boldsymbol{\theta} d\boldsymbol{\beta} d\sigma_v^2 d\rho$ désirée est finie.

A.2 Renseignements sur le modèle autorégressif simultané (SAR)

Prenons maintenant $k = 3$ pour le modèle SAR. Suivant $\mathbf{W}_* = \mathbf{L}^{-1/2}\mathbf{W}\mathbf{L}^{-1/2}$, nous avons :

$$\begin{aligned}\boldsymbol{\Omega}_3 &= (\mathbf{I}_m - \rho\tilde{\mathbf{W}})^\top(\mathbf{I}_m - \rho\tilde{\mathbf{W}}) \\ &= (\mathbf{L} - \rho\mathbf{W})^\top \mathbf{L}^{-2}(\mathbf{L} - \rho\mathbf{W}) \\ &= \mathbf{L}^{1/2}(\mathbf{I}_m - \rho\mathbf{W}_*)\,\mathbf{L}^{-1}(\mathbf{I}_m - \rho\mathbf{W}_*)\mathbf{L}^{1/2}.\end{aligned}$$

D'abord, $\operatorname{tr}\boldsymbol{\Omega}_3 = m + \rho^2 \sum_i \sum_j \tilde{w}_{ij}^2 \le m + \rho^2 \sum_i \sum_j \tilde{w}_{ij} = m + \rho^2 m < 2m$ puisque $0 \le \tilde{w}_{ij} \le 1$, $\sum_j \tilde{w}_{ij} = 1$, et $-1 < \rho < 1$.

Il convient de souligner que les valeurs propres $\nu_1, \ldots, \nu_m$ de $\mathbf{W}_*$ sont toutes réelles (car $\mathbf{W}_*$ est symétrique). De plus, $\mathbf{W}_*$ et $\tilde{\mathbf{W}}$ ont des valeurs propres identiques. Étant une matrice stochastique, $\tilde{\mathbf{W}}$ a au moins une valeur propre qui est 1 et les valeurs propres restantes ont une borne supérieure égale à 1, c'est-à-dire que $|\nu_i| \le 1$ et $\max_i \nu_i = 1$. Comme $-1 < \rho < 1$ et $1 - \rho\nu_i > 0$, $|\boldsymbol{\Omega}_3| = \prod_{i=1}^m (1 - \rho\nu_i)^2 > 0$. Ainsi, les valeurs propres de $\boldsymbol{\Omega}_3$ sont positives, et elles ont une borne supérieure égale à $2m$. Supposons que $l_{(1)} = \min l_i$ et $l_{(m)} = \max l_i$, où $\mathbf{L} = \operatorname{diag}\{l_i\}_{i=1}^m$. Alors $l_{(1)} > 0$ et $l_{(m)}$ ont une borne supérieure. Si l'on écrit :

$$\boldsymbol{\Sigma}_3 = \delta\mathbf{I}_m + \sigma_v^2\,\boldsymbol{\Omega}_3^{-1} = \mathbf{L}^{-1/2}\left\{\delta\mathbf{L} + \sigma_v^2(\mathbf{I}_m - \rho\mathbf{W}_*)^{-1}\mathbf{L}(\mathbf{I} - \rho\mathbf{W}_*)^{-1}\right\}\mathbf{L}^{-1/2},$$

nous avons :

$$\begin{aligned}|\boldsymbol{\Sigma}_3| &= |\mathbf{L}|^{-1}\left|\delta\mathbf{L} + \sigma_v^2(\mathbf{I}_m - \rho\mathbf{W}_*)^{-1}\mathbf{L}(\mathbf{I}_m - \rho\mathbf{W}_*)^{-1}\right| \ge |\mathbf{L}|^{-1} l_{(1)}^m \left|\delta\mathbf{I}_m + \sigma_v^2(\mathbf{I}_m - \rho\mathbf{W}_*)^{-2}\right| \\ &= |\mathbf{L}|^{-1} l_{(1)}^m \prod_{i=1}^m \left\{\delta + \sigma_v^2(1 - \rho\nu_i)^{-2}\right\}, \qquad \text{(A.10)}\end{aligned}$$

Supposons que $\mathbf{P}_{\mathbf{W}_*}$ est la matrice des vecteurs propres de $\mathbf{W}_*$, de sorte que $\mathbf{P}_{\mathbf{W}_*}^\top \mathbf{W}_* \mathbf{P}_{\mathbf{W}_*} = \operatorname{diag}\{\nu_i\}_{i=1}^m = \mathbf{N}_*$, nous avons aussi :

$$\begin{aligned}(\mathbf{d} - \mathbf{G}\boldsymbol{\phi})^\top \boldsymbol{\Sigma}_3^{-1}(\mathbf{d} - \mathbf{G}\boldsymbol{\phi}) &= (\mathbf{L}^{1/2}\mathbf{d} - \mathbf{L}^{1/2}\mathbf{G}\boldsymbol{\phi})^\top\left\{\delta\mathbf{L} + \sigma_v^2(\mathbf{I}_m - \rho\mathbf{W}_*)^{-1}\mathbf{L}(\mathbf{I}_m - \rho\mathbf{W}_*)^{-1}\right\}^{-1}(\mathbf{L}^{1/2}\mathbf{d} - \mathbf{L}^{1/2}\mathbf{G}\boldsymbol{\phi}) \\ &= (\mathbf{r} - \mathbf{S}\boldsymbol{\phi})^\top\left\{\delta\mathbf{L} + \sigma_v^2(\mathbf{I}_m - \rho\mathbf{W}_*)^{-1}\mathbf{L}(\mathbf{I}_m - \rho\mathbf{W}_*)^{-1}\right\}^{-1}(\mathbf{r} - \mathbf{S}\boldsymbol{\phi}) \\ &\ge (l_{(m)}^{-1/2}\mathbf{r} - l_{(m)}^{-1/2}\mathbf{S}\boldsymbol{\phi})^\top\left\{\delta\mathbf{I}_m + \sigma_v^2(\mathbf{I}_m - \rho\mathbf{W}_*)^{-2}\right\}^{-1}(l_{(m)}^{-1/2}\mathbf{r} - l_{(m)}^{-1/2}\mathbf{S}\boldsymbol{\phi}) \\ &= (\tilde{\mathbf{r}} - \tilde{\mathbf{S}}\boldsymbol{\phi})^\top\left\{\delta\mathbf{I}_m + \sigma_v^2(\mathbf{I}_m - \rho\mathbf{N}_*)^{-2}\right\}^{-1}(\tilde{\mathbf{r}} - \tilde{\mathbf{S}}\boldsymbol{\phi}) \\ &\ge \sum_{k=1}^q \frac{(\tilde{r}_{i_k} - \tilde{s}_{i_k}^\top\boldsymbol{\phi})^2}{\delta + \sigma_v^2(1 - \rho\nu_{i_k})^{-2}}, \qquad \text{(A.11)}\end{aligned}$$

où $\mathbf{r} = \mathbf{L}^{1/2}\mathbf{d}$, $\mathbf{S} = \mathbf{L}^{1/2}\mathbf{G}$, $\tilde{\mathbf{r}} = l_{(m)}^{-1/2}\mathbf{P}_{\mathbf{W}_*}\mathbf{r}$, $\tilde{\mathbf{S}} = l_{(m)}^{-1/2}\mathbf{P}_{\mathbf{W}_*}\mathbf{S}$, et $\{i_1, \ldots, i_q\}$ forment un sous-ensemble de $\{1, \ldots, m\}$ de telle sorte que la matrice $q \times q$ $[\tilde{\mathbf{s}}_{i_1}, \ldots, \tilde{\mathbf{s}}_{i_q}]^\top = \tilde{\mathbf{S}}_1$, une sous-matrice de $\tilde{\mathbf{S}}$, est non singulaire. Il convient de mentionner que $\tilde{\mathbf{S}}_1$ est déterminé par $\mathbf{W}$. En utilisant (A.11), nous obtenons :

$$\int \exp\left\{-\frac{1}{2}(\mathbf{d}-\mathbf{G}\boldsymbol{\phi})^\top \boldsymbol{\Sigma}_3^{-1}(\mathbf{d}-\mathbf{G}\boldsymbol{\phi})\right\} d\boldsymbol{\phi} \le (2\pi)^{q/2} \left|\tilde{\mathbf{S}}_1^\top \tilde{\mathbf{S}}_1\right|^{-1/2} \prod_{k=1}^{q} \left\{\delta + \sigma_v^2 (1-\rho\nu_{i_k})^{-2}\right\}^{1/2}. \quad \text{(A.12)}$$

Suivant (A.10) et (A.12), nous obtenons :

$$\begin{aligned}\int |\boldsymbol{\Sigma}_3|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{d}-\mathbf{G}\boldsymbol{\phi})^\top \boldsymbol{\Sigma}_3^{-1}(\mathbf{d}-\mathbf{G}\boldsymbol{\phi})\right\} d\boldsymbol{\phi} &\le K \prod_{i \notin \{i_1,\ldots,i_q\}} \left\{\delta + \sigma_v^2 (1-\rho\nu_i)^{-2}\right\}^{-1/2} \\ &\le K\left\{I(\sigma_v^2 < N) + (\sigma_v^2)^{-(m-q)/2} I(\sigma_v^2 > N) \prod_{i \notin \{i_1,\ldots,i_q\}} (1-\rho\nu_i)\right\} \\ &\le K\left\{I(\sigma_v^2 < N) + (\sigma_v^2)^{-(m-q)/2} I(\sigma_v^2 > N)\right\}, \end{aligned} \quad \text{(A.13)}$$

où nous nous appuyons sur le fait que $-1 < \rho < 1$ et que $-1 \le \nu_i \le 1$ pour faire valoir que $0 < 1 - \rho\nu_i < 2$. Suivant (A.5) et (A.13), nous procédons donc comme nous l'avons fait pour le modèle CAR, c'est-à-dire que l'intégrale $\int f(\mathbf{y}_{(2)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \rho)\, d\boldsymbol{\theta} d\boldsymbol{\beta} d\sigma_v^2 d\rho$ désirée est finie dans les conditions du théorème.

A.3 Renseignements sur le modèle autorégressif conditionnel (CAR)

Prenons maintenant $k = 4$ pour le modèle autorégressif intrinsèque (IAR), où :

$$\boldsymbol{\Omega}_4 = \mathbf{L} - \rho\mathbf{W} = \mathbf{L}^{1/2}(\mathbf{I}_m - \rho\mathbf{W}_*)\mathbf{L}^{1/2}.$$

Soit $\boldsymbol{\Sigma}_4 = \delta\mathbf{I}_m + \sigma_v^2 \boldsymbol{\Omega}_4^{-1} = \mathbf{L}^{-1/2}\left\{\delta\mathbf{L} + \sigma_v^2(\mathbf{I}_m - \rho\mathbf{W}_*)^{-1}\right\}\mathbf{L}^{-1/2}$. Alors :

$$|\boldsymbol{\Sigma}_4| \ge |\mathbf{L}|^{-1} k_*^m \prod_{i=1}^{m} \left\{\delta + \sigma_v^2 (1-\rho\nu_i)^{-1}\right\}, \quad \text{(A.14)}$$

où $k_*^m = \min\{l_{(1)}, 1\}$. En procédant de la même façon que dans (A.11), nous obtenons ceci :

$$(\mathbf{d}-\mathbf{G}\boldsymbol{\phi})^\top \boldsymbol{\Sigma}_4^{-1}(\mathbf{d}-\mathbf{G}\boldsymbol{\phi}) \ge \sum_{k=1}^{q} \frac{(\tilde{r}_{i_k} - \tilde{\mathbf{s}}_{i_k}^\top \boldsymbol{\phi})^2}{\delta + \sigma_v^2 (1-\rho\nu_{i_k})^{-1}}. \quad \text{(A.15)}$$

Encore une fois, comme nous l'avons fait pour les deux cas précédents, nous pouvons utiliser (A.14) et (A.15) pour établir que l'intégrale désirée est finie dans les conditions énoncées dans le théorème.

A.4 Renseignements sur le modèle autorégressif conditionnel de Leroux (LCAR)

Enfin, nous considérons $k = 5$, où pour le cas LCAR nous avons :

$$\boldsymbol{\Omega}_5 = \rho\mathbf{R} + (1-\rho)\mathbf{I}_m.$$

Supposons que $r_1, \ldots, r_m$ sont les valeurs propres de $\mathbf{R}$ et que $\mathbf{P}_\mathbf{R}$ est une matrice orthogonale de sorte que $\mathbf{P}_\mathbf{R}^\top \mathbf{R} \mathbf{P}_\mathbf{R} = \text{diag}\{r_i\}_{i=1}^m$. Étant donné que $\mathbf{R}$ est une matrice définie non négative, $r_i \geq 0, \quad i = 1, \ldots, m$ et $\sum_{i=1}^m r_i = \text{tr}\,\mathbf{R} = \sum_{i=1}^m l_i,$ ce qui suppose que $r_1, \ldots, r_m$ sont tous compris entre 0 et $l = \sum_{i=1}^m l_i$. Alors on peut écrire :

$$\mathbf{\Omega}_5 \;=\; \mathbf{P}_\mathbf{R} \left\{ \text{diag}\{\rho r_i + 1 - \rho\}_{i=1}^m \right\} \mathbf{P}_\mathbf{R}^\top,$$

et prétendre que pour $0 < \rho < 1,$ les valeurs propres de $\mathbf{\Omega}_5$ sont toutes positives et possèdent une borne supérieure égale à $\sum_{i=1}^m r_i + 1 = l + 1.$ Ensuite, étant donné $\tilde{\mathbf{r}} = \mathbf{P}_\mathbf{R}^\top \mathbf{d}$ et $\tilde{\mathbf{S}} = \mathbf{P}_\mathbf{R}^\top \mathbf{G},$ nous pouvons établir une inégalité similaire à (A.11). Il convient de mentionner que la matrice non singulière $\tilde{\mathbf{S}}_1$ est une sous-matrice de $\tilde{\mathbf{S}}$ et est exempte de ρ. Le caractère limitatif des valeurs propres de $\mathbf{\Omega}_5$ entraînera une inégalité semblable à (A.14). Enfin, nous obtenons l'intégrale désirée qui est finie dans les conditions du théorème.

Bibliographie

Banerjee, S., Carlin, B.P., et Gelfand, A.E. (2003). *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall/CRC.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, Springer Science & Business Media.

Besag, J., et Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82, 733-746.

Brown, D.A., Datta, G.S. et Lazar, N.A. (2017). A Bayesian generalized CAR model for correlated signal detection. *Statistica Sinica*, 27, 1125-1153.

Datta, G.S., et Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10, 613-627.

Datta, G.S., et Smith, D.D. (2003). On propriety of posterior distributions of variance components in small area estimation. *Journal of Statistical Planning and Inference*, 112, 175-183.

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, 362-374.

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, 183-196.

Datta, G.S., Lahiri, P., Maiti, T. et Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.

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, 269-277.

Gelman, A., et Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. et Rubin, D.B. (2013). *Bayesian Data Analysis*, CRC Press, 3rd ed.

Ghosh, M. (1992). Hierarchical and empirical Bayes multivariate estimation. Dans *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, (Éds., M. Ghosh et P.K. Pathak), Institute of Mathematical Statistics, 151-177.

Hodges, J.S. (2019). *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects*, Chapman and Hall/CRC.

Leroux, B.G., Lei, X. et Breslow, N. (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence. Dans *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, Springer, 179-191.

MacNab, Y.C. (2003). Hierarchical Bayesian spatial modelling of small-area rates of non-rare disease. *Statistics in Medicine*, 22, 1761-1773.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. et Breidt, F. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 265-286.

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, 15-29.

Porter, A.T., Holan, S.H., Wikle, C.K. et Cressie, N. (2014). Spatial Fay-Herriot models for small area estimation with functional covariates. *Spatial Statistics*, 10, 27-42.

Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

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

Rao, J.N.K., Sinha, S.K. et Dumitrescu, L. (2014). Robust small area estimation under semi-parametric mixed models. *Canadian Journal of Statistics*, 42, 126-141.

Speckman, P.L., et Sun, D. (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika*, 90, 289-302.

Stan Development Team (2018). `RStan`: The R interface to Stan. R package version 2.17.3.

Sun, D., Tsutakawa, R.K. et Speckman, P.L. (1999). Posterior distribution of hierarchical models using CAR (1) distributions. *Biometrika*, 86, 341-350.

Torabi, M. (2012). Hierarchical Bayes estimation of spatial statistics for rates. *Journal of Statistical Planning and Inference*, 142, 358-365.

Trevisani, M., et Gelfand, A. (2013). Spatial misalignment models for small area estimation: A simulation study. Dans *Advances in Theoretical and Applied Statistics*, Springer, 269-279.

Watanabe, S., et Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41, 434-449.

You, Y., et Zhou, Q.M. (2011). Estimation sur petits domaines hiérarchique bayésienne sous un modèle spatial avec application à des données d'enquête sur la santé. *Techniques d'enquête*, 37, 1, 31-44. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2011001/article/11445-fra.pdf.

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 2022.

· C. Bayart, *University Claude Bernard Lyon 1*
· A. Béliveau, *University of Waterloo*
· D. Bonnery, *University of Cambridge*
· H.J. Boonstra, *Statistics Netherlands*
· J. Breidt, *NORC at the University of Chicago*
· J.M. Brick, *Westat Inc.*
· J. Brown, *University of Technology Sydney*
· E. Bueno, *Stockholm University*
· S. Cai, *Carleton University*
· P.J. Cantwell, *U.S. Census Bureau*
· G. Chauvet, *École nationale de la statistique et de l'analyse de l'information*
· Q. Chen, *Columbia University*
· S. Chen, *University of Oklahoma Health Sciences Center*
· Y. Cheng, *USDA National Agricultural Statistics Service*
· J. Chipperfield, *Australian Bureau of Statistics*
· S. Cho, *Iowa State University*
· R. Clark, *Australian National University*
· M. Cohen, *American Institutes for Research*
· M. Couper, *University of Michigan*
· M. Dagdoug, *Laboratoire de Mathématiques de Besançon*
· G. Datta, *University of Georgia*
· T. De Waal, *Centraal Bureau voor de Statistiek*
· J. Dever, *RTI International*
· M. Elliott, *University of Michigan*
· J.L. Eltinge, *U.S. Bureau of Labor Statistics*
· V. Estevao, *Statistique Canada*
· K. Foster, *East Tennesse State University*
· W.A. Fuller, *Iowa State University*
· K.C.M. Gonçalves, *Universidade Federal do Rio de Janeiro*
· W. Gross, *University of Wollongong*
· P. Han, *University of Michigan*
· S. Hasslet, *Australian National University*
· D. Haziza, *University of Ottawa*
· D. Hedlin, *Stockholm University*
· M.A. Hidiroglou, *Statistique Canada*
· B. Hulliger, *University of Applied and Arts Sciences Northwestern, Switzerland*
· M. Hyllienmark, *Stockholms universitet Samhallsvetenskapliga Fakulteten*
· D. Judkins, *ABT Associates Inc Bethesda*
· M.G.M. Khan, *University of the South Pacific*
· J.K. Kim, *Iowa State University*
· P.S. Kott, *RTI International*
· T. Kubokawa, *The University of Tokyo*
· P. Lahiri, *University of Maryland*
· R. Lehtonen, *University of Helsinki*
· É. Lesage, *L'Institut national de la statistique et des études économiques*
· A. Matei, *Université de Neuchâtel*
· K. McConville, *Reed College*
· F. Mecatti, *University of Milan-Bicocca*
· I. Molina, *Universidad Complutense de Madrid*
· B. Nandram, *Worcester Polytechnic Institute*
· T. Nayak, *The George Washington University*
· J. Opsomer, *Westat Inc*
· D. Pfeffermann, *University of Southampton*
· J.N.K. Rao, *Carleton University*
· L.-P. Rivest, *Université Laval*
· L. Rizzo, *WESTAT*
· F.J. Scheuren, *National Opinion Research Center*
· Z. Seeskin, *NORC*
· P.L.d.N. Silva, *Escola Nacional de Ciências Estatísticas*
· P. Smith, *University of Southampton*
· D. Steel, *University of Wollongong*
· L. Stokes, *Southern Methodist University Dedman College of Humanities and Sciences*
· S. Sugasawa, *The University of Tokyo*
· N. Thuan, *Oregon Health and Science University*
· Y. Tillé, *University of Neuchatel Faculty of Sciences*
· M. Torabi, *University of Manitoba*
· D. Toth, *U.S. Bureau of Labor Statistics*
· R. Valliant, *University of Michigan*
· J. van den Brakel, *Statistics Netherlands*
· F. Verret, *Statistique Canada*
· J. Wagner, *University of Michigan*
· L. Wang, *National Cancer Institute*
· H. Wang, *University of Connecticut*
· M. Witt, *Abt Associates*
· C. Wu, *University of Waterloo*
· C. Xu, *The University of Oklahoma Health Sciences Center*
· D. Yang, *US Bureau of Labor Statistics*
· C. Yu, *Iowa State University*
· W. Yung, *Statistique Canada*
· L.-C. Zhang, *University of Southampton*
· P. Zhao, *Yunnan University*

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2022 : Céline Ethier du Centre de collaboration internationale et d'innovation en méthodologie; Patrick O'Leary de la Division des méthodes de la statistique sociale; l'équipe de la Division de la diffusion, en particulier Sophie Pilon, Eric Chénier, Camil Séguin, Kathy Charbonneau, Travis Robinson et Karo-Lynn Audy et de même que nos partenaires de la Division des communications.

ANNONCES

Demande de candidatures pour le prix Waksberg 2024

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 2024 au Symposium de Statistique Canada prévu à l'automne 2024. L'article paraîtra dans un numéro futur de *Techniques d'enquête* (prévu pour décembre 2024).

L'auteur de l'article Waksberg 2024 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 2023 à la présidente du comité, Maria Giovanna Ranalli, (maria.ranalli@unipg.it). 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 (2022-2023)

Maria Giovanna Ranalli, *University of Perugia* (Présidente)
Denise Silva, *Brazilian Institute of Geography and Statistics*
Jae-Kwang Kim, *Iowa State University*
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)

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 38, No. 2, June 2022

In Memory of Dr. Lars Lyberg Remembering a Giant in Survey Research 1944–2021 353

Spatial Sampling Design to Improve the Efficiency of the Estimation of the Critical Parameters of the SARS-CoV-2 Epidemic
Giorgio Alleva, Giuseppe Arbia, Piero Demetrio Falorsi, Vincenzo Nardelli and Alberto Zuliani 367

Assessing Residual Seasonality in the U.S. National Income and Product Accounts Aggregates
Baoline Chen, Tucker S. McElroy and Osbert C. Pang 399

Improved Assessment of the Accuracy of Record Linkage via an Extended MaCSim Approach
Shovanur Haque and Kerrie Mengersen 429

If They Don't Understand the Question, They Don't answer. Language Mismatch in Face-to-Face Interviews
Jannes Jacobsen 453

Improving the Output Quality of Official Statistics Based on Machine Learning Algorithms
Q.A. Meertens, C.G.H. Diks, H.J. van den Herik and F.W. Takes 485

Data Fusion for Joining Income and Consumtion Information using Different Donor-Recipient Distance Metrics
Florian Meinfelder and Jannik Schaller 509

Total Process Error: An Approach for Assessing and Monitoring the Quality of Multisource Processes
Fabiana Rocci, Roberta Varriale and Orietta Luzi 533

Some Thoughts on Official Statistics and its Future (with discussion)
Yves Tillé, Marc Debusschere, Henri Luomaranta, Martin Axelson, Eva Elvers, Anders Holmberg and Richard Valliant 557

Iterative Kernel Density Estimation Applied to Grouped Data: Estimating Poverty and Inequality Indicators from the German Microcensus
Paul Walter, Marcus Groß, Timo Schmid and Katja Weimer 599

Data Collection Expert Prior Elicitation in Survey Design: Two Case Studies
Shiya Wu, Barry Schouten, Ralph Meijers and Mirjam Moerbeek 637

Rejoinder: Measuring Inflation under Pandemic Conditions
W. Erwin Diewert and Kevin J. Fox 663

Book Review
Ann-Marie Flygare and Ingegerd Jansson 669

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 38, No. 3, September 2022

Construction of Databases for Small Area Estimation
Emily Berg 673

Hierarchical Bayesian Model with Inequality Constraints for US County Estimates
Lu Chen, Balgobin Nandram and Nathan B. Cruze 709

Timely Estimates of the Monthly Mexican Economic Activity
Francisco Corona, Graciela González-Farías and Jesús López-Pérez 733

Small Domain Estimation of Census Coverage – A Case Study in Bayesian Analysis of Complex Survey Data
Joane S. Elleouet, Patrick Graham, Nikolai Kondratev, Abby K. Morgan and Rebecca M. Green 767

Identifying Data Quality Challenges in Online Opt-In Panels Using Cognitive Interviews in English and Spanish
Yazmín García Trejo, Mikelyn Meyers, Mandi Martinez, Angela O'Brien, Patricia Goerman and Betsarí Otero Class 793

Measuring and Mapping Micro Level Earning Inequality towards Addressing the Sustainable Development Goals – A Multivariate Small Area Modelling Approach
Saurav Guha and Hukum Chandra 823

Economic Nowcasting with Long Short-Term Memory Artificial Neural Networks (LSTM)
Daniel Hopp 847

Variable Inclusion Strategies through Directed Acyclic Graphs to adjust Health Surveys subject to Selection Bias for Producing National Estimates
Yan Li, Katherine E. Irimata, Yulei He and Jennifer Parker 875

Pseudo Bayesian Mixed Models under Informative Sampling
Terrance D. Savitsky and Matthew R. Williams 901

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 50, No. 1, March/mars 2022

Special Issue

Functional and object data analysis/L'analyse de données fonctionnelles et d'objets

Issue Information 1

Editorial
Johanna G. Nešlehová 5

Introduction

Special issue on "Functional and object data analysis": Guest Editor's introduction
Hans-Georg Müller 8

Research Articles

Bayesian spline smoothing with ambiguous penalties
Xinlian Zhang, Gauri S. Datta, Ping Ma, Wenxuan Zhong 20

Direct local linear estimation for Sharpe ratio function
Hongmei Lin, Tiejun Tong, Yuedong Wang, Wenchao Xu, Riquan Zhang 36

A nonlinear sparse neural ordinary differential equation model for multiple functional processes
Yijia Liu, Lexin Li, Xiao Wang 59

Adaptive estimation for functional data: Using a framelet block-thresholding method
Kun Cheng, Di-Rong Chen 86

Recovering the underlying trajectory from sparse and irregular longitudinal data
Yunlong Nie, Yuping Yang, Liangliang Wang, Jiguo Cao 122

Variable selection in nonparametric functional concurrent regression
Rahul Ghosal, Arnab Maity 142

Robust estimation and variable selection for function-on-scalar regression
Xiong Cai, Liugen Xue, Jiguo Cao, for the Alzheimer's Disease Neuroimaging Initiative 162

Simultaneous variable selection, clustering, and smoothing in function-on-scalar regression
Suchit Mehrotra, Arnab Maity 180

Continuum centroid classifier for functional data
Zhiyang Zhou, Peijun Sang 200

Cluster analysis with regression of non-Gaussian functional data on covariates
Jiakun Jiang, Huazhen Lin, Heng Peng, Gang-Zhi Fan, Yi Li 221

Reproducing kernel-based functional linear expectile regression
Meichen Liu, Matthew Pietrosanu, Peng Liu, Bei Jiang, Xingcai Zhou, Linglong Kong, for the Alzheimer's Disease Neuroimaging Initiative 241

Extreme quantile estimation for partial functional linear regression models with heavy-tailed distributions
Hanbing Zhu, Yehua Li, Baisen Liu, Weixin Yao, Riquan Zhang 267

Doubly stochastic models for spatio-temporal covariation of replicated point processes
Daniel Gervini 287

Testing normality of spatially indexed functional data
Siegfried Hörmann, Piotr Kokoszka, Thomas Kuenzer 304

Modelling multi-scale, state-switching functional data with hidden Markov models
Evan Sidrow, Nancy Heckman, Sarah M. E. Fortune, Andrew W. Trites, Ian Murphy, Marie Auger-Méthé 327

Not all long-memory estimators are born equal: The case of nonstationary functional time series
Han Lin Shang 357

Erratum

Authors' corrigenda/corrections des auteurs on synthetic data method to incorporate external information into a current study
Tian Gu, Jeremy M. G. Taylor, Wenting Cheng, Bhramar Mukherjee 381

Acknowledgement

Acknowledgement of Referees' Services Remerciements aux lecteurs critiques 385

CONTENTS TABLE DES MATIÈRES

Volume 50, No. 2, June/juin 2022

Issue Information 391

Research Articles

Characterizing the COVID-19 dynamics with a new epidemic model: Susceptible-exposed-asymptomatic-symptomatic-active-removed
Grace Y. Yi, Pingbo Hu, Wenqing He 395

Estimating design operating characteristics in Bayesian adaptive clinical trials
Shirin Golchi 417

The linear Lasso: A location model approach
D.A.S. Fraser, Mylène Bédard 437

Model selection properties of forward selection and sequential cross-validation for high-dimensional regression
Jerzy Wieczorek, Jing Lei 454

Recovery of sums of sparse and dense signals by incorporating graphical structure among predictors
Yiyun Luo, Yufeng Liu, for the Alzheimer's Disease Neuroimaging Initiative 471

Avoiding prior–data conflict in regression models via mixture priors
Leonardo Egidi, Francesco Pauli, Nicola Torelli 491

Additive hazard regression of event history studies with intermittently measured covariates
Xiaowei Sun, Xinyuan Song, Liuquan Sun 511

Dynamic estimation with random forests for discrete-time survival data
Hoora Moradian, Weichi Yao, Denis Larocque, Jeffrey S. Simonoff, Halina Frydman 533

Semiparametric additive frailty hazard model for clustered failure time data
Peng Liu, Shanshan Song, Yong Zhou 549

Assessing the calibration of subdistribution hazard models in discrete time
Moritz Berger, Matthias Schmid 572

Synthetic estimation for the complier average causal effect
Denis Agniel, Matthew Cefalu, Bing Han 592

A modified expectation-maximization algorithm for latent Gaussian graphical model
Chaowen Zheng, Jingfang Huang, Ian A. Wood, Yichao Wu 612

Nonparametric estimation of unrestricted distributions and their jumps
Kairat Mynbaev, Carlos Martins-Filho, Daniel J. Henderson 638

On the spectral coherence between two periodically correlated processes
Mahnaz Khalafi, Ahmad Reza Soltani, Masoud Golalipour, Majid Azimmohseni, Farzad Najafiamiri 663

The estimation of the fractal dimension of a Gaussian field via Euler characteristic
Ali Reza Taheriyoun, Khalil Shafie 681

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.